ROBUST BAYESIAN ANALYSIS, AN ATTEMPT TO IMPROVE BAYESIAN SEQUENCING

Franz Weninger[1] • Peter Steier • Walter Kutschera • Eva Maria Wild

Vienna Environmental Research Accelerator (VERA), Faculty of Physics, Isotope Research, University of Vienna, Währinger Straße 17, A-1090 Vienna, Austria

[1]Corresponding author. Email: franz.weninger@univie.ac.at

ABSTRACT. Bayesian sequencing of radiocarbon dates deals with the problem that in most cases there does not exist an unambiguous way to define the so-called prior function, which represents information in addition to the result of the radiocarbon measurements alone. However, a random choice of a particular prior function can lead to biased results. In this paper 'robust Bayesian analysis', which uses a whole set of prior functions, is introduced as a more reliable method. The most important aspects of the mathematical foundation and of the practical realization of the method are described. As a general result, robust Bayesian analysis leads to a higher accuracy, however paid for with a reduced precision. Our investigations indicate that it seems possible to establish robust analysis for practical applications.

INTRODUCTION

Motivation

Bayesian sequencing has become a generally accepted and very successful tool to reduce the uncertainty of calibrated radiocarbon ages, when additional information (the so called prior information) about the temporal relationship within a series of radiocarbon dates from an archaeological stratigraphy exist. However, there is a well-known fundamental problem in applying Bayesian statistics: The prior information, which is the archaeological evidence in this application, is often not sufficient to unambiguously define a prior function. Nevertheless, in the standard approach, a specific prior function is selected in a canonical way. The gaps in the prior knowledge are filled with assumed information, which is intended to have no influence on the final result. This may, however, not necessarily be the case. A possible way to overcome this uncertainty is the use of various prior functions that are all in agreement with the archaeological information, and then unify the results; this is called 'robust Bayesian analysis'.

After introducing the basics of Bayesian sequencing below, the idea of robust analysis will be described in detail. However, the actual realization of the method is not so simple as the principle idea suggests. Two main problems will be discussed in the paper: The need of discarding corrupt priors and how to deal with an infinite set of prior functions. A comparison of robust analysis with the usual sequencing is performed with the help of two specific examples.

A remark to the used notation

The basic method of Bayesian sequencing will be illustrated by the help of three different probability densities, denoted as prior function, likelihood function and posterior function, as described in detail below. In full notation these densities can be distinguished by their arguments, because the prior function is a probability density of the real ages $p(t_1,...,t_n)$, the likelihood function is a conditional density of the radiocarbon ages (denoted $x_1,...,x_n$) at given real ages $p(x_1,...,x_n \mid t_1,...,t_n)$ and the posterior function is a conditional density of the real ages at given radiocarbon ages $p(t_1,...,t_n \mid x_1,...,x_n)$. However, for clearness sometime it is useful to simplify the arguments or omit them completely. To distinguish the different densities independently of their arguments we use the letter a to denote the prior function (from 'a priori', which is the origin of 'prior'), *l* for the likelihood function and *p* for the posterior function.

Basic description of Bayesian sequencing

The basic single sample calibration in radiocarbon dating has a serious drawback. Depending on the shape of the calibration curve (e.g. for sections with large wiggles) the procedure may produce calibrated ages with high uncertainty. Fortunately in many cases a series of samples with various known relations between their ages deduced from the excavation site is available. Including this archaeological information within the calibration process by means of Bayesian statistics can improve the accuracy of the resulting ages. The procedure of Bayesian sequencing (or 'multi sample calibration' more generally spoken) is given in the following in its basic form. (Since the calculations are performed for the particular determined set of radiocarbon ages, a simplified notation without indicating the radiocarbon ages is used.) The method was introduced by Buck et al. (1991 and 1992) and is described in detail in Buck et al. (1996); a description focusing on the a practical realization is given by Bronk Ramsey (2009). Weninger et al. (2006) may be an adequate basic description for readers that are not familiar with the field.

For a given set of samples, first - as in single sample calibration - each determined radiocarbon age is transformed into a probability density distribution for the real sample age. This distribution is called single sample likelihood function and characterizes the probability distribution for the real sample age, based only on the information from the measurements so far. (Exactly spoken this is only true under the assumption that any real age is previously equally probable. Actually, as mentioned above, the likelihood function is the conditional probability density to obtain a particular radiocarbon age if a sample of given real age is measured.) In the most simplified form (neglecting the error of the calibration curve and without taking account of standardization) the single sample likelihood function l_i is given by equation 1:

$$l_i(t) \propto \exp\left(-\frac{(x_i - c(t))^2}{2 \cdot \sigma_i^2}\right) \qquad (1)$$

Where x_i is the determined radiocarbon age of the *i*-th sample and σ_i its uncertainty; *t* is the unknown real age of the sample; $c(t)$ gives the calibration curve.
The single sample likelihood functions are combined to a multi-dimensional likelihood function $l(t_1,...,t_n)$ that gives the probability density for any particular combination of the real sample ages $t_1,...,t_n$, still based on the measurements only; see equation 2

$$l(t_1, \dots, t_n) \propto l_1(t_1) \cdot \dots \cdot l_n(t_n) \qquad (2)$$

The available 'a priori' information from archaeology is introduced as a further n-dimensional function, the prior function $a(t_1,...,t_n)$, which is the probability density considering exclusively the archaeological information now.
The product of both gives the so called (multi-dimensional) posterior function $p(t_1,...,t_n)$ (see equation 3) which finally gives the probability density for any particular sample age combination, now considering both the information from the radiocarbon measurement and the archaeological information (deduced e.g. from stratigraphy).

$$p(t_1, \dots, t_n) \propto l(t_1, \dots, t_n) \cdot a(t_1, \dots, t_n) \qquad (3)$$

The probability distributions of the individual real sample ages, the so called marginal posterior densities $p_i(t_i)$, are calculated by projecting the posterior probability to the individual sample co-ordinates, by integration over all other co-ordinates (excluding that of the sample actually evaluated), as shown in equation 4.

$$p_i\,(t_i) \;\propto\; \int\limits_{-\infty}^{+\infty}\!\!\dots\!\int\limits_{-\infty}^{+\infty} p\,(t_1\,,\dots,t_n)\;\;\mathrm{d}t_1\dots\mathrm{d}t_{i-1}\;\mathrm{d}t_{i+1}\dots\mathrm{d}t_n \qquad (4)$$

These resulting probability distributions can be, similar to the case of single-sample calibration, reduced to highest posterior density ranges (hpd-ranges) by collecting the most probable years until the required confidence level (e.g. 95%) is reached.

Equations 1 to 4 are sufficient to perform Bayesian sequencing in its straightforward form. Although the numerical integration imposes some difficulties, because there are too many points in the multi-dimensional space even for a very low resolution grid, it can be performed well with the help of Monte Carlo methods.

Frequently additional statistical parameters beyond the sample ages t_i are used, mainly boundaries of archaeological phases. This can be implemented without difficulties, but will not be shown in this brief introduction.

To perform our investigations we developed a program based on the mathematical language Matlab (by The MathWorks, Inc., Natick, Massachusetts), that runs the calculations by Gibbs sampling, which is a very basic Markov Chain Monte Carlo method; see e.g. Gilks et al. (1996) or Krause (1994).

The simplified notation used here to demonstrate the method most clearly somewhat hides the Bayes theorem on which the method is based. The Bayes theorem - expressed above by equation 3 - looks as follows when given in exact notation (equation 5):

$$p\,(t_1\,,\dots,t_n \,|\, x_1\,,\dots,x_n) \;=\; \frac{l\,(x_1\,,\dots,x_n \,|\, t_1\,,\dots,t_n)\cdot a\,(t_1\,,\dots,t_n)}{\int\limits_{-\infty}^{+\infty}\!\!\dots\!\int\limits_{-\infty}^{+\infty} l\,(x_1\,,\dots,x_n \,|\, t_1\,,\dots,t_n)\cdot a\,(t_1\,,\dots,t_n)\;\;\mathrm{d}t_1\dots\mathrm{d}t_n} \qquad (5)$$

In this notation one can see that the likelihood function, which is the conditional probability density to get the set of radiocarbon ages for a given set of real ages, is converted into a conditional probability density for a set of real ages for given radiocarbon ages - what we are looking for - with the help of the prior probability density. Detailed descriptions of the Bayes theorem and statistical foundations can be found e.g. in Jeffreys (1961) (theoretical background) and in Sivia (1996) (easier to understand).

