Chapter

Bayesian Approach for X-Ray and Neutron Scattering Spectroscopy

Alessio De Francesco, Alessandro Cunsolo and Luisa Scaccia

Abstract

The rapidly improving performance of inelastic scattering instruments has prompted tremendous advances in our knowledge of the high-frequency dynamics of disordered systems, yet also imposing new demands to the data analysis and interpretation. This ongoing effort is likely to reach soon an impasse, unless new protocols are developed in the data modeling. This need stems from the increasingly detailed information sought for in typical line shape measurements, which often touches or crosses the boundaries imposed by the limited experimental accuracy. Given this scenario, the risk of a bias and an over-parametrized data modeling represents a concrete threat for further advances in the field. Being aware of the severity of the problem, we illustrate here the new hopes brought in this area by Bayesian inference methods. Making reference to recent literature results, we demonstrate the superior ability of these methods in providing a probabilistic and evidence-based modeling of experimental data. Most importantly, this approach can enable hypothesis test involving competitive line shape models and is intrinsically equipped with natural antidotes against the risk of over-parametrization as it naturally enforces the Occam maximum parsimony principle, which favors intrinsically simple models over overly complex ones.

Keywords: inelastic X-ray scattering, inelastic neutron scattering, Bayes analysis, MCMC methods, model choice

1. Introduction

In the last decade, a large amount of inelastic neutron and X-ray scattering measurements focused on the study of the collective atomic dynamics of disordered system [1–5]. Although, across the years, the analysis of the line shape reported in these measurements seldom benefited from the support of a Bayesian inference analysis, the need of this statistical tool is becoming increasingly urgent. As a general premise, it is worth stressing that a scattering measurement somehow resembles a microscope pointed on the dynamics, whose "focus" can be adjusted by suitable choice of the momentum $\hbar Q$ and the energy $E = h\omega$ exchanged between the particle beam and the target sample in the scattering event, where $\hbar$ is the reduced Planck constant, Q is the wave vector transfer, and ω is the angular frequency. Specifically, upon increasing Q the probe "perceives" the response of the system as an average over smaller and smaller distances $\approx 2\pi/Q$ and over times $\approx 2\pi/\omega$ including a decreasing number of elementary microscopic events, e.g., mutual interatomic collisions. The observable accessed by these spectroscopic

methods is the spectrum associated to density fluctuations, either spontaneous or scattering induced. When quasi-macroscopic distances are probed, i.e., in the $Q \rightarrow 0$ limit, the detail of atomic structure is lost, and the target sample is perceived as a continuum medium, whose dynamic behavior is recorded as an average over many elementary events [6]. Being the mass conserved over macroscopic scales, at these distances the liquid density tends to become a constant of motion, i.e., a time-invariant. For this reason, quasi-macroscopic density fluctuations relax very slowly to equilibrium, and collective density oscillations are correspondingly very long-living. The typical spectral signature of this so-called hydrodynamic behavior is a very sharp triplet reflecting the quasi-conserved nature of hydrodynamic density fluctuations. A striking example of such sharp triplet shape is provided in panel A of **Figure 1**, where the low $-(Q, \omega)$ spectrum on liquid argon at the triple point is reported as measured by Brillouin visible light scattering (BVLS) [7].

One could guess that such a sharp spectral shape does not leave any room for interpretative doubts, also considering that the limiting hydrodynamic spectral profile is exactly known as analytically treatable starting from the application of mass, momentum, and energy conservation laws. Although these statements appear partly true, the very concept of "interpretative doubt" sounds grossly ill-defined before spelling out explicitly the accuracy required to the interpretation one alludes to. Despite its pioneering nature, the quality of the measurements in panel A seems certainly adequate for a precise determination of the side-peak position, probably not much so for a detailed analysis of the spectral tails, which are dominated by the slowly decaying resolution wings. Nonetheless such a shape might still appear a more encouraging candidate for a line shape analysis than its counterpart reported in panel B of **Figure 1** which is featured by broad and loosely resolved spectral features, besides a definitely poorer count statistics. Given that the latter result is fairly prototypical of terahertz spectroscopic measurements on simple disordered systems, one might wonder why, thus far, the analysis of these measurements failed to benefit from Bayesian inference methods as routine line shape analysis tools. Aside of hardly justifiable initial mistrusts, a likely explanation is that only recently these spectroscopic techniques transitioned to a mature age in which the very detection of collective modes in amorphous systems can no longer be considered a discovery in itself, and detailed analyses of the spectral shape are more and more common and required. Again, the take-on message of this course of events is that

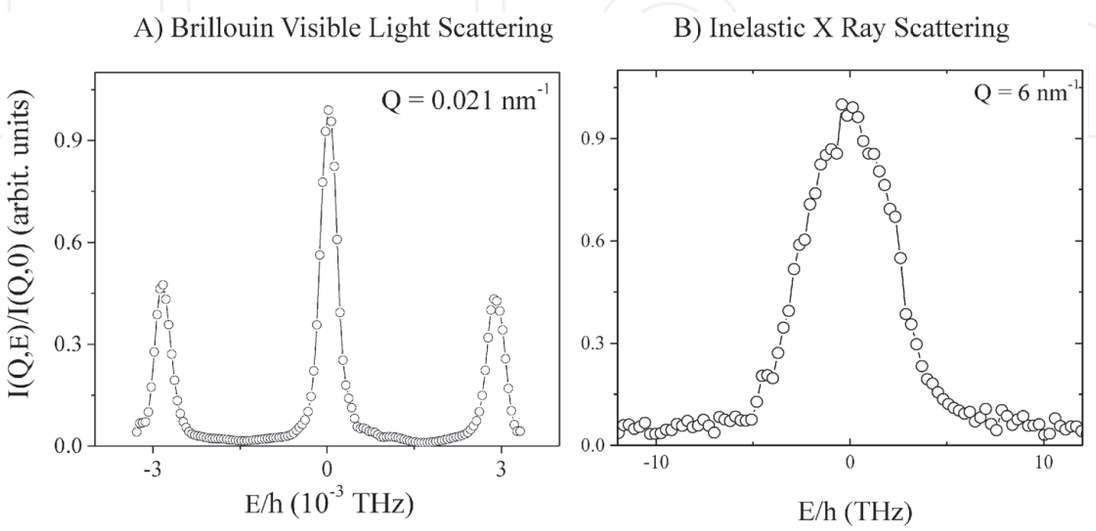

Figure 1.
Panel A: The Brillouin light scattering spectral intensity, $I(Q, \mathrm{E})$, measured in liquid argon at the triple point, redrawn from Ref. [7]. Panel B: The inelastic X-ray scattering spectrum of another noble gas: neon at ambient temperature and 0.3 GPa pressure [8]. Spectral profiles are normalized to their maxima.

the pivotal issue is the adequacy of a given measurement to provide the sought for information, rather than the quality of the measurement in itself. The unbalance between an unavoidably limited experimental performance and the rapidly increasing interpretative needs dramatically enhances the risk of "good faith over-interpretations" representing a lethal threat for the progress of knowledge.

Listen, the data talk! Every time we need to proceed with a data analysis, we could be induced or even tempted, on the basis of our prior knowledge or intuition, to somehow suggest the data what they should tell us about the properties of the system we are investigating. Being driven by acquired knowledge is not necessarily a wrong attitude, it is actually a natural demeanor which effectively drives the cognitive process and the progress of knowledge. However it could become deceiving if we do not have well-consolidated insight about the system under investigation and the observed data are not accurate enough or barely informative. In such cases, in fact, it is highly probable that we just adapt a model to the data, which fits them as well as many other possible models, with the only advantage to deliver results and solutions we feel more at ease with, as they confirm our prior beliefs. This model, of course, can be really plausible and reasonably pondered, and the solution adopted can accidentally be the right one; however, it would be desirable a robust method to quantify how much we can trust such a solution, either in itself or in comparison with alternative ones. We surely want to avoid an aprioristic reliance in a model, which might coerce data to confirm certain results preventing them from providing new insights on the investigated system.

When dealing with neutron or X-ray scattering, the statistical accuracy of spectral acquisition is the primary concern. For the most varied reasons, e.g., relating to the scattering properties of the sample, the integration time, or the count rate of the measurement, the achieved count statistics may either be adequate for a rigorous data analysis or, as often happens, not as good as we would like it to be. In the latter case, the experimental data might not be accurate enough to tell us everything about the physical problem under scrutiny. They could tell us something, but not everything! This is why we need a solid inferential method capable of extracting the maximum amount of information from the data acquired and possibly providing us with a quantitative probabilistic evaluation of the different models that are compatible with the data at hand. Especially when nothing or very little is known about a specific sample or system, the point is, given the observed data, how plausible is a specific model? What is the precision of the conclusions drawn from this model? Are there other possible interpretations of the data at hand? To what extent are different models and interpretations supported by the observed data?

A Bayesian inferential approach provides answers to all these questions on a probabilistic basis, along with a sound criterion to integrate any prior knowledge in the process of data analysis. Bayesian inference, in fact, recognizes the importance of including prior knowledge in the analysis. When we do have well-established prior knowledge about a sample property or a general law a physical phenomenon must comply with, it would be insane and pointless not to use this information. Such a prior knowledge, in fact, can protect us from the risk of making mistakes in the description of experimental data, hence in their interpretation. In the Bayesian framework, prior knowledge takes the form of probability statements so that different probabilities, ranging from zero to one, can be attributed to competitive explanations of the data. In this way, less probable explanations are not excluded a priori but simply given a smaller prior probability. The a priori probability of different explanations is then updated, through the Bayes theorem, based on the new information provided by the data. The results of this analysis, thus, assume the form of posterior probabilities. On this basis, one can easily establish which model is most supported by both data and prior knowledge, what are the posterior

probabilities of alternative models and those of their parameters, and which provides a ground to appreciate the precision of their estimates. In addition, Bayesian methods naturally embody the Occam's razor principle, thus favoring simpler models over unnecessarily complex ones. Last but not least, Bayesian estimation algorithms are generally less affected by the presence of local optima in the parameter space and are not sensitive to the starting values used to initialize the estimation process.

The aim of this chapter is to illustrate how Bayesian inference can be used in X-ray and neutron scattering applications. The Bayesian approach proposed here is implemented through an estimation algorithm, which makes use of Markov chains Monte Carlo (MCMC) methods [9, 10] integrated, where appropriate, with a reversible jump (RJ) extension [11]. This Bayesian method has been already successfully applied in a series of Brillouin inelastic neutron scattering works [12], as well as inelastic X-ray scattering ones [13, 14] and, very recently, in the description of the time correlation function decay in the time domain as measured by spin echo neutron scattering [15, 16]. The rest of the work is organized as follows: Section 2 provides a motivating example; Section 3 revises the Bayes theorem and discusses its different components, as well as some advantages inherent in the Bayesian method; Section 4 applies the Bayesian inference to X-ray and neutron scattering spectroscopy with special emphasis on model choice, parameter estimation, and results interpretation.

2. An example: searching for differences

Depending on the problem at hand, our approach to data analysis can be very different. Imagine that we want, as a toy or teaching example, to measure either the neutron or the X-ray scattering spectrum from a system whose spectrum is well-known and its interpretation unanimously agreed. For instance, we aim at extracting the phonon dispersion curve from the thoroughly measured spectral density $S(Q, E)$ of a given sample. In our replica of past measurements, it is possible that the proper discernment of the excitation lines is hampered by both the course instrumental resolution and the limited statistical accuracy. The poor quality of data could prevent us from easily identifying the spectral features (peaks, bumps, shoulders), already measured and characterized by others. For instance, it could be overly difficult to establish how many excitations are present in the spectra. Unless we deliberately refute the conclusions previously reached by other scientists, it is natural to enforce a line shape modeling well-established in the kinematic range spanned and to verify ex post if the resulting spectral features are consistent with those known from literature.

