Parameter Estimation as a Problem in Statistical Thermodynamics

Keith A. Earle* and David J. Schneider†

*Physics Department, University at Albany (SUNY)
kearle@albany.edu, http://earlelab.rit.albany.edu
†USDA Agricultural Research Service and Department of Plant Pathology
Cornell University, Ithaca NY 14853

Abstract. In this work, we explore the connections between parameter fitting and statistical thermodynamics using the maxent principle of Jaynes as a starting point. In particular, we show how signal averaging may be described by a suitable one particle partition function, modified for the case of a variable number of particles. These modifications lead to an entropy that is extensive in the number of measurements in the average. Systematic error may be interpreted as a departure from ideal gas behavior. In addition, we show how to combine measurements from different experiments in an unbiased way in order to maximize the entropy of simultaneous parameter fitting. We suggest that fit parameters may be interpreted as generalized coordinates and the forces conjugate to them may be derived from the system partition function. From this perspective, the parameter fitting problem may be interpreted as a process where the system (spectrum) does work against internal stresses (non-optimum model parameters) to achieve a state of minimum free energy/maximum entropy. Finally, we show how the distribution function allows us to define a geometry on parameter space, building on previous work[1, 2]. This geometry has implications for error estimation and we outline a program for incorporating these geometrical insights into an automated parameter fitting algorithm.

Keywords: Maximum Entropy, Parameter Optimization, Statistical Physics
PACS: 02.60.Pn,02.70.Rn,05.60.Cd

INTRODUCTION

In order to compare experimental data to a model it is crucial to have a robust means for performing parameter estimation. Our approach to this problem has been strongly influenced by the observation that the sum of square differences or residuals between the experimentally observed signal and a parameter-dependent model is intuitively analogous to a distortion energy. As the fitting procedure progresses, the sum of square residuals is reduced in magnitude until some minimum value is reached. The residual, or perhaps more accurately, a probability density function derived from the residual would be maximally uninformative[3, 4] with respect to any missing information in the absence of systematic error. The present work is an attempt to assess how far one may press these observations in order to address some significant issues in data analysis that have arisen in the authors' laboratories. In particular, data analysis of magnetic resonance spectra from different frequency bands may be known on physical grounds to satisfy a common model of dynamics, yet the individual spectra may have very different lineshapes due to the frequency dependent interplay of magnetic field dependent and magnetic field independent interactions. It is our intention to suggest a procedure that

exploits insights from statistical physics and information theory to address the question of how one should weight contributions from different spectral bands when the model has frequency-dependent parameter sensitivity and the noise residual is also frequency-dependent.

SPECTRAL FUNCTIONS AND PARAMETER STIFFNESS

If one wishes to pursue the analogy that the spectral residual plays the role of a displacement in a distortion energy, then it is useful to define a notion of stiffness. We note at the outset that if one is only analyzing data from a single frequency band, then identification of a stiffness parameter is not strictly necessary. For multifrequency fits however, the sensitivity of the spectrum to changes in parameters will in general be frequency dependent, thus the stiffness of the model to changes in parameters is something that can vary across spectral frequency bands. A magnetic resonance absorption spectrum, at least in the linear response regime[1], is a normalizable absorption cross-section and may be treated as a probability density function (PDF). This observation allows one to compute the Fisher information matrix associated with the spectral lineshape function

$$g_{ij}(\theta) = \int d\omega \left(\frac{\partial \ln p(\omega|\theta)}{\partial \theta^i} \right) \left(\frac{\partial \ln p(\omega|\theta)}{\partial \theta^j} \right) p(\omega|\theta) \tag{1}$$

The notation used here is discussed elsewhere[1] in detail. Here, $p(\omega|\theta)$ is our model for the spectral absorption at frequency ω depending on the parameters $\{\theta\}$. We note that ω may be treated as random variable, as is done in the stochastic resonance experiment[5, 6]. We also note that the spectrum