THE AMBIGUITY PROBLEM AND THE BASIC IDEA OF 'ROBUST BAYESIAN ANALYSIS'

A critical step within the procedure of Bayesian multi sample calibration is the transformation of the archaeological information into a prior function, which has to be a probability density in the real age space $a(t_1,\dots,t_n)$. 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 stratigraphy can be described by various differently shaped prior functions. A simple example can illustrate the nature of the problem: The knowledge that e.g. sample B is older than sample A is described correctly by both functions displayed in figure 1, which show the probability densities for the age difference age(B) minus age(A).

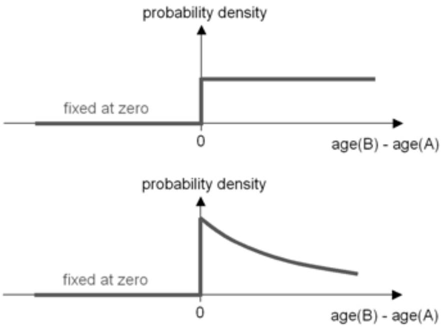

Figure 1 Both prior functions are in full agreement with the prior information that sample B is older then sample A. There is no way to prefer one of them based only on this information.

One knows that age B must not be younger than age A, so the prior function has to be zero on the negative left part of the axis. But any shape of the function one chooses on the positive right side defines a particular probability density for any given age difference. This choice can not be based on available information, because it is only known that sample B is older than A and nothing more. However, the procedure of Bayesian sequencing as given above requires the use of one particular prior function. That means, that Bayesian modeling has to assume information that is actually without foundation. The problem is discussed e.g. by Steier and Rom (2000), Steier et al. (2001) and Bronk Ramsey (2000).

It should be briefly mentioned here, that there are methods to find priors that bias the result as little as possible by reducing their unwanted information content. These priors are then called 'non-informative' priors. One way to minimize unwanted information within a prior function is to characterize its non-informative behavior by an entropy measure (principle of maximum entropy; see e.g. Sivia, 1996). However, we do not follow this approach in the present work.

A very general approach to deal with the ambiguity problem is the so-called robust Bayesian analysis. Robustness in Bayesian analysis has become a theoretically complex field. The theory (summarized e.g. by Berger, 1994 or Rios Insua and Ruggeri, 2000) is not finally settled, as there are ongoing developments and discussions, see e.g. Berger (2006). So we do not try to analyze the problem in a theoretical way, but only introduce the following particular concept to the archaeological application.

The basic idea is to use a (theoretically infinite) set of prior functions including all possible shapes consistent with the available information (figure 2). The Bayesian calculations are 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 (hpd-ranges) to a chosen confidence level. 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. The meaning of these unified hpd-ranges is then a bit different from the usual hpd-ranges: The statement 'the age falls into the interval with a probability of 95 %' is changed to 'the age falls into the interval with a probability of at least 95 %'. Of course, the first statement for the usual interval is only true if the correct prior was used; the second statement for the interval from robust analysis is true if the correct prior was among the used set.

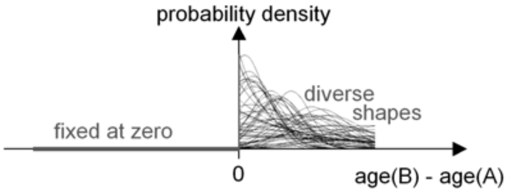

Figure 2 The principle idea of 'robust Bayesian analysis': Using all possible shapes of priors that are consistent with the available information: sample B is older than sample A.

Although robust analysis is very simple in principle, there are some difficulties in the mathematical concept and also in the numerical methods that will be discussed in this work. Before entering into methodical and technical details, a comparison between usual sequencing and robust analysis by the use of a simple example shall clarify the reason why there is a need for improvement.

AN ILLUSTRATIVE EXAMPLE: DATING THE ICEMAN AND HIS AXE

In the year 1991 the famous, very well preserved body of an Early Bronze Age man was found in the European Alps on the Italian side near the Austrian-Italian border in the Ötztaler Alps, released by a melting ice shield. Samples from the body were radiocarbon dated at the AMS laboratories in Oxford and Zürich; see Hedges et al. (1992), Bonani et al. (1994). The actual value used in our work (see figure 3) is taken from Kutschera and Rom (2000), which is a combined age based on those data including tissue as well as bone samples. (A possible effect of inbuilt age in the bone samples is neglected for this example.) In the surrounding of the iceman various parts of his equipment were found and many samples were dated at the AMS lab in Vienna; see Rom et al. (1999); Kutschera and Müller (2003).
To demonstrate the differences of usual sequencing and robust analysis, the direct dates of the iceman and the dates of the wooden shaft of his axe are evaluated with a simple model. The determined mean radiocarbon dates and their un-modeled calibration are shown in figure 3.

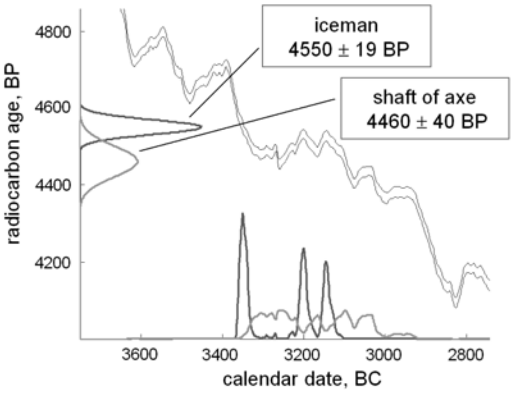

Figure 3 Mean radiocarbon ages and the corresponding un-modeled calibrations of the iceman's body and of the wooden shaft of the axe. The calibration curve (Reimer et al., 2004) is given with two lines indicating the one-sigma precision band.

If one accepts that the axe found close to the iceman was actually used by him, then the wood of the axe shaft cannot be younger than the iceman himself. Maybe one could get more detailed information by further archaeological analysis, which could then be used to construct an improved prior function: e.g. that an axe shaft typically may not be produced of very young wood, or similar considerations. But as such investigations have not been carried out by a qualified expert in this case, the only reliable prior information is the fact that the shaft has to be older than or of equal age as the iceman.

Uniform prior:
The simplest way to formulate the knowledge from above in functional form is the so-called uniform prior, which is zero for age combinations with wrong chronological order, and constant for all allowed cases:

$$a_{uniform}(t_{man}, t_{axe}) \propto \left\{ \begin{matrix} 1 \ \ if \ \ t_{axe} \geq t_{man} \\ 0 \ \ if \ \ t_{axe} < t_{man} \end{matrix} \right\}$$

It can be shown, that in this case where only two ages are involved the uniform prior implies also equal initial probabilities for all age difference t_{axe}-t_{man}, where t_{axe}-$t_{man} \geq 0$ (correct chronological order). Assuming equal probability for any age difference seems reasonable (except for very large age differences) as there is no knowledge whether the shaft is made from young or old wood. However, there are theoretical aspects to favor decreasing functions (exponential, 1/*x*) as non-informative priors for positive definite numbers (see e.g. Jeffreys, 1961).

Uniform span prior:

An alternative possible prior function results from the assumption that the ages of man and axe lie within a particular possible time span with unknown length. This time span is modeled by two outer boundaries (t_{old} and t_{young}) and the prior function is constructed in such a way, that any age difference between these boundaries has the same initial probability. This prior function, denoted as uniform span prior in the following, is mentioned because in a generalized form it is the commonly used prior for conventional archaeological sequences (see Bronk Ramsey, 2001). It can be shown that its functional form is this:

$$a_{unif.span}(t_{man}, t_{axe}, t_{old}, t_{young}) \propto \left\{\begin{matrix} 1/(t_{old}-t_{young})^2 & \text{if} \quad t_{old} \geq t_{axe} \geq t_{man} \geq t_{young} \\ 0 & \text{otherwise} \end{matrix}\right\}$$

Prior set:

Our preferred approach is robust analysis. The finite prior set used as approximation for the infinite set of all possible functions is of the following simple form:

$$a_{robust,\alpha}(t_{man}, t_{axe}) \propto \left\{\begin{matrix} e^{(t_{axe}-t_{man})/\alpha} & \text{if} \quad t_{axe} \geq t_{man} \\ 0 & \text{otherwise} \end{matrix}\right\}$$

Where the parameter α steps through negative values simulating probabilities decreasing with increasing age difference with various slopes, and also through positive values simulating increasing probabilities. The uniform prior ($\alpha = \infty$) is also included.

