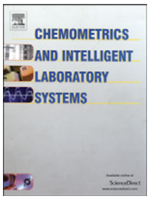
Chemometrics and Intelligent Laboratory Systems 97 (2009) 194–210

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab



CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS

An introduction to Bayesian methods for analyzing chemistry data Part 1: An introduction to Bayesian theory and methods

N. Armstrong *, D.B. Hibbert

School of Chemistry, University of New South Wales, Sydney NSW 2052, Australia

ARTICLE INFO

Article history:
Received 30 June 2008
Received in revised form 2 April 2009
Accepted 4 April 2009
Available online 9 April 2009

Keywords:
Bayesian analysis
Chemistry data
Parameter estimation
Model selection
Data analysis

ABSTRACT

In this tutorial paper, we outline the application of Bayesian theory and methods for analysing experimental chemistry data. We provide an overview of the background theory and the essential rules necessary for manipulating conditional probabilities and density functions (pdfs) i.e. the product and marginalisation rules. Drawing on these rules we demonstrate, using a variety of examples from chemistry, how Bayes theorem can be adapted to analyse and interpret experimental data for a wide range of typical chemistry experiments, including basic model selection (i.e. hypothesis testing), parameter estimation, peak refinement and advanced model selection. An outline of the steps and underlying assumptions are presented, while necessary mathematics are also discussed.

1. Introduction

The availability of data acquisition software and hardware means that experimental data can be collected relatively continuously and in a timely manner. In addition, if computational modeling is necessary, it can also be carried out to complement the experimental analysis. The implications of this is that scientific models can be analysed and tested quickly. Although this analysis is carried out routinely, it still requires knowledge, understanding and experience of the scientist to interpret the data and subsequent results.

This underlying reasoning process is bundled up into the "scientific method", in which hypothesis are formulated and tested, in the presence of the experimental data. Bayesian analysis captures this reasoning process [1–3], by defining probabilities as quantifying the plausibility or truth of a proposition, which are conditional on the available data and information. This theory and treatment of probabilities was originally formulated in the eighteenth century by Thomas Bayes (1702–1761). However, it has been rediscovered and adopted by the wider scientific community as an analysis tool only in the last two decades. A possible reason for this "slow uptake" is that the necessary computational processing (both hardware and software) and numerical techniques, such as Monte Carlo methods, are readily available and widely applied (see [3–9]). In addition, the application of the maximum entropy methods and its formulation in a Bayesian context has demonstrated its versatility in solving a wide range of scientific problems (see [3], ch. 6 and references therein).

In this tutorial paper, we outline some of the key components of Bayesian analysis, with particular attention directed towards demonstrating its application in analysing experimental chemistry data. One of the advantages of Bayesian analysis over orthodox data analysis techniques, is that a firm knowledge of the basic rules of probability theory enables complex data analysis problems to be solved. Using minimal assumptions, the corresponding posterior probability density functions (pdf) can be derived. The posterior pdf captures all the necessary information about a parameter, such as the most probable value, average value, standard deviation and confidence (or credibility) intervals for a parameter. However, its greatest advantage is its application is in *model selection* or *hypothesis testing*, where the definition of probabilities as conditional quantities is critical.

The central aim of this paper is to introduce these key components by analysing "typical" experimental chemistry data. We also attempt to maintain the mathematical component, while outlining all the necessary steps in deriving the corresponding posterior pdfs. The present tutorial is based on a number of well known discussions about Bayesian theory, including D'Agostini [1], Sivia and Skilling [3], Gregory [8], Ó Ruanaidh and Fitzgerald [9], Dose [10], and MacKay [11].

The outline of the paper includes an overview of probabilities, the rules needed for manipulating probabilities which are presented in Section 2, includes the *product* and *addition* rules, from which Bayes theorem is derived (see Sections 2.1–2.5). Some introductory applications of Bayes theorem are presented in Section 2.5. These include determining the posterior pdf for the mass fraction of calcium in milk powder using two different experimental techniques

* Corresponding author. Former address: Department of Physics and Advanced Materials, University of Technology Sydney, NSW 2007, Australia.
E-mail address: Nicholas.G.Armstrong@gmail.com (N. Armstrong).

0169-7439/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2009.04.001

(see Section 3.2), while the underlying question of whether the experimental techniques produce consistent results is addressed in Section 3. An example of parameter estimation related to determining the concentration of lead in solution is presented in Section 5. Section 6 discusses how Bayesian and maximum entropy methods can be applied to deconvoluting spectra, which is illustrated by determining the parameters and most probable number of peaks in X-ray photoelectron spectroscopy (XPS) data. The analysis of the XPS data provides an advanced example of parameter estimation and model selection which employ Markov Chain Monte Carlo (MCMC) to sample the posterior pdfs. A summary of the tutorial is presented in Section 7. Appendix A lists the notation throughout the paper, while Appendix B summarizes MCMC methods for sampling posterior pdfs. In a companion paper [12], we critically review reports of the use of Bayesian methods in the chemical literature.

2. Rules for manipulating probabilities

2.1. Probability theory as a process of reasoning

In the following discussion, we offer key elements needed to establish a process of reasoning. It raises the following question, “How do we formulate this process and quantify it?” This is a scientific question which Cox [13] addressed using Boolean logic to establish the rules for consistent and logical reasoning [1,3] and demonstrated that this reasoning follows the rules for manipulating probabilities. That is, Cox [13] makes the connection between the proposition and values of probabilities [1]. The probability of a proposition is quantified as a real, positive number in the range of [0, 1]. Here probability defines the “degree of belief” or *plausibility* that a given proposition is true [1,2,8,11,13]. This definition is not the same as the commonly quoted definition of probability, which is the relative frequency of a proposition from a large number of repeated trials. The *frequentist* definition raises some interesting epistemological issues regarding probability theory. For example, how do we define the probability density function of a parameter from a single data set, such as a intensity pattern or spectrum? These issues are resolved if a Bayesian interpretation of probability is adopted.

In the following we outline the rules for manipulating probabilities, beginning with the terminology used in defining them [1,3].

Let x represent a proposition, such as “will it rain today?” or “what is the average increase in ambient temperature resulting from global climate change?” or “what is the average mass fraction of calcium in a sample of milk powder?”.

The probability that a proposition x has a particular value (for the examples given “yes”, “(3.4 ± 0.3) K”, “2.69 ± 0.43 mg g^{-1}”) is denoted by $\Pr(x|\mathcal{I})$, where “$|\mathcal{I}$” indicates that the probability is *conditional on* some background information, $\mathcal{I}$. That is, the probability of anything to the left of “|” is conditional on anything to the right of “|”. In addition, care is taken to declare the appropriate information which the probability is based. This underscores an important property of consistent reasoning process, that in order to make an inference concerning a particular proposition, the available and appropriate information must be declared and taken into account [3]. An interesting example of this is to be found in the debate about the notorious 1995 ‘O. J. Simpson’ trial, where the distinction was made between the probability of a wife beater murdering his wife (i.e. $\mathcal{I}$ = *wife beater*), and the probability that given a dead wife and given a of wife beating, a husband is guilty (i.e. $\mathcal{I}$ = *wife beater and dead wife*) [14].

An important property of probabilities, $\Pr(x|\mathcal{I})$, is that they are bounded below by 0 and above by 1, such that $0\leq\Pr(x|\mathcal{I})\leq1$, where $x\in\Theta$ for a *complete* set of propositions or hypotheses, Θ [3,13]. This property asserts the *positivity property* of probabilities. For example, suppose $x\cup\overline{x}$, where $\overline{x}$ represents the proposition of x *not* true, then the probabilities must be normalised to unity over Θ to satisfy the abovementioned property,

$$\Pr(x|\mathcal{I})+\Pr(\overline{x}|\mathcal{I})=1. \tag{1}$$

This condition in turn asserts the *additivity property* of probabilities — see Section 2.3 for more details. In general, for a complete set of N independent probabilities, we expect them to be normalised, such that $\sum_{i=1}^{N}\Pr(x_i|\mathcal{I})=1$.

2.2. Product rule

The product rule for probabilities defines the joint probability of two or more propositions. Given the two propositions, x and y, the probability that x and y are both true is the probability that one, say x, is true multiplied by the probability that y is true, given the truth of x,

$$\Pr(x,y|\mathcal{I})=\Pr(x|\mathcal{I})\Pr(y|x,\mathcal{I}), \tag{2a}$$

and because x and y are interchangeable, we also have,

$$\Pr(x,y|\mathcal{I})=\Pr(y|\mathcal{I})\Pr(x|y,\mathcal{I}), \tag{2b}$$

where $\Pr(x,y|\mathcal{I})$ reads the probability of “x and y” conditional on $\mathcal{I}$. The product rule enables a joint probability to be decomposed into its constituents, and demonstrates that such conditions are in fact symmetrical. In the case of probabilistic independence, where $\Pr(x|y,\mathcal{I})=\Pr(x|\mathcal{I})$ and $\Pr(y|x,\mathcal{I})=\Pr(y|\mathcal{I})$, the equations become,

$$\Pr(x,y|\mathcal{I})=\Pr(y|\mathcal{I})\Pr(x|\mathcal{I}). \tag{3}$$

An alternative way of expressing Eqs. (2a) and (2b) is,

$$\begin{aligned}\Pr(x\cap y|\mathcal{I})&\equiv\Pr(x,y|\mathcal{I})\\&=\Pr(x|\mathcal{I})+\Pr(y|\mathcal{I})-\Pr(x\cup y|\mathcal{I}),\end{aligned} \tag{4}$$

where $\Pr(x\cap y|\mathcal{I})$ and $\Pr(x\cup y|\mathcal{I})$ define the intersection (logical AND) and union (logical OR) of the propositions of x and y.

2.3. Addition rule or ‘marginalisation’

Now suppose we have a set of propositions $\{x_i;i=1,2,3,\ldots,M\}$ and $\{y_j;i=1,2,3,\ldots,N\}$, respectively and where $M\neq N$. The generalisation of the additivity properties can be expressed as

$$\Pr\Big(\{y_j\}|\mathcal{I}\Big)=\sum_{i=1}^{M}\Pr\Big(\{y_j\},\{x_i\}|\mathcal{I}\Big) \tag{5a}$$

$$=\sum_{i=1}^{M}\Pr(\{x_i\}|\mathcal{I})\Pr\Big(\{y_j\}|\{x_i\},\mathcal{I}\Big), \tag{5b}$$

where we have used the product rule to decompose the joint probability in going from Eq. (5a) to Eq. (5b). Also by symmetry, we have

$$\Pr(\{x_i\}|\mathcal{I})=\sum_{j=1}^{N}\Pr\Big(\{y_j\},\{x_i\}|\mathcal{I}\Big) \tag{6a}$$

$$=\sum_{j=1}^{N}\Pr\Big(\{y_j\}|\mathcal{I}\Big)\Pr\Big(\{x_i\}|\{y_j\},\mathcal{I}\Big). \tag{6b}$$

In the above cases, we have defined *marginalisation* of joint probabilities for a discrete set of variables. Both Eqs. (5b) and (6b) are also known as the Total Probability Theorem. The same holds for continuous variables, such that $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$,

$$\Pr(x|\mathcal{I}) = \int_{\mathcal{Y}} \Pr(x, y|\mathcal{I})\mathrm{d}y \tag{7a}$$

and

$$\Pr(y|\mathcal{I}) = \int_{\mathcal{X}} \Pr(x, y|\mathcal{I})\mathrm{d}x. \tag{7b}$$

In the above mentioned case, marginalisation can be considered as "integrating out" unnecessary variables.

2.4. Background information

In the present discussion we have defined conditional probabilities with respect to the background information, $\mathcal{I}$. Many presentations and discussions on Bayesian theory neglect this term, which is usually taken for granted. This undermines an important feature of Bayesian analysis, in that it quantifies our present state of knowledge based on the available information. By including the background information, $\mathcal{I}$, we are stating the underlying assumptions which specify the probabilities [1,3].

For example, suppose a data set was collected for particular instrument settings, these settings would specify the background information, $\mathcal{I}_1$, on which any subsequent probabilities would be conditional. If the data collection were repeated for different instrument settings, this information would correspond to another state of background information, $\mathcal{I}_2$. Given these different background information, then in general the corresponding conditional probabilities will not be equal, $\Pr(x|\mathcal{I}_1) \neq \Pr(\mathrm{x}|\mathcal{I}_2)$ [1,3]. Intuitively, this makes sense in that when we perform experiments, we attempt to ensure that the experimental conditions are consistent and thus we can draw sensible conclusions from our data.

2.5. Bayes theorem

Two fundamental rules of probability theory have been presented, namely the addition and product rules. However, we need a final rule which "unifies" the two rules into a single rule, and tells us how to treat conditional probabilities. This is essentially the role of Bayes theorem as it follows directly from the product and addition rules [3]. Suppose, we want to determine the probability of a continuous parameter, θ, given the available data, D. For example θ could be the value of a quantity that is estimated as the mean of independent measurements. We can use the joint probability rules to express the probability of θ and D, given the background information, $\mathcal{I}$, as

$$\Pr(\theta, D|\mathcal{I}) = \Pr(\theta|\mathcal{I})\Pr(D|\theta, \mathcal{I}) \tag{8a}$$

$$= \Pr(D|\mathcal{I})\Pr(\theta|D, \mathcal{I}). \tag{8b}$$

We notice the symmetry in Eqs. (8a) and (8b), and by equating the right-hand sides, we obtain Bayes theorem,

$$\Pr(\theta|D, \mathcal{I}) = \frac{\Pr(\theta|\mathcal{I})\Pr(D|\theta, \mathcal{I})}{\Pr(D|\mathcal{I})}, \tag{9}$$

where the denominator is $\Pr(D|\mathcal{I}) = \int_{\Theta} \Pr(\theta, D|\mathcal{I})\mathrm{d}\theta$ with $\Theta \subseteq \mathbb{R}$. Each term in Eq. (9) may be interpreted as follows.

The left-hand side of Eq. (9) asserts the plausibility of θ given the data, D. This is termed the posterior probability of θ conditional on D and $\mathcal{I}$. In other words, $\Pr(\theta|D, \mathcal{I})$ quantifies a scientist's state of knowledge "after" performing the experiment and obtaining the data. On the right-hand side, the prior information is defined by $\Pr(\theta|\mathcal{I})$, which asserts the plausibility of θ "before" the experiment has been conducted. The likelihood probability, $\Pr(D|\theta, \mathcal{I})$ quantifies the plausibility of the data, D, conditional on θ and $\mathcal{I}$. An important observation to make regarding Eq. (9) is the role the prior term plays in reversing the statement asserting "the plausibility of the data, D, given θ and $\mathcal{I}$" to asserting "the plausibility of θ conditional on D and $\mathcal{I}$". Logically, these statements are *not* equal, and it is the assertions of the prior which plays a critical role. In many applications of Bayesian analysis, such as parameter estimation, it is sufficient to state,

$$\Pr(\theta|D, \mathcal{I}) \propto \Pr(\theta|\mathcal{I})\Pr(D|\theta, \mathcal{I}), \tag{10}$$

and ignore the denominator, since it plays the role of normalisation constant. The resulting equation can be used to determine the most probable parameter value, say $\hat{\theta}$, such that $\frac{\mathrm{d}\Pr(\theta|D,\mathcal{I})}{\mathrm{d}\theta} = 0$ for $\theta = \hat{\theta}$. Alternatively, Eq. (10) can be normalised for unit area from which the average and standard deviation, μ_θ (or $\langle\theta\rangle$) and σ_θ, respectively for θ can be determined. By using probabilities, the plausibility of the parameter value (i.e. μ_θ) is being quantified as a real number and mapped into the region of [0, 1].

Bayes discovered the way to combine prior information for a hypothesis with the probability of the observed data if the hypothesis were true to give the probability of the hypothesis itself. This is precisely what a scientist wants to do. Taking background information and expertise, $\mathcal{I}$, together with the outcome of experiments, D, to obtain the probability of his/her hypothesis, $\Pr(\theta|D, \mathcal{I})$. Merely being able to state that "if my hypothesis is correct then the probability I would find these data is ..." is only half the problem and does not answer "given these data what is the probability that my hypothesis is correct?" Note that it is the former probability that is calculated in a traditional frequentist test, such as a Student *t*-test of two means. In this particular case, the null hypothesis that two data sets being compared come from populations with the same mean ($\mu_1 = \mu_2$) and the probability that values greater than the observed difference in means would be found in repeated measurements is calculated. This probability is not the probability of the null hypothesis.