More often, we face a different problem, as we want to measure for the first time a certain system on which we might not have previous knowledge. Alternatively, we could have prior knowledge about that same system, yet in different thermodynamic or environmental conditions— for instance, a liquid either in bulk or confinement geometries—and possible effects of these peculiar conditions are under scrutiny. Changes could also be very small, and, since detecting them is the focus of our research, it is essential to take the most impartial and noninvasive approach. In this second situation, it would be desirable not to rely too heavily on previous results when choosing the model and to allow the measurement to reveal possible new features.

The two situations mentioned above notably differ in the amount of information available on the system before analyzing the data. In the first case, we have a complete knowledge of the system, while, in the second case, this knowledge is partial or even lacking at all. In this second situation, a traditional approach would

consist in either best fitting a model we deem adequate for the data, e.g., well-assessed for the considered sample, albeit only in different thermodynamic or environmental conditions, or fitting competing models to establish the one best performing based on criteria agreed upon, e.g., the chi-square value. Following the first path, we hinge too much on a specific model and on previous partial knowledge, thus jeopardizing the chance of new findings. On the other hand, the second path would be less coercive at the cost of completely ignoring previous partial knowledge. In addition, the model chosen would be simply the one providing the best fit, but no assessment can be made on the plausibility of this or any other fitting model, based on the data measured. Conversely, a Bayesian approach to data analysis would, instead, allow to assign a different prior probability to the different models (accounting for the uncertainty of available information on the system) and, then, revise these probabilities in the light of the data to deliver the posterior probability of each model, conditional on the data at hand.

3. Bayesian inference

3.1 The Bayes theorem

The Bayes theorem stems from the theorem of compound probability and from the definition of conditional probability. If we consider two events A and B, the compound probability theorem states that the probability of the two events occurring simultaneously is given by:

$$P(A,B) = P(B|A)P(A) = P(A|B)\mathrm{P(B)}, \tag{1}$$

where $P(B|A)$ is the probability of observing B, once A has been observed. Obviously, if A and B are independent, so that the occurrence of one of them does not affect the probability of occurrence of the other one, the compound probability theorem reduces to:

$$P(A,B) = P(A)P(B). \tag{2}$$

From Eq. (1), we immediately get:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

which is nothing else than the Bayes theorem.

Let us now consider A as the ensemble of the parameters of a certain model (or class of models) we choose to describe experimental data. In a slightly different notation, let this ensemble be denoted, from now on, as the vector $\Theta = (\theta_1, \theta_2, \ldots, \theta_m)$, where each vector component θ_m is a parameter. Notice that a component of Θ might also be associated to a parameter that designates a particular model among several proposed. Also, consider B as the entire set of experimental data. Let us indicate this dataset with the vector $y = (y_1, y_2, \ldots, y_n)$, with n being the sample size. In this new notation, the Bayes theorem reads obviously as:

$$P(\Theta|y) = \frac{P(y|\Theta)P(\Theta)}{P(y)}, \tag{4}$$

where $P(\Theta|y)$ is the posterior distribution of the parameters given the observed data; $P(\Theta)$ is the prior distribution of the parameters before observing the data;

$P(y|\Theta)$ is the likelihood of the data, i.e., the probability of observing the data conditional on a certain parameter vector; and $P(y)$ is the marginal probability of the data, which plays the role of normalizing constant so that Eq. (4) has a unit integral over the variable Θ. The different elements of Eq. (4) are thoroughly discussed in the following sections.

3.2 The prior distribution

Let us consider the different elements of Eq. (4), starting with the prior distribution (or simply prior) $P(\Theta)$. This is the distribution function *elicited* for the parameters, given the information at our disposal *before* data collection. Using a slightly more redundant notation, the prior can be explicitly denoted as $P(\Theta|I)$, where I represents the a priori information. This prior probability includes all prior knowledge (or lack of it) we might have, and it can be more or less informative depending on the amount of information on the problem under investigation. Using the same explicit notation, the Bayes formula in Eq. (4) can be rewritten as:

$$P(\Theta|y,I) = \frac{P(y|\Theta,I)P(\Theta|I)}{P(y|I)}. \tag{5}$$

Just to make a few examples, it might be possible that a certain parameter θ included in the model is known, or either already measured or somehow evaluated independently, and its value is $\theta^\star$. In this case, we can assume that the parameter takes the specific value $\theta^\star$ with probability equal to one. Otherwise, if we want to be less coercive, we can adopt for the parameter a Gaussian prior centered on $\theta^\star$ and with a variance opportunely chosen to limit the parameter variability to a certain interval around $\theta^\star$. In this way, values closer to $\theta^\star$ will be given a higher a priori probability.

In other situations, the information available on the parameters might be more vague. For example, we might simply know that a certain parameter must be nonnegative or that it must range in a limited interval, as often the case of neutron scattering hampered by severe kinematic constraints. Nonnegative parameters can be a priori assumed to follow, for example, a truncated Gaussian or a gamma distribution, and, if no other information is available, the prior distribution will be adjusted to make allowance for a large parameter variability, reflecting the noninformative initial guess. Parameters having random or hardly quantifiable variations within limited windows can be assumed to approximately follow a uniform distribution over such a window. Also, whenever feasible, any mutual entanglement between parameters, as well as any selection, conservation, or sum rule, should be embodied in a usable distribution function complementing our prior knowledge I in the cognitive process.

Notice that, even if it is common practice to assume that the parameters are a priori independently distributed, correlation between them can be naturally induced by the data, through the combination of the likelihood and the prior. Parameters can be a posteriori correlated, even if they are a priori independent.

3.3 The likelihood function

The likelihood function is the joint probability of the observed data, conditional on the model adopted and its parameter values. Notice that for continuous data, the likelihood becomes a density of probability. Let $y = (y_1, y_2, \ldots y_n)$ be a sample of data. Each datum y_i can be portrayed as a particular realization of a random variable

Y_i distributed as $f(Y_i;\Theta)$. In fact, if we had to collect data again, even under the same experimental conditions, we would obtain a different sample of data. This means that, before collecting data, the i-th result is to be considered a random variable Y_i. Once the data have been collected, y_i is the particular realization observed of the random variable Y_i, and the sample $y = (y_1, y_2, \ldots y_n)$ is the particular realization observed of the multiple random variable $Y = (Y_1, Y_2, \ldots Y_n)$, whose components Y_i are independent and identically distributed as $f(Y_i;\Theta)$. Then, the joint (density of) probability of the observed data is the probability that, simultaneously, each variable Y_i takes the value y_i (or takes a value in the interval $[y_i, y_i + \Delta y_i]$), for $i \in 1, \ldots, n$, i.e., $f(y_1, y_2, \ldots y_n, \Theta)$. Given the independence of the variables Y_i, for $i \in 1, \ldots, n$, and using the compound probability theorem, we obtain:

$$f(y_1, y_2 \cdots y_n; \Theta) = f(y_1; \Theta) f(y_2; \Theta) \cdots f(y_n; \Theta). \tag{6}$$

The left side of Eq. (6) is the likelihood function for the observed sample $y = y_1, y_2, \cdots y_n$, which depends on the unknown parameter vector Θ. If we condition on a particular value of Θ, we can compute the probability (or density) of the observed sample, conditional on Θ, i.e., $P(y|\Theta)$ in Eq. (3).

To be more specific, we can consider spectroscopic data. The observable directly accessed by a spectroscopic measurement is the spectrum of the correlation function of density fluctuation, or dynamic structure factor $S(Q, E)$, which, in a scattering experiment, is a unique function of the energy, $E = \hbar\omega$, and the momentum, $\hbar Q$, exchanged between the probe particles and the target sample in the scattering process. One has:

$$y_i = S(Q, E_j) + \varepsilon_i, \tag{7}$$

where $S(Q, E)$ is the model used for the dynamic structure factor, depending on a vector of unknown parameters Θ, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)$ is a vector of random errors, here assumed to be independently and normally distributed, i.e., $\varepsilon_i \sim N(0, \sigma_i^2)$, for $i \in 1, \ldots, n$. Notice that assuming heteroscedastic errors, we are not imposing any restriction other than normality on the error term. The heteroscedastic model embeds the homoscedastic one, and since the parameters σ_i^2 are estimated from the data, it might reduce to it if the data were compatible with the homoscedasticity constraint.

Under the assumption above, the likelihood function is:

$$P(y|\Theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{[y_i - S_i(Q,E_i)]^2}{2\sigma_i^2}} = cost \cdot e^{-\sum_{i=1}^{n} \frac{[y_i - S(Q,E_i)]^2}{2\sigma_i^2}} \tag{8}$$

Conditional on a certain value of the parameter vector Θ (which might also include the variance σ_i^2 of the error term), we can compute $S(Q, E_i)$ and, thus, $P(y|\Theta)$.

3.4 The posterior distribution and its normalizing constant

The term on the left-hand side of Eq. (3) is the joint posterior distribution of the model parameters, given prior knowledge and measured data, i.e., *after* data collection. It incorporates both prior knowledge and the information conveyed by the data, and Bayesian inference completely relies on it. In practice, prior knowledge about the investigated problem is modified by the data evidence (through the likelihood function) to provide the final posterior distribution (**Figure 2**). Estimates for a single parameter θ_k can be obtained by marginalizing, i.e., by integrating

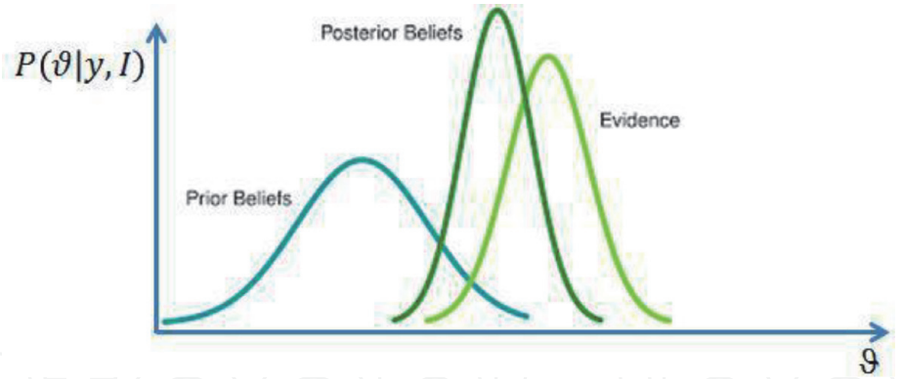

Figure 2.
Sketch of how the prior distribution and therefore our prior knowledge about a model parameter are changed by the data evidence.

(summing) the posterior over all the other parameters to get $P(\theta_k|y) = \int_{\Theta_{-k}} P(\Theta|y)d\Theta_{-k}$, where Θ_{-k} is the whole parameter vector except θ_k. Then, the mean of $P(\theta_k|y)$ is taken as a point estimate of θ_k, while the square root of its variance provides a measure of the estimation error. Also, the probability that the parameter θ_k belongs to a certain interval can be inferred from its marginal posterior.