Figure 4 gives a comparison of the outcome using the two different single priors with that of robust analysis. Always the hpd-ranges at 95.4% confidence level are shown.

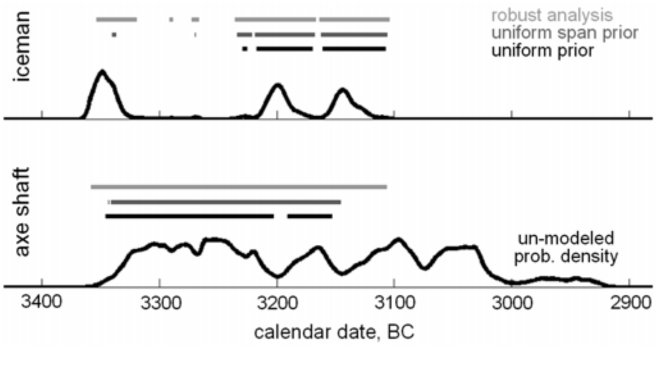

Figure 4 Comparison of the hpd-ranges at a 95.4% confidence level resulting from two different single prior models (uniform prior and uniform span prior) and from robust Bayesian analysis. The un-modeled calibrations are given as probability density distributions.

The difference between robust Bayesian analysis and the use of a single prior can be seen most clearly when we consider the possibility that the axe shaft could be of the same age or only a few years older than the iceman. (This is included in our prior set for robust analysis by priors with $\alpha \approx$ 10a or less.) In that case iceman ages at the oldest peak of the single sample calibration of the iceman are possible, because the un-modeled probability densities overlap considerably. In full accordance with this fact, robust analysis includes the oldest peak of the iceman's probability density within the 95.4% hpd-range. In contrast, when using the uniform prior, iceman ages at the oldest peak are misleadingly excluded, although the uniform prior seems a reasonable choice for this case as mentioned above. The result with the uniform span prior is a bit closer to that of robust analysis; the oldest peak is not excluded totally as with the uniform prior. However, the uniform span prior is still only one particular choice out of an infinite number of possible shapes, and so it can still exclude possible age bands, as indicated by the comparison with the result of robust analysis.

Looking at this different results, one could get the impression that the results of Bayesian sequencing are in some way arbitrary and depend only on the choice of the prior. This was a reason for us to introduce robust analysis. But on the other hand, figure 4 also shows results which are independent from the choice of the prior: ages for the axe shaft younger than about 3100 BC are not possible. Robust analysis preserves this common improvement of all Bayesian methods over single sample calibration.

Remark: 'hpd-range envelopes' instead of probability densities

The final result of robust Bayesian analysis is only available in the form of highest posterior density ranges (hpd-ranges). There are no resulting marginal posterior probability densities calculated, only the hpd-ranges for the various different priors are unified. However, density curves bear valuable information, and they are appreciated by the users of the method. We therefore have developed analogues curves for robust analysis: The unified hpd-ranges can be calculated for all confidence levels. Enveloping all these hpd-ranges results in a continuous function again, that offers information comparable with the marginal posterior density. This 'hpd-range envelopes' are shown in figure 5 for the current example.

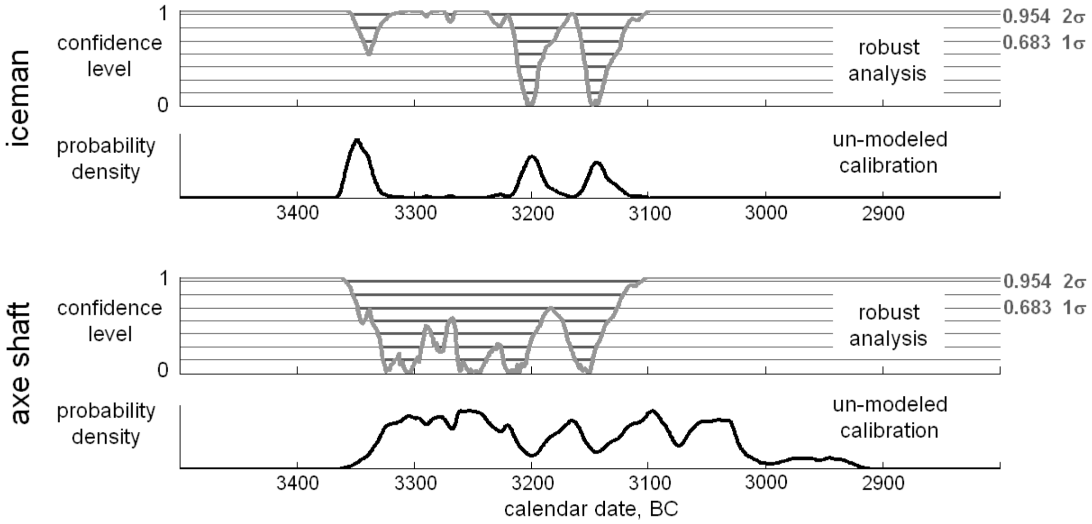

Figure 5 Resulting 'hpd-range envelopes' from robust analysis of the iceman example. These curves (light gray) envelope the hpd-ranges at all different confidence levels. For clarification hpd-ranges to some particular levels (thin lines) are also given (thick lines). The black curves give again the un-modeled probability densities as in figure 4. To avoid misunderstandings: The 'hpd-range envelopes' are not probability density distributions, they only show the limits of the hpd-ranges to a corresponding confidence level.

DISCARDING INCOMPATIBLE PRIOR FUNCTIONS

The idea of robust analysis in principle is the use of all shapes of priors that are consistent with the available information. The problem of this theoretical concept is that one can always find extremely shaped prior functions that, although still consistent with the given information, can 'damage' the result by producing posterior probabilities that are in disagreement with the measurements. This are usually priors which have a strong 'bias' for a specific result, essentially the opposite of the 'non-informative' priors mentioned above. E.g. if we take the iceman example, and assume a prior function that gives a very high probability density for age differences between 1000 and 1100 years (axe older than man) and very low probability for all other age differences, this prior forces the posterior of either the iceman or the axe shaft to lie far apart, completely out of the range obtained by un-modeled calibration. But still this prior is not in disagreement with the (only used) prior information that the axe shaft has to be older (or equal aged) as the iceman. Less artificial, even a prior decreasing exponentially with increasing age differences (as used in the example above) can damage the result if the slope is much steeper as consistent with the real, but unknown, historical situation.

The literature suggests 'prior elicitation', i.e. manual identification and rejection of such priors from the prior sets, see e.g. Berger, 1994. However, we think that this is a error-prone and cumbersome procedure, which would prevent broad acceptance of the method. The manual rejection of priors could be used to 'tune' the result towards the expectations of the user, and thus introduces a unwanted subjectivity. We are looking for a more objective and automatic way to identify corrupt prior functions, focusing on the agreement of the prior with the measured radiocarbon ages (assuming that the radiocarbon measurements are correct). As a prerequisite one has to find a good measure for the agreement of model and data. This will typically be a so-called agreement index, i.e. a number calculated by some mathematical formalism, and a threshold for this number, below which the prior is rejected.

Discarding corrupt prior functions is essential for using robust Bayesian sequencing in practice. However, it must not be ignored that this could be seen as a mutilation of the pure theoretical concept: the measured radiocarbon data now enters the mathematical procedure from two sides, through the reduction of the prior set and through the likelihood functions. Additionally, the choice of the agreement criterion and the threshold level could introduce again a subjectivity similar to the choice of the prior for the 'non-robust' multi-sample calibration. We thus have put significant effort into the development of this criterion.

Measuring the agreement of model and data

The most fundamental measure for the agreement of model (i.e. prior) and determined radiocarbon ages is the so-called prior predictive probability distribution, which will be shortly denoted as prior-prediction in the following. It occurs as 'standardization term' in the denominator in equation 5 - the Bayes theorem - and is given by the multi-dimensional volume (therefore denoted with v) of the product of likelihood and prior function:

$$v(x_1, \dots, x_n) \quad = \quad \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} l(x_1, \dots, x_n | t_1, \dots, t_n) \cdot a(t_1, \dots, t_n) \; \mathrm{d}t_1 \dots \mathrm{d}t_n \qquad (6)$$

The prior-prediction $v(x_1,...,x_n)$ is the probability density to determine a particular set of radiocarbon ages for a given prior function $a(t_1,...,t_n)$. In the specific application here, one is only interested in one particular value of v, that of the actually measured set of radiocarbon values. This value can be compared for different priors used. The ratio of the v-values gives the ratio of the probabilities to get the measured data for different a-priori probability densities for the real ages. This ratio is well known as Bayes-factor and established for model comparison (see e.g. Garcia-Donato and Chen, 2005), and has been already used for archaeological applications too (Jones and Nicholls, 2002). Priors with a higher information content (i.e. more stratigraphic information) are preferred against unspecific ones, since they are less likely to fit to the measured data by chance.