In the above discussion, we have introduced Bayes theorem in terms of probabilities. The same rules also apply for probability density functions (pdf). The probability (Pr) and pdf (p) are related in the following definition [1,3],

$$\Pr(x_1 \leq x \leq x_2|D, \mathcal{I}) = \int_{x_1}^{x_2} p(x|D, \mathcal{I})\mathrm{d}x, \tag{11}$$

where we have assumed continuous parameters. In the case of a discrete pdf [1,3],

$$\Pr(X = x_i|D, \mathcal{I}) = p(x_i|D, \mathcal{I}). \tag{12}$$

In summary, the basic properties and rules of probability theory have been introduced, while the importance of the background information has been outlined. As stated above the probabilities reflect our state of belief or knowledge given the available information. Our ability as scientists to make consistent decisions depends on the available information. The basic rules of probability theory provide the means to assert the plausibility of a particular proposition by application of Bayes theorem. How can this theorem be used to address typical chemical problems?

3. What can we do with Bayes theorem? Some simple examples

In this section we outline some applications of Bayes theorem that relate to chemistry. First we develop the theory for estimating of population mean and standard deviation from repeated measurement results. This allows examples to be developed: estimation of optimum parameters for a model, whether two sets of data have been measured on the same sample.

3.1. Estimating the mean and standard deviation of a sample

Here we present the general problem of estimating both mean, μ, and standard deviation, σ from a series of measurement results. This example will show the influence of prior pdf has on the estimated values of μ and σ.

3.1.1. Bayes theorem for μ and σ

Suppose we make a series of measurements of mass fraction for an element in a substance, such as calcium in milk powder, and record this data as an array, $\mathbf{D}=\{D_i;i=1,2,3,\ldots,N\}$. From the set of measurement results, we want to estimate the true value of a quantity, μ, and standard deviation, σ, mass fraction. The measurements are assumed to be independent of each other and an appropriate model is,

$$D_i = \mu + n_i, \quad \forall i, \tag{13}$$

where n_i is noise that may be considered Gaussian with a zero mean and standard deviation, σ.

The initial steps expressing Bayes theorem for μ and σ conditional on $\mathbf{D}$ and $\mathcal{I}$, and define the prior and likelihood term in the theorem. Essentially, we want to determine the posterior pdfs for μ and σ, from which the most probable values can be determined. The joint posterior pdf for μ and σ can be expressed as,

$$p(\mu,\sigma|\mathbf{D},\mathcal{I}) = \frac{p(\mu,\sigma|\mathcal{I})p(\mathbf{D}|\mu,\sigma,\mathcal{I})}{p(\mathbf{D}|\mathcal{I})}, \tag{14}$$

$p(\mu,\sigma|\mathcal{I})$ is the joint-prior density function; $p(\mathbf{D}|\mu,\sigma,\mathcal{I})$ is the likelihood term and conditional on the parameters of the model; $p(\mathbf{D}|\mathcal{I})$ is a normalising term and involves integrating out all the parameter values,

$$p(\mathbf{D}|\mathcal{I}) = \int_\Sigma\int_\Upsilon p(\mu,\sigma,\mathbf{D}|\mathcal{I})\mathrm{d}\mu\,\mathrm{d}\sigma, \tag{15}$$

where Σ and Υ define the region of integration for σ and μ, with $\sigma\in\Sigma\subseteq\mathbb{R}^+$ and $\mu\in\Upsilon\subseteq\mathbb{R}$, respectively.

From Eq. (14), the posterior pdf for μ and σ, respectively, can be determined by integrating out the respective terms, such as,

$$p(\mu|\mathbf{D},\mathcal{I}) = \int_\Sigma p(\mu,\sigma|\mathbf{D},\mathcal{I})\mathrm{d}\sigma, \tag{16}$$

and similarly, the pdf for σ

$$p(\sigma|\mathbf{D},\mathcal{I}) = \int_\Upsilon p(\mu,\sigma|\mathbf{D},\mathcal{I})\mathrm{d}\mu. \tag{17}$$

All necessary quantities, such as the mean, standard deviation and confidence intervals (or credibility regions — see footnote 1 and references) can be evaluated from Eqs. (16) and (17). The important point, regarding Eqs. (16) and (17) is that both posterior pdfs are conditional on the data, $\mathbf{D}$. In the above case, the prior pdfs for both μ and σ have been defined and resulted in the joint-posterior pdf. On the other hand, if μ (or σ) were known, the subsequent posterior pdf for σ (or μ) would be conditional on the known μ (or σ) and $\mathbf{D}$.

3.1.2. Defining the likelihood function

An aspect that many newcomers find mystifying or even frustrating about Bayesian theory is that the theorem does not tell the researcher how to define the prior pdf and the likelihood function. It essentially tells us how the manipulate probabilities and conditional density functions, while specifying their dependency on background information, and assumptions. On the other hand, this is the underlying strength of the theory, in that it enables posterior pdfs to be defined which take into consideration the assumptions. The posterior pdfs given in Eqs. (14), (16) and (17) can be defined specifically to reflect the available information and data.

For the moment, we will write the likelihood function, and explore the influence of various prior pdfs. By assuming the noise is a Gaussian pdf and independent, the likelihood function is the product of the individual Gaussian pdf for D_i,

$$p(\mathbf{D}|\mu,\sigma,\mathcal{I}) = \prod_{i=1}^{N} p(D_i|\mu,\sigma,\mathcal{I}) \tag{18a}$$

$$= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}(D_i-\mu)^2/\sigma^2\right] \tag{18b}$$

$$= \left(2\pi\sigma^2\right)^{-N/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{N}(D_i-\mu)^2/\sigma^2\right] \tag{18c}$$

$$= \left(2\pi\sigma^2\right)^{-N/2}\exp\left[-\frac{Ns^2}{2\sigma^2}\right]\exp\left[-\frac{N(\mu-\overline{D})^2}{2\sigma^2}\right] \tag{18d}$$

where $(D_i-\mu)/\sigma$ in Eq. (18c) are the weighted-residuals between the data and the model; in Eq. (18d) $\overline{D}=\Sigma_{i=1}^{N} D_i/N$ and $s^2 = \frac{1}{N}\sum_i(D_i-\overline{D})^2$. In general, Eqs. (18a)–(18d) quantify the plausibility (in terms of probabilities) of generating the data, $\mathbf{D}$ conditional on (i.e. "|") the model, μ and σ. As outlined, Eqs. (18a)–(18d) introduces the experimental data, and assumptions regarding the distribution which describe the data.

3.1.3. Something about priors

Now we introduce the prior pdfs, which plays an important role of reversing the likelihood statement, to reflect the plausibility of the parameters μ and σ with respects to the data, $\mathbf{D}$. Assigning prior pdfs is the most controversial aspect of Bayesian analysis. Historically, is has been viewed 'subjective'. Here, we assert that the prior is necessary in turning around the logical statement to reflect scientific reasoning, "the plausibility of the hypothesis or proposition conditional on the data etc..". The prior pdf declares assumptions or beliefs about the proposition. For example, Fig. 1 shows different priors for μ and σ. In Fig. 1(a), the range over which μ is thought to exist is [2.00, 5.00] mg g^{-1}. The uniform prior, discussed below, assigns an equal probability for all values of μ over this range. On the other hand, the Gaussian prior states that the most probable value lies in the (2.70 ± 0.27) mg g^{-1}. Its functional property and tails specify the concentration of probability over which μ exists. The Lorentzian pdf is located in the same region with the same width, but attributes a greater probability to the tails compared to the Gaussian pdf. Outside this specified region the prior assigns zero probability for μ. Fig. 1(b) shows possible prior candidates for σ. Essentially, the both priors quantify the plausibility of σ over a large range, from 0.001 mg g^{-1} to 4 mg g^{-1}. The first prior, $\propto 1/\sigma$, is a uniform prior over the logarithmic scale, while the second prior, $\propto 1/\sigma^2$, attributes a greater probability to small values of σ, while penalising large values of σ.

Our position is that scientific inference and reasoning cannot be made without making assumptions, and Bayesian analysis enables these assumptions to be included explicitly in the reasoning process.

3.1.4. Using uniform priors

Uniform pdfs are the simplest form of priors for μ and σ. The uniformity in the structure of the pdfs reflects ignorance, as it does not

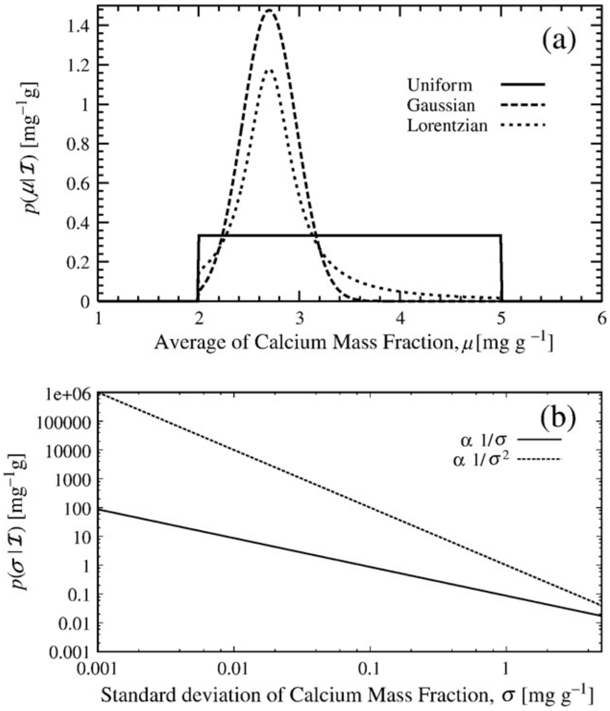


Fig. 1. Possible prior pdfs for μ and σ. (a). Compares uniform, Gaussian and Lorentzian priors, respectively. The Gaussian and Lorentzian pdfs have the same location and width, but their tails attribute different probabilities for large values of μ. (b) Two different priors for σ over a large range. The first prior, ~$1/\sigma$, is a uniform density function over the logarithmic-scale, while the second prior, ~$1/\sigma^2$, assigns greater probability for small of σ and less to large values of σ. (All pdfs have been normalised for unit area.)

place any preference on a particular parameter value. However, the priors do qualify the range in which the parameters are defined. In this case, the prior term in Eq. (14) for μ and σ can be expressed as independent pdfs,

$$p(\mu,\sigma|\mathcal{I}) = p(\mu|\mathcal{I})p(\sigma|\mathcal{I}). \tag{19}$$

The prior for μ is dependent on specified range, $\mu \in [\mu_{\min}, \mu_{\max}]$ — see Fig. 1(a),

$$p(\mu|\mu_{min},\mu_{max},\mathcal{I}) = \begin{cases} \dfrac{1}{\mu_{max} - \mu_{min}}, & \mu_{min} \le \mu \le \mu_{max} \\ 0, & \text{otherwise}. \end{cases} \tag{20}$$

At first we would expect the prior for σ to be defined in a similar manner i.e. uniform over a linear range. However, we note that the prior for μ is shift invariant, while the prior for σ should be scale invariant. In other words, σ can range over orders of magnitude and a Jeffrey's prior defines a uniform pdf over orders of magnitude — see Fig. 1(b),

$$p(\sigma|\sigma_{\min},\sigma_{\max},\mathcal{I}) = \begin{cases} \dfrac{1}{\ln(\sigma_{\max}/\sigma_{\min})\sigma}, & \sigma_{\min} \le \sigma \le \sigma_{\max} \\ 0, & \text{otherwise}. \end{cases} \tag{21}$$

It is also worth noting that Sivia and Skilling [3] and Sivia et al. [15] suggest some useful modifications to Eq. (21), which can be used to address the issue of outliers. For notational convenience, we specify the ranges for μ and σ as $r_\mu = [\mu_{min}, \mu_{max}]$ and $r_\sigma = [\sigma_{min}, \sigma_{max}]$, respectively. Using Eqs. (21) and (20) in Eq. (14),

$$p\left(\mu,\sigma|\mathbf{D},r_\mu,r_\sigma,\mathcal{I}\right) = \frac{\left(2\pi\sigma^2\right)^{-N/2}}{\mu_{\max}-\mu_{\min}} \frac{1}{\ln(\sigma_{\max}/\sigma_{\min})\sigma} \frac{1}{p(\mathbf{D}|\mathcal{I})} \times \exp\left[-\frac{1}{2}\sum_{i=1}^{N}(D_i-\mu)^2/2\sigma^2\right] \tag{22a}$$

$$= \frac{(2\pi)^{-N/2}\sigma^{-(1+N)}}{\mu_{\max}-\mu_{\min}} \frac{1}{\ln(\sigma_{\max}/\sigma_{\min})} \frac{1}{p(\mathbf{D}|\mathcal{I})} \times \exp\left[-\frac{N(\mu-\overline{D})^2}{2\sigma^2}\right] \exp\left[-\frac{Ns^2}{2\sigma^2}\right] \tag{22b}$$

$$p\left(\mu,\sigma|\mathbf{D},r_\mu,r_\sigma,\mathcal{I}\right) = \sigma^{-(1+N)}Z^{-1}\exp\left[-\frac{Q}{2\sigma^2}\right], \tag{22c}$$

where in Eq. (22c), all the constant terms have been adsorbed into the normalisation term, Z, and finally $Q = N(\mu-\overline{D})^2 + Ns^2$.

Given Eqs. (22a)–(22c), we are in a position to explicitly determine Eq. (16). The region of integration is specified by limits $[\sigma_{\min}, \sigma_{\max}]$,

$$p\left(\mu|\mathbf{D},r_\mu,r_\sigma,\mathcal{I}\right) = \int_{\sigma_{\min}}^{\sigma_{\max}} p\left(\mu,\sigma|\mathbf{D},r_\mu,r_\sigma,\mathcal{I}\right)\mathrm{d}\sigma \tag{23a}$$

$$= Z^{-1}Q^{-N/2}\left[\Gamma\left(N/2,\frac{Q}{2\sigma_{\max}^2}\right) - \Gamma\left(N/2,\frac{Q}{2\sigma_{\min}^2}\right)\right], \tag{23b}$$

where Eq. (23b) the limits for σ have been imposed, and the resulting integral is in the form of an incomplete Gamma function, $\Gamma(\xi,x) = \frac{1}{\Gamma(\xi)}\int_x^\infty e^{-t}t^{\xi-1}\mathrm{d}t$. Also note that all the 'unimportant' terms have been absorbed into the normalisation term, Z. If we are willing to extend the limits in Eq. (23b) to ±∞, then,

$$p\left(\mu|\mathbf{D},r_\mu,r_\sigma\mathcal{I}\right) \approx 2^{-1+N/2}Z^{-1}\Gamma(N/2)Q^{-N/2}. \tag{24}$$

It is clear that Eq. (24) is not the typical Gaussian distribution, and is closely related to a Student-t distribution [3,9]. For small N (i.e. $N \lesssim 10$), the tails of Eq. (24) will be extended, and as the number of measurements increase, Eq. (24) approaches a Gaussian distribution. The location of the maximum for $p(\mu|\mathbf{D},\mathcal{I})$ can be found by determining the value of μ where $\mathrm{d}\ln p(\mu|\mathbf{D},\mathcal{I})/\mathrm{d}\mu = 0$, which results in $\hat{\mu} = \overline{D} = \sum_{i=1}^{N} D_i/N$ i.e. this is simply the average. Also note that Eqs. (23a), (23b), and (24) are normalised for unit area. A credibility region or what typically called in orthodox statistics a confidence interval,[1] defines a region in which an estimated quantity exists with probability, C. In the context of posterior pdfs, it defines the

[1] It should be noted that Bayesianist and frequentist have different interpretations regarding credibility regions. Bayesian interpretation makes a probabilistic statement by defining the probability density for the parameter and the probability that the estimated value of the parameter exist within a specified region. On the other hand, orthodox statistics interprets the region as the range in which the parameter's value would exists from repeated samples being randomly drawn from some hypothetical population, see [8]. More specifically, the frequentist would argue that the 95% confidence interval from an analysis of repeated measurements has the property that 95% of the 95% confidence intervals would include the population mean.

lower- and upper-bounds about $\hat{\mu}$ with limits $[\hat{\mu}-\delta_{\mu1}, \hat{\mu}+\delta_{\mu2}]$, such that $\int_{\hat{\mu}-\delta_{\mu1}}^{\hat{\mu}+\delta_{\mu2}} p(\mu|\mathbf{D},\mathcal{I})\mathrm{d}\mu = C$. This case assumes the pdf is symmetric about $\hat{\mu}$. It also assumes that $\hat{\mu}-\delta_{\mu1}>\mu_{\min}$ and $\hat{\mu}+\delta_{\mu2}<\mu_{\max}$, respectively. That is, the lower- and upper-limits are within the limits defined by prior pdf. On the other hand, in Eq. (23b), $\delta_{\mu1}\neq\delta_{\mu2}$, since Eqs. (23a) and (23b) are an asymmetric distribution. In the case of Eq. (24), is a symmetrical distribution and $\delta_{\mu1}=\delta_{\mu2}$. In the former case, we can quote the range as $\hat{\mu}^{+\delta_{\mu2}}_{-\delta_{\mu1}}$, while in the latter case the uncertainties can be specified as $\hat{\mu}\pm\delta_{\mu1}$ for a credibility region, C.

