Bayesian inference in physics

Udo von Toussaint*

Max-Planck-Institute for Plasmaphysics, Boltzmannstrasse 2, 85748 Garching, Germany

(Received 8 December 2009; published 19 September 2011)

Bayesian inference provides a consistent method for the extraction of information from physics experiments even in ill-conditioned circumstances. The approach provides a unified rationale for data analysis, which both justifies many of the commonly used analysis procedures and reveals some of the implicit underlying assumptions. This review summarizes the general ideas of the Bayesian probability theory with emphasis on the application to the evaluation of experimental data. As case studies for Bayesian parameter estimation techniques examples ranging from extra-solar planet detection to the deconvolution of the apparatus functions for improving the energy resolution and change point estimation in time series are discussed. Special attention is paid to the numerical techniques suited for Bayesian analysis, with a focus on recent developments of Markov chain Monte Carlo algorithms for high-dimensional integration problems. Bayesian model comparison, the quantitative ranking of models for the explanation of a given data set, is illustrated with examples collected from cosmology, mass spectroscopy, and surface physics, covering problems such as background subtraction and automated outlier detection. Additionally the Bayesian inference techniques for the design and optimization of future experiments are introduced. Experiments, instead of being merely passive recording devices, can now be designed to adapt to measured data and to change the measurement strategy on the fly to maximize the information of an experiment. The applied key concepts and necessary numerical tools which provide the means of designing such inference chains and the crucial aspects of data fusion are summarized and some of the expected implications are highlighted.

DOI: 10.1103/RevModPhys.83.943 PACS numbers: 02.50.Tt, 07.05.Fb, 82.80.−d, 89.20.Ff

CONTENTS

*udo.v.toussaint@ipp.mpg.de

I. INTRODUCTION

In physics one is constantly faced with the task of having to draw conclusions from imperfect information. Physics experiments are always affected by instrumental restrictions and limited measurement time. Therefore, the typical problems encountered in the analysis of physics measurements involve incomplete and noisy data. In addition, the reasoning about the interesting quantities is often hampered by the ill-conditioned nature of the underlying inversion problem.

Bayesian methods have been developed as a tool for reasoning quantitatively in situations where arguments cannot be made with certainty. Historically, the connections between physics and Bayesian inference had been strong, with contributions, e.g., from de Laplace (1812), already in the early 19th century. Later on the frequentist interpretation was dominating the field of data analysis, especially over large periods of the 20th century. One main reason for this development was the significantly smaller numerical effort of the frequentist approach (requiring the optimization of model parameters) compared to the Bayesian approach (requiring the integration of model parameters) in many problems of interest. However, the situation is about to change: The application of Bayesian inference in physics has flourished over the past two decades, driven by the rapid increase of computer power and theoretical progress.

In addition, the landscape of data analysis problems is broadening, in many respects favoring the use of Bayesian approaches. Although classical parameter estimation still dominates the field, model comparison is increasingly coming into focus. Here the Bayesian method provides a conceptual simple and transparent approach which allows for a quantitative ranking of competing physical models.

A further data analysis challenge is raised by the increasing number of complex, multidiagnostic experiments [see, e.g., Wendelstein 7-X (Dinklage *et al.*, 2004)]. The interrelated measurements of different diagnostics have to be evaluated in a coherent way. For this problem of data fusion with potentially many nuisance parameters, boundary conditions, and different noise statistics, Bayes' theorem provides a consistent and scalable approach.

New approaches for data analysis are also required for next-generation experiments, providing several terabyte of data per day [see, e.g., Large Synoptic Survey Telescope (Loredo *et al.*, 2009)]. This amount of data requires automated analysis systems that can both act and react with minimal human intervention. Data driven experiment forecast and optimization will therefore become increasingly important to identify optimal measurements and strategies that are likely to give the highest scientific return. One such vision is based on a cycle of hypothesis building, inquiry, and inference. However, since the accessible information is almost always incomplete and noisy, the inference has to process uncertain knowledge. Bayesian probability theory provides a consistent conceptual basis for this problem of induction in the presence of uncertainty.

The remainder of the paper is structured as follows: In Sec. II the basics of Bayesian inference are introduced and Bayes' theorem is derived from the sum and product rule. Next various ways are considered how to encode a given state of information into the form of a probability distribution. Section III applies the concept to several parameter estimation examples, covering extra-solar planet detection, climate time-series analysis, experiments with very low counting statistic, and deconvolution of broadened spectra in Rutherford backscattering analysis. Starting from Bayes' theorem the solution to a problem is often very simple in principle. However, many calculations, e.g., the marginalization of parameters, require integrals over the model parameter space which can be very time consuming. An overview of the available numerical methods to perform these high-dimensional integrations is given in Sec. IV. The focus is on recent developments of Markov chain Monte Carlo (MCMC) algorithms, in many situations the only way to integrate over the parameter space. Methods to suppress random-walk behavior of standard-MCMC algorithms are presented. Several new approaches to compute model probabilities are discussed in more detail. The Bayesian approach to model comparison is illustrated in Sec. V with examples from cosmology and the analysis of mass spectrometer data with limited information about the cracking matrix. Further case studies are about the joint analysis of discordant data sets from plasma experiments and signal-background separation in electron spectroscopy using mixture models. The analysis of data from large experiments such as the stellarator Wendelstein 7-X (Dinklage *et al.*, 2004) requires new, integrated data analysis approaches. How to combine data from several different diagnostic systems in a coherent way is discussed in Sec. VI and applied to examples from fusion research. Bayesian graphical networks as a tool to handle complex interdependencies are introduced and approximation methods are discussed. The simultaneous localization and mapping problem (SLAM), one of the key problems in robotics, is used as another example of joint data analysis. In Sec. VII the common way of using experimental diagnostics as passive recording devices is challenged. Using Bayesian experimental design techniques the measurement strategy can be adapted on the fly, in many situations with a tremendous increase in performance. In addition, experimental designs can be optimized with respect to different measurement scenarios using a quantitative measure of performance. Although these techniques are still in an early state and their potential still needs to be explored, first applications already yielded very promising results. The main concepts underlying all these diverse data analysis problems are introduced in the next section.

II. BAYESIAN CONCEPT

Bayesian probability theory provides a simple and straightforward recipe to data analysis problems in physics, as will be detailed with several examples. Although this review tried to be as self-contained as possible, for brevity several technical details had to be omitted and the interested

reader is encouraged to consult the available literature. Two highly recommended textbooks on the subject written by physicists are Gregory (2005a) and Sivia (2006). Bayesian inference from a statisticians and mathematicians point of view is presented by Berger (1985), Bernardo and Smith (2000), and Harney (2003). Bernardo and Smith also provide a good overview of the statistics literature. The use of Bayesian inference in the field of artificial intelligence is presented by Russell and Norvig (2003), the connection of inference and information theory is made by MacKay (2003), and state-of-the-art applications of Bayesian algorithms are presented in the book by Bishop (2006) about machine learning. The work of E. T. Jaynes has inspired many physicists to use Bayesian methods; his recommended book (Jaynes and Bretthorst, 2003) still reflects the fierce debates between frequentist and Bayesian statisticians up to the 1980s. Eloquent introductions to Bayesian inference and illustrative examples can be found in Loredo (1990, 1992, 1994). The applied statisticians point of view is given by Box and Tiao (1992) and O'Hagan (1994) and more recently by Robert (1994) and Leonard and Hsu (1999), while Carlin and Louis (1998) and especially Gelman *et al.* (2004) put emphasis on Markov chain Monte Carlo methods for Bayesian inference. Furthermore, there are several read-worthy reviews of Bayesian methods applied in various areas of physics, see, e.g., D'Agostini (2003), Dose (2003a), and Trotta (2008). An overview of applied Bayesian analysis in areas outside of physics is given by O'Hagan and West (2010). The latest developments in Bayesian inference are topics of several conference series, e.g., the annual conference on "Bayesian Inference and Maximum Entropy Methods in Science and Engineering," the "Valencia International Meetings on Bayesian Statistics" every 4 years, and the biannual "ISBA meetings."

A. Basics

In Bayesian probability theory (BPT), the viability of a hypothesis H is assessed by calculating the probability of the hypothesis given the observed data D and any background information I. Following Jeffreys (1961) such a probability is written as $p(H|D, I)$. In the Bayesian framework the frequentist interpretation of probability as long-run relative frequency of occurrence of an event in an ensemble of identically prepared systems is generalized. In BPT, probability is commonly regarded as a measure of the degree of belief about a proposition (Bernardo and Smith, 2000; Jaynes and Bretthorst, 2003). This definition has the advantage that systematic and statistical uncertainty can be treated with the same formalism and also situations where identical ensembles are hard to imagine, e.g., "What is the age of the Universe?," thus significantly extending the range of application.

The BPT rests on two rules (Cox, 1946) for manipulating conditional probabilities. The sum rule states that the probabilities of a proposition H and the proposition for not H (denoted by $\bar{H}$) add up to unity:

$$p(H|I) + p(\bar{H}|I) = 1. \tag{1}$$

Throughout this work, only exclusive and exhaustive hypotheses will be considered, so that if one particular hypothesis is true, all the others are false. For such hypotheses the normalization rule

$$\sum_i p(H_i|I) = 1 \tag{2}$$

applies. The second rule is the product rule which states that a joint probability or probability density function $p(H, D|I)$ can be factorized such that one of the propositions becomes part of the condition (i.e., moves right of the vertical bar). Because of the symmetry with respect to H and D, this can be done in two ways

$$p(H, D|I) = p(H|I)p(D|H, I) = p(D|I)p(H|D, I). \tag{3}$$

Comparison of the two equivalent decompositions in Eq. (3) yields Bayes' theorem

$$p(H|D, I) = p(H|I)p(D|H, I)/p(D|I). \tag{4}$$

Bayes' theorem relates the product of prior $p(H|I)$ and likelihood $p(D|H, I)$ to the posterior probability $p(H|D, I)$. The prior distribution represents the state of knowledge before seeing the data. The normalization constant in the denominator is the *marginal likelihood* or *evidence*. The evidence will be crucial for model comparison but is usually of less importance in parameter estimation problems. The posterior probability distributions provide the full description of the state of knowledge about H. In most cases the hypothesis H will be a combination of parameters $(\theta_1, \theta_2, \ldots, \theta_N)$ and possible models $(M_1, \ldots, M_K)$ which depend on the problem to be analyzed.

Equation (4) also reveals that the maximum-likelihood (ML) estimate is usually different from the posterior estimate except for the special case of a constant prior. The maximum-likelihood estimate obtained by maximizing the likelihood function is often mistaken as the most probable estimate given the data. This is not so: The obtained hypothesis is the one that would make the observed data most probable. This is logically quite different. An example taken from Sivia and David (1994) highlights this distinction. The probability of rain given that there are clouds overhead and the probability of clouds overhead given that it is raining are clearly not the same. The quantity that is required (the most probable estimate given the data) is instead given by the posterior probability $p(H|D, I)$. It is related to the likelihood function through the prior probability $p(H|I)$. From a different point of view Bayes' theorem is a recipe for learning. Initially available prior knowledge about the hypothesis H coded in the distribution $p(H|I)$ is modified by the new information provided by the measured data D to its posterior distribution $p(H|D, I)$.

It is often necessary to summarize the posterior distribution of a parameter $p(\theta|D, I)$ in terms of a few numbers. A convenient description is given by the moments of the posterior

$$\langle \theta^n \rangle = \int d\theta \theta^n p(\theta|D, I), \tag{5}$$

where the mean $\langle \theta \rangle$ is obtained with $n = 1$. Other possible choices are the position of the most probable value of the posterior [also termed maximum *a posteriori* (MAP) estimate] or the median of the posterior distribution. For a

symmetric distribution the mean value and the median coincide. However, all those numbers may be strongly misleading in the case of skew or multimodal distributions. Furthermore, the moments and the MAP estimate depend on the chosen parametrization.

The Bayesian analog of a frequentist confidence interval is usually referred to as a credible region or also simply as a (Bayesian) confidence interval. The credible region R corresponding to some probability mass C (typically $C = 68\%$ or $C = 95\%$) is defined as the part of the parameter space such that

$$\int_R d\theta\, p(\theta|D, I) = C, \tag{6}$$

with the posterior density inside R everywhere greater than outside it. This definition enables direct statements about the likelihood of θ falling in R, i.e., "The probability that θ lies in R given the observed data D is at least C," in accord with common practice in physics (Carlin and Louis, 1998). In contrast the corresponding statement of the frequentist confidence interval could be phrased as "If we could recompute R for a large number of data sets collected in the same way as ours, about C would contain the true value of θ." However, only the measured data set is available; therefore the computed confidence interval R will either contain θ or it will not. This may lead to the problem of *recognizable subclasses* (Cornfield, 1969): A statistic that is good in the long run may be poor in cases that can be identified from the measured data. Loredo (1992) provided the following example: The probability density that an event (e.g., photon) will arrive at time t is given by a truncated exponential distribution,

$$p(t|\tau, t_0, I) = \begin{cases} 0, & \text{if } t < t_0, \\ \frac{1}{\tau}\exp(-\frac{t-t_0}{\tau}), & \text{if } t \geq t_0, \end{cases} \tag{7}$$

with $\tau = 1$ known. The time t_0 is to be estimated from three observed events $t_1 = 12$, $t_2 = 14$, and $t_3 = 16$. Using an unbiased (frequentist) estimator $\hat{t}$ of t_0:

$$\hat{t} = \frac{1}{N}\sum_{i=1}^{N}(t_i - \tau), \tag{8}$$

the shortest 90% confidence interval for t_0 can be calculated to be $12.15 < t_0 < 13.83$ [cf. Jaynes (1983), p. 173]. However, as the first recorded event was observed at $t_1 = 12$ it is certain that t_0 was earlier and that the confidence interval is centered around a value where it is impossible for t_0 to lie. Using a uniform prior for t_0 the Bayesian 90% credible region for the same problem is $11.23 \leq t_0 \leq 12.0$ which is entirely in the allowed range and about one-half the size of the confidence interval. For the details of the straightforward calculation and a further discussion about recognizable subclasses see Loredo (1992). As pointed out by Jaynes (1976): The value of an inference lies in its usefulness in the individual case and not in its long-run frequency of success; they are not necessarily even positively correlated.

B. An example

The concepts presented in the previous section are best illustrated by an example. The mass of a new elementary particle has to be determined. From previous measurements and theory it is known that the mass has to be non-negative and is not larger than an upper limit $0 \leq m \leq m_{\text{upper}}$. This defines the prior information. The prior distribution has to encode the existing knowledge about the restricted range of the possible mass of the particle. A sensible suggestion is

$$p(m|I) = \begin{cases} 1/m_{\text{upper}}, & \text{if } 0 \leq m \leq m_{\text{upper}}, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

which assigns zero probability to the mass parameter values excluded *a priori* and otherwise does not prefer any value within the allowed range. Furthermore, the experiment measures the mass distorted by Gaussian noise of known variance σ^2. The physical model, relating the parameter of interest and the ideal undistorted signal, is in this example the identity, because the parameter (mass) and signal (mass) are identical. The likelihood for measuring d is then given by

$$p(d|m, \sigma, I) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(d-m)^2}{\sigma^2}\right). \tag{10}$$

Once likelihood and prior distributions are specified, the problem is reduced to computing the posterior. The posterior distribution $p(m|\mathbf{d}, \sigma, I)$ follows from Bayes' theorem

$$p(m|\mathbf{d}, \sigma, I) = \frac{p(m|I)\prod_{i=1}^{N} p(d_i|m, \sigma, I)}{Z} \tag{11}$$

for N independent measurements (bold typeface such as $\mathbf{d}$ is used throughout the review to indicate vectors and matrices). The variable Z (called evidence) normalizes the posterior distribution $p(m|\mathbf{d}, \sigma, I)$ and is of importance in model comparison problems. The computation of the evidence Z is treated next [cf. Eq. (18)]. In Fig. 1 the data points of three measurements $\{d_1 = 0.09 \pm 0.15,\ d_2 = -0.2 \pm 0.15,\ d_3 = 0.05 \pm 0.15\}$ are shown. Using the maximum-likelihood method to estimate the most likely mass would result in $m = -0.02$, an estimate outside of the physically sensible range of $0 \leq m \leq 0.2$. Also the confidence interval would cover to a large extent negative (impossible) mass values. Simply discarding the negative data point (on what grounds)

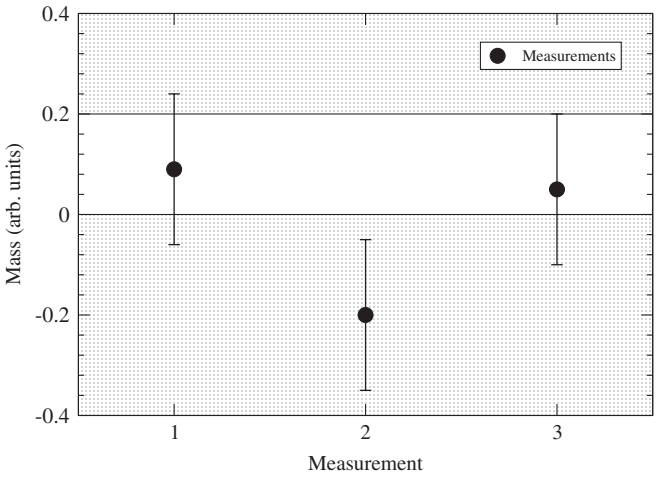

FIG. 1. Three measurements of a physical quantity (e.g., mass) with Gaussian uncertainty, the allowed range being limited to [0; 0.2].

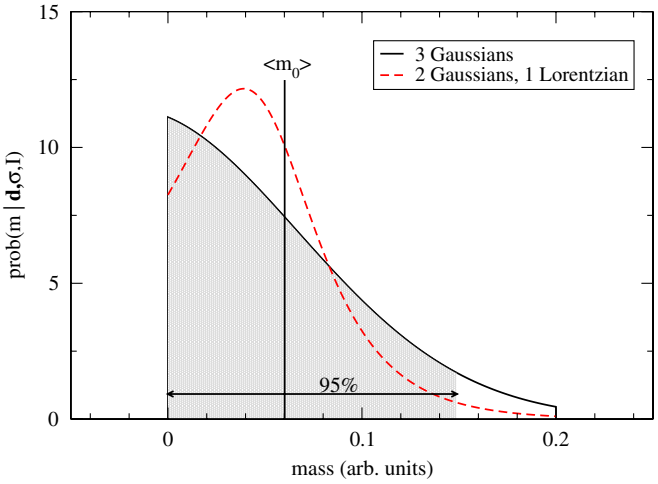

FIG. 2 (color online). Posterior distribution $p(m|\mathbf{d}, \sigma, I)$ for three measurements with Gaussian uncertainty. The mean value $\langle m_0 \rangle$ is denoted by a vertical bar and the extent of the smallest 95% credibility region is indicated by the shaded area underneath the posterior distribution. Additionally the posterior distribution for the case of one measurement with a Cauchy (Lorentzian-) error distribution instead of a Gaussian distribution is given by a dashed line.

would neglect the information gained by this measurement and lead to a biased (wrong) estimate. A more sensible result is obtained when the available prior information is incorporated. Figure 2 shows the Bayesian solution calculated by Eq. (11). The posterior distribution is positive only within the allowed range and has a maximum for $m = 0$. The mean value $\langle m_0 \rangle$ is given by

$$\langle m_0 \rangle = \int dm\, m\, p(m|\mathbf{d}, \sigma, I) = 0.06 \tag{12}$$

and the smallest 95% interval (credibility region) for the mass is [0, 0.145]. The interested reader is invited to compare this straightforward incorporation of prior knowledge with the complexity of frequentist based approaches to the problem of parameter estimation in truncated parameter spaces (Katz, 1961; Bickel, 1981; van Eeden, 1995; Marchand and Strawderman, 2004). A slight change in the example, e.g., that the experiment providing d_3 suffers from noise given by a Cauchy-(Lorentzian) distribution,

$$p(d|m, \beta, I) = \frac{\beta}{\pi[\beta^2 + (d-m)^2]}, \tag{13}$$

with full width at half maximum (FWHM) of $2\beta = 0.1$ is easily accommodated in the Bayesian approach by updating the corresponding likelihood term in Eq. (11). The effect on the posterior estimate of the mass is shown as a dashed line in Fig. 2 which now exhibits a maximum around $m = 0.04$ and the 95% credibility interval for the mass can easily be determined using the shortest interval $[m_{\min}, m_{\max}]$ which fulfills

$$\int_{m_{\min}}^{m_{\max}} dm\, p(m|\mathbf{d}, \sigma, I) = 0.95. \tag{14}$$

However, within the frequentist framework the design of appropriate confidence intervals for the altered estimation problem is still an active research area.

C. Marginalization and evidence

The last quantity to be explained in Eq. (4) in more detail is the term $p(D|I)$. As a consequence of the sum and product rules it follows that for mutually exclusive and exhaustive hypothesis H_i

$$p(D|I) = \sum_i p(H_i|I)p(D|H_i, I) = \sum_i p(H_i, D|I) \tag{15}$$

holds. Summations of this kind are often applied in Bayesian inference and are called *marginalizations*. The important marginalization rule, Eq. (15), provides a recipe how to remove unwanted nuisance variables from a Bayesian calculation. In the continuum limit the summation of Eq. (15) is replaced by integration over the nuisance variable (here y),

$$p(x|I) = \int dy p(x, y|I) = \int dy p(y|I)p(x|y, I). \tag{16}$$

Comparing Eq. (15) and Bayes' theorem Eq. (4) it follows that the denominator of Bayes' theorem $p(D|I)$ (which does not depend on H) plays the role of a normalization constant, which ensures

$$\sum_i p(H_i|D, I) = 1. \tag{17}$$

This denominator, obtained by marginalization of all variables (hypotheses) is an uninteresting normalization constant from the perspective of parameter estimation. However, it is of vital importance for model comparison. For this reason, it is often referred to as the evidence for a model, but also other names, such as *prior-predictive value* or marginal likelihood are common. For the example in the previous section the evidence is easily obtained by integration of the nominator of Eq. (11) with respect to the one-dimensional parameter m according to Eq. (16):

$$Z = p(\boldsymbol{d}|, \sigma, I) = \int dm p(m|I) \prod_{i=1}^{3} p(d_i|m, \sigma, I) \approx 69, \tag{18}$$

but in general integrations over large (multidimensional) parameter spaces are required. The evidence value on its own is of no particular relevance and is just normalizing the posterior distribution. Therefore, its computation is often omitted in parameter estimation problems. However, comparison of evidence values of competing models for a given data set is the basis of Bayesian model comparison and is further discussed in Sec. V.

D. Prior probability distributions

All of the rules written down so far show how to manipulate known probabilities to find the values of other probabilities, and they skipped the problem of how to formulate a distribution given certain prior knowledge. But to be useful in applications, rules are needed that assign numerical values or functions to the initial (prior) probabilities that will be manipulated. Bayesian methods are sometimes said to be especially subjective because they depend on prior distributions. However, in most physical problems scientific judgment is required to decide on a model, to specify the likelihood, and to take into account further available information

(e.g., positivity of parameters). Indeed, one of the advantages of Bayesian analysis is that it not only explicitly admits the existence of prior information but also tries to exploit it as much as possible. In other types of analysis it is often not easy to recognize the specific assumptions made by the analyst and (even worse) the implicit assumptions of the method (the latter assumptions are often unknown to the average practitioner). Prior information can consist of numerical values for the maximum, width, or moments as mean or variance. Alternatively prior information can consist of properties which are expected for the posterior distribution of a problem. The prior probability distribution should reflect the state of knowledge about the relevant parameters before the current experiment is analyzed. This prior distribution is modified by the new experimental data through the likelihood function and yields the posterior distribution, thus representing the state of knowledge in light of the new data. This posterior distribution can (and should) be used as prior distribution for further measurements. It may be tempting to use this modified prior distribution for a reanalysis of the same data. This will lead to a wrong (too narrow) posterior distribution (Sivia, 2006). The most promising approach in physical data analysis problems is the comprehensive elicitation of available expert knowledge as basis for prior distributions [see Oakley and O'Hagan (2007) and references therein on methods of how to construct prior distributions that represent the expert's belief]. In addition there are several guiding principles to derive a prior distribution.

1. Transformation invariance

E. T. Jaynes, 1968 [but also others, see, e.g., Kendall and Moran (1963) and Harney (2003)] applied group-theoretical methods to the problem of assigning priors. He demonstrated for a number of simple but practically important cases that, even if one is completely ignorant about the numerical values of the estimated parameters, the symmetry of the problem often determines the prior unambiguously. Prominent examples are priors for scale parameters, location parameters, or even priors for hyperplanes which are essential for Bayesian neural networks (Dose, 2003b; von Toussaint, Gori, and Dose, 2004b). However, in most cases these priors are not proper probability distributions and the derivation requires care to avoid pitfalls [see, e.g., Harney (2003), Chap. 12].

For concreteness the specific case of a prior for straight lines through the origin $y = ax$ is considered in more detail. A possible, naive prior for the slope of the straight lines would be $p(a|I) = \text{const}$. On the other hand, the only sensible transformation of the coordinate system is in our specific case a rotation. $p(a|I)da$ is then an element of probability mass which must be independent of the system of coordinates that is used to evaluate its numerical value. Hence, for a different system of coordinates a' it must be required that

$$p(a|I)da = p(a'|I)da', \tag{19}$$

yielding the functional equation

$$p(a|I) = p(a'|I)\left|\frac{\partial a'}{\partial a}\right|, \tag{20}$$

where the determinant on the right-hand side is the Jacobian of the transformation $a \to a'$. Since any finite transformation $a \to a'$ can be constructed from an appropriate sequence of infinitesimal transformations it is sufficient to consider those. Denoting the infinitesimal transformation which maps a onto a' as $T_\epsilon(a)$ it is possible to rewrite the functional equation (20) as a partial differential equation (Dose, 2003b):

$$\frac{\partial}{\partial \epsilon}\left\{p[T_\epsilon(a)]\left|\frac{\partial T_\epsilon(a)}{\partial a}\right|\right\}_{\epsilon=0} = 0. \tag{21}$$

In the case of the linear relation $y = ax$ and invariance under rotation of the coordinate system this differential equation is solved by

$$p(a|I) = \frac{1}{\pi}\frac{1}{1+a^2}, \tag{22}$$

which is not an obvious prior for the slope. But a visualization of both priors (Fig. 3) shows that the prior Eq. (22) is in agreement with our intuition: The angles of the straight lines with the x axis are equally distributed (Fig. 3, right-hand panel), whereas the constant slope prior strongly favors large angles (Fig. 3, left-hand panel) (Dose, 2002).

If a location parameter is to be estimated, for instance, the mean μ of a Gaussian, the prior must be invariant under a shift b of the location. The solution of Eq. (21) is in this case

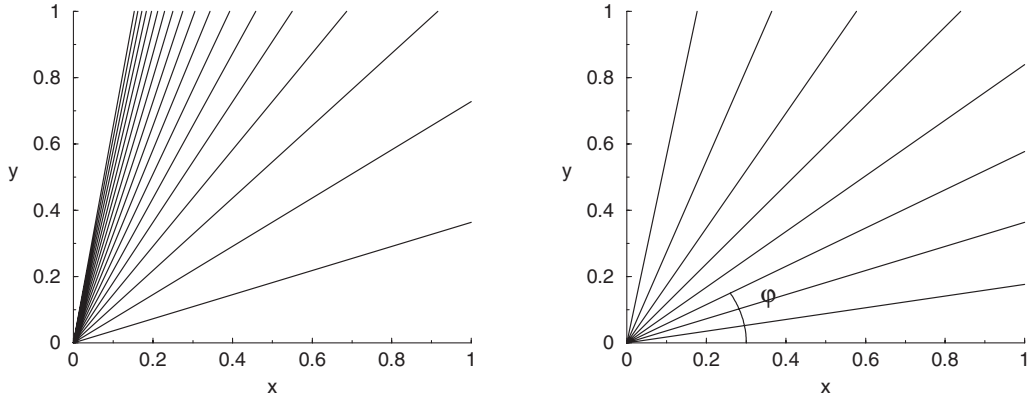

FIG. 3. Left panel: The density of the slope is constant. This results in a nonuniform distribution for the angle between the straight lines and the x axis. Right panel: The angular density is kept constant. Adapted from Dose, 2002.

a constant prior $p(\mu|I) = \text{const}$. If we are ignorant about a scale parameter σ such as the decay length of an exponential or the width of a Gaussian, the appropriate prior satisfying transformation invariance is Jeffreys's prior (Jeffreys, 1961)

$$p(\sigma|I) \propto 1/\sigma. \tag{23}$$

Both priors $p(\mu|I) = \text{const}$ and $p(\sigma|I) \propto 1/\sigma$ are called improper because they are not normalizable on their respective supports $-\infty < \mu < \infty$ and $0 \leq \sigma < \infty$. Improper priors can lead to paradoxes and should not be used. In addition, Bayesian model comparison depends on the use of proper priors in almost all cases. However, in some situations improper priors simplify equations and allow for analytical solutions. Therefore, the common procedure in most physical applications is to consider, e.g., Jeffreys's prior as the limit of properly normalized priors on the support $1/B \leq \sigma \leq B$:

$$p(\sigma|B, I) = \frac{1}{2\ln B}\frac{1}{\sigma}. \tag{24}$$

Inferences from the posterior are then considered for $B \to \infty$. If the inference depends on B in this limit, the improper prior Eq. (23) can lead to inconsistencies (indicating that the information provided by the data is insufficient to enforce a proper posterior) and the whole problem must be reassessed. For other approaches to handle improper priors see, e.g., Harney (2003) (cf. Sec. II.E).

2. The maximum entropy principle

A principle-based approach for coding numerical information into prior probability densities is the maximum entropy (ME) principle (Jaynes, 1957a, 1957b; Kapur and Kesavan, 1992). It is a rule for converting certain types of information, called testable information, to a probability assignment. The information $Q(\boldsymbol{\theta})$ is testable if, given a probability distribution $p(\boldsymbol{\theta}|M, I)$, it can be determined unambiguously whether or not the distribution $p(\boldsymbol{\theta}|M, I)$ is consistent with the information $Q(\boldsymbol{\theta})$. Q may be the already mentioned maximum or mean of a distribution. But, in general, there may be many distributions consistent with the given testable information Q. For example, it may be known that the mean value of many rolls of a (biased) die was 2.5 and we want to use this knowledge to assign probabilities to the six possible outcomes of the next role of the die. This information is testable; one can calculate the mean value of any probability distribution for the six possible outcomes of a roll and see if it is 2.5 or not, but it does not single out one distribution. The basic idea is to choose the prior probability distribution that is compatible with the given information yet has minimal information content otherwise. A function satisfying this requirement is the entropy

$$S = -\sum_i p_i \ln p_i, \tag{25}$$

subject to the constraining information. For continuous probability distributions the entropy is given by

$$S = -\int dx p(x|I) \ln \frac{p(x|I)}{m(x|I)}, \tag{26}$$

where $m(x|I)$ represents the invariant measure required for a proper transformation property of the entropy under coordinate transformations. $m(x|I)$ is sometimes used as a "default distribution" and is often chosen to be uniform $m(x|I) = \text{const}$ (Harney, 2003; Jaynes and Bretthorst, 2003). If the only information at hand is that the probability distribution is normalized to 1 in an interval $[a, b]$ then the ME principle provides a uniform distribution over the interval,

$$p(\theta|Q_0 = 1, M, I) = 1/(b - a). \tag{27}$$

If additionally the expectation value θ_0 of the distribution is given, then the most uninformative distribution for positive variables $0 \leq \theta < \infty$ compatible with those constraints is

$$p(\theta|Q_0 = 1, Q_1 = \theta_0, M, I) = \frac{1}{\theta_0}\exp\left(-\frac{\theta}{\theta_0}\right). \tag{28}$$

As a final example assume that the point estimate θ_0 of θ and also its variance $\langle \Delta\theta^2 \rangle = \sigma^2$ are known. In this case maximum entropy selects as the least informative distribution a Gaussian in $-\infty < \theta < \infty$:

$$p(\theta|Q_0 = 1, Q_1 = \theta_0, Q_2 = \sigma^2, M, I) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(\theta - \theta_0)^2\right). \tag{29}$$

For recent research on the connections between Bayesian inference and the maximum entropy principle see, e.g., Caticha (2008).

3. Reference priors

A different approach for deriving priors has been pursued by Bernardo (1979b). He defined a *reference prior* in such a way that the contribution of the data to the resulting posterior is maximized, i.e., the prior is designed as noninfluential as possible (Bernardo, 2005). For a formal definition of reference priors in terms of a limiting process see, e.g., Bernardo and Smith (2000), p. 307. There are, however, mathematical and philosophical difficulties with this approach (Cox, 2006), e.g., the reference prior may depend on the order in which a set of nuisance parameters is considered or which a parameter is of primary interest (Berger and Bernardo, 1992). As detailed by Bernardo and Smith (2000), reference priors should be considered as a mathematical tool. However, within the statistics community the use of reference priors is common (Bernardo, 2005). To a certain extent the reference prior approach is disposing one of the main features of the Bayesian probability theory, the possibility to incorporate available information. Therefore, the unreflected use of reference priors in the statistics community has been criticized (D'Agostini, 1999; O'Hagan, 2006).

E. Research on foundations of Bayesian inference

Cox (1946, 1961) proved in 1946 that the rules of probability constitute the only consistent extension of ordinary logic in which degrees of belief are represented by real numbers (Bernardo and Smith, 2000). As a consequence all conditional distributions must be proper and normalized (Harney, 2003)

$$\int dx p(x|y, I) = 1. \tag{30}$$

Therefore, Jaynes and Bretthorst (2003) recommended that one approach non-normalizable priors only as a well-defined limit of a sequence of proper priors. This advice is strongly supported by several others, e.g., Dose, 2003a; MacKay, 2003; P. Gregory, 2005; Sivia, 2006). However, given the fact that generally priors based on group invariance arguments as well as most reference priors are improper, one of the current research areas is the extension of Bayesian inference to handle improper priors [see, e.g., Harney (2003) and Bernardo (2005)], which typically requires a measure theoretic approach.

Progress is also made with respect to the extension of Cox's technique to other algebras, which emphasizes as a common concept *partially ordered sets* (Knuth, 2004; Skilling, 2010), generalizing the concept of inference.

Another active research topic within the statistics community is the (re-)analysis of frequentist procedures from a Bayesian point of view, which sometimes reveals interesting relationships and insights [see, e.g., Berger *et al.* (1994) for a comparison of posterior probabilities and conditional type I frequentist error probability or Berger *et al.* (1997) for hypothesis testing].

Also in focus is research about the robustness of model comparison results to model (mis-)specifications. In the Bayesian framework model comparison is conceptually simple. However, the assumption that the correct model is a member of the compared models need not be correct. Here ideas from frequentist significance tests are scrutinized. Along the same lines, there is active research about the effect of prior or likelihood (mis-)specifications on the derived posterior, especially in many dimensions (O'Hagan and Berger, 1988; Hagan and Le, 1994; Bernardo and Smith, 2000).

III. PARAMETER ESTIMATION

A. Introduction

The Bayesian formalism has been known for more than two centuries and it is extensively used in many fields such as robotics (Russell and Norvig, 2003), astronomy (Gregory, 1999), and geology (Malinverno and Briggs, 2004; Tarantola, 2005; Gallagher *et al.*, 2009). The routine use of Bayesian methods in the analysis of physics data, however, is still to come (Fröhner, 2000; Dose, 2003a). The formalism is simple and matches common sense, only the application is sometimes computationally demanding. The general problem of Bayesian parameter inference can be decomposed in several well-defined tasks. First, a model equation $f(\boldsymbol{\theta})$ is needed which relates the parameters $\boldsymbol{\theta}$ to the ideal (undisturbed) signal $s = f(\boldsymbol{\theta})$. This model equation incorporates the physical knowledge about the system and is assumed to be correct. The next step is to construct the likelihood function for the measurement. If the true parameters are known for a deterministic model, then the difference between the measured data d and the signal s would be just noise. The noise distribution depends on the experiment. For example, measurements with Gaussian noise will result in a normal distribution, while counting experiments will have a Poisson distribution as likelihood. Finally the prior distribution for the parameters has to made explicit and should summarize the state of knowledge about the parameters before considering the result of the new measurement. The posterior distribution of the parameters is subsequently obtained by applying Bayes' theorem [Eq. (4)]. This well-defined way of tackling parameter estimation problems is illustrated in the following by recent examples chosen from different areas of physics.

B. Case studies

The case studies have been selected to cover several typical data analysis problems. Arguably the most common task in physical data analysis is the estimation of parameters in nonlinear models (Dose, 2003a). A prototypical example is given by the analysis of radial velocity data in extra-solar planet search, which is often also constrained by limited observation time. The following example outlines a possible approach to detect change points in data sets, a common topic in time-series analysis. The use of a non-Gaussian (i.e., Poisson) likelihood is demonstrated in an example based on a study of neutrino-antineutrino oscillations. Finally, the results of Bayesian inference applied to ill-conditioned linear and nonlinear deconvolution problems are given.

1. Extra-solar planet search

About 15 years ago the first confirmed discovery of a planet orbiting a main-sequence star outside of the solar system was made (Mayor and Queloz, 1995). Since then spectroscopic high-precision radial velocity measurements of small velocity fluctuations in the movement of stars have led to the detection of many new exoplanets. However, the properties of a system with exoplanets are still an area of active research (Baraffe *et al.*, 2010; Gregory and Fischer, 2010). One of the debated issues is the mass distribution of extra-solar planets (Chambers, 2010) because the radial-velocity method is most sensitive to large planets on small orbits resulting in a severe observation bias. The detection of small planets is much more challenging and severely affected by the low signal-to-noise ratio. Therefore, besides the technological advances in present and future missions [e.g., KEPLER (Koch *et al.*, 2010)], improved detection algorithms are of central importance. P. C. Gregory (2005) analyzed radial velocimetry data of the extra-solar planet HD 73526 (Tinney *et al.*, 2003) using a Bayesian parameter estimation. The data set is displayed in Fig. 4. Note the sparse, nonuniform sampling of the data, with a minimum sample interval of ≈ 1 day and an average sample interval of 73 days.

The starting point is the model equation for the radial velocity f_i, which involves six unknowns:

$$f_i = V_0 + K\{e\cos\omega + \cos[\phi(t_i + \chi P) + \omega]\}, \tag{31}$$

where V_0 is a constant velocity, K is the velocity amplitude, P is the orbital period, e is the orbital eccentricity, ω is the longitude of the periastron, and χ is the fraction of an orbit, prior to the start of data taking, at which periastron occurred. The conservation of angular momentum allows one to express ϕ as function of the other parameters

$$d\phi/dt = 2\pi[1 + e\cos\phi(t_i + \chi P)]^2/P(1 - e^2)^{3/2}. \tag{32}$$

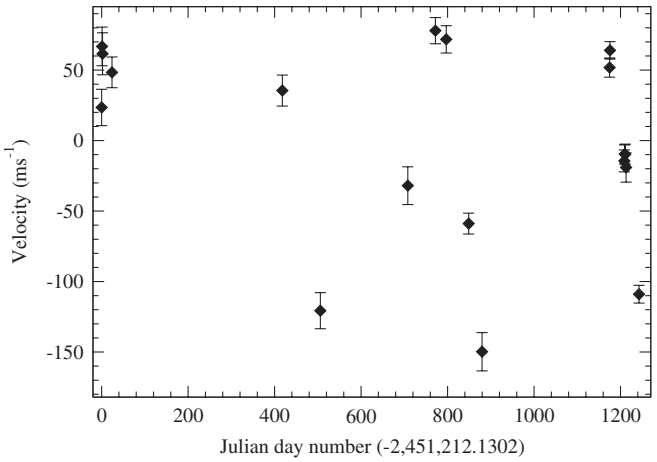

FIG. 4. HD 73526 radial velocity measurement plotted from data given by Tinney *et al.* (2003).

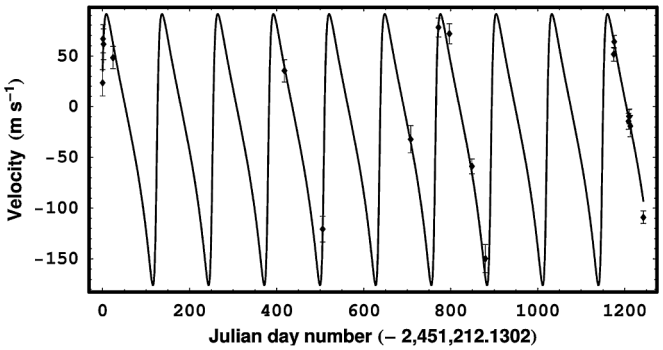

FIG. 5. HD 73526 radial velocity measurement data superimposed with the best-fit model radial velocity. Adapted from Gregory, 2005b.

The measured velocities v_i are related to the model prediction f_i by

$$v_i = f_i + \epsilon_i + \epsilon_{0i}, \tag{33}$$

where the measurement errors ϵ_i have been assumed to be Gaussian with known but unequal standard deviation σ_i. The term ϵ_{0i} accounts for additional measurement errors such as "jitter," which is due in part to flows and inhomogeneities on the stellar surface (Wright, 2005). The distribution of ϵ_{0i} is again assumed to be Gaussian but with a common variance s for all data points $\mathbf{D} = v_1, \ldots, v_N$. With $\boldsymbol{\theta} = \{V_0, K, P, e, \omega, \chi, s\}$ the likelihood distribution is given by the product of N Gaussian, one for each data point

$$p(\mathbf{D}|\boldsymbol{\theta}, I) = (2\pi)^{-N/2}\left[\prod_{i=1}^{N}(\sigma_i^2 + s^2)^{-1/2}\right] \times \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(v_i - f_i)^2}{\sigma_i^2 + s^2}\right]. \tag{34}$$

Bounded, independent, and normalized priors were chosen for the parameters, for example, the prior for the longitude of the periastron was chosen to be uniform in $[0; 2\pi]$: $p(\omega|I) = 1/(2\pi)$. The joint prior is then given by the product of the individual prior distributions

$$p(\boldsymbol{\theta}|I) = p(V_0|I)p(K|I)p(P|I)p(e|I)p(\omega|I)p(\chi|I)p(s|I). \tag{35}$$

The posterior distribution for $\boldsymbol{\theta}$ is obtained using Bayes' theorem

$$p(\boldsymbol{\theta}|\mathbf{D}, I) = p(\mathbf{D}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)/Z, \tag{36}$$

where Z is a normalization constant which can be neglected for parameter estimation purposes. However, this posterior distribution [Eq. (36)] still depends on the additional parameter s, which accommodates additional noise. Application of the marginalization rule finally yields the estimation of the physical model parameters

$$p(V_0, K, P, e, \omega, \chi|\mathbf{D}, I) = \int ds p(\boldsymbol{\theta}|\mathbf{D}, I). \tag{37}$$

In the work of Gregory (2005b) the integration of the parameter space was performed using a parallel tempering Markov chain Monte Carlo algorithm, which is superior to the standard Metropolis-Hastings MCMC algorithm in situations with multimodal distributions (cf. IV.E.3.b) because transitions between different modes are facilitated (Liu, 2001). The best-fit result for a 128-day orbit is shown in Fig. 5, in reasonable agreement with the data. A detailed analysis revealed that three different periods of 128, 190, and 376 days are compatible with the measured data and that the longest period of 376 days has the highest evidence. This conclusion differed from the original result of an orbital period of 190.5 days derived by Tinney *et al.* (2003) and resulted in further investigations of the system. Eventually, with additional data at hand it was discovered that two planets orbit HD 73526 in a 2:1 resonant orbit, one with an orbital period of 377 days and a second one with 188 days (Tinney *et al.*, 2006). The search for extra-solar planets is revisited in Sec. VII.C, where the optimization of observational resources is discussed in the framework of Bayesian experimental design.