Before we give the formalism to discard prior functions which do not agree with the data with the help of the Bayes-factor, an additional problem has to be solved: In practical applications of Bayesian sequencing technical difficulties arise. Often priors are used that are significantly different from zero on an unrestricted domain and thus cannot be standardized (they have an infinite integral). Examples are both the uniform prior and the uniform span prior mentioned above. From a theoretical point of view these priors are forbidden. However, the infinite domain does not affect the usual multi-sample calibration procedure, since the tails of the likelihood functions used in radiocarbon calibration decrease very fast to zero (essentially proportional to exp(-x^2)). Therefore, unrestricted priors are used in almost all examples in the literature, and we have to be able to cope with them. However, a proper definition of their prior-prediction is impossible (the standard definition results in a factor of zero, i.e. they do not fit to any measured data). To circumvent the problem of standardization the integration could be confined to a restricted domain. Unfortunately, the prior-prediction obtained this way will depend (in indirect proportion) on the volume of this arbitrarily chosen domain.

Our (disputable) approach to calculate the prior-predictions is to transform the unrestricted priors into standardizable ones, by multiplying them with a standardizable weighting function. This can be understood also as restricting the integration to a domain, but with gradual boundaries instead of sharp edges. The function should be significantly different from zero in the region where the radiocarbon measurements are situated, but should not extend much farther, since this will reduce the prior-prediction by increasing the 'domain'. Consequently, we use the sample ages measured, respectively their likelihood functions, to define the weighting function. The straightforward approach is to simply add all likelihood functions of the individual samples and use this as the weighting function for all coordinates. The total n-dimensional weighting function $\lambda(t_1,...,t_n)$ is given by equation 7 (in the following the simplified notation for the fixed set of determined radiocarbon ages, not indicating the radiocarbon ages explicitly, is used again):

$$\lambda(t_1,\ldots,t_n) \quad \propto \quad \prod_{j=1}^{n}\left(\sum_{i=1}^{n}\left(\frac{l_i(t_j)}{\int_{-\infty}^{+\infty} l_i(t)\,\mathrm{d}t}\right)\right) \qquad (7)$$

Where l_i are the single sample likelihood functions as defined in equation 1.

Since all priors can be standardized now and the prior-predictions can be calculated, it is possible to define an agreement factor B (which is a Bayes-factor as mentioned above) that describes quantitatively the agreement of model and data regarding to a reference prior a_{ref}; see equation 8:

$$B \;=\; \frac{v^{(\lambda)}}{v^{(\lambda)}_{ref}} \;=\; \frac{\int_{vol} l\cdot\dfrac{a\cdot\lambda}{\int_{vol} a\cdot\lambda\,\mathrm{d}\mathbf{t}'}\;\mathrm{d}\mathbf{t}}{\int_{vol} l\cdot\dfrac{a_{ref}\cdot\lambda}{\int_{vol} a_{ref}\cdot\lambda\,\mathrm{d}\mathbf{t}'}\;\mathrm{d}\mathbf{t}} \;=\; \frac{\int_{vol} l\cdot a\cdot\lambda\;\mathrm{d}\mathbf{t}\;\Big/\;\int_{vol} a\cdot\lambda\;\mathrm{d}\mathbf{t}}{\int_{vol} l\cdot a_{ref}\cdot\lambda\;\mathrm{d}\mathbf{t}\;\Big/\;\int_{vol} a_{ref}\cdot\lambda\;\mathrm{d}\mathbf{t}} \qquad (8)$$

$\int_{vol} \ldots \mathrm{d}\mathbf{t}$ means a volume integration over all age dimensions; a, a_{ref} and λ denote non-standardized functions here, the needed standardizations are explicitly given in the equation. One can see that the prior functions a and a_{ref} need not to be standardizable any more, it is sufficient if $a\cdot\lambda$ and $a_{ref}\cdot\lambda$ have finite integrals, what is usually granted by the fast decrease of the likelihood functions.

For the special case using a constant reference prior the agreement factor B is the Bayes-factor that compares the prior-prediction $v^{(\lambda)}$, calculated with a prior that combines the tested prior function with the weighting function, with the prior-prediction $v^{(\lambda)}_{ref}$, using a prior that is the weighting function only. The value roughly spoken gives the increase (B>1) or decrease (B<1) of the probability to get the determined set of radiocarbon ages resulting from the use of the prior information. Therefore B is a good quantitative measure of the agreement of the prior function with the radiocarbon ages.

To illustrate the potential of the agreement factor B we look at an ordered sequence of n samples, which are separated sufficiently in time so that the likelihood functions do not overlap. Again the constant function is used as reference prior, so for the constant prior B is equal one. If one tests the uniform prior for a sequence of samples (which is one for age combinations that are in the order given by the prior information, and zero elsewhere), then two results are possible: If the assumed prior order is in disagreement with the order of the radiocarbon ages, the agreement factor would be zero (for non-overlapping likelihoods). If the assumed prior order is in agreement with the order of the radiocarbon ages, the resulting value is the faculty of n. This result is reasonable, because $1/n!$ is the probability to get the right order by chance.

It has to be acknowledged that the restriction of the domain by the function λ as given by equation 7 is not the only possible way. A different definition (e.g. including also all regions between determined radiocarbon ages) would lead to different results for the agreement factor B. However, we see no way to compare priors with finite and infinite integrals in an unambiguous way.

It should be remarked, that there is a relation of the agreement factor B to an agreement measure used in the well-known program OxCal (Bronk Ramsey 1995, 2001, 2009): If one would use the normal multi-dimensional likelihood function as weighting function, by replacing λ with l (and still using a constant a_{ref}), B would become the same as 'F_{model}'.

Finding a reasonable threshold level

The threshold level of B for rejecting a model should be chosen relative to that of a model that is in good agreement with the real ages. A 'perfect' model per definition is one that knows the real ages of the samples and constrains the prior function sharply around them. Within this 'perfect' model only the statistical variances of the measurements reduce the agreement factor. So one can use this case to get an estimate for the threshold of B so that the model is always accepted, except for a allowed percentage of statistical outliers. We estimate the threshold for the most simplest situation considering n samples, all with the same real age and measured with the same accuracy, generating a n-dimensional Gaussian likelihood function (assuming a linear calibration curve). The posterior function resulting from the 'perfect' model is a n-dimensional δ-function at the position of the set of the real sample ages. For this simple case the corresponding agreement factor B, assuming a constant reference prior, can be analytically calculated and is denoted $B_{perfect}$:

$$B_{perfect} \quad = \quad 2^{\frac{n}{2}} \cdot e^{-\frac{1}{2}\cdot\sum_i \left(\frac{x_i - x^{(0)}}{\sigma}\right)^2} \qquad (9)$$

x_i are the measured radiocarbon ages with their common value of uncertainty σ; $x^{(0)}$ is the radiocarbon age corresponding to the common real age of the samples when assuming no measurement error. The sum in equation 9 is the well known chi-square distribution. Analogous to a conventional chi-square test, for a certain confidence level P (e.g. 95.4%), we obtain a corresponding threshold level. The threshold depends on the number of samples n:

$$B_{threshold} \quad = \quad 2^{\frac{n}{2}} \cdot e^{-\frac{1}{2}\cdot\chi^2(P,n)} \qquad (10)$$

So with $B_{threshold}$ a characteristic threshold level has been estimated that accepts models that agree with the real sample ages, in most cases of possible measured radiocarbon data sets.