The term in the denominator of Eq. (3):

$$P(y) = \int P(y|\Theta)P(\Theta)d\Theta \tag{9}$$

is generally called the marginal likelihood and represents the probability of observing the measured data $y_i (i = 1, \cdots n)$, averaged over all possible values of the model parameters. It represents the normalization constant for the posterior distribution, and it is required in the evaluation of $P(\Theta|y)$. However, in most cases, $P(y)$ does not have a closed analytical expression, as its determination would require the computation of high-dimensional integrals. Hence, the posterior distribution can only be obtained up to a normalizing constant, namely:

$$P(\Theta|y) \propto P(y|\Theta)P(\Theta). \tag{10}$$

For this reason, Bayesian inference usually needs to resort to MCMC methods to simulate the joint posterior distribution. MCMC algorithms, in fact, allow to draw values from distributions known up to a normalizing constant, as is often the case for $P(\Theta|y)$. Inference is then carried out on the basis of the simulated, rather than analytical, joint posterior distribution. More details on these methods will be given in Section 3.6 (see also Refs. [9, 10]).

To illustrate an interesting point, let us go back to the example considered before, in which we want to analyze spectroscopic data that can be modeled as in Eq. (7) and for which the likelihood is given in Eq. (8). Imagine to have no prior information at all on the parameters of the model so that the only sensible choice for the prior is a uniform distribution on the parameter space. Then, from Eqs. (8) and (10), it follows that:

$$P(\Theta|y) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left[y_i - S(Q,E_i)\right]^2}{2\sigma_i^2}\right) \propto \exp\left(-\sum_{i=1}^{n} \frac{\left[y_i - S(Q,E_i)\right]^2}{2\sigma_i^2}\right) \tag{11}$$

which implies that the posterior distribution is a multivariate Gaussian. As already mentioned, parameters can be estimated taking the mean of the posterior distribution, which, for a Gaussian distribution, corresponds to the median, mode, and maximum of the distribution. Therefore Bayesian parameter estimates are obtained as those values of Θ that maximize $\exp\left(-\Sigma_{i=1}^{n}\frac{\left[y_i - S(Q,\mathrm{E}_i)\right]^2}{2\sigma_i^2}\right)$. This maximization is equivalent to the minimization of the $\chi^2 = \Sigma_{i=1}^{n}\frac{\left[y_i - \mathrm{S}(Q,\mathrm{E}_i)\right]^2}{2\sigma_i^2}$ function and thus provides the same estimates we would obtain through standard fitting procedures [13]. Therefore, whenever no prior information is available, which translates into a uniform prior, and a normal error distribution is assumed, the posterior distribution coincides up to a constant to the classical likelihood function, and Bayesian and classical estimates are equivalent. This result can be extended to the case of an informative prior, for which, again, Bayesian and traditional approaches provide asymptotically the same results. In particular, as sample size increases, the posterior distribution of the parameter vector approaches a multivariate normal distribution, which is independent of the prior distribution. These posterior asymptotic results [17] formalize the notion that the importance of the prior diminishes as n increases. Only when n is small, the prior choice is an important part of the specification of the model. In such situations it is essential that the prior truly reflects existing and well-documented information on the parameters so that its use can significantly improve the precision of the estimates.

Despite the asymptotic equivalence, sometimes parameters are much easier estimated in a Bayesian rather than in a frequentist perspective. Frequentist estimation, in fact, is generally based on least squares or maximum likelihood methods, and this might be a problem in the presence of local optima. If, for example, the starting values of the parameters, needed to initialize the optimization algorithm, are close to a local optimum, the algorithm might be trapped in this suboptimal solution. As a consequence, different starting values might determine different solutions and, thus, parameter estimates. The Bayesian estimate of a parameter, as stated before, is instead obtained as the mean of its posterior distribution, marginalized with respect to all other parameters. This estimation procedure does not involve any minimization or maximization, and, thus, the fitting algorithm does not risk to get trapped in local optima, and the results are independent from starting values used in the MCMC algorithm used to simulate the posterior distribution (see Section 3.6). It might happen, obviously, that the posterior of one or more parameters is bimodal or multimodal. The presence of different parameter regions with high posterior density might suggest that the data show some evidence in favor of a more complex model but not enough for this model to have the highest posterior probability. In this case, it is not reasonable to use the mean as a point estimate for the parameters, since it might fall in a low posterior density region, and the mode of the posterior distribution can be used in its place. In such situations of posterior multimodality, it is evident how the whole posterior distribution conveys a much richer information than the simple parameter estimate.

3.5 The Occam's razor principle

Even if Bayesian and classical analysis asymptotically give the same results, Bayesian results always have a probabilistic interpretation, and this is particularly relevant when we need to compare different models and determine, for instance, the number of spectral excitations (in the frequency domain) or the number of relaxations (in the time domain). In addition, the Bayesian method represents a natural implementation of the Occam's razor [18–20]: this principle is intrinsic to

Bayesian inference and is a simple consequence of the adoption of the Bayes theorem. In model choice problems, in fact, the posterior probabilities of the different models naturally penalize complex solutions with respect to simple ones, thus conforming to the parsimony principle.

To see this, consider Eq. (4), and imagine that the parameter vector also includes a model indicator parameter M_j, for $j = 1, \ldots, k$. To make this more explicit, we can rewrite Eq. (4) as $P(\Theta, M_j|y) \propto P(y|\Theta, M_j)P(\Theta, M_j)$. Then, Bayesian model choice simply consists in choosing the model with the highest posterior probability $P(M_j|y) = \int_\Theta P(\Theta, M_j|y)d\Theta \propto \int_\Theta P(y|\Theta, M_j)P(\Theta, M_j)d\Theta = P(M_j)\int_\Theta P(y|\Theta, M_j)$ $P(\Theta|M_j)d\Theta = P(y|M_j)P(M_j)$. Thus, if the same a priori probability is attributed to the models, i.e., $P(M_1) = P(M_2) = \cdots = P(M_k)$, the posterior probability $P(M_j|y)$ is simply proportional to the marginal likelihood:

$$P(y|M_j) = \int_\Theta P(y|\Theta, M_j)P(\Theta|M_j)d\Theta \tag{12}$$

Now, consider for simplicity just two possible models, the first one, denoted as M_1, more complex and characterized by a larger number of parameters and the second one, denoted as M_2, simpler and characterized by a smaller number of parameters. Clearly, the more complex model is able to generate a much wider range of possible datasets (i.e., for which the model would provide a reasonable fit) than the smaller model. Therefore, the marginal likelihood $P(y|M_1)$ is more dispersed than $P(y|M_2)$ (cf. Figure 28.3 of Ref. [20]). This implies that dataset in accordance with both M_1 and M_2 have $P(y|M_2) > P(y|M_1)$, while those in accordance with just the more complex model M_1 have $P(y|M_2) < P(y|M_1)$ (with $P(y|M_2) \cong 0$). If the two models are a priori given the same probability, for datasets in accordance with both models, the inequality $P(M_2|y) > P(M_1|y)$ holds for the posterior probabilities, determining the choice of the simplest model to represent the data.

3.6 Bayesian computation of model parameters

As already stated, Bayesian inference completely relies on the joint posterior distribution $P(\Theta|y)$. However, for a complex model, it is often impossible to compute this posterior distribution analytically, and the latter is only known up to a normalizing constant. The MCMC methods allow to draw values from distributions known up to a normalizing constant and, thus, to obtain the simulated joint posterior distribution. In practice, MCMC methods consist in constructing an ergodic Markov chain (**Figure 3**) with states Θ^m, $m = 1 \cdots M$, and stationary distribution corresponding to the joint posterior distribution. M is the number of states, i.e., the number of

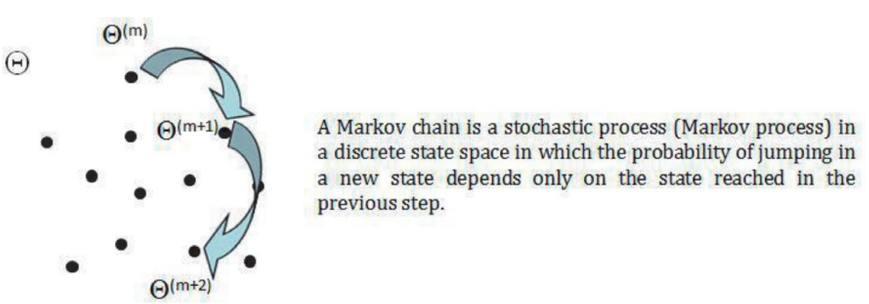

Figure 3.
Parameter updating. Θ is the parameter vector. $\Theta^{(m)}$ is a particular set of parameter values in the parameter hyperspace.

updating of the parameter values, and is generally called the number of sweeps of the MCMC algorithm. At each sweep of the algorithm, a new draw of Θ from its posterior is obtained updating all the parameters in turn, drawing each of them from its posterior distribution, conditional on the value of all the other parameters. If this posterior conditional distribution is known, the parameter is updated using a Gibbs sampling step, which simply draws the new value of the parameter from this known distribution. Otherwise, if this posterior conditional distribution is known only up to a normalizing constant, the parameter needs to be updated through a Metropolis-Hasting move [21]. This move is built as follows. Suppose that, at a given sweep of the algorithm, the current value of a certain parameter is θ. A new *candidate* value θ' can be drawn from an opportunely chosen *proposal* distribution $q(\cdot|\theta)$, which generally depends on the current value θ. The new value θ' is then accepted with a probability equal to $\min(1, R)$, where R is given by:

$$R = \frac{P(y|\Theta')}{P(y|\Theta)} \frac{P(\theta')}{P(\theta)} \frac{q(\theta|\theta')}{q(\theta|\theta)} \tag{13}$$

where Θ' is the whole parameter vector with the parameter θ replaced by the new value θ', $P(y|\Theta)$ is the likelihood, and finally $P(\theta)$ is the prior on that parameter. In other words, R is nothing else but the ratio between the joint posterior distribution calculated with the updated parameter values and the posterior distribution calculated with the current ones, multiplied by the ratio between the proposals, $q(\theta|\theta')/q(\theta'|\theta)$. The higher the posterior ratio, the larger R and hence the probability to move to the new parameter value. In practice, to decide whether or not a candidate value is accepted, a random number is drawn from a uniform distribution defined between 0 and 1 and compared with the calculated value for R. If the random number is less than R, the parameter is updated to the new value; otherwise the new value is rejected. The way the acceptation rule in Eq. (13) is built ensures that the resulting Markov chain has the joint posterior distribution $P(\Theta|y)$ as stationary distribution.

Concerning the proposal distribution, this should be chosen as a distribution from which it is easy to sample. It could be, for instance, a normal distribution centered on the current value of the parameter and with a certain variance which can be adjusted and used as a tuning parameter. This locution alludes to the circumstance that adjustments of this parameter can literally tune the step of the parameter updates. For a normal proposal distribution, a large variance allows the new value θ' to substantially change from the current value. However, if we already are in a high posterior distribution region for the parameter, values far from the current one will fall in low-density regions and are accepted with a very low probability. As a consequence, the algorithm will remain stuck on the same value of the parameter for a long time, causing an inefficient exploration of the parameter space. On the contrary, a small variance will constrain θ' to be close to θ. In this case, the new value has a high probability of being accepted, but the algorithm would move slowly and take a long time to reach convergence to the stationary distribution. The tuning parameters can be appropriately chosen so that the algorithm explores the parameter space efficiently. A rule of thumb states that this happens when the acceptance ratio for each parameter is about 30% [22].