2. Change point analysis

The identification of changes in measured data and the extraction of the underlying parameter changes are at the core of many physical investigations. Very often sudden changes indicate the presence of an unknown effect or a transition in the properties of the system of interest. The problem of detecting and locating abrupt changes in data sequences has been studied under the name change point detection for decades, and a large number of methods have been developed for this problem; see, e.g., Carlin *et al.* (1992), Muller (1992), Basseville and Nikiforov (1993), Stephens (1994), Chen and Gupta (2000), and Garnett *et al.* (2009). Common applications are time-series prediction (Garnett *et al.*, 2009), investigation of stock-market trends (Hsu, 1982; Loschi *et al.*, 2008), or analysis of environmental changes (Bradley *et al.*, 1999; Perreault *et al.*, 2000; Zhao and Chu, 2010). But also in areas as diverse as material science (Rudoy *et al.*, 2010), surface physics (von der Linden *et al.*, 1998), and plasma physics (Preuss *et al.*, 2003) the identification of change points is a recurring theme.

The following example from Dose and Menzel (2004) addresses the question if there are indications for changes in the blossom time series collected for several species over

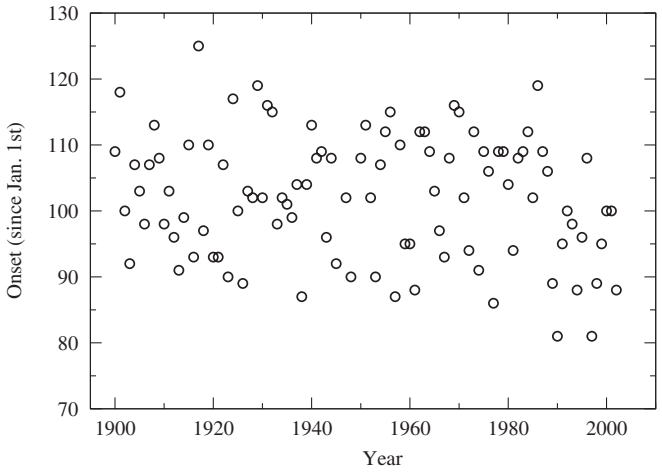

FIG. 6. Observations of the onset of cherry blossoms at Geisenheim, Germany from 1900 to 2002 in terms of days after the beginning of the year. Two observations (1946 and 1949) are missing. Adapted from Dose and Menzel, 2004.

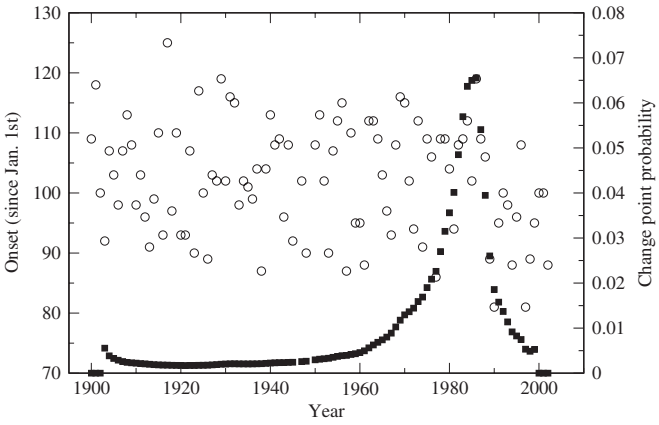

FIG. 7. Data of cherry blossom onset with normalized change point probability (full squares) superimposed (see right vertical scale). Adapted from Dose and Menzel, 2004.

the last century. The observations on cherry blossoms at a specific location (Geisenheim, Germany) are shown in Fig. 6. The cherry blossom (*Prunus* avium L.) is known to flower in midspring and Fig. 6 shows the occurrence of cherry blossoms in terms of days after the beginning of the year. The first impression is that the observations suffer from a considerable scatter and no obvious trend is visible. A possible continuous change-point model consists of piecewise linear sections, the simplest case would be one change point separating two linear sections. This model contains four parameters: the design values at the boundaries and at the change point, and in addition, the parameter that specifies the position of the change point. The model equation for the general case is given by (Dose and Menzel, 2004)

$$y_i = \left(\frac{x_{k+1} - \xi_i}{x_{k+1} - x_k} f_k + \frac{\xi_i - x_k}{x_{k+1} - x_k} f_{k+1}\right), \tag{38}$$

$x_k \leq \xi < x_{k+1}$, which can be expressed in matrix notation as

$$\mathbf{y} = \mathbf{A}(\mathbf{E})\mathbf{f}, \tag{39}$$

where the matrix $\mathbf{A}$ depends on the K change-point positions $\mathbf{E} = \{x_k\}, k = 1, \ldots, K$. Using the maximum entropy principle a Gaussian distribution is derived as an appropriate likelihood for the data $\mathbf{d}$ (Dose and Menzel, 2004):

$$p(\mathbf{d}|\xi, \sigma, \mathbf{f}, \mathbf{E}, I) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left[-\frac{1}{2\sigma^2}[\mathbf{d} - \mathbf{A}(\mathbf{E})\mathbf{f}]^T \times [\mathbf{d} - \mathbf{A}(\mathbf{E})\mathbf{f}]\right]. \tag{40}$$

The prior distribution for σ was chosen as a normalized form of Jeffreys's prior

$$p(\sigma|\beta, I) = \frac{1}{2\ln\beta}\frac{1}{\sigma}, \qquad \frac{1}{\beta} < \sigma < \beta \tag{41}$$

and a constant bounded prior $p(\mathbf{f}|I)$ for the design values $\mathbf{f}$. The prior for the change-point positions $p(\mathbf{E}|I)$ was chosen to be uniform. For the one change point model the probability distribution of the change point location is given by Bayes' theorem

$$p(E|\mathbf{d}, \xi, I) = p(\mathbf{d}|\xi, E, I)p(E|I)/Z. \tag{42}$$

The required distribution $p(\mathbf{d}|\xi, E, I)$ is computed from Eq. (40) using the marginalization rule

$$p(\mathbf{d}|\xi, E, I) = \int d\mathbf{f} d\sigma p(\mathbf{d}|\xi, \sigma, \mathbf{f}, \mathbf{E}, I)p(\sigma|\beta, I)p(\mathbf{f}|I). \tag{43}$$

The normalized probability distribution of E is represented in Fig. 7 by solid squares. The maximum probability for a change point is reached near 1985 and has a value of about 0.07. This result is in good agreement with change point positions computed for several other species (Dose and Menzel, 2004). However, the distribution is very broad, so that every change point position has to be taken into account for subsequent predictions, thus requiring model averaging. For further details about the comparison with other models (no trend, no change point, several change points, etc.) and the estimation of change rates, see Dose and Menzel (2004, 2006).

It should be pointed out that multiple change point estimations usually require extensive use of Markov chain Monte Carlo methods due to the combinatorial increase of possibilities to position the change points (Green, 1995; Zhao and Chu, 2010), rendering an exact summation impossible. However, for a class of problems with certain independence properties between the posterior distributions of individual segments recursive algorithms have been developed with a much better scaling of the computational complexity (Fearnhead, 2006). For the related problem of sequential change point detection see, e.g., Adams and MacKay (2007) and Garnett *et al.* (2009).

3. Counting experiments

In many areas of physics, notably high-energy and elementary particle physics, experiments with small numbers of events are common. If the events obey a Poisson distribution with rate r, the probability of n events in a time interval t is given by

$$p(n|r, t, I) = \frac{(rt)^n}{n!}\exp(-rt). \tag{44}$$

Please note that, although the likelihood contains only the product of r and t, the incorporation of prior knowledge

(e.g., precise knowledge of the measurement time t) allows a separate estimate of r and t using the marginalization rule. In a study of neutrino-antineutrino oscillations (Prosper, 1984, 2007) an on-off experiment was performed. A beam of cold neutrons with effective temperature of 1.5 K was impinging on a 100 μm thick graphite target. By modulating an external magnetic field the amplitude of the putative oscillation effect could be influenced. When the magnetic field was turned on, the oscillation should be suppressed by a factor of a million compared to field-off conditions. By switching back and forth between the different field conditions the background events n_{off} and independently the number of signal plus background events n_{on} could be recorded. For a large number of events the Gaussian approximation for the Poisson distribution is commonly used, yielding a rate for the sum of signal and background $\hat{r} = n_{\mathrm{on}}/t$ with standard deviation $\sigma_r = \sqrt{n_{\mathrm{on}}}/t$. The values for the background measurement are $\hat{b} = n_{\mathrm{off}}/t$ and $\sigma_b = n_{\mathrm{off}}/t$. The signal s can then be estimated as

$$\hat{s} = \hat{r} - \hat{b} \text{ with variance } \sigma_s^2 = \sigma_r^2 + \sigma_b^2. \tag{45}$$

However, in the present case the number of background events was $n_{\mathrm{off}} = 7$, exceeding the number of signal and background events $n_{\mathrm{on}} = 3$, thus leading to negative results for the signal rate s if Eq. (45) is applied. For such a low number of events the Gaussian approximation fails. Prosper (1985, 1988), and Loredo (1992) used a Bayesian approach to estimate the signal rate without Gaussian approximation. In the first step the probability distribution for the background rate is inferred. The likelihood is given by Eq. (44). As a prior distribution for the background rate a uniform prior $p(b|I) =$ const is chosen. Using Bayes' theorem, Eq. (4), the posterior distribution is given by

$$p(b|n_{\mathrm{off}}, t, I) = \frac{p(n_{\mathrm{off}}|b, t, I)p(b|I)}{\int db\, p(n_{\mathrm{off}}|b, t, I)p(b|I)} = \frac{t(bt)^{n_{\mathrm{off}}} \exp(-bt)}{n_{\mathrm{off}}!}. \tag{46}$$

For the on measurement, the joint probability of source and background rate is

$$p(s, b|n_{\mathrm{on}}, t, I) = \frac{p(n_{\mathrm{on}}|s, b, t, I)p(s, b|I)}{p(n_{\mathrm{on}}|t, I)} = \frac{p(n_{\mathrm{on}}|s, b, t, I)p(s|b, I)p(b|I)}{p(n_{\mathrm{on}}|t, I)}. \tag{47}$$

As prior for the signal rate, similar to the background rate, a uniform prior $p(s|b, I) =$ const is chosen. The term $p(b|I)$ encodes our knowledge about the distribution of the background rate. Here the knowledge gained from the background measurement enters [Eq. (46)]. The likelihood is the Poisson distribution for a source with strength $s + b$:

$$p(n_{\mathrm{on}}|s, b, t, I) = [(s + b)t]^{n_{\mathrm{on}}} \exp[-(s + b)t]/n_{\mathrm{on}}!. \tag{48}$$

With these assignments the joint probability distribution $p(s, b|n_{\mathrm{on}}, n_{\mathrm{off}}, t, I)$ can be computed. The marginalization with respect to b

$$p(s|n_{\mathrm{on}}, n_{\mathrm{off}}, t, I) = \int db\, p(s, b|n_{\mathrm{on}}, n_{\mathrm{off}}, t, I) \tag{49}$$

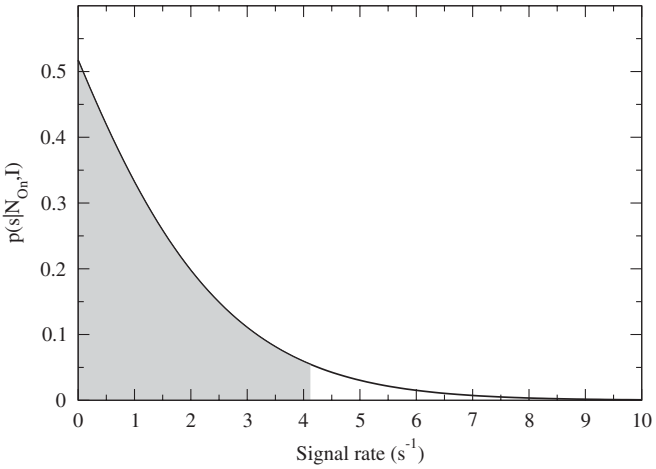

FIG. 8. Posterior density for the signal rate s for $N_{\mathrm{on}} = 3$, $N_{\mathrm{off}} = 7$, and $t = 1$ s and uniform priors. The maximum is for $s = 0$, corresponding to a vanishing signal rate, but the gray shaded 95% credible interval extends up to $s = 4$ s^{-1}.

yields the posterior distribution for the signal s, without explicit dependence on the background rate. Nevertheless, the uncertainty of the background estimation is taken into account. The integration of Eq. (49) is given in detail by P. Gregory (2005) and results in

$$p(s|n_{\mathrm{on}}, n_{\mathrm{off}}, t, I) = \sum_{i=0}^{n_{\mathrm{on}}} C_i \frac{t(st)^i \exp(-st)}{i!}. \tag{50}$$

The posterior probability density of Eq. (50) is shown for the cases of $n_{\mathrm{off}} = 7$, $n_{\mathrm{on}} = 3$, and $t = 1$ s in Fig. 8. Although the number of background events is higher than the number of on events, the low counting statistics does not fully exclude the possibility of a signal. The highest probability of the signal rate is at zero; however, the 95% posterior density region for the signal rate extends up to 4 s^{-1}. Using Eq. (50) the possible gains of measurement extensions can be studied (P. Gregory, 2005). A discussion of various approaches (frequentist and Bayesian) to the background subtraction problem can be found by Loredo (1992).

The related question of source detection in the case of on-off measurements is addressed by P. Gregory (2005). In passing it should be noted that the Poisson distribution does not belong to the class of stable distributions (Feller, 1991): Although the sum of independent Poisson random variables follows again a Poisson distribution [in Eq. (48) advantage was taken of this well-known property], the distribution of the difference is given by a Skellam distribution (Skellam, 1946). Therefore, the use of Eq. (45) for estimation of the difference of independent Poisson distributed measurements may also lead to inferior results. Various approaches to detect and analyze periodic signals are given by Bretthorst (1990a, 1990b, 1990c, 1991), Gregory and Loredo (1993, 1996), Bretthorst (2001), and Gregory (2002).

4. Rutherford backscattering

In the following sections the BPT will be applied to various ill-conditioned inverse problems, encountered in analyzing experimental data from surface physics experiments. The first

example is the deconvolution of Rutherford backscattering (RBS) data which is also of importance for the improved depth resolution of RBS measurements in the subsequent paragraph. Rutherford backscattering is a surface analytical technique which is routinely used to determine surface compositions and depth profiles (Tesmer and Nastasi, 1995). Its importance is derived from its quantitative nature. In RBS, the energy distribution $d(E)$ of backscattered ions is measured for a fixed scattering angle ϕ. In the lower MeV range an elastic Coulomb collision model can be employed, in which the energy of the backscattered ions, usually either protons or helium nuclei, is determined by the incident energy E_0, the scattering angle ϕ, and the mass ratio $r_i = m_0/m_i$ of projectile ion m_0 and target atoms m_i. Since the projectile-target interaction is based on Coulomb interaction, the scattering cross section is the quantitatively known Rutherford scattering cross section and only the mass ratio is unknown. The energy E of the backscattered ions is given by

$$E = E_0\left(\frac{\sqrt{1 - r_i^2 \sin^2\phi} + r_i \cos\phi}{1 + r_i}\right)^2. \tag{51}$$

From Eq. (51) it follows that ions undergoing a collision with a heavy target atom lose less energy than ions colliding with target atoms of lower atomic mass, as long as the projectile is lighter than the target atom ($m_0 < m_i$). In an ideal RBS experiment the energy distribution of an infinitely thin film sample $\tilde{d}(E)$ would be composed of delta peaks for the different masses. However, the resolution is limited due to the apparatus function and finite sample size. In a thick sample both primary ions and scattered ions lose energy on their way through the sample, depending on the stopping power. This enables RBS to be depth sensitive but may also give rise to overlapping peaks in the spectrum.

a. Deconvolution of apparatus functions

Small, cheap, and easy-to-use semiconductor based detectors are used in most RBS experiments for the energy analysis of the backscattered particles. Their performance is hampered by the energy loss straggling in the Au entrance electrode of the detector and in the dead layer of the detector and by the statistics of the electron-hole pair creation. Together with additional contributions to the energy broadening, namely, energy spread of the incident beam, electronic noise of the detector-preamplifier system and, for higher fluxes, pileup the achievable resolution is limited. Therefore, the energy distribution for fixed target mass is rather broad. As long as the masses, or rather the respective backscattering energy distributions, are well separated, it is straightforward to extract the mass composition from the bare experimental data. If, however, the masses are similar, particularly in the case of isotopes, the information is not readily accessible. The different contributions to the energy broadening can be summarized in a transfer function of the whole system, the apparatus function $A(E)$. The measured spectrum $d(E)$ is given by the convolution of the ideal spectrum $\tilde{d}(E)$ with the apparatus function

$$d(E_i) = \int_{-\infty}^{\infty} dE' \tilde{d}(E') A(E_i - E') \approx \sum_{j=1}^{N_d} A_{ij} \tilde{d}(E_j). \tag{52}$$

The matrix A_{ij} represents the discretized apparatus function, taking into account that the measured spectrum is binned. The convoluted spectrum $d(E)$ can be calculated easily if $\tilde{d}(E)$ and A are known. The inversion of Eq. (52) yields the ideal spectrum. Unfortunately, the inversion is frequently utterly ill conditioned if the eigenvalue spectrum of A_{ij} varies over orders of magnitude (von der Linden, 1995). This is generally the case for Gaussian apparatus functions and entails a strong amplification of experimental errors. Dose (2003a) gave an example where a spectrum distortion of the order of 10^{-5} led to reconstruction errors of several 100%. To overcome this problem the statistical nature of the error has to be taken into account properly in conjunction with the intrinsic properties of the objective solution. The goal is to determine the posterior probability density $p(\mathbf{f}|\mathbf{d}, \boldsymbol{\sigma}, I)$ for the RBS spectrum f_j at the N energies E_j, given N_d experimental data d_i and the respective errors σ_i. Prior knowledge to be incorporated is that signal intensity of neighboring channels is usually related: A random permutation of the energy channels results in a spectrum not accepted as RBS spectrum by any expert. However, this does not imply a (statistical) correlation via the likelihood: The noise is independent for each channel. The correlations are imposed on $\mathbf{f}$ through a convolution of a *hidden* density $\mathbf{h}$ with a smoothing kernel B with spatially varying widths. The image f is then obtained from

$$f(x, h, b) = \int dy B\left(\frac{x - y}{b(y)}\right) h(y). \tag{53}$$

Fischer *et al.* (1997) used a Gaussian kernel

$$B\left(\frac{x - y}{b(y)}\right) = \frac{1}{b(y)\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - y}{b(y)}\right)^2\right]. \tag{54}$$

In the Bayesian approach, since the interest is in $\mathbf{f}$, the nuisance parameters $\mathbf{h}$ and $\mathbf{b}$ have to be marginalized,

$$p(\mathbf{f}|\mathbf{d}, \boldsymbol{\sigma}, I) = \int d^N h d^N b\, p(\mathbf{f}, \mathbf{h}, \mathbf{b}|\mathbf{d}, \boldsymbol{\sigma}, I). \tag{55}$$

Bayes theorem relates the yet unknown $p(\mathbf{f}, \mathbf{h}, \mathbf{b}|\mathbf{d}, \boldsymbol{\sigma}, I)$ to known quantities, namely, the likelihood $p(\mathbf{d}|\mathbf{f}, \boldsymbol{\sigma}, I)$ and the prior probability densities $p(\mathbf{h}|I)$ and $p(\mathbf{b}|I)$ via

$$p(\mathbf{f}, \mathbf{h}, \mathbf{b}|\mathbf{d}, \boldsymbol{\sigma}, I) \propto p(\mathbf{d}|\mathbf{f}, \boldsymbol{\sigma}, I) p(\mathbf{f}|\mathbf{h}, \mathbf{b}, I) p(\mathbf{h}|I) p(\mathbf{b}|I). \tag{56}$$

An uninformative prior $p(\mathbf{h}|I)$ for a positive and additive distribution is the entropic prior (Skilling, 1991)

$$p(\mathbf{h}|\mathbf{m}, \alpha, I) \propto \frac{1}{\prod_i \sqrt{h_i}} \exp\left(\alpha \sum_i h_i - m_i - h_i \ln\frac{h_i}{m_i}\right). \tag{57}$$

The default model $\mathbf{m}$ is chosen to be flat and α is a scale parameter for which Jeffreys's prior is used. The prior $p(\mathbf{b}|I)$ constrains the kernel widths to a sensible range. Finally, the probability density $p(\mathbf{f}|\mathbf{h}, \mathbf{b}, I)$ is given by $\delta(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}, \mathbf{h}, \mathbf{b}))$, because the knowledge of $\mathbf{h}$ and $\mathbf{b}$ uniquely determines the value of $\mathbf{f}$. The application of the adaptive deconvolution method is shown in Fig. 9. The spectrum was measured with 2.6 MeV ^{4}He at a scattering angle of 165°. The apparatus function (left peak) was determined by measuring an RBS spectrum of a thin cobalt layer of about 0.75 nm thickness on a silicon substrate (cobalt is isotopically

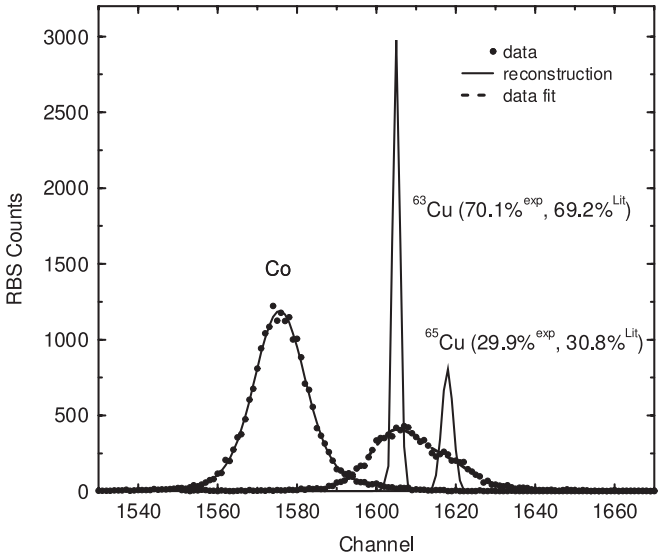

FIG. 9. RBS spectra of thin Co and Cu films on a Si substrate, measured with 2.6 MeV He. The Cu spectrum (right-hand side) is deconvolved with the apparatus function obtained from the Co spectrum (left-hand side). The two Cu isotopes are clearly resolved with measured abundances close to the natural abundances. Adapted from Fischer *et al.*, 1998.

pure). The width of the Co peak is about 19 keV FWHM which reflects the limited resolution since the intrinsic energy spread due to energy loss and energy-loss straggling in the thin Co layer is only about 3 keV. The apparatus function is slightly asymmetric. Using this measured apparatus function $\mathbf{A}$ with its pointwise uncertainty $\boldsymbol{\sigma}_A$ due to the counting statistics, the likelihood function $p(\mathbf{d}|\mathbf{f}, \boldsymbol{\sigma}, I)$ of counting experiments obeys the Poisson statistics. In the case of a large number of counts the Poisson distribution is well approximated by a Gaussian distribution

$$p(\mathbf{d}|\mathbf{f},I)=\frac{1}{\prod_{i=1}^{N_d}\sqrt{2\pi\sigma_{\mathrm{eff},i}^2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{N_d}\frac{(d_i-\sum_{j=1}^{N}A_{ij}f_j)^2}{\sigma_{\mathrm{eff},i}^2}\right] \tag{58}$$

with $\sigma_{\mathrm{eff},i}^2=\sigma_i^2+\sum_{j=1}^{N}\sigma_{A,ij}^2 f_j^2$ (Dose *et al.*, 1998). Using Eq. (55) and the measured apparatus function for cobalt the copper signal on the right-hand side was deconvolved (Fischer *et al.*, 1998). After deconvolution, the two isotopes ^{63}Cu and ^{65}Cu are clearly resolved. The FWHM of the dominant ^{63}Cu peak after deconvolution is 3.0 keV, which is about 6 times better than the achieved experimental resolution and far beyond any conceivable experimental resolution with the available setup. The measured abundances of the isotopes are 70.1% ^{63}Cu and 29.9% ^{65}Cu. This compares favorably to the natural abundance of 69.2 % ^{63}Cu and 30.8% ^{65}Cu.

A different nonparametric approach to ill-conditioned inverse problems, such as Fredholm integral equations with smooth kernels, has been investigated by Wolpert and Ickstadt (2004). Mixtures of Lévy random fields have been used to design prior distributions on functions (Clyde and Wolpert, 2007), leading to increased numerical complexity. However, the large flexibility of Lévy random fields makes up for the computational effort. The approach has been applied to derive molecular weight distributions of polymers from rheological measurements.

b. Depth profiles

Backscattering spectroscopy using ion beams with energies in the MeV range is used extensively to determine the distribution of target elements in the sample as a function of depth below the surface. The ideal RBS spectrum f is a linear superposition of the spectra of the individual elemental depth profiles $c_i(x)$. But the energy of the penetrating and backscattered particles depends in a complicated, nonlinear way on the sample composition and morphology. Therefore simulation codes [such as SIMNRA (Mayer, 1999)] are required to simulate an RBS spectrum for a given sample. The depth profiles are then obtained by varying the sample parameters until a minimum quadratic misfit is achieved (maximum-likelihood solution). Although this approach, guided by the experimentalist's experience, often succeeds it has the severe shortcoming that it does not solve the inverse problem (Mayer *et al.*, 2005). A good fit is a necessary but not sufficient condition: Different depth profiles can result in very similar fits. The posterior expectation (the mean) for the concentration c is

$$\langle c\rangle=\int dc\, c\, p(c|d,I) \tag{59}$$

and the variance is

$$\langle \delta c^2\rangle=\int dc (c-\langle c\rangle)^2 p(c|d,I). \tag{60}$$

The analysis is completely analogous to the one in Sec. III.B.4.a. Only the linear relationship given by Eq. (52) is now replaced by the forward calculation of the simulation codes

$$d'(E_i)=g(c(x)) \tag{61}$$

given a depth profile $c(x)$. A prior which incorporates the knowledge of the concentrations being larger than 0 and allows in addition for the inclusion of a default model is given by the entropic prior (Skilling, 1991).

An example is provided by a study of first-wall materials for fusion experiments. Carbon is considered as a first-wall material for fusion reactors, in particular, for plasma-facing components subject to exceptionally high thermal heat loads. Apart from the lifetime of a material under such conditions, a critical issue in the case of carbon is the possible formation of significant tritium inventories by codeposition with redeposited carbon atoms (Brooks *et al.*, 1999). Both issues are mainly determined by the carbon erosion rate resulting from physical sputtering and chemical erosion (Krieger *et al.*, 1999). To estimate the carbon erosion rates in the divertor of the fusion experiment ASDEX Upgrade graphite probes were covered with a 150 nm layer of ^{13}C and exposed to a single plasma discharge. The RBS spectrum of the sample before exposure is shown in Fig. 10 as the solid line. The right peak indicates the ^{13}C layer on top of the ^{12}C sample. After exposure the high-energy edge is shifted toward lower energies, indicating the absence of ^{13}C at the surface. The increased intensity in the channels at 500 keV indicates a mixture of ^{12}C and ^{13}C, but no further information is easily

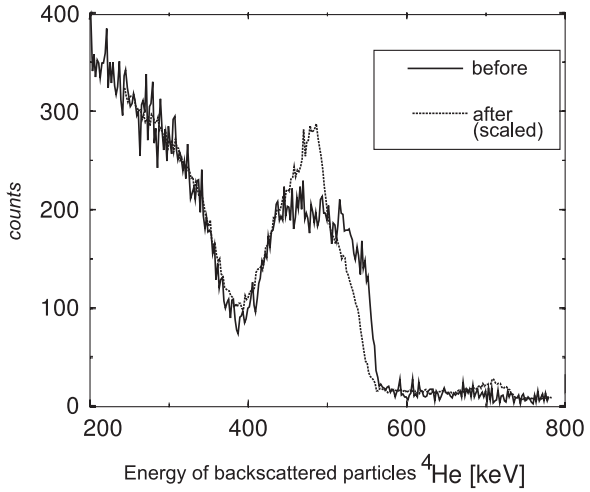

FIG. 10. RBS data of the sample before and after plasma exposition. The signal edge position at around 550 keV is shifted toward lower energies due to the plasma exposition. At the same time the peak intensity increases relative to the bulk signal. From von Toussaint and Dose, 2005.

extracted from the spectrum. The results of the Bayesian depth profile reconstruction are given in Fig. 11 (von Toussaint, Krieger *et al.*, 1999). Before exposure a ^{13}C layer can be seen, $\approx 2.2 \times 10^{18}$ atoms/cm^2 thick, but with an average contribution of 20% ^{12}C. After exposure most of the ^{13}C is still present but there is an additional layer of ^{12}C deposited on top of it. The surprising result is the coexistence of erosion and deposition at the area where the outermost closed magnetic surface intersects the divertor (von Toussaint, Fischer *et al.*, 1999). This so-called ''strike-point'' area experiences extremely high thermal loads

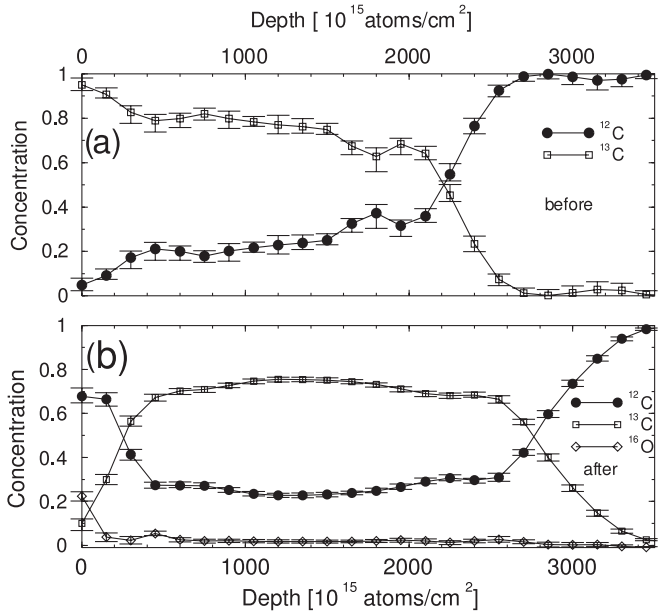

FIG. 11. Reconstructed depth profiles and asymmetric confidence intervals from the RBS spectra shown in Fig. 10. (a) The sample composition before exposure (lines are added to guide the eye), and (b) the sample composition after exposure. From von Toussaint *et al.*, 1999b.

and was considered as erosion dominated. At the same time this measurement shows that conclusions based on net changes in sample thickness may strongly underestimate the dynamical modifications. For further applications of Bayesian parameter estimation in physics see, e.g., Bretthorst (1988, 2001), Dose (2003a), and Meier *et al.* (2003).

IV. NUMERICAL METHODS

A. Overview

Once the likelihood and prior distributions are specified Bayes' theorem, Eq. (4), allows one to derive the posterior probability for every specified parameter vector. However, in most situations the posterior distribution is required primarily for the purpose of evaluating expectation values of a function of interest $f(\boldsymbol{\theta})$ with respect to the posterior,

$$\langle f(\boldsymbol{\theta})\rangle = \int d\boldsymbol{\theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}, I) = \int d\boldsymbol{\theta} f(\boldsymbol{\theta}) \frac{p(\mathbf{D}|\boldsymbol{\theta}, I) p(\boldsymbol{\theta}|I)}{Z}$$
$$= \int d\boldsymbol{\theta} f(\boldsymbol{\theta}) \frac{p^*(\boldsymbol{\theta})}{Z} = \int d\boldsymbol{\theta} g(\boldsymbol{\theta}). \tag{62}$$

The normalization constant of the unnormalized distribution $p^*(\boldsymbol{\theta})$ is given by

$$Z = \int d\boldsymbol{\theta} p^*(\boldsymbol{\theta}). \tag{63}$$

These integrals over the parameter space are commonly high dimensional and analytically intractable, except in very rare circumstances, so that typically neither the expectation value nor the normalization constant are at hand—the latter is the key quantity for Bayesian model comparison, which will be discussed in Sec. V. Also the important marginalization of parameters requires integration in often high-dimensional spaces. There are two different ways to proceed. Either the integrant of Eq. (62) is approximated by a different, more easily accessible function or the integral itself is approximated by numerical integration or by sampling techniques. A note on notation: Throughout this review all probabilities are considered as conditional probabilities (there is always some, however vague, background information). However, in this section the focus is on integration techniques instead of probabilistic inference. Therefore, to keep the notation uncluttered, the I denoting background information is omitted at several places.

B. Approximation methods

First approximation methods are addressed since they provide in many circumstances a fast and convenient way to obtain approximations to the expectation values. If the underlying assumptions hold (which has to be verified), the computations are often fast enough for use in monitoring or even real-time applications.

1. Laplace approximation

The Laplace approximation (also known as saddle-point approximation) substitutes the distribution $g(\boldsymbol{\theta})$ by its asymptotic normal form around its mode $\boldsymbol{\theta}_0$, i.e., the value of $\boldsymbol{\theta}$

maximizing $g(\boldsymbol{\theta})$. Expanding the logarithm of $g(\boldsymbol{\theta})$ around this point yields

$$\ln g(\boldsymbol{\theta}) \approx \ln g(\boldsymbol{\theta}_0) - \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{A} (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{64}$$

where the elements of the $N \times N$ Hessian matrix $\mathbf{A}$ are defined by

$$A_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln g(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{65}$$

Taking the exponential of Eq. (64) provides an N-dimensional Gaussian function

$$g(\boldsymbol{\theta}) \approx g(\boldsymbol{\theta}_0) \exp[-\tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{A} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)] \tag{66}$$

and the analytical integration of Eq. (62) finally yields

$$\langle f(\boldsymbol{\theta}) \rangle \approx g(\boldsymbol{\theta}_0) \sqrt{\frac{(2\pi)^N}{|\mathbf{A}|}}. \tag{67}$$

In order to apply the Laplace approximation the mode has to be localized, typically using a gradient-based numerical optimization algorithm [e.g., the conjugate gradient method (Press *et al.*, 1996)], followed by the evaluation of the Hessian matrix at the mode. The Laplace approximation will be accurate if the data vector consists of a suitably large number of observations such that the central limit theorem applies. However, especially in problems of medium to high dimensionality this is rarely the case and the dependence of the approximation on a single point (the mode) of the distribution may be fatal as generic properties of the distribution may be completely missed. On the other hand, an approximate value of the evidence is easily accessible with the Laplace approximation, a difference to most sampling methods. Extensions to higher order approximations and estimates of the asymptotic accuracy are given by Lindley *et al.* (1980), Tierney and Kadane (1986), and Kass *et al.* (1988). Several special geometries, e.g., spheres are considered by Bagchi and Kadane (1991).

2. Variational methods

Only recently variational methods have been used to approximate complex posterior distributions. The basic idea is to introduce a tractable and flexible parametric test distribution $q(\boldsymbol{\theta}, \boldsymbol{w})$ and to optimize the parameter vector $\boldsymbol{w}$ to provide the best possible approximation to the true posterior distribution. The most commonly used objective function to measure the quality of the approximation is the relative entropy between the test distribution and the (unnormalized) posterior distribution as target distribution

$$\begin{aligned} F(\boldsymbol{w}) &= \int d\boldsymbol{\theta} q(\boldsymbol{\theta}, \boldsymbol{w}) \ln \frac{q(\boldsymbol{\theta}, \boldsymbol{w})}{p(\mathbf{D}|\boldsymbol{\theta}, I) p(\boldsymbol{\theta}|I)} \\ &= \int d\boldsymbol{\theta} q(\boldsymbol{\theta}, \boldsymbol{w}) \ln \frac{q(\boldsymbol{\theta}, \boldsymbol{w})}{p(\boldsymbol{\theta}|\mathbf{D}, I)} - \ln p(\mathbf{D}|I). \end{aligned} \tag{68}$$

Since the relative entropy is never negative the objective function is bounded below by $-\ln p(\mathbf{D}|I)$ and the minimum occurs when the test function equals the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D}, I)$. Although the test function may be of arbitrary complexity (the problem of overfitting does not exist), often factorized distributions are used as test functions (Jordan *et al.*, 1999; Jaakkola and Jordan, 2000; Jaakkola, 2001) to allow for an efficient (convex) optimization of the parameters $\boldsymbol{w}$. A convenient software package (VIBES) for variational inference in Bayesian networks exists (Bishop *et al.*, 2003). It may be tempting to use an optimized test function as a proposal function for importance sampling. However, the optimized test function is usually more compact than the target distribution which is a severe disadvantage for a proposal function.

C. Quadrature

In many problems the dimension of the parameter space is small, i.e., of the order of 1 to 10. In this situation the classical numerical integration, also called quadrature, is often the method of choice for Bayesian computations. This holds especially for the computation of the evidence [cf. Eq. (63)] which is very challenging to estimate with MCMC methods. For the quadrature an extensive literature exists [see, e.g., Davis and Rabinowitz (1984) and Press *et al.* (1996) and references therein]. The one-dimensional integral

$$I = \int d\theta g(\theta) \tag{69}$$

is approximated by a weighted average of the function g evaluated at a number of design points $\theta_{i=1,\ldots,n}$

$$I \approx \sum_{i=1}^{n} w_i g(\theta_i), \tag{70}$$

where the different quadrature schemes are distinguished by using different sets of design points and weights $w_{i=1,\ldots,n}$. Commonly Gauss-Hermite quadrature rules are especially advantageous for integrals of probability distributions, since often $g(\theta)$ is approximately normal and therefore closely approximated by $h(\theta)\exp(\theta^2/2)$, where $h(\theta)$ is a polynomial in θ and the design points and weights of the Gauss-Hermite quadrature are now such that Eq. (70) yields the exact integral if $\exp(\theta^2/2)g(\theta)$ is a polynomial up to order $2n-1$ on the support of $]-\infty, \infty[$. Tables of design points and weights for different quadrature rules can be found in Abramowitz and Stegun (1965). In one-dimensional cases the efficiency of quadrature rules is unsurpassed. The situation changes if the number of dimensions increases. Assuming an integration scheme with n design points in one dimension, the same coverage in m dimensions requires n^m design points, which will be impractically large unless m is sufficiently small ($m \lesssim 10$). This exponential increase in the number of function evaluations with the dimensionality of the problem is often called the curse of dimensions and is the driving force toward Monte Carlo methods. Nevertheless, quadrature, despite suffering from the curse of dimension, is indispensable for several reasons: Well-tested implementations of the various algorithms exist [see, e.g., Press *et al.* (1996), GSL (2008), and NAG (2008)] and reliable error estimates are available. Furthermore, the quadrature algorithms are very robust and can be used to validate MCMC codes.

D. Monte Carlo methods

The idea of Monte Carlo simulation is to obtain (by various means described below) a set of samples $\{\boldsymbol{\theta}^r\}_{r=1}^R$, where the samples are distributed according to $p(\boldsymbol{\theta})$. (The process of generating a random sample according to a probability distribution is commonly called a "draw." Therefore the aim of Monte Carlo simulations is a set of samples drawn from the target distribution.) This allows the expectation value

$$\langle f(\boldsymbol{\theta})\rangle = \int d\boldsymbol{\theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{71}$$

to be approximated by an estimator

$$\hat{f} = \frac{1}{R}\sum_{r=1}^{R} f(\boldsymbol{\theta}^r). \tag{72}$$

The estimator is unbiased and generally $\langle \hat{f}\rangle = \langle f\rangle$. Using the standard definition of the variance σ^2 of $f(\boldsymbol{\theta})$ under the distribution $p(\boldsymbol{\theta})$

$$\sigma^2 = \int d\boldsymbol{\theta} p(\boldsymbol{\theta})[f(\boldsymbol{\theta}) - \langle f\rangle]^2, \tag{73}$$

the variance of the estimator is given by

$$\mathrm{var}(\hat{f}) = \sigma^2/R. \tag{74}$$

Equation (74) is the corner stone of all sampling methods: The accuracy of the Monte Carlo estimate, Eq. (72), does not depend on the dimensionality of $\boldsymbol{\theta}$ (Chen *et al.*, 2001). Therefore, several tens of independent samples may be sufficient to estimate expectation values of a high-dimensional problem with reasonable accuracy. The problem, however, is to get independent samples from $p(\boldsymbol{\theta})$.

1. Standard distributions

If $p(\boldsymbol{\theta})$ is of a standard form then it is straightforward to sample from it using available algorithms [see, e.g., Devroye (1986)[1] or Ripley (1987) for an extensive overview], most of which are based on nonlinear transformations of uniformly distributed random numbers (Press *et al.*, 1996). These methods of generating independent random samples from nonuniform distributions such as Cauchy distribution, student-t, or the famous Gaussian distribution are often used as building blocks in more general strategies as, e.g., rejection sampling or MCMC. However, for the generation of random samples from nonstandard, arbitrary distributions (such as the one in Fig. 12), no algorithms are available.

2. Rejection sampling

Rejection sampling is a conceptually simple method to compute independent samples from a target distribution $p_t(\boldsymbol{\theta}) = p_t^*(\boldsymbol{\theta})/Z_{pt}$, where p_t^* can readily be evaluated and Z_{pt} is unknown. First a proposal density $p_p(\boldsymbol{\theta})$ has to be selected from which independent samples can be generated, the simplest choice being a uniform distribution. For this probability density a constant k has to be determined such

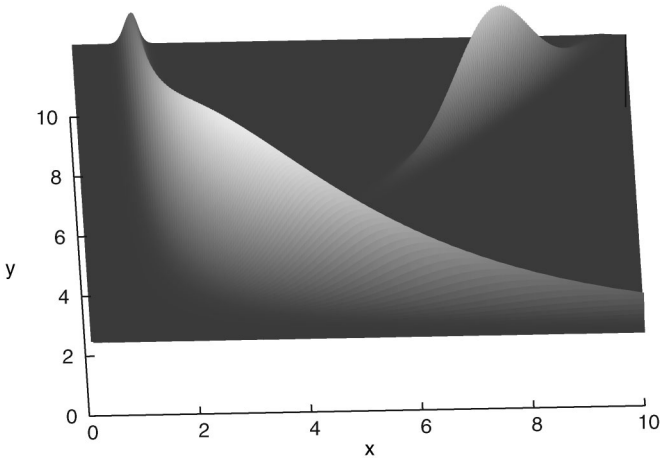

FIG. 12. Example of a two-dimensional probability distribution with properties complicating the sampling: More than one maximum, regions of high probability are separated, non-Gaussian shape, no alignment with the coordinate system.

that $kp_p(\boldsymbol{\theta}) \geq p_t^*(\boldsymbol{\theta})$ for all values of $\boldsymbol{\theta}$ thus forming an envelope to $p_t^*(\boldsymbol{\theta})$. This may be difficult for multimodal and/or multidimensional distributions $p_t(\boldsymbol{\theta})$. A schematic picture of the scaled proposal function and the target distribution is shown in Fig. 13 for the univariate case. Then a random sample $\boldsymbol{\theta}^{(0)}$ is generated from the proposal density $p_p(\boldsymbol{\theta})$. Next a random number u is generated from the uniform distribution [0, 1]. If $ukp_p(\boldsymbol{\theta}^{(0)}) > p_t(\boldsymbol{\theta}^{(0)})$, then the sample $\boldsymbol{\theta}^{(0)}$ is rejected; otherwise it is accepted, which means that $\boldsymbol{\theta}^{(0)}$ is added to the set of samples $\{\boldsymbol{\theta}^r\}$. Rejection sampling has the invaluable advantage to yield independent samples, a feature lacking the various methods introduced below.

A typical sample obtained by rejection sampling is shown in Fig. 14. The density of the sample points matches nicely the shape of the underlying probability density function, shown in Fig. 12.