Similarly, we can follow the same approach for $p(\sigma|\mathbf{D},\mathcal{I})$, over the region $[\mu_{\min}, \mu_{\max}]$,

$$p\Big(\sigma|\mathbf{D},r_\mu,r_\sigma,\mathcal{I}\Big) = \int_{\mu_{\min}}^{\mu_{\max}} p(\sigma,\mu|\mathbf{D},\mathcal{I})\mathrm{d}\mu \tag{25a}$$

$$= Z^{-1}\exp\left[-Ns^2/2\sigma^2\right] \times \left\{\mathrm{erf}\left[\frac{\sqrt{N}(\mu_o-\mu_{\max})}{\sqrt{2}\sigma}\right]-\mathrm{erf}\left[\frac{\sqrt{N}(\mu_o-\mu_{\min})}{\sqrt{2}\sigma}\right]\right\}, \tag{25b}$$

where in Eq. (26) we have relaxed the limits to [0, ∞),

$$p\Big(\sigma|\mathbf{D},r_\mu,r_\sigma,\mathcal{I}\Big)\approx Z^{-1}\sigma^{-N}\exp\left[-\sum_i(D_i-\overline{D})^2/2\sigma^2\right]. \tag{26}$$

As in the above case for μ, Eq. (26) is obviously not a Gaussian distribution. This is significant for small N, but as N increases Eq. (26) approaches a Gaussian distribution. The maximum corresponds to $\hat{\sigma} = \sqrt{\sum_i(D_i-\overline{D})^2/N}$ and can be contrasted with other estimates, such as $\overline{\sigma} = \int_\Sigma \sigma\, p(\sigma|\mathbf{D},\mathcal{I})\mathrm{d}\sigma = \frac{\sqrt{N}s}{\sqrt{2}}\frac{\Gamma((N-2)/2)}{\Gamma((N-1)/2)}$ and $\langle\sigma^2\rangle = \int_\Sigma \sigma^2 p(\sigma|\mathbf{D},\mathcal{I}))\mathrm{d}\sigma = \frac{Ns^2}{N-1}$. The latter result is the typical result which orthodox statistics uses to estimate the population variance from independent and identical distributed Gaussian distributions, see [3,8]. As pointed above, a credibility region, C, can be defined about $\hat{\sigma}$ — in general this region will be asymmetrical.

It is worthwhile pointing out, that it is tempting to take the above results and introduce them as non-uniform priors in to Eq. (14). This procedure results in $\overline{D}$, but reduces the variance. Repeating the procedure, results in $\sigma\rightarrow 0$, which is a nonsense result [10].

3.1.5. Using non-uniform priors

Suppose we have good reason to be believe that the prior pdf for μ can be defined by a Gaussian distribution located at μ_0 and with a standard deviation, δ_0, such that,

$$p(\mu|\mu_0,\delta_0,\mathcal{I}) = \frac{1}{\sqrt{2\pi\delta_0^2}}\exp\left[-\frac{1}{2}(\mu-\mu_0)^2/\delta_0^2\right]. \tag{27}$$

This Gaussian prior states that the true value with in a specified range i.e. ~95% probability in the interval $\mu_0\pm 2\delta_0$ — see Fig. 1(a). We also suppose that the prior for σ is expressed as [3],

$$p(\sigma|\sigma_0,\mathcal{I}) = \frac{\sigma_0}{\sigma^2}, \tag{28}$$

where $\sigma\in[\sigma_0, \infty)$ and zero otherwise — see Fig. 1(b). This prior assumes that the estimated value, σ_0, represents a lower-bound for the standard deviation and its true value is greater.

Bayes theorem for μ and σ can be expressed in terms of the prior values as,

$$p(\mu,\sigma|\mathbf{D},\mu_0,\sigma_0,\delta_0,\mathcal{I}) = \frac{p(\mu|\mu_0,\delta_0,\mathcal{I})p(\sigma|\sigma_0,\mathcal{I})p(\mathbf{D}|\mu,\sigma,\mathcal{I})}{p(\mathbf{D}|\mu_0,\sigma_0,\delta_0,\mathcal{I})}. \tag{29}$$

Following the same steps as for Eqs. (24) and (26), the posterior pdf for μ and σ can be expressed analytically and also evaluated using numerical integration techniques and/or MCMC methods.

In Eq. (29), the evidence or denominator term is conditional on μ_0 and σ_0, and can be used to evaluate the plausibility of the prior information,

$$p(\mathbf{D}|\mu_0,\sigma_0,\mathcal{I}) = \int_{-\infty}^{+\infty}\int_{\sigma_0}^{\infty} p(\mathbf{D},\mu,\sigma|\mu_0,\sigma_0,\mathcal{I})\mathrm{d}\sigma. \tag{30}$$

For different values of μ_0 and σ_0, the evidence will change. That is, Eq. (20) quantifies the predictability of the model parameter, μ_0 and σ_0. If there were different opinions for values of μ_0 and σ_0 described in Eqs. (27) and (28), they can also be tested, by evaluating Eq. (30).

In the above results the posterior pdf was derived analytically. However, in many cases this is not possible and approximation methods or numerical methods for sampling the posterior pdf are necessary. In the former case, approximation can be carried out in the Gaussian limit, while in the latter case, Markov Chain Monte Carlo (MCMC) methods can be applied — see [3–9]. The advantage of using MCMC methods for sampling the posterior pdf and evidence is particularly acute when using non-uniform priors, and when the number of parameters is large (>4). The advanced example in Section 6.1 outlines how both methods can be applied to experimental data. Also Appendix B summarizes the key aspects of applying MCMC methods in sampling the posterior pdf and evidence.

In summary, we have shown how, using both uniform and non-uniform priors, posterior pdfs can be derived. In the case of the uniform prior the results are similar to the orthodox statistical interpretation. However, the orthodox statistical approaches provide no scope for the introduction of priors or the estimation of the abovementioned quantities from a single set of measurements.

3.2. Example 1: a simple worked case

3.2.1. Outline of the problem

To demonstrate the application of Bayes theorem in determining the posterior pdf for μ and σ, consider the data given in Table 1 presented and analysed by Hibbert and Gooding [16], p. 95. The data represents calcium mass fractions, w_{Ca}, in milk powder measured using two different experimental techniques, namely atomic absorption spectroscopy (AAS) and complexometric titration (CT) by second year undergraduate students. The underlying problem of whether the two samples have been drawn from a population with the same mean and standard deviation or whether they have a distinct mean and standard deviation will be discussed in Section 4. For the moment, we determine the average and standard deviation for each sample, and for the combined data set.

Table 1
Measurement of the mass fraction of calcium, w_{Ca}, in a sample of milk powder from two different methods, atomic absorption spectroscopy (AAS), and complexometric titration (CT).

Experimental technique	Group performing measurement								
	1	2	3	4	5	6	7	8	9
w_{Ca} by AAS/mg g^{-1}	3.01	2.58	2.52	1.00	1.81	2.83	2.13	5.14	3.20
w_{Ca} by CT/mg g^{-1}	2.81	3.20	3.20	3.20	3.35	3.86	3.88	4.13	4.86

Data presented in Hibbert and Gooding [16].

Table 2
Measurement of the mass fraction of calcium, w_{Ca}, in a sample of milk powder.

Experimental technique	N	$[\mu_{min}, \mu_{max}]$/mg g^{-1}	$[\sigma_{min}, \sigma_{max}]$/mg g^{-1}
w_{Ca} by ASS/mg g^{-1}	9	[1.00, 5.14]	[0.01, 4.00]
w_{Ca} by CT/mg g^{-1}	9	[2.00, 5.00]	[0.01, 4.00]
Combined w_{Ca}/mg g^{-1}	18	[1.00, 5.14]	[0.01, 4.00]

A priori ranges used in the prior pdfs for μ and σ, respectively. N represents the number of data values in a data set.

Table 3
Average and standard deviation of mass fraction of calcium in milk powder, evaluated using the uniform priors, Eqs. (20) and (21), respectively.

Experimental technique	$\hat{\mu} \pm \delta_\mu$/ mg g^{-1}	$\hat{\sigma} \pm \delta_\sigma$/ mg g^{-1}	$\langle\sigma\rangle \pm \delta_{\langle\sigma\rangle}$/ mg g^{-1}	$\sqrt{\langle\sigma^2\rangle}$/ mg g^{-1}
w_{Ca} by ASS/mg g^{-1}	2.69 ± 0.87	$1.07^{+1.00}_{-0.33}$	$1.26^{+0.91}_{-0.49}$	1.31
w_{Ca} by CT/mg g^{-1}	3.61 ± 0.48	$0.70^{+0.51}_{-0.27}$	$0.69^{+0.56}_{-0.17}$	0.73
Combined w_{Ca}/mg g^{-1}	3.15 ± 0.50	$0.98^{+0.50}_{-0.24}$	$1.06^{+0.46}_{-0.27}$	1.08

The credibility regions correspond to the ~95%-region.

3.2.2. Applying uniform priors

For this particular example, we start by defining the prior information. Table 2 summarizes the prior information used in Eqs. (20) and (21) for the calculation of posterior density for μ and σ, given by Eqs. (23a), (23b) and (25a), (25b). In this case the prior terms have a large range which incorporates the range of measurement values. This in turn, reflects our lack of understanding of the values of the average and standard deviation. In Table 2, N defines the number of data for each sample. We assume that the experimental model is given by Eqs. (18a)–(18d), with Gaussian noise. This assumption results in a Gaussian likelihood function, Eq. (13). This is a common assumption, but if we are uncertain about the type of noise, the likelihood function can be modified to reflect this uncertainty — see Sivia and Skilling [3]. Moreover, there is no reason why Bayesian model selection principles cannot be applied to determine the appropriate likelihood function.

Fig. 2 shows the posterior distribution pdf for μ and σ. It is evident from the figure that the overlap between the posterior pdfs for the ASS and CT samples occurs in low probability regions. The 'combined' sample, attempts to straddle the overlapping region between the ASS and CT samples. Using the posterior pdfs, the average and standard deviation can be estimated for each sample. These results are given in Table 3. The symmetry in the posterior pdfs for μ, see Fig. 2(a), implies that the most probable and average μ correspond to the same quantity. The estimated uncertainty for μ correspond to the 95% credibility region. In case of the posterior pdfs for σ, see Fig. 2(b), the asymmetry results in different estimates for σ. The $\hat{\sigma}$, $\overline{\sigma}$ and $\sqrt{\langle\sigma^2\rangle}$-estimates are also given in Table 3, while the credibility region is asymmetric and also denotes the 95% credibility region. As demonstrated in Fig. 2, the results in Table 3, suggest that μ and σ are distinctly different. This simple example illustrates the how the posterior pdf for μ and σ can be defined given the available information and explicit assumptions concerning the ranges of the prior pdfs.

3.2.3. Applying non-uniform priors

Fig. 3 demonstrates the influence of the priors, Eqs. (27) and (28), respectively, for estimating μ and σ. In this calculation, values for the Gaussian prior, Eq. (27), were (2.34 ± 0.47) mg g^{-1} and the prior value for Eq. (28) with $\sigma_0 = 0.01$ mg g^{-1} was used. The first three, six and all nine data values for CT measurements in Table 3 were used in the likelihood function, Eqs. (18a)–(18d).

For a small number of data values, the posterior pdfs for both μ and σ are spread between the prior values and the available data. For example, in Fig. 3(a) for the case of three data values the posterior pdf for μ has extended left-tails or negative skewness and its maximum is

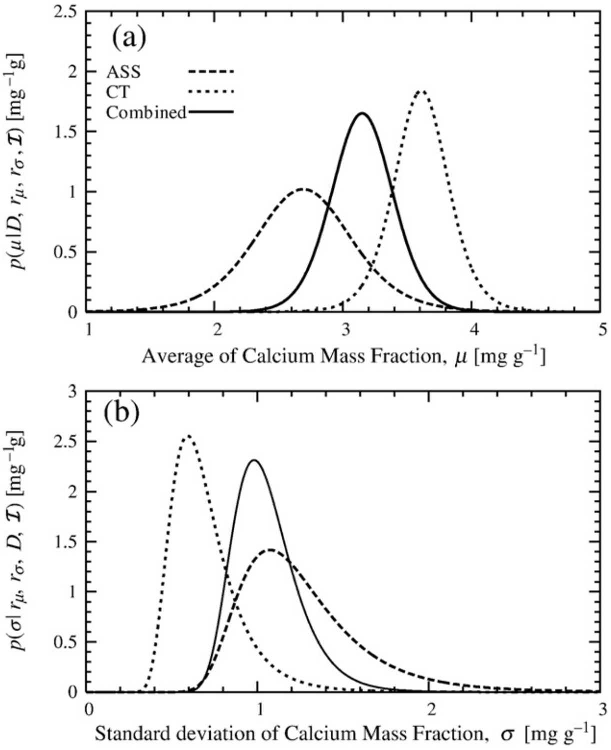


Fig. 2. Posterior pdfs for μ and σ using uniform priors, (20) and (21), respectively, for the ASS, CT and combined data sets. (a) Posterior pdfs for μ using Eq. (23) for ASS, CT and combined data sets. (b) Posterior pdfs for σ using Eq. (25) for ASS, CT and combined data sets. (All pdfs have been normalised for unit area.)

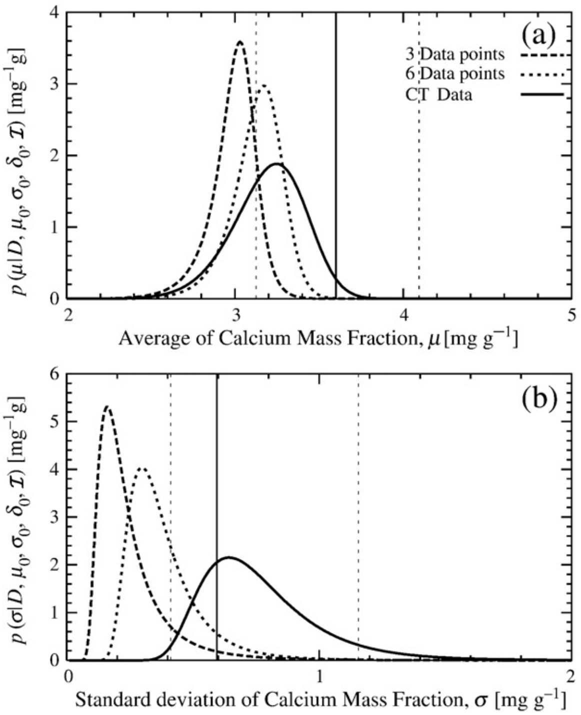


Fig. 3. Posterior pdfs for μ and σ by integrating out the corresponding variable in Eq. (29), respectively. (a) Posterior pdfs for μ using the non-uniform prior, Eq. (27), for first three, six and entire CT data from Table 3. (b) Similarly for σ using the non-uniform prior given by Eq. (28) using the first three, six and entire CT data. The solid and dashed vertical lines represent the $\hat{\sigma}$ and $\hat{\sigma}$, and 95% credibility region for the CT data given in Table 3. (All pdfs have been normalised for unit area.)

located approximately at 3.04 mg g^{-1}. As the number of measurements increase, 'the center of mass' between the prior and the likelihood functions shifts towards the likelihood, but is qualified by the uncertainty in the data. This accounts for the broad density function when all the data values are used, where the maximum located approximately at 3.14 mg g^{-1} — within the ~95% credibility region for $\hat{\mu}$ given in Table 3. Similar observations can be made for σ in Fig. 3(b).