The actual applied agreement factor

The power of using a Bayes-factor as agreement measure is the fact that it compares prior-predictions that are absolute measures for the quality of a prior. Unfortunately this absolute character can be a disadvantage in practical applications. More often than not, there are outliers (beyond the statistical variance) in the radiocarbon measurements and/or mistakes in the model definition (e.g. a sample is misplaced in stratigraphy). Both would lead to a general decrease of the agreement factors (when defined with a constant reference prior) for all individual priors used, a behavior which is for sure correct. However, in our method, now more priors would be excluded (in the worst case, all), leading to an artificially small prior set. To bypass this problem one has to use a reference prior in the definition of the agreement factor (equation 8) that is as well inflected of the described faults but represents a model that is for sure not extreme, so that it should be discarded itself. Our approach is the use of the uniform prior (a_{unif}) as reference prior, which achieves the mentioned requirements. So the actual applied agreement factor is denoted $B^{(unif)}$ and is given by equation 11 (according to the definition by equation 8):

$$B^{(unif)} \quad = \quad \frac{\int_{vol} l \cdot a \cdot \lambda \ \mathrm{d}\mathbf{t} \ \Big/ \ \int_{vol} a \cdot \lambda \ \mathrm{d}\mathbf{t}}{\int_{vol} l \cdot a_{unif} \cdot \lambda \ \mathrm{d}\mathbf{t} \ \Big/ \ \int_{vol} a_{unif} \cdot \lambda \ \mathrm{d}\mathbf{t}} \qquad (11)$$

The threshold defined in equation 10 is used unchanged for $B^{(unif)}$. This is reliable, because in the considerations to estimate a threshold level, the 'perfect model' was compared with the 'neutral model' represented by the constant reference prior. In the actual applications all priors in the set have to agree with the given stratigraphic constrains, so that the uniform prior can be seen as 'neutral prior'. Naturally, this is just a rough estimate, because as pointed out clearly earlier in this work, whether the constant nor the uniform prior are really 'neutral' in the sense of carrying no information or only the information of the constraints respectively. However, the choice of the threshold cannot be objective anyway, we can only define a consistent procedure.

The equations 11, 10 and 7 establish the needed system for discarding corrupt prior functions that would damage the result of robust Bayesian analysis. It should be noted that the definition of the agreement factor can be generalized for the use of additional statistical parameters like phase boundaries as well.

An additional aspect that has to be mentioned shortly is the fact, that the numerical evaluation of the multi-dimensional volume integrals in equation 11 is not trivial. Actually, we were not able to find an applicable method in the literature. We thus developed a method based on a comparison of the volume to be estimated with a reference volume, executed by a modified kind of Gibbs sampling. The method is sufficient up to about thirty samples or dimensions, but gets convergence problems beyond that number.

A tentative procedure for large sequences

Due to the numerical problems mentioned above, for large sequences we tentatively base the criterion to discard corrupt priors on the single sample agreements indices, as described shortly: For the one-dimensional case l and λ are the same and (using the Bayes theorem; equation 5) equation 11 can be expressed as

$$B_{onedim}^{(unif)} \quad = \quad \frac{\int\limits_{-\infty}^{\infty} l \cdot p \ \ \mathrm{d}t}{\int\limits_{-\infty}^{\infty} l \cdot p_{unif} \ \ \mathrm{d}t} \quad ,$$

where p and p_{unif} are the posterior functions. So for the one-dimensional case the degrees of overlap of likelihood and posterior function are compared. For the multi-dimensional case individual single sample agreement indices are defined analogously by using the single sample likelihood functions l_i and the marginal posterior densities p_i (see equations 1 and 4):

$$B_i^{(unif)} \quad = \quad \frac{\int\limits_{-\infty}^{\infty} l_i \cdot p_i \ \ \mathrm{d}t}{\int\limits_{-\infty}^{\infty} l_i \cdot p_{i,\,unif} \ \ \mathrm{d}t}$$

($\int l_i p_i \mathrm{d}t / \int l_i l_i \mathrm{d}t$ is the usual definition of the single sample agreement index A_i, see Bronk Ramsey, 1995; the second part is canceled out in the equation above.) Similar considerations as above show that also for these single sample agreement indices the threshold from equation 10, evaluated for n=1, is reliable (i.e. $\sqrt{2}/e^2$ for P=0.954 '2σ'). We accept a prior if none of the $B_i^{(unif)}$ drops under that level. This has turned out to be more efficient than using an 'overall index' deduced from the individual agreement indices. However, the use of a single-sample-based agreement criterion is not what we finally intend, and it should be replaced by the multi-dimensional agreement factor (equation 11), as soon as we are able to calculate the latter for large sequences too.

'PARAMETRIC PRIORS' AND 'MODEL AVERAGING'

To perform robust Bayesian analysis, the calculations have to be performed individually for each function from a complete set of prior functions. Theoretically there is an infinite number of possible prior functions, but one can find ways to reduce the set to a practical manageable number (see chapter 'Second example: a large archaeological sequence'). However, it would be advantageous if there was a method that could make the use of an infinite prior set actually possible. The obvious idea is to use a set of prior functions of the same mathematical form, but generated by a set of free parameters. These parameters are then treated as additional dimensions within the formalism, simulating various shapes of prior functions by only one single evaluation of the posteriors. For example, a prior set with exponential decreasing probability densities for an age difference (a similar set with decreasing and increasing functions was used in the iceman example above) is generated by the following parameterization:

$$prior \ = \ \frac{1}{\alpha} \cdot e^{-(t_B - t_A)/\alpha} \quad \text{for } t_B \geq t_A, \ \text{ and 0 otherwise;} \quad (\alpha > 0)$$

It can be shown, that Gibbs sampling with this parametric prior function leads to a summation of the posterior probability densities for all possible shapes of the prior function due to various parameter values, automatically weighted by the prior-prediction. So the method would have the potential to suppress corrupt shaped priors intrinsically by weighting them low. However, there are fundamental differences compared to the independent calculations for the various different prior functions.

First, the generating parameters now need a prior function too. This brings along all the problems of the choice of a proper prior, e.g. that the result now depends on the scaling of the parameters, which changes with the form of the parameterization; e.g.

$$\frac{1}{\alpha} \cdot e^{-(t_B - t_A)/\alpha} \quad \text{and} \quad \alpha \cdot e^{-(t_B - t_A) \cdot \alpha}$$

describe exactly the same set of functions, but give different results, when assuming a constant prior for the parameter α.
And second, this approach is mathematically equivalent to the calculation with a certain, single prior, which can be obtained by integrating first over all prior parameters. While one could argue that this effective prior might now be better - less informative - than the commonly used priors, the search for non-informative priors is not the target of our research. The equivalence to a single prior can be seen as follows:
The final posterior probability density $p(\mathbf{t})$ when using a set of free prior parameters $\boldsymbol{\alpha}$, results from integrating the posterior density including the parameters $p(\mathbf{t},\boldsymbol{\alpha})$ (a probability density within the combined vector space of sample ages and prior parameters) along all parameter dimensions (projection to the sample age sub-space), see equation 12. (Once again the simplified notation without indicating the set of radiocarbon ages is used.)

$$p(\mathbf{t}) \; = \; \int_{vol} p(\mathbf{t},\boldsymbol{\alpha}) \; \mathrm{d}\boldsymbol{\alpha} \qquad (12)$$

Where $p(\mathbf{t},\boldsymbol{\alpha})$ results from the Bayes theorem:

$$p(\mathbf{t},\boldsymbol{\alpha}) \; = \; \frac{l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha})}{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha}) \; \mathrm{d}\mathbf{t} \, \mathrm{d}\boldsymbol{\alpha}} \qquad (13)$$

So the final posterior function $p(\mathbf{t})$ when integrating over the parameters results in:

$$p(\mathbf{t}) \; = \; \int_{vol} \frac{l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha})}{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha}') \; \mathrm{d}\mathbf{t} \, \mathrm{d}\boldsymbol{\alpha}'} \, \mathrm{d}\boldsymbol{\alpha} \; = \; \frac{l(\mathbf{t}) \cdot \int_{vol} a(\mathbf{t},\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\alpha}}{\int_{vol} l(\mathbf{t}) \cdot (\int_{vol} a(\mathbf{t},\boldsymbol{\alpha}') \, \mathrm{d}\boldsymbol{\alpha}') \; \mathrm{d}\mathbf{t}} \qquad (14)$$

Denoting the integral of the prior function along the parameters - which is a particular function in $\mathbf{t}$ again - as effective prior function $a_{eff}(\mathbf{t})$, the final posterior results from a single application of the Bayes theorem again:

$$p(\mathbf{t}) \; = \; \frac{l(\mathbf{t}) \cdot a_{eff}(\mathbf{t})}{\int_{vol} l(\mathbf{t}) \cdot a_{eff}(\mathbf{t}) \; \mathrm{d}\mathbf{t}} \qquad (15)$$

Thus, calculation with parametric priors is equivalent to calculating with a single effective prior, and its shape depends only on the chosen kind of parameterization.