However, the requirement of $kp_p(\boldsymbol{\theta})$ being an envelope for $p_t^*(\boldsymbol{\theta})$ will generally lead to exponentially decreasing

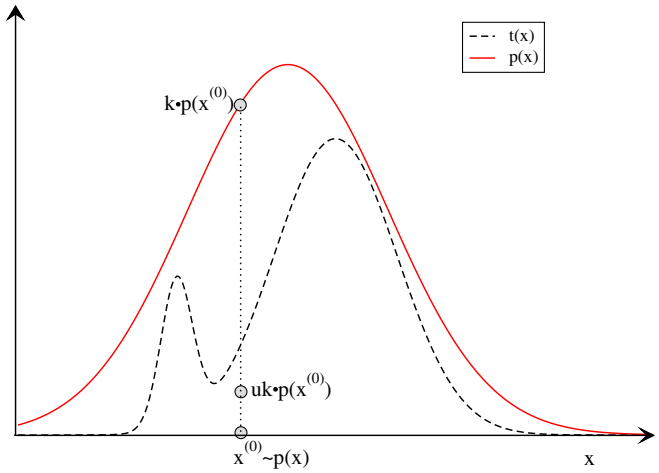

FIG. 13 (color online). Rejection sampling: First, sample a candidate $x^{(0)}$ from $p(x)$ and a uniform variable u. Then accept this candidate if $t(x^{(0)}) \geq ukp(x^{(0)})$ (as in the situation shown here); otherwise reject the candidate.

[1]Available online at http://cg.scs.carleton.ca/~luc/rnbookindex.html.

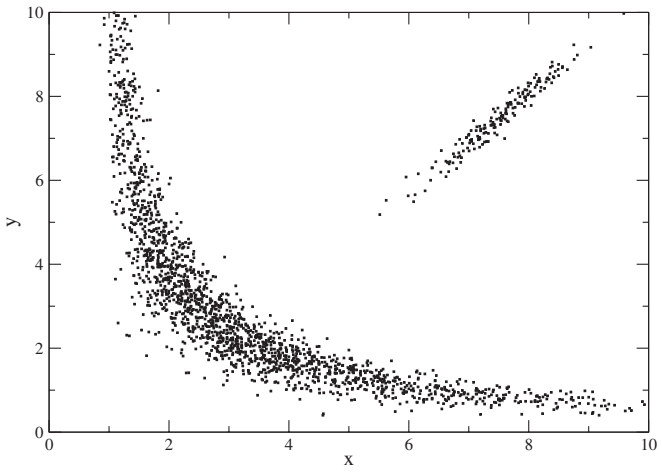

FIG. 14. Sample of 2000 independent samples drawn by rejection sampling from the 2D test distribution shown in Fig. 12.

acceptance probabilities with increasing dimensionality of the problem. As an example consider the problem of generating random samples according to the hypersphere distribution $p_t^*(\boldsymbol{\theta}) = \|\mathbf{r}^2 - \boldsymbol{\theta}^2\|$, for $\|\boldsymbol{\theta}\| < r$ and the enclosing hypercube as a proposal distribution. Then the acceptance probability is given by the ratio of the hypersphere volume V_s and the hypercube volume V_c. As a function of the dimension of the problem this ratio

$$r = V_s/V_c = \frac{2(\sqrt{\pi}r)^d}{d\Gamma(d/2)} \Big/ (2r)^d \tag{75}$$

decays exponentially, e.g., from $\pi/4$ for $d = 2$ to 2.5×10^{-8} for $d = 20$. To improve the match of the envelope function to the target distribution and therefore the acceptance ratio several adaptive methods have been suggested (Gilks and Wild, 1992; Gilks *et al.*, 1995). In the adaptive rejection sampling method (Gilks and Wild, 1992) piecewise exponential distributions are adapted to log-concave target distributions on the fly. In later work the restriction to log-concave distributions was removed (Gilks *et al.*, 1995). Nevertheless, the exponential decrease of the acceptance with the dimensionality of the problem, although alleviated, still persists.

3. Importance sampling

Importance sampling is a useful technique to approximate expectation values [see Eq. (71)] without being able to sample from $p_t(\boldsymbol{\theta})$. As in the case of rejection sampling, importance sampling rests on the availability of a proposal distribution $p_p(\boldsymbol{\theta})$ from which it is easy to sample. Then the expectation value can be approximated by a finite set of R samples

$$\langle f(\boldsymbol{\theta})\rangle = \int d\boldsymbol{\theta} f(\boldsymbol{\theta}) p_t(\boldsymbol{\theta}) = \int d\boldsymbol{\theta} f(\boldsymbol{\theta}) \frac{p_t(\boldsymbol{\theta})}{p_p(\boldsymbol{\theta})} p_p(\boldsymbol{\theta})$$

$$\approx \frac{1}{R}\sum_{r=1}^{R} f(\boldsymbol{\theta}^{(r)}) \frac{p_t(\boldsymbol{\theta}^{(r)})}{p_p(\boldsymbol{\theta}^{(r)})} \tag{76}$$

$$= \frac{1}{R}\sum_{r=1}^{R} f(\boldsymbol{\theta}^{(r)}) w_r, \tag{77}$$

where the quantities w_r are the importance weights, which are used to correct the bias introduced by sampling from $p_p(\boldsymbol{\theta})$ instead of $p_t(\boldsymbol{\theta})$. If $p_t(\boldsymbol{\theta})$ can only be evaluated except for a normalization constant, so that $p_t(\boldsymbol{\theta}) = p_t^*(\boldsymbol{\theta})/Z_t$, and, as is commonly the case, $p_t^*(\boldsymbol{\theta})$ can be evaluated easily, then it is also possible to evaluate Z_t by

$$Z_t = \frac{1}{R}\sum_{r=1}^{R} \frac{p_t^*(\boldsymbol{\theta})}{p_p(\boldsymbol{\theta})}. \tag{78}$$

Similar to rejection sampling importance sampling suffers in high dimensions from any mismatch of the proposal and the target distribution. Then the set of importance weights w_r are dominated by a few weights, thus reducing the sample size (R) effectively to a very small number. Furthermore, there may be the problem that none of the samples generated from $p_p(\boldsymbol{\theta})$ are in regions where $f(\boldsymbol{\theta})p_t(\boldsymbol{\theta})$ is large. Then the estimate of the expectation value may be severely wrong without any visible indication. On the other hand, the concept of importance sampling has the advantage of being extremely flexible: In principle the proposal distribution p_p can be continuously adapted based on the samples obtained thus far and still the method yields an unbiased estimator of the expectation value (Cappé *et al.*, 2004). This has been exploited by a number of algorithms as, e.g., sampling-importance resampling or sequential-importance sampling (Doucet *et al.*, 2001; Moral *et al.*, 2006; Cappé *et al.*, 2008; Cornebise *et al.*, 2008). A convenient implementation of the importance sampling method for general integration is the computer code VEGAS (Lepage, 1980; Press *et al.*, 1996).

However, all methods presented thus far are not suited for problems of high dimensionality. Although in principle Eq. (74) guarantees that a small number of independent samples are sufficient to estimate the expectation value with a reasonable accuracy, the generation of these samples by rejection sampling is impractical for all but modest problems. A way around this problem is offered by MCMC methods.

E. Markov chain Monte Carlo

The first MCMC algorithm was introduced by Metropolis *et al.* (1953) as a method for the simulation of fluids. The physics community immediately noticed the potential of the new algorithm. However, it took about 35 years until the MCMC methods were rediscovered by (Bayesian) statisticians (Hitchcock, 2003) in a series of publications (Tanner and Wong, 1987; Gelfand *et al.*, 1990; Gelfand and Smith, 1990). MCMC techniques immediately enabled the use of complicated (i.e., realistic) models and the estimation of posterior distributions and arbitrary functions of the model parameters without the need for approximations. Since then MCMC methods became the main computational tools in Bayesian inference because of their unique potential in evaluating the multidimensional integrals required in a Bayesian analysis of models with many parameters.

The main difference of MCMC algorithms to the Monte Carlo methods described above is that each generated point depends on its predecessor instead of being independent. Abandoning the independence requirement allows a wide variety of algorithms to be designed with the only

requirement that the distribution the samples are generated from converges to the target distribution. The individual algorithms differ widely in complexity, efficiency, and number of parameters to be adjusted.

1. MCMC basics

a. Markov chains

First the key concepts underlying the MCMC approach are sketched. A first-order Markov chain is a series of random variables $x^{(0)}, \ldots, x^{(T)}$ with the property

$$p(x^{(t+1)}|x^{(0)}, \ldots, x^{(t)}) = p(x^{(t+1)}|x^{(t)}), \tag{79}$$

in which the influence of the values $x^{(0)}, \ldots, x^{(t)}$ on the distribution of $x^{(t+1)}$ is mediated entirely by the value of $x^{(t)}$ (Neal, 1993) A homogenous Markov chain can be specified by giving the probability distribution for the initial variable $p(x^{(0)})$ together with the transition probabilities $T(x', x) = p(x'|x)$ for one state x' to follow another state x. An *invariant* distribution $\hat{p}(x)$ with respect to a Markov chain persists forever once it is reached

$$\hat{p}(x') = \sum_x T(x', x)\hat{p}(x). \tag{80}$$

The Markov chain is said to be *ergodic*, i.e., it converges to its invariant distribution if

$$p^{(t)}(x) \to \hat{p}(x) \quad \text{as} \quad t \to \infty \tag{81}$$

holds regardless of the initial probabilities $p^{(0)}(x)$. Necessary conditions for ergodicity are irreducibility and aperiodicity. Irreducibility guarantees that from any state there is a positive probability of visiting all other states of the Markov chain. The second condition ensures that the Markov chain is not trapped in periodic cycles. MCMC algorithms are based on irreducible and aperiodic Markov chains that have the target distribution $p_t(\boldsymbol{\theta})$ as the invariant distribution. The task of designing such a Markov chain is greatly simplified if the transition probabilities satisfy the *detailed balance* property for the target distribution

$$T(x, x')\hat{p}(x') = T(x', x)\hat{p}(x). \tag{82}$$

It is easily seen that such a transition probability will leave the distribution invariant, because

$$\sum_{x'} T(x, x')\hat{p}(x') = \sum_{x'} T(x', x)\hat{p}(x) = \hat{p}(x)\sum_{x'} p(x'|x) = \hat{p}(x), \tag{83}$$

which is a necessary condition for the convergence of the Markov chain toward the target distribution.

The Metropolis algorithm (Metropolis *et al.*, 1953) and its generalization, the Metropolis-Hastings (Hastings, 1970) algorithm, are the most popular MCMC methods. Most of the later presented algorithms can be considered special cases or extensions of these algorithms. In the basic Metropolis algorithm the proposal function is symmetric $p_p(x'|x^{(t)}) = p_p(x^{(t)}|x')$, quite often a Gaussian (or student-t) distribution centered on the present state of the Markov chain. The candidate sample x' is then accepted with probability

$$A(x', x^{(t)}) = \min(1, p(x')/p(x^{(t)})). \tag{84}$$

If the step is accepted the Markov chain is updated by setting $x^{(t+1)} = x'$, or else by $x^{(t+1)} = x^{(t)}$. Note that if the step from $x^{(t)}$ to x' increases the value of $p(x)$, the candidate sample is always accepted. The target distribution is indeed an invariant distribution of the Markov chain defined by the Metropolis algorithm since the detailed balance criterion is satisfied,

$$\begin{aligned} T(x, x')p_t(x') &= A(x, x')p_t(x') = \min(p_t(x'), p(x')) \\ &= \min(p(x'), p_t(x')) = A(x', x)p_t(x) \\ &= T(x', x)p_t(x). \end{aligned} \tag{85}$$

Metropolis sampling is illustrated for the two-dimensional case in Fig. 15. On each iteration we start from the current state $x^{(t)}$ and a tentative new state $\mathbf{x}' = (x_1, x_2)$ is generated from a two-dimensional random distribution (in this example a 2D Gaussian distribution was used). All samples which are accepted are indicated by a sphere and connected by a solid line, visualizing the exploration of the parameter space. Rejected samples are connected by a dashed line with the state the Markov chain was in when the rejection happened.

The Metropolis-Hastings algorithm uses the generalization

$$A(x', x^{(t)}) = \min\left(1, \frac{p(x')p_p(x^{(t)}|x')}{p(x^{(t)})p_p(x'|x^{(t)})}\right) \tag{86}$$

of the acceptance criterion, Eq. (84), to accommodate also asymmetric proposal functions, giving even more flexibility in the design of the Markov chain. Although the asymptotic convergence of the probability distribution of the samples to the target distribution is guaranteed the rate of convergence depends crucially on the proposal function and the underlying structure of the problem. For more information on the theory of MCMC a wealth of information can be found in, e.g., Neal (1993), Gilks *et al.* (1996), Gamerman (1997), Robert and Casella (1999), and Liu (2001).

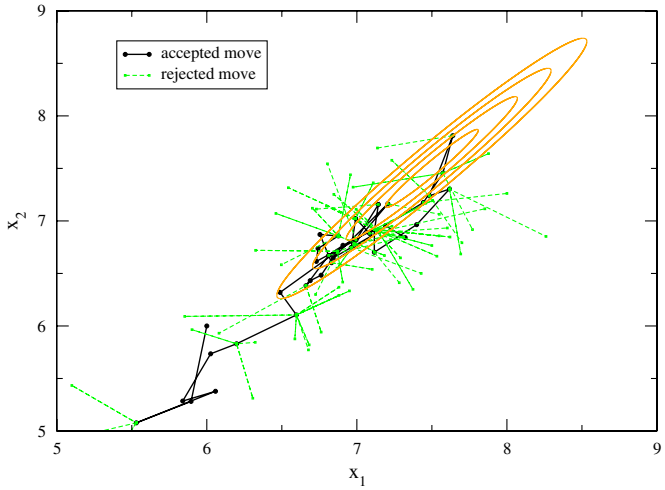

FIG. 15 (color online). Example of a typical MCMC sample, showing indication of random-walk behavior. Solid lines indicate accepted proposals, dashed lines rejected proposals. The underlying potential is the one displayed in Fig. 12, upper right part, and represented by ellipsoidal contour lines.

b. Efficiency

Generally the most crucial choice for the efficiency of MCMC samples is the proposal function. However, some theoretical considerations and experience is available to act as a guideline. A narrow proposal function, suggesting only very small changes of the actual position in the parameter space, has a high acceptance ratio since the ratio of the function values will be close to 1. However, the position of the Markov chain hardly changes, the parameter space is not explored, and the obtained samples are highly correlated. On the other hand, a very broad proposal function will yield an extremely high rejection rate especially in higher dimensions since it is very unlikely to select a suitable position by chance. In both cases the autocorrelation ρ of the samples, defined as

$$\rho(j) = \frac{\sum_t [(x^{(t)} - \langle x\rangle)(x^{(t+j)} - \langle x\rangle)]}{\sqrt{\sum_t (x^{(t)} - \langle x\rangle)^2}\sqrt{\sum_t (x^{(t+j)} - \langle x\rangle)^2}}, \tag{87}$$

decays (much) slower with increasing lag j than for better adapted settings. An example is given in Fig. 16 where the logarithm of the autocorrelation is plotted as a function of the lag for MCMC runs with a Gaussian proposal distribution with three different standard deviations. Typically ρ_j decays exponentially and is therefore well approximated by

$$\rho(j) \propto \exp(-j/\tau_{\text{exp}}). \tag{88}$$

A small autocorrelation time constant τ_{exp} indicates a fast mixing MCMC sampler which also reduces the variance of the expectation value $\hat{f}$ [Eq. (72)] which is given by

$$\text{var}(\hat{f}) = \frac{\sigma^2}{R}\left(1 + 2\sum_{j=1}^{\infty}\rho(j)\right) \tag{89}$$

for a converged Markov chain (Tierney, 1994; Liu, 2001). Please note how the correlation of the samples increases the variance of the expectation value compared to the case of independent samples [Eq. (74)]. Peskun (1973) conducted a comparison of the asymptotic variance of Eq. (89) for a finite state space and found that the Metropolis acceptance rate has the smallest asymptotic variance of the estimates for a given proposal rate. Although this result is reassuring, it does not answer the question about the choice of good proposal functions. Unfortunately those distributions depend very strongly on the problem at hand, so that only very general recommendations are possible: The proposal function should be chosen in such a way that the acceptance rate is between 25% for high-dimensional models and about 50% for models of dimension 1 or 2 (Gelman *et al.*, 1997; Roberts and Rosenthal, 2001). Furthermore, in the case of multimodal target distributions the proposal distribution should exhibit heavy tails (such as a student-t distribution) to facilitate the exploration of the parameter space. There are two fundamental difficulties of designing an effective MCMC sampler.

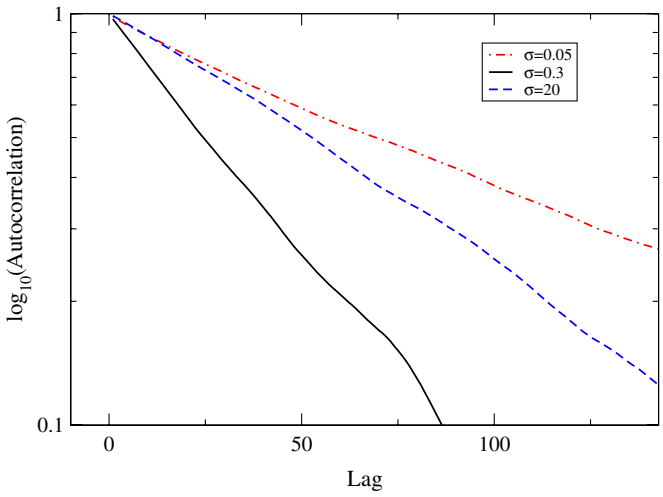

FIG. 16 (color online). Influence of the width σ of a Gaussian proposal function on the autocorrelation of the samples. Too small and too large widths result in a very slow decay of the autocorrelation. If the width of the proposal function is too small then the chain hardly explores the parameter space. If instead the width of the proposal function is too large then almost all proposed changes are rejected again causing insufficient exploration.

The first challenge is due to target distributions with widely varying properties, e.g., localized structures (maxima, ridges) in some regions of the parameter space and broad features in others. In those situations it is unlikely that a fixed proposal distribution is efficiently exploring the parameter space. Unfortunately, adaptive proposal distributions easily destroy the Markov property. In this case the convergence to the desired target distribution is not guaranteed (Gelfand and Sahu, 1994). Nevertheless, several converging adaptive MCMC methods have been developed, including slice sampling (Neal, 2003) (see Sec. IV.E.3.c), parallel chains (Gilks *et al.*, 1996), regeneration (Gilks *et al.*, 1998), delayed rejection (Tierney and Mira, 1999; Haario *et al.*, 2006), and differential evolution Markov chain (Ter Braak and Vrugt, 2008). Overviews of recent developments are given by Andrieu and Thoms (2008) and Roberts and Rosenthal (2009). So far only limited experience with adaptive methods is available. Issues such as robustness or the choice of the adaption strategy for particular contexts still have to be investigated in more detail.

The second reason for inefficiency is the inherent random-walk property of standard-MCMC samplers, which causes a slow exploration of the parameter space [although in some cases properly chosen proposal distributions may provide significant gains efficiency (Wu *et al.*, 2007)]. Here Hamiltonian Monte Carlo (see Sec. IV.E.3.d) and multistate methods are possible remedies.

c. Convergence diagnostic

The estimate of the variance of the expectation value given by Eq. (89) is valid only if the Markov chain has achieved stationarity; that is, if it samples from the target distribution. Unfortunately, for realistic problems no fail-safe convergence diagnostics exist. For that reason it is wise to have several complementary diagnostics on hand to decrease the chances to be fooled. Reviews of more than one dozen different convergence diagnostics are given by Brooks and Roberts (1999), Cowles and Carlin (1996), Mengersen *et al.* (1999), and Robert and Casella (1999). Among the most revealing diagnostics are trace plots: plots of some function of the state of the Markov chain (e.g., one of the parameters) against the iteration number. In Fig. 17 two typical traces are displayed.

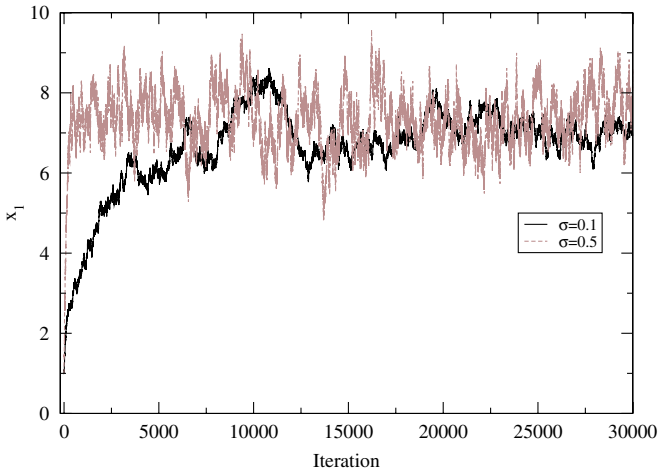

FIG. 17 (color online). Influence of the width of the proposal function on the burn-in and convergence of the Markov chains. The simulation with a narrow proposal function (solid black line) exhibits a larger autocorrelation length of an arbitrarily chosen parameter x_1 and a slower convergence to the stationary distribution compared to the simulation with a 5 times larger proposal width.

The dashed line indicates a chain which appears to have reached stationarity after 5000 iterations (which is confirmed by the result of a much longer simulation run). The other Markov chain (solid black line) is slowly increasing up to 10 000 iterations, indicating a long burn-in time which is a consequence of the narrow width of the proposal distribution. In practice a number of different chains are run starting at different initial values which increases the probability to cover relevant regions of the stationary distribution and to detect insufficient mixing of the chains. In the latter case one (or more) of the chains often displays a different behavior compared to the other chains, although seemingly having achieved stationarity given the properties of this chain alone. This may be the case if there are several well-separated local modes of the target distribution. Based on this observation, Gelman and Rubin (1992) proposed a popular and simple convergence diagnostic which compares the variance within each chain and the interchain variance, often referred to as the $\hat{R}$ statistic (Gelman *et al.*, 2004). However, if the exploration of the parameter space is insufficient, false results of the convergence diagnostic routines are inevitable [see, e.g., Sollom *et al.* (2009) for a recent example]. Early criticism of the multiple chain approach (Geyer, 1992), advocating the advantages of a single, very long MC run, no longer seem compelling in view of today's parallel computing capabilities.

2. MCMC methods I: Standard-MCMC algorithms

a. Gibbs sampling

In the statistical community the Gibbs sampler (Geman and Geman, 1984; Gelfand *et al.*, 1990), a special case of the Metropolis-Hastings sampler, is the most widely used method. It is only applicable if sampling from all the one-dimensional conditional distributions of the target distribution is feasible or can be emulated by rejection sampling methods (Gilks, 1992). The algorithm starts in an N-dimensional parameter space with an arbitrary starting point $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_N^{(0)})$ and the sequence of random points is generated by repeated cycling through the steps

$$\begin{aligned} x_1^{(t+1)} &\propto p_c(x_1|x_2^{(t)}, x_3^{(t)}, \ldots, x_N^{(t)}), \\ x_2^{(t+1)} &\propto p_c(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_N^{(t)}), \\ &\vdots \quad \vdots \\ x_N^{(t+1)} &\propto p_c(x_N|x_1^{(t)}, x_2^{(t+1)}, \ldots, x_{N-1}^{(t+1)}), \end{aligned} \tag{90}$$

where $p_c(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N)$ denotes the one-dimensional conditional distribution of x_i conditioned on the actual values of all the other variables. The Gibbs sampling method has several advantages: First, every sample is accepted as there is no rejection step. More important is the fact that no adjustable parameters are present. This allows the design of multipurpose computer packages for Gibbs sampling such as BUGS (Bayesian inference using Gibbs sampling) (Thomas *et al.*, 1992), where only the conditional distributions have to be specified to get started. Special care with respect to the ergodicity of the Gibbs sampler is in order if one or more of the conditional distributions are somewhere zero. O'Hagan (1994) gave a simple example where the Gibbs sampler fails to explore the parameter space in that case. Figure 18 illustrates the random-walk behavior of a Gibbs sampler for a two-dimensional correlated Gaussian distribution. As each update of a variable corresponds to a movement parallel to the corresponding coordinate axes only small steps are accepted if the individual variables are strongly correlated, thus yielding a target distribution which is elongated and rotated (misaligned) with respect to the coordinate system of the parameters. This may cause extremely long correlation times of the samples (Matthews, 1993; Belisle, 1998). Sometimes a possible remedy is the introduction of auxiliary variables (Polsen, 1996) or a joint update ("blocking") of highly correlated variables (Jensen *et al.*, 1995). An approach to reduce the random-walk

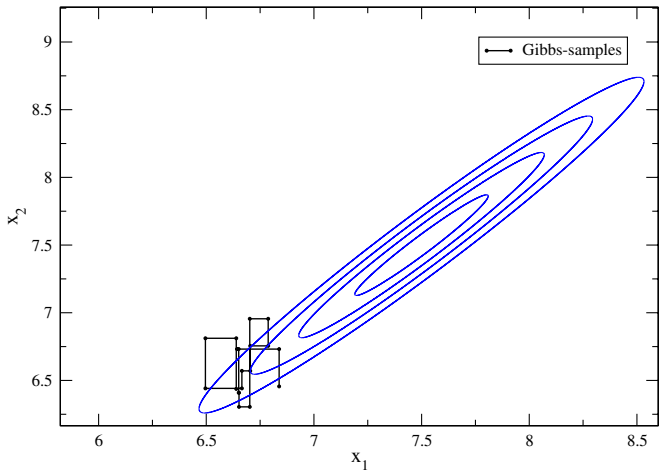

FIG. 18 (color online). Gibbs sampling of a correlated two-dimensional probability distribution. The alternating updates of the two variables are responsible for the "rectangular" movement of the chain. The step size is governed by the respective conditional distributions. Unlike in standard-MCMC methods all proposals are accepted. The underlying potential is the one displayed in Fig. 12, upper right part, and represented by ellipsoidal contour lines.

behavior in order to speed up the exploration of the distribution was suggested by Neal (1999b) with his ordered overrelaxation method. The practical applicability of Gibbs sampling depends on the ease with which samples can be drawn from the conditional distributions. In many settings with likelihoods given by physical models the conditional distributions will be more or less intractable, thus excluding the use of Gibbs sampling.

b. Particle filters: Sequential Monte Carlo

Sequential Monte Carlo methods are algorithms optimized for sampling from a sequence of probability distributions. A typical example is tracking of dynamical systems, where new measurements should be included in the inference process as soon as the data are available. Since the pioneering paper of Gordon *et al.* (1993) introducing *particle filters* as a first instance of sequential Monte Carlo methods, these techniques are now commonly applied in signal processing, robotics, and Bayesian dynamical models. Sequential Monte Carlo methods approximate the sequence of probability distributions of interest using a large set of random samples, named particles. These particles are propagated over time using various importance sampling and resampling mechanisms. The basic scheme proceeds as follows: Initially, at time 0 a population of N samples is created by sampling from a prior distribution,

$$x_{i,0} \sim p(x|0),\ i = 1, \ldots, N. \tag{91}$$

Using a transition model $p(x_{t+1}|x_t)$ each sample is propagated by one step

$$x_{i,t+1} \sim p(x|x_{i,t}),\ i = 1, \ldots, N, \tag{92}$$

where the particle parameters are updated probabilistically according to the transition model. Then all samples are weighted proportional to the likelihood $w_i = p(d_{t+1}|x_{i,t+1})$ and the new population is selected: Each new sample is selected from the existing population. The probability of a sample to be selected is proportional to its weight.

Preserving a sufficient coverage of the probability distribution with a reasonable number of particles is a key issue in sequential Monte Carlo algorithms. Good coverage of sequential Monte Carlo methods and applications is given in several recent review papers (Cappé *et al.*, 2007; Doucet and Johansen, 2008).

3. MCMC methods II: Specialized algorithms

The methods presented in the previous section are efficient general purpose methods (Gilks *et al.*, 1996). However, they exhibit some weaknesses which can be addressed by more specialized methods. The key issues are as follows: multimodality of distributions, variable dimension of the parameter space, adaptive proposal distributions, and faster exploration of the distribution. The most common and severe problem is multimodality of the sampling distribution, i.e., several maxima well separated by regions of low probability. The probability to cross such low-probability regions decays exponentially with their extension. Furthermore, such incomplete sampling of the parameter space is often hard to recognize. If the Markov chains are trapped in an extended but isolated maximum all the convergence diagnostics indicate proper sampling. Only by chance, e.g., by a Markov chain sampling a different maximum or by prior knowledge (e.g., symmetries of the problem), can such problematic behavior be recognized. A (partial) solution of that problem is the use of auxiliary tempered distributions, which facilitate the crossing between different modes of the target distribution. Mainly, there are two different approaches: In the *simulated* tempering approach the parameter space is augmented by an additional tempering parameter β, whereas in the *parallel* tempering algorithm the joint parameter space of a set of temperature altered distributions is used as sample space.

a. Simulated tempering

In order to explore the parameter space more freely than in the standard-MCMC scheme, Marinari and Parisi (1992) and Geyer and Thompson (1995) proposed using a discrete set of progressively flatter versions of the target distribution by varying an additional single parameter, the tempering parameter β, in the target distribution. If exact samples can be generated from the prior distribution then the tempering parameter is commonly applied to the likelihood only:

$$p(\theta|D, \beta_m, I) = p(D|\theta, I)^{\beta_m} p(\theta|I), \tag{93}$$

for $\beta_m \in \{0 = \beta_0, \beta_1, \beta_2, \ldots, \beta_M = 1\}$. For β values close to zero the target distribution is nearly flat, allowing the sampler to escape local modes and therefore increasing its chance of reaching other modes of the distribution. The actual algorithm alternates between a standard-MCMC step in the θ space with constant β_m and moves in the β direction, where $m' = m \pm 1$ is proposed with equal probability and accepted with probability

$$\min\{1, c_{m'} p(\theta|D, \beta_{m'}, I)/c_m p(\theta|D, \beta_m, I)\}; \tag{94}$$

otherwise, m' is set equal to m. The relative frequency between the temperature changing moves and the standard-MCMC steps can be adjusted using an additional parameter α_0. The c_i and α_0 are constants that can be controlled by the user and should be tuned so that each of the β_i distributions has roughly equal chance to be visited (Geyer and Thompson, 1995). A special case arises for $m' = 0$: In this case an independent sample can be generated from $p(\theta|I)$. The time for one exact sample is therefore given by the time needed by the Markov chain to migrate through the set of tempered distribution from β_0 to β_M and back again. The expectation values for the target distribution $p(\theta|D, I)$ are calculated from all the (correlated and uncorrelated) samples with $\beta = 1$, where the correlated samples are from Markov sequences which repeatedly reach $\beta = 1$ before drawing a new and independent sample from the $\beta = 0$ distribution. The independent samples provide an expedient method to compute reliable error estimates (Geyer and Thompson, 1995; Daghofer *et al.*, 2004) of the expectation values. However, the expected waiting time for an independent sample is of the order of M^2 even for the ideal situation of a symmetric random walk in temperature space. Therefore, the number of temperature levels should be kept low while still ensuring a sufficient overlap between two adjacent distributions. The choice of the coefficients c_m is crucial for a reasonable performance of simulated tempering. Ideally, the coefficients c_m should be proportional to the reciprocal of the respective

(unknown) partition function Z_m (Marinari and Parisi, 1992). Since the Z_m often vary over orders of magnitude, appropriate adjustment is mandatory. Use of parallel tempering (see Sec. IV.E.3.b) preruns is often advocated (Geyer and Thompson, 1995; Daghofer *et al.*, 2004) but also iterative schemes have been suggested [see, e.g., Kerler and Rehberg (1994)]. Fiore and da Luz (2010) used the simulated tempering approach to simulate the behavior of a Blume-Emery-Griffiths model near a first-order phase-transition state.

b. Parallel tempering

The parallel tempering technique (Geyer, 1991), reinvented later under the name "exchange Monte Carlo" (Hukushima *et al.*, 1996), is a very efficient method to tackle multimodal probability distributions.

In the parallel tempering method several Monte Carlo simulations ("replicas") are run in parallel in which the posterior distribution [Eq. (93)] is tempered with a series of different temperatures $\beta_m \in \{0 = \beta_0, \beta_1, \beta_2, \ldots, \beta_M = 1\}$. As in simulated tempering the simulation with $\beta = 1$ is the desired target distribution; the other distributions for lower values of β are auxiliary distributions to facilitate an effective exploration of the configuration space. The parallel tempering algorithm alternates between parallel updates of the individual replicas using their respective MCMC scheme, commonly the standard Metropolis-Hasting algorithm, and swap moves. Here a pair of adjacent simulations is chosen at random and a proposal is made to swap their parameter states. Suppose the replicas m and m' are chosen. Then the swap is accepted with probability

$$\min\left\{1, \frac{p(\theta_m|D, \beta_{m'}, I)p(\theta_{m'}|D, \beta_m, I)}{p(\theta_m|D, \beta_m, I)p(\theta_{m'}|D, \beta_{m'}, I)}\right\}. \tag{95}$$

For a visualization of the swapping process, see Fig. 19. These swaps allow for a propagation of information across the simulation chains. The continued suggestion of new configurations from the more freely moving chains with low β allows the chains with larger β values to sample configurations much more efficiently than with local Metropolis updates only.

Contrary to simulated tempering only the number of temperature levels and spacings but no weighting coefficients c_m need to be adjusted: In simulated tempering the weighting coefficients were needed to prevent the chain from getting stuck in a state with intermediate β. In parallel tempering, however, there is a fixed number of chains, one for each β value; thus the system cannot "collapse" toward the most likely value of β. Several suggestions for the number of temperature levels M and the temperature levels β_m are offered in the literature: Rathore *et al.* (2005) found in their case studies that adjusting the number of temperature levels such that an acceptance ratio of 20% is achieved was optimal. Maximizing the mean square displacement of the random walk between the temperature levels Kone and Kofke (2005) arrived at a very similar recommendation of 23%. However, as Katzgraber *et al.* (2006) subsequently pointed out the rate of round trips between low and high β values should be optimized instead.

Parallel tempering has meanwhile proven to be an efficient method for multimodal problems and it is widely acknowledged that the additional workload of running M chains in parallel is more than compensated by the increased sampling efficiency. Parallel tempering can be considered a workhorse in physical chemistry (Earl and Deem, 2005) and data analysis (Habeck *et al.* (2004, 2005) while still being continuously improved (Coluzza and Frenkel, 2005; Bittner *et al.*, 2008).

c. Slice sampling

Slice sampling is a recently proposed Markov chain Monte Carlo method (Higdon, 1998; Neal, 2003) which circumvents the strong dependence of the efficiency of the Metropolis algorithm on the step size providing an adaptive step size adjustment. In Fig. 20 the basic idea is sketched. For a given value $x^{(t)}$ a uniform random value u is sampled

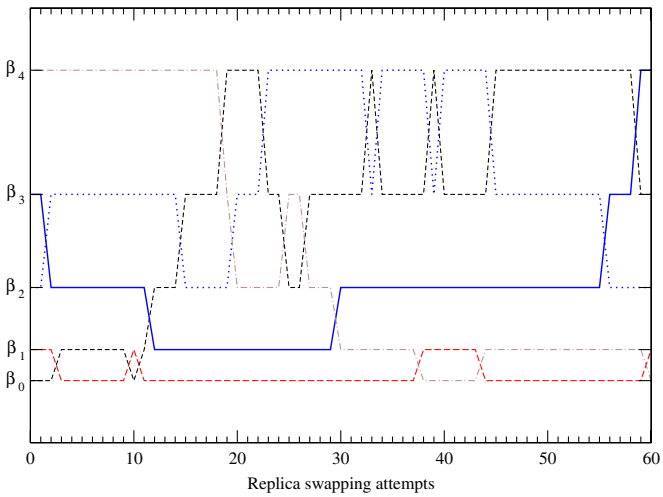

FIG. 19 (color online). Example of a typical parallel tempering run with five replicas at different temperatures $0 = \beta_0 < \beta_1 < \cdots < \beta_4 = 1$. The average time of the replicas to swap from $\beta = 0$ to $\beta = 1$ and back is a measure of the efficiency of a parallel tempering scheme.

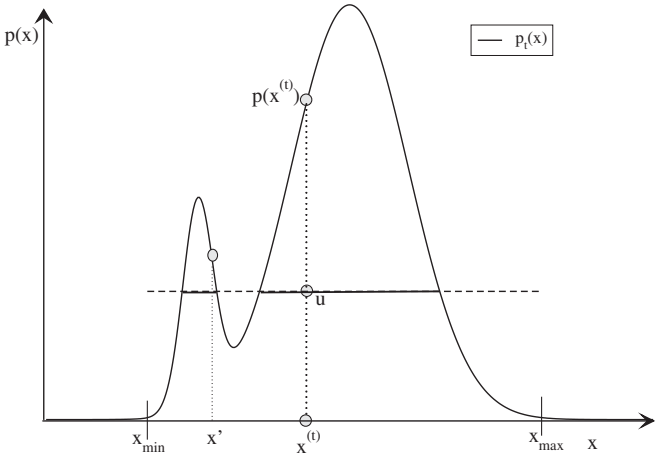

FIG. 20. Schematics of one-dimensional slice sampling: For a given location $x^{(t)}$ a uniform random number u is sampled from $[0, p(x^{(t)})]$ and an interval $[x_{\min}, x_{\max}]$ is determined such that at both locations $\{x_{\min}, x_{\max}\}$ the probability distribution is smaller than u. In this interval a position x' is accepted as a new location if $p(x') > u$.

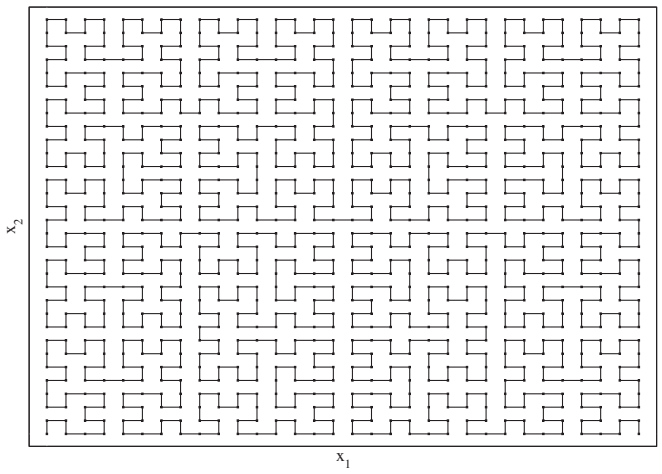

FIG. 21. Example of a two-dimensional Hilbert curve on a 32^2 grid. The space-filling curve can be used to map distributions into a lower-dimensional space with arbitrary precision.

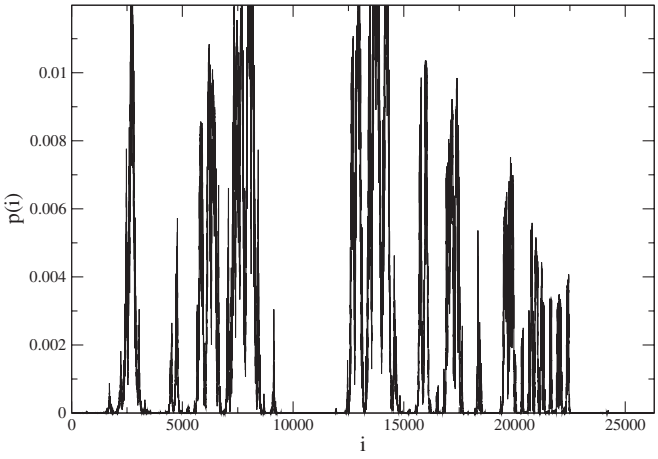

FIG. 22. Part of the two-dimensional probability distribution of Fig. 12 being mapped by a 256^2 Hilbert curve (similar to the one in Fig. 21) into one dimension. The index i labels the points of the discretized 2D Hilbert curve in consecutive order, and $p(i)$ gives the probability at the corresponding location $[x_1(i), x_2(i)]$. The mapping results in a ragged shape of the distribution; however, this is no obstacle for most MCMC algorithms.

between 0 and $p(x^{(t)})$, the value of u defining a horizontal slice which is extended to both sides until $p(x_{\min}) < p(x^{(t)})$ and $p(x_{\max}) < p(x^{(t)})$. Then a new point x' is drawn from $[x_{\min}, x_{\max}]$. If $p(x') < u$ then the slice is shrunk in such a way that x' forms an end point and the original point $x^{(t)}$ is still in the slice. Then another proposal point x' is drawn from this reduced region until a value x' is found for which $p(x') >= u$ and which is therefore accepted as $x^{(t+1)}$. Slice sampling can also be applied to multivariate distributions (Neal, 2003) but loses some of its appeal. A different approach suggested by Skilling and MacKay (2003) may therefore be advantageous: mapping the (discretized) N-dimensional parameter space $\mathbb{R}^N$ on $\mathbb{R}^1$ using Hilbert curves allows one to apply the 1D slice sampler to problems of arbitrary dimension. A low-order example (16^2) of a two-dimensional Hilbert curve is shown in Fig. 21. Using a higher resolution version with (256^2) points the two-dimensional probability distribution of Fig. 12 has been mapped into a one-dimensional probability (Fig. 22). Hilbert curves are as neighborhood preserving as possible but some distortion is unavoidable. A function which is smooth in several dimensions will look jagged when mapped into one dimension. However, if the function is already multimodal and twisted in several dimensions, the mapping does not make it look appreciably worse and the unavoidable discontinuities are no obstacle for most Monte Carlo methods. Several algorithms taking advantage of this fact in combination with adaptive step sizes are given by Skilling (2004b). Slice sampling was used for inference in a geospatial context by Agarwal and Gelfand (2005) and is implemented, e.g., in a software package for Bayesian inference of phylogenies from DNA sequences[2].

d. Hamiltonian Monte Carlo method

Hamiltonian Monte Carlo or hybrid Monte Carlo methods (Duane *et al.*, 1987; Toussaint, 1989; Kennedy *et al.*, 1990; Neal, 1996; Neal, 2011) are Markov chain Monte Carlo methods designed to suppress the random-walk nature of standard Markov chain algorithms by taking into account not only the function value at a given point $\mathbf{x}$ but also its gradient.

Any probability density that is nowhere zero can be written in the form

$$p_t(\mathbf{x}) \propto \exp[-E(\mathbf{x})], \tag{96}$$

which yields the gradient

$$\mathbf{g}(\mathbf{x}) = \nabla \log p_t(\mathbf{x}) = -E(\mathbf{x})/\partial \mathbf{x}. \tag{97}$$

After introducing an auxiliary "momentum" variable $\mathbf{q}$ of the same dimension as $\mathbf{x}$ a Hamiltonian can be defined by

$$H(\mathbf{x}, \mathbf{q}) = E(\mathbf{x}) + K(\mathbf{q}) = E(\mathbf{x}) + \mathbf{q}^T\mathbf{q}/2. \tag{98}$$

Then an extended target distribution $p_t(\mathbf{x}, \mathbf{q}) = p_t(\mathbf{x})p_N(\mathbf{q})$ is introduced. This density is separable, so simply discarding the momentum variables from the obtained samples yields the desired distribution $p_t(\mathbf{x})$. The Hamiltonian Monte Carlo algorithm proceeds by an alternating sequence of Gibbs sampling updates of the momentum from a multivariate Gaussian distribution and a dynamical evolution of the system for a finite time. Starting with the previous value of $\mathbf{x}$ the momentum variable $\mathbf{q}$ is generated from a multivariate Gaussian distribution. Then the dynamical evolution of the system is followed for L time steps of duration ϵ using the leapfrog scheme (which is known to preserve the phase space) for the numerical integration:

$$\begin{aligned} \mathbf{q}(\tau + \epsilon/2) &= \mathbf{q}(\tau) + \frac{\epsilon}{2}\mathbf{g}(\mathbf{x}(\tau)), \\ \mathbf{x}'(\tau + \epsilon) &= \mathbf{x}(\tau) + \epsilon\mathbf{q}(\tau + \epsilon/2), \\ \mathbf{q}'(\tau + \epsilon) &= \mathbf{q}(\tau + \epsilon/2) + \frac{\epsilon}{2}\mathbf{g}(\mathbf{x}(\tau + \epsilon)). \end{aligned} \tag{99}$$

The final candidate state is then accepted with probability

$$A((\mathbf{x}', \mathbf{q}'), (\mathbf{x}, \mathbf{q})) = \min(1, \exp[H(\mathbf{x}, \mathbf{q}) - H(\mathbf{x}', \mathbf{q}')]). \tag{100}$$

[2]http://www.phycas.org.