4. Introduction to Bayesian model selection

4.1. Overview of model selection

The problem of model selection is paramount in any analysis of the scientific data. The task is often complicated by the quality of the experimental data and can be corrupted by statistical noise, as modeled in Eq. (13). The data presented in Table 1 for mass fraction of calcium in milk powder using two different experimental techniques is a good example to illustrate how Bayesian hypothesis testing can be applied. Without the use of a standard, whose mass fraction of calcium is known with a high degree of certainty, do the two experimental methods produce equivalent results? If the two techniques are equivalent, then their averages and standard deviations are the same [16]. In this example, we follow the approach of Sivia and Skilling [3], and Gregory [8]. It is also important to note that other cases could also be examined, such as whether the data sets have different averages, but same standard deviations etc. These cases are addressed by Gregory [8] (see pp. 228–241).

4.2. Defining and evaluating hypotheses

Let us denote the null hypothesis, $\mathcal{H}_o$, to state that measurements of equivalent samples by the two experimental techniques, ASS and CT, produce results with the same average, μ_o, and standard deviation, σ_o. The background information for $\mathcal{H}_o$, is defined as $\mathcal{I}_o=\mathcal{I}_1\cup\mathcal{I}_2$, where $\mathcal{I}_1$ and $\mathcal{I}_2$ are the background information for ASS and CT experimental techniques, respectively. In other words, in stating $\mathcal{H}_o$, we are also assuming that the experimental techniques are entirely equivalent, otherwise we would not expect the same results. The alternative hypothesis, $\mathcal{H}_1$, denotes that the two data sets have different means and standard deviations, μ_1, σ_1 and μ_2, σ_2 for ASS and CT, respectively. That is, μ_1, σ_1, μ_2 and σ_2 are not known or determined from the data, as shown below — see Eqs. (33a) and (33b). It also follows that $\mathcal{I}_1\neq\mathcal{I}_2$.

We start by gathering the probabilistic evidence for $\mathcal{H}_o$. This involves defining the joint-pdf for $\mathbf{D}$, μ_o, σ_o and integrating out the latter two parameters, given as,

$$p(\mathbf{D}|\mathcal{H}_o,\mathcal{I}_o)=\int\int p(\mathbf{D},\mu_o,\sigma_o|\mathcal{H}_o,\mathcal{I}_o)\mathrm{d}\mu_o\mathrm{d}\sigma_o \tag{31a}$$

$$=\int\int p(\mu_o|\mathcal{H}_o,\mathcal{I}_o)p(\sigma_o|\mathcal{H}_o,\mathcal{I}_o)p(\mathbf{D}|,\mu_o,\sigma_o,\mathcal{H}_o,\mathcal{I}_o)\mathrm{d}\mu_o\mathrm{d}\sigma_o, \tag{31b}$$

where we have assumed independence between the priors.

Evaluating Eqs. (31a) and (31b) involves a two-dimensional integration. The result can be approximated by expanding Eqs. (31a) and (31b) about the optimum values of μ_o and σ_o and extending the limits of integration, to produce (see [3]),

$$p(\mathbf{D}|\mathcal{H}_o,\mathcal{I}_o)\approx\frac{\left(\sigma_o\sqrt{2\pi}\right)^{2-N}e^{-N/2}}{\sqrt{2}N(\mu_{\max}-\mu_{\min})\sigma_{\max}}, \tag{32}$$

where in Eqs. (31a), (31b) and (32), $\mathbf{D}$ denotes the combined data set of ASS and CT.

Table 4
The values for the probabilistic evidence evaluated using (ie. "Numerical") numerical integration of Eqs. (31) and (33), respectively.

Method	$p(D\|\mathcal{H}_o,\mathcal{I})$	$p(D_1\|\mathcal{H}_1,\mathcal{I}_1)$	$p(D_2\|\mathcal{H}_1,\mathcal{I}_2)$	$p(D_1,D_2\|\mathcal{H}_1,\mathcal{I}_1,\mathcal{I}_2)$
"Numerical"	1.16×10^{-13}	3.57×10^{-8}	5.58×10^{-6}	1.99×10^{-13}
"Approx."	1.62×10^{-13}	5.16×10^{-8}	4.46×10^{-6}	2.30×10^{-13}

These results are compared with the approximations (ie. "Approx.") given by Eqs. (32) and (35) (see [3]).

In the case of $\mathcal{H}_1$, we have,

$$p(\mathbf{D}_1,\mathbf{D}_2|\mathcal{H}_1,\mathcal{I}_1,\mathcal{I}_2)=p(\mathbf{D}_1|\mathcal{H}_1,\mathcal{I}_1)p(\mathbf{D}_2|\mathcal{H}_1,\mathcal{I}_2) \tag{33a}$$

$$=\int\int\int\int p(\mathbf{D}_1,\mu_1,\sigma_1|\mathcal{H}_1,\mathcal{I}_1)p(\mathbf{D}_2,\mu_2,\sigma_2|\mathcal{H}_1,\mathcal{I}_2)\times\mathrm{d}\mu_1\mathrm{d}\sigma_1\mathrm{d}\mu_2\mathrm{d}\sigma_2 \tag{33b}$$

where,

$$p(\mathbf{D}_1,\mu_1,\sigma_1|\mathcal{H}_1,\mathcal{I}_2)=p(\mu_1|\mathcal{H}_1,\mathcal{I}_1)p(\sigma_1|\mathcal{H}_1,\mathcal{I}_1)p(\mathbf{D}|,\mu_1,\sigma_1,\mathcal{H}_1,\mathcal{I}_1), \tag{34}$$

and similarly for $p(\mathbf{D}_2,\mu_2,\sigma_2|\mathcal{H}_1,\mathcal{I}_2)$. In this case, the evaluation of Eqs. (33a) and (33b) involves a four-dimensional integration. The approximation established for Eq. (32) is also applicable for Eqs. (33a) and (33b) (also see [3]),

$$p(\mathbf{D}_1,\mathbf{D}_2|\mathcal{H}_1,\mathcal{I}_1,\mathcal{I}_2)\approx\frac{\left(\sigma_{o1}\sqrt{2\pi}\right)^{2-N_1}e^{-N_1/2}}{\sqrt{2}N_1(\mu_{\max1}-\mu_{\min1})\sigma_{\max1}}\times\frac{\left(\sigma_{o2}\sqrt{2\pi}\right)^{2-N_2}e^{-N_2/2}}{\sqrt{2}N_2(\mu_{\max2}-\mu_{\min2})\sigma_{\max2}}. \tag{35}$$

The prior pdfs for μ_o and σ_o etc are given by Eqs. (20) and (21), respectively, for the values given in Table 2. The corresponding the likelihood functions are assumed to be Gaussian, given by Eqs. (18a)–(18d). Using numerical integration,[2] the probabilistic evidences for $\mathcal{H}_o$ and $\mathcal{H}_1$ are given in Table 4. These values are compared with the approximations for the evidence in Eqs. (32) and (35).

It is interesting to note that the ratio of $p(\mathbf{D}|\mathcal{H}_o,\mathcal{I}_o)$ to $p(\mathbf{D}_1,\mathbf{D}_2|\mathcal{H}_1,\mathcal{I}_1,\mathcal{I}_2)$ is about ~0.6, which suggests that $\mathcal{H}_o$ is less probable — as demonstrated below. The same ratio of approximate values of the evidence is about ~0.7.

Using Bayes theorem, the probability for $\mathcal{H}_o$ can be expressed as,

$$p(\mathcal{H}_o|\mathbf{D},\mathcal{I}_o)=\frac{p(\mathcal{H}_o|\mathcal{I}_o)p(\mathbf{D}|\mathcal{H}_o,\mathcal{I}_o)}{Z} \tag{36}$$

and for $\mathcal{H}_1$,

$$p(\mathcal{H}_1|\mathbf{D}_1,\mathbf{D}_2,\mathcal{I}_1,\mathcal{I}_2)=\frac{p(\mathcal{H}_1|\mathcal{I}_1,\mathcal{I}_2)p(\mathbf{D}_1,\mathbf{D}_2|\mathcal{H}_1,\mathcal{I}_1,\mathcal{I}_2)}{Z}, \tag{37}$$

where the normalization term, Z, requires summing over the hypotheses,

$$Z=p(\mathcal{H}_o|\mathcal{I}_o)p(\mathbf{D}|\mathcal{H}_o,\mathcal{I}_o)+p(\mathcal{H}_1|\mathcal{I}_1,\mathcal{I}_2)p(\mathbf{D}_1,\mathbf{D}_2|\mathcal{H}_1,\mathcal{I}_1,\mathcal{I}_2). \tag{38}$$

It is reasonable to assume that each hypothesis are equally probable, so the prior probabilities for $\mathcal{H}_o$ and $\mathcal{H}_1$ in Eqs. (36) and (37) can be set to 0.5, respectively. Following from this the approximate calculations produces $Z\approx1.96\times10^{-13}$ and for the full numerical results, $Z\approx1.58\times10^{-13}$. It is also worth noting that Eq. (36) reverses the

[2] The numerical integration used NIntergate in *Mathematica* software package.

Table 5
Probabilities values for $\mathcal{H}_o$ and $\mathcal{H}_1$ for each method using the prior ranges in Table 2.

Method	$p(\mathcal{H}_o\|D,\mathcal{I})$	$p(\mathcal{H}_1\|D_1,D_2,\mathcal{I}_1,\mathcal{I}_2)$
“Numerical”	0.37	0.63
“Approx.”	0.41	0.59

statement given by Eqs. (31a) and (31b) to read how “what is the probability of $\mathcal{H}_o$ *given* the data **D**, and background information $\mathcal{I}_o$?”.

The final posterior probabilities for $\mathcal{H}_o$ and $\mathcal{H}_1$ are given in Table 5 for both the approximate and full numerical methods. Concentrating on the full results, $\mathcal{H}_o$ can be rejected, and the inference can be made that the experimental techniques are not equivalent. In addition, it is also worth noting that the ranges for the priors for μ and σ for all data sets are relatively large. This in turn, implies our lack of knowledge concerning the values for these parameters. As a result this lack of knowledge is reflected in the relatively large probability value of $\mathcal{H}_o$. The approximate results are better for $\mathcal{H}_o$, they provide some hint of which hypothesis is the most likely, but also suggest the need for full evaluation of the evidence.

It is interesting to investigate the influence of the prior ranges on the values of the probabilities. Suppose, for argument's sake, we had some good reason to reduce the ranges of the priors for the means to ±5% about $\hat{\mu}$, then for all data sets, the probabilities become more decisive, as given in Table 6. Our increased knowledge regarding the parameter values (i.e. decreased prior ranges for μ and σ) is reflected in the considerably greater differences between the probabilities of $\mathcal{H}_o$ and $\mathcal{H}_1$.

In Tables 5 and 6, we have demonstrated how the values of competing probabilities can be updated when new information, such as decreased ranges in the prior pdfs can be introduced via Bayes theorem.

5. Parameter estimation and curve fitting

In the physical sciences, a set of measurements, $\mathbf{D}=\{D_i;i=1,\ldots,N\}$, are usually recorded as a function of an independent variable, $\mathbf{x}=\{x_i;i=1,\ldots,N\}$. Using this data a model is fitted to the experimental data, where the parameters are determined and related to some underlying physical and/or chemical process. Curve fitting techniques are widely used and available, and commonly found in various plotting and analysis software packages. There are two cases a chemist will encounter. The first and typical case, is where the statistical uncertainty is assumed to exist in only the dependent variable, **D**. The analysis essentially involves determining the optimum parameter values and corresponding uncertainty from the data. This is particularly important if the data is suspected of having outliers. Recently, Sivia and Skilling [3] have provided an outline of the Bayesian treatment of outliers. The advantage of this approach is that outliers can be treated in a consistent manner, by manipulating the likelihood function to reflect our understanding of the problem.

The second case, assumes statistical uncertainty exists in both the independent variable, **x**, and the dependent variable, **D**. This example has been discussed by Gregory [8], Gull [17] and Jaynes [18]. Again, in order to understand this example, the likelihood function must be manipulated to reflect the available experimental data and model.

In this section, we provide two examples. The first example, demonstrates how a Bayesian approach can be used to determine the posterior pdfs for the parameters of a non-linear calibration equation for the electrochemical detection of lead ions [19]. The second example is a special case of the first example in which the uncertainties exist in both coordinates for the linear approximation of the Langmuir equation. This example follows the discussion of Gregory [8] (see pp. 92–93).

Table 6
Probabilities values for $\mathcal{H}_o$ and $\mathcal{H}_1$ for each method assuming the prior range for the means are ±5% about their optimum values for $\hat{\mu}$ given in Table 3.

Method	$p(\mathcal{H}_o\|D,\mathcal{I})$	$p(\mathcal{H}_1\|D_1,D_2,\mathcal{I}_1,\mathcal{I}_2)$
“Numerical”	0.16	0.84
“Approx.”	0.05	0.95

Table 7
Measurement of current density as a function of lead [Pb$^+$] concentration.

c [Pb^{2+}]/nM	1.9	5.3	11.0	16.0	24.0	37.0	45.0	92.0	232.0	458.0
I/μA cm^{-2}	0.11	0.40	1.10	1.40	2.10	2.70	3.20	4.30	6.30	7.20

Data presented in Chow et al. [19].

5.1. Example 3: determining the concentration of lead

The presence of heavy metals, such as lead and cadmium, in drinking water or when released into the environment can have serious health and environmental consequences. The application of biosensors to detect and monitor low concentrations of such metals is an important development [19]. This example is drawn from Chow et al. [19], where a electrochemical biosensor consisting of a gold substrate modified with a self assembled monolayer (SAM) of a peptide is used to measure the concentration of lead ions. The device is calibrated by measuring the current density, I (units, μA cm^{-2}) for reduction of the ions accumulated from solution as a function of concentration, c (units, nM^{-1}). Experimental data collected by [19] are presented in Table 7. For these data the precision of the current density is unknown, but for the purpose of this example we assume that it is constant for values of I. Furthermore, we will assume that a Gaussian the likelihood function accounts for the precision, such that $\langle e\rangle=0$ with a standard deviation, $\langle e^2\rangle=\sigma^2$.

The current density for a given solution concentration can be analysed using a rectangular hyperbolic equation,

$$I_i = I_{\mathrm{calc}\,i} + e_i \tag{39a}$$

$$= \frac{I_\infty K c_i}{1+Kc_i} + e_i, \quad \forall i = 1,\ldots,N \tag{39b}$$

where unknown parameter values are I_∞, the saturation current density (units, μA cm^{-2}); and K, the equilibrium constant for the metal binding to peptide (units, nM^{-1}). In the following analysis we treat the data set as one dimensional arrays, with $\mathbf{I}=\{I_i;i=1,\ldots,N\}$ and $\mathbf{c}=\{c_i;i=1,\ldots,N\}$.