When the parameter vector also includes a model indicator parameter, a further move needs to be considered to update this parameter and to allow the algorithm to explore different models. This move is a reversible jump [11] step, which is specifically designed to allow the Markov chain to move between states having different dimensions (since the dimension of the parameter space varies accordingly to the model considered).

As a final remark, consider that when the MCMC algorithm reaches convergence, after a so called "burn-in" period, the draws not only effectively represent samples from the joint posterior distribution but are also theoretically independent from the starting values of each parameter. Few examples about this point are shown in **Table 1** of Ref. [12]. Notice, however, that the time required to reach convergence might vary a lot depending on the data and the prior. For example, peaked unimodal posterior distributions (i.e., highly informative data) generally speed up convergence, as well as the availability of an important prior information, which reduces the size of the effectively accessible parameter space. On the contrary, the presence of many high posterior density regions can hinder and slow down convergence.

4. The Bayesian approach in neutron and X-ray scattering spectroscopy

4.1 Neutron and X-ray Brillouin scattering

One of the models commonly used to analyze either neutron or X-ray scattering data is the so-called damped harmonic oscillator (DHO) profile, which we report here below:

$$S(Q,E) = A_e(Q)\delta(E) + \frac{E}{k_B T}[n(E)+1]\left\{ L_{A_0 z_0}(Q,E) + \sum_{j=1}^{k} \frac{2}{\pi} \frac{A_j(Q)\Omega_j^2(Q)\Gamma_j(Q)}{\left[E^2 - \Omega_j^2(Q)\right]^2 + 4\left[\Gamma_j(Q)E\right]^2} \right\} \tag{14}$$

where $\delta(E)$ is the Dirac delta function describing the elastic response of the system modulated by an intensity factor $A_e(Q)$, $n(E) = \left(e^{\mathrm{E}/k_B T} - 1\right)^{-1}$ is the Bose population factor expressing the detailed balance condition, and the term in curly brackets is the sum of a Lorentzian central contribution, characterized by the parameters A_0 and z_0, and the contribution of k pairs of peaks, the DHO doublets, symmetrically shifted from the elastic $(E = 0)$ position. The generic j-th DHO is characterized by its undamped oscillation frequency $\Omega_j(Q)$, damping $\Gamma_j(Q)$, and intensity factor $A_j(Q)$. The Lorentzian contribution, not necessarily present, accounts for the quasielastic response of the system. We have intentionally expressed the inelastic contribution as an indefinite sum of k terms, as the scattering signal from amorphous systems is often poorly structured and the number of inelastic modes contributing to it is often hard to guess. A series of concomitant factors, such as the instrument energy resolution, the limited statistical accuracy, and the intrinsically weak scattering signal, can make the line shape modeling not straightforward. In a Bayesian perspective, the number of inelastic features can be treated as a parameter to be estimated along with the other model parameters.

To fit the experimental data, the model in Eq. (14) needs to be convoluted with the instrument resolution, and it can conceivably sit on top of an unknown linear background. Overall, the final model used to approximate the measured line shape is given by:

$$\tilde{S}(Q,E) = R(Q,E) \otimes S(Q,E) + (\beta_0 + \beta_1 E). \tag{15}$$

where " $\otimes$ " represents the convolution operator. For neutron scattering, the instrument resolution function has often a Gaussian shape; thus the final model reads as:

$$\tilde{S}(Q,E) = \left[\frac{1}{\sqrt{2\pi}\zeta}\exp\left(-\frac{E^2}{2\zeta^2}\right)\right] \otimes S(Q,E) + (\beta_0 + \beta_1 E). \tag{16}$$

For IXS, Eqs. (14–16) are still formally valid although the instrument resolution function has usually a slightly more complex shape which appears in the convolution of Eq. (15) either as approximated by an analytical model or measured from the signal of an almost elastic scatterer; obviously, in the latter case, the convolution is computed numerically. The final model is further corrupted by an additive Gaussian noise, having a variance that, for instance, can be taken proportional to each data point. Thus, the experimental data points are given by:

$$y_i = \tilde{S}(Q,E_i) + \varepsilon(Q,E_i), \tag{17}$$

with

$$\varepsilon(Q,E_i) \overset{i.i.d}{\sim} \mathcal{N}\left(0, \sigma^2 \tilde{S}(Q,E)\right), \tag{18}$$

where σ^2 is the proportionality constant. Thus, the likelihood for model in Eq. (17) is simply given in Eq. (8), with $S(Q,E_i)$ replaced by $\tilde{S}(Q,E_i)$ defined in Eq. (16) and $\sigma_i^2 = \sigma^2\tilde{S}(Q,E_i)$.

The whole parameter vector for the model in Eq. (17) is $\Theta = (k, A, \Omega, \Gamma, A_e, A_0, z_0, \beta_0, \beta_1, \sigma^2)$, with $A = (A_1, \cdots, A_k)$, $\Omega = (\Omega_1, \cdots, \Omega_k)$, and $\Gamma = (\Gamma_1, \cdots, \Gamma_k)$, so that the dimension of the parameter vector depends on the number of inelastic modes, k. In a Bayesian perspective, suitable priors need to be chosen for each component of Θ. For example, k can be safely assumed as uniformly distributed between 1 and a certain value $k_{\max}$ opportunely fixed so that all models are a priori given the same probability. All parameters only attaining nonnegative values such as $(A, \Omega, \Gamma, A_e, A_0, z_0$ and $\sigma^2)$ can, instead, be assumed distributed according to a Gamma distribution or a Gaussian distribution truncated in zero. Finally, β_0 and β_1 are assumed to follow a normal distribution, centered in zero and with a large variance, to keep the priors scarcely informative.

Bayesian inference is, then, based on the joint posterior of the whole parameter vector Θ. However, as mentioned, given the complexity of the model $\tilde{S}(Q,E_i)$ in Eq. (17), the normalizing constant in Eq. (9) cannot be analytically evaluated, and it is necessary to resort to MCMC methods to obtain a simulated joint posterior. Since the parameter space dimension depends on the number of inelastic modes, k, a RJ step needs to be added to allow the exploration of a parameter space of variable dimension. The updating of the parameter k can be implemented according to different types of moves, which, for instance, can enable either the creation (the birth) of a new component in Eq. (14) or the suppression (the death) of an existing one, i.e., the so-called birth-death moves; or they can promote the splitting of an existing component into two components or the combination of two existing components into one (split-combine move). These moves are described in Ref. [12]. In practice, at each step, the algorithm tries to jump to another value of k (from 1 to 2, from to 2 to 1 or 3, from 3 to 2 or 4, and so on). The new value of k is accepted with an acceptance probability that guarantees the convergence of the algorithm to the joint posterior distribution.

Once the convergence is attained, after a burn-in period, at each sweep $m = 1, \cdots, M$, the RJ MCMC algorithm draws a vector:

$$\left(k^{(m)}, A^{(m)}, \Omega^{(m)}, \Gamma^{(m)}, A_e^{(m)}, A_0^{(m)}, z_0^{(m)}, \beta_0^{(m)}, \beta_1^{(m)}, \sigma^{2(m)}\right), \tag{19}$$

from the joint posterior $P(\Theta|y)$. In practice, the output of the algorithm is a matrix of the form:

$$\begin{pmatrix} k^{(1)} & A^{(1)} & \Omega^{(1)} & \Gamma^{(1)} & A_e^{(1)} & A_0^{(1)} & z_0^{(1)} & \beta_0^{(1)} & \beta_1^{(1)} & \sigma^{2(1)} \\ k^{(2)} & A^{(2)} & \Omega^{(2)} & \Gamma^{(2)} & A_e^{(2)} & A_0^{(2)} & z_0^{(2)} & \beta_0^{(2)} & \beta_1^{(2)} & \sigma^{2(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ k^{(m)} & A^{(m)} & \Omega^{(m)} & \Gamma^{(m)} & A_e^{(m)} & A_0^{(m)} & z^{(m)} & \beta_0^{(m)} & \beta_1^{(m)} & \sigma^{2(m)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ k^{(M)} & A^{(M)} & \Omega^{(M)} & \Gamma^{(M)} & A_e^{(M)} & A_0^{(M)} & z_0^{(M)} & \beta_0^{(M)} & \beta_1^{(M)} & \sigma^{2(M)} \end{pmatrix} \tag{20}$$

where each row is a particular draw of the whole parameter vector Θ from its joint posterior $P(\Theta|y)$, while each column refers to a particular parameter and represents the whole simulated marginal posterior distribution for that parameter, independently from the values observed for all the other parameters. Model choice can, then, be accomplished considering the first column of the matrix in Eq. (20), that is, the simulated marginal posterior $P(k|y)$. This column contains a string of values for k (e.g., 1, 1, 1, 2, 2, 3, 2, 3, 4, 3,3,3, $2_{\dots\dots\dots}$4,5,4,3, 4, 3, $2_{\dots}$). Therefore, the posterior probability that the number of modes is equal to a specific value ℓ, $P(k = \ell|y)$, is given by the relative frequency of occurrence of the value ℓ in the strings, and the model chosen will be the one corresponding to the value of k with the highest occurrence.

Once a particular model with, let us say, $k = \ell$ inelastic modes has been chosen, the parameters of this model can be estimated conditionally on $k = \ell$. This means that we only need to consider a submatrix of the matrix in Eq. (20), made up of those rows for which the first column is equal to ℓ. Then, a certain parameter θ can be estimated taking the mean (or the mode) of the corresponding column of this sub-matrix, which represents the simulated posterior distribution for θ, conditionally on the model with ℓ modes and marginalized with respect to all the other parameters, i.e., $P(\theta|y, k = \ell)$.

In assessing convergence, a valuable tool is provided by trace plots, which show the sampled values of a parameter over the sweeps of the algorithm. Ideally, a trace plot should exhibit rapid up-and-down variation with no long-term trends or drifts. Imagining to break up this plot into a few horizontal sections, the trace within any section should not look much different from the trace in any other section. This indicates that the algorithm has converged to the posterior distribution of the parameters. Other convergence criteria can be found, for example, in Ref. [23]. **Figure 4** shows the trace plots of three DHO-mode frequencies ($\Omega_{1,2,3}$) the algorithm found, fitting a spectrum relative to IXS data on pure water recently measured (data not published) at room temperature and at a wave vector transfer $Q = 3\text{nm}^{-1}$, after the first 1000 (a), 10,000 (b), and 100,000 (c) sweeps. In plot (a), it can be seen how rapidly Ω_2 and Ω_3 reach their respective high-density regions, while Ω_1 has more problems in exploring the parameter space. Plot (b) shows that, after nearly 2000 sweeps, also Ω_1 finally starts oscillating around its mean, according to its posterior distribution. Plot (c) illustrates that a burn in of, for example, 10,000 sweeps is large enough to ensure convergence of the algorithm: the trace plots for the three parameters stabilize well before the end of the burn in period.

In **Figure 5**, we report an example of Bayesian analysis applied to neutron Brillouin scattering data from liquid gold [12] at different values of the momentum transfer Q. In this work, after a proper removal of spurious effects such as

background, self-absorption, and multiple scattering, the data look indeed rather structured so that inferring the number of inelastic components seems rather obvious and the result confirms the findings of a previous work [24]. Estimates were obtained from 10^5 sweeps of the algorithm, after a burn in of 10^4 sweeps, and the running time for the algorithm was of approximately 5/10 minutes for each spectrum. We chose a precautionary large value for the burn-in, but convergence was normally achieved in a few hundreds of sweeps.