Equation 12 can also be interpreted as a weighted sum of posteriors within the normal space of sample ages (when the parameters are not treated as statistical variables for the Bayes theorem), denoted as $p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha})$. The weighting factors are the prior-predictions within the normal sample age space, what can be seen as follows:
In the sample age space the Bayes theorem can be expressed as:

$$p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha}) \; = \; \frac{l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha})}{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha}) \; \mathrm{d}\mathbf{t}} \qquad (16)$$

Different from equation 13 there is no integration over the parameters, and the prior-prediction remains parameter dependent. Combining equation 13 and 16 one gets:

$$p(\mathbf{t},\boldsymbol{\alpha}) \;=\; p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha}) \cdot \frac{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha})\ \mathrm{d}\mathbf{t}}{\int_{vol} l(\mathbf{t}) \cdot a(\mathbf{t},\boldsymbol{\alpha})\ \mathrm{d}\mathbf{t}\,\mathrm{d}\boldsymbol{\alpha}} \;=\; p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha}) \cdot \frac{v^{(\mathbf{t})}(\boldsymbol{\alpha})}{v} \qquad (17)$$

Where $v^{(\mathbf{t})}(\alpha)$ 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 $\boldsymbol{\alpha}$. So putting $p(\mathbf{t},\boldsymbol{\alpha})$ into equation 12 one gets equation 18, an integration over the individual posteriors weighted with $v^{(\mathbf{t})}(\alpha)$:

$$p(\mathbf{t}) \;=\; \int_{vol} p^{(\mathbf{t})}(\mathbf{t},\boldsymbol{\alpha}) \cdot \frac{v^{(\mathbf{t})}(\boldsymbol{\alpha})}{v}\ \mathrm{d}\boldsymbol{\alpha} \qquad (18)$$

It should be remarked here that this kind of weighted summation of the posteriors for different priors is in principle a method already used and denoted as 'model averaging' (see e.g. Hoeting et al., 1999). It can be seen as a kind of robust analysis, however, it has to cope with the a significant problem: Equation 18 shows clearly that the weighting depends unfortunately not only on the fixed value of $v^{(\mathbf{t})}(\alpha)$ for a particular parameter set $\boldsymbol{\alpha}$, but also on the arbitrary definable scaling of the parameters, mentioned above: a change to another parameter set $\boldsymbol{\beta}(\boldsymbol{\alpha})$ would require to introduce factors like $\partial\alpha_i/\partial\beta_j$ to keep the integration constant; however, no such factors will be present if the set $\boldsymbol{\beta}$ is used from the beginning.

So robust Bayesian analysis that wants to check actually all possible prior functions in our understanding has to calculate each prior function individually. Therefore we have to reduce the theoretical infinite prior set to a manageable finite one, which is shortly discussed within the archaeological example demonstrated next.

SECOND EXAMPLE: A LARGE ARCHAEOLOGICAL SEQUENCE

In this section robust Bayesian sequencing and the usual Bayesian method will be compared using a sequence taken from the article 'Chronology for the Aegean Late Bronze Age 1700-1400 B.C.' by Manning et al. (2006). This large sequence, that links together related contexts of various sites, is very well documented and therefore highly suitable as an example. The left side of figure 6 gives the structure of the sequence in a slightly simplified illustration. The sequence consists mainly of four phases; some of them include substructures. Two tree-ring wiggle matches are used as terminus post quem (TPQ). The sequence includes about 100 radiocarbon data; many of the shown ages of contexts are based on averages of several ^{14}C data.

One has to find a finite set of priors that is a good approximation to the theoretical infinite set of differently shaped prior functions. Although we yet do not have a general rule to find such a good and small prior set, the challenge seems to be manageable in practice. Often critical prior shapes that influence the result considerably are obvious. We want to note that in principle the final result of robust analysis is only a set of n highest posterior density ranges for n samples and parameters. Each interval limit is finally determined by only one single 'extreme', but acceptable prior. If we ignore split hpd-ranges, 2n such extreme priors account for the complete result. Our task is thus not to calculate as many different, but redundant priors as possible, but to find these extreme priors.

Considering the main structure of the sequence we put priors on the lengths of the phases. Similar to the iceman example above, exponential decreasing and increasing probability densities were used. It turned out to be sufficient to use only two different slopes of decreasing and two different slopes of increasing functions, and additionally the constant function for each phase length. A complete n-dimensional prior for the whole sequence is an arbitrary combination of these individual single-phase priors. So for five different slopes and four phases one would get $5^4 = 625$ different total prior functions. Fortunately in practice it turns out that this number can be decreased considerable, because phases which are far from each other seem to behave fairly independent, so there is no need to try all respective combinations of their priors. This is at least the case for the two outer phases in this example. Additionally most of the effect of robust analysis apparently can be achieved by focusing on extreme combinations, as combing a prior with decreasing probability for a phase length with one with increasing density of the neighboring phase. By these considerations the prior set was reduced to 45 differently shaped prior functions, what made the calculation time manageable, although the computer program developed by us is based on the mathematical programming language Matlab and not optimized with regard to runtime, and therefore considerably slower than technically possible. Additionally the commonly used uniform span prior was also included in the prior set.

As mentioned earlier the priors are sieved for their agreement with the measured radiocarbon ages. In this example 26 of the initial 46 priors where excluded by the agreement criterion. It is no disadvantage that a considerable number of prior shapes is discarded. If no prior would fail the agreement criterion, this could be an indicator that the prior set does not cover the full range of possible prior shapes, and the resulting hpd-ranges could be to small.

The resulting hpd-ranges at 95.4% confidence level for robust Bayesian analysis and additionally that for the common uniform span prior are given in figure 6.

When comparing the results of robust analysis with that of the uniform span prior alone, one can see that there are no considerable systematic shifts; the positions of the hpd-ranges are not changed in principle. This indicates that the choice of the uniform span prior does not produce a misleading result in general for this example. However, it is also obvious that most of the intervals become noticeable and sometimes considerable wider. For some contexts (e.g. 'Trianda late LMIA twig' or 'Thera VDL') minor peaks within the un-modeled calibration, that are excluded with the uniform span prior, are included within the 95.4 % interval by robust analysis. The hpd-ranges of phase boundaries are generally changed more than that of contexts, because they are not so directly related to measured radiocarbon ages (e.g. 'Start Mature LMIA' or 'LMIB/LMII transition').

Even though the structure of the sequence is not too simple, the change of some hpd-ranges by robust analysis can be understood qualitatively. For example the Thera VDL has to be younger than the contexts in the preceding phase 'Pre-VDL LMIA'. Especially the context '65/N001/I2' within this phase is very young and so tends to shift the Thera VDL date to younger ages. On the other hand, the subsequent phase 'LMIB Destructions Crete', that starts earliest (with 95.4 % confidence) around 1650 BC when using the uniform span prior, forces the Thera VDL in the opposite direction towards older dates, which tends to compensate the shift from '65/N001/I2'. Robust analysis tries to consider all possible priors, and so also the case where the probability increases highly for very short lengths of the 'LMIB Destructions Crete' phase. Thus a very short 'LMIB Destructions Crete' phase has to start and end somewhere around 1530 BC where all un-modeled calibrations of the contexts within the phase overlap. But if the phase starts that late, it does not tend to shift the Thera VDL age towards older ages any more. So only the shift to younger ages remains and enhances the small tail - in detail two small peaks - at the young side of the un-modeled Thera VDL probability distribution, so that they are included in the resulting 95.4 % hpd-range. Thus, when focusing only on the 95.4 % (2σ) confidence level, the result for the Thera VDL from robust analysis differs remarkable from this using the uniform span prior. However, at the 68.3 % (1σ) confidence level (we calculate the intervals at all levels of confidence simultaneously) the minor peaks are excluded again by both methods in common.