The corresponding joint posterior pdf for I_∞ and K is,

$$p(I_\infty,K,\sigma|\mathbf{I},\mathcal{I}) = Z^{-1}\,p\left(I_\infty|r_{\mathcal{I}_\infty},\mathcal{I}\right)p(K|r_K,\mathcal{I})\,p(\sigma|r_\sigma,\mathcal{I})\,p(\mathbf{I}|I_\infty,K,\sigma,\mathcal{I}). \tag{40}$$

The prior pdfs for I_∞ and K are defined by uniform distributions with the form,

$$p\left(I_\infty|r_{I_\infty},\mathcal{I}\right) = \frac{1}{I_{\infty\max}-I_{\infty\min}}, \tag{41}$$

and similarly for $p(K|r_K,\mathcal{I})$. The ranges for I_∞ and K used in the present example were [5.00, 13.00] μA cm^{-2} and [0.008, 0.018] mM^{-1}, respectively. The prior pdf for σ is defined as a Jeffery's prior,

$$p(\sigma|r_\sigma,\mathcal{I}) = \frac{1}{\ln(\sigma_{\max}/\sigma_{\min})\sigma}, \tag{42}$$

and its ranges were specified by $r_\sigma\in[10^{-3},1]$ nM. The corresponding likelihood function was defined as,

$$p(\mathbf{I}|I_\infty,K,\sigma,\mathcal{I}) = \left(2\pi\sigma^2\right)^{-N/2}\exp\left[-\frac{1}{2}\sum_i^N\left(I_i-I_{\mathrm{calc}i}\right)^2/\sigma^2\right] \tag{43a}$$

$$= \left(2\pi\sigma^2\right)^{-N/2}\exp\left[-\frac{1}{2}Q(I_\infty,K)/\sigma^2\right], \tag{43b}$$

where σ has been decoupled, and Q in Eq. (43b) is the function of I_∞ and K, which can be optimised using a nonlinear optimisation

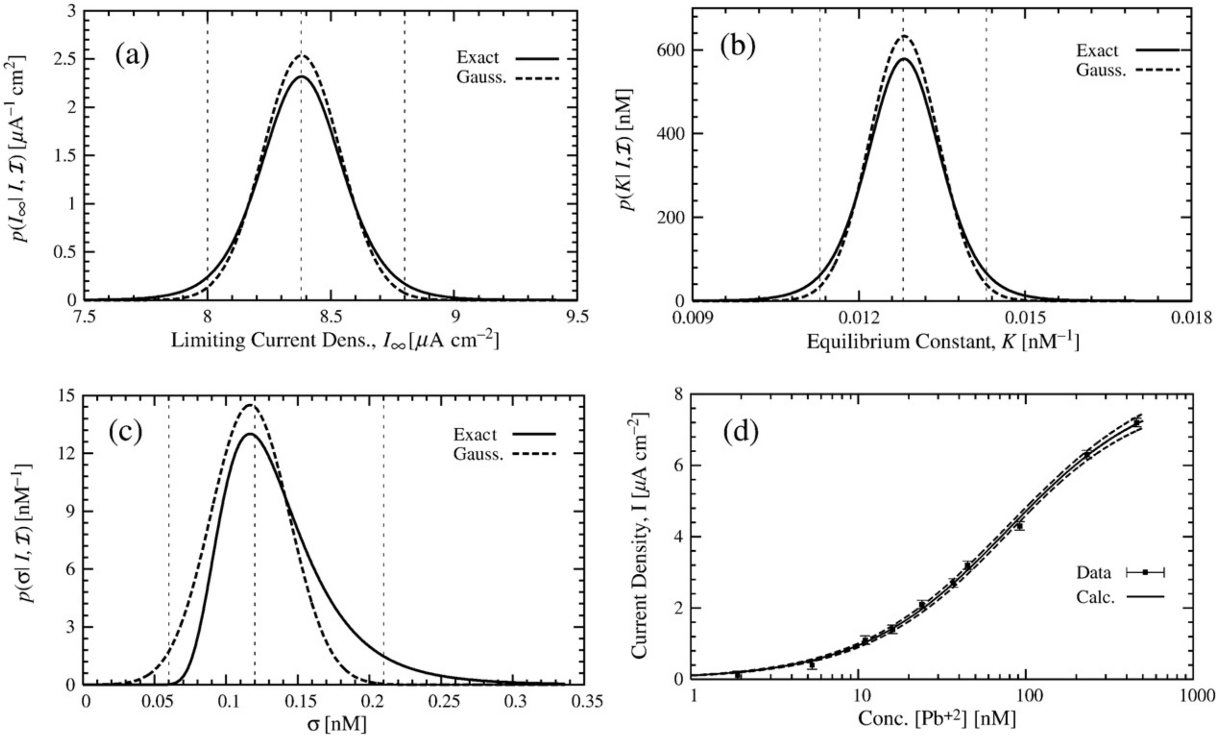


Fig. 4. Posterior pdfs for I_∞, K, σ, and the experimental and calculated I as a function of concentration, c. (a–c) Posterior pdfs for I_∞, K, σ, respectively, conditional on the experimental data, I, and the corresponding Gaussian approximations. The vertical lines correspond to the lower-, optimum- and upper-values for the parameters of the 'exact posterior pdfs'. (d) Experimental I with calculated uncertainty-bars, σ, from (c), and the calculated I with lower- and upper- values for the ~95% credibility region, respectively. (All pdfs have been normalised for unit area.)

algorithm.[3] The marginalisation rule can be applied to the joint-pdf, Eq. (40), to determine the posterior pdfs for the individual for I_∞, K and σ.

The posterior pdfs for each of the quantities are plotted in Fig. 4(a–c), while Fig. 4(d) shows the experimental data (see Table 7) with the Bayesian estimate of the standard deviation determined from Fig. 4(c), and calculated I curve. Fig. 4(a–c) demonstrate the differences between the 'exact' and Gaussian approximation. The main difference can be seen in the tails of the posterior pdfs, which are broader than the Gaussian pdfs, and consequently the amplitudes are also lower relative to the Gaussian approximation. The optimum estimate values for I_∞, K and σ are given in Table 8. The estimated uncertainties correspond to ~95% credibility region.

By determining the posterior pdf for each parameter, we have demonstrated the spread parameter values and skewness in terms of the density function and its shape — for example see Fig. 4(c). That is, all necessary 'observables' can evaluated from the posterior pdf.

5.2. Examples 4: errors in both coordinates

In order to demonstrate the general case of uncertainties in both the independent and dependent variables, we use simulated data modelled on Eqs. (39a) and (39b) in the linear region or for low concentrations, c. This assumes that the linear model is 'correct' and the uncertainties in both variables are known prior. The linear approximation of Eqs. (39a) and (39b) with uncertainties in both coordinates is,

$$I_i \approx I_\infty K(c_i + e_{ci}) + e_{Ii}, \tag{44a}$$

$$= m(c_i + e_{ci}) + e_{Ii}, \quad \forall i = 1, \ldots, N \tag{44b}$$

where m ≡ $I_\infty K$ (units, μA cm^{-2} nM^{-1}). We also assume that the statistical uncertainties for independent and dependent variables are drawn from Gaussian distributions, such that $\langle e_{ci} \rangle = 0$ and $\langle e_{ci}^2 \rangle = \sigma_{ci}^2$, and similarly for e_{Ii}.

Following the discussion of Gregory [8] (see Section 4.8, pp. 89–93), the likelihood function for the data, I can be expressed as,

$$p(\mathbf{I}|\mathbf{c}, m, \{\sigma_{ci}\}, \{\sigma_{Ii}\}, \mathcal{I}) = (2\pi)^{-N/2} \left(\prod_{i=1}^{N} \left(\sigma_{Ii}^2 + m^2 \sigma_{ci}^2 \right)^{-1/2} \right) \times \exp\left[-\sum_{i=1}^{N} \frac{(I_i - mc_i)^2}{2(\sigma_{Ii}^2 + m^2 \sigma_{ci}^2)} \right], \tag{45}$$

and the posterior pdf for m becomes,

$$p(m|\mathbf{I}, \mathbf{c}, \{\sigma_{ci}\}, \{\sigma_{Ii}\}, r_m, \mathcal{I}) = Z^{-1} p(m|r_m, \mathcal{I}) p\left(\mathbf{I}|m, \left\{\sigma_{ci}^2\right\}, \left\{\sigma_{Ii}^2\right\}, \mathcal{I}\right), \tag{46}$$

where a uniform prior pdf was defined for m. For the simulated I data, the slope in Eq. (44b) was set to $m = 0.10$ μA cm^{-2} nM^{-1}. The standard deviations for the statistical uncertainties in c and I were defined by $\sigma_{ci} = \beta_c c_i$ and $\sigma_{Ii} = \beta_I I_i$, with $\beta_c = 0.1$ and $\beta_I = 0.07$, for all i.

Fig. 5(a) shows the Bayesian predictions for m assuming uncertainties in both independent and dependent variables ('Calc. 1') and in only the dependent variable ('Calc. 2'). Fig. 5(b) compares the lines of best fit for both cases with the simulated data. The most probable values for Calc. 1 (i.e. uncertainties in both c and I) was found to be

Table 8
Bayesian estimates for I_∞, K and σ taken from Fig. 4.

I_∞	(8.4 ± 0.4) μA cm^{-2}
K	(0.0128 ± 0.0015) nM^{-1}
σ	$0.12^{+0.09}_{-0.06}$ nM

The estimated uncertainties correspond to ~95% credibility region.

[3] In the present calculations the *Mathematica* command NMinimize was used to optimise Q.

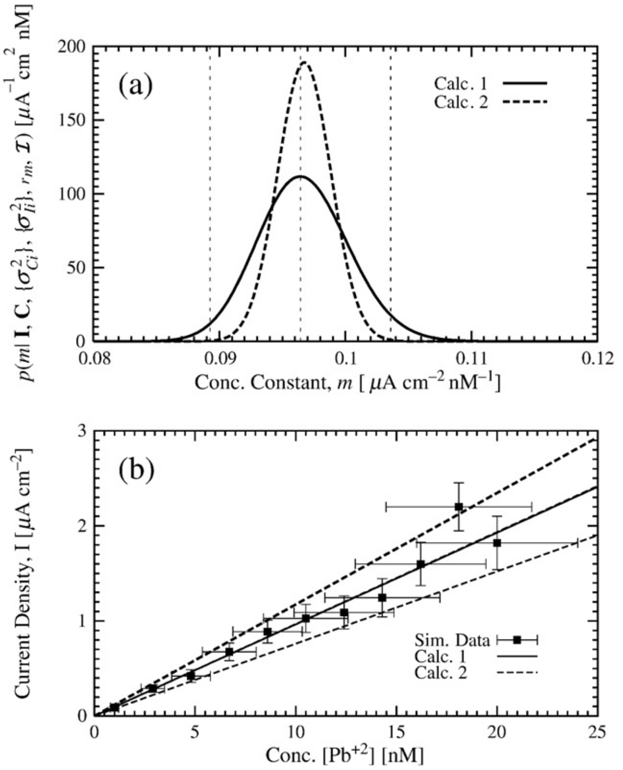


Fig. 5. Bayesian predictions for the concentration constant, m: (a) Posterior pdf for m with uncertainties in both coordinated (Calc. 1) and for uncertainties only in I-values (Calc. 2).The vertical line represents the ~95% credibility region for Cal. 1. (b) Simulated current-concentration data in the linear region, and the line of best fit for both Calc. 1 and 2. The uncertainty-bars represent $\pm 2\sigma$ and the lower- and upper- trend line representing the 95% credibility region for Calc. 1.

$(0.096^{+0.007}_{-0.008})$ μA cm^{-2} nM^{-1} and for Calc. 2 was found to be (0.097 ± 0.004) μA cm^{-2} nM^{-1}, where the uncertainty-bars represent the ~95% credibility region.

The posterior pdf for m in Fig. 5(a) illustrates the difference between Bayesian with and without uncertainties in both coordinates. When the uncertainties in c are taken into consideration, the resulting posterior pdf is broader compared to the posterior with uncertainties only in I. That is, both Calc. 1 and 2 predict the same values of m, but the uncertainty for m in Calc. 1 encompasses the true value of 0.1 μA cm^{-2} nM^{-1} in the ~95% credibility region. In addition, Eq. (46) suggests that asymmetry will occur in the posterior distribution. This asymmetry is present in Fig. 5(a) for Calc. 1, although it is small and is reflected in the uncertainty for m — see above. The asymmetry in Eq. (46) increases as the uncertainties in c increase.

The present discussion for uncertainties in both coordinates, assumes that the underlying model is linear. This is also demonstrated to some extent in the Bayesian literature — for example see [8,17,18]. However, when this is applied to a nonlinear model, such as Eqs. (39a) and (39b), difficulties are encountered in the producing a analytical result from the convolution of the likelihood functions for independent and dependent variables (see [8] Section 4.8, pp. 89–93). In principle this could be resolved numerically. Essentially, it involves evaluating the convolution equation numerically, and sampling the resulting likelihood function using Markov Chain Monte Carlo methods (as outlined in Gregory [8] and Sivia and Skilling [3]) with an underlying assumption that the model is monotonic and differentiable function for over the range of the independent variable. Finally, in deriving the posterior pdf for m, Eq. (46), a uniform prior was used. However, there is no reason why the priors outlined by Sivia and Skilling [3] (see chapter 8) could not be applied with sampling of the posterior pdf by Markov Chain Monte Carlo methods.

6. Peak fitting and deconvolution

6.1. Example 5: Bayesian peak fitting and model selection XPS data

X-ray photoelectron spectroscopy (XPS) is a surface analytical technique in which core electrons in atoms are excited by X-ray photons. Excited electrons are emitted with characteristic energies that allows identification of the element, and from second order effects (chemical shifts) allows determination of the bonding state of the atom. An XPS spectrum is a plot of electron current against electron energy, and the area under a peak is proportional to the fraction of that atom on the surface studied. When atoms of a given kind on a surface have different bonding (e.g. C might be C–C, C–O, C=O) the signals from a particular energy level are convolutions of overlapped Gaussian peaks that correspond to emission from the different species. The objectives of a spectral analysis are to determine the number of independent species, their distributions of electron energies and therefore their relative coverages on the surface.

Bayesian theory can be applied to determine the posterior density for the number of peaks in the spectroscopy data set. The present example draws on the discussion by Sivia and Carlyle [20] and Siva and Skilling [3] (see Section 4.2). A Bayesian/Markov Chain Monte Carlo (MCMC) approach, based on a Metropolis–Hastings algorithm, was applied to determine the posterior pdf for the number of peaks — see [1,3,8–10] for outline of algorithm and applications. In addition, Bayesian theory was also used to estimate the parameter values and their corresponding uncertainties. There are two levels in which Bayesian theory can be applied, the first is the parameter estimation, and the second is model selection by determining the number of peaks which appropriately describe the data. The Bayesian results are also compared with orthodox statistical methods for model selection. These are typically hierarchical approaches and rely on estimating the likelihood function and include the Akaike Information Criteria (AIC)[4], Schwarz Information Criteria (SIC)[5], and Deviance Information Criteria (DIC)[6] — also see [9,21].

In the present case we express the data, $D(x)$, in terms a set of N-Gaussian peaks superimposed on a background level and statistical noise,

$$D(x) = G_N(x) + b(x) + n(x) \tag{47a}$$

$$= \sum_{j=1}^{N} A_j \exp\left[-\frac{1}{2}\left(x - x_{oj}\right)^2 / \omega_j^2\right] + b(x) + n(x), \tag{47b}$$

where x represents the binding energy, $\{A_j\}$, $\{x_{oj}\}$, and $\{\omega_{oj}\}$ are the set of parameters that describe the Gaussian peak: amplitude, position and full-width at half-maximum, respectively; $b(x)$ is background level as a result of diffuse scattering events, and $n(x)$ is statistical noise imparted onto the recorded spectrum as a result of counting uncertainties in the measurement process. The recorded counts, $D(x)$ and corresponding binding energy, x, are treated as a vector of M data points, such that $\mathbf{D} = \{D_i;\ \forall i = 1,2,3,\ldots,M\}$ and $\mathbf{x} = \{x_i;\ \forall,\ i\}$, respectively. The standard deviation for the data is assumed to be near-Gaussian, since the counts are large and the Poisson distribution for large counts ($\gg 10$) can be approximate as Gaussian,

$$\sigma_i = \beta\sqrt{D_i + b_i} \tag{48a}$$

$$= \beta\sigma_{0i}, \forall i, \tag{48b}$$

where $\beta \in [0, \infty)$ is the proportionality constant for the uncertainty in the data. In other words, we are not entirely sure if uncertainty in the

[4] See URL:http://en.wikipedia.org/wiki/Akaike_information_criterion.

[5] See URL:http://en.wikipedia.org/wiki/Bayesian_information_criterion.