Even in this straightforward case, however, additional insights can be obtained from the posterior distributions delivered by the Bayesian inference. For example, in **Figure 6**, it can be noticed that, as the value of Q increases, the posterior probability of $k = 2$ also increases. This trend in the discrete distribution for k as a function of Q could possibly convey interesting insights on the actual onset of a second excitation or simply indicate a progressive degradation of the experimental data or, still, suggest that, as the damping becomes more and more effective, the determination of the number of inelastic features becomes more controversial.

To investigate these issues, one can look, for example, at the posterior distributions for the excitation frequency Ω, conditional to $k = 1$ and to $k = 2$, respectively (see **Figure 7** for an example on the same data of **Figure 5** for a Q value of 16 nm^{-1}). Considering the matrix in Eq. (20), as explained above, for $k = 1$, all the values in the column referring to Ω_1, and in correspondence with the rows for which $k = 1$, are draws from $P(\Omega_1|y, k = 1)$, and a histogram of these values can be used to visualize the marginal posterior distribution of Ω_1, conditional to $k = 1$. In the same way, for $k = 2$, all the values in the column referring to Ω_1, and in correspondence with the rows for which $k = 2$, are draws from $P(\Omega_1|y, k = 2)$, while those in the column referring to Ω_2 and in the same rows represent draws from $P(\Omega_2|y, k = 2)$. **Figure 7** illustrates, from left to right, the distributions $P(\Omega_1|y, k = 1)$, $P(\Omega_1|y, k = 2)$, and $P(\Omega_2|y, k = 2)$ at $Q = 16\text{nm}^{-1}$.

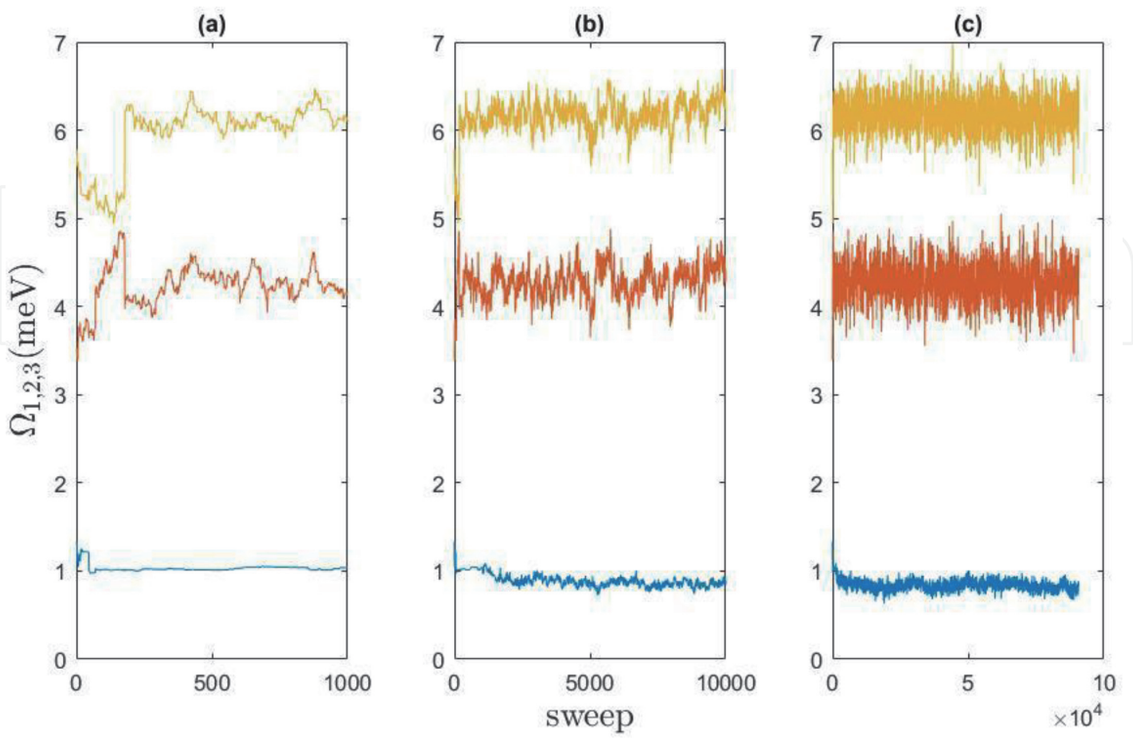

Figure 4.
Typical trace plots for the three DHO-mode frequencies as obtained after the first 1000 (a), 10000 (b) and 100000 (c) sweeps of the algorithm for IXS data on pure water at room temperature and at a wave vector transfer $Q = 3\text{nm}^{-1}$. The frequencies are indexed in increasing order with respect to their value.

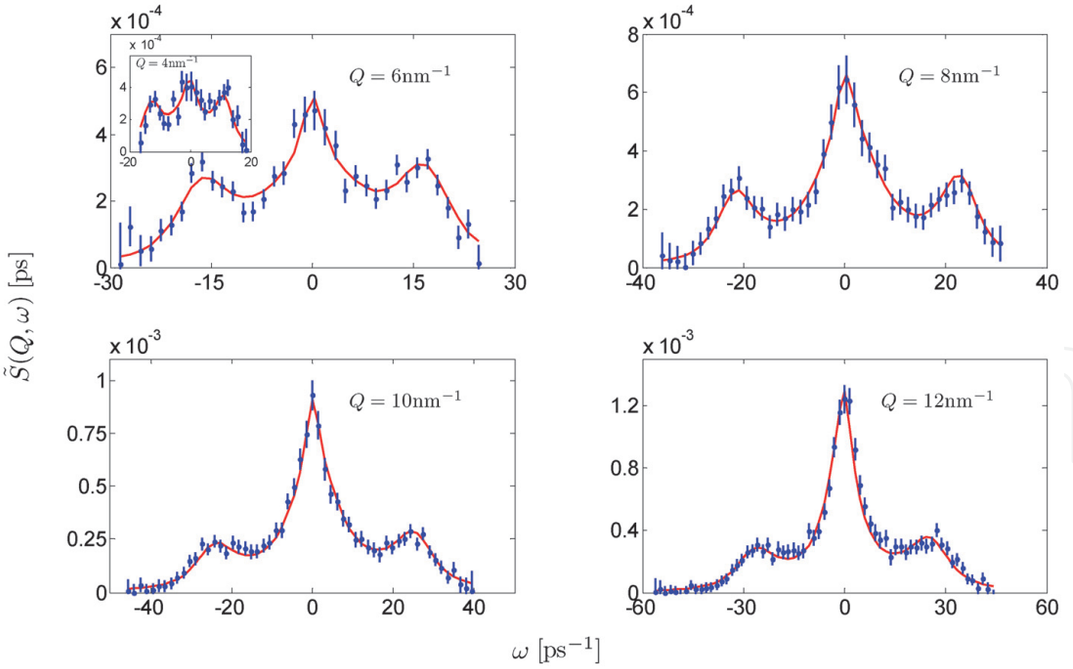

Figure 5.
Dynamic structure factor of liquid gold at five Q values measured on the Brillouin neutron spectrometer BRISP at ILL (Grenoble, France). The experimental data (blue dots) are broadened by the instrumental energy resolution. The RJ-MCMC best fit (red line) takes detailed-balance asymmetry and resolution into account. Reproduced from Ref. [12], Copyright (2016) of American Physical Society.

The shape of these posterior distributions provides a measure of the precision with which the parameter is estimated. For example, $P(\Omega_1|y, k = 1)$ is well-shaped, i.e., unimodal and approximately symmetric, yet quite dispersed. Its mean is equal to 23.8 meV, but there is a 95% probability that the value of Ω_1 lies in the large interva1 22.3–25.2 meV. This large interval tells us that many different values of Ω_1 are compatible with the data, signifying that the inelastic mode at $Q = 16$ nm^{-1} is largely damped—as confirmed also by the large Γ_1 value $(= 7.5\text{meV})$—and less defined, which reveals the large uncertainty in the estimation of the undamped oscillation frequency of the *DHO* excitation. If we now look at the posteriors for $P(\Omega_1|y, k = 2)$ and $P(\Omega_2|y, k = 2)$, we can see that these are much worse shaped than $P(\Omega_1|y, k = 1)$, with unreasonably large or small values having nonvanishing probability. Their mean are, respectively, 17.6 and 25.5 meV and are outside the probability interval obtained for Ω_1, when $k = 1$. Therefore, based on these findings, the Q-evolution of the posterior probability of k seems to simply reveal the increasingly elusive discernment of distinct inelastic features as their damping, or broadening, increases. In practice, at the highest Q explored (16 nm^{-1}), the oscillation mode becomes so highly damped that it can be fitted equally well either by two distinct DHO peaks or by a (broader) single one in the middle of the two. At this stage, the Occam's razor comes into play, naturally integrated in the Bayesian model choice, which ultimately privileges the model with only one DHO, as it involves fewer free parameters. Imagine, instead, that $P(\Omega_1|y, k = 1)$ were bimodal, with the two modes corresponding to the single modes of $P(\Omega_1|y, k = 2)$ and $P(\Omega_2|y, k = 2)$, respectively, as observed, for instance, in Ref. [25]. In this case, the bimodality of $P(\Omega_1|y, k = 1)$ would have provided stronger support to the actual presence of two DHOs, thus suggesting that the finding $P(k = 1|y) > P(k = 2|y)$ only stemmed from the scarcity of data. Should this have been the case, additional observations would have probably led to privilege a more complex model.

Data discussed in Ref. [25] provide another example of the efficacy of Bayesian inference in enforcing the parsimony principle. Specifically, we refer to the case of