Rejections can occur only if there are numerical errors; otherwise $H(\mathbf{x}, \mathbf{q})$ is an invariant quantity. In order that the discrete leapfrog integration introduces only a reasonably small error, it is necessary for the step width ϵ to be smaller than the shortest length scale over which the potential is varying significantly. An efficient exploration can therefore be achieved by a reasonably large L. The momentum term within a cycle ensures that a substantial distance is covered, thus suppressing random-walk behavior. A detailed comparison of hybrid Monte Carlo methods with other methods can be found in Neal (1996). See, e.g., Hansmann *et al.* (1996) for the use of Hamiltonian Monte Carlo methods for the problem of protein folding with multiple energy minima. A noteworthy variation of hybrid Monte Carlo methods is based on a suggestion by Horowitz (1991) that partially keeps the momentum variables instead of a full replacement by a Gibbs sampling step will further reduce random-walk behavior (Neal, 1996). For recent attempts to incorporate geometric information of the target density into the Hamiltonian Monte Carlo algorithm see Girolami and Calderhead (2011).

e. Reversible jump Markov chain Monte Carlo (RJMCMC) methods

All the Monte Carlo methods considered so far keep the number of parameters constant and update only the parameter values. In a number of settings, most notably in model selection problems, the number of parameters is unknown. Green (1995) introduced an extension of the conventional Metropolis-Hastings acceptance criterion which allows one to apply the Metropolis algorithm also to parameter spaces of varying dimension or to a set of different models simultaneously (Sisson, 2005). This coupling of different models using RJMCMC has the additional benefit of automatically incorporating Occam's razor (cf. Sec. V.A), a natural preference for "simpler" models (all else being equal).

Given a set of different models M_k, $k = 1, \ldots, K$ with parameter vectors $\mathbf{x}_k$ of dimension d_k the straightforward comparison of densities as in the Metropolis-Hastings case is no longer possible, since the dimensionality of the models varies. The key aspect of the reversible jump approach is the introduction of additional random variables $\mathbf{u}$ that enable the matching of the parameter space dimensions across models. The transition probability is generalized from $T(\mathbf{x}_1, \mathbf{x}_2)$ to $T((\mathbf{x}_1, \mathbf{u}_1), (\mathbf{x}_2, \mathbf{u}_2))$ with the requirement of dimension matching $d_1 + \dim(\mathbf{u}_1) = d_2 + \dim(\mathbf{u}_2)$. An example is given in Fig. 23. The one-dimensional probability density is augmented by an additional random variable u to match the dimension of the two-dimensional model displayed in the upper right part of Fig. 23, so that both models have a common measure.

The key steps of the model changing part of a reversible jump Markov chain are as follows:

(1) Starting from model M_i with parameter vector $\mathbf{x}_i$ with probability $J_{j,i}$ a jump to a new model class M_j is proposed (with still undetermined parameter vector), and an augmenting random variable vector $\mathbf{u}$ is generated from a proposal density $J(\mathbf{u}_i|\mathbf{x}_i, j, i)$.

(2) With that expanded parameter vector $(\mathbf{x}_i, \mathbf{u}_i)$ a parameter vector for M_j is proposed $(\mathbf{x}_j, \mathbf{u}_j) = g_{(j,i)}(\mathbf{x}_i, \mathbf{u}_i)$.

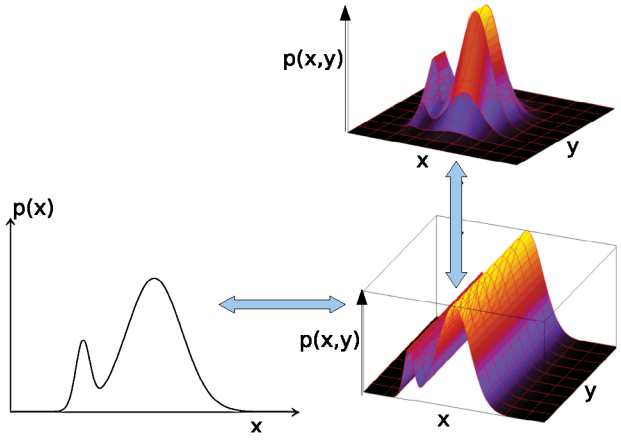

FIG. 23 (color online). Illustration of the dimension matching in RJMCMC. The lower-dimensional parameter spaces are padded with additional dimensions to allow a comparison with the higher-dimensional models.

(3) The proposal is accepted with the probability

$$\alpha = \min\left\{1, \frac{p(y|\mathbf{x}_j, M_j)p(\mathbf{x}_j|M_j)p(M_j|I)}{p(y|\mathbf{x}_i, M_i)p(\mathbf{x}_i|M_i)p(M_i|I)} \times \frac{J_{j,i}J(\mathbf{u}_i|\mathbf{x}_i, j, i)}{J_{i,j}J(\mathbf{u}_j|\mathbf{x}_j, j, i)} \left|\frac{\partial(\mathbf{x}_j, \mathbf{u}_j)}{\partial(\mathbf{x}_i, \mathbf{u}_i)}\right|\right\}, \quad (101)$$

where the final term in the ratio is the Jacobian arising from the change of variables from $(\mathbf{x}_i, \mathbf{u}_i)$ to $(\mathbf{x}_j, \mathbf{u}_j)$.

A good exploration of the individual models is possible only if the Markov chain mixes well between the models. This requires well-chosen jump proposals between the different parameter spaces in addition to the conventional update proposals of the parameter vector of each model. This indicates one of the major difficulties for RJMCMC methods: An efficient construction of reversible jump proposal distributions may be challenging and is complicated by the fact that sometimes a natural neighborhood structure (e.g., Euclidean space) between different models does not exist (Brooks *et al.*, 2003). Nevertheless, especially for nested models RJMCMC is often the most straightforward approach and the application of RJMCMC to mixtures of Gaussians (Richardson and Green, 1997) has been widely adopted. Andrieu and Doucet (1999) presented an interesting application of the RJMCMC algorithm to the detection of an unknown number of sinusoids with low signal-to-noise ratio. Also in inverse problems commonly occurring in geophysics (e.g., profile estimation) RJMCMC is often used (Jasra *et al.*, 2006; Charvin *et al.*, 2009; Gallagher *et al.*, 2009). However, the construction of the proposal densities may sometimes be subtle as a subsequent correction (Richardson and Green, 1998) of the original paper highlights. Another complication is convergence assessment for multimodal Markov chains. As with the unimodal case, useful *a priori* convergence bounds do not exist and most convergence diagnostics assess only necessary indicators of chain convergence (Cowles and Carlin, 1996). Furthermore, even if all the subchains (i.e., the individual models) have converged this does not imply that the full density has converged. Even one result given in

the ground-breaking paper of Green (1995), the change point estimates of the coal-mining time series, was based on simulations which had failed to converge (Green, 2003). Sometimes the symmetry of the posterior distribution provides the possibility of basic control checks. Consider one of the typical applications of RJMCMC: a mixture distribution with an unknown number of components K,

$$p(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k f(x|\boldsymbol{\theta}_k), \tag{102}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, w_1, \ldots, w_K)$ and the non-negative weights satisfy $w_1 + \cdots + w_K = 1$ and the $f(x|\boldsymbol{\theta}_k)$'s are from some parametric family, e.g., the normal distribution with mean μ_k and variance σ_k: $\boldsymbol{\theta}_k = (\mu_k, \sigma_k)$. Since the mixture distribution Eq. (102) is invariant under permutation of the indices k, monitoring of the MCMC samples should reveal a uniform exploration of the $K!$ equivalent modes. As this is rarely the case even for moderate values of K, Celeux *et al.* (2000) concluded "...that almost the entirety of MCMC samplers implemented for mixture models has failed to converge." In addition, the invariance of the posterior distribution under relabeling of some parameters results in the so-called label-switching problem (Redner and Walker, 1984), even for a fully converged Markov chain: The usual practice of summarizing the results by marginal posterior distributions of the individual parameters is often inappropriate due to the multimodality of the joint posterior distribution. The obvious approach to introduce artificial identifiability constraints (Dieboldt and Robert, 1994) on the parameter space $\boldsymbol{\theta}$ such as partial ordering $(\mu_1 < \cdots < \mu_K)$ or relabeling may affect the estimates (Celeux *et al.*, 2000; Stephens, 2000). A review of various approaches to the label-switching problem is given by Jasra *et al.* (2005). Nevertheless, despite these technical challenges RJMCMC is in many instances (i.e., with an unknown number of model parameters) the method of choice.

4. MCMC methods III: Evaluating the marginal likelihood

In Sec. IV.D several algorithms for the computation of expectation values of the form

$$\langle f(\boldsymbol{\theta}) \rangle = \int d\boldsymbol{\theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta}|I) \tag{103}$$

were introduced. These algorithms are adequate for the situation shown in Fig. 24(a), which is commonly the case. A prominent exception is the computation of the evidence, also called prior-predictive value [cf. Eq. (15)]

$$Z = p(\mathbf{d}|M, I) = \int d\boldsymbol{\theta} p(\mathbf{d}|\boldsymbol{\theta}, M, I) p(\boldsymbol{\theta}|M, I) \tag{104}$$

which is often the single most important number in a problem. It represents the probability of the observed data $\mathbf{d}$ given a model M and is the key quantity for model comparison. Comparison of Eqs. (103) and (104) reveals that here the expectation value of the likelihood with respect to the prior has to be computed. Typically the likelihood is much more structured than the prior, so that here the situation shown in Fig. 24(b) applies. Straightforward sampling from the prior $p_p(\boldsymbol{\theta}|M, I)$ is ineligible since the huge variations

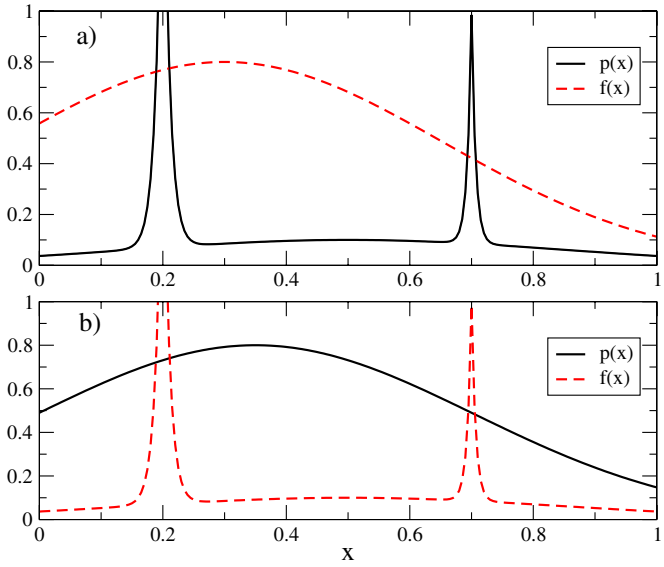

FIG. 24 (color online). Example of the two different cases for expectation value computation as discussed. (a) The probability distribution $p(x)$ displays more structure than the function $f(x)$ of which the expectation value is taken. This is the easier case. (b) The expectation value is computed from a function which is not well matched to the probability distribution, often the case in the evaluation of the marginal likelihood, requiring the use of advanced methods (see Sec. IV.E.3).

in the likelihood lead to large variances [cf. Eq. (74)] and correspondingly to an extremely large number of required MCMC samples. See Fig. 24, lower panel, for a visualization of that case. von der Linden, Preuss, and Dose (1999) discussed a realistic test case which would require 10^{138} independent samples from the prior distribution for an accuracy of 10%. Below several methods are presented which represent different approaches to cope with this problem. It should be noted that sometimes the ratio of the evidence, the Bayes factor, is easier to compute than the individual evidence values. For example, the ratio of the residence time of a RJMCMC run in the different models provides a direct estimation of the Bayes factors. The drawback of this approach is the insufficient exploration of less likely models leading to large uncertainties in the ratio. Recent overviews of methods for the computation of Bayes factors are given, e.g., by DiCiccio *et al.* (1997), Gelman and Meng (1998), and Han and Carlin (2001). In the following a thermodynamic integration is presented as an example for a well-established technique, a nested sampling as a technique using a quite different approach, and finally a promising nonequilibrium technique, highlighting the ongoing development.

a. Thermodynamic integration

In the thermodynamic integration (Frenkel, 1986; Ogata, 1989) [or thin MCMC (Neal, 1993)] an auxiliary quantity $Z(\beta)$

$$Z(\beta) = \int d\boldsymbol{\theta} p(\mathbf{d}|\boldsymbol{\theta}, M, I)^{\beta} p(\boldsymbol{\theta}|M, I) \tag{105}$$

is introduced with $Z(1) = p(\mathbf{d}|M)$ and $Z(0) = 1$ due to the normalization of the prior. The derivative of $\ln Z(\beta)$ yields

$$\frac{\partial}{\partial\beta}\ln(Z(\beta)) = \frac{1}{Z(\beta)}\int \mathrm{d}\boldsymbol{\theta}\,\ln[p(\mathbf{d}|\boldsymbol{\theta},M,I)]p(\mathbf{d}|\boldsymbol{\theta},M,I)^{\beta}p(\boldsymbol{\theta}|M,I) = \langle\ln[p_l(\mathbf{d}|\boldsymbol{\theta},M,I)]\rangle_{\beta}, \tag{106}$$

where $\langle\ln[p(\mathbf{d}|\boldsymbol{\theta},M,I)]\rangle_{\beta}$ is the expectation value of $\ln[p(\mathbf{d}|\boldsymbol{\theta},M,I)]$ with respect to the distribution of

$$p_{\beta}(\boldsymbol{\theta}) = \frac{1}{Z(\beta)}p(\mathbf{d}|\boldsymbol{\theta},M,I)^{\beta}p(\boldsymbol{\theta}|M,I). \tag{107}$$

The reformulation with auxiliary parameter β made the problem more tractable since now only the expectation value of the *logarithm* of the likelihood has to be computed and additionally the probability density $p_{\beta}(\boldsymbol{\theta})$ contains some structure of likelihood as well. From Eq. (107) it follows that the evidence can be obtained by integration over β

$$\ln[p(\mathbf{d}|M,I)] = \ln[Z(1)] - \ln[Z(0)] = \int_0^1 d\beta\frac{\partial}{\partial\beta}\ln[Z(\beta)] = \int_0^1 d\beta\langle\ln[p(\mathbf{d}|\boldsymbol{\theta},M,I)]\rangle_{\beta}. \tag{108}$$

This integral can be approximated by a sequence of $\langle\ln[p(\mathbf{d}|\boldsymbol{\theta},M,I)]\rangle_{\beta_i}$ values for $0=\beta_1<\beta_2<\cdots<\beta_I=1$, where all expectation values are computed by individual MCMC runs. In the example considered by von der Linden, Preuss, and Dose (1999) the required sample size for an estimate of the evidence with 10% accuracy was estimated to be around 10^6 using thermodynamic integration compared to 10^{138} for straight sampling from the prior.

b. Nested sampling

The nested sampling algorithm (Skilling, 2004a, 2006) was developed specifically to compute evidence integrals. The first idea underlying nested sampling is shown in Fig. 25. The multidimensional integral over the parameter space is transformed to a simple one-dimensional integration over the normalized prior mass ξ:

$$Z = \int_0^1 d\xi L(\xi), \tag{109}$$

where $L(\xi^*)$ is the likelihood value such that the volume of the prior where $L \geq L(\xi^*)$ is ξ^*. However, the sorted likelihood function $L(\xi)$ is usually not accessible since the information about the enclosed prior mass ξ of a given sample point θ with its associated $L(\theta)$ is unknown. Nested sampling circumvents this difficulty by replacing the exact values of ξ by estimates of ξ using an iterative random mechanism.

The method is to start with an initial set of N samples uniformly sampled from the full prior mass range $[0, \xi_0 = 1]$, corresponding to no restriction on the likelihood values $L \in [0, \infty[$. The samples are ordered in terms of likelihood and the sample with the smallest likelihood is discarded and its likelihood value L_1 is used as a new lower threshold for the replacement sample, again uniformly sampled from the prior distribution subject to $L \geq L_1$ and thus implying a reduced prior mass range $[0, \xi_1]$. The updated set is ordered again and the new threshold and the sample to be replaced are determined for the next cycle. Since the distribution of the sorted samples is given by order statistics of order N of a uniform distribution, the probability distribution of the shrinkage ratio t_i of each cycle is given by the β distribution

$$p(t_i) = p(\xi_i/\xi_{i-1}) = B(t_i; N, 1) \tag{110}$$

with

$$\langle p(t_i)\rangle = N/(N+1) \tag{111}$$

thus implying geometric shrinkage of the ξ_i with increasing i. Then the evidence can be approximated by the sum

$$Z = \frac{1}{2}\sum_{i=1}^{M} L_i(\xi_{i+1} - \xi_{i-1}). \tag{112}$$

The probabilistic estimation of the ξ_i values used for the summation in Eq. (112) initially raised some concerns about the convergence properties of the method (Chopin and Robert, 2007); in subsequent papers, however, convergence was proven (Evans, 2007; Chopin and Robert, 2008, 2010). Since the introduction of nested sampling, it has found widespread use especially in astrophysics [see, e.g., Mukherjee, Parkinson, and Liddle (2006)] and cosmology (Shaw *et al.*, 2007).

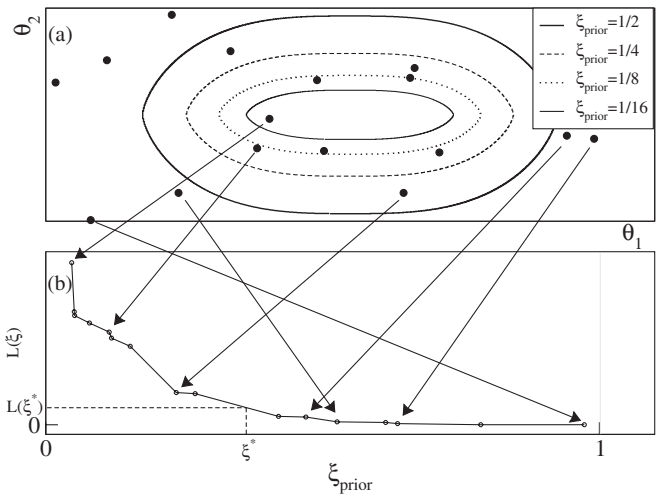

FIG. 25. Nested sampling: (a) The sample points are uniformly sampled from the prior distribution. (b) The sample points are sorted with respect to prior volume enclosing likelihood mass $\leq L(\xi)$.

c. Nonequilibrium MC

A recently proposed method (Ahlers and Engel, 2008) based on nonequilibrium equality for free energy differences (Jarzynski, 1997) highlights the continuing fruitful exchange of ideas and concepts between statistical physics and statistics. The method can be considered as interpolation between the simple but inefficient sampling from the prior distribution and thermodynamic integration. It is similar to thermodynamic integration in which a sequence of temperature values β_m is used to bridge from the prior distribution to the posterior distribution. However, the (time-consuming) equilibrium requirements for the Monte Carlo chains at the

various temperatures in the thermodynamic integration method are avoided by building on recent progress in the understanding of nonequilibrium processes (Jarzynski, 1997; Seifert, 2005).

For the nonstationary Markov process a time interval $t = 1, \ldots, N$ is chosen with $M \leq N$ intermediate time points t_m, where β changes by $\Delta\beta_m = \beta_{m+1} - \beta_m$ such that β changes from 0 to 1 within the time interval. Defining the β-tempered distribution as

$$p_\beta(\boldsymbol{\theta}) = \frac{1}{Z(\beta)} p(\mathbf{d}|\boldsymbol{\theta}, M, I)^\beta p(\boldsymbol{\theta}|M, I) \tag{113}$$

and the transition probability $T(\boldsymbol{\theta}', \boldsymbol{\theta}; \beta_m)$

$$\hat{p}_{\beta_m}(\boldsymbol{\theta}') = \int d\boldsymbol{\theta} T(\boldsymbol{\theta}', \boldsymbol{\theta}; \beta_m) \hat{p}_{\beta_m}(\boldsymbol{\theta}), \tag{114}$$

the probability for a trajectory is given by

$$p(\{\boldsymbol{\theta}_t\}) = p(\boldsymbol{\theta}|M, I) \prod_{t=1}^{N-1} T(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t; \beta_m). \tag{115}$$

Ahlers and Engel (2008) showed that the exponential average of the trajectory dependent functional

$$R(\{\boldsymbol{\theta}_t\}) = \sum_{m=1}^{M-1} \Delta\beta_m \ln p(\boldsymbol{d}|\boldsymbol{\theta}_{t_m}, I) \tag{116}$$

with respect to the trajectory probability yields the evidence

$$Z = p(\mathbf{d}|M, I) = \langle \exp(R) \rangle = \int \sum_{t=1}^{N} d\boldsymbol{\theta}_t p(\{\boldsymbol{\theta}_t\}) e^{R(\{\boldsymbol{\theta}_t\})}. \tag{117}$$

Although the exponential average may be dominated by rare events (Jarzynski, 2006) a comparison with thermodynamic integration in bimodal test cases turned out in favor of the nonequilibrium method. It is obvious that further research is needed to fully explore the potential of this (and potentially other) nonequilibrium technique(s). However, the flexibility in the choice of the β sequences offers promising optimization potential.

5. Concluding remarks

The enormous potential of Markov chain Monte Carlo methods to compute high-dimensional integrals has led to the development of a wide variety of different algorithms. Combined with today's increasing (parallel) computing capabilities the researcher has a well-equipped toolbox available even for difficult integration problems. Several well-developed MCMC program packages are available for Bayesian inference. The most well-known software package for Bayesian inference is BUGS (Thomas *et al.*, 1992), which is also available as WINBUGS[3] for the Windows operating system (Lunn *et al.*, 2000) and as an open-source version OPENBUGS[4]. Several books provide worked examples of statistical inference using BUGS, see, e.g., Gelman *et al.* (2004) and Gamerman and Lopes (2006). A variety of different MCMC algorithms have been implemented and the source code is available on the Web.[5] As a general purpose high-level language for statistical inference and postprocessing *R*[6] is in widespread use. It is an open source and is developed under the GNU license[7]. Manuals and FAQs are available on the *R* project Web site. Examples of the use of *R* can also be found by Gelman *et al.* (2004), Albert (2009), and Robert and Casella (2010).

The crucial question remains of the convergence of the Markov chains. Essentially all algorithms rely on convergence diagnostics which only supply necessary criteria for convergence and may be misled by a very slowly converging algorithm (which misses an isolated peak).

The best chance to detect such a failure is provided by the complementary use of several MCMC algorithms with different properties (Clyde *et al.*, 2007; Preuss and von Toussaint, 2007). Unlike the case of convergence diagnostics where several quantities are routinely monitored, this is, unfortunately, still not common.

V. MODEL COMPARISON

So far the Bayesian approach to the parametric estimation has been demonstrated (cf. Sec. III) which is essentially centered around the computation of the posterior distribution which in most cases provides an easy access to all desired quantities, e.g., mean, variance, or median of the parameters of a given model. A more complex situation arises when there are several models M_i, each of which might depend on several, possibly different parameters (Bretthorst, 1996). A cautionary note about the use of improper priors in model comparison: In many cases of parameter estimation the (convenient) use of improper priors is good natured and reduces Bayesian estimation problems to a maximum-likelihood problem in the case of an unbounded uniform prior. By contrast, the use of improper priors is inappropriate in problems of model comparison where the range and volume of the prior is of decisive importance [cf. Eq. (120)].

A. Basics

The formal Bayesian approach to model comparison is very similar to the one of parameter estimation,

$$\begin{aligned} p(M|D, I) &= p(D|M, I) p(M|I) / p(D|I) \\ &\propto p(D|M, I) p(M|I). \end{aligned} \tag{118}$$

The term $p(D|M, I)$ can be computed with the help of the marginalization rule Eq. (15) [see Eq. (120)]. If there is no reason to prefer a model, equal prior probabilities can be assigned to all models $p(M_i|I) = \text{const}$. This is a frequently arising situation, but it should be kept in mind that more precise prior knowledge can be incorporated and should be used if available. With an ignorant state of knowledge about the prior model probability $p(M_i|I)$ the ratio of the posterior model probabilities $p(M_i|D, I)$ and $p(M_j|D, I)$ reduces to the

[3] http://www.mrc-bsu.cam.ac.uk/bugs/
[4] http://www.openbugs.info/.
[5] http://www.cs.toronto.edu/~radford/software-online.html.
[6] http://www.r-project.org.
[7] http://www.gnu.org.

ratio of the evidences, the so-called Bayes factor (Kass and Raftery, 1995)

$$B_{ij} = p(D|M_j, I)/p(D|M_i, I). \quad (119)$$

A simple interpretation of the evidence and the way *Ockham's razor* (avoiding unnecessarily complex models) is incorporated in the model comparison can be given (Jaynes, 1979; Gregory and Loredo, 1992; MacKay, 1992a): first, the marginal likelihood $p(D|M, I)$ is written in the form

$$p(D|M, I) = \int d\boldsymbol{\theta}\, p(D|M, \boldsymbol{\theta}, I) p(\boldsymbol{\theta}|I). \quad (120)$$

Under the assumption that the prior is much more diffuse than the likelihood its variations over the range where the likelihood peaks can be neglected. Therefore the prior term can be taken at θ_{ML}, the point where the likelihood attains its maximum value, outside the integral

$$p(D|M, I) \approx p(\boldsymbol{\theta}_{\mathrm{ML}}|I) \int d\boldsymbol{\theta}\, p(D|M, \boldsymbol{\theta}, I). \quad (121)$$

The remaining θ integral over the likelihood may be further approximated by

$$p(D|M, I) \approx p(D|M, \boldsymbol{\theta}_{\mathrm{ML}}, I) p(\boldsymbol{\theta}_{\mathrm{ML}}|I) (\Delta\theta_{\mathrm{like}})^{N_\theta}, \quad (122)$$

where N_θ is the number of model parameters and $(\Delta\theta_{\mathrm{like}})^{N_\theta}$ is the approximate likelihood volume. Taking advantage of the fact that the prior is approximately uniform over some interval $\Delta\theta_{\mathrm{prior}}$ larger than the posterior peak and that the prior is normalized to 1, an approximation is given by $p(\boldsymbol{\theta}_{\mathrm{ML}}|I) \approx 1/(\Delta\theta_{\mathrm{prior}})^{N_\theta}$. Equation (122) becomes

$$p(D|M, I) \approx p(D|M, \theta_{\mathrm{ML}}, I)(\Delta\theta_{\mathrm{like}}/\Delta\theta_{\mathrm{prior}})^{N_\theta}. \quad (123)$$

Under these assumptions the evidence is approximately equal to the maximum-likelihood solution penalized by the second term, which is referred to as an Ockham factor. Since by assumption $\Delta\theta_{\mathrm{like}} \ll \Delta\theta_{\mathrm{prior}}$, the Ockham factor is $\ll 1$. With an increasing number of model parameters N_θ the improvements in the likelihood will eventually be counterbalanced by the decreasing second term in Eq. (123) thus defining an optimal model complexity.

A particular property of the evidence is that it does not penalize parameters which are unconstrained by the data (Spiegelhalter, 2002; Liddle, 2007), essentially penalizing only relevant parameters. If the likelihood is not affected by a parameter, the evidence integral is unchanged since the prior distribution is normalized. Liddle (2007) provided further comments on this property.

1. Other measures of model complexity

The necessary integrations to compute the evidence can be very demanding, even with state-of-the art equipment and algorithms. For that reason many simpler surrogates for the evidence are in use which try to balance between fit quality and model complexity. The most common ones are (Stoica and Selén, 2004) as follows:

- Akaike information criterion (AIC): the AIC is defined as

$$\mathrm{AIC} = -2\ln p(d|\boldsymbol{\theta}_{\mathrm{ML}}, I) + 2k, \quad (124)$$

  where k is the number of adjustable parameters of the model (Akaike, 1974; Smith and Spiegelhalter, 1980) and $p(d|\boldsymbol{\theta}_{\mathrm{ML}}, I)$ the maximum-likelihood value.
- Bayesian information criterion (BIC): the BIC, also known as Schwarz criterion (Schwarz, 1978) is defined as

$$\mathrm{BIC} = -2\ln p(d|\boldsymbol{\theta}_{\mathrm{ML}}, I) + k\ln N, \quad (125)$$

  where k is the number of parameters of the model and N is the number of data points. The data points are assumed to be independent and identically distributed. Note that, compared to AIC, this penalizes model complexity more heavily for a moderate number of data points.

These local criteria do not take account of the uncertainty (and possible degeneracy) in the model parameters. In practice the performance varies (Stoica and Selén, 2004). Spiegelhalter (2002) introduced a measure for the *effective* number of parameters in a model and developed another criterion, the *deviance information criterion*. Liddle (2007) used this and several other information criteria for the ranking of cosmological models and got significantly different conclusions from the data. In practice the (asymptotic) assumptions underlying the different information criteria are nearly always violated (Berger *et al.*, 2003) and this may influence the results. For that reason, model comparison should be based on Bayesian evidence whenever possible.

2. A note on significance tests

Bayesian model comparison is always based on the comparison of different (at least two) proposed models. There is no counterpart to the frequentist significance tests which evaluate a model based on only a single model. However, from a Bayesian point of view the frequentist significance test (i.e., the use of p values for model evaluation) has weaknesses; one deficit is the violation of the likelihood principle (Birnbaum, 1962; Berger and Wolpert, 1988). The likelihood principle is implied by the generally accepted sufficiency principle (Huzurbazar, 1976) conditionally on the acceptance of a second principle, the conditionality principle: If two experiments on the parameter θ, E_1 and E_2 are available and if one of these two experiments is selected with probability 0.5, the resulting inference on θ should depend only on the selected measurement. This principle seems difficult to reject (Robert, 1994). The violation of the likelihood principle introduces a dependence of the significance test result on unobserved data or stopping rules, which is strongly criticized by Bayesian proponents (Jeffreys, 1939; Berger and Sellke, 1987; Berger and Berry, 1988; Loredo, 1992; Jaynes and Bretthorst, 2003). The Bayesian model selection is considered to be far more flexible with respect to multiple hypothesis testing, nonstandard distributions, consistency, and overfitting (Berger and Pericchi, 1988). Furthermore, as pointed out by Berger and Sellke (1987) and also observed in practice (Davidoff, 1999; Goodman, 1999a, 1999b), frequentist significance levels (P values) can be a highly misleading measure of the evidence provided by the data against a null hypothesis. For recent attempts to reconcile Bayesian evidence measures and frequentist hypothesis test, see Sellke *et al.* (2001).

B. Model averaging

In many situations the focus is less on singling out a specific model but to make predictions. A common approach in this situation is the following one: A model is selected from some class of models on the basis of available data and then this model is used for predictions. However, proceeding this way ignores the uncertainty in the model selection, leading to overconfident estimates of uncertainty about the quantities of interest (Draper, 1995). Basing inferences on a single model alone is risky; ambiguity about the correct model should affect the predictions (Miller, 1984; Hoeting *et al.*, 1999). In principle, the Bayesian approach can handle this difficulty simply by replacing the model choice with model averaging. Suppose there are K models $(M_1, \ldots, M_K)$, and prior probabilities for the models $p(M_k|I)$ and for the respective parameters $p(\boldsymbol{\theta}_k|M_k, I)$ are given. Then the posterior distribution for a quantity of interest ω is computable as (Leamer, 1978; Stewart, 1987)

$$p(\omega|\mathbf{d}, I) = \sum_{k=1}^{K} p(\omega|\mathbf{d}, M_k, I)\, p(M_k|\mathbf{d}, I), \tag{126}$$

where $p(\omega|\mathbf{d}, M_k, I)$ is the posterior for ω under the kth model. Each term is weighted by the posterior model probability,

$$p(M_k|\mathbf{d}, I) = \frac{p(\mathbf{d}|M_k, I)\, p(M_k|I)}{\sum_k p(\mathbf{d}|M_k, I)\, p(M_k|I)}, \tag{127}$$

where

$$p(\mathbf{d}|M_k, I) = \int d\boldsymbol{\theta}\, p(\mathbf{d}|\boldsymbol{\theta}, M_k, I)\, p(\boldsymbol{\theta}_k|M_k, I) \tag{128}$$

is the evidence (marginal likelihood) of M_k. $p(\mathbf{d}|M_k, I)$ can be considered as the probability that the data are generated from model M_k (Clyde and George, 2004). If one model is overwhelmingly more probable than the others, the model-averaged posterior distribution $p(\omega|\mathbf{d}, I)$ is close to the model-specific distribution $p(\omega|\mathbf{d}, M_k, I)$, but otherwise the distributions can be significantly different. Madigan and Raftery (1994) showed that averaging over all models results in a better (log) predictive score than using any one of the models individually. A major difficulty in implementing this approach is that the number of models to be considered are often large and the computation of the evidence is in many cases very time consuming. For nonlinear models it is only now becoming feasible, by virtue of recent computational advances and approximations (Raftery *et al.*, 1997). Often the RJMCMC algorithm (Green, 1995, 2003) is the most convenient one to compute the model evidence (Clyde and George, 2004). The model averaging approach is widely used for variable selection (George, 2000), mixture modeling (Richardson and Green, 1997, 1998), and nonparametric regression [see, e.g., Denison *et al.* (1998), DiMatteo *et al.* (2001), and Clyde and George (2004) for additional references].

C. Case studies

1. The primordial power spectrum

The cosmic microwave background (CMB) was discovered by Penzias and Wilson and explained by Dicke and collaborators in 1965. The derived big-bang model is able to explain the primordial abundances of light elements and the origin of the cosmic microwave background (Kolb and Turner, 1990; Durrer, 2008). The measured CMB spectrum precisely matches the spectrum of a blackbody with a temperature of 2.725 K and has the same temperature to high precision in all directions of the CMB sky. However, this homogenous temperature cannot be explained within the big-bang model. Regions which have been causally connected at the time of decoupling of matter and radiation (about 380 000 years after the big bang) correspond nowadays to an angle of order 1°. But if areas which are farther apart than 1° had no causal contact before the last scattering then there is no way to establish thermal equilibrium (Bassett *et al.*, 2006). Theoretical considerations suggested the existence of relative amplitudes near 10^{-4}. These predicted fluctuations, however, were not observed, although cosmologists kept searching increasingly desperate for decades after 1965 (Uson and Wilkinson, 1984). In the early 1980s Peebles (1982) suggested that the reduced CMB fluctuation level could be explained if a kind of "dark matter," not interacting with light, is present. Other problems of standard big-bang cosmology [e.g., the relic density problem (Bassett *et al.*, 2006)] then led to the introduction of an inflationary model (Starobinsky, 1982; Linde, 1994): Quantum fluctuations of the field responsible for inflation, called the inflation, are stretched on macroscopic scales by the accelerated expansion (Linde, 1983). The COBE mission (Mather, 2007) then discovered the primordial density fluctuations in the CMB and subsequent precise measurements of cosmic microwave background fluctuations by various experiments [e.g., the Wilkinson microwave anisotropy probe (WMAP) (Bennett *et al.*, 2003)] have helped to establish a standard cosmology, the hot-big-bang model followed by an inflationary phase (Bartelmann, 2010). It is therefore not exaggerated to state that the measurement and analysis of CMB data led to a revolution in our understanding of the Universe. Since the detailed shape of the CMB power spectrum depends sensitively on the cosmological models and parameters, these models can be in turn constrained by high-precision measurements. In the following an example of a Bayesian model comparison on different possible cosmological models is presented. A recently published map of fluctuations of the cosmic microwave background based on 5-year WMAP data (Hinshaw *et al.*, 2009) is shown in Fig. 26. The fluctuations of the temperature around the mean value of $T_0 = 2.725$ K are on the order of 200 μK. It is convenient to express these temperature fluctuations δT in spherical harmonics,

$$\frac{\delta T(\vartheta, \phi)}{\mathrm{T}_0} = \sum_{l=2}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\vartheta, \phi), \tag{129}$$

where the monopole and dipole terms have been subtracted out (Lidsay *et al.*, 1997).

Inflation theory predicts that the a_{lm} are Gaussian random variables. In a rotationally invariant case this translates to a simplified representation of the angular correlation function of the temperature fluctuations $\langle a^*_{l'm'}, a_{lm}\rangle = C_l \delta_{l'l} \delta_{mm'}$, in terms of multipole moments C_l. For Gaussian fluctuations, the set of C_l's completely characterizes the temperature anisotropy. If the fluctuations are non-Gaussian, higher order

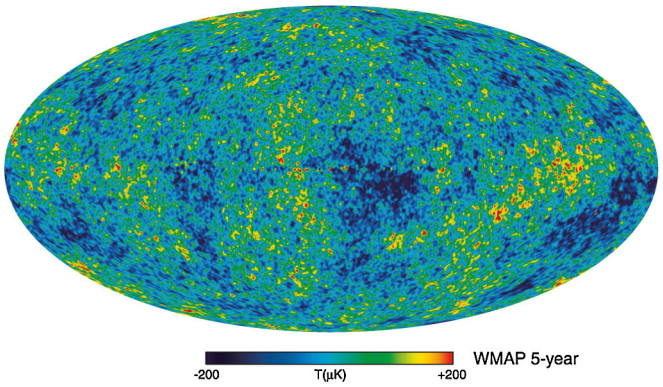

FIG. 26 (color online). Sky map of the cosmic microwave temperature fluctuations around the mean value of 2.725 K in galactic coordinates (Mollweide projection) after subtraction of foreground sources. The foreground-reduced internal linear combination map is based on the 5-yr WMAP data. From Hinshaw *et al.*, 2009.

correlation functions are necessary to fully characterize the anisotropy (Kinney, 2001; Komatsu *et al.*, 2009). The radiation power spectrum is defined to be $l(l+1)C_l$. Accurate calculations of the power spectrum from cosmological models require the numerical solution of the coupled Einstein-Boltzmann equations [e.g., CAMB[8] (Lewis *et al.*, 2000) or CMBFAST[9]], usually coupled with MCMC codes such as COSMOMC[10] or COSMONEST[11] for parameter estimation. The computation of $C_l(\theta)$ as a function of the cosmological parameters θ can now be done with high accuracy of around 1% precision or better but is very time consuming. The power of the CMB in constraining cosmological parameters comes from the fact that a large number of data (the C_l spectrum) is used to constrain about a dozen cosmological parameters (such as total matter density or cosmological constant) and a set of parameters to describe the inflaton potential (Lidsay *et al.*, 1997). The primordial power density spectra (not to be confused with the radiation power spectrum) predicted by many inflationary models is often written as (Kosowsky and Turner, 1995; Leach *et al.*, 2002; Bassett *et al.*, 2006)

$$P(k) \propto (k/k_0)^{n_s-1+(1/2)\ln(k/k_0)n_{\mathrm{run}}+\cdots}. \tag{130}$$

This parametrization encompasses the most commonly tested power spectra, namely, the scale-invariant spectrum ($n_{\mathrm{run}} = n_s - 1 = 0$), the tilted spectrum ($n_{\mathrm{run}} = 0$), and a running spectrum in which the tilt becomes a function of scale ($n_{\mathrm{run}} = dn_s/d\ln k \neq 0$).

Spergel *et al.* (2007) fitted a large variety of power spectra models to the data of the 3-year WMAP mission. The reference model was the Λ-cold dark matter model (ΛCDM), which predicts a tilted power spectrum

$$P(k) \propto k^{n_s-1}, \tag{131}$$

with $n_s < 1$. The best-fit tilt parameter is $n_s = 0.958 \pm 0.016$ based on the 3-year WMAP data. This is in good agreement with the inflationary paradigm (Spergel *et al.*, 2007). The comparison of the different models was based on the relative goodness of fit

$$\Delta\chi^2_{\mathrm{eff}} = 2\ln L(\Lambda\mathrm{CDM}) - 2\ln L(\mathrm{model}), \tag{132}$$

where L denotes the respective likelihood functions. The differences in the likelihood values were relatively small [$|\Delta\chi^2_{\mathrm{eff}}| \sim O(1)$], except for the model without dark matter ($\Delta\chi^2_{\mathrm{eff}} = 248$), which was clearly ruled out.

The best-fitting model was a form-free one with 15 logarithmically spaced support points (see Fig. 27) which improved the likelihood by $\Delta\chi^2_{\mathrm{eff}} = -22$ compared to the Λ-cold dark matter model. However, it is not immediately obvious if the model is fully adequate (Liddle, 2004; Trotta, 2007a) or if it is missing some structure which is supported by measurements or if it is already overfitting the noisy data. For that reason Bridges *et al.* (2009) used Bayesian model selection to reconstruct the optimal structure in the spectrum. Similar to the approach of Spergel *et al.* (2007) the spectrum was modeled as piecewise linear between the support points in k space whose amplitudes are allowed to vary. The number of support points and their k-space positions were chosen by Bayesian evidence. If there is initially no reason to prefer model M_j over M_1 the Bayes factor is given by [cf. Eq. (119)]

$$B_{1j} = \frac{p(\mathbf{D}|M_j,I)}{p(\mathbf{D}|M_1,I)} = \frac{\int d\boldsymbol{\theta}_j\, p(\mathbf{D}|\boldsymbol{\theta}_j,M_j,I)\,p(\boldsymbol{\theta}_j|I)}{\int d\boldsymbol{\theta}_1\, p(\mathbf{D}|\boldsymbol{\theta}_1,M_1,I)\,p(\boldsymbol{\theta}_1|I)}. \tag{133}$$

The evaluation of the multidimensional integrals is not trivial. Several possible integration methods are suggested in Sec. IV.E.4. Bridges *et al.* (2009) applied the method of nested sampling (Skilling, 2006; Feroz *et al.*, 2009). The considered data included the 5-year release from WMAP (Hinshaw *et al.*, 2009) and results from the arcminute

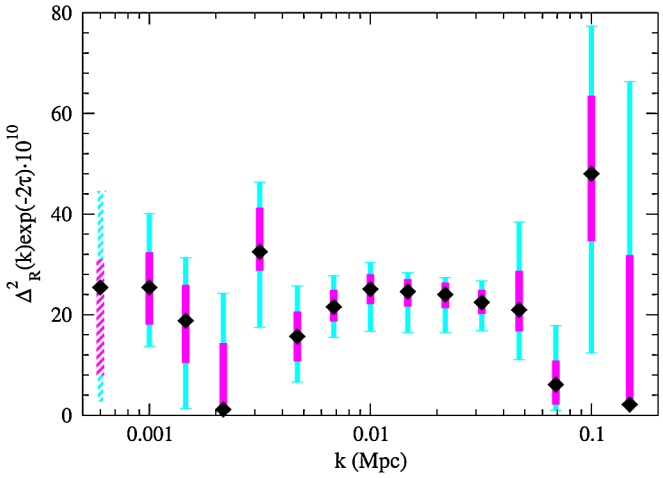

FIG. 27 (color online). Linear interpolated reconstruction of the primordial curvature fluctuation spectrum. The bins are logarithmically spaced. The errors show the 68% and 95% constraints and the black diamonds show the mode of the likelihood. The dashed vertical line on the left-hand side shows the values for $k = 0$. Only the amplitude was allowed to vary at each of the nodes. From Spergel *et al.*, 2007.

[8]http://www.camb.info.
[9]http://lambda.gsfc.nasa.gov/toolbox/tb_cmbfast_ov.cfm.
[10]http://cosmologist.info/cosmomc/.
[11]http://www.cosmonest.org.