[6] See URL:http://en.wikipedia.org/wiki/Bayesian_information_criterion.

counting statistics follows the square-root of the counts. The recorded counts could have been amplified or modified by the electronics of the detector. We treat β as a Bayesian variable and determine its corresponding posterior pdf. In Eq. (48a) the uncertainties in estimating the background level also need to be taken into account. The background signal is assumed to be linear and given by, $b(x_i)=b_0+b_1x_1$, $\forall i$. In Eqs. (47a) and (47b), we have not taken into account an instrumental effect, and Eqs. (47a) and (47b) should be treated as a convolution equation. However, since we are only interested in the number of peaks this task is not necessary. If, on the other hand, it was necessary to relate the broadening of the peaks to some underlying microstructural property, Bayesian deconvolution would be applicable. A Bayesian and maximum entropy approach for solving this problems is outlined in Section 6.2.

Using Bayes' theorem, the joint-posterior pdf for $\{A_j\}$, $\{x_{oj}\}$, $\{\omega_{oj}\}$, $\{b_o, b_1\}$, β and N, can be expressed [3,20]. For convenience, we express the parameters for the peaks by,

$$\mathbf{X} = \{A_1, \dots, A_N, x_{o1}, \dots, x_{oN}, \omega_{o1}, \dots, \omega_{oN}, b_o, b_1\}.$$

For a model with N-Gaussian peaks, there are $\mathcal{P} = 3N + 2$ parameters, where the additional two parameters correspond to coefficients b_o, b_1 for the linear background level. Bayes theorem is given by,

$$p(\mathbf{X}, \beta, N|\mathbf{D}, \mathcal{I}) = p(\mathbf{X}|\mathcal{I})p(\beta|\mathcal{I})p(N|\mathcal{I})\frac{p(\mathbf{D}|\mathbf{X}, \beta, N, \mathcal{I})}{p(\mathbf{D}|\mathcal{I})}. \tag{49}$$

The prior pdf, $p(\mathbf{X}|\mathcal{I})$, for each parameter can be expressed as a product of uniform pdfs, defined over a pre-defined range,

$$p(\mathbf{X}|\mathcal{I}) = \prod_{j=1}^{P} \frac{1}{X_{\max j} - X_{\min j}} \tag{50a}$$

$$\approx \frac{1}{[A_{\max}(x_{\max} - x_{\min})(\omega_{\max} - \omega_{\min})]^N} \times \frac{1}{(b_{o\max} - b_{o\min})(b_{1\max} - b_{1\min})}. \tag{50b}$$

The pre-defined range, $\{X_{\min j}, X_{\max j}\}$ in Eq. (50a) reflects an underlying understanding of the approximate range for the parameters. The approximation, Eq. (50b), corresponds to the suggestion of Sivia and Skilling [3]. The prior for N, $p(N|\mathcal{I})$, can also be expressed as a uniform pdf over a range of number of peaks. On the other hand, if there were some theory specifying a greater probability to particular number peaks, there is no reason why this proposition cannot be tested. In the case of β, a Jeffery's prior, Eq. (21), can be defined.

The likelihood function in Eq. (49) is treated as Gaussian pdf and includes the N-peaks and background level,

$$p(\mathbf{D}|\mathbf{X}, \beta, N, \mathcal{I}) = \left[\prod_{i=1}^{M}\left(2\pi\sigma_i^2\right)^{-\frac{1}{2}}\right]\exp\left[-\frac{1}{2}\sum_{i=1}^{M}(D_i - F_i)^2/\sigma_i^2\right] \tag{51a}$$

$$= \left[\prod_{i=1}^{M}\left(2\pi\sigma_{0i}^2\right)^{-\frac{1}{2}}\right]\beta^{-M}\exp\left[-\frac{1}{2}Q(\mathbf{X})/\beta^2\right], \tag{51b}$$

where we have used Eqs. (48a) and (48b) in going from Eq. (51a) to Eq. (51b). In Eq. (51b), we have,

$$Q(\mathbf{X}) = \sum_{i=1}^{M}(D_i - F_i)^2/\sigma_{0i}^2, \tag{51c}$$

where $F_i = (x_i, \mathbf{X})$ such that,

$$F_i = \sum_{j=1}^{N} A_j \exp\left[-\frac{1}{2}\left(x_i - x_{oj}\right)^2/\omega_j^2\right] + b_0 + x_i b_1, \forall i. \tag{51d}$$

In Eq. (51b), the optimisation function $Q(\mathbf{X})$ has been expanded in terms of a Taylor series about the optimum parameter values, $\hat{\mathbf{X}}$. Using the rules of marginalisation, the posterior density for N can be determined in the Gaussian limit as [3,20],

$$p(N|\mathbf{D}, \mathcal{I}) = \int_0^\infty \int_{X\in\mathbb{R}^{\mathcal{P}}} p(\mathbf{X}, \beta, N|\mathbf{D}, \mathcal{I})\mathrm{d}^{\mathcal{P}}\mathbf{X}\,\mathrm{d}\beta \tag{52a}$$

$$\approx p(\mathbf{X}|\mathcal{I})p(N|\mathcal{I})N!\left[\prod_{i=1}^{M}\left(2\pi\sigma_{oi}^2\right)^{-\frac{1}{2}}\right] \times \int_0^\infty \int_{\mathbf{X}\in\mathbb{R}^{\mathcal{P}}} \beta^{-(M+1)} \mathrm{e}^{-\frac{1}{2}Q(\hat{\mathbf{X}})/\beta^2} \mathrm{e}^{-\frac{1}{2}(\mathbf{X}-\hat{\mathbf{X}})^{\mathrm{T}}\frac{\nabla\nabla Q}{2\beta^2}(\mathbf{X}-\hat{\mathbf{X}})}\mathrm{d}^{\mathcal{P}}\mathbf{X}\,\mathrm{d}\beta \tag{52b}$$

$$\approx p(\mathbf{X}|\mathcal{I})p(N|\mathcal{I})N!\left[\prod_{i=1}^{M}\left(2\pi\sigma_{oi}^2\right)^{-\frac{1}{2}}\right]2^{\frac{1}{2}(M+\mathcal{P}-2)}\pi^{\frac{\mathcal{P}}{2}}\Gamma\left(\frac{M-\mathcal{P}}{2}\right) \times \nabla\nabla Q^{-\frac{1}{2}}Q(\hat{\mathbf{X}})^{-\frac{1}{2}(M-\mathcal{P})}, \tag{52c}$$

where $M > P$, $\nabla\nabla Q$ is evaluated at $\hat{\mathbf{X}}$, $N!$ corresponds to the number of ways N-peaks can be arranged [3], and $\Gamma(\cdot)$ is a Gamma function (see Appendix A). In going from Eq. (52a) to Eq. (52b), requires expanding the likelihood function about $\hat{\mathbf{X}}$ i.e. in the Gaussian limit. The integration over β is a Gamma density function and the final approximation can evaluated in a closed form. This result differs from Sivia and Skilling [3] in that it includes the width of the peaks and background level. Eq. (52a) can be evaluated using MCMC method, while Eq. (52c) can be evaluated relatively quickly and provides an indication of the most probable model.

MCMC methods have become an important tool in the application of Bayesian theory, especially for sampling posterior pdfs with non-uniform priors and where the parameter space is large. Eq. (52a) is a good example as the parameter space is large (>4). Appendix B

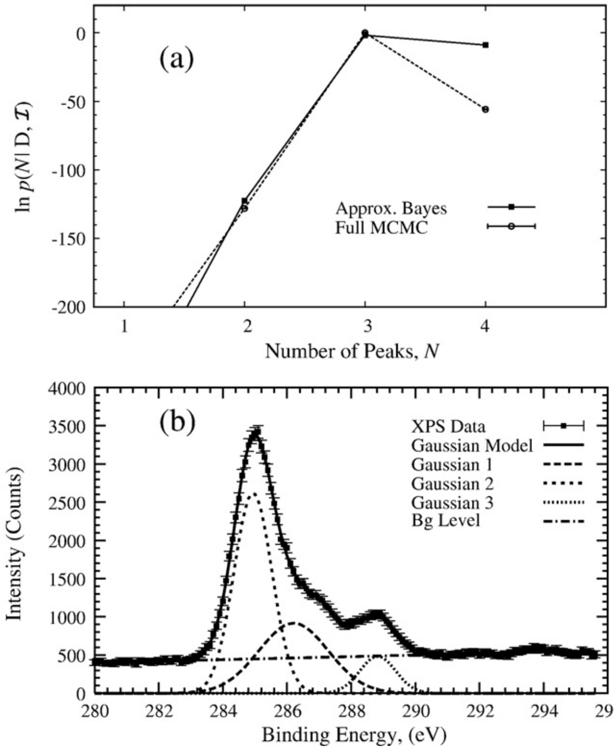


Fig. 6. Bayesian spectral analysis. (a) Bayesian estimate of the number peaks in the data. The uncertainties for the MCMC results correspond to $\pm 2\sigma$. (b) XPS data with uncertainty-bars estimated from the Bayesian analysis (see Eq. (49)), Gaussian peaks and estimated background level. Uncertainty-bars represent $\pm 2\sigma$.

Table 9
Gaussian parameters for the most probable number of peaks, $N=3$.

Peak	(a) A_j (counts)	x_{oj} (eV)	w_j (eV)	(b) A_j (counts)	x_{oj} (eV)	w_j (eV)
1	2619 ± 256	284.94 ± 0.02	0.59 ± 0.02	2619^{+109}_{-114}	284.938 ± 0.009	0.594 ± 0.009
2	876 ± 58	286.4 ± 0.2	1.0 ± 0.2	876 ± 22	286.34 ± 0.11	1.00 ± 0.10
3	483 ± 42	288.83 ± 0.06	0.56 ± 0.04	483^{+15}_{-16}	288.83 ± 0.02	0.560 ± 0.013

(a) and (b) Parameters estimated from approximate and Monte Carlo Markov Chain method (MCMC), respectively. The uncertainty-bars correspond to the ~95% credibility region.

outlines the Metropolis–Hastings algorithm used in the paper and the necessary background information for analysing the MCMC samples.

The parameter ranges for the priors in MCMC method were set to be $\hat{\mathbf{X}}\pm2\sigma_{\hat{\mathbf{X}}}$, about the parameter's least-square estimate, $\hat{\mathbf{X}}$. In addition, the least-square estimates were also used as the starting values for the MCMC method. A total of 60×10^4 samples were used, with a 'burn-in' of 20% — see Appendix B. The autocorrelation function was used to determine the lag-length and the MCMC sample for each parameter was re-sampled using the optimal lag-length. Using the re-sampled MCMC data, the parameter's posterior distribution, and evidence were determined. In addition, the uncertainty for the evidence and corresponding posterior pdf for the number of peaks was also estimated — see Fig. 6(a). The evidence for each model was evaluated using the reverse importance sampling approach (see O'Ruanaidh and Fitzgerald [9], Appendix E). This approach has the advantage of using the MCMC sample to estimate the evidence from which the probabilities for the number of peaks were determined. The posterior distribution for each parameter, except β, were near-Gaussian pdfs with the $\pm2\sigma$ region corresponding to the ~95% credibility region. This was determined from the inverse cumulative function for each parameter at the 0.25 and 0.975 percentiles, respectively — see Table 9.

In Fig. 6(a) the approximate Bayesian approach demonstrates that there is moderate evidence for $N=3$ Gaussian peaks, while the MCMC method clearly indicates strong evidence for $N=3$ Gaussian peaks. For the latter, the proposition that there are greater than three peaks has little evidence. These calculations are dependent on the assumption of a linear background level, and the uncertainty in the XPS data being described by Gaussian pdf. Both of these assumption are valid, in that the binding energy range is sufficiently narrow that a linear background level is appropriate, and that large counting statistics can be approximated by Gaussian the likelihood function. The moderate evidence given by approximate Bayesian approach suggest that a additional analysis is necessary to confirm the result. In addition, the approximate Bayesian approach makes broad assumptions regarding the priors, see Eq. (50b), while the MCMC method does not make such assumptions, see Eq. (50a).

Fig. 6(b) compares the $N=3$ Gaussian model with the experimental XPS data. In addition, the uncertainty-bars were determined using Eqs. (48a) and (48b) in the approximate Bayesian analysis, resulting in $\beta=0.68\pm0.12$ for the ~95% credibility region, compared with $0.685^{+0.081}_{-0.071}$ using MCMC methods for the same credibility region. It can be seen from Fig. 6(b) that the Gaussian model fits the experimental data very well. Lorentzian peaks were also tested and resulted in very low probabilities for a given N, suggesting that Gaussian peaks are an appropriate choice. A full model selection outlined in Sivia and Skilling [3] could be used to test other peak profile functions, such as Pearson VII, pseudo-Lorentzian and Voigt profiles.

The parameters for the Gaussian model, Eqs. (47a) and (47b) are given in Table (9). The background coefficients were found to be $b_o=(-2363\pm248)$ counts and $b_1=(9.87\pm0.86)$ counts/eV for the ~95% credibility region for the approximate Bayesian analysis. This approach assumes the posterior distribution for the parameters is symmetrical. The MCMC analysis values are also given in Table (9). The posterior distributions from the MCMC method consisted of both symmetric and asymmetric pdfs, and the uncertainties values for the parameters are quoted accordingly i.e. $\hat{\mathbf{X}}\pm\ 2\sigma_{\hat{\mathbf{X}}}$ and $\hat{\mathbf{X}}^{+\delta_2}_{-\delta_1}$, respectively. For example, the background parameters had symmetrical posterior distributions and estimated values are given by $b_o=(-2373\pm197)$ counts and $b_1=(9.91\pm0.69)$ counts/eV for the ~95% credibility region.

Fig. 7 shows the results for most probable number of peaks using the AIC, BIC and DIC methods. In this particular case, the minimum values of these hierarchical methods indicates most probable model. There is no decisive selection between models with $N=3$ and $N=4$ peaks. The AIC, SIC and DIC results for these cases are very similar and there is only $\lesssim1\%$ difference between the results. The SIC selects $N=3$ peaks, while AIC and DIC marginally selects $N=4$ peaks as the most probable model. The results from using these methods are essentially indecisive.

The application of Bayesian and MCMC methods have provided a decisive result concerning the number of peaks in the XPS data, while orthodox methods have provided indecisive conclusions. The inference that can be made from the Bayesian results are that theC–C, O–C, O=C–O components are present in the XPS spectrum.

6.2. Bayesian instrumental deconvolution — an outline

In the previous example, our spectral peak analysis did not take into account any instrumental effects. The analysis was only concerned with determining the number of overlapped peaks in the XPS data and does not require a deconvolution.

However, there are many cases in chemistry and physics where data is recorded as intensity spectra or patterns, such as X-ray, electron and neutron diffraction, mass spectroscopy, Rutherford backscattering, and liquid and gas chromatography. Intensity spectra are corrupted by a number of sources, such as background level, finite instrument resolution and counting uncertainty. It is often desirable to remove these sources of data corruption and determine the intrinsic or underlying specimen profile. The shape and broadening of the specimen profile is a characteristic of underlying physical and/or chemical property/process. For example in the case of X-ray diffraction from crystalline materials, the shape and broadening of the specimen profile is related to the crystallite size ($\lesssim$100 nm), size distribution and density of lattice defects [22,23]. In the case of high

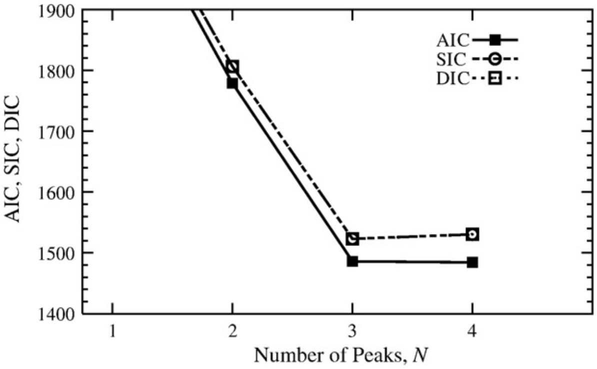


Fig. 7. Using orthodox methods for model selection to determine the most probable number of peaks in the XPS data, including the AIC, SIC and DIC methods. The smallest value of each of these criteria produces the most probable model.

performance liquid chromatography, the shape and broadening of the spectra is directly related to the diffusion processes [21]. The contributions can be defined in a integral equation [3,10,22–26],

$$D(x) = \int_{\Xi} R(x,\xi) f(\xi) d\xi + b(x) + n(x), \tag{53}$$

where $D(x)$ is the recorded data; $R(x)$ is the instrument response function or kernel; $b(x)$ is the background level and $n(x)$ is the imparted noise. In Eq. (53), we have assumed the general case of R being a shift-variant function over $\Xi \subseteq \mathbb{R}$, while in the shift-invariant case $R(x,\ \xi) = R(x-\xi)$ only depends on the difference, $x-\xi$. The kernel, $R(x)$, can usually be modelled or determined from the use of a reference material (see [22,23]). The integral equation, Eq. (53) can be expressed in matrix–vector notation as [3,10,22–26],

$$\mathbf{D} = \mathbf{Rf} + \mathbf{b} + \mathbf{n}, \tag{54}$$

where $\mathbf{D}$ is a $[M \times 1]$ vector, $\mathbf{R}$ is a $[M \times N]$ matrix; and $\mathbf{f}$ is $[N \times 1]$ vector, while $\mathbf{b}$ and $\mathbf{n}$ are $[M \times 1]$ vectors. The typical matrix inversion, $\hat{\mathbf{f}} = (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T(\mathbf{D}-\mathbf{b})$ will usually result in an ill-conditioned solution, where the noise is amplified by the near-singular properties of $\mathbf{R}^T\mathbf{R}$ matrix (see [3,10,22–29]). Furthermore, the ill-conditioning and noise amplification usually produces specious oscillations in $\hat{\mathbf{f}}$ which result in the violation of the positivity and additivity of properties of $\mathbf{f}$. This in turn renders any physical or chemical interpretation meaningless. In other words, it is often desirable to determine solution to the inverse problem (53) and (54), while retaining the positivity and additivity properties of the spectra.