Nevertheless, to avoid premature interpretations of this or any other detailed result of the sequence, one should be aware of the following fact: When evaluating the sequence with robust analysis we use in principle all possible prior functions that are in agreement with the constraints as given in simplified form on the left side of figure 6, without asking whether archaeological reasons could be found further that exclude a part of the particularly used prior functions within the used set. If we actually could exclude priors by further archaeological arguments (e.g. expectation values on some phase length or maxima for some time spans, etc.) this would naturally reduce the hpd-ranges. Therefore, robust Bayesian analysis forces the user to collect and use actually all available archaeological information, to avoid getting wider hpd-ranges as necessary.

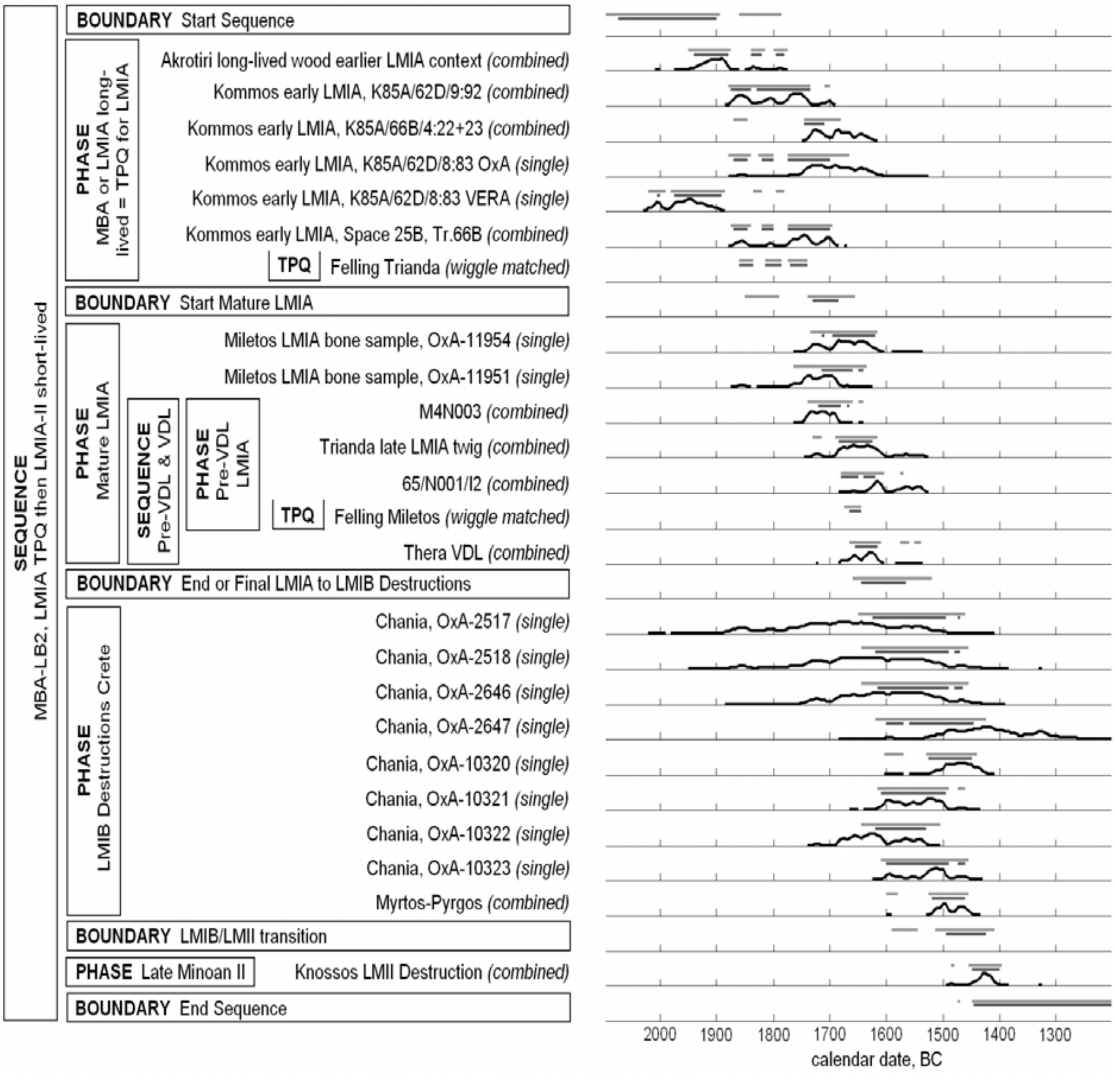

Figure 6 Second example: a large archaeological sequence taken from the article 'Chronology for the Aegean Late Bronze Age 1700-1400 B.C.', Manning et al. (2006). A comparison of the hpd-ranges at a 95.4% confidence level resulting from robust Bayesian analysis (bars in light gray) and from the usual uniform span prior (bars in dark gray) is given. The un-modeled calibrations are given as probability density distributions (curves in black).

DISCUSSION

For sure, we do not see the procedures introduced by this work as a completed method, we rather want to introduce ideas that are hopefully discussed within the community. Our view of what we consider to be the most critical aspects is summarized in the following.

What is goal of our 'robust analysis', and what it is not:

The idea is to avoid (or at least to reduce) the subjectivity of the chosen prior function for a given archaeological setting. This means, we assume that the archeologists came to a final description of the archaeological facts, which allows one to define a complete set of possible different prior functions that are all in agreement with this constraints. Our results are subsequently deduced from this set as described in detail in this work. Of course, this is an idealization of the real scenario, where the archaeological facts can hardly be fixed in an unchangeable form. Archaeologists will rather discuss the stratigraphy and find different possible interpretations and perspectives. They are able to assess different assumptions but can not assign quantitative probability values. To handle this complexity remains with the archaeologists. The goal of our method is not to deal with this part of the problem but to avoid the need of selecting a particular prior function, as there remain an infinite number of different possible functions, even when the archaeological constraints are fixed.

However, the calculation of Bayes factors to compare models in an quantitative way, could support the selection of the archaeological model (the fixation of the archaeological constraints) too. Although, in our opinion, the archaeological constraints should better be defined based on archeological information only, without considering the measurements.

What is the rigorous meaning of the resulting unified highest posterior density ranges:

Usually in Bayesian statistics the result is expressed as posterior probability density, which is not possible any more within the method introduced here. Instead we use the unified highest posterior density ranges (unified hpd-ranges). Since these are not commonly used, it is important to elaborate there rigorous meaning very carefully: For given constraints based on the archaeological evidence, there is an infinite number of consistent realizations of the prior function, each resulting in different marginal posterior distributions. If we assume that the formulated archaeological constraints are correct, which means that there is no statement included that is not historical true, then the prior set will definitely include a function that correctly represents the actual probability to find samples of particular ages. This prior function would be the ideal choice to get an unbiased posterior density. To unify the hpd-ranges of the posterior marginals for all priors guaranties that the hpd-ranges for the ideal prior are included. As the real sample ages lie within the hpd-ranges for the ideal prior with the corresponding probability (e.g. 95%), they lie within the unified ranges with at least the corresponding probability. So the meaning of the unified ranges is this: One can be sure that the real sample ages lie within those intervals with at least the specified probability. The conditions therefore are just that the archaeological constraints and the used accuracies of the measurements are correct. A similar statement is not possible if one uses a single particular prior, because one does not know how strong the result is biased by the prior. (The deviation from this ideal concept caused by the need of discarding 'corrupt' priors does not change the considerations here in principle.)

The kind of robust analysis used by us is based on Berger (1994; section 1.3 and 4.1): 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 our case, we think that the union of all hpd-ranges is the straight forward implementation of this method.

Finally it has to be remarked, that the use of hpd-ranges instead of probability densities has a little disadvantage: If the marginal posterior density is very flat, the excluded parts can have densities nearly as high as the included ones. However, this problem is not specific to robust analysis, it is present whenever hpd-ranges are used. Regardless of this, hpd-rages are highly accepted to characterize the results of Bayesian analysis.

Is this emphasis on 'safe' results really necessary?

It is disputable whether it is really necessary that the results have to be 'correct' in a rigorous sense as provided by robust analysis; or are results even meaningful although they may be biased by the choice of a particular prior function. Naturally, it is reasonable to look at results calculated with the best assumption for the prior function, but one has always to be aware that this results are only accurate if the prior function reflects the unbiased functional representation of the constraints given by the archaeological facts. The meaning of 'an unbiased functional representation' can be illustrated by the Iceman example from above. Imagine one would know the real ages (times of death) of a large number of Early Bronze Age men and the ages of the wood of the axes carried by them too. Now one could plot a point for each pair of ages within a two-dimensional co-ordinate system and fit the points with a corresponding probability density function. This density would represent an unbiased prior function, because now the prediction on the probability of a particular age difference of 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. So deciding to use a particular single prior or robust analysis instead, is the decision between a high precision but also a higher risk of error and lower precision but a lower risk of error too.