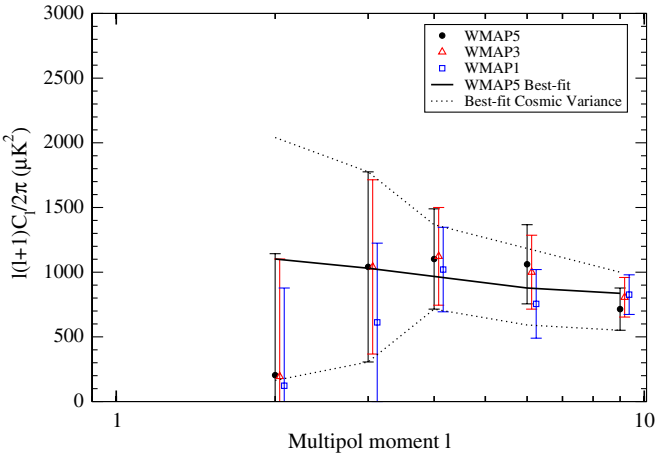

FIG. 28 (color online). Low-l multipoles and 1σ error bars from three releases of WMAP*WMAP* data. The best-fitting fiducial power spectrum based on WMAP5 inferences is indicated together with the associated cosmic variance limits. The multipol values are slightly shifted for clarity. From Bridges *et al.*, 2009.

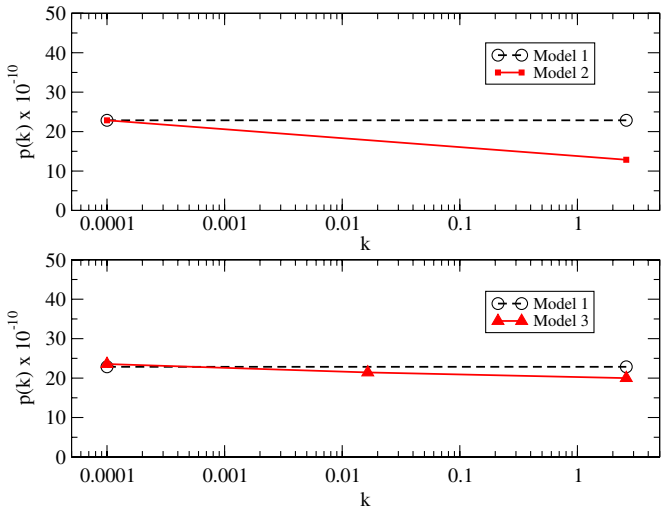

FIG. 29 (color online). Piecewise linear interpolated reconstruction of the primordial spectrum with different degrees of flexibility. Only the amplitude was allowed to vary at each of the support points (indicated by filled symbols). In model 1 only the amplitude of a flat spectrum could be varied (shown as a dashed line in both graphs), in model 2 the slope could also be adjusted (two support points, solid line, upper graph), and model 3 corresponds to a piecewise linear model with two segments (corresponding to three support points, solid line, lower graph). The Bayes factors are $B_{21} = 2$ and $B_{31} = 3$ in favor of the more complex models compared to the flat model 1. Adapted from Bridges *et al.*, 2009.

cosmology bolometer array (ACBAR) (Reichardt *et al.*, 2009). For the full list of included data, see Bridges *et al.* (2009).

In Fig. 28 the measured C_l values at low-l multipoles and 1σ error bars from three releases of WMAP data (1 year, 3 years, and 5 years) are displayed. Especially at the low-l side the measured values are at the edge of the 1σ limit of the best-fit model of the WMAP5 data. Please note the shift of the octupole moment (second data column from the left) between the first year data and the 3-year data.

To determine the degree of structure that can be usefully constrained by the measured data Bridges *et al.* (2009) first computed the evidence for the constant power spectrum, which corresponds to the scale-invariant spectrum ($n_s - 1 = 0$). In the next step, two support points at the edges of the k space emulated a tilted spectrum. The next model had three nodes, the new node added between two existing nodes. This process was continued up to six support points. The evidence of each model was computed by marginalizing the model parameters, i.e., the amplitude parameters. In Fig. 29 the results for one up to three support points are shown. The mean amplitude values at the support points are indicated by symbols. A comparison of the Bayes factors B_{j1} [cf. Eq. (133)] of the different models reveals that all models with more than three support points are less likely than the constant model since the increased number of parameters is not compensated by a significantly better fit. The power spectrum structure shown in Fig. 27 is therefore highly susceptible to overfitting resulting in artificial structures. The models with two and three support points are the most likely ones, with Bayes factors of $B_{21} = 2$ and $B_{31} = 3$, dominating the scale-invariant model and instead slightly favoring a tilted primordial power spectrum. For the latest results on the cosmic microwave background see, e.g., Jarosik *et al.* (2011).

Bayesian model selection techniques have also been applied to estimate the discriminative power of planned experiments (Mukherje, Parkinson *et al.*, 2006; Pahud *et al.*, 2007; Trotta, 2007b), the first step toward Bayesian experimental design (cf. Sec. VII) of future missions. Bayesian techniques are also extensively used in the analysis of cosmological data; see, e.g., Dickinson *et al.* (2009) and Dunkley *et al.* (2009). For a recent review of Bayesian inference in cosmology, see Trotta (2008).

2. Mass spectroscopy

Plasma-based surface processing is widely used in the microchip and display industry, where many manufacturing processes occur in plasma reactors. The identification and quantification of plasma products for processing control have become one of the urgent topics for plasma physicists. Detailed knowledge of concentrations of reactive particles such as free radicals is needed to understand the underlying microprocesses (von Keudell, 2002). Mass spectroscopy is a convenient technique to directly monitor the particle fluxes at the substrate sites. Traditional quadrupole spectrometers are widely used due to high sensitivity, reasonable stability, and low costs. To be filtered in the quadrupole field, neutral gases have to be ionized, most commonly by electron impact. At a typical electron energy of 50–100 eV (used to achieve a high ionization efficiency) stable molecules decompose in a variety of fragment ions leading to the so-called *cracking pattern*. For overlapping cracking patterns subtraction methods have been devised to disentangle the measured spectra (Dobrozemsky and Schwarzinger, 1992). These methods suffer from excessive error buildup and are not applicable, when unstable constituents such as radicals are assessed, due to the lack of knowledge of cracking patterns. Furthermore, the fragmentation is also an instrument-specific property and thus requires an instrument-specific calibration. A rigorous analysis of composite mass spectra employs BPT which also

succeeds without exact cracking patterns (Schwarz-Selinger *et al.*, 2001; Preuss *et al.*, 2002).

Assuming a linear response of the mass spectrometer the mass signal vector of measurement j, $\mathbf{d}_j$ is the sum of the contributions of all species in the mixture

$$\mathbf{d}_j = \mathbf{C}\mathbf{x}_j + \boldsymbol{\epsilon}_j \tag{134}$$

with Gaussian noise $\boldsymbol{\epsilon}$. The goal is to determine the posterior distribution of the cracking matrix elements $\mathbf{C}$, the vector $\mathbf{x}_j$ of species concentrations in measurement j, and also the number of species E. $\boldsymbol{\epsilon}_j$ is the vector of measurement errors associated with the signal vector $\mathbf{d}_j$. The cracking column vectors are normalized to sum up to 1. The likelihood is given as

$$p(\mathbf{D}|\mathbf{C},\mathbf{X},\{\mathbf{S}\},E,I) = \prod_j \frac{1}{\prod_i \sqrt{2\pi}s_{ij}} \exp\left(-\frac{1}{2}(\mathbf{d}_j - \mathbf{C}\mathbf{x}_j)^T \times \mathbf{S}_j^{-1}(\mathbf{d}_j - \mathbf{C}\mathbf{x}_j)\right). \tag{135}$$

$\{\mathbf{S}\}$ denotes the ensemble of diagonal matrices $\mathbf{S}_j$ with components $(\mathbf{S}_j)_{ii} = s_{ij}^2$, given by the measurement error of the jth measurement in the ith mass channel. The only components which still need to be specified to start the Bayesian inference are the prior distributions for the number of components $p(E|I)$, the concentration matrix $p(\mathbf{X}|E,I)$, and finally the cracking matrix elements $p(\mathbf{C}|E,I)$. For the prior probability of E a constant prior is chosen $p(E|I) = 1/E_{\max}$. Cracking patterns of stable molecules are listed as point estimates, e.g., in the tables of Cornu and Massot (1979), with the dominant component being normalized to 1000. Together with the requirement that the cracking coefficients are confined to the interval [0, 1] this allows the computation of an exponential prior for the cracking coefficients $p(\mathbf{C}|E,I)$. Note, however, that this prior, though still exponential, is more complicated than Eq. (28) since the support of the cracking coefficients is not infinite but rather confined to the interval [0, 1] (Schwarz-Selinger *et al.*, 2001). Prior knowledge about the components of a CH_4 plasma is chosen from experimental experience. Common knowledge is that H_2 and CH_4 are the main neutral constituents and all other species remain below a few percent with declining intensity as the carbon content of a species rises. This allows again the assignment of exponential prior distributions for the concentrations. The probability for a particular set of E species in the model is given in terms of the data $\mathbf{D}$ and variances $\{\mathbf{S}\}$ by Bayes theorem,

$$p(E|\mathbf{D},\{\mathbf{S}\},I) = p(E|I)p(\mathbf{D}|\{\mathbf{S}\},E,I)/p(\mathbf{D}|\{\mathbf{S}\},I). \tag{136}$$

The marginal likelihood $p(\mathbf{D}|\{\mathbf{S}\},I)$ is obtained from

$$p(\mathbf{D}|\{\mathbf{S},E\},I) = \int d\mathbf{C}d\mathbf{X}\, p(\mathbf{C}|E,I)p(\mathbf{X}|E,I) \times p(\mathbf{D}|\mathbf{C},\mathbf{X},\{\mathbf{S}\},E,I). \tag{137}$$

The dimension of the integral is high and increases with the number of data sets represented by $\mathbf{D}$ and the number of species chosen to model the observations. Such high-dimensional integrals (for interpretation of the spectrum shown in Fig. 30 the dimension exceeds 400) can be tackled either by Markov chain Monte Carlo techniques (using thermodynamic integration for a faster convergence) or by saddle-point approximations which may not always exist in the analysis of mass spectra. A low temperature methane plasma was analyzed with respect to H atoms and H_2 and C_nH_x, $n = 1,\ldots,4$ molecules. In particular, the identification of the relevant radicals and their concentrations was of interest.

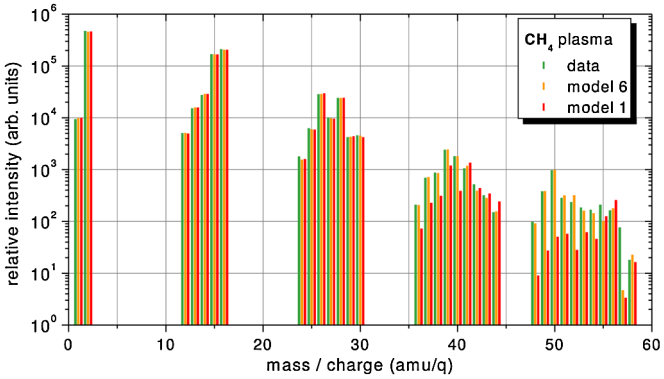

FIG. 30 (color online). Comparison of data computed by two different models (but with the same number of radicals) with measured mass spectrometer data. Model 6 additionally incorporates the species C_4H_2 and C_4H_6 compared to model 1 which contains only up to C_3H_x species. The former model provides a near perfect fit to the measured data even at high masses.

As can be seen from Fig. 31 a model taking into account only nonradical molecules cannot describe the measured data well. The misfit decreases monotonously as more radicals are incorporated into the model. By contrast, the evidence attains a maximum for inclusion of three radicals (C_2H_5, CH_3) and H and decays slowly for more complicated models. This result is rather reasonable since these radicals are produced by

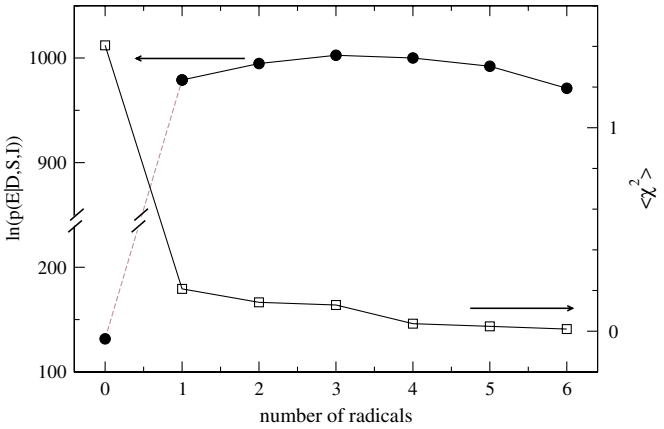

FIG. 31 (color online). The natural logarithm of $p(E|\mathbf{D},\mathbf{S},I)$ is displayed using full dots; the scale is given on the left ordinate (indicated by the upper arrow). The misfit between data and model for different combinations of six free radicals (C_2H_5, CH_3, H, C_2H_3, CH, and CH_2) (while keeping the set of nonradical species constant) is indicated with open symbols, the corresponding scale is given on the right ordinate as the averaged χ^2 normalized by the number of data points, indicated by the lower arrow. The abscissa shows the number of radicals involved in the model, which were taken in the given order. Lines are to guide the eye. From Kang and Dose, 2003.

breaking just one atomic bond from the stable and abundant molecules H_2, CH_4, and C_2H_6. The next step after the identification of the number of species contributing to the set of measurements is the estimation of the concentrations and the cracking coefficients. The required posterior probability distribution is given by

$$p(\mathbf{X},\mathbf{C}|\mathbf{D},\mathbf{S},E,I)=\frac{p(\mathbf{X}|E,I)p(\mathbf{C}|E,I)p(\mathbf{D}|\mathbf{C},\mathbf{X},\mathbf{S},E,I)}{p(\mathbf{D}|\mathbf{S},E,I)}. \tag{138}$$

Detection and quantification of radicals is one attractive result of the Bayesian analysis of a beam shutter (on or off) experiment in the diagnostic of a low temperature plasma. Equally important and equally demanding is the analysis of the neutral gas mass spectra, in particular, for plasmas with hydrocarbon fuel gases. Figure 30 shows a result from a comprehensive data set from 34 mass channels for 27 different plasma conditions of an inductively coupled pulsed plasma discharge, together with calibration measurements for 11 species. Two models with a different number of hydrocarbon molecules are compared. The modeled data agree extremely well with the measurements for masses below 35. For higher masses there is a discrepancy between model 1 and the data, whereas model 6 gives a nearly perfect match, indicating the presence of C_4H_2 and C_4H_6 in the plasma. The large number of different species possibly present in the plasma lead to a large number of different models to be compared. A detailed discussion of the results is beyond the scope of this paper and has been given elsewhere. Nevertheless, the algorithmic implementation of the Bayesian method is so efficient that CPU time is no longer a valid argument to digress to less-reliable methods, except for monitoring purposes (von Toussaint, Does, and Golan, 2004).

3. Discordant data sets

Experimental data from different sources may suffer from discordant calibrations and possibly cover different regions of the independent variables. Here an example is given of how to treat unknown scale factors of different data sets.

Chemical erosion due to hydrogen ion bombardment is the dominant erosion process for carbon-based plasma-facing materials in fusion experiments. In the low flux regime, i.e., $\theta < 10^{19}/\mathrm{m}^2\mathrm{s}$, the mechanism of chemical erosion is reasonably well understood. At high fluxes θ, such as experienced in fusion devices, there was indication from various data that the chemical sputtering yield decreases with ion flux above a certain threshold (Roth and Garcia-Rosales, 1996). Weight loss measurements are available for the low flux regime (Balden and Roth, 2000). Those measurements are the most reliable ones, since these data require no further calibration factors. The function for the chemical erosion yield is taken from Roth (1999). For the weight loss measurements it is assumed that the erosion yield $\phi(\theta,\theta_0)$ depends on flux θ through

$$\phi(\theta,\theta_0)=Y_{\mathrm{chem}}\frac{1}{1+\theta/\theta_0}, \tag{139}$$

with the threshold determined by the parameter θ_0. In contrast, calibration factors are necessary for mass spectroscopy and optical emission spectroscopy. For the high flux data the eroded molecule flux was determined spectroscopically from the CH band intensity. The reduction of the CH band emission to a total erosion yield requires accurate knowledge of the CH optical transition rates. To allow here for an uncertainty of the measured erosion data an unknown calibration factor γ is introduced. However, with erosion data collected in fusion machines the situation may also be different. The optical system used to record hydrogen and CH band emissions may suffer from a calibration error which translates into a common recalibration factor γ for both the hydrogen flux θ and the erosion yield. In this case the appropriate description is given by

$$\phi(\theta,\theta_0,\gamma)=Y_{\mathrm{chem}}\cdot\frac{1}{1+\gamma\theta/\theta_0}. \tag{140}$$

The first term Y_{chem} varies very weakly with flux θ and is considered constant. In the end it has to be distinguished between a data set from weight loss measurements δ considered to be scaled correctly

$$\delta_i = c\phi(\theta_i,\theta_0)+e_i \tag{141}$$

and the data sets from optical measurements Δ_j with a possible scale factor γ either only for the erosion yield

$$\gamma\Delta_j = c\phi(\Theta_j,\theta_0)+E_j \tag{142}$$

or also for the incoming flux

$$\gamma\Delta_j = c\phi(\Theta_j,\theta_0,\gamma)+E_j. \tag{143}$$

In the following only the two models Eqs. (141) and (143) for the high flux regime are considered. Assuming the expectation value of the errors $\langle e\rangle$ and $\langle E\rangle$ to be zero and the variance given by s_i^2 and S_i^2, respectively, the likelihood functions for the two data sets read

$$p(\boldsymbol{\delta}|\boldsymbol{\theta},\mathbf{s},\theta_0,c,I)=\prod_i\frac{1}{s_i\sqrt{2\pi}}\times\exp\left\{-\frac{1}{2}\left[\frac{\delta_i-c\phi(\theta_i,\theta_0)}{s_i}\right]^2\right\}, \tag{144}$$

$$p(\Delta|,\mathbf{S},\theta_0,\gamma,c,I)=\prod_j\frac{\gamma}{S_j\sqrt{2\pi}}\times\exp\left\{-\frac{1}{2}\left[\frac{\gamma\Delta_j-c\phi(\Theta_j,\theta_0,\gamma)}{S_j}\right]^2\right\}. \tag{145}$$

Unfortunately, the experimental error estimates for the available data sets are not compatible with the observed scatter of the data (so-called outliers are present). Outlier tolerance may be obtained in the following way (Dose and von der Linden, 1999). Assume that the probability density for the true error σ is given by a distribution which allows for large discrepancies between scattered data and specified errors,

$$p(\sigma_i|s_i,I)=(2/\pi)(s_i/\sigma_i)^2\exp(s_i^2/\sigma_i^2), \tag{146}$$

but with mean $\langle\sigma\rangle = s$ of the error estimate. Marginalization of σ yields a modified likelihood [cf. Eq. (146)]

$$p(\boldsymbol{\delta}|\boldsymbol{\theta},\mathbf{s},\theta_0,c,I)=\prod_i \frac{1}{s_i 2\pi\sqrt{2}}\left\{\frac{1}{\pi}+\frac{1}{2}[\delta_i - c\phi(\theta_i,\theta_0)]\right\}^{-3/2} \tag{147}$$

and similarly for Eq. (145). First the expectation value of the scale parameter γ is of interest. It is obtained by

$$\langle\gamma\rangle = \frac{\int d\gamma d\theta_0 \gamma p(\gamma,\theta_0|\boldsymbol{\delta},\boldsymbol{\Delta},\boldsymbol{\theta},\boldsymbol{\Theta},\mathbf{s},\mathbf{S},c,I)}{\int d\gamma d\theta_0 p(\gamma,\theta_0|\boldsymbol{\delta},\boldsymbol{\Delta},\boldsymbol{\theta},\boldsymbol{\Theta},\mathbf{s},\mathbf{S},c,I)} \tag{148}$$

and can be rewritten using Bayes' theorem

$$p(\gamma,\theta_0|\boldsymbol{\delta},\boldsymbol{\Delta},\boldsymbol{\theta},\boldsymbol{\Theta},\mathbf{s},\mathbf{S},c,I) = \frac{p(\gamma,\theta_0,c|I)}{p(\boldsymbol{\delta},\boldsymbol{\Delta}|\boldsymbol{\theta},\boldsymbol{\Theta},\mathbf{s},\mathbf{S},c,I)} p(\boldsymbol{\delta},\boldsymbol{\Delta}|\gamma,\theta_0,\boldsymbol{\theta},\boldsymbol{\Theta},\mathbf{s},\mathbf{S},c,I). \tag{149}$$

The denominator in Eq. (148) equals 1 (as normalized probability distribution) and could be omitted. However, using MCMC the estimate $\langle\gamma\rangle$ is computed from the samples obtained from the distribution $(\boldsymbol{\delta},\boldsymbol{\Delta}|\gamma,\theta_0,\boldsymbol{\theta},\boldsymbol{\Theta},\mathbf{s},\mathbf{S},c,I)$ using Eq. (148), where the integrals over γ and θ_0 are replaced by a summation over the MCMC samples.

The last term in Eq. (149) is the product of the two likelihoods. Assuming the independence of the two data sets $\boldsymbol{\delta}$ and $\boldsymbol{\Delta}$:

$$p(\boldsymbol{\delta},\boldsymbol{\Delta}|\gamma,\theta_0,\boldsymbol{\theta},\boldsymbol{\Theta},\mathbf{s},\mathbf{S},c,I) = p(\boldsymbol{\delta}|\boldsymbol{\theta},\mathbf{s},\theta_0,I)p(\boldsymbol{\Delta}|\boldsymbol{\Theta},\mathbf{S},\theta_0,\gamma,c,I). \tag{150}$$

The prior distributions of

$$p(\gamma,\theta_0,c|I) = p(\gamma|I)p(\theta_0|I)p(c|I) \tag{151}$$

are taken to be flat for c, and a Jeffreys's prior is used for θ_0. For $p(\gamma|I)$ one can assume an expectation value for the scale factor $\langle\gamma\rangle = 1$. Any other choice would imply a deliberately introduced bias in the calibrations used to obtain data set $\boldsymbol{\Delta}$. By virtue of the principle of maximum entropy this results in an exponential prior

$$p(\gamma|I) = \exp(-\gamma). \tag{152}$$

The Bayes factor for model M_1 [see Eq. (139) for the description of the low flux data combined with Eq. (142) for the high flux data] versus model M_2 [Eq. (139) with Eq. (143)] is given by the ratio of the marginalized likelihoods

$$p(\boldsymbol{\delta},\boldsymbol{\Delta}|M_k,\boldsymbol{\theta},\boldsymbol{\Theta},\mathbf{s},\mathbf{S},I) = \int d\gamma d\theta_0 dc\, p(\boldsymbol{\delta},\boldsymbol{\Delta}|M_k,\gamma,\theta_0,\boldsymbol{\theta},\boldsymbol{\Theta},\mathbf{s},\mathbf{S},c,I) \times p(\gamma,\theta_0,c|I) \tag{153}$$

when no model is preferred *a priori*.

Computing the odds ratio reveals that the model given in Eq. (139) for the low flux data ($< 10^{20}/\mathrm{m}^2\,\mathrm{s}$) and Eq. (142) for the high flux data is to be preferred by a factor of 10 over model M_2 for the data sets shown in Fig. 32. This does not give any reason for deferring from the statement of the experimentalists that the calibration for the incident hydrogen flux is quite reliable and that the correction factor should be applied to the eroded atom flux only, rather than to the incident hydrogen flux and eroded atom flux (Dose *et al.*, 2001; Preuss *et al.*, 2001). Therefore, the results shown in Fig. 32 refer to the first model. The mean of the threshold value is $\theta_0 = 28.8\times10^{-23}\ \mathrm{m}^2\,\mathrm{s}$ with scale factors of $\gamma = 0.72$ and 0.32 for the data from Grote *et al.* (1999) and Tynan (1998), respectively. The values for γ have been derived using Eq. (148).

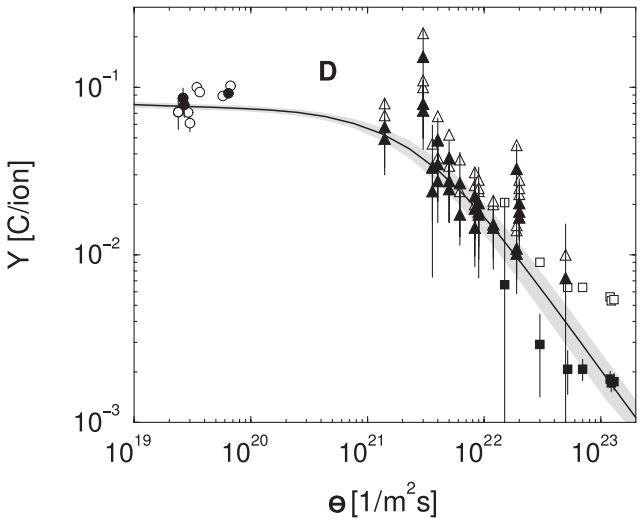

FIG. 32. Flux dependence of the chemical erosion yield of graphite under hydrogen irradiation. The abscissa shows the flux of deuterium, and the ordinate shows the carbon erosion yield in eroded carbon atoms per impinging deuterium ion. The data set δ is represented by circles. Filled circles correspond to the subset for which the fitting curve (solid line) is valid ($E_0 = 30$ eV, $T = 600$ K). Open triangles and squares represent data from the experiments PSI-I (Grote *et al.*, 1999) and PISCES (Whyte *et al.*, 2001), respectively, while the full symbols show the data sets after multiplication with the corresponding scale factors (0.72 and 0.32). Error bars show the assigned experimental error. The gray shaded area is the confidence range. From (Preuss *et al.*, 2001.

4. Mixture modeling

A mixture distribution $g(x|I)$ is given by any convex combination,

$$g(x|I) = \sum_{k=1}^{K} p_k f_k(x|I), \qquad \sum_{k=1}^{K} p_k = 1 k > 1, \tag{154}$$

of other probability distributions (Marin *et al.*, 2005). In most cases, the $f_k(x|I)$ distributions are from the same parametric family (e.g., Gaussian distributions with different means and variances), leading to a parametric mixture model

$$\sum_{k=1}^{K} p_k f(x|\boldsymbol{\theta}_k, I). \tag{155}$$

Because of their flexibility mixture models (Marin *et al.*, 2005; Bishop, 2006) are an ideal tool to solve the ubiquitous problem of background and source separation. Examples are particle induced x-ray emission measurements (Padayachee *et al.*, 1999) and Auger data (Fischer *et al.*, 2000), but also x-ray images in high-energy astrophysics (Guglielmetti *et al.*, 2004, 2009). The basic idea is simple. The background

is relatively slowly varying compared to the signal. Therefore, the background is represented by a smooth function. Data points that are significantly separated from the background here are considered as outliers (i.e., coming from a different distribution), as data points containing background and signal contributions. Given an observed data set $\mathbf{d} = \{d_i\}$ two complementary hypotheses can be formulated

$$B_i\colon\ d_i = b_i + \epsilon_i \tag{156}$$

and

$$\bar{B}_i\colon\ d_i = b_i + s_i + \epsilon_i. \tag{157}$$

Hypothesis B_i specifies that d_i consists only of background b_i and noise ϵ_i and hypothesis $\bar{B}_i$ that an additional source contribution is present. For counting experiments (and only positive signal contributions) the likelihood for the two distributions is given by the Poisson distribution,

$$p(d_i|B_i, b_i, I) = (b_i^{d_i}/d_i!)\exp(-b_i), \tag{158}$$

and

$$p(d_i|B_i, b_i, I) = [(b_i + s_i)^{d_i}/d_i!]\exp[-(b_i + s_i)]. \tag{159}$$

Since the signal intensities are unknown, they are marginalized (integrated out). The average signal intensity s_0 of the data set can be used as a reasonable expectation value of the prior distribution of the signal (von der Linden, Dose *et al.*, 1999)

$$p(s_i|s_0, I) = \exp(-s_i/s_0)/s_0. \tag{160}$$

Then the marginal Poisson likelihood for the hypothesis $\bar{B}_i$ is given by

$$p(d_i|\bar{B}_i, b_i, s_0, I) = \frac{\exp(b_i/s_0)}{s_0(1 + 1/s_0)^{(d_i+1)}} \times \frac{\Gamma[(d_i + 1), b_i(1 + 1/s_0)]}{\Gamma(d_i + 1)}. \tag{161}$$

The two different likelihoods for the propositions B_i and $\bar{B}_i$ are combined in the likelihood for the mixture model

$$p(\mathbf{d}|b, s_0, \beta, I) = \prod_i [\beta p(d_i|B_i, b_i) + (1 - \beta)p(d_i|\bar{B}_i, b_i, s_0)], \tag{162}$$

where β is the probability that a data point contains no signal contribution. $\beta = 0.5$ is a noncommittal but unrealistic value, stating whether or not each datum is equally likely to contain a signal contribution. So far the appropriate basic functions for the background model have not been specified. An obvious choice in one dimension is to use cubic splines. Fischer *et al.* (2000) trepresented he background by a cubic spline together with a smoothness prior for the background,

$$p(b|\mu, I) = \frac{1}{Z}\exp\left(-\mu \int dx|b''|^2\right), \tag{163}$$

and applied to an Auger spectrum obtained with four-grid low-energy electron-diffraction optics in the retarding field mode. Such spectra constitute the superposition of the energy derivative of the sum of the Auger electron energy distribution, the signal, and the much larger secondary electron energy distribution, the background. The latter is known to be rapidly varying in the low-energy region, as seen in Fig. 33(a). The peaks at 47 eV come from an $M_{2,3}VV$ Auger transition. While the background may be smooth, it varies quite rapidly at low energies. The variation of the data can be reduced by a logarithmic transformation of the signal $y' = \log(a - y)$. The estimated background is given in Fig. 33(b) as a solid line together with the transformed data. After plotting the difference between the original spectrum and its estimated background shown in Fig. 33(c), a possible secondary peak is observed at (47 + 39) eV, which is further substantiated by the autocorrelation of the background subtracted spectrum. The peak at 86 eV with an amplitude of about 2% of the main signal corresponds to the M_1VV Auger transition for iron. In this case a proper background subtraction reveals the presence of less apparent signals in the Auger spectrum.

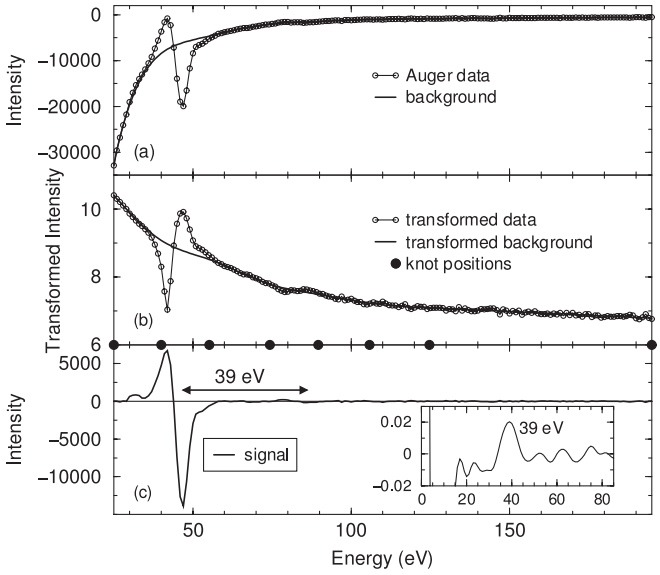

FIG. 33. (a) An MVV Auger spectrum for iron. The estimated background shown is obtained for the transformed spectrum shown in (b). A logarithmic transformation of the Auger spectrum reduces the curvature of the background. The estimated background is shown as a solid line. The eight support points of the spline are indicated by filled circles. (c) The signal obtained by subtracting the data and the background. A secondary peak is present at an energy of 86, 39 eV above the $M_{2,3}VV$ Auger transition, substantiated by the autocorrelation of the signal vs energy difference (see inset). Adapted from Fischer *et al.*, 2000.

VI. INTEGRATED DATA ANALYSIS

A. Introduction

Technological progress has had a tremendous impact on the setup of most physics experiments. Not only has the data acquisition rate increased but also the number of diagnostics. However, when there is a multiplicity of diagnostics, the problem of combining their information arises. Typically, a wide variety of diagnostics is employed simultaneously to collect data covering complementary aspects of the system under investigation. Therefore, many different,

heterogeneous information sources have to be linked and often the amount of data requires automated analysis.

Essentially there are two main approaches used to integrate information: data fusion in data space and data fusion in parameter space. Both approaches will be addressed in the following.

1. Data fusion in data space

The combination of data directly in data space is most often applied in the field of image fusion. Here images recorded by different sensors or at different times are fused using algorithms operating on pixel or feature level. Typical operations are rescaling of images to a common pixel spacing or co-registering of images using extracted features. A review of various techniques used in remote sensing is given by Pohl and Van Genderen (1998). Although data integration in data space may be computationally very efficient, the range of applicability is restricted to data sets of sufficient homogeneity. In all other cases a second approach has to be used.

2. Data fusion in parameter space

In the case of simultaneous recording of data using heterogeneous diagnostics the linkage between the measured data is provided by the state $\boldsymbol{\theta}$ of the observed physical system. This is exploited by several data fusion techniques operating in data space. Assuming that K diagnostics with corresponding individual measurements $\mathbf{d}^{(k)}$, $k=1,\ldots,K$, likelihoods $p_{(k)}(\mathbf{d}^{(k)}|\boldsymbol{\theta})$, and posterior distributions $p_{(k)}(\boldsymbol{\theta}|\mathbf{d}^{(k)})$, as well as prior distributions $p_{(k)}(\boldsymbol{\theta}|I)$ are available, several methods for fusing the information have been proposed:

a. Linear opinion pool

In the linear opinion pool approach (Manyika and Durrant-Whyte, 1995; Punska, 1999), weights w_k are assigned to the posterior distribution of each information source, reflecting the reliability and relevance of each diagnostic, yielding

$$p(\boldsymbol{\theta}|\mathbf{d}^{(1)},\ldots,\mathbf{d}^{(K)},I)=\sum_{k=1}^{K} w_k p_{(k)}(\boldsymbol{\theta}|\mathbf{d}^{(k)},I), \tag{164}$$

with $0\le w_k\le 1$ and $\sum_{k=1}^{K} w_k=1$ (Stone, 1961). This approach to sensor fusion is often used to combine sensors using simple rules ["in the near field rely on sensor A or else use sensor B"; see, e.g.,Flynn (1988)]. However, the linear opinion pool may assign a high probability to parameters which are fully excluded by some of the diagnostics $p_{(j)}(\boldsymbol{\theta}|\mathbf{d}^{(j)},I)=0$. This drawback naturally leads to the next approach.

b. Independent opinion pool

In the independent opinion pool (Manyika and Durrant-Whyte, 1995) the posterior distributions from the different sources are multiplied,

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{d}^{(1)},\ldots,\mathbf{d}^{(K)},I) &\propto \prod_{k=1}^{K} p_{(k)}(\boldsymbol{\theta}|\mathbf{d}^{(k)},I) \\ &\propto \prod_{k=1}^{K} p_{(k)}(\mathbf{d}^{(k)}|\boldsymbol{\theta},I)p_{(k)}(\boldsymbol{\theta}|I). \end{aligned} \tag{165}$$

This overcomes the problem in linear opinion pooling of accidentally assigning a large probability to a parameter vector θ which is contradicted by one or more sensor recordings. On the other hand, the simple multiplication of posterior distributions gives undue credence to the prior distribution in the standard case when there is common prior information about the physical object under investigation $p_{(k)}(\theta|I)=p(\theta|I)$, since it enters K times.

c. Pragmatic approach

In traditional diagnostic data analysis, physical parameters are evaluated using separate models tied to the individual diagnostics. Interdependencies between the diagnostic models are then treated in an iterative fashion, where the output from one diagnostic model is used as the input for the other models in the next iteration (see Fig. 34) after taking into account additional constraints (e.g., positivity). This cycle is repeated until convergence (i.e., consistency) is achieved. However, this common approach has several drawbacks:

- The same data set may lead to different results depending on the ordering of the parameter updates.
- When many diagnostic models are interdependent through common physical parameters, those diagnostic models all provide information about (and thus modify) parameters that in the traditional approach are merely used as fixed inputs from the previous iteration.
- Reliable uncertainties (and correlations) of the parameter estimates are hardly accessible due to the lack of a common model.
- The iteration cycle is very time consuming and often requires repeated human input.

Especially the last point leads to a growing mismatch between data collection and data evaluation.

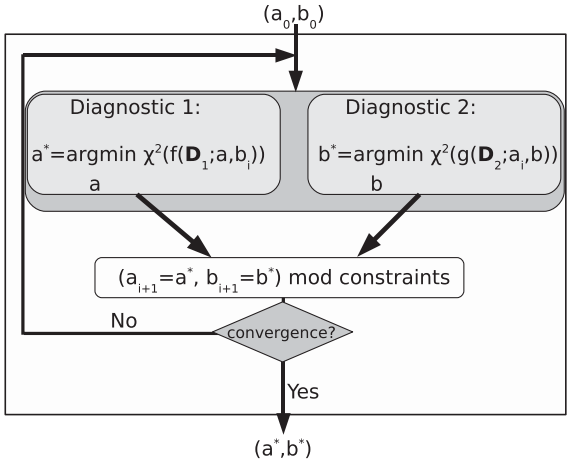

FIG. 34. Traditional approach to data fusion; based on some initial guess of the required parameters the respective data of each diagnostic are evaluated to yield individual best-fit parameters. These are combined, and additional constraints (e.g., positive density) are taken into account and used as starting parameters for the next iteration cycle. This process is repeated until a self-consistent solution is obtained.

d. Independent likelihood pool

From a Bayesian point of view the data fusion problem can be addressed using Bayes' theorem (Manyika and Durrant-Whyte, 1995; Fischer *et al.*, 2003). The likelihoods of the individual diagnostics depend on the state of the physical object to be investigated $p_{(k)}(\mathbf{d}^{(k)}|\boldsymbol{\theta}, I)$; therefore, the prior information about the physical object is the same for all diagnostics $p_{(k)}(\boldsymbol{\theta}|I) = p(\boldsymbol{\theta}|I)$ yielding as posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(K)}, I) \propto p(\boldsymbol{\theta}|I) \prod_{k=1}^{K} p_{(k)}(\mathbf{d}^{(k)}|\boldsymbol{\theta}, I) \qquad (166)$$

For the independent likelihood pool to be valid the conditional independence of the observations has to be verified, i.e., no hidden shared parameter dependences should exist. This requires careful specification of the likelihood models.

The difficulties of proper integration of data within the iterative approach increase tremendously with the number of heterogeneous diagnostics to be combined. Starting around 1985 in many areas [e.g., astrophysics (Obric *et al.*, 2006), geophysics or remote sensing (Pohl and Van Genderen, 1998; Wald, 1998; Quartulli and Datcu, 2003), robotics (Thomopoulos, 1990; Manyika and Durrant-Whyte, 1995), and defense (Hall, 2004)], the requirement of combining multisensor data was recognized as vital and became a strong focus of research. Also the fusion physics community with the requirement of linking on the order of 100 widely different diagnostic instruments was experiencing a strong need to develop a systematic approach for joint data evaluation. Similar to earlier experiences in the robotics community it soon emerged that the combination of heterogeneous diagnostics within a Bayesian framework was often conceptually the easiest one, although the actual computations can be demanding. In the following, two applications of Bayesian data integration to fusion research are presented.

B. Application in fusion research

Starting around 2000 a Bayesian framework named integrated data analysis (IDA) for magnetic confinement fusion experiments (such as ASDEX-Upgrade or W7-X) was developed (Fischer *et al.*, 2003; Dinklage *et al.*, 2004; Fischer and Dinklage, 2004) which several years later was also extended to JET (Arshad *et al.*, 2007; Svensson and Werner, 2007). The key idea of the IDA framework is to reformulate the set of individual inference problems of each diagnostic as a single (Bayesian) inference problem on the unknown state of the plasma being investigated, therefore essentially implementing the independent likelihood pool concept. This approach is schematically visualized in Fig. 35: Starting with a prior distribution of the unknown physics parameters the posterior distribution is computed using the combined data of the available different diagnostics taking advantage of all interdependencies. A stringent prerequisite of this approach is a careful assessment of all systematic and statistical errors which have to be incorporated into the model, often leading to non-Gaussian likelihoods. But, in turn, the combination of different diagnostics has the potential to validate measurements or detect insufficient or incomplete models.

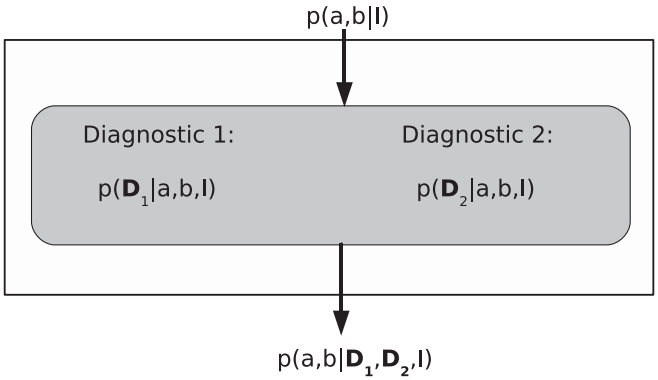

FIG. 35. Bayesian approach to data fusion; based on the prior distribution of the required parameters the respective likelihood distribution of each diagnostic is evaluated. Then, using Bayes theorem the posterior distribution of the parameters is computed, automatically taking into account all interdependencies of the diagnostics.

1. Thomson scattering and soft x ray at W7-AS

A striking example with a, at first glance, counterintuitive result is provided by the combination of the soft x ray and the Thomson scattering diagnostics of the stellarator W7-AS (Fischer *et al.*, 2002, 2003). Two of the key quantities in the description of confined plasmas are the electron temperature T_e and the electron density n_e. Both are accessible using Thomson scattering of intense laser light (Sheffield, 1975). The forward model is given by

$$d_{\mathrm{Th}} = c_{\mathrm{geom}} P n_e \sigma_{\mathrm{Th}} \int d\lambda \tau(\lambda) S(\lambda, T_e, \theta), \qquad (167)$$

where P is the laser power and $\sigma_{\mathrm{Th}} = (8\pi/3) r_e^2$ is the total Thomson scattering cross section for a single electron and r_e corresponds to the classical electron radius. The geometry factor c_{geom} considers both imaging effects and the overall sensitivity of the detection system. $\tau(\lambda)$ is the wavelength dependent spectral transmission of the interference filters, and $S(\lambda, T_e, \theta)$ is the scattering function of the scattered light (Fischer *et al.*, 2002) approximated by the analytical formula of Matoba *et al.* (1979). In Fig. 36 a typical posterior distribution of the electron density derived from Thomson scattering data is displayed as a dashed line. For the case at hand the electron density is asymmetric with a mode at $n_e = 15 \times 10^{19}$ m^3 and a pronounced tail toward higher densities. A joint evaluation with the soft x-ray diagnostic which provides (in first approximation) only information about the electron temperature (represented by the dotted line in Fig. 36) results in a reduction of the uncertainty of the electron density by 30% due to the strong suppression of the high density part of the distribution. The explanation of this surprising result is based on the correlation between electron density and electron temperature. A closer inspection of Eq. (167) reveals that the data are determined by the product of electron density and the spectral distribution S, which depends on electron temperature. This correlation is clearly displayed in the 2D posterior distribution $p(n_e, T_e|d)$ (Fig. 37) derived from Thomson scattering data of a Wendelstein 7-AS plasma discharge. The hyperbolic shape of the posterior distribution explains the empirical

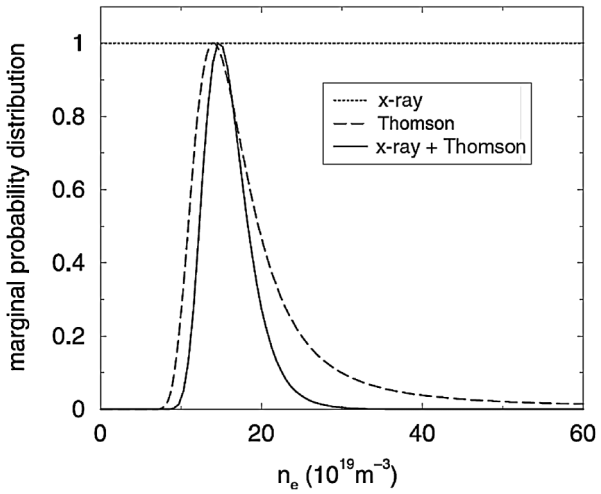

FIG. 36. Posterior distribution of electron density at the position $z = 6$ cm for a plasma discharge derived from combined evaluation of Thomson scattering data and x-ray diagnostic.