The application of Bayesian/maximum entropy (MaxEnt) methods have proven to be very successful in restoring intrinsic spectra [3,10,30]. This approach employs Bayes theorem to define the posterior distribution for $\mathbf{f}$ and other terms in Eq. (54) such as background level, $\mathbf{b}$, and noise characteristics of the data. The MaxEnt component ensures the positivity and additivity of $\mathbf{f}$, from which the physical properties of the sample can be inferred. In this section we outline the application of Bayesian/MaxEnt methods for the numerical solution of Eq. (54). For convenience, we will not include a discussion concerning the estimation of the background level, but direct the reader to a number of useful references [3,10,30]. Starting with Bayes theorem for $\mathbf{f}$,

$$p(\mathbf{f}|\mathbf{D},\mathbf{m},\alpha,\mathcal{I}) = \frac{p(\mathbf{f}|\mathbf{m},\alpha,\mathcal{I})p(\mathbf{D}|\mathbf{f},\mathcal{I})}{p(\mathbf{D}|\mathbf{m},\mathcal{I})}, \tag{55}$$

where $p(\mathbf{f}|\mathbf{D},\mathbf{m},\alpha,\mathcal{I})$ is the posterior pdf for $\mathbf{f}$ conditional on the recorded data, $\mathbf{D}$; prior model, $\mathbf{m}$, positive constant α and $\mathcal{I}$. The prior pdf specifies the positivity and additivity properties of $\mathbf{f}$, relative to a default model, $\mathbf{m}$, is expressed in terms of entropy function (see [3]),

$$p(\mathbf{f}|\mathbf{m},\alpha,\mathcal{I}) = \frac{1}{Z_S}\exp\left[-\alpha S(\mathbf{f},\mathbf{m})\right], \tag{56}$$

where Z_S is the normalisation term; the entropy function, $S(\mathbf{f},\ \mathbf{m})$ defines how assign the positive and additive values to $\mathbf{f}$, relative to $\mathbf{m}$ (see [3]),

$$S(\mathbf{f},\mathbf{m}) = \sum_{j=1}^{N} f_j - m_j - f_j \ln\left(f_j / m_j\right). \tag{57}$$

In Eq. (57), $\mathbf{m}$, denotes the prior model or the default for $\mathbf{f}$. Given no other data or information about $\mathbf{f}$, the prior model, $\mathbf{m}$, maximises Eq. (57). From a functional analysis point of view, $\mathbf{m}$ is a measure that ensures Eq. (57) is invariant regardless of the underlying coordinate system or bin-width (see [3]). The role of the likelihood function is to constrain Eq. (57) to the available data, $\mathbf{D}$. In this case, we assume that Gaussian pdf is appropriately counts for the likelihood term,

$$p(\mathbf{D}|\mathbf{f},\mathcal{I}) = (2\pi)^{-M/2}\det\sum{}^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{D}-\mathbf{Rf})^T\left[\sum{}^{-2}\right](\mathbf{D}-\mathbf{Rf})\right] \tag{58a}$$

$$\equiv (2\pi)^{-M/2}\det\sum{}^{-1/2}\exp[-L(\mathbf{D},\mathbf{f})], \tag{58b}$$

where $\sum = [\sigma_{ii}^2;\ \forall i]$ is a $[M \times M]$ diagonal matrix which defines the variance for the likelihood function as $\sigma_{ii}^2 = D_i,\ \forall i$. Using Eqs. (56), (58a) and (58b) in Eq. (55),

$$p(\mathbf{f}|\mathbf{D},\mathbf{m},\alpha,\mathcal{I}) \propto \exp\left[-\alpha S(\mathbf{f},\mathbf{m}) - L(\mathbf{D},\mathbf{f})\right] \tag{59a}$$

$$\propto \exp\left[-Q(\mathbf{D},\mathbf{f},\mathbf{m},\alpha)\right] \tag{59b}$$

where α can be interpreted as a Lagrangian multiplier which couples the entropy and likelihood functions. Strictly speaking, we should also treat α as a Bayesian variables and define its pdf [3,10,22–25,27–29],

$$p(\alpha|\mathbf{D},\mathbf{m},\mathcal{I}) = \int p(\mathbf{f},\alpha|\mathbf{D},\mathbf{m},\mathcal{I})\mathcal{D}\mathbf{f}, \tag{60}$$

where $\mathcal{D}$f represents the entropic measure, $\mathcal{D}\mathbf{f} = \prod_{j=1}^{N} f_j^{-\frac{1}{2}} df_j$, since it corresponds to $\sqrt{\det[-\nabla\nabla S]}$. Using Eq. (60) enables $\hat{\alpha}$ to be evaluated which can be used to determine $\hat{\mathbf{f}}$. Alternatively, Eq. (60) can be used to average over the solution to produce $\langle \mathbf{f} \rangle$.

This discussion outlines the importance of a using an appropriately defined prior, such as the entropy function. That is, typical the intrinsic profile is a positive and additive distribution [25,27–29]. Finally, using the basic rules of probability theory and Bayesian analysis, a detailed analysis of spectra has been outlined.

7. Summary

In this tutorial paper we have demonstrated how Bayesian theory can be used to analyse a wide variety of experimental chemistry data. Each case draws on Bayes theorem to express the posterior probability density function conditional on the experimental data and all necessary prior information. The manipulation of Bayes theorem involves applying the product and marginalisation rules [3]. Here in lies the strength and simplicity of Bayesian analysis techniques, whereby a small number of rules can be produced in detailed and sophisticated analysis of experimental data. Moreover, these rules are ideally suited to the scientific reasoning process, since the underlying hypothesis or parameter is analysed with respect to experimental data, and quantified in terms of a probability.

Acknowledgment

The authors would like to acknowledge Drs E. Chow and Bin Gong, School of Chemistry, UNSW for use of lead voltammetry data presented in Section 5.1 and carbon XPS presented in Section 6.1, respectively.

Appendix A. List of symbols

$\mathbb{Z}$	Set of integers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^+$	Set of all positive real numbers
$\mathbb{R}^n$	Set of all ordered n-tuples of real numbers
$\mathbb{N}$	Set of nonnegative integers
$\in$	“belongs to”
$\forall$	for all (values of)
$\sim A$	“of the order” or “similar” to A
$\lesssim$, ($\gtrsim$)	less than and similar to, (greater than and similar to)
$\cup$, ($\cap$)	Union, (Intersection)

$A \equiv B$	A is equivalent to B
$A \subset B$	A is subset of B
$A \subseteq B$	A is a subset of (or is included in) B
$\langle x \rangle$ or $\overline{x}$	is the mean of x
det$\mathbf{A}$	Determinant of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of matrix $\mathbf{A}$
$\mathbf{A}^{\top}$	Transpose of matrix $\mathbf{A}$
$\nabla \mathbf{f}(\mathbf{x})$	Del operator given by $\frac{\partial f}{\partial x_i}$, where $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_N\}$
$\nabla\nabla Q(\mathbf{x})$	$= \frac{\partial^2 Q}{\partial x_i \partial x_j}, \forall i, j$
$\mathrm{erf}(x)$	$= \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t$ is the error function
$\Gamma(x)$	$= \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$ is the Gamma function
$\Pr(\cdot \vert \cdot)$	is conditional probability
$p(\cdot \vert \cdot)$	is conditional probability density function

Appendix B. Summary of Markov Chain Monte Carlo method for Bayesian analysis

B.1. Overview

In this section an outline of Markov Chain Monte Carlo (MCMC) methods is presented. MCMC methods are particularly important in Bayesian analysis for numerically sampling the posterior pdf and evaluating the probabilistic evidence. MCMC methods draw samples from the posterior pdf by constructing a random walk in the model's parameter space [8]. The random walk is determined by the probability of stepping in a region of the parameter space, which is proportional to the posterior pdf. The Metropolis–Hastings (MH) algorithm and Gibbs sampler are two popular MCMC methods applied in Bayesian analysis, where the latter is a special case of the former [9]. An outline of the properties of Markov chains is presented in Section B.2, pseudo-code for the MH-algorithm and the Gibbs sampler are given in Section B.3, while Section B.4 to Section B.6 discuss details necessary for analysing the data from MCMC methods.

The literature regarding MCMC methods, MH algorithm and Gibbs sampler is considerable (see [3–9]). Chen et al. [4] demonstrate the application of Bayesian and MCMC methods in chemometrics. In particular, Gamerman et al. [5], Gelman et al. [6], Gilks et al. [7] and O'Ruanaidh [9] are very good references that discuss the implementation and application of MCMC methods. In addition, Sivia and Skilling [3] and Gregory [8] provide a “hands on” application of MCMC methods.

B.2. Properties of Markov Chain

Essentially, a Markov chain is a sequence of random variables (or *state variable*), $\{X_1, X_2, X_3, \ldots, X_N\}$, where the current value, X_i, is the dependent on the previous value X_{i-1} and not any other previous values, such that

$$p(X_{N+1} \vert X_1, X_2, X_3, \cdots, X_N, \mathcal{I}) = p(X_{N+1} \vert X_N, \mathcal{I}). \tag{61}$$

A Markov chain is defined by its transition kernel, which specifies the probability of transition from X' to X [5,9],

$$p_{N+1}(X) = \sum_{\{X'\}} p_N(X') T_N(X', X), \tag{62}$$

where $T_N(X', X)$ is the transition kernel.

In order to use Markov chain to sample the posterior pdf, the chain must have two properties [9], namely, *detailed balance* and *ergodicity*. Detailed balance refers to the probability of transition from state X' to X is equal to the probability of transition from state X to X' [9],

$$p(X') T(X', X) = p(X) T(X, X'). \tag{63}$$

This property ensures that the random walk reaches equilibrium, which in turn implies that the distribution is stationary. In practice, once the random walk has reached equilibrium, it represents a sample of the posterior pdf.

The second condition, ergodicity, implies that independent of the initial starting distribution, the chain converges to a stationary distribution [5], $\pi(x)$,

$$\lim_{N \to \infty} p_N(x) = \pi(x) \tag{64}$$

The ergodic property of the Markov chain ensures that probability density of the samples converges to the posterior pdf and is an equilibrium distribution [9]. Determining the rate of convergence of the Markov chain is dependent on a number of factors, such as the properties of the *proposal pdf* and the *burn in* [9] — see Section B.4.

B.3. Outline of Metropolis–Hastings algorithm and Gibbs sampler

In this section, the full MH-algorithm is presented, while Gibbs sampler is outlined and necessary modification, where the MH-algorithm can be adapted as presented.

B.3.1. Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm is widely used in generating a random sample from a posterior density function, where the distribution of the sample converges to the posterior pdf. It is particularly useful when sampling posterior densities, which include non-uniform prior pdfs. In addition, it is also applicable to problems, where the dimensions of the parameter space is large and the parameters can be sampled as a single vector. However, often in these types of problems an estimate of the parameter's covariance matrix is necessary, which can be determined using a non-linear optimization.

The pseudo-code for the MH-algorithm is given by Algorithm 1 — see below. It is used to sample a single posterior pdf, $p(\mathbf{X} \vert \mathbf{D}, \mathcal{I})$, such that $\mathbf{X} \in \mathbb{R}^N$. The proposal distribution, $q(\mathbf{Y} \vert \mathbf{X}_n)$, and posterior pdf, $p(\mathbf{Y} \vert \mathbf{D}, \mathcal{I})$, where $\mathbf{Y} \equiv \mathbf{X}_{n+1}$, must be defined prior to the simulation, as well as initial starting values, $\mathbf{X}_0$ — note the subscript denotes the number of iterations. The proposal distribution specifies the probability of the proposed value, $\mathbf{Y}$, conditional on the previous value, $\mathbf{X}_n$. Also a counter, a, is also used to count the number of accepted steps, which is necessary in tuning the algorithm. The total number of iterations are given by $L_{\max}$. In Section B.4 to Section B.6, the necessary details needed in applying the MH or more generally MCMC algorithms are discussed.

B.3.2. Gibbs sampler

The Gibbs sampler reduces the problem of sampling a complex multivariate posterior pdf to a univariate conditional posterior pdf, where a single parameter value is drawn from a pdf, conditional on the remaining parameters with fixed values [5,7,9]. In other words for an N-dimensional parameter-space and for a single iteration, the N-parameters are drawn sequentially from N conditional posterior pdfs. This is in contrast to the MH-algorithm, given above, where for single iteration an N-dimensional vector is drawn from an multivariate posterior pdf. The underlying assumption of the Gibbs sampler is that conditional posterior pdf are easier to sample compared to a multivariate pdf, and can be expressed in a simple standard form [5,7,9]. This is particularly useful when using conjugate priors, since the posterior pdf will have the same functional form as the likelihood function.