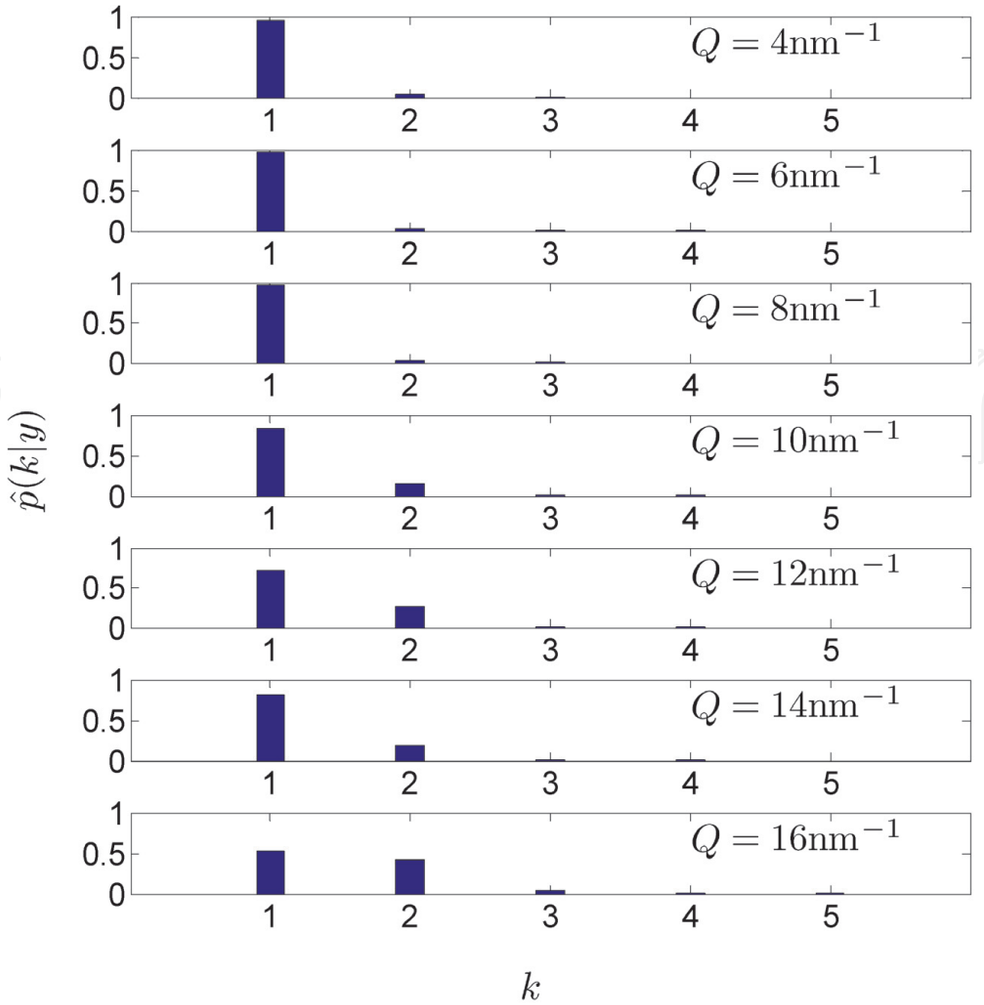

Figure 6.
Posterior probability for the number of modes k at different values of Q for the spectra of ***Figure 5****. Reproduced from Ref. [12], Copyright (2016) of American Physical Society.*

an IXS measurement from a suspension of gold nanoparticles in water which has been analyzed with a model similar to the one in Eq. (14), yet with the DHO terms replaced by Dirac delta functions, due to the extremely narrow width of the measured excitations. For all Q's explored, the posterior distributions for the number of inelastic modes have a maximum (**Figure 8**), which is smaller than $k_{\max}$. In particular, we can also observe that the most probable number of modes and the related probability change from one dataset to the other; this partly reflects the physics of the phenomenon under study but also drawbacks of the modeling, such as the limited count statistics and the increasingly intertwined nature of spectral features at high Q's.

As a further remark, we would like to stress again the fact that results from Bayesian inference are always to be interpreted in a probabilistic nuance. For instance, we stated before that the oscillation mode Ω_1 lies in the interval $(22.3, 25.2)$, with a probability of 95%. This interval, called *credibility interval*, is obtained by sorting the values of Ω_1, drawn from its posterior conditional to $k = 1$, and taking the two values below which we can find, respectively, the 2.5% and 97.5% of all simulated values of Ω_1. In practice, the values inside the interval are those with the highest density given the observed data and so the most credible. Classical confidence interval, obtained in the frequentist approach, does not have such a probabilistic interpretation. The interpretation of confidence intervals is that, if we imagine to repeat data sampling indefinitely under the same conditions and to build a confidence interval at

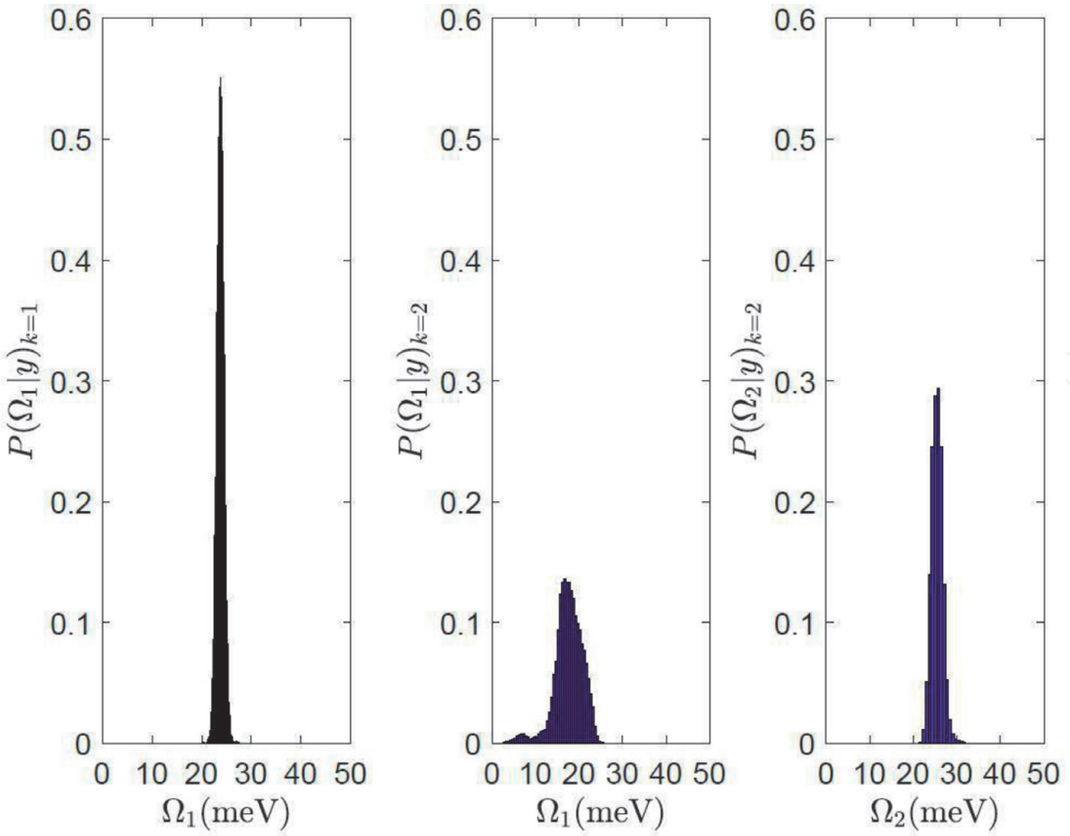

Figure 7.
Simulated posterior distributions for the excitation frequency Ω_1 and Ω_2 in the case of the model with $k = 1$ (panel on the left) and $k = 2$ (central and right panel) for liquid gold at a momentum transfer of $Q = 16\,\mathrm{nm}^{-1}$.

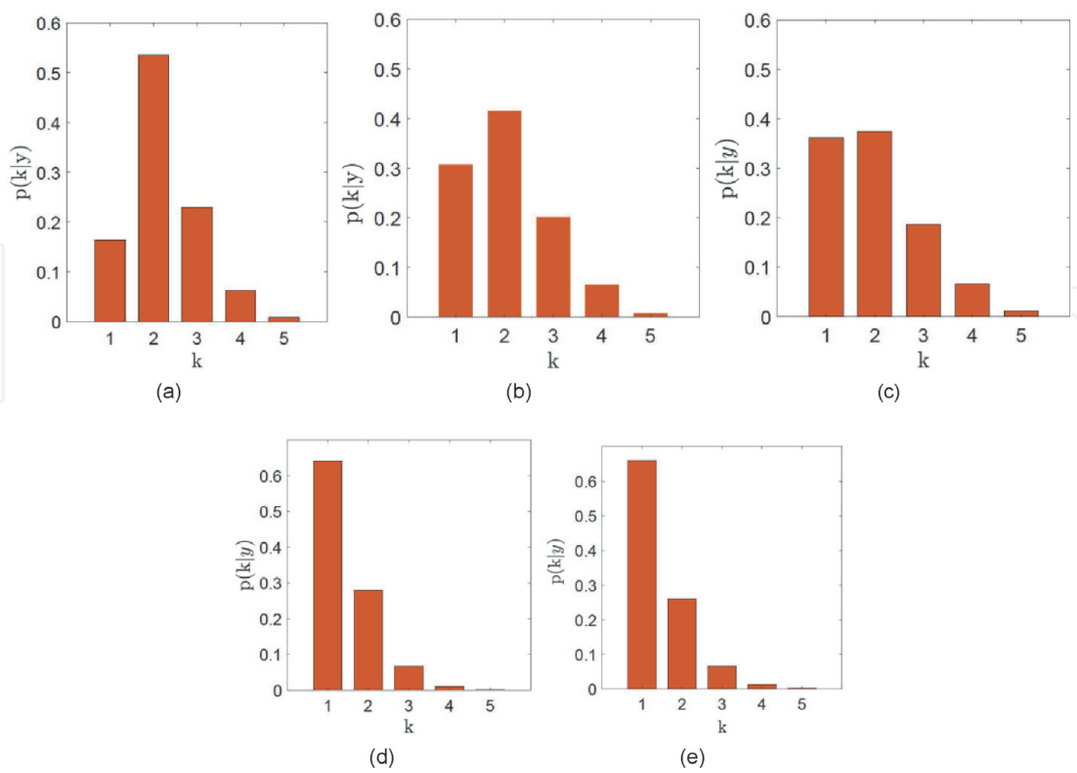

Figure 8.
Posterior probability for the number of k modes at different values of the momentum transfer Q in an inelastic scattering experiment performed on a gold nanoparticle suspension in water: (a) $Q = 3.5\,nm^{-1}$, (b) $Q = 5.5\,nm^{-1}$, (c) $Q = 7.5\,nm^{-1}$, (d) $Q = 9.5\,nm^{-1}$, and (e) $Q = 13.5\,nm^{-1}$. *Adapted with permission from Ref. [13], Copyright (2018) American Chemical Society.*

a certain $1-\alpha$ confidence level for each of the datasets, then (1–α)% of these confidence intervals will contain the true fixed value of the parameter. However, we have no guarantee that the single confidence interval, calculated on the basis of the only dataset actually observed, contains the true parameter value. We can only be confident [at the (1–α)% level] that it does so, since it comes from a set of intervals, (1–α)% of which do contain the parameter value. In practice, under a frequentist approach, data are random variables and give rise to random intervals that have a specific probability of containing the fixed, but unknown, value of the parameter. The single interval is also fixed and might or not contain the fixed parameter, but we cannot associate any probability measure to this possibility. In the Bayesian approach, the parameter is random in the sense that we have a prior belief about its value, while the interval can be thought of as fixed, once the data have been observed. In summary, the frequentist approach do provide a definition of confidence intervals, which, however, are endowed with a robust probabilistic ground only with respect to the hypothetic space of all possible repetitions of the measurement experiment but not with respect to the unique dataset at hand.

4.2 Bayesian inference in the time domain

Time correlation function decays can be modeled in terms of an expansion of the intermediate scattering function $I(Q,t)$ in exponentials, and the aim is often to determine the number of time decay channels that could be envisaged in the relaxation of $I(Q,t)$. In Ref. [15], the dynamics of polymer-coated gold nanoparticles in D_2O was tackled by neutron spin echo (NSE) scattering and analyzed within a Bayesian approach with the goal of establishing how many characteristic relaxations were present in a given spin echo time window and if they could be described by either simple or stretched exponentials or by a combination of the two. The data were assumed to be sampled by the following model:

$$y_i = \gamma \sum_{j=1}^{k} A_j \exp\left(-\left(\frac{t_i}{\tau_j}\right)^{\beta_j}\right) + \varepsilon_i, \qquad \text{for } i = 1, \ldots, n \tag{21}$$

where γ is a proportionality constant possibly enabling a data normalization, k represents the number of exponential relaxations, A_j is the weight of the j-th component of the exponential mixture, τ_j its relaxation time, and β_j its stretching parameter. The ε_i, for $i = 1, \ldots, n$, are random noises, accounting for statistical errors in the measurements. These are assumed to be independent and identically distributed with a normal distribution $\mathcal{N}\left(0, v\sigma_i^2\right)$, where σ_i is the measurement error corresponding to the i-th observation and v is a proportionality constant. As a consequence, the likelihood of the data is a product of normal densities, each having mean $\gamma \Sigma_{j=1}^{k} A_j \exp\left(-\left(\frac{t_i}{\tau_j}\right)^{\beta_j}\right)$ and variance $v\sigma_i^2$.