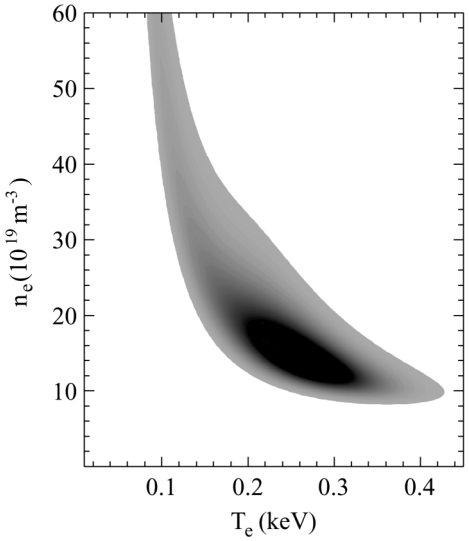

FIG. 37. Posterior distribution of electron temperature and electron density at the position $z = 6$ cm for a plasma discharge derived from Thomson scattering data.

observation that the plasma pressure $p_e = n_e T_e$ is often well reflected in the Thomson data despite displaying large uncertainties in both electron temperature and electron density. At low electron temperatures of around 100 eV the electron density is highly uncertain and extends beyond $n_e = 60 \times 10^{19}$ m^3, yielding the heavy tail in the marginal electron density distribution shown in Fig. 36 as a dashed line. Figure 38 shows the two-dimensional posterior distribution from soft x-ray measurements, where the electron temperature is determined by comparing the soft x-ray continuum emissivity measured with two different spectral edge filters (Fischer *et al.*, 2003). The temperature is determined to be 0.25 ± 0.02 keV (2σ). Since the soft x-ray measurement does not provide any information about n_e, the probability density in Fig. 38 is a vertical bar which is invariant in the n_e direction. Combining the Thomson scattering results and the soft x-ray data using the independent likelihood pool

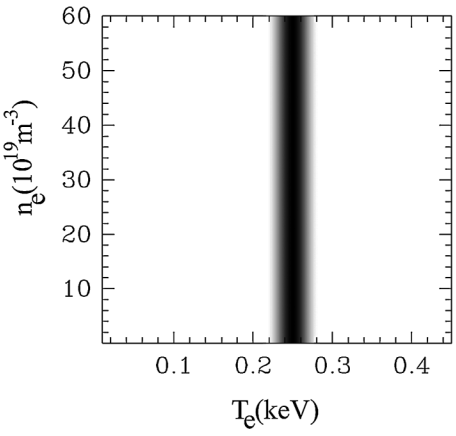

FIG. 38. Posterior distribution of electron temperature given by a soft-x-ray diagnostic. The constant values along the vertical direction indicate that the measurement provides no information about the electron density n_e.

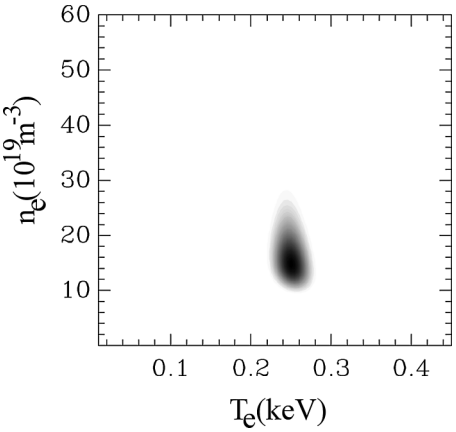

FIG. 39. Two-dimensional posterior distribution of electron temperature and electron density at the position $z = 6$ cm for a plasma discharge derived from Thomson scattering data.

(or IDA) approach of multiplying the likelihood distributions yields the two-dimensional posterior distribution displayed in Fig. 39. As expected, the uncertainty of the electron temperature estimate is reduced by the inclusion of the precise soft x-ray data. However, the density estimation is also affected by the temperature measurement of the soft x-ray diagnostic since the additional T_e data lead to the suppression of the low-T_e, high-n_e part of the probability density function. In this example the low dimensionality of the parameter space provides easy access to the benefits of a proper consideration of parameter correlations. In problems with more parameters the implications of the correlation structure are harder to grasp, but are automatically taken into account by the IDA approach.

2. Bayesian graphical models for diagnostics

When attempting to integrate large sets of diagnostics, the number of interdependencies between the parameters increases rapidly. Here the representation of the system by Bayesian graphical models can be very helpful and also provide some insight into less obvious interrelations,

although all inference on proper probability distributions in Bayesian probability theory, no matter how complex, amounts to repeated application of the sum rule and the product rule. Therefore, purely algebraic manipulation is sufficient to solve complicated probabilistic models (Bishop, 2006). However, Bayesian graphical models (Pearl, 1988; Lauritzen, 1996; Cowell *et al.*, 1999; Pearl, 2000; Jensen, 2001; Jordan, 2004; Darwiche, 2009) (also called Bayesian networks) provide an efficient language to visualize the dependencies in complex models and may be used to express complex computations in terms of intuitive graphical manipulations of the graph structure. Bayesian networks consist of nodes connected by directed links, where each node represents a random variable and the directed links (arrows) express the probabilistic relationship between these variables. If there is an arrow from node x to node y, x is said to be a parent of y. Each node x_i has a conditional probability distribution $p(x_i|q_i)$ that quantifies the effect of the parents on the node q_i representing the set of parents of x_i. Bayesian graphical models are restricted to directed acyclic graphs (DAG). A directed graph is called acyclic if there is no directed path between the nodes A_i such that $A_1 \rightarrow \cdots \rightarrow A_n$ subject to $A_1 = A_n$ (Jensen, 2001). In this case the joint distribution defined by a graph is given by the product over all of the nodes of the graph

$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i|q_i), \tag{168}$$

where $\mathbf{x} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$. A simple Bayesian graphical model (DAG) with 4 nodes is shown in Fig. 40. In general the joint probability distribution $p(x_1, x_2, x_3, x_4)$ can be expressed as

$$p(\mathbf{x}) = p(x_4|x_1, x_2, x_3)p(x_3|x_1, x_2)p(x_2|x_1)p(x_1) \tag{169}$$

using the chain rule of probability. By using the conditional independence relationships encoded in the network shown in Fig. 40 a slightly simpler representation of the joint probability distribution is obtained with the help of Eq. (168).

$$p(\mathbf{x}) = p(x_4|x_2, x_3)p(x_3|x_1)p(x_2|x_1)p(x_1), \tag{170}$$

a slightly simpler representation of the joint probability distribution. However, for large sparse network huge savings may be realized.

In most physics based data analysis problems the specification of the topology of the corresponding Bayesian network is straightforward: The intuitive meaning of an arrow in a properly constructed network is usually that the parent node has a direct influence on the successor nodes (Russell and Norvig, 2003). Therefore, the natural sequence of building a graphical model is to add the root causes (the underlying physical quantities such as, e.g., electron density) first, then the parameters they influence until the last nodes (the measured data) are reached, which have no direct causal influence on the other variables. Once the network has been formulated conditional independence (Dawid, 1979) between parts of the network can be determined using concepts such as d separation (Pearl, 1988) or Markov blankets. In favorable circumstances the exploitation of conditional independence may reduce the numerical effort of inference in Bayesian networks by orders of magnitude. The most common algorithm used for exact inference in general DAGs is the junction tree algorithm (Lauritzen and Spiegelhalter, 1988; Aji and McEliece, 2000; Kschischang *et al.*, 2001), which is based on a conversion of the DAG into an undirected graph with treelike topology (Huang and Darwiche, 1996; Cowell *et al.*, 1999) by clustering nodes together.

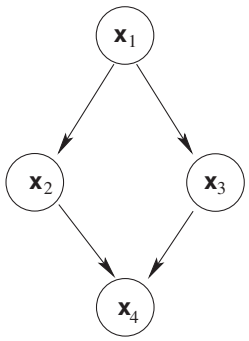

FIG. 40. Representation of a Bayesian network as a directed acyclic graph (DAG), the causal directions indicated by arrows.

However, exact inference in Bayesian networks can have exponential complexity. It can be shown that exact inference in Bayesian networks is NP hard (Dagum and Luby, 1993) or even number P hard, (number P hard, strictly harder than NP-complete problems) (Cooper, 1990; Mertens, 2002; Russell and Norvig, 2003). For many problems of practical interest, it is therefore necessary to exploit various approximation methods as follows:

- Loopy belief propagation. For treelike graphs effective and exact local message-passing algorithms for inference exist (belief propagation) (Pearl, 1988; Peot and Shachter, 1991). The idea of loopy belief propagation is simply to apply the belief propagation algorithm also to networks which are not treelike. In some applications this approach has empirically proven to be very effective (McEliece *et al.*, 1998; Murphy *et al.*, 1999), which was subsequently partly made transparent by theoretical analysis (Weiss, 2000).
- Variational methods. The true probability distributions p are replaced by (factorized) variational distributions q and the Kullback-Leibler (KL) divergence $(p \parallel q)$ is minimized (Jordan, 1999). The variational framework may be applied on a global level (approximating the full posterior distribution over all random variables) or on the conditional distributions of individual nodes or groups of nodes until a tractable approximation is obtained. For more details see Sec. IV.B.2 or Jordan *et al.* (1999), Jaakkola and Jordan (2000), Jaakkola (2001), and Wainwright and Jordan (2003).
- Sampling (Monte Carlo) methods. The simplest kind is importance sampling, where random samples are generated from the distribution of the "root nodes," and then weighted by their likelihood (Shachter and Peot, 1989; Cheng and Druzdzel, 2000). Other approaches are based on various MCMC techniques (cf. Sec. IV.D) (Pearl, 1987), especially Gibbs sampling (Gilks *et al.*, 1994, 1996) has been widely used.

The graphical framework provides a way to view many different algorithms (e.g., mixture models, factor analysis,

Kalman filters, hidden Markov models) and inference problems as special instances of a common underlying formalism (Jordan, 1999; Roweis and Ghahramani, 1999) and to take advantage of the specialized techniques developed for graph structures. However, so far the use of graph-theoretical algorithms (besides simple visualization of the problem as a Bayesian network) in physics based problems is very limited. On the one hand, this is due to a good understanding of the problems [the causal structure is very often known, in contrast to other areas such as sociology (Pearl, 2000; Spirtes *et al.*, 2000)] thus ''automatically'' leading to a very good representation with very little room for improvement. On the other hand, many problems of interest were simply too large to be handled. Here the increase of computing capabilities is providing some perspective, where especially the demanding problem of learning the underlying graph structure from observations (Murphy, 2001; Friedman and Koller, 2003) may benefit. One of the few examples of the application of a Bayesian graphical model on a larger scale is given by Dinklage, Fischer, and Svensson (2003). There the directed acyclic Bayesian graphical model for several diagnostics at the W7-AS stellarator is derived. The model includes the Thomson scattering diagnostic, a microwave interferometer, and diamagnetic energy measurements. Despite the relatively small number of diagnostics, the derived graphical model covers a whole page in Dinklage, Fischer, and Svensson (2003) and is of surprising complexity. A fundamental cause of the complexity of the acyclic graph is the required mapping of different diagnostic signals to a common magnetic coordinate system. This mapping is obtained from plasma equilibrium calculations which in turn depend on the measured data. The mapping will therefore be uncertain and increase the uncertainty of the parameter estimation. Thus to obtain a self-consistent solution the integrated model must include a module for calculating the magnetic coordinate system from estimated profiles (which themselves rely on the estimated mapping). It is worth noting that despite the seemingly circular reasoning the actual computation is going straight from the values for electron density, electron temperature, and ion pressure along the directed edges to the terminal nodes of the diagnostics without any feedback loops. The outcome of the graphical model is the joint probability distribution of model parameters and measured data. Inference of specific parameters has been done using standard-MCMC techniques. Figure 41 shows the closed flux surfaces together with their uncertainty (represented by the width of the flux surface) derived from a large number of posterior samples (Dinklage, Fischer, and Svensson, 2003; Svensson *et al.*, 2004) based on the analysis of the graphical model of the W7-AS diagnostics. Neglect of the position and angle dependent uncertainty of the flux surfaces results in a strong underestimation of the total uncertainty of the free parameters (electron density, electron temperature, and ion pressure).

Since the integration of different diagnostics within the independent likelihood approach is straightforward, once adequate (forward) models exist an increasing number of fusion experiments [e.g., ASDEX-UG (Fischer *et al.*, 2008), W7-X (Dinklage *et al.*, 2003), JET (Arshad *et al.*, 2007; Svensson and Werner, 2007), and MAST (Hole *et al.*, 2009)] are developing frameworks for Bayesian integrated data analysis. Additional challenges arise if real-time or monitoring requirements are present. Especially for superconducting tokamaks with longer pulse lengths or stellerators with quasicontinuous operation offline processing is not sufficient (Dinklage *et al.*, 2003). Here suited simplifications (e.g., Laplace approximation) have to be considered because in most cases the evaluation of the exact joint posterior distribution takes too long.

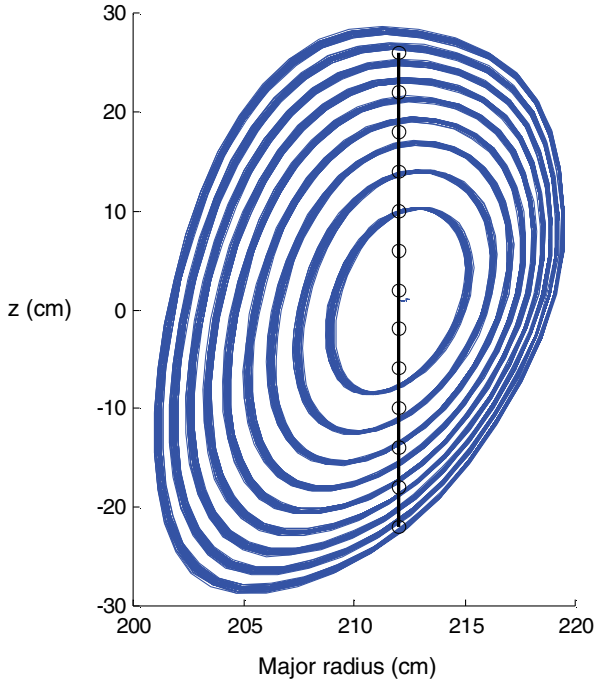

FIG. 41 (color online). Uncertainty of the magnetic flux surfaces represented by posterior samples. Adapted from Dinklage *et al.*, 2003b.

C. Application in robotics: SLAM

Simultaneous localization and mapping (SLAM) is one of the most fundamental problems in mobile, autonomous robotics and has been the subject of extensive research in the past decades (Leonard and Durrant-Whyte, 1992; Borenstein *et al.*, 1996; Bailey and Durrant-Whyte, 2006; Durrant-Whyte and Bailey, 2006). In the SLAM setting the robot is required to derive a map from its (noisy) perceptions and simultaneously determine its own position in the map. Since robot motion (or the perception) is inaccurate, the general problem of map building is an example of a chicken-and-egg problem: To determine the location of the landmarks, the robot needs to know where it is. To determine where it is, the robot needs to know the location of the landmarks (Thrun *et al.*, 1998). In the key paper by Smith *et al.* (1990) the SLAM problem was reformulated as a probabilistic state-estimation problem and it was shown that as the robot moves through an unknown environment taking relative measurements of points of interests, the estimates are correlated because of the common error in the estimated robot location. Subsequently it was derived that the structure of the SLAM problem critically depends on those cross correlations between landmark estimates (Dissanayake *et al.*, 2001), giving

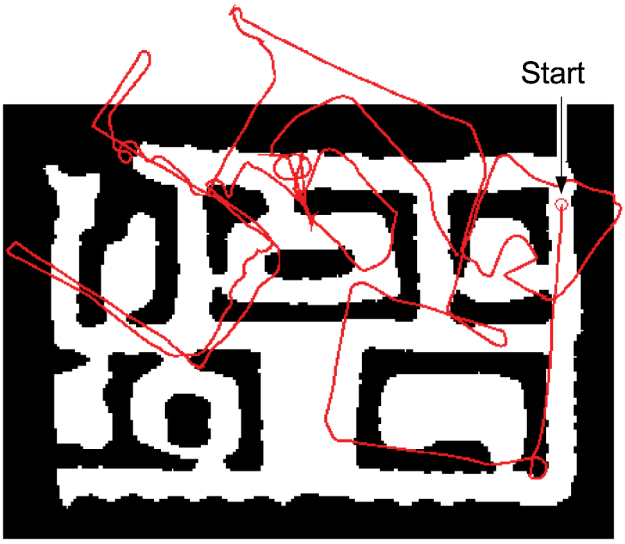

FIG. 42 (color online). Internal representation of the trajectory (continuous line) derived from odometer data of a robot exploring a building superimposed onto a map of the building (dark areas represent walls). The starting point is in the upper right corridor of the building, indicated by an open circle. The mismatch is continuously increasing due to accumulation of orientation angle uncertainties. Adapted from Fox *et al.*, 1999.

rise to a structure which may be paraphrased as certainty of relations despite uncertainty of positions (Frese, 2006). Figure 42 shows this behavior. Starting in the upper right hallway (indicated by an open circle) the trajectory recorded by a mobile robot using odometric measurements is displayed as a continuous line superimposed on a map of the building. Since the measurement uncertainties (i.e., the orientation of the robot) add up, the global map shows a very large mismatch although the local relations are well preserved. From a statistical perspective the SLAM problem can be viewed as a temporal Bayesian inference problem, estimating the joint posterior

$$p(\mathbf{X}_{t+1}, M|\mathbf{z}_{1:t+1}, a_{1:t}) \tag{171}$$

for the actual position $\mathbf{X}_{t+1}$ and the map M from sequences of actions a (robot movements) and the $t+1$ measurements $\mathbf{z}_{1:t+1}$.

Often the model of the robot movement is simplified to two dimensions. Then the pose $\mathbf{X}_t = (x_t, y_t, \theta_t)$ of the robot is defined by its two Cartesian coordinates (x, y) and its heading with value θ at instant t. An action a (a motion command) consists of a combination of rotational and translational motion. Since the robot's motion is inaccurate, the effect of a control signal a on the robot's location $\mathbf{X}$ is given by a probability density

$$p(\mathbf{X}_{t+1}|\mathbf{X}_t, a_t), \tag{172}$$

which is commonly modeled by triangular (Thrun *et al.*, 1998) or Gaussian distributions (Russell and Norvig, 2003). In Fig. 43, the probability density of the robot's location after some actions a is visualized using 2D projections onto the x-y plane. The particular shape of the distribution results from accumulated translational and rotational uncertainties as the robot moves.

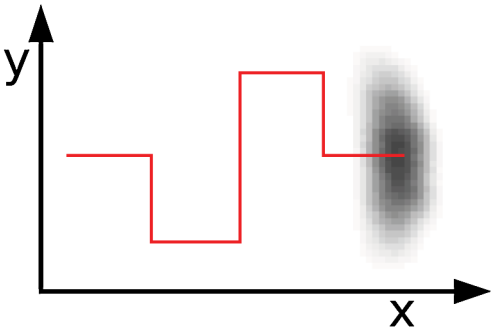

FIG. 43 (color online). 2D projection of the 3D-probability distribution visualizing the accumulated localization uncertainty after moving from a precisely known initial position with several 90° turns. The robot is more likely to be in darker areas. The line describes the ideal trajectory of the robot. Adapted from Fox *et al.*, 1999.

As a model for the robot's perception it is assumed that the robot can observe and identify landmarks, and the sensors report the range and bearing of the landmarks. It is also assumed that the perceptual component suffers from Gaussian noise. Then the model of the robot's perception can be described by

$$p(\mathbf{z}_t|\mathbf{x}_t, M) = N(\hat{\mathbf{z}}_t, \Sigma_z), \tag{173}$$

where $\hat{\mathbf{z}}_t = \binom{r}{\Delta\theta}$ is given by the column vector of the distance r between the robot and the landmark and the relative angle $\Delta\theta$

$$\hat{\mathbf{z}}_t = h(\mathbf{x}_t) = \begin{pmatrix} r \\ \Delta\theta \end{pmatrix} = \begin{pmatrix} \sqrt{(x_t - x_i)^2 + (y_t - y_i)^2} \\ \arctan[(y_i - y_t)/(x_i - x_t)] - \theta_t \end{pmatrix} \tag{174}$$

for a robot at $\mathbf{x}_t = (x_t, y_t, \theta_t)^T$ and a landmark at (x_i, y_i). The new estimate $p(\mathbf{X}_{t+1}, M|\mathbf{z}_{1:t+1}, a_{1:t})$ can then be computed from the current state $p(\mathbf{X}_t, M|\mathbf{z}_{1:t}, a_{1:t-1})$ and the observation $\mathbf{z}_{t+1}$ using

$$p(\mathbf{X}_{t+1}, M|\mathbf{z}_{1:t+1}, a_{1:t}) \propto p(\mathbf{z}_{t+1}|\mathbf{X}_{t+1}, M) \int d\mathbf{x}_t p(\mathbf{X}_{t+1}|\mathbf{x}_t, a_t) \times p(\mathbf{x}_t, M|\mathbf{z}_{1:t}, a_{1:t-1}). \tag{175}$$

Equation (175) provides a straightforward recursive recipe for updating. However, the dimensionality of the state space is large: The number of landmarks (n) and robot poses (p) may be of the order of thousands yielding $N = 2n + 3p$ unknown parameters. Furthermore, n and p may depend on the unknown size of the environment to be explored. Thus, although Eq. (175) provides a formal solution to the SLAM problem, the search for efficient algorithms suited for autonomous robots is still ongoing.

Several methods have been proposed for approximate solutions of Eq. (175), most notably the extended Kalman filter (EKF) and Monte Carlo particle filter.

In the extended Kalman filter technique the joint probability density is approximated by a single (high-dimensional) Gaussian distribution and the nonlinear motion and sensor

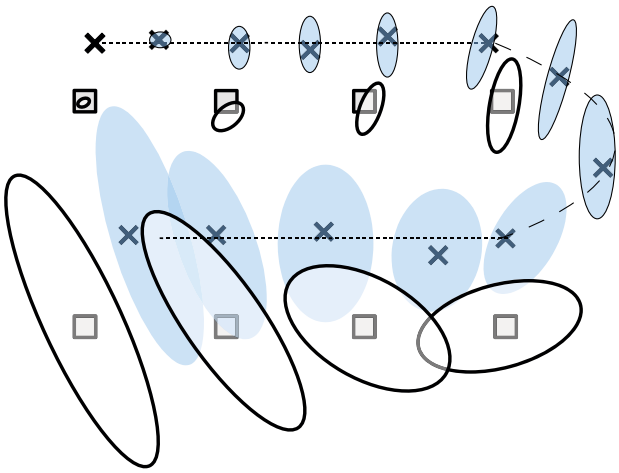

FIG. 44 (color online). The extended Kalman filter (EKF) applied to the robot mapping problem. The robot's path is indicated by the dotted line, and its estimations of its own position are shown as shaded ellipses which may deviate from the true position (indicated by crosses). Eight distinguishable landmarks of unknown location are indicated by small squares, and their location estimates by the robot are shown as open ellipses. During the path the robot's positional uncertainty is increasing, as is its uncertainty about the landmarks it encounters. Adapted from Russell and Norvig, 2003.

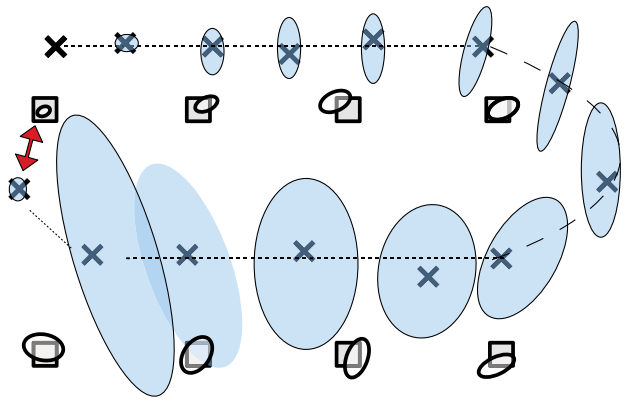

FIG. 45 (color online). Once the robot senses again the first landmark (indicated by an arrow) the uncertainty of all landmark estimates decreases (indicated by the reduced size of the unfilled ellipses), thanks to the fact that all the estimates are correlated. In addition the new position estimate of the robot itself is also much more precise. Adapted from Russell and Norvig, 2003.

models are linearized using Taylor expansion. Using this first-order approximation the integration in Eq. (175) can be done analytically, yielding a Gaussian probability distribution with modified mean and covariance for the map and position estimates in each step. This approach maintains all dependencies between the variables requiring updates of an $N \times N$-covariance matrix and resulting in computing times of $O(N^2)$. To avoid the quadratic complexity other approaches exploit, e.g., the limited field of view due to obstacles or sensor range. Thus, at any point in the environment only a few landmarks in the vicinity of the robot are observable. This number (k) depends on the environment and the sensor, but it does not grow when the maps get larger. An overview of different recent algorithms has been given by Frese (2006).

An example of the EKF in a SLAM setting is shown in Fig. 44: It shows an environment with eight landmarks, arranged in two rows. The robot starts at a well-defined location, its own positional uncertainty indicated by the shaded error ellipses. As the robot moves it gradually loses certainty as to where it is. This uncertainty is also transferred to the position of the detected landmarks (open error ellipses). When the first landmark is detected again, the position uncertainty of all landmarks decreases since the estimates of robot and landmark positions are highly correlated (cf. Fig. 45). Another algorithmic approach to a computational efficient estimate of Eq. (175) is based on particle filters (cf. Sec. IV.E.2.b). The insight that the landmark estimates are conditionally independent given the robot's pose was used by Montemerlo *et al.* (2002) to derive an algorithm called fastSLAM. The conditional independence allows the representation of the probability distribution by particles where each particle represents a sampled robot trajectory and associated Gaussian distributions of the landmark positions without requiring a full-system-size covariance matrix. This reduces the dimensionality of the probability space to be covered by the particles. The time to integrate a new measurement scales as $M \log n$, with M the number of particles used to represent the probability distribution. Additionally, the fastSLAM algorithm can accommodate nonlinear sensor and motion models. However, it may suffer from a loss of particle diversity (Bailey and Durrant-Whyte, 2006): The particle filter is working in the space of the robot trajectory and the number of particles required to maintain a certain coverage of this space is therefore exponential in the length of the trajectory. A smaller number of particles may no longer represent the probability distribution with sufficient accuracy and may become inconsistent. Therefore, hybrid approaches have been suggested to combine the strengths of EKF (long-term memory) and the flexibility of the particle filter approach (Brooks and Bailey, 2008), enabling the mapping also in complex environments with a large number of landmarks. Other work focuses on the integration of the Global Positioning System based data into the SLAM inference problem using a sparse graphical network (Golfarelli *et al.*, 1998) representation, generating maps with up to $O(10^8)$ features (Thrun and Montemerlo, 2006). Proper handling of dynamic environments (e.g., crowded places) is still an active research topic although some progress has been made (Fox *et al.*, 1999; Thrun and Fox, 2005). Finally, there are quite recent attempts to use decision-theory based reasoning when active loop closing is beneficial for improved mapping (Ji *et al.*, 2009), a special case of experimental design which will be discussed in the next section.

VII. BAYESIAN EXPERIMENTAL DESIGN

In the previous sections about parameter estimation and model comparison many examples demonstrated how Bayesian probability theory can be used for quantitative inference based on prior knowledge and measured data. However, Bayesian probability theory is not a magic black box guaranteed to compensate for badly designed experiments. Information absent in the data cannot be revealed by

any kind of data analysis. This immediately raises the question of how the information provided by a measurement can be quantified and, in the next step, how to optimize experiments to maximize the information gain. Here one of the very recent areas of applied Bayesian data analysis is entered: Bayesian experimental design is an increasingly important topic driven by progress in computer power and algorithmic improvements (Loredo and Chernoff, 2003a).

So far it has been implicitly assumed that there is little choice in the actual execution of the experiment; in other words, the data to be analyzed were assumed to be given. While this is the most widespread use of data analysis, the *active* selection of data is holding great promise to improve the measurement process. There are several scenarios in which an active selection of the data to be collected or evaluated is obviously very advantageous, e.g., expensive and/or time-consuming measurements; thus one wants to know where to look next to learn as much as possible, or when to stop performing further experiments; design of a future experiment to obtain the best performance (information gain) within the scheduled experimental scenarios; selection of the most useful data points from a large amount of data; and an optimal (most informative) combination of different experiments for the quantity of interest.

A. Overview of the Bayesian approach

The theory of frequentist experimental design dates back to the 1970s (Chernoff, 1972; Fedorov, 1972). About the same time the Bayesian approach was put forward by the influential review of Lindley (1972). Lindley's decision-theoretic approach involves the specification of a suitable utility function $U(\mathbf{D}, \eta)$ which depends on the result (data) of an experiment and the design parameters η. Design parameters are understood as parameters of an experiment which are accessible and adjustable. Examples are the point in time for the next measurement or the analysis beam energy. The utility function has to be defined with respect to the goals of the experiment and cannot be derived from first principles. It may contain considerations about the cost of an experiment or the value of a reduced uncertainty of a parameter estimation. For a discussion about the formal requirements for utility functions, see Bernardo and Smith (2000). The experimental design decision η^* (e.g., where to measure next) which maximizes the chosen utility function $U(\mathbf{D}, \eta)$ is the optimal design. However, the data $\mathbf{D}$ are uncertain before the actual measurement due to statistical or systematic uncertainties and an incomplete knowledge about the parameters of the physical system. Therefore, the Bayesian experimental design has to take into account all possible data sets and the utility function has to be marginalized over the data space, which results in the expected utility (EU)

$$\mathrm{EU}(\eta) = \int d\mathbf{D}\, p(\mathbf{D}|\eta, I) U(\mathbf{D}, \eta). \tag{176}$$

As can be seen from Eq. (176) the expected utility is the integral over all possible data weighted by the probability of the data under the design decision η and the utility of the corresponding data. The required predictive distribution $p(\mathbf{D}|\eta, I)$ is not immediately accessible but can be expressed in terms of likelihood and prior, both are depending on the parameter vector $\boldsymbol{\theta}$

$$p(\mathbf{D}|\eta, I) = \int d\boldsymbol{\theta}\, p(\mathbf{D}|\boldsymbol{\theta}, \eta, I) p(\boldsymbol{\theta}|I). \tag{177}$$

Substituting $p(\mathbf{D}|\eta, I)$ into Eq. (176) by Eq. (177) yields

$$\mathrm{EU}(\eta^*) = \max_{\eta} \int d\mathbf{D} \int d\boldsymbol{\theta}\, p(\mathbf{D}|\theta, \eta, I) p(\boldsymbol{\theta}|I) U(\mathbf{D}, \eta) \tag{178}$$

for the best experimental design decision. Therefore, the expected utility can be expressed in terms of likelihood and prior distributions combined with a suitable utility function $U(\mathbf{D}, \eta)$. The evaluation, however, requires nested integrations. Only for very few cases (almost always involving Gaussian likelihoods and linear models) can the integration over parameter space and data space be performed analytically.

B. Optimality criteria and utility functions

The most widely used optimality criteria for experimental design are derived from various desirable properties of parameter estimates of linear models: Minimizing the average variance of the best estimates of the regression coefficients by minimizing the trace of the variance-covariance matrix is called *A* optimality (Fedorov, 1972; Steinerg and Hunter, 1984). The sometimes harmful neglect of the parameter covariances in *A* optimality motivates *D* optimality, where the determinant of the variance-covariance matrix is minimized (Steinerg and Hunter, 1984). Other optimality criteria focus on the variance of predictions instead of the variance of the parameter estimates. For an overview of the various optimality criteria see, e.g., Pukelsheim (1993) and Atkinson *et al.* (2007) and the relationship between frequentist and Bayesian optimality criteria is discussed by Chaloner and Verdinelli (1995) and DasGupta (1996). However, the focus on the best estimate $\boldsymbol{\theta}^*$ only as a basis for experimental design does not take into account the full information content of the probability distribution of the parameters in nonlinear settings. The suggestion of Lindley (1956) to use the information gain of an experiment as a utility function has been followed by several (Stone, 1959; DeGroot, 1962; Bernardo, 1979a; Luttrell, 1985; DeGroot, 1986; Loredo and Chernoff, 2003a). The information gain is given by the expected Kullback-Leibler divergence between the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D}, \eta, I)$ and the prior distribution $p(\boldsymbol{\theta}|I)$:

$$\mathrm{U}_{\mathrm{KL}}(\mathbf{D}, \eta) = \int d\boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{D}, \eta, I) \log \frac{p(\boldsymbol{\theta}|\mathbf{D}, \eta, I)}{p(\boldsymbol{\theta}|I)}. \tag{179}$$

For the standard Gaussian linear regression model this utility function yields the same results as using a Bayes *D*-optimality criteria for design (Chaloner and Verdinelli, 1995). A discussion of some of the properties of the Kullback-Leibler divergence as a utility function in experimental design has been given by MacKay (1992b). From a theoretical point of view the decision-theoretic formulation of experimental design is well understood. Nevertheless, nonlinear experimental design methods received only little

attention within and outside the physics community [see, e.g., statements about lack of real applications by Goldstein (1992), Chaloner and Verdinelli (1995), and Toman (1999)], until Loredo and Chernoff (2003b) published an illustrative example about optimization of observation times in astronomy highlighting the potential and feasibility of nonlinear Bayesian experimental design,

C. Adaptive exploration for extra-solar planets

The search for extra-solar planets is one of the foci of astrophysical research programs, supported by space based missions such as Kepler (Koch *et al.* (2010)) or by ground instruments such as HARPS (high-accuracy radial velocity planet searcher). But the necessary high-accuracy measurements are time consuming, seriously restricting the number of stars that can be examined in search of extra-solar planets. Observation time is thus a precious resource that must be carefully allocated. Therefore, observations of stars with companions should be scheduled optimally to determine the orbital parameters with the fewest number of observations. Loredo and Chernoff (2003a) addressed the problem of determining the best time for the next measurement of the radial velocity of a star known to have a single planetary companion. The time-dependent radial velocity is a nonlinear function given by

$$v(t) = v_0 + K\{e\cos\omega + \cos[\omega + v(t)]\}, \tag{180}$$

where the true anomaly $v(t)$ can be computed by the joint solution of two nonlinear equations for the eccentric anomaly [an angular parameter that defines the position of a body that is moving along an elliptic Kepler orbit (Murray and Dermott, 1999; Wright and Howard, 2009)],

$$E(t) - e\sin[E(t)] = 2\pi t/\tau - M_0, \tag{181}$$

and

$$\tan\frac{v(t)}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E(t)}{2}. \tag{182}$$

The six parameters of the model are the orbital period τ, the orbital eccentricity e, the velocity amplitude K, the center-of-mass velocity of the system v_0, the mean anomaly at $t = 0$, M_0, and the argument of the pericenter ω. Please note that the parametrization differs slightly from the one used in the parameter estimation example (see Sec. III.B.1). Loredo and Chernoff (2003a) simplified the treatment, taking into account only three of the six parameters and assuming Gaussian additive noise ϵ with a standard deviation σ, so that the measured datum d_i at time t_i is given by

$$d_i = v(t_i;\tau, e, K) + \epsilon_i. \tag{183}$$

For parameter values of $\tau = 800$ d, $e = 0.5$, $K = 50$ m/s, and $\sigma = 8$ m/s a vector $\mathbf{d}$ of ten simulated observations was computed. In Fig. 46 the ten data points with error bars are displayed together with the true velocity curve. The posterior distribution of the parameters is given by Bayes' theorem as

$$p(\tau, e, K|\mathbf{d}, I) \propto p(\mathbf{d}|\tau, e, K, I)p(\tau, e, K|I), \tag{184}$$

where $p(\tau, e, K|I)$ is the prior probability density for the orbital parameters and $(\mathbf{d}|\tau, e, K, I)$ is the Gaussian likelihood function. For the data set shown in Fig. 46 rejection sampling

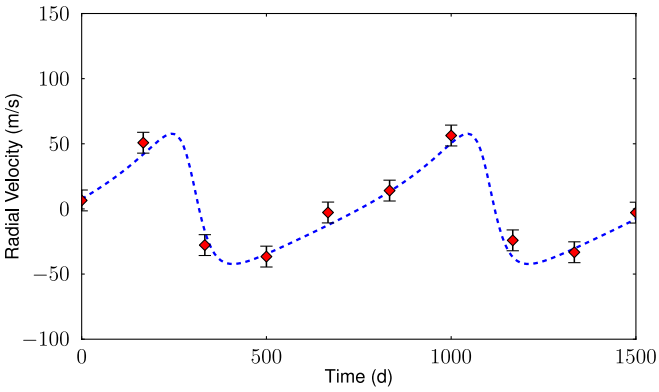

FIG. 46 (color online). The true velocity curve used for the generation of mock data is displayed as a dashed line. From this velocity curve ten simulated observations distorted with Gaussian noise were generated. These observations are given as points with error bars. From Loredo and Chernoff, 2003b.

was used to obtain independent samples from the posterior distribution. Figure 47 shows the τ and e coordinates of $N = 100$ such samples, displaying the two-dimensional marginal distribution $p(\tau, e|\mathbf{d}, I)$. The distribution is roughly located at the true values of $(\tau, e) = (800, 0.5)$ but with an asymmetric shape and still significant uncertainty. Based on the posterior distribution the predictive distribution $p(D|t, \mathbf{d}, I)$ for a datum D at a future time t can be computed. For given values of (τ, e, and K) the predictive probability density for D is a Gaussian centered at $v(t;\tau, e, K)$. The predictive distribution $p(D|t, \tau, e, K, \mathbf{d}, I)$ is thus given by the product of the Gaussian likelihood for D and the posterior distribution $p(\tau, e, K|\mathbf{d}, I)$. To account for the parameter uncertainty the model parameters have to be marginalized. Use of the posterior samples circumvents the time-consuming integration over the parameter space, since $p(D|t, \mathbf{d}, I)$ can be expressed as

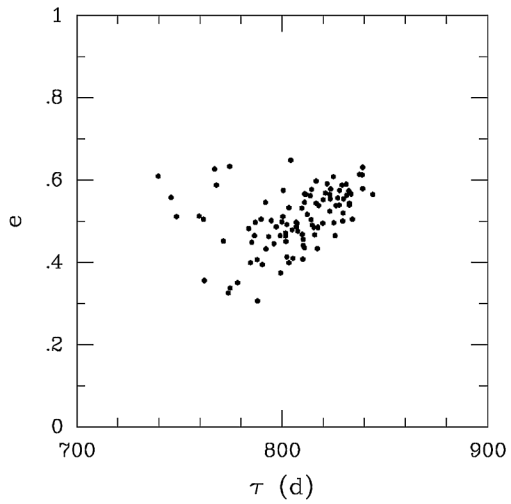

FIG. 47. Posterior samples from the probability distribution for two velocity curve parameters, the spread indicating the uncertainty in the marginalized parameters. The time for an orbital period is given by τ. The orbital eccentricity e describes the amount by which an orbit deviates from a perfect circle: $e = 0$ is a perfectly circular orbit and $e = 1$ corresponds to an open parabolic orbit. From Loredo and Chernoff, 2003b.

$$p(D|t,\mathbf{d},I) = \int d\tau \int de \int dK\, p(\tau,e,K|\mathbf{d},I) \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{[D-v(t;\tau,e,K)]^2}{\sigma^2}\right) \approx \frac{1}{N}\sum_{\tau_i,e_i,K_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{[D-v(t;\tau_i,e_i,K_i)]^2}{\sigma^2}\right).$$

The last line gives a Monte Carlo integration estimate of the predictive distribution using the independent samples from the posterior distribution. In Fig. 48 the velocity $v(t;\tau, e, K)$ is shown for the first 15 sampled parameter values (t; τ_i, e_i, and K_i) as thin solid lines. The spread of the velocity functions represents the uncertainty in the predictive distribution. The uncertainty is largest where the velocity changes most quickly and is slowly increasing with time since predictions with different periods fall increasingly out of synchronization. Once the predictive distribution is available the expected utility can be computed as a function of time using Eq. (176),

$$\mathrm{EU}(t) = \int dD\, p(D|t,\mathbf{d},I)U(D,t). \tag{185}$$

For the present problem where the width of the noise distribution does not depend on the underlying signal, it can be shown that the expected information gain is equal to the entropy of the predictive distribution (Sebastiani and Wynn, 2000)

$$\mathrm{EU}(t) = \int dD\, p(D|t,\mathbf{d},I)\log[p(D|t,\mathbf{d},I)]. \tag{186}$$

This equality is saving one (possibly high-dimensional) integration over the parameter space otherwise needed for the computation of the information-based utility function. Thus the best sampling time is the time at which the entropy (uncertainty) of the predictive distribution is largest. The thick line in Fig. 48 shows the estimate of EU(t) using base-2 logarithms so that the relative information gain is measured in bits (ordinate on the right-hand side). It is largest near the periastron crossings, thus recommending an additional observation at the maximum at $t = 1925$ d. Incorporating the new noisy datum measured at $t = 1925$ d into the posterior yields a significantly reduced uncertainty in the period estimate as a comparison of the posterior distributions without (see Fig. 47) and with the new data point (Fig. 49) reveals. The optimization procedure can be repeated and the well-chosen data points yield an increase in precision exceeding the rule-of-thumb $\sqrt{n}$ dependence often seen for random sampling (Loredo and Chernoff, 2003a).

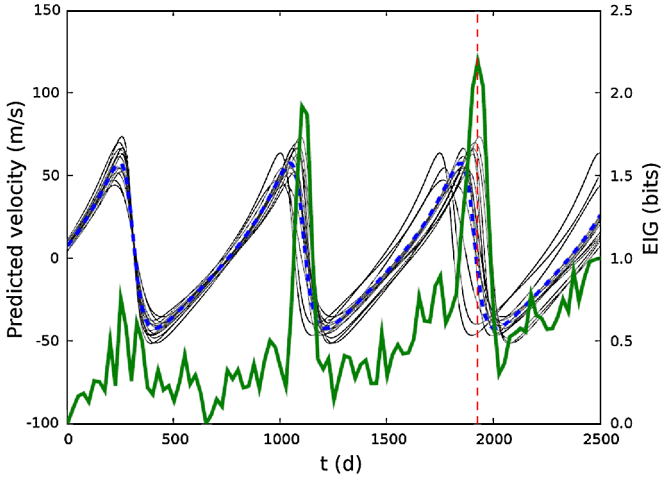

FIG. 48 (color online). The true velocity curve used for the simulation of the data is given as a dashed line. A posterior sample of 15 predicted velocity curves based on the observations is given by thin solid lines; their spread is a measure of the uncertainty of the predicted velocity. The expected information gain for a further measurement at each time is indicated by a thick solid curve; the information gain is given on the right axis.Note that the positions of largest uncertainty and largest information gain coincide. From Loredo and Chernoff, 2003b.

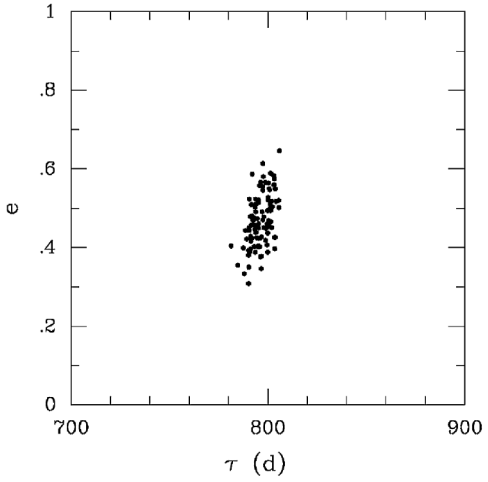

FIG. 49. Samples from the probability distribution for two velocity curve parameters after one additional measurement at the best time. Comparison with Fig. 47 reveals that the parameters have a significantly increased precision, the samples are less disperse. The time for an orbital period is given by τ. The orbital eccentricity e describes the amount by which an orbit deviates from a perfect circle: $e = 0$ is a perfectly circular orbit and $e = 1$ corresponds to an open parabolic orbit. From Loredo and Chernoff, 2003b.

D. Optimizing NRA measurement protocols

Nuclear reaction analysis (NRA) is a well-known technique for depth profiling of light elements in the near-surface region of solids (up to depths of several μm) using ion beams with energies in the MeV range. NRA measurements yield quantitative information on the isotopic depth distribution within the target and are highly sensitive. For an introduction to NRA for material analysis see, e.g.. Amsel and Lanford (1984) and Tesmer and Nastasi (1995).

The basic principle of NRA is straightforward: The sample is subjected to an energetic ion beam with an initial energy E_i^0 and an angle of incidence ϕ, which reacts with the species of interest. The products of the reaction are measured under a specified angle θ. Given the total number of impinging ions N_i, the energy dependent cross section of the reaction $\sigma(E)$, the efficiency of the detection, and the geometry of the setup μ, the measured total signal counts d_i depend (in the limit of

small concentrations) linearly on the concentration profile $c(x)$ of the species in depth x:

$$d_i = d(E_i^0) = \mu N_i \int_0^{x(E_i^0)} dx\sigma(E)c(x) + \epsilon_i$$
$$= \mu N_i \int_0^{x(E_i^0)} dx\sigma(E(x, E_i^0))c(x) + \epsilon_i. \quad (187)$$

The measurement uncertainty ϵ_i is approximated by a Gaussian distribution $\epsilon_i \propto N(0, \sigma_i)$. Repeated measurements with different initial energies E_i^0 provide increasing information about the depth profile of the species of interest. The optimization of NRA measurements of deuterium profiles for the weakly resonant nuclear reaction D + ^{3}He $\rightarrow$ p + ^{4}He + 18.352 MeV (Amsel and Lanford, 1984) has been studied by von Toussaint *et al.* (2008). The high Q value of 18.352 MeV provides an analysis depth for deuterium of several μm even in high-Z materials such as tungsten. Therefore, the reaction is commonly used to study the hydrogen isotope retention in plasma-facing components of fusion experiments (Skinner *et al.*, 2008). However, measurements are time consuming and the extraction of the concentration depth profile from the measured data is an ill-conditioned inversion problem due to the very broad cross section of the D(^{3}He, p)^{4}He reaction (Möller and Besenbacher, 1980). Therefore, the experimental setup (i.e., the choice of the analysis energies) should be optimized to provide a maximum of information about the depth profile. To evaluate Eq. (187) the energy $E(x)$ of the incident particle on its path through the sample for a given initial energy E^0 is required. The energy loss of the impinging ^{3}He ion in the sample is determined by the stopping power $S(E)$ of the sample

$$dE/dx = -S(E), \quad (188)$$

which can be solved to get the depth-dependent energy $E_i(x)$ for different initial energies E_i^0. Parametrizations and tables of S for different elements are given by Tesmer and Nastasi (1995). Since the amount of hydrogen in the sample is usually well below 1% (with the exception of a very thin surface layer), the influence of the hydrogen concentration on the stopping power can be neglected in most cases. A parametrization for the cross section $\sigma(E)$ (Alimov *et al.*, 2005) is provided by von Toussaint *et al.* (2008).

A tungsten sample (ρ = 19.3 g/cm^3) with a (high) surface concentration of 12% deuterium, followed by an exponentially decaying deuterium concentration profile down to a constant background level, described by

$$c(x) = a_0 \exp(-x/a_1) + a_2 \quad (189)$$

has been used. The parameter values are $a_0 = 0.1$, $a_1 =$ 395 nm, and $a_2 = 0.02$. The corresponding mock data for a set of initial energies $E^0 = \{500, 700, 1000, 1300, 1600, 2000, 2500, 3000\}$ keV is shown in Fig. 50. The variations in the detected yields reflect the interplay of the increasing range of the ions with increasing energy and the reduced cross section at higher energies modulated with the decreasing deuterium concentration at larger depths. The increase of the signal by raising the initial energy from 2500 to 3000 keV is caused by the constant deuterium background of 2%. The time needed to obtain the eight data points is around one working day taking into account necessary

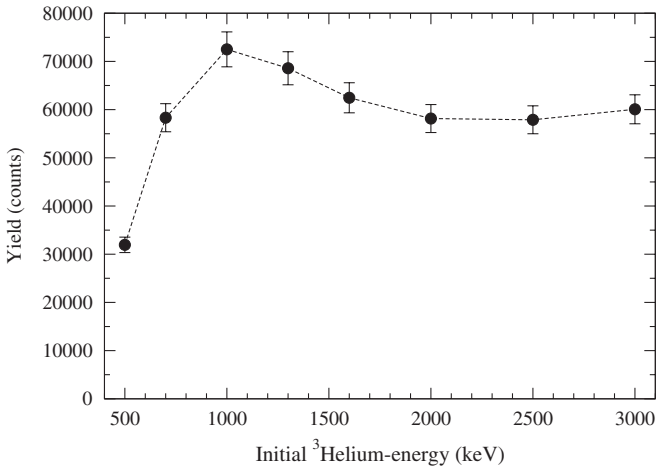

FIG. 50. Simulated yield data of a D(^{3}He, p)^{4}He nuclear reaction analysis of a tungsten sample with an exponentially decaying deuterium concentration profile. The varying intensity reflects the interplay of an energy dependent cross section, ion beam range, and concentration profile. Adapted from von Toussaint *et al.*, 2008.

calibration measurements. The uncertainty of the measurement is given by Poisson statistics. However, fluctuations in the beam current measurements are very often the dominating factor, affecting the prefactor N_i in Eq. (187). An accuracy of up to 3% can be achieved in favorable circumstances (e.g., by using the number of Rutherford-scattered ^{3}He ions on a thin gold coating on top of the sample as reference). Therefore, a realistic estimate of the measurement uncertainty was assumed to be $\sigma_i = \max(5\% d_i, \sqrt{d_i})$. von Toussaint *et al.* (2008) favorably compared the best experimental design for a linear setting (assuming a piecewise linear concentration profile) with the established experimental technique of an equidistant choice of the beam energies. For the nonlinear design the Kullback-Leibler divergence was optimized. In the experimental design approach for the Kepler orbit measurements, the computational effort could be reduced exploiting the maximum entropy sampling. Instead, in the present case the data dependent uncertainty $\sigma_i = \max(5\% d_i, \sqrt{d_i})$ requires an additional parameter space integration to compute the information gain of a measurement using the Kullback-Leibler divergence [Eq. (179)], increasing computation time. However, the optimal next accelerator energy can be computed only after the result of the previous measurement is available. Therefore, long computing times are not compatible with an efficient operation. To circumvent this problem the posterior sampling method (Loredo, 1999) was used, reducing the computation of the next measurement energy to less than 5 minutes using a standard PC with 2 GHz CPU (von Toussaint *et al.*, 2008). In Fig. 51 three cycles of nonlinear Bayesian experimental design are shown: After the first measurement at 500 keV the posterior distribution of $\{a_1, a_2\}$ can be seen in the upper left graph by the posterior sample. The single measurement does not allow one to distinguish between a large decay constant a_1 and low constant offset a_2 or vice versa. The EU, plotted in the upper right graph, now favors a measurement at the other end of the energy range (the maximum of the utility function is encircled). After a measurement with 3 MeV ^{3}He the "area"

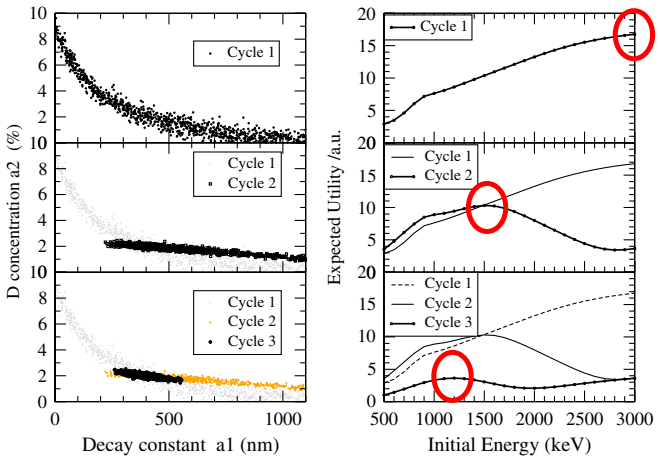

FIG. 51 (color online). Three cycles of the experimental design process. On the left-hand side 1000 samples drawn from the posterior distribution $p(a_1, a_2|\mathbf{d}, \eta)$ are displayed. On the right-hand side the expected utility is plotted and the maximum is indicated by a circle. The corresponding abscissa value is the suggested next measurement energy. Performing that measurement yields the posterior samples (black dots) given in the second row, left-hand side. The previous posterior samples are given in gray for comparison. On the right-hand side the new expected utility as a function of beam energy is given. Adding a new measurement with the proposed energy results in the posterior sample displayed in the lower left panel. The previous posterior samples are also given in different colors. The posterior volume has already decreased significantly. Therefore the best EU for the next measurement is lower than before. Adapted from von Toussaint *et al.*, 2008.

of the posterior distribution is significantly reduced (middle row, left graph): The background concentration is below 3% but the decay length is still quite undetermined. The EU has a maximum at 1500 keV, still with a pretty high EU. Performing a measurement with 1500 keV localizes the posterior distribution around the true, but unknown, value of $a_1 = 395$ nm and $a_2 = 0.02$. The next measurement should be performed at 1200 keV but the EU is significantly lower than before; subsequent measurements are predominantly improving the statistics. A second measurement at 3 MeV provides nearly the same information.

E. Autonomous exploration

Most present day advanced remote science operations use semiautomated systems that can carry out basic tasks such as locomotion and directed data collection, but require human intervention when it comes to deciding where to go or which experiment to perform. However, for many applications instruments that can both act and react with minimal human intervention would be advantageous. Knuth *et al.* (2007) described a simple robot that collects data in an automated fashion, and based on what it learns, decides which new measurement to take, thus, pursuing the learning cycle of observation, inference, and hypothesis refinement.

The experimental problem addressed is the localization (x, y position) and characterization (radius) of a white disk on a black plane using a robot arm equipped with a light sensor capable of *noisy point* measurements only. This toy problem can be considered as a crude representation of a land mine search problem (Goggans and Chi, 2007). The parameter vector of the disk consists of the disk center coordinates (x_0, y_0) and the disk radius r_0

$$\mathbf{C} = \{(x_0, y_0), r_0\} \tag{190}$$

and the data vector is given by a set of N light measurements

$$\mathbf{d} = \{d_1, d_2, \ldots, d_N\} \tag{191}$$

recorded at positions

$$\mathbf{X} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}. \tag{192}$$

When the positions are assumed to be known with certainty the posterior probability for the disk parameters is given by

$$p(\mathbf{C}|\mathbf{d}, \mathbf{X}, I) = p(\mathbf{d}|\mathbf{C}, \mathbf{X}, I)p(\mathbf{C}|I)/p(\mathbf{d}|I). \tag{193}$$

A uniform prior probability is assigned for the disk parameters

$$p(\mathbf{C}|I) = \left(\frac{1}{x_{\max} - x_{\min}}\right)\left(\frac{1}{y_{\max} - y_{\min}}\right)\left(\frac{1}{r_{\max} - r_{\min}}\right) \tag{194}$$

with $r_{\min} = 1$ cm and $r_{\max} = 15$ cm. The likelihood function for one measurement d_i taken at (x_i, y_i) can be written as

$$\begin{aligned} &p(d_i|\mathbf{C}, (x_i, y_i), I) \\ &\quad = p(d_i|\{(x_0, y_0), r_0\}, (x_i, y_i), I) \\ &\quad = \begin{cases} N(d_W, \sigma), & \text{if } (x_i - x_0)^2 + (y_i - y_0)^2 \le r_0^2, \\ N(d_B, \sigma), & \text{if } (x_i - x_0)^2 + (y_i - y_0)^2 > r_0^2. \end{cases} \end{aligned} \tag{195}$$

The expected value μ of a light measurement on the white disk is d_W and d_B is the expected value of a light measurement on the black background. The uncertainty of the intensity measurement is given by a Gaussian distribution $N(\mu, \sigma)$ with uncertainty σ, centered around the expected value μ. The information gain (Shannon entropy) of a measurement has been taken as a utility function. As the noise level is independent from the sampling location the maximum entropy sampling (Sebastiani and Wynn, 2000) can be used for an efficient computation of the expected utility based on posterior samples

$$\begin{aligned} (\hat{x}_e, \hat{y}_e) = \arg\max_{(x_e, y_e)} \Big(&- \int dd_e p(d_e|\mathbf{d}, (x_e, y_e), I) \\ &\times \log p(d_e|\mathbf{d}, (x_e, y_e), I) \Big). \end{aligned} \tag{196}$$

For an efficient computation of the posterior samples the nested sampling algorithm has been used. To find the next measurement position a grid on the space of possible measurement locations is considered and Eq. (196) is only computed at the grid points. The alignment of this grid is randomly jittered so that a greater variety of points can be considered during the measurement process. In Fig. 52(a) the initial stage of the inference process is shown: The first measurement has been taken [indicated by the black mark in the upper right part of Fig. 52(a)]. The white disk has not yet been located. For this reason there are large regions of the measurement space that are potentially equally informative, indicated by the homogeneous areas in Fig. 52(b), where the entropy gain of a further measurement at that location is

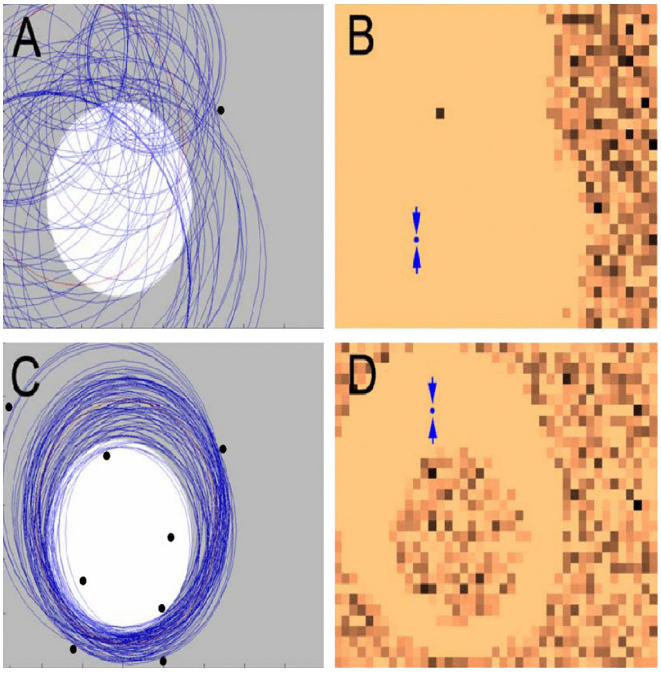

FIG. 52 (color online). (A), (C) The area to be searched together with the position of the white disk to be detected. In (A) a first measurement has been taken (small filled black circle in the upper right). Since at that location background only was detected, a number of possible positions of the white disk can immediately be excluded. This is visualized by 150 circles sampled from the updated posterior. These represent possible disk positions and sizes consistent with the measurement(s). (C) The situation after several measurements: The algorithm is getting close to a solution. The possible circle parameters are already narrowed down. The scatter of possible disk positions is already pretty small. (B) and (D) The selection algorithm for the location of the next measurement. The panels display the information gain for each measurement location (dark colors indicate uninformative locations) and the location with the highest information gain is selected for the next measurement (indicated by arrows). Further details are given in the text. From Knuth *et al.*, 2007.

displayed. Locations with a darker color provide less information (e.g., already measured locations). After several iterations, the robot will eventually find a white area belonging to the disk, thus immediately constraining the possible parameter space considerably. In Fig. 52(c) the set of circles in agreement with all measurements is now already constrained to the vicinity of the true disk location [Fig. 52(c)]. The measurement with the highest expected utility [indicated by two arrows in Fig. 52(d)] is in the region with the highest scattering of the posterior samples. It is essentially asking a binary question that rules out half of the models. This results in a rapid convergence significantly reducing the number of necessary measurements. These binary questions are not hard wired into the system but are a natural consequence of the selection of the most informative measurement (Knuth *et al.*, 2007).

F. Optimizing interferometric diagnostics

In a series of papers Dreier *et al.* (Fischer *et al.*, 2005; Dreier *et al.*, 2006a, 2006b, 2008a, 2008b) studied the design of a multichannel interferometer at the Wendelstein 7-X stellarator with respect to beam-line configuration, the number of beam lines, and joint evaluation with other diagnostics. Dreier *et al.* (2008b) investigated the impact of technical boundary conditions on the measurement of plasma electron density distributions using a four-channel two-color interferometer. For the interferometry system at W7-X three entrance ports into the vacuum vessel are reserved, allowing different beam-line configurations from vertical to horizontal optical paths (Kornejew *et al.*, 2006). Because no opposite ports are available, the probing beams have to be reflected by corner cube retroreflectors mounted on the opposite wall. These reflectors have to fit the structure of the in-vessel components. In combination with other constraints (e.g., limited port size) the number of realizable beam lines is 101. One of the physical questions to be addressed in the W7-X stellarator is the variation of the plasma density profiles at various confinement regimes (*H* mode and high density *H* mode). Here especially the maximum density within the plasma, the edge gradient, and the position of the edge are of interest (see Fig. 53). Maximizing the expected utility using the information gain of measurements (Kullback-Leibler divergence) as a utility function yielded an optimal design (Fig. 54, right panel) with an expected utility of EU $= 28.3 \pm 0.2$ bit. The best design taking into account the technical boundary conditions (Fig. 54, left panel) leads to an expected utility of only EU $= 8.53 \pm 0.01$ bit. A comparison of the two different designs reveals the reason for the large difference of the expected information gains. In the unconstrained design (Fig. 54, right panel) two lines of sight are localized at the very edge of the plasma, additionally passing the plasma on a very long path. This provides a good signal-to-noise ratio and at the same time a high sensitivity to small shifts in the position of the plasma edge. In both aspects the design where the port system had to be taken into account is inferior.

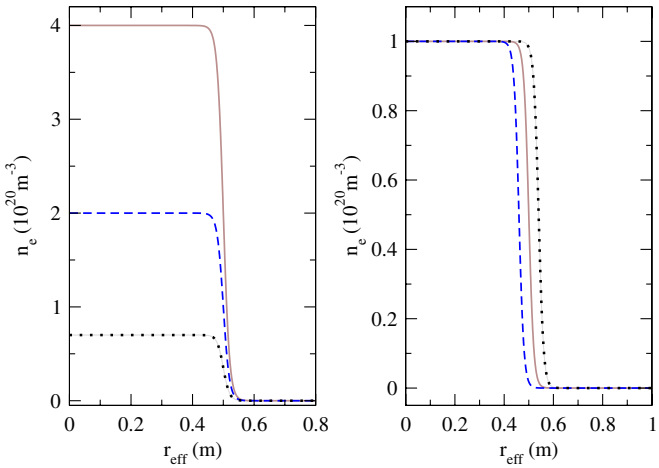

FIG. 53 (color online). Parameters of interest for the design of a multichannel interferometer for W7-X. The parameters are varied according to different high confinement regimes. The different lines (solid, dashed, and dotted) represent different realizations of physical scenarios On the left-hand side the maximum density of the plasma is varied, keeping the edge position constant. On the right-hand side the maximum density and the maximum slope are kept constant, varying the position of the plasma edge only. Adapted from Dreier *et al.*, 2008a.

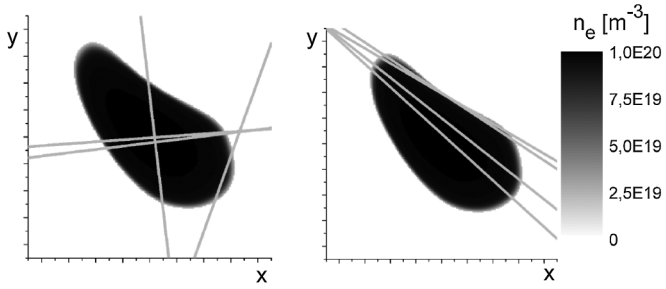

FIG. 54. Design result for four interferometry chords with respect to the measurement of high confinement regimes: Four beam lines in the interferometry plane (left) and optimal four beam configuration without technical constraints (right) superimposed onto a cross section of the W7-X plasma. Adapted from Dreier *et al.*, 2008b.

Therefore, the expected utility of modifications of the experimental setup has been investigated yielding a different design suggestion based on an out-of-plane setup (Dreier *et al.*, 2008b).

Other applications of (nonlinear) Bayesian experimental design are optimized material testing schemes (Ryan, 2003), filter design for Thomson scattering diagnostics (Fischer, 2004), optimized experiment proposals based on scaling laws (Preuss *et al.*, 2008), and the optimal design of heart defibrillators (Clyde *et al.*, 1993).

G. *N*-step ahead designs

All the design policies considered so far are greedy, selecting the best action using a one-step ahead approach, i.e., the best action is chosen as if the next measurement would be the last one. In practice, however, most experimental design optimizations are used in a repetitive manner. This may lead to less than optimal designs as can be demonstrated with a simple example. The interval [1, 4] has to be segmented with two support points such that the largest segment is minimized (e.g., for efficient regression). A one-step ahead algorithm selects $x_1 = 2.5$ as the best segmentation value for the first support point but in the subsequent optimization no placement of the second support point can reduce the size of the largest segment below 1.5. A two-step ahead algorithm would position the support points instead at $x_1 = 2$ and $x_2 = 3$ achieving an upper limit on the segment length of $s = 1$.

N-step ahead designs, also known as *full-sequential designs*, correspond to stochastic dynamic programming problems (SDP) (Bellman, 1957). The dependence of the later experiments on previous actions and observations (here for a two-step ahead design)

$$\mathrm{EU}(\eta_1) = \max_{\eta_1}\Big\{\int d\mathbf{D}_1\, p(\mathbf{D}_1|\eta_1, I) \times \max_{\eta_2}\Big[\int d\mathbf{D}_2\, p(\mathbf{D}_2|\mathbf{D}_1, \eta_1, \eta_2, I) \times U(\mathbf{D}_1, \mathbf{D}_2, \eta_1, \eta_2)\Big]\Big\} \tag{197}$$

introduces a feedback of information. This feedback, leading to the repeated embedded maximizations and integrations in Eq. (197), is the reason for the extreme difficulty of full-sequential designs with $N > 1$. Approximate solution methods of equations of similar structure are discussed in the areas of feedback control (Gautier and Pronzato, 1998; Pronzato, 2008) and partially observable Markov decision processes (Kaelbling *et al.*, 1996, 1998; Ng and Jordan, 2000). The computational complexity of the SDP problem has so far mostly precluded the use of full-sequential designs in experimental design [Kulcsar *et al.* (1994) provides one of the few attempts in this direction]. On the other hand, there is some evidence that in many cases the largest benefit is already provided by the step from $N = 1$ to $N = 2$ [see Pronzato (2008) and references therein (von Toussaint, 2011)]. The emerging computing power widens the range of models for which this limited increase in the prediction horizon is feasible.

H. Experimental design: Outlook

In the preceding sections the Bayesian approach to experimental design was illustrated with several examples, all of them focusing on the best strategy for parameter estimation $p(\boldsymbol{\theta}|\mathbf{D}, \mathbf{d}, M_1)$ for a given model M_1. In contrast the closely related approach of experimental design for model identification (Toman, 2008), i.e., the selection of measurements which best discriminates between a set of models M_k, $k = 1, \ldots, K$ has so far only rarely been applied for nonlinear models, most likely due to the increased numerical complexity to compute $p(M_k|\mathbf{D}, \mathbf{d})$. Some further aspects of experimental design which are the focus of current research are as follows:

- The optimization procedure assumes that the model is correct. This may (especially for linear models) lead to design suggestions which appear strange and are not robust with respect to minor deviations from the model. This is reflected in optimized designs which suggest measurements only at the end points of the design interval (DasGupta, 1996) or repeated measurements with the same settings (sometimes referred to as *thinly supported designs*). Averaging the expected utility over a set of plausible models (the mixture approach) may provide more robust designs (Chaloner and Verdinelli, 1995).
- The majority of experimental design techniques focuses either on the estimation of parameters of a given model or on model identification. For both cases appropriate utility functions are known (Pukelsheim, 1993). Relatively little work has been done to develop experimental design criteria to jointly improve parameter estimates and model identification. Some ideas are given by Borth (1975) and Chick and Ng (2002), but these ideas still wait to be tested in physics applications.
- The applicability of Bayesian experimental design depends on the feasibility of the necessary integrations. Several efficient algorithms have already been proposed (Müller and Parmigiani, 1995; Müller, 1999; Sebastiani and Wynn, 2000; Müller *et al.*, 2004), but the special structure of (sequential) expected utility computation still provides possibilities for further optimization. This is an active area of research (Brockwell and Kadane, 2003; Müller *et al.*, 2007).

- Once a joint environment-sensor model is created, the act of calibration becomes another potential experiment. In such a system, the instrument can decide to interact with either the environment via measurements or itself via calibration giving rise to an instrument that actively self-calibrates during an experiment (Thrun and Fox, 2005; Knuth *et al.*, 2007). The potential of these and similar ideas still waits to be explored.

Another noteworthy application of experimental design is the optimization of computer experiments. Elaborate computer models are progressively used in scientific research. As surrogates for physical systems, computer models can be subjected to experimentation, the goal being to predict how the real system would behave or to validate the computer model. Complex models often require long running times thus severely limiting the size and scope of computer experiments. A frequently used approach to circumvent these restrictions is based on fitting a cheaper predictor [e.g., a response surface model (Myers *et al.*, 1989)] of the simulation code output $y(t)$ to the input data t. The predictor is then used for parametric parameter studies instead of the original computer code. The experimental design is concerned with the best prediction of the simulation code output $y(t)$ using an optimized selection of sites $\{t_1, t_2, \ldots, t_n\}$ (Sacks *et al.*, 1989; Currin *et al.*, 1991; Kennedy and O'Hagan, 2001) and the efficient identification of the most relevant input parameters (Saltelli *et al.*, 2000). Bayesian approaches based on Gaussian processes (Neal, 1999a) may require a far smaller number of model runs than standard Monte Carlo approaches (Oakley and O'Hagan, 2004). A transfer of ideas from control theory for dynamic model systems (Hjalmarsson, 2005) and correlated, multidimensional response variables may provide further progress.

VIII. CONCLUSION AND OUTLOOK

It was demonstrated that Bayesian probability theory is a powerful tool for inference from physical data and uncertain information. It allows the extraction of the most convincing conclusions implied by given data and any prior knowledge in a systematic way.

This was first noticed in observational branches as in biometrics and astronomy where the data sets cannot be augmented at will and have to be exploited as far as possible. Fortunately in other branches of physics the situation expressed by Mackenzie (2004): ''We use fantastic telescopes, the best physical models, and the best computers. The weak link in this chain is interpreting our data using 100-year-old-mathematics'' is steadily improving. Thanks to the ongoing increase of computing power, it now becomes possible to handle problems using modern MCMC techniques which were infeasible 10 years ago. This naturally broadens the range of possible applications of Bayesian inference. The areas most likely to benefit most are those where statistical model building has so far been hampered by a lack of knowledge or insight into the system, e.g., at the interface of biophysics and biology. Here the increasing amount of available molecular (low level) data calls for methods and tools for comprehensive, unsupervised model selection and model design (Huelsenbeck *et al.*, 2001) in large model spaces.

Within the area of physics, the increasing complexity of simulation codes (e.g., climate simulation codes, plasma simulations) suggests to compute and analyze the obtained data in an optimized way, using Bayesian design and prediction methods. So far, most approaches to this problem have treated the simulation codes as black boxes. The challenge is to find ways to take advantage of the available insight into the complex system in the best possible way. The emerging computational resources also provide, for the first time, the possibility to implement nontrivial prediction-testing-inference cycles. Although still limited to greedy algorithms, *automated exploration algorithms* are now becoming feasible. The exploitation of the full potential of these algorithms is only just now beginning.

Given the steady development of Bayesian foundations, the progress in variational and sampling algorithms and, most importantly, the tremendous increase of complex information gathered from physical experiments, Bayesian inference can expect a bright future.

ACKNOWLEDGMENTS

I am grateful to the anonymous referees for their valuable comments and suggestions that have helped to improve the manuscript substantially. I wish to thank Volker Dose for his continuous support and the introduction to Bayesian inference. I am grateful to many colleagues for inspiring discussions, ideas, and suggestions, in particular, to Wolfgang Jacob, Rainer Fischer, Silvio Gori, Roland Preuss, Jochen Roth, Christian Hopf, Thomas Schwarz-Selinger, Wolfgang von der Linden, Martin Kern, Tom Loredo, and Katharina Diller.

REFERENCES

Abramowitz, M., and I. Stegun, 1965, *NBS Handbook of Mathematical Functions* (U.S. GPO, Washington, DC).

Adams, R. P., and D. J. MacKay, 2007, Bayesian Online Changepoint Detection, Technical Report arXiv:0710.3742v1 [stat.ML], University of Cambridge.

Agarwal, D. K., and A. E. Gelfand, 2005, Stat. Comput. **15**, 61.

Ahlers, H., and A. Engel, 2008, Eur. Phys. J. B **62**, 357.

Aji, S. M., and R. J. McEliece, 2000, IEEE Trans. Inf. Theory **46**, 325.

Akaike, H., 1974, IEEE Trans. Autom. Control **19**, 716.

Albert, J., 2009, *Bayesian Computation with R* (Springer, New York).

Alimov, V. K., M. Mayer, and J. Roth, 2005, Nucl. Instrum. Methods Phys. Res., Sect. B **234**, 169.

Amsel, G., and W. A. Lanford, 1984, Annu. Rev. Nucl. Part. Sci. **34**, 435.

Andrieu, C., and A. Doucet, 1999, IEEE Trans. Signal Process. **47**, 2667.

Andrieu, C., and J. Thoms, 2008, Stat. Comput. **18**, 343.

Arshad, S., J. Cordey, D. McDonald, A. Dinklage, J. Farthing, R. Fischer, E. Joffrin, M. von Hellermann, C. Roach, and JET EFDA contributors, 2007, EFDA-JET-Report EFDA-JET-PR(06)26.

Atkinson, A. C., A. Donev, and R. Tobias, 2007, *Optimum Experimental Designs, with SAS* (Oxford University Press, Oxford, UK).

Bagchi, P., and J. Kadane, 1991, Can. J. Stat. **19**, 67.

Bailey, T., and H. Durrant-Whyte, 2006, Robotics and Automation Magazine **13**, 108.

Balden, M., and J. Roth, 2000, J. Nucl. Mater. **280**, 39.
Baraffe, I., G. Chabrier, and T. Barman, 2010, Rep. Prog. Phys. **73**, 016901.
Bartelmann, M., 2010, Rev. Mod. Phys. **82**, 331.
Bassett, B. A., S. Tsujikawa, and D. Wands, 2006, Rev. Mod. Phys. **78**, 537.
Basseville, M., and I. Nikiforov, 1993, *Detection of Abrupt Changes: Theory and Application* (Prentice-Hall, Englewood Cliffs, NJ).
Belisle, C., 1998, Can. J. Stat. **26**, 629.
Bellman, R. E., 1957, *Dynamic Programming* (Princeton University Press, Princeton, NJ).
Bennett, C. L., *et al.*, 2003, Astr. Phys. J. **583**, 1.
Berger, J. O., 1985, *Statistical Decision Theory and Bayesian Analysis* (Springer, New York).
Berger, J. O., and J. M. Bernardo, 1992, in *Bayesian Statistics*, edited by J. Bernardo, J. Berger, A. Dawid, and A. Smith (Oxford University Press, Oxford), Vol. 4, p. 35.
Berger, J. O., and D. A. Berry, 1988, Am. Sci. **76**, 159 [http://www.jstor.org/stable/27855070].
Berger, J. O., B. Boukai, and Y. Wang, 1997, Stat. Sci. **12**, 133.
Berger, J. O., L. D. Brown, and R. L. Wolpert, 1994, Ann. Stat. **22**, 1787.
Berger, J. O., J. K. Gosh, and N. Mukhopadhyay, 2003, J. Stat. Plann. Infer. **112**, 241.
Berger, J. O., and L. R. Pericchi, 1988, in *Model Selection*, edited by P. Lahiri, IMS Lecture Notes—Monograph Series Vol. 38 (IMS, Beachwood, OH), p. 135.
Berger, J. O., and T. Sellke, 1987, J. Am. Stat. Assoc. **82**, 112.
Berger, J. O., and R. L. Wolpert, 1988, *The Likelihood Principle* (IMS, Hayward, CA), 2nd ed.
Bernardo, J. M., 1979a, Ann. Stat. **7**, 686.
Bernardo, J. M., 1979b, J. R. Stat. Soc. Ser. B **41**, 113 [http://www.jstor.org/stable/2985028].
Bernardo, J. M., 2005, *Reference Analysis* (Elsevier, Amsterdam).
Bernardo, J. M., and A. F. M. Smith, 2000, *Bayesian Theory* (Wiley, Chichester, UK).
Bickel, P., 1981, Ann. Stat. **9**, 1301.
Birnbaum, A., 1962, J. Am. Stat. Assoc. **57**, 269.
Bishop, C., 2006, *Pattern Recognition and Machine Learning* (Springer, New York).
Bishop, C., D. Spiegelhalter, and J. Winn, 2003, in *Advances in Neural Information Processing Systems*, edited by S. Becker, S. Thrun, and K. Obermeyer (MIT Press, Cambridge, MA), Vol. 15, p. 793.
Bittner, E., A. Nussbaumer, and W. Janke, 2008, Phys. Rev. Lett. **101**, 130603.
Borenstein, J., B. Everett, and L. Feng, 1996, *Navigating Mobile Robots: Systems and Techniques* (A.K. Peters, Ltd, Wellesley, MA).
Borth, D. M., 1975, J. R. Stat. Soc. Ser. B **37**, 77 [http://www.jstor.org/stable/2984993].
Box, G. E. P., and G. C. Tiao, 1992, *Bayesian Inference in Statistical Analysis* (John Wiley, New York) (reprint from 1973).
Bradley, N. L., A. C. Leopold, J. Ross, and W. Huffaker, 1999, Proc. Natl. Acad. Sci. U.S.A. **96**, 9701.
Bretthorst, G. L., 1988, *Bayesian Spectrum Analysis and Parameter Estimation* (Springer, New York).
Bretthorst, G. L., 1990a, J. Magn. Reson. **88**, 533.
Bretthorst, G. L., 1990b, J. Magn. Reson. **88**, 552.
Bretthorst, G. L., 1990c, J. Magn. Reson. **88**, 571.
Bretthorst, G. L., 1991, J. Magn. Reson. **93**, 369.
Bretthorst, G. L., 1996, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by G. R. Heidbreder (Kluwer Academic Publishers, Dordrecht), p. 1.
Bretthorst, G. L., 2001, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by A. Mohammad-Djafari, AIP Conf. Proc. 568 (AIP, New York), p. 241.
Bridges, M., F. Feroz, M. P. Hobson, and A. N. Lasenby, 2009, Mon. Not. R. Astron. Soc. **400**, 1075.
Brockwell, A. E., and J. B. Kadane, 2003, J. Comput. Graph. Stat. **12**, 566.
Brooks, A., and T. Bailey, 2008, Workshop on the Algorithmic Fundamentals of Robotics 30.11.2008.
Brooks, J., D. Alman, G. Federici, D. Ruzic, and D. White, 1999, J. Nucl. Mater. **266–269**, 58.
Brooks, S., P. Giudici, and G. Roberts, 2003, J. R. Stat. Soc. Ser. B **65**, 3.
Brooks, S., and G. Roberts, 1999, Biometrika **86**, 710.
Cappé, O., R. Douc, A. Guillin, J. M. Marin, and C. P. Robert, 2008, Stat. Comput. **18**, 447.
Cappé, O., S. Godsill, and E. Moulines, 2007, Proc. IEEE **95**, 899.
Cappé, O., A. Guillin, J. Marin, and C. Robert, 2004, J. Comput. Graph. Stat. **13**, 907.
Carlin, B. P., A. E. Gelfand, and A. F. M. Smith, 1992, J. Roy. Stat. Soc. C **41**, 389, http://www.jstor.org/stable/2347570.
Carlin, B. P., and T. A. Louis, 1998, *Bayes and Empirical bayes Methods for Data Analysis* (Chapman & Hall/CRC, Boca Raton).
Caticha, A., 2008, Lectures on probability, entropy and statistical physics, http://arxiv.org/pdf/0808.0012.
Celeux, G., M. Hurn, and C. Robert, 2000, J. Am. Stat. Assoc. **95**, 957.
Chaloner, K., and I. Verdinelli, 1995, Stat. Sci. **10**, 273.
Chambers, J., 2010, Nature (London) **467**, 405.
Charvin, K., K. Gallagher, G. Hampson, and R. Labourdette, 2009, Basin Research **21**, 5.
Chen, J., and A. K. Gupta, 2000, *Parametric Statistical Change Point Analysis* (Birkhäuser Verlag, Boston).
Chen, M. H., Q. M. Shao, and J. G. Ibrahim, 2001, *Monte Carlo Methods for Bayesian Computation* (Springer, New York).
Cheng, J., and M. J. Druzdzel, 2000, J. Artif. Intell. Res. **13**, 155.
Chernoff, H., 1972, *Sequential Analysis and Optimal Design* (SIAM, Philadelphia).
Chick, S. E., and S. H. Ng, 2002, in *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes (IEEE, San Diego, CA), p. 400.
Chopin, N., and C. Robert, 2007, in *Bayesian Statistics*, edited by J. Bernardo, M. J. Bayarri, J. Berger, A. Dawid, D. Heckermann, A. Smith, and M. West (Oxford University Press, Oxford), Vol. 8, p. 491.
Chopin, N., and C. Robert, 2008, arXiv:0801.3887.
Chopin, N., and C. Robert, 2010, Biometrika **97**, 741.
Clyde, M., and E. I. George, 2004, Stat. Sci. **19**, 81.
Clyde, M., P. Müller, and G. Parmigiani, 1993, in *Case Studies in Bayesian Statistics, II* (Springer-Verlag, Berlin), p. 278.
Clyde, M. A., J. O. Berger, F. Bullard, E. B. Ford, W. H. Jefferys, R. Luo, R. Paulo, and T. Loredo, 2007, in *Statistical Challenges in Modern Astronomy IV*, edited by G. J. Babu and E. D. Feigelson (ASP), Vol. 371, p. 224.
Clyde, M. A., and R. L. Wolpert, 2007, in *Bayesian Statistics*, edited by J. Bernardo, M. J. Bayarri, J. Berger, A. Dawid, D. Heckermann, A. Smith, and M. West (Oxford University Press, Oxford), Vol. 8, p. 1.
Coluzza, I., and D. Frenkel, 2005, Chem. Phys. Chem. **6**, 1779.
Cooper, G., 1990, Artif. Intell. **42**, 393.
Cornebise, J., E. Moulines, and J. Olsson, 2008, Stat. Comput. **18**, 461.
Cornfield, J., 1969, Biometrics **25**, 617.

Cornu, A., and R. Massot, 1979, *Compilation of Mass Spectral Data* (Heyden, London).
Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, 1999, *Probabilistic Networks and Expert Systems* (Springer, New York).
Cowles, M., and B. Carlin, 1996, J. Am. Stat. Assoc. **91**, 883.
Cox, D. R., 2006, *Principles of Statistical Inference* (Cambridge University Press, Cambridge).
Cox, R. T., 1946, Am. J. Phys. **14**, 1.
Cox, R. T., 1961, *The Algebra of Probable Inference* (John Hopkins Press, Baltimore, MD).
Currin, C., T. Mitchell, M. Moris, and D. Ylvisaker, 1991, J. Am. Stat. Assoc. **86**, 953.
Daghofer, M., and W. von der Linden, 2004, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by R. Fischer, R. Preuss, and U. von Toussaint, AIP Conf. Proc. No. 735 (AIP, Melville, NY), p. 355.
D'Agostini, G., 1999, in *Bayesian Reasoning in High-Energy Physics: Principles and Applications*, CERN 99-03 (CERN, Geneva).
D'Agostini, G., 2003, Rep. Prog. Phys. **66**, 1383.
Dagum, P., and M. Luby, 1993, Artif. Intell. **60**, 141.
Darwiche, A., 2009, *Modeling and Reasoning with Bayesian Networks* (Cambridge University Press, Cambridge, UK).
DasGupta, A., 1996, in *Handbook of Statistics 13: Design and Analysis of Experiments*, edited by S. Gosh and C. R. Rao (Elsevier, Amsterdam).
Davidoff, F., 1999, Ann. Intern. Med. **130**, 1019.
Davis, P., and P. Rabinowitz, 1984, *Methods of Numerical Integration* (Academic Press, Orlando, FL).
Dawid, A., 1979, J. R. Stat. Soc. Ser. B **41**, 1 [http://www.jstor.org/stable/2984718].
DeGroot, M. H., 1962, Ann. Math. Stat. **33**, 404.
DeGroot, M. H., 1986, in *Recent Developments in the Foundations of Utility and Risk Theory*, edited by L. Daboni, A. Montesano, and M. Lines (Reidel, Dordrecht), p. 265.
de Laplace, P. S., 1812, *Theorie Analytique des Probabilites* (Courcier Imprimeur, Paris).
Denison, D. G. T., B. K. Mallick, and A. F. M. Smith, 1998, J. R. Stat. Soc. Ser. B **60**, 333.
Devroye, L., 1986, *Non-uniform Random Variate Generation* (Springer, New York).
DiCiccio, T., R. Kass, A. Raftery, and L. Wassermann, 1997, J. Am. Stat. Assoc. **92**, 903.
Dickinson, C., H. K. Eriksen, A. J. Banday, J. B. Jewell, K. M. Gorski, G. Huey, C. R. Lawrence, I. J. O'Dwyer, and B. D. Wandelt, 2009, Astrophys. J. **705**, 1607.
Dieboldt, J., and C. Robert, 1994, J. R. Stat. Soc. Ser. B **56**, 363 [http://www.jstor.org/stable/2345907].
DiMatteo, I., C. R. Genovese, and R. E. Kass, 2001, Biometrika **88**, 1055.
Dinklage, A., R. Fischer, J. Geiger, G. Kühner, H. Maassberg, J. Svensson, and U. von Toussaint, 2003, in *30th EPS Conference on Controlled Fusion and Plasma Physics*, edited by R. Koch and S. Lebedev (Europ. Phys. Soc., Geneva), Vol. ECA 27A, p. P-4.80.
Dinklage, A., R. Fischer, and J. Svensson, 2003, in *Proceedings of PLASMA 2003 'Research and Applications of Plasmas*, p. I-1.1.
Dinklage, A., R. Fischer, and J. Svensson, 2004, Fusion Sci. Technol. **46**, 355.
Dissanayake, G., P. Newman, S. Clark, H. Durrant-Whyte, and M. Csorba, 2001, IEEE Transactions on Robotics and Automation **17**, 229.
Dobrozemsky, R., and G. Schwarzinger, 1992, J. Vac. Sci. Technol. A **10**, 2661.
Dose, V., 2002, Bayes in five days.
Dose, V., 2003a, Rep. Prog. Phys. **66**, 1421 [stacks.iop.org/RoPP/66/1421].
Dose, V., 2003b, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by C. J. Williams, AIP Conf. Proc. No. 659 (AIP, Melville, NY), p. 350.
Dose, V., R. Fischer, and W. von der Linden, 1998, in *Maximum Entropy and Bayesian Methods*, edited by G. Erickson (Kluwer Academic, Dordrecht), p. 147.
Dose, V., and A. Menzel, 2004, Global Change Biology **10**, 259.
Dose, V., and A. Menzel, 2006, Global Change Biology **12**, 1451.
Dose, V., R. Preuss, and J. Roth, 2001, J. Nucl. Mater. **288**, 153.
Dose, V., and W. von der Linden, 1999, in *Maximum Entropy and Bayesian Methods*, edited by W. von der Linden, V. Dose, R. Fischer, and R. Preuss (Kluwer Academic Publishers, Dordrecht).
Doucet, A., N. de Freitas, and N. Gordon, 2001, Eds., *Sequential Monte Carlo in Practice* (Springer, New York).
Doucet, A., and A. Johansen, 2008, in *Oxford Handbook of Nonlinear Filtering*, edited by D. Crisan and B. Rozovsky (Oxford University Press) [www.cs.ubc.ca/~arnaud/doucet_johansen_tutorialPF.pdf].
Draper, D., 1995, J. R. Stat. Soc. Ser. B **57**, 45 [http://www.jstor.org/stable/2346087].
Dreier, H., A. Dinklage, R. Fischer, M. Hirsch, and P. Kornejew, 2006a, Rev. Sci. Instrum. **77**, 10F323.
Dreier, H., A. Dinklage, R. Fischer, M. Hirsch, and P. Kornejew, 2008a, Rev. Sci. Instrum. **79**, 10E712.
Dreier, H., A. Dinklage, R. Fischer, M. Hirsch, and P. Kornejew, 2008b, in *PLASMA 2007*, edited by H. J. Hartfuss, M. Dudeck, J. Musielok, and M. J. Sadowski, AIP Conf. Proc. No. 993 (AIP, Melville, NY), p. 183.
Dreier, H., A. Dinklage, R. Fischer, M. Hirsch, P. Kornejew, and E. Pasch, 2006b, Fusion Sci. Technol. **50**, 262.
Duane, S., A. Kennedy, B. Pendleton, and D. Roweth, 1987, Phys. Lett. B **195**, 216.
Dunkley, J., *et al.*, 2009, Astrophys. J. **701**, 1804.
Durrant-Whyte, H., and T. Bailey, 2006, IEEE Robotics & Automation Magazine **13**, 99.
Durrer, R., 2008, *The Cosmic Microwave Background* (Cambridge University Press, Cambridge, UK).
Earl, D., and M. Deem, 2005, Phys. Chem. Chem. Phys. **7**, 3910.
Evans, M., 2007, in *Bayesian Statistics*, edited by J. Bernardo, M. J. Bayarri, J. Berger, A. Dawid, D. Heckermann, A. Smith, and M. West (Oxford University Press, Oxford), Vol. 8, p. 491.
Fearnhead, P., 2006, Stat. Comput. **16**, 203.
Fedorov, V. V., 1972, *Theory of Optimal Experiments* (Academic, New York).
Feller, W., 1991, *An Introduction to Probability Theory and Its Applications* (Wiley, Chichester), Vol. 2.
Feroz, F., M. P. Hobson, and M. Bridges, 2009, Mon. Not. R. Astron. Soc. **398**, 1601.
Fiore, C. E., and M. G. E. da Luz, 2010, Phys. Rev. E **82**, 031104.
Fischer, R., 2004, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by R. Fischer, R. Preuss, and U. von Toussaint, AIP Conf. Proc. No. 735 (AIP, Melville, NY), p. 76.
Fischer, R., and A. Dinklage, 2004, Rev. Sci. Instrum. **75**, 4237.
Fischer, R., A. Dinklage, and E. Pasch, 2003, Plasma Phys. Controlled Fusion **45**, 1095.
Fischer, R., H. Dreier, A. Dinklage, B. Kurzan, and E. Pasch, 2005, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by K. Knuth, A. Abbas, R. Morris, and J. Castle, AIP Conf. Proc. No. 803 (AIP, Melville, NY), p. 440.

Fischer, R., K. M. Hanson, V. Dose, and W. von der Linden, 2000, Phys. Rev. E **61**, 1152.
Fischer, R., M. Mayer, W. von der Linden, and V. Dose, 1997, Phys. Rev. E **55**, 6667.
Fischer, R., M. Mayer, W. von der Linden, and V. Dose, 1998, Nucl. Instrum. Methods Phys. Res., Sect. B **136–138**, 1140.
Fischer, R., C. Wendland, A. Dinklage, S. Gori, V. Dose, and The W7-AS team, 2002, Plasma Phys. Controlled Fusion **44**, 1501.
Fischer, R., E. Wolfrum, J. Schweinzer, and The ASDEX Upgrade Team, 2008, Plasma Phys. Controlled Fusion **50**, 085009.
Flynn, A. M., 1988, Int. J. Robotics Research **7**, 5.
Fox, D., W. Burgard, and S. Thrun, 1999, J. Artif. Intell. Res. **11**, 391.
Frenkel, D., 1986, in *Molecular Dynamics Simulation of Statistical-Mechanical Systems*, edited by G. Ciccotti and W. Hoover (North-Holland, Amsterdam), p. 151.
Frese, U., 2006, Autonomous Robots **20**, 25.
Friedman, N., and D. Koller, 2003, Mach. Learn. **50**, 95.
Fröhner, F., 2000, Evaluation and Analysis of Nuclear Resonance Data, JEFF Report 18 (OECD Nuclear Energy Agency, Paris).
Gallagher, K., K. Charvin, S. Nielsen, M. Sambridge, and J. Stephenson, 2009, Marine and Petroleum Geology **26**, 525.
Gamerman, D., 1997, *Markov Chain Monte Carlo* (Chapman & Hall, London).
Gamerman, D., and H. F. Lopes, 2006, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* (Chapman & Hall, London).
Garnett, R., M. A. Osborne, and S. J. Roberts, 2009, in *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning* (ACM, New York), p. 345, ISBN 978-1-60558-516-1.
Gautier, R., and L. Pronzato, 1998, in *New Developments and Applications in Experimental Design* (Institute of Mathematical Statistics, Hayward, CA), Vol. 34, pp. 138–151 [http://www.jstor.org/stable/4356069].
Gelfand, A. E., S. E. Hills, A. Racine-Poon, and A. F. M. Smith, 1990, J. Am. Stat. Assoc. **85**, 972.
Gelfand, A. E., and S. K. Sahu, 1994, J. Comp. Graph. Statist. **3**, 261 [http://www.jstor.org/stable/1390911].
Gelfand, A. E., and A. F. M. Smith, 1990, J. Am. Stat. Assoc. **85**, 398.
Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004, *Bayesian Data Analysis* (Chapman and Hall/CRC, Boca Raton).
Gelman, A., and X. Meng, 1998, Stat. Sci. **13**, 163.
Gelman, A., and D. Rubin, 1992, Stat. Sci. **7**, 457.
Geman, S., and D. Geman, 1984, IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**, 721.
George, E. I., 2000, J. Am. Stat. Assoc. **95**, 1304.
Geyer, C., 1991, in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (American Statistical Association, New York), p. 156.
Geyer, C., 1992, Stat. Sci. **7**, 473.
Geyer, C., and E. Thompson, 1995, J. Am. Stat. Assoc. **90**, 909.
Gilks, W. R., 1992, in *Bayesian Statistics*, edited by J. Bernardo, J. Berger, A. Dawid, and A. Smith (Oxford University Press, Oxford), Vol. 4, p. 641.
Gilks, W. R., N. Best, and K. Tan, 1995, Appl. Statist. **44**, 455 [http://www.jstor.org/stable/2986138].
Gilks, W. R., S. Richardson, and D. Spiegelhalter, 1996, *Markov Chain Monte Carlo in Practice* (Chapman & Hall, London).
Gilks, W. R., G. Roberts, and S. Sahu, 1998, J. Am. Stat. Assoc. **93**, 1045.
Gilks, W. R., and P. Wild, 1992, Appl. Statist. **41**, 337 [http://www.jstor.org/stable/2347565].
Gilks, W. R., A. Thomas, and D. J. Spiegelhalter, 1994, The Statistician **43**, 169 [http://www.jstor.org/stable/2348941].
Girolami, M., and B. Calderhead, 2011, J. R. Stat. Soc. Ser. B **73**, 1.
Goggans, P. M., and Y. Chi, 2007, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by A. M. Djafari, AIP Conf. Proc. No. 872 (AIP, Melville, NY), p. 533.
Goldstein, M., 1992, in *Bayesian Statistics*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford University Press, Oxford), Vol. 4, p. 477.
Golfarelli, M., D. Maio, and S. Rizzi, 1998, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, New York), p. 905.
Goodman, S. N., 1999a, Ann. Intern. Med. **130**, 995 [http://www.annals.org/content/130/12/995.abstract].
Goodman, S. N., 1999b, Ann. Intern. Med. **130**, 1005 [http://www.annals.org/content/130/12/1005.abstract].
Gordon, N., D. Salmond, and A. Smith, 1993, *Radar and Signal Processing, IEE Proceedings F* Vol. 140, p. 107 [ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=210672].
Green, P. J., 1995, Biometrika **82**, 711.
Green, P. J., 2003, in *Highly Structured Stochastic Systems*, edited by P. J. Green, N. L. Hjort, and S. Richardson (Oxford University Press, Oxford), p. 179.
Gregory, P. C., 1999, Astrophys. J. **520**, 361.
Gregory, P. C., 2005, *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge University Press, Cambridge).
Gregory, P. C., and T. Loredo, 1992, Astrophys. J. **398**, 146.
Gregory, P. C., 2002, Astrophys. J. **575**, 427.
Gregory, P. C., 2005, Astrophys. J. **631**, 1198.
Gregory, P. C., and D. A. Fischer, 2010, Mon. Not. R. Astron. Soc. **403**, 731.
Gregory, P. C., and T. J. Loredo, 1993, in *Maximum Entropy Method and Bayesian Methods*, edited by A. Mohammad-Djafari, and G. Demoment (Kluwer Academic Press, Dordrecht), p. 225.
Gregory, P. C., and T. J. Loredo, 1996, Astrophys. J. **473**, 1059.
Grote, H., *et al.*, 1999, J. Nucl. Mater. **266–269**, 1059.
GSL, 2008, GNU Scientific Library [http://www.gnu.org/software/gsl/].
Guglielmetti, F., R. Fischer, and V. Dose, 2009, Mon. Not. R. Astron. Soc. **396**, 165.
Guglielmetti, F., R. Fischer, V. Dose, W. Voges, and G. Boese, 2004, in *ASP Conference Series Volumes, Astronomical Data Analysis Software and Systems (ADASS) XIII*, edited by M. G. A. Francois Ochsenbein and D. Egret (Astronomical Society of the Pacific, San Francisco), Vol. CS 314, p. O03.3.
Haario, H., M. Laine, A. Mira, and E. Saksman, 2006, Stat. Comput. **16**, 339.
Habeck, M., M. Nilges, and W. Rieping, 2005, Phys. Rev. Lett. **94**, 018105.
Habeck, M., W. Rieping, and M. Nilges, 2004, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by G. Erickson and Y. Zhai, AIP Conf. Proc. No. 707 (AIP, Melville, NY), p. 157.
Hagan, A. O., and H. Le, 1994, in *Aspects of Uncertainty: A Tribute to D. V. Lindley*, edited by P. R. Freeman, and A. F. M. Smith (Wiley, Chichester), p. 311.
Hall, D. L., 2004, *Mathematical Techniques in Multisensor Data Fusion* (Artech House, Norwood).
Han, C., and B. P. Carlin, 2001, J. Am. Stat. Assoc. **96**, 1122.
Hansmann, U. H. E., Y. Okamoto, and F. Eisenmenger, 1996, Chem. Phys. Lett. **259**, 321.
Harney, H. L., 2003, *Bayesian Inference: Parameter Estimation and Decisions* (Springer, Berlin).

Hastings, W., 1970, Biometrika **57**, 97.
Higdon, D., 1998, J. Am. Stat. Assoc. **93**, 585.
Hinshaw, G., *et al.*, 2009, Astrophys. J. Suppl. Ser. **180**, 225.
Hitchcock, D. B., 2003, The American Statistician **57**, 254.
Hjalmarsson, H., 2005, Automatica **41**, 393.
Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999, Stat. Sci. **14**, 382.
Hole, M., G. von Nessi, J. Bertram, J. Svensson, L. C. Appel, B. D. Blackwell, R. L. Dewar, and J. Howard, 2009, J. Plasma Fusion Res. Series **9**, 479.
Horowitz, A., 1991, Phys. Lett. B **268**, 247.
Hsu, D. A., 1982, J. Am. Stat. Assoc. **77**, 29.
Huang, C., and A. Darwiche, 1996, Intl. J. Approx. Reasoning **15**, 225.
Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback, 2001, Science **294**, 2310.
Hukushima, K., H. Takayama, and K. Nemoto, 1996, Int. J. Mod. Phys. C **7**, 337.
Huzurbazar, V. S., 1976, *Sufficient Statistics* (Marcel Dekker, New York).
Jaakkola, T., 2001, in *Advances in Mean Field Methods*, edited by M. Opper and D. Saad (MIT Press, Cambridge, MA), p. 129.
Jaakkola, T., and M. Jordan, 2000, Stat. Comput. **10**, 25.
Jarosik, N., *et al.*, 2011, Astrophys. J. Suppl. Ser. **192**, 14.
Jarzynski, C., 1997, Phys. Rev. Lett. **78**, 2690.
Jarzynski, C., 2006, Phys. Rev. E **73**, 046105.
Jasra, A., C. Holmes, and D. Stephens, 2005, Stat. Sci. **20**, 50.
Jasra, A., D. Stephens, K. Gallagher, and C. Holmes, 2006, Math. Geol. **38**, 269.
Jaynes, E. T., 1957a, Phys. Rev. **106**, 620.
Jaynes, E. T., 1957b, Phys. Rev. **108**, 171.
Jaynes, E. T., 1968, IEEE Trans. Syst. Sci. Cybern. **4**, 227.
Jaynes, E. T., 1983, in *Papers on Probability, Statistics and Statistical Physics*, edited by W. L. Harper and C. A. Hooker (Reidel, Dordrecht).
Jaynes, E. T., 1976, in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, edited by W. L. harper and C. A. Hooker (Reidel, Dordrecht), p. 252.
Jaynes, E. T., 1979, J. Am. Stat. Assoc. **74**, pp. 740.
Jaynes, E. T., and L. Bretthorst, 2003, *Probability Theory, The Logic of Science* (Oxford University Press, Oxford).
Jeffreys, H., 1939, *Theory of Probability* (Oxford University Press, Oxford), 3rd ed., revised edition 1961.
Jeffreys, H., 1961, *Theory of Probability* (Oxford University Press, Oxford).
Jensen, C., A. Kong, and U. Kjaerulff, 1995, Int. J. Human-Comput. Stud., Special Issue on Real-World Applications of Uncertain Reasoning, **42**, 647.
Jensen, F. V., 2001, *Bayesian Networks and Decision Graphs* (Springer, New York).
Ji, X., *et al.*, 2009, in *RoboCup*, edited by L. Iocchi , Lecture Notes in Computer Science Vol. 5399 (Springer, Berlin), p. 507.
Jordan, M. I., 1999, *Learning in Graphical Models* (MIT Press, Cambridge, MA).
Jordan, M. I., 2004, Stat. Sci. **19**, 140.
Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, 1999, in *Learning in Graphical Models*, edited by M. I. Jordan (MIT Press, Cambridge, MA), p. 105.
Kaelbling, L. P., M. L. Littman, and A. R. Cassandra, 1998, Artif. Intell. **101**, 99.
Kaelbling, L. P., M. L. Littman, and A. W. Moore, 1996, J. Artif. Intell. Res. **4**, 237.
Kang, H., and V. Dose, 2003, J. Vac. Sci. Technol. A **21**, 1978.
Kapur, J., and H. Kesavan, 1992, *Entropy Optimization Principles with Applications* (Academic, Boston).
Kass, R. E., *et al.*, 1988, in *Bayesian Statistics*, edited by J. Bernardo (Oxford University Press, New York), Vol. 3, p. 261.
Kass, R. E., and A. E. Raftery, 1995, J. Am. Stat. Assoc. **90**, 773.
Katz, M., 1961, Ann. Math. Stat. **32**, 136.
Katzgraber, H., S. Trebst, D. Huse, and M. Troyer, 2006, J. Stat. Mech., P03018.
Kendall, M., and P. Moran, 1963, *Geometrical Probability* (Griffin, London).
Kennedy, A., *et al.*, 1990, in *Probabilistic Methods in Quantum Field Theory and Quantum Gravity*, edited by P. D. Damgaard, H. Hüffel, and A. Rosenblum (Springer, New York).
Kennedy, M. C., and A. O'Hagan, 2001, J. R. Stat. Soc. Ser. B **63**, 425.
Kerler, W., and P. Rehberg, 1994, Phys. Rev. E **50**, 4220.
Kinney, W. H., 2001, Phys. Rev. D **63**, 043001.
Knuth, K. H., 2004, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by G. Erickson and Y. Zhai, AIP Conf. Proc. No. 707 (AIP, Melville, NY), p. 204.
Knuth, K. H., P. M. Erner, and S. Frasso, 2007, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by K. Knuth, A. Caticha, J. Center, A. Giffin, and C. Rodrigues, AIP Conf. Proc. No. 954 (AIP, Melville, NY), p. 203.
Koch, D. G., *et al.*, 2010, Astrophys. J. Lett. **713**, L79.
Kolb, E. W., and M. S. Turner, 1990, *The Early Universe* (Addison-Wesley, Redwood City).
Komatsu, E., *et al.*, 2009, Astrophys. J. Suppl. Ser. **180**, 330.
Kone, A., and D. Kofke, 2005, J. Chem. Phys. **122**, 206101.
Kornejew, P., M. Hirsch, T. Bindemann, A. Dinklage, H. Dreier, and H. J. Hartfuss, 2006, Rev. Sci. Instrum. **77**, 10F128.
Kosowsky, A., and M. S. Turner, 1995, Phys. Rev. D **52**, R1739.
Krieger, K., U. von Toussaint, and The ASDEX-Upgrade team, 1999, in *Proceedings of the 26th EPS Conference on Controlled Fusion and Plasma Physics* (European Physical Society, Maastricht), Vol. ECA 23J, pp. 1529.
Kschischang, F. R., B. J. Frey, and H. A. Loeliger, 2001, IEEE Trans. Inf. Theory **47**, 498.
Kulcsar, C., L. Pronzato, and E. Walter, 1994, Int. J. Bio-Medical Computing **36**, 95.
Lauritzen, S. L., 1996, *Graphical Models* (Oxford University Press, Oxford, UK).
Lauritzen, S. L., and D. J. Spiegelhalter, 1988, J. R. Stat. Soc. Ser. B **50**, 157 [http://www.jstor.org/stable/2345762].
Leach, S. M., A. R. Liddle, J. Martin, and D. J. Schwarz, 2002, Phys. Rev. D **66**, 023515.
Leamer, E. E., 1978, *Specification Searches: Ad Hoc Inference With Nonexperimental Data* (John Wiley, New York).
Leonard, J. J., and H. Durrant-Whyte, 1992, *Directed Sonar Sensing for Mobile Robot Navigation* (Kluwer, Dordrecht).
Leonard, T., and J. Hsu, 1999, *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers* (Cambridge University Press, Cambridge).
Lepage, G., 1980, VEGAS: An Adaptive Multidimensional Integration Program, Technical Report No. CLNS-80/447, Department of Engineering, Cornell University.
Lewis, A., A. Challinor, and A. Lasenby, 2000, Astrophys. J. **538**, 473.
Liddle, A. R., 2004, Mon. Not. R. Astron. Soc. **351**, L49.
Liddle, A. R., 2007, Mon. Not. R. Astron. Soc. **377**, L74.
Lidsay, J. E., A. R. Liddle, E. W. Kolb, E. J. Copeland, and T. Barreiro, 1997, Rev. Mod. Phys. **69**, 373.
Linde, A. D., 1983, Phys. Lett. B **129**, 177.
Linde, A. D., 1994, Phys. Rev. D **49**, 748.
Lindley, D., *et al.*, 1980, in *Bayesian Statistics*, edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Valencia University Press, Valencia, Spain), p. 223.

Lindley, D. V., 1956, Ann. Math. Stat. **27**, 986.
Lindley, D. V., 1972, *Bayesian Statistics—A Review* (SIAM, Philadelphia).
Liu, J., 2001, *Monte Carlo Strategies in Scientific Computing* (Springer, New York).
Loredo, T. J., 1990, in *Maximum Entropy and Bayesian Methods*, edited by P. Fougere (Kluwer Academic Publishers, Dordrecht), p. 81.
Loredo, T. J., 1992, in *Statistical Challenges in Modern Astronomy*, edited by E. Feigelson and G. Babu (Springer, New York), p. 275.
Loredo, T. J., and D. Chernoff, 2003a, in *Statistical Challenges in Astronomy*, edited by E. Feigelson and G. Babu (Springer, Berlin), p. 57.
Loredo, T. J., and D. Chernoff, 2003b, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by G. Erickson and Y. Zhai, AIP Conf. Proc. No. 707 (AIP, Melville, NY), p. 330.
Loredo, T. J., 1994, The Return of the Prodigal: Bayesian Inference in Astrophysics, manuscript, 37 pages [http://www.astro.cornell.edu/staff/loredo/bayes/return.pdf].
Loredo, T. J., 1999, in *ASP Conference Series 172: Astronomical Data Analysis Software and Systems VIII*, edited by D. M. Mehringer, R. L. Plante, and D. A. Roberts (ASP, San Francisco), Vol. 8, p. 297.
Loredo, T. J., J. Rice, and M. L. Stein, 2009, Ann. Appl. Stat. **3**, 1.
Loschi, R. H., F. R. B. Cruz, R. H. C. Takahashi, P. L. Iglesias, R. B. Arellano-Valle, and J. MacGregor Smith, 2008, Comput. Oper. Res. **35**, 156.
Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter, 2000, Stat. Comput. **10**, 325.
Luttrell, S. P., 1985, Inverse Probl. **1**, 199.
MacKay, D., 2003, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge).
MacKay, D. J. C., 1992a, Neural Comput. **4**, 415.
MacKay, D. J. C., 1992b, Neural Comput. **4**, 590.
Mackenzie, D., 2004, New Sci. **2453**, 36 [http://www.newscientist.com/article/mg18224535.500-vital-statistics.html].
Madigan, D., and A. E. Raftery, 1994, J. Am. Stat. Assoc. **89**, 1535.
Malinverno, A., and V. Briggs, 2004, Geophysics **69**, 1005.
Manyika, J., and H. Durrant-Whyte, 1995, *Data Fusion and Sensor Management: A Decentralized Information-Theoretic Approach* (Prentice Hall PTR, Upper Saddle River, NJ).
Marchand, E., and W. Strawderman, 2004, in *A Festschrift for Herman Rubin*, edited by A. DasGupta, Monograph Series 45 (Institute of Mathematical Statistics Lecture Notes, Beachwood, OH), p. 21.
Marin, J., K. Mengersen, and C. P. Robert, 2005, in *Handbook of Statistics*, edited by D. Dey and C. R. Rao (Elsevier, Amsterdam), Vol. 25.
Marinari, E., and G. Parisi, 1992, Europhys. Lett. **19**, 451.
Mather, J. C., 2007, Rev. Mod. Phys. **79**, 1331.
Matoba, T., T. Itagaki, T. Yamauchi, and A. Funahashi, 1979, Jpn. J. Appl. Phys. **18**, 1127.
Matthews, P., 1993, Stat. Prob. Lett. **17**, 231.
Mayer, M., 1999, in *Proceedings of the 15th International Conference on the Application of Accelerators in Research and Industry*, edited by J. Duggan and I. Morgan, AIP Conf. Proc. No. 475 (AIP, Melville, NY), p. 541.
Mayer, M., R. Fischer, S. Lindig, U. v. Toussaint, R. W. Stark, and V. Dose, 2005, Nucl. Instrum. Methods Phys. Res., Sect. B **228**, 349.
Mayor, M., and D. Queloz, 1995, Nature (London) **378**, 355.
McEliece, R. J., D. J. C. MacKay, and J. F. Cheng, 1998, IEEE J. Selected Areas Comm. **16**, 140.
Meier, M., R. Preuss, and V. Dose, 2003, New J. Phys. **5**, 133.
Mengersen, K., C. Robert, and C. Cuihenneuc-Jouyaux, 1999, in *Bayesian Statistics*, edited by J. Bernardo, J. Berger, A. Dawid, and A. Smith (Clarendon Press, Oxford), Vol. 6, p. 415.
Mertens, S., 2002, Comput. Sci. Eng. **4**, 31.
Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, 1953, J. Chem. Phys. **21**, 1087.
Miller, A. J., 1984, J. R. Stat. Soc. Ser. A (General) **147**, 389.
Möller, W., and F. Besenbacher, 1980, Nucl. Instrum. Methods **168**, 111.
Montemerlo, M., S. Thrun, D. Koller, and B. Wegbreit, 2002, in *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02)* (AAAI Press, Edmonton, Alberta, Canada).
Moral, P. D., A. Doucet, and A. Jasra, 2006, J. R. Stat. Soc. Ser. B **68**, 411.
Mukherjee, P., D. Parkinson, P. S. Corasaniti, A. R. Liddle, and M. Kunz, 2006, Mon. Not. R. Astron. Soc. **369**, 1725.
Mukherjee, P., D. Parkinson, and A. Liddle, 2006, Astrophys. J. **638**, L51.
Muller, H., 1992, Ann. Stat. **20**, 737.
Müller, P., 1999, in *Bayesian Statistics*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford University Press, Oxford), Vol. 6, p. 459.
Müller, P., D. Berry, A. Grieve, M. Smith, and M. Krams, 2007, J. Stat. Plann. Infer. **137**, 3140.
Müller, P., and G. Parmigiani, 1995, in *Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner*, edited by D. A. Berry, K. M. Chaloner, and J. F. Geweke (Wiley, New York), p. 397.
Müller, P., B. Sansó, and M. Delorio, 2004, J. Am. Stat. Assoc. **99**, 788.
Murphy, K., 2001, Learning Bayes Net Structure from Sparse Data Sets, Technical Report, Comp. Sci. Division, UC Berkeley.
Murphy, K., Y. Weiss, and M. I. Jordan, 1999, in *Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference* (Morgan and Kaufmann, Stockholm), p. 467.
Murray, C. D., and S. F. Dermott, 1999, Solar System Dynamics (Cambridge University Press, Cambridge).
Myers, R. H., A. I. Khuri, and W. H. Carter, 1989, Technometrics **31**, 137.
NAG, 2008, *The Numerical Algorithms Group Ltd* (Barnett House, 53 Fountain Street, Manchester).
Neal, R., 1993, Probabilistic Inference Using Markov Chain Monte Carlo Methods, Technical Report No. CRG-TR-93-1, University of Toronto.
Neal, R., 1996, *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics Vol. 18 (Springer, New York).
Neal, R., 1999a, in *Bayesian Statistics*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford University Press, Oxford), Vol. 6, p. 69.
Neal, R., 1999b, in *Learning in Graphical Models*, edited by M. Jordan (MIT Press, MIT), p. 205.
Neal, R., 2003, Ann. Stat. **31**, 705.
Neal, R., 2011, in *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. Jones, and X. Meng (CRC Press, Boca Raton) [http://www.cs.utoronto.ca/~radford/ftp/ham-mcmc.ps].
Ng, A. Y., and M. I. Jordan, 2000, in *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference* (Morgan Kaufmann, Stanford, CA), p. 406.
Oakley, J. E., and A. O'Hagan, 2004, J. R. Stat. Soc. Ser. B **66**, 751.
Oakley, J. E., and A. O'Hagan, 2007, Biometrika **94**, 427.
Obric, M., *et al.*, 2006, Mon. Not. R. Astron. Soc. **370**, 1677.
Ogata, Y., 1989, Numer. Math. **55**, 137.

O'Hagan, A., 1994, *Advanced Theory of Statistics*, Bayesian Inference (Arnold, London), Vol. 2B.
O'Hagan, A., 2006, Bayesian Analysis **1**, 445.
O'Hagan, A., and J. O. Berger, 1988, J. Am. Stat. Assoc. **83**, 503.
O'Hagan, T., and M. West, 2010, *The Oxford Handbook of Applied Bayesian Analysis* (Oxford University Press, Oxford, UK).
Padayachee, J., V. M. Prozesky, W. von der Linden, M. S. Nkwinika, and V. Dose, 1999, Nucl. Instrum. Methods Phys. Res., Sect. B **150**, 129.
Pahud, C., A. R. Liddle, P. Mukherjee, and D. Parkinson, 2007, Mon. Not. R. Astron. Soc. **381**, 489.
Pearl, J., 1987, Artif. Intell. **32**, 245.
Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems* (Morgan and Kaufmann, San Mateo, CA).
Pearl, J., 2000, *Causality* (Cambridge University Press, Cambridge, UK).
Peebles, P. J. E., 1982, Astrophys. J. **263**, L1.
Peot, M., and R. Shachter, 1991, Artif. Intell. **48**, 299.
Perreault, L., J. Bernier, B. Bobee, and E. Parent, 2000, J. Hydrology **235**, 221.
Peskun, P., 1973, Biometrika **60**, 607.
Pohl, C., and J. Van Genderen, 1998, Int. J. Remote Sensing **19**, 823.
Polsen, N., 1996, in *Bayesian Statistics*, edited by J. Bernardo, J. Berger, A. Dawid, and A. Smith (Oxford University Press, Oxford), Vol. 5, p. 297.
Press, W., S. Teukolsky, W. Vetterlin, and B. Flannery, 1996, *Numerical Recipes in Fortran 90* (Cambridge University Press, Cambridge, England).
Preuss, R., H. Dreier, A. Dinklage, and V. Dose, 2008, Europhys. Lett. **81**, 55 001.
Preuss, R., H. Kang, T. Schwarz-Selinger, and V. Dose, 2002, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by R. L. Fry, AIP Conf. Proc. No. 617 (AIP, Melville, NY), p. 155.
Preuss, R., M. Maraschek, H. Zohm, and V. Dose, 2003, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by C. Williams, AIP Conf. Proc. No. 659 (AIP, Melville, NY), p. 124.
Preuss, R., P. Pecher, and V. Dose, 2001, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by C. R. S. J. Rychert and G. Erickson, AIP Conf. Proc. No. 567 (AIP, Melville, NY), p. 213.
Preuss, R., and U. von Toussaint, 2007, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by K. Knuth, A. Caticha, J. Center, A. Giffin, and C. Rodriguez, AIP Conf. Proc. No. 954 (AIP, Melville, NY), p. 221.
Pronzato, L., 2008, Automatica **44**, 303.
Prosper, H. B., 1984, J. Phys. Colloques **45**, C3-185.
Prosper, H. B., 1985, Nucl. Instrum. Methods Phys. Res., Sect. A **241**, 236.
Prosper, H. B., 1988, Phys. Rev. D **37**, 1153.
Prosper, H. B., 2007, in *Statistical Challenges in Modern Astronomy IV*, edited by G. J. Babu and E. D. Feigelson (ASP, Orem, UT), Vol. 371, p. 87.
Pukelsheim, F., 1993, *Optimal Design of Experiments* (Wiley, New York).
Punska, O., 1999, Bayesian Approaches to Multi-Sensor Data Fusion, Ph.D. thesis, University of Cambridge, Cambridge, UK.
Quartulli, M., and M. Datcu, 2003, IEEE Transactions on Geoscience and Remote Sensing **41**, 1976.
Raftery, A. E., D. Madigan, and J. A. Hoeting, 1997, J. Am. Stat. Assoc. **92**, 179.
Rathore, N., M. Chopra, and J. de Pablo, 2005, J. Chem. Phys. **122**, 024111.
Redner, R., and H. Walker, 1984, SIAM Rev. **26**, 195.
Reichardt, C. L., P. A. R. Ade, J. J. Bock, J. R. Bond, and J. A. Brevik, 2009, Astrophys. J. **694**, 1200.
Richardson, S., and P. Green, 1997, J. R. Stat. Soc. Ser. B **59**, 731.
Richardson, S., and P. Green, 1998, J. R. Stat. Soc. Ser. B **60**, U3.
Ripley, B., 1987, *Stochastic Simulation* (Wiley, New York).
Robert, C. P., and G. Casella, 2010, *Introducing Monte Carlo Methods with R* (Springer, New York).
Robert, C. P., 1994, *The Bayesian Choice* (Springer, Berlin).
Robert, C. P., and G. Casella, 1999, *Monte Carlo Statistical Methods* (Springer, New York).
Gelman, A., W. R. Gilks, and G. O. Roberts, 1997, Ann. Appl. Probab. **7**, 110.
Roberts, G. O., and J. S. Rosenthal, 2001, Stat. Sci. **16**, 351.
Roberts, G. O., and J. S. Rosenthal, 2009, J. Comput. Graph. Stat. **18**, 349.
Roth, J., 1999, J. Nucl. Mater. **266–269**, 51.
Roth, J., and C. Garcia-Rosales, 1996, Nucl. Fusion **36**, 1647.
Roweis, S., and Z. Ghahramani, 1999, Neural Comput. **11**, 305.
Rudoy, D., S. G. Yuen, R. D. Howe, and P. J. Wolfe, 2010, J. Roy. Stat. Soc. Ser. C **59**, 573.
Russell, S., and P. Norvig, 2003, *Artifical Intelligence: A Modern Approach* (Prentice-Hall, Englewood Cliffs, NJ).
Ryan, K. J., 2003, J. Comput. Graph. Stat. **12**, 585.
Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn, 1989, Stat. Sci. **4**, 409.
Saltelli, A., S. Tarantola, and F. Campolongo, 2000, Stat. Sci. **15**, 377.
Schwarz, G., 1978, Ann. Stat. **6**, 461.
Schwarz-Selinger, T., R. Preuss, V. Dose, and W. von der Linden, 2001, J. Mass Spectrom. **36**, 866.
Sebastiani, P., and H. P. Wynn, 2000, J. R. Stat. Soc. Ser. B **62**, 145.
Seifert, U., 2005, Phys. Rev. Lett. **95**, 040602.
Sellke, T., M. J. Bayarri, and J. O. Berger, 2001, The American Statistician **55**, 62.
Shachter, R. D., and M. Peot, 1989, in *Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence (UAI-89)* (Morgan and Kaufmann, Windsor, Ontario, Canada).
Shaw, J., M. Bridges, and M. Hobson, 2007, Mon. Not. R. Astron. Soc. **378**, 1365.
Sheffield, J., 1975, *Plasma Scattering of Electromagnetic Radiation* (Academic Press, New York).
Sisson, S., 2005, J. Am. Stat. Assoc. **100**, 1077.
Sivia, D., 2006, *Data Analysis—A Baysian Tutorial* (Oxford University Press, Oxford).
Sivia, D., and W. David, 1994, Acta Crystallogr. Sect. A **50**, 703.
Skellam, J. G., 1946, J. R. Stat. Soc. **109**, 296.
Skilling, J., 1991, in *Fundamentals of Maxent in Data Analysis*, edited by B. Buck and V. Macauley (Clarendon Press, Oxford), p. 19.
Skilling, J., 2004a, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by G. Erickson, J. Rychert, and C. Smith, AIP Conf. Proc. No. 735 (AIP, Melville, NY), p. 395.
Skilling, J., 2004b, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by G. Erickson and Y. Zhai, AIP Conf. Proc. No. 707 (AIP, Melville, NY), p. 388.
Skilling, J., 2006, Bayesian Analysis **1**, 833.
Skilling, J., 2010, in *Bayesian Methods in Cosmology*, edited by M. P. Hobson, A. H. Jaffe, A. R. Liddle, P. Mukherjee, and D. Parkinson (Cambridge University Press, Cambridge, UK), p. 3.
Skilling, J., and D. MacKay, 2003, Ann. Stat. **31**, 753.
Skinner, C. H., *et al.*, 2008, Fusion Sci. Technol. **54**, 891.
Smith, A. F. M., and D. J. Spiegelhalter, 1980, J. R. Stat. Soc. Ser. B **42**, 213 [http://www.jstor.org/stable/2984964].

Smith, R., M. Self, and P. Cheeseman, 1990, in *Autonomous Robot Vehicles*, edited by I. J. Cox and G. T. Wilfon (Springer, New York), p. 167.
Sollom, I., A. Challinor, and M. P. Hobson, 2009, Phys. Rev. D **79**, 123521.
Spergel, D. N., *et al.*, 2007, Astrophys. J. Suppl. Ser. **170**, 377.
Spiegelhalter, D. J., 2002, J. R. Stat. Soc. Ser. B **64**, 583.
Spirtes, P., C. Glymour, and R. Scheines, 2000, *Causation, Prediction, and Search* (MIT Press, Cambridge, MA), 2nd ed.
Starobinsky, A. A., 1982, Phys. Lett. **117B**, 175.
Steinerg, D. M., and W. G. Hunter, 1984, Technometrics **26**, 71.
Stephens, D. A., 1994, Appl. Statist. **43**, 159 [http://www.jstor.org/stable/2986119].
Stephens, M., 2000, J. R. Stat. Soc. Ser. B **62**, 795.
Stewart, L., 1987, J. R. Stat. Soc. Series D **36**, 211 [http://www.jstor.org/stable/2348514].
Stoica, P., and Y. Selén, 2004, IEEE Signal Process. Mag. **21**, 36.
Stone, M., 1959, Ann. Math. Stat. **30**, 55.
Stone, M., 1961, Ann. Math. Stat. **32**, 1339.
Svensson, J., A. Dinklage, J. Geiger, A. Werner, and R. Fischer, 2004, Rev. Sci. Instrum. **75**, 4219.
Svensson, J., and A. Werner, 2007, in *Proceedings of the IEEE International Symposium on Intelligent Signal Processing* (IEEE, New York), p. 1.
Tanner, M. A., and W. H. Wong, 1987, J. Am. Stat. Assoc. **82**, 528.
Tarantola, A., 2005, *Inverse Problem Theory and Methods for Model Parameter Estimation* (SIAM, Philadelphia).
Ter Braak, C. J., and J. A. Vrugt, 2008, Stat. Comput. **18**, 435.
Tesmer, J., and M. Nastasi, 1995, Eds., *Handbook of Modern Ion Beam Analysis* (Material Research Society, Pittsburgh, PA).
Thomas, A., D. Spiegelhalter, and W. Gilks, 1992, in *Bayesian Statistics*, edited by J. Bernardo, J. Berger, A. Dawid, and A. Smith (Oxford University Press, Oxford), Vol. 4, p. 837.
Thomopoulos, S., 1990, Journal of Robotic Systems **7**, 337.
Thrun, S., and D. Fox, 2005, *Probabilistic Robotics* (MIT Press, Cambridge, MA).
Thrun, S., W. Burgard, and D. Fox, 1998, Autonomous Robots **5**, 253.
Thrun, S., and M. Montemerlo, 2006, Int. J. Robotics Res. **25**, 403.
Tierney, L., 1994, Ann. Stat. **22**, 1701.
Tierney, L., and J. Kadane, 1986, J. Am. Stat. Assoc. **81**, 82.
Tierney, L., and A. Mira, 1999, Stat. Med. **18**, 2507.
Tinney, C. G., R. P. Butler, G. W. Marcy, H. R. A. Jones, G. Laughlin, B. D. Carter, J. A. Bailey, and S. O'Toole, 2006, Astrophys. J. **647**, 594.
Tinney, C. G., R. P. Butler, G. W. Marcy, H. R. A. Jones, A. J. Penny, C. McCarthy, B. D. Carter, and J. Bond, 2003, Astrophys. J. **587**, 423.
Toman, B., 1999, in *Encyclopedia of Statistical Sciences Update*, edited by S. Kotz, and N. L. Johnson (Wiley, New York), Vol. 3, p. 35.
Toman, B., 1996, J. Am. Stat. Assoc. **91**, 185.
Toussaint, D., 1989, Comput. Phys. Commun. **56**, 69.
Trotta, R., 2007a, Mon. Not. R. Astron. Soc. **378**, 72.
Trotta, R., 2007b, Mon. Not. R. Astron. Soc. **378**, 819.
Trotta, R., 2008, Contemp. Phys. **49**, 71.
Tynan, G., 1998, New Orleans, APS Meeting.
Uson, J. M., and D. T. Wilkinson, 1984, Astrophys. J. **277**, L1.
van Eeden, C., 1995, Can. J. Stat. **23**, 245.
von der Linden, W., 1995, Appl. Phys. A **60**, 155.
von der Linden, W., V. Dose, N. Memmel, and R. Fischer, 1998, Surf. Sci. **409**, 290.
von der Linden, W., V. Dose, J. Padayachee, and V. Prozesky, 1999, Phys. Rev. E **59**, 6527.
von der Linden, W., R. Preuss, and V. Dose, 1999b, in *Maximum Entropy and Bayesian Methods*, edited by W. von der Linden, V. Dose, R. Fischer, and R. Preuss (Kluwer Academic Publishers, Dordrecht).
von Keudell, A., 2002, Thin Solid Films **402**, 1.
von Toussaint, U., 2011 (unpublished).
von Toussaint, U., and V. Dose, 2005, Appl. Phys. A **82**, 403.
von Toussaint, U., V. Dose, and A. Golan, 2004, J. Vac. Sci. Technol. A **22**, 401.
von Toussaint, U., R. Fischer, K. Krieger, and V. Dose, 1999, New J. Phys. **1**, 11.
von Toussaint, U., S. Gori, and V. Dose, 2004, Appl. Opt. **43**, 5356.
von Toussaint, U., K. Krieger, R. Fischer, and V. Dose, 1999, in *Maximum Entropy and Bayesian Methods*, edited by W. von der Linden, V. Dose, R. Fischer, and R. Preuss (Kluwer Academic Publishers, Dordrecht).
von Toussaint, U., T. Schwarz-Selinger, and S. Gori, 2008, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by M. S. Lauretto, C. A. B. Pereira, and J. M. Stern, AIP Conf. Proc. No. 1073 (AIP, Melville, NY), p. 348.
Wainwright, M. J., and M. I. Jordan, 2003, Graphical Models, Exponential Families, and Variational Inference, Technical Report No. 649, Department Statistics, UC Berkeley.
Wald, L., 1998, in *Proceedings EARSeL Symposium 1997: Future Trends in Remote Sensing*, edited by P. Gudmansen (A. A. Balkema, Rotterdam), p. 385.
Weiss, Y., 2000, Neural Comput. **12**, 1.
Whyte, D., G. Tynan, R. Doerner, and J. Brooks, 2001, Nucl. Fusion **41**, 47.
Wolpert, R. L., and K. Ickstadt, 2004, Inverse Probl. **20**, 1759.
Wright, J. T., 2005, Publ. Astron. Soc. Pac. **117**, 657.
Wright, J. T., and A. W. Howard, 2009, Astrophys. J. Suppl. Ser. **182**, 205.
Y. Wu, H. Tjelmeland, and M. West, 2007, J. Comput. Graph. Stat. **16**, 44.
Zhao, X., and P. Chu, 2010, J. Clim. **23**, 1034.