If the conditional posterior pdf cannot be expressed in a simple standard form, the MH-algorithm can be applied to sample the conditional probability density [5,9]. In fact, Algorithm 1 can be adapted to include Gibbs sampling steps, since it is a special case of the MH-algorithm [6,7,9]. That is, Gibbs steps are applied with a local

MH-algorithm [9]. For example, given $\mathbf{X}=\{X_1, X_2, \ldots, X_N\}$ parameters, and assuming starting values of $\{X_1^{(0)}, X_2^{(0)}, \ldots, x_N^{(0)}\}$ the iterations consist of [5–7,9]:

Algorithm 1
Pseudocode for Metropolis–Hastings algorithm

```
Initialize $\mathbf{X}_0$                                    ◁ Initialize starting value
Initialize $a=0$                                             ◁ Initialize acceptance counter
Initialize $q(\mathbf{Y}|\mathbf{X}_n)$                      ◁ Initialize proposal distribution
Initialize $p(\mathbf{Y}|\mathbf{D},\mathcal{I})$            ◁ Initialize the posterior pdf
for $n \leftarrow 1, L_{\max}$ do
    $\mathbf{Y} \leftarrow q(\mathbf{Y}|\mathbf{X}_n)$       ◁ Randomly draw Y
    $w \leftarrow \frac{p(\mathbf{Y}|\mathbf{D},\mathcal{I})q(\mathbf{X}_n|\mathbf{Y})}{p(\mathbf{X}_n|\mathbf{D},\mathcal{I})q(\mathbf{Y}|\mathbf{Y}_n)}$   ◁ Evaluate the ratio, w, to compare Y with $\mathbf{X}_n$
    if $w \geq 1$ then
        $\mathbf{X}_{n+1}=\mathbf{Y}$                        ◁ Accept Y unconditionally as the new value
        $a=a+1$
  else
      $r \leftarrow U[0, 1]$                                 ◁ Randomly draw r from a uniform pdf [0, 1]
  end if
  if $r \leq w$ then
      $\mathbf{X}_{n+1}=\mathbf{Y}$                          ◁ Only accept Y with probability r, else reject.
      $a=a+1$
  else
      $\mathbf{X}_{n+1}=\mathbf{X}_n$
      end if
end for
$aratio = \frac{a}{L_{\max}}$                                ◁ Evaluate the acceptance ratio
```

First iteration:

$$X_1^{(1)} \leftarrow p\left(X_1|\mathbf{D}, X_2^{(0)}, \ldots, X_N^{(0)}, \mathcal{I}\right)$$
$$X_2^{(1)} \leftarrow p\left(X_2|\mathbf{D}, X_1^{(1)}, X_3^{(0)}, \ldots, X_N^{(0)}, \mathcal{I}\right)$$
$$X_3^{(1)} \leftarrow p\left(X_3|\mathbf{D}, X_1^{(1)}, X_2^{(1)}, X_4^{(0)}, \ldots, X_N^{(0)}, \mathcal{I}\right)$$
$$\vdots$$
$$X_N^{(1)} \leftarrow p\left(X_N|\mathbf{D}, X_1^{(1)}, X_2^{(1)}, X_3^{(1)}, \ldots, X_{N-1}^{(1)}, \mathcal{I}\right)$$

Second iteration:

$$X_1^{(2)} \leftarrow p\left(X_1|\mathbf{D}, X_2^{(1)}, \ldots, X_N^{(1)}, \mathcal{I}\right)$$
$$X_2^{(2)} \leftarrow p\left(X_2|\mathbf{D}, X_1^{(2)}, X_3^{(1)}, \ldots, X_N^{(1)}, \mathcal{I}\right)$$
$$X_3^{(2)} \leftarrow p\left(X_3|\mathbf{D}, X_1^{(2)}, X_2^{(2)}, X_4^{(1)}, \ldots, X_N^{(1)}, \mathcal{I}\right)$$
$$\vdots$$
$$X_N^{(2)} \leftarrow p\left(X_N|\mathbf{D}, X_1^{(2)}, X_2^{(2)}, X_3^{(2)}, \ldots, X_{N-1}^{(2)}, \mathcal{I}\right)$$

*k*th iteration:

$$X_1^{(k)} \leftarrow p\left(X_1|\mathbf{D}, X_2^{(k-1)}, \ldots, X_N^{(k-1)}, \mathcal{I}\right)$$
$$X_2^{(k)} \leftarrow p\left(X_2|\mathbf{D}, X_1^{(k)}, X_3^{(k-1)}, \ldots, X_N^{(k-1)}, \mathcal{I}\right)$$
$$X_3^{(k)} \leftarrow p\left(X_3|\mathbf{D}, X_1^{(k)}, X_2^{(k)}, X_4^{(k-1)}, \ldots, X_N^{(k-1)}, \mathcal{I}\right)$$
$$\vdots$$
$$X_N^{(k)} \leftarrow p\left(X_N|\mathbf{D}, X_1^{(k)}, X_2^{(k)}, X_3^{(k)}, \ldots, X_{N-1}^{(k-1)}, \mathcal{I}\right)$$

For each parameter, the new parameter values is drawn using a local MH-algorithm, such that w in Algorithm 1 becomes,

$$w_i = \frac{p\left(X_i^{(k)}|\mathbf{D}, X_1^{(k)}, X_2^{(k)}, \ldots, X_{i-1}^{(k-1)}, X_{i+1}^{(k-1)}, \mathcal{I}\right)}{p\left(X_i^{(k-1)}|\mathbf{D}, X_1^{(k)}, X_2^{(k)}, \ldots, X_{i-1}^{(k-1)}, X_{i+1}^{(k-1)}, \mathcal{I}\right)} \tag{65}$$

for the *i*th-parameter on the *k*-iteration. The proposal pdf for a single parameter is symmetric and as a consequence $X_i^{(k)}=X_i^{(k-1)}+\xi_i$, where ξ_i is drawn from density function, typically a normal density function, with zero mean and standard deviation, σ_{ξ_i}. The standard deviation, σ_{ξ_i}, is proportional to the standard deviation of the parameter's posterior density function and typically has to be tuned — see Section B.4. For the case of a symmetric proposal density, the Metropolis–Hastings algorithm reduces to a Metropolis algorithm.

As pointed out above, the Gibbs sampler is ideally suited when the conditional posterior pdfs can be expressed in a standard form. However, when the dimensions of the parameter space are large and the conditional cannot be expressed in standard form, such as for the XPS problem given in Section 6.1, an equivalent number of samples must be drawn before a single iteration is completed. For Eq. (65), where a local Metropolis algorithm is applied, it may be necessary to tune σ_{ξ_i} for each parameter, which becomes difficult if there exist strong correlations between the parameters. On the other hand, the full MH-algorithm, such as Algorithm 1, requires the covariance matrix of the parameters, which can be difficult to determine prior to simulations. In 'pre-analysis', the Gibbs sampler can be used to generated samples for each parameter from which the covariance matrix can determined and the parameter's expected values can be used as the starting values in the full MH-algorithm.

B.4. Choosing a proposal pdf, acceptance ratio, tuning proposal pdf and 'burn in'

B.4.1. Choosing a proposal pdf

Choosing the best proposal pdf in Algorithm 1 can be challenging. The final MCMC sample of the posterior pdf is independent to the proposal pdf. However, the proposal pdf does affect the rate of convergence of the MCMC to the equilibrium distribution. That is, a large class of distribution functions can be used for the proposal pdf, but the underlying issue becomes how to determine the standard deviation (or covariance matrix) of the proposal pdf — see below. In most cases, a multivariate (Gaussian) distribution is sufficient. However, O'Ruanaidh [9] suggests using a fat-tailed distribution which allows for large transitions.

B.4.2. Acceptance ratio

In Algorithm 1 the number of accepted steps are counted, and after the MCMC simulations are complete, the acceptance ratio is quantified. The acceptance ratio provides an indication of whether the MCMC algorithm is tuned. That is, the ratio indicates whether the scaling of the variance (i.e. for one-dimensional sampling) or covariance matrix (i.e. for multi-dimensional sampling) of the proposal distribution is appropriate to sample the posterior pdf. The present simulations tuned the proposal pdf to have an acceptance ratio of ~50%. An outline of the tuning of the proposal pdf is given below.

If the acceptance ratio is ~0, it implies that the random walk is spending substantial time waiting, with the occasional abrupt change. This implies that the standard deviation (or covariance matrix) of the proposal pdf is very much larger than the standard deviation (or covariance matrix) of the posterior pdf. On the other hand, if the acceptance ratio is ~1, it suggests that the standard deviation (or covariance matrix) of the proposal pdf very much smaller than the standard deviation (covariance matrix) of the posterior pdf. The random walk takes many small steps and is slow to converge. Essentially, the standard deviation (or covariance matrix) of the proposal pdf is proportional to the standard deviation (or covariance matrix) of the posterior pdf and can be scaled to achieve the shortest correlation length, $\hat{l}$ — see below.

B.4.3. Tuning the proposal pdf

The tuning of the proposal pdf can be a trial and error process. However, 'pre-MCMC' sampling can be used to tune the proposal pdf, such that the final MCMC simulations result in the desired acceptance ratio. Once the proposal distribution is chosen and the covariance matrix estimated for the parameters, the covariance matrix of the proposal distribution can be scaled, such that

$$\mathbf{\Sigma}_{\mathrm{pd}} = c\mathbf{\Sigma}, \tag{66}$$

where $\mathbf{\Sigma}_{\mathrm{pd}}$ is the covariance matrix of the proposal pdf and $\mathbf{\Sigma}$ is the covariance matrix of the parameters. A number of pre-MCMC simulations are completed for increasing c-values, for $c\in\{c_{\min}, c_{\max}\}$. For each value of c the resulting acceptance ratio is determined. A calibration plot for the set of acceptance ratio versus c-values was generated, and used to determine the c-value, $\hat{c}$, that produces an acceptance ratio of ~0.5.

This tuning approach used 15 to 20 values over the range $[c_{\min}, c_{\max}]$, each requiring ~2000 samples for each simulation to estimate the acceptance ratio. The determined $\hat{c}$-values were used in the 'final' MCMC simulations consisting of 60×10^4 samples, with a burn in of ~20%. The final acceptance ratio were about ± 1% to 3% from the initial estimates of the acceptance ratio in the pre-MCMC simulations.

B.4.4. Burn in

The MCMC simulations requires an initial phase or 'burn in' for the MCMC simulations to equilibrate. A rule of thumb is to reject the initial ~15 to 20% of $L_{\max}$. This can be confirmed by plotting the MCMC samples for each parameter as a function of the number of samples and determining the region where the MCMC simulations begin to equilibrate.

B.5. Independent sample

In general the raw MCMC data is correlated and requires re-sampling at the correlation length scale or lag length, $\hat{l}$. The correlation length scale can be determined from autocorrelation function and the sampling efficiency of the MCMC process. The autocorrelation function is given by,

$$\rho(l) = \frac{\sum_{i=1}^{N-l}\left(X_i - \hat{X}\right)\left(X_{i+l} - \hat{X}\right)}{\sqrt{\sum_{i=1}^{N-l}\left(X_i - \hat{X}\right)^2}\sqrt{\sum_{i=1}^{N-l}\left(X_{i+l} - \hat{X}\right)^2}}, \tag{67}$$

where l is the lag-length, and is evaluated over a range of values. The optimum correlation length can be estimated by fitting Eq. (67) with

$$\rho(l) = e^{-l/\hat{l}}, \tag{68}$$

where $\hat{l}$ is the correlation length scale. Using $\hat{l}$, the MCMC data can be re-sampled to produce an independent sample from which subsequent analysis can be carried out. In addition, using $\hat{l}$, the efficiency of the MCMC simulation can be estimated [31,32],

$$\eta\left(\hat{l}\right) \approx \frac{1}{1 + 2\hat{l}}, \tag{69}$$

which provides another indicator of the effectiveness of the tuning process.

B.6. Descriptive statistics

The mean, median and standard deviation can be determined from the re-sampled data. The re-sampled data can also be binned and the histogram of the data can be produced.

The re-sampled data also can be used to produce the inverse cumulative distribution (i.e. percentile *versus* re-sampled data value). This distribution enables the credibility regions from the re-sampled data to be estimated. The distribution consists of an *xy*-pair of percentile value versus sorted data. The percentile-range is given by i/N, where $i = 1,2,3,\ldots,N$ and N is the total number of re-sampled data. The re-sampled data are sorted in ascending order and a plot of percentile value versus sorted data can be generated. Similarly, the cumulative distribution is given by the *xy*-pair of the sorted data versus percentile value. In both cases, the values can be tabulated. In the former case, for the 0.025 and 0.975-percentiles the corresponding parameter values can be read directly from tabulated values or plot.

References

[1] G. D' Agostini, Bayesian inference in processing experimental data: principles and basic applications, Rep. Prog. Phys. 66 (2003) 1383–1419.

[2] E.T. Jaynes, Probability Theory: The Logic of Science, Cambridge Uni. Press, Cambridge, 2003.

[3] D.S. Sivia, J. Skilling, Data Analysis: A Bayesian Tutorial, Oxford Uni. Press, Oxford, 2006.

[4] H. Chen, B.R. Bakshi, P.K. Goel, Toward Bayesian chemometrics — a tutorial on some recent advances, Anal. Chim. Acta 602 (2007) 1–16.

[5] D. Gamerman, H.F. Loes, Markov Chain Monte Carlo: Stochatic Simulations for Bayesian Inference, 2nd editionChapman and Hall, New York, 2006.

[6] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, 2nd edition Chapman and Hall, New York, 2003.

[7] W.R. Gilks, S. Richardson, D.J. Spiegelhalter, Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics, Chapman and Hall, New York, 1996.

[8] P.C. Gregory, Bayesian Logical Data Analysis for the Physical Sciences, Cambridge Uni. Press, Cambridge, 2005.

[9] J.J.K. O'Ruanaidh, W.J. Fitzgerald, Numerical Bayesian Methods Applied to Signal Processing, Springer, New York, 1996.

[10] V. Dose, Bayesian inference in physics: case studies, Rep. Prog. Phys. 66 (2003) 1421–1461.

[11] D.J.C. MacKay, Information Theory, Inference and Learning Algorithms, Cambridge Uni. Press, Cambridge, 2003.

[12] D.B. Hibbert, N. Armstrong, An introduction to Bayesian methods for analyzing chemistry data. Part II: A review of applications of Bayesian methods in chemistry, Chemom. Intell. Lab. Syst. 97 (2) (2009) 211–220, doi:10.1016/j.chemolab.2009.03.009.

[13] R.T. Cox, Probability, frequency and reasonable expectation, Am. J. Phys. 14 (1) (1946) 1–13.

[14] I.J. Good, When a batter turns murderer, Nature 375 (1995) 541.

[15] D.S. Sivia, C. Burbidge, R.C. Roberts, R.M. Bailey, A Bayesian approach to the evaluation of equivalent dose in sediment mixtures for luminescence dating, Maximum Entropy and Bayesian Methods, 2004, 24th Internation Workshop, Garching bei Muchen, Germany, 2004.

[16] D.B. Hibbert, J.J. Gooding, Data Analysis for Chemistry: Introductory Guide for Students and Laboratory Scientists, Oxford Uni. Press, New York, 2005.

[17] S.F. Gull, Bayesian data analysis: straight-line fitting, in: J. Skilling (Ed.), Maximum Entropy and Bayesian Methods, Kluwer Academic Pub., 1989, pp. 511–518.

[18] E.T. Jaynes, Straight line fitting — a Bayesian solution. In Maximum Entropy and Bayesian Methods. Kluwer Acad. Press, Netherlands. http://citeseer.ist.psu.edu/69286.html.

[19] E. Chow, D.B. Hibbert, J.J. Gooding, Electrochemical detection of lead ions via the covalent attachment of human angiotensin I to mercaptopropionic acid and thioctic acid self-assembled monolayers, Anal. Chim. Acta 543 (2005) 167–176.

[20] D.S. Sivia, C.J. Carlie, Molecular spectroscopy and Bayesian spectral analysis — how many lines are there? J. Chem. Phys. 1 (1) (1992) 170–178.

[21] N. Armstrong, Bayesian analysis of band-broadening models in high performance liquid chromatography, Chemometr. Intell. Lab. Syst. 81 (2006) 188–201.

[22] N. Armstrong, W. Kalceff, J.P. Cline, J. Bonevich, Bayesian inference of nanoparticle broadened X-ray line profiles, J. Res. Nat. Inst. Stand. Techn. 109 (1) (2004) 155–178.

[23] N. Armstrong, W. Kalceff, J.P. Cline, J. Bonevich, A Bayesian/maximum entropy method for certification of a nanocrystallite-size NIST standard reference material, Diffraction Analysis of the Microstructure of Materials, Chapter 8, Springer-Verlag, Berlin, ISBN: 3-540-40519-4, 2004, pp. 187–227.

[24] W. von der Linen, Maximum entropy data analysis, Appl. Phys., A 60 (1995) 155–165.

[25] S.F. Gull, Developments in maximum entropy data analysis, in: J. Skilling (Ed.), Maximum Entropy and Bayesian Methods, Kluwer Acad. Publ., Netherlands, 1989, pp. 53–71.

[26] J. Skilling, R.K. Bryan, Maximum entropy image reconstruction: general algorithm, Mon. Not. R. Astron. Soc. 211 (1984) 111–124.

[27] R.K. Bryan, Maximum entropy analysis of oversampled data problems, Eur. Biophys. J. 18 (1990) 165–174.

[28] R.K. Bryan, Solving oversampled data problems by maximum entropy, in: P.F. Fougére (Ed.), Maximum Entropy and Bayesian Methods, Kluwer Acad. Publ., Netherlands, 1990, pp. 221–232.

[29] J. Skilling, Quantified maximum entropy, in: P.F. Fougére (Ed.), Maximum Entropy and Bayesian Methods, Kluwer Acad. Pub., Netherlands, 1990, pp. 341–350.

[30] W. von der Linen, V. Dose, Signal and background separation, Phys. Rev. E 59 (6) (1999) 6527–6534.

[31] K.M. Hanson, G.S. Cunningham, Posterior sampling, Phys. Rev. E 59 (6) (1999) 6527–6534.

[32] K.M. Hanson, G.S. Cunningham, Posterior sampling with improved efficiency, SPIE 3338 (1998) 371–382.