The value of k is, obviously, unknown, and its determination is of great relevance. Therefore, also in this case, k is considered a stochastic variable to be estimated based on the data and conditional to all the other model parameters. Imagine that we have no clue about how many relaxations are necessary to describe the observed behavior of the time correlation function. However, we are aware that, in a case like this, the risk to over-parametrize the model is high, and we certainly know that, given the finite time window covered by the experiment and the limited number of experimental data, the number of relaxations should not be too large;

otherwise the results could be meaningless, hardly justifiable, and unlikely. Therefore, it seems a priori reasonable that k has a uniform distribution on the discrete values $k = 1, \cdots, k_{\max}$, where $k_{\max}$ is a small integer, as previously assumed when dealing with the number of excitations in the energy spectrum. Also, the relaxation times τ_j are supposed uniformly distributed on a continuous range of nonnegative values. The prior on A_j is tailored to ensure that the combination of relaxation terms fulfills the constraints $\Sigma_{j=1}^{k} A_j = 1$ and $A_j \geq 0$. The natural choice for the prior of $A = (A_1, A_2, \cdots, A_k)$ is, then, a Dirichlet density, which takes values on the standard simplex. A crucial prior is that of the stretching parameter β_j. This is specifically meant, in fact, to discern whether the relaxations in the given time window are simple exponential decays, stretched exponential decays, or a combination of the two. A simple exponential decay corresponds to $\beta_j = 1$, and thus a positive probability mass can be assigned to this specific value. The remaining probability can be assigned to β_j values within the interval $(0; 1)$. Therefore, a reasonable prior for β_j can be a mixed distribution made up of a probability mass in 1 and a continuous beta density, i.e., $\beta_j \sim \zeta \mathcal{B}(\kappa, \psi) + (1 - \zeta)\delta_{\beta_j,1}$, independently for $j = 1, \cdots, k$, where κ and ψ are parameters of the beta density, $\delta_{\beta_j,1}$ is an indicator function equal to 1 when $\beta_j = 1$ and 0 otherwise, and ζ is a weight denoting our prior support in favor of a stretched, rather than simple, exponential components. Once the $\zeta = 0$ and $\zeta = 1$ weights are, respectively, assigned to the sums of simple and stretched exponential terms in Eq. (21), other $0 < \zeta < 1$ weights will be associated to mixed combinations of these decay terms. In particular, a $\zeta = 0.5$ means that the j-th exponential can be either stretched or not with a priori the same probability, for all $j = 1, \cdots, k$. In addition, setting $\kappa = 1$ and $\psi = 1$ allows to assume that the stretching parameters are uniformly distributed on the interval $(0; 1)$. This corresponds to an uninformative prior giving a probability of 0.5 to both a stretched or unstretched component and, for a stretched component, assigning the same density to any value of β_j in $(0; 1)$. Obviously, more informative priors can be chosen, e.g., by assigning different values to κ, ψ and ζ, so to favor, for example, a Zimm or Rouse model specification (see discussion in Ref. [15]) when dealing with polymer dynamics. A similar prior probability can be adopted for the proportionality constant γ, i.e., mixed distribution made up of a continuous beta density and a probability mass in 1, corresponding to no need for a refinement of the data normalization process. Finally, the proportionality constant in the error variance, v, can be, for example, assumed to have a priori a gamma density so that only nonnegative values are allowed.

Let us consider one of the datasets in Ref. [15], representing the time correlation decay of a polymer solution of polyethylene glycol with a molecular weight of 2000D (PEG2000) as measured in a NSE scattering experiment and collected at a momentum transfer $Q = 0.091$ Å. Also in this case, we allowed for 10^5 sweeps of the algorithm and a burn-in of 10^4 sweeps, resulting in approximately 5/10 minutes of computing time. From the output of the MCMC-RJ algorithm, values for the discrete posterior distribution function of k are found in **Table 1**.

The most visited model is the one with two exponential functions. The fit is shown in the figure below (**Figure 9**).

The values reported in **Table 1** clearly show that the posterior distribution of k has a maximum. In fact, this is a general result (see, e.g., **Figures 8** and **10**).

When we model a spectroscopic dataset through a homogeneous mixture, e.g., a linear combination of exponentials, Lorentzians or DHO functions, the posterior distribution for the number of components always has at least a maximum, unless the data are so scarcely informative that the posterior for k simply reproduces the

k	$P(k\|y)\%$
1	8.47
2	61.83
3	23.91
4	4.41
5	1.12
6	0.26

Table 1.
Posterior distribution for the number of time correlation decay channels for a polymer solution of polyethylene glycol with a molecular weight of 2000D(PEG2000) *as measured in a NSE experiment and collected at a momentum transfer* $Q = 0.091$ Å.

prior, which might be uniform. In principle, when jumping in a more complicated model characterized by a larger number of parameters, the χ^2 tends to decrease, and the likelihood tends to increase. However, according to the Bayes theorem, the posterior for k is computed averaging the likelihood over all the parameters value (see Eq. (12)). Therefore, models that are under-parametrized will perform poorly on average since they just cannot fit the data well enough and have a small likelihood, while models that are over-parametrized will also perform poorly on average, because the subset of the parameter space that fit the data well (and where the likelihood is high) becomes tiny compared to the whole volume of the parameter space. This means that adding components to the mixture model increases the posterior distribution of k only until the increment in the likelihood more than compensates for the augmentation of the "wasted" parameter space; overall the competition of these effects ensures the presence of a maximum in $P(k|y)$. It is worth noticing that assuming a model with more free parameters does not necessarily mean a better fit, once the likelihood has saturated. To see this, we report here below (**Figure 11**) the fit we get with a number of relaxation channels $k \neq 2$. We can observe how the fit with three relaxation components or more is not better than the one more supported by the available data and estimated by the MCMC-RJ algorithm. Moreover it is *insane* and hopeless to confer a distinct physical meaning to each one of the corresponding characteristic relaxation times.

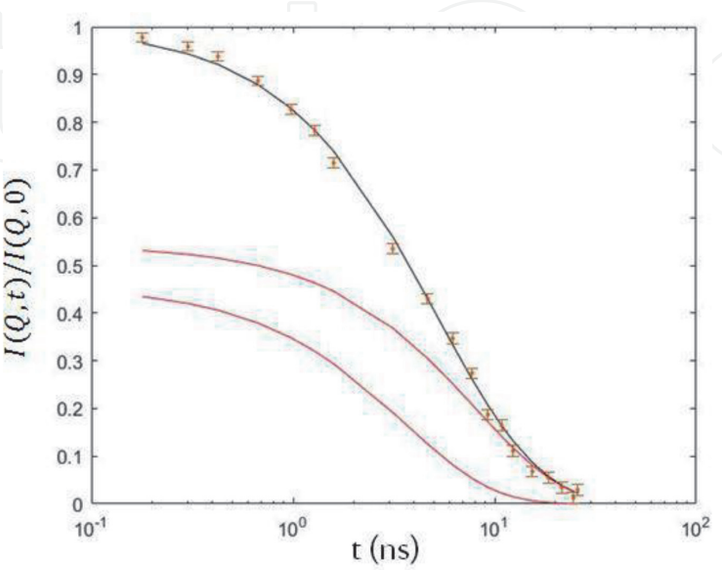

Figure 9.
$I(Q,t)/I(Q,0)$ *vs. time* (ns). *The black line is the best fit as determined with the RJ-MCMC. The two red lines are the two exponential components.*

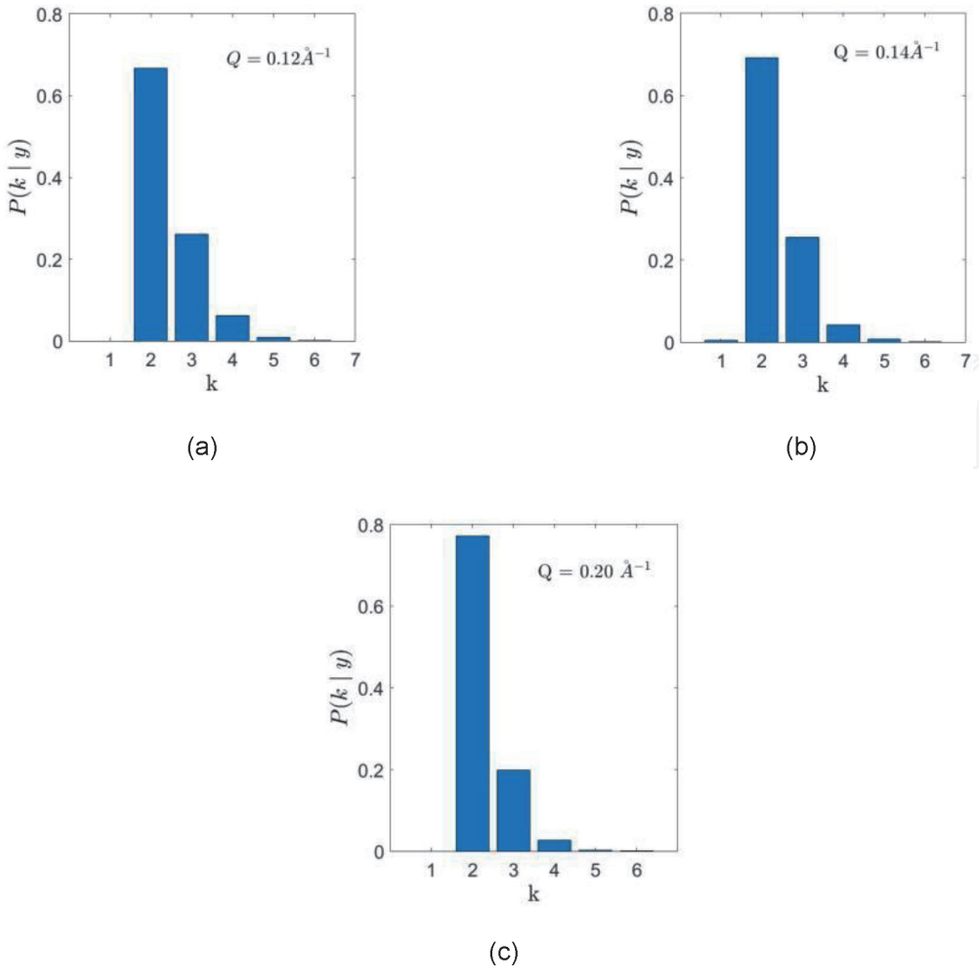

Figure 10.
Posterior probability for the number of k modes at different values of the momentum transfer Q in an NSE experiment performed on polymer solution of polyethylene glycol with a molecular weight of 2000D (PEG2000) *in* D_2O.