It should be noted, that the perception expressed by the 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.

The loss of a particular prior or 'model':

It can be seen as drawback of the used method that one only focuses on the union of all priors and does not analyze the influence of the individual priors - which are usually seen as different models - on the result. First of all it should be remarked that the calculation performed by us provides all individual marginal posteriors to any prior used within the prior set. So it is possible to analyze the influence of individual priors on the result manually. Of course if there are hundreds of different prior functions used - which can be reasonable - this will become cumbersome. However, our perception of this topic is the following: We think, what should be actually seen as 'model' for this considerations is not the individual prior function, but the available archeological constraints or information. And, as pointed out earlier in this paper, this information can usually not be expressed by a single function, but by a set of prior functions. So the entirety of the prior functions characterizes the model, and therefore the result for this entirety has a more fundamental meaning than the results for the individual prior functions.

The subjectivity of the finite prior set:

In fact, it is a disadvantage of robust Bayesian analysis as performed by us, that the theoretical infinite set of possible prior functions has to be approximated by a finite set. This could lead to the impression, that the choice of the actual used priors brings in the subjectivity again, that the method tries to avoid. Fortunately the problem turns out to be solvable in practice, because one is allowed to include every prior one can think about within the set, if just consistent with the archaeological constraints. Neither redundant priors nor meaningless priors will (in principle) degrade the result. Naturally 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.

The arbitrary procedure for prior standardization:

There could be the legitimate criticism that we use an arbitrary method to standardize prior functions with originally infinite integrals, for the purpose to be able to calculate Bayes factors. The problem is not easy to solve, because it is a fundamental one. It arises from the use of unrestricted priors that have no defined probability density values (their density would be zero anywhere, if standardized on the infinite domain). On the other hand, it would not make sense to discard unrestricted priors generally, because many archaeological constraints are most directly realized by this kind of priors. Thus, to do a quantitative model comparison, one has to restrict these priors in an appropriate way. This restriction will always have to be based on the available calibrated radiocarbon ages or on the likelihood function in other words. The particular way to realize this restriction is arbitrary in fact. However, it does not make a difference in principle if one defines a sharp domain around the likelihood or uses a 'graduated domain' derived from the sum of all single sample calibrations, as we do in this work.

Although it is not possible to avoid this arbitrariness, it is important to make clear, that it does only influence the process of discarding 'corrupt' priors (that are in serious disagreement with the data). The calculation of the posterior densities for the different priors used can be performed without standardization of the prior function. This is the benefit of our method where only the hpd-ranges are unified rather than the densities are summarized, which makes the result independent of a weighting of the individual priors.

The problem with 'model averaging':

We presented a detailed discussion in our work on defining a 'parametric' prior set and handle the parameters as additional dimensions within Bayesian calculation, which is a possible technical realization of an approach that is known as model averaging. Of course, this method has the great advantage, that it can be performed fully embedded in the Bayesian framework, resulting in marginal posterior densities for the original parameters (sample ages, boundaries, ...) and for the prior parameters too. When starting our work we assumed that we will realize our goal with this method, which is mathematically well-defined and clear. It was not easy to accept a concept with discrete prior functions that have to approximate an infinite set of functions. However, 'model averaging' has its serious drawback too: The different functions have to be weighted now by a prior probability density for the parameters. In our opinion it is very difficult to assign well-grounded probabilities for differently shaped prior functions. This problem induces serious arbitrariness again and it is not possible to solve it pragmatically by postulating that all priors should have the same probability, because the chosen kind of parameterization (scaling) is an intrinsic arbitrary weighting anyhow. Aside from this complication the method delivers just an effective (or 'average') prior function, which again can be 'right' or 'wrong' in the sense that it can approximate well the unbiased functional representation of the constraints given by the archaeological facts or it does not (as illustrated by the example in 'Is this emphasis on 'safe' results really necessary?' above).

However, we do not want to deny that model averaging is a meaningful method. The method assists powerfully the process of finding an appropriate average prior function and the posterior density of the prior parameters shows furthermore to which extend the individual prior shapes contribute to the average prior. So in our understanding, model averaging is a kind of robust analysis that is based on the principle of finding a non-informative prior function; an approach that is different in principle from that we favor in this work.

CONCLUSION

It seems obvious, that there are applications where usual Bayesian sequencing suppresses possible ranges for the real sample ages. These results from the use of a single, particular prior function, which is commonly not fully defined by the available archaeological information, and therefore is actually only one of various (infinitely many) different possible functional realizations of the prior information. Robust Bayesian analysis, that in principle considers all possible different shaped prior functions, could be a way to avoid these artifacts, thus leading to a higher reliability, even though paid for with a reduced precision.

However, there are difficulties that have to be investigated further: Not completely solved is the problem that the use of 'corrupt' priors can destroy the result. Discarding the priors in respect to their agreement with the radiocarbon measurements by means of a threshold level is possible, but not finally satisfying, because this technique breaks the absolute objectivity of robust analysis again. A further difficulty is the fact that the calculations have to be done for each prior individually, so that only a finite number of prior functions can be considered. Fortunately it seems to be not difficult in practice to find a finite prior set that is a sufficient approximation for the theoretical infinite set of prior functions.

In conclusion, we think that there are no unsolvable obstacles to establish robust Bayesian analysis as a safe sequencing method for radiocarbon dates which are related through archaeological evidence.

ACKNOWLEDGEMENTS

This work is supported by 'SCIEM2000 - Synchronization of Civilizations in the Eastern Mediterranean in the 2nd Millennium BC'; a Special Research Program (SFB) of the Austrian Academy at the Austrian Science Fund and under Patronage of the UNESCO.

We thank two anonymous reviewers for critical and helpful comments.

REFERENCES

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

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

Bonani G, Ivy S, Hajdas I., Niklaus TR, Suter M. 1994. AMS ^{14}C age determinations of tissue, bone and grass samples from the Ötztal Ice Man. *Radiocarbon* 36(2):247-250.

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. Development of the radiocarbon calibration program. *Radiocarbon* 43(2A):355-363.

Bronk Ramsey C. 2009. Bayesian analysis of radiocarbon dates. *Radiocarbon* 51(1):337-360.

Buck CE, Cavanagh WG, Litton CD. 1996. *Bayesian Approach to Interpreting Archaeological Data*. John Wiley & Sons Ltd, Chichester, England.

Buck CE, Kenworthy JB, Litton CD, Smith AFM. 1991. Combining archaeological and radiocarbon information: a Bayesian approach to calibration. *Antiquity* 65:808-821.

Buck CE, Litton CD, Smith AFM. 1992. Calibration of radiocarbon results pertaining to related archaeological results. *Journal of Archaeological Science* 19:497-512.

Garcia-Donato G, Chen M-H. 2005. Calibrating Bayes factor under prior predictive distributions. *Statistica Sinica* 15:359-380.

Gilks WR, Richardsion S, Spiegelhalter DJ (editors). 1996. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

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, 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.

Jones MD, Nicholls GK. 2002. New radiocarbon calibration software. *Radiocarbon* 44(3):663-674.

Krause A. 1994. *Computerintensive statistische Methoden: Gibbs sampling in Regressionsmodellen.* Gustav Fischer Verlag, Stuttgart.

Kutschera W, 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.

Kutschera W, Rom W. 2000. Ötzi, the prehistoric Iceman. *Nuclear Instruments and Methods in Physics Research B* 164-165:12-22.

Manning SW, Bronk Ramsey C, Kutschera W, Higham T, Kromer B, Steier P, Wild EM. 2006. Chronology for the Aegean Late Bronze Age 1700-1400 B.C. *Science* 312:565-569.

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.

Rios Insua D, Ruggeri F. 2000. *Robust Bayesian Analysis*. Springer, New York.

Rom W, Golser R, Kutschera W, Priller A, Steier P, 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.

Sivia DS. 1996. *Data Analysis - A Bayesian Tutorial*. Oxford University Press Inc., New York.

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

Weninger F, Steier P, Kutschera W, Wild EM. 2006. The Principle of the Bayesian Method. *Egypt and the Levant* 16:317-324.