Let us introduce a quantity which could resemble the χ^2, namely:

$$s^2 = \sum_{i=1}^{n} \frac{\left(y_i - y_{calc}\right)^2}{\sigma_i^2}, \tag{22}$$

which measures the distance between the experimental data and the best fit determined with the RJ-MCMC algorithm, where n is the number of experimental observations, y_i are the experimental data, y_{calc} are the best fit calculated values, σ_i are the experimental errors. This variable differs from the usual χ^2 as the model parameters are not estimated by least squares minimization, but are the averages, of the corresponding marginal posterior distributions. Nevertheless we can use this quantity to show what follows. If we calculate the quantity in Eq. (22) for each value of k, we get for s^2 the values reported in **Table 2** which indicates an overall decrease upon increasing the number of exponentials. Actually, s^2 does not strictly decrease with the numbers of parameters, because, as mentioned before, the fit is not calculated with parameter values which minimize the χ^2. If, for example, in particular situations (e.g., for $k = 3$), the algorithm faces some challenges in determining a parameter and its posterior distribution is *very broad* and slowly decaying, the average of this parameter could be severely affected by the presence of these sizable distribution tails. In these cases, the mode of the distribution should be used instead to estimate the parameter.

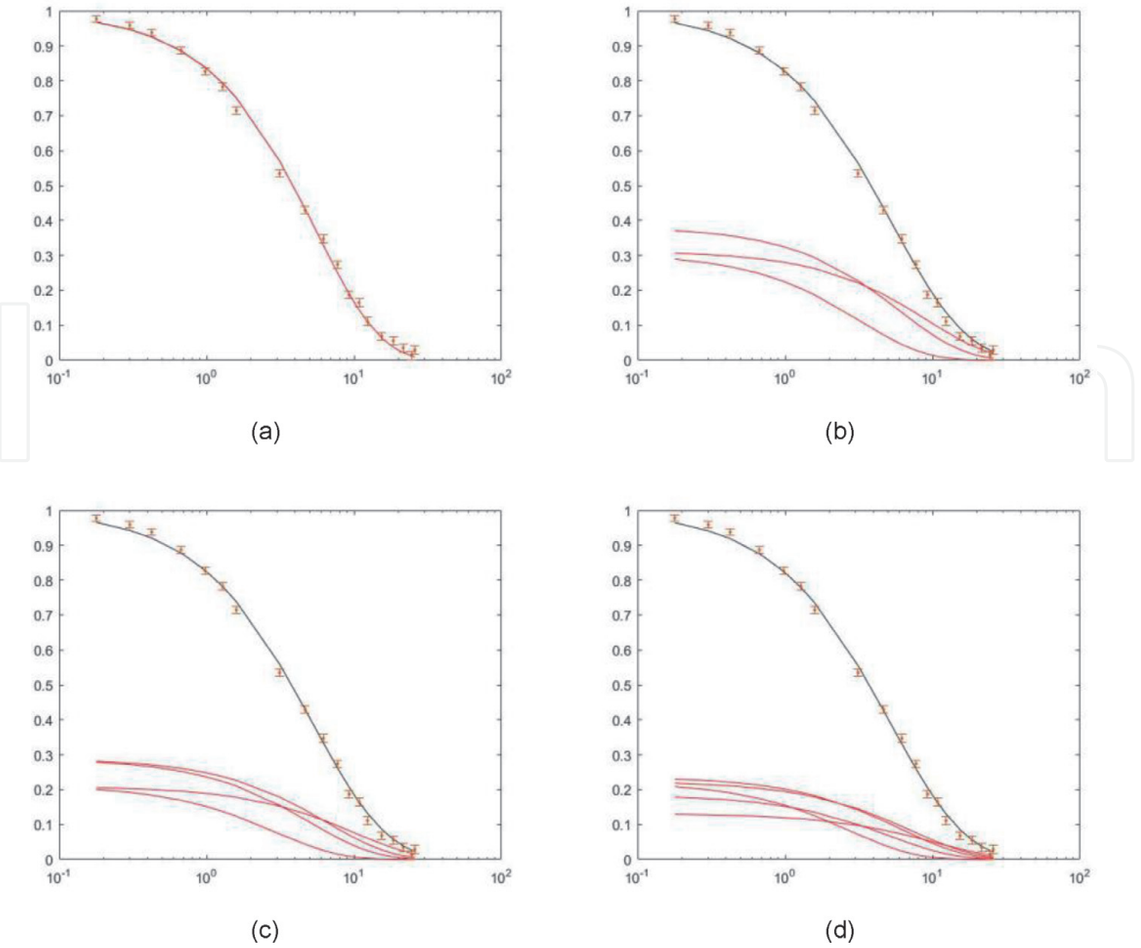

Figure 11.
$I(Q,t)/I(Q,0)$ vs. line (ns). The black line is the best fit as determined with the RJ-MCMC. The two red lines are the two exponential components. (a) $k = 1$, (b) $k = 3$, (c) $k = 4$, and (d) $k = 6$.

k	s^2
1	13.26
2	9.07
3	10.92
4	9.00
5	8.57
6	8.57

Table 2.
Values of the quantity s^2 as defined in Eq. (22) calculated for the different values of k and considering the averages of the model parameter posterior distribution.

Nevertheless, s^2 shows that even with a distance between experimental and fitted values which is effectively decreasing as the number of parameters increases, the most probable model from **Table 1** is the one with $k = 2$ and not the one with $k = k_{\max}$. The effect of the razor is evident. It can also be noted that the fit with $k = 2$ is not only the most probable but it is also the best (in the sense that it is much better than the one with $k = 1$ and it is not worse than those obtained using a larger number of exponentials). More interestingly, we observe also that increasing the number of parameters, the χ^2 (or any other measure of the distance between the fitted and the observed data) is not decreasing much for large values of k, because

obviously at the end, this quantity is going to saturate (and so does the likelihood). The fit with $k = 2$ determines a value of s^2, which is not too different from the one we get with $k = 6$. Incidentally, as it is largely discussed in Ref. [15], the model with two relaxation channels has also a perfectly plausible and consistent explanation, which would not be possible if a more complicated model were chosen.

In summary, we have here shown some of the opportunity offered by a Bayesian inference analysis of experimental results and, in particular, those obtained with spectroscopic methods. As possible future development, it appears very promising the opportunity of applying similar methods to the joint analysis of complementary time or frequency-resolved measurements. Also, we can envisage the use of more informative priors implementing the fulfillment of sum rules of the spectra or any other known physical constraint of the measurement. We are confident that, in the long run, these methods will improve the rigor of routine data analysis protocols, supporting a probability-based, unprejudiced interpretation of the experimental outcome.

This work used resources of the National Synchrotron Light Source II, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC 0012704. The open access fee was covered by FILL2030, a European Union project within the European Commission's Horizon 2020 Research and Innovation programme under grant agreement N°731096.

Acknowledgements

We would like to thank U. Bafile, E. Guarini, F. Formisano, and M. Maccarini for the very stimulating discussions.

Author details

Alessio De Francesco[1,2*], Alessandro Cunsolo[3] and Luisa Scaccia[4]

1 CNR—IOM, Grenoble, France

2 Institut Laue-Langevin, Grenoble, France

3 National Synchrotron Light Source-II, Brookhaven National Laboratory, Upton, NY, USA

4 Università di Macerata, Macerata, Italy

*Address all correspondence to: defrance@ill.fr

References

[1] Boon JP, Yip S. Molecular Hydrodynamics. Mineola, NY: Dover Publication Inc.; 1980

[2] Hansen J-P, McDonald IR. Theory of Simple Liquids. New York: Academic Press; 1976

[3] Balucani U, Zoppi M. Dynamics of the Liquid State. Vol. 10. Oxford: Clarendon Press; 1995

[4] Copley J, Lovesey S. The dynamic properties of monatomic liquids. Reports on Progress in Physics. 1975;**38**:461

[5] Scopigno T, Ruocco G, Sette F. Microscopic dynamics in liquid metals: The experimental point of view. Reviews of Modern Physics. 2005;**77**:881

[6] Berne BJ, Pecora R. Dynamic Light Scattering: With Applications to Chemistry, Biology, and Physics. New York: Dover Publications, Inc.; 2000

[7] Fleury PA, Boon JP. Brillouin scattering in simple liquids—argon and neon. Physical Review. 1969;**186**:244

[8] Cunsolo A, Pratesi G, Verbeni R, Colognesi D, Masciovecchio C, Monaco G, et al. Microscopic relaxation in supercritical and liquid neon. The Journal of Chemical Physics. 2001;**114**: 2259

[9] Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. London: Chapman & Hall/CRC; 1996

[10] Tierney L. Markov chains for exploring posterior distributions. The Annals of Statistics. 1994;**22**:1701

[11] Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995;**82**:711

[12] De Francesco A, Guarini E, Bafile U, Formisano F, Scaccia L. Bayesian approach to the analysis of neutron Brillouin scattering data on liquid metals. Physical Review E. 2016;**94**: 023305

[13] De Francesco A, Scaccia L, Maccarini M, Formisano F, Zhang Y, Gang O, et al. Damping off terahertz sound modes of a liquid upon immersion of nanoparticles. ACS Nano. 2018;**12**:8867

[14] De Francesco A, Scaccia L, Formisano F, Maccarini M, De Luca F, Parmentier A, et al. Shaping the terahertz sound propagation in water under highly directional confinement. Physical Review B. 2020; **101**:05436

[15] De Francesco A, Scaccia L, Lennox RB, Guarini E, Bafile U, Falus P, et al. Model-free description of polymer-coated gold nanoparticle dynamics in aqueous solutions obtained by Bayesian analysis of neutron spin echo data. Physical Review E. 2019;**99**: 052504

[16] Parmentier A, Maccarini M, De Francesco A, Scaccia L, Rogati G, Czakkel O, et al. Neutron spin echo monitoring of segmental-like diffusion of water confined in the cores of carbon nanotubes. Physical Chemistry Chemical Physics. 2019;**21**:21456

[17] Bemardo JM. Philosophy of statistics. In: Bandyopadhyay PS, Forster MR, editors. Handbook of the Philosophy of Science. Vol. 7. Amsterdam: North-Holland; 2011. pp. 263-306

[18] Berger JO, Jefferys WHA. The application of robust Bayesian analysis to hypothesis testing and Occam's razor. Journal of the Royal Statistical Society, Series A. 1992;**1**:17

[19] Jefferys WH, Berger JO. Ockham's Razor and Bayesian Analysis. American Scientist. 1992;**80**:64

[20] MacKay D. Information Theory, Inference and Learning Algorithms. Cambridge: Cambridge University Press; 2003

[21] Chib S, Greenberg E. Understanding the metropolis-hastings algorithm. The American Statistician. 1995;**49**:327

[22] Roberts GO, Gelman A, Gilks WR. Weak convergence and optimal scaling of random walk metropolis algorithms. The Annals of Applied Probability. 1997; **7**:110

[23] Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics: A comparative review. Journal of the American Statistical Association. 1996;**91**:883

[24] Guarini E, Bafile U, Barocchi F, De Francesco A, Farhi E, Formisano F, et al. Dynamics of liquid Au from neutron Brillouin scattering and ab initio simulations: Analogies in the behavior of metallic and insulating liquids. Physics Review. 2013;**B88**: 104201

[25] De Francesco A, Scaccia L, Maccarini M, Formisano F, Guarini E, Bafile U, et al. Interpreting the terahertz spectrum of complex materials: The unique contribution of the Bayesian analysis. Materials. 2019;**12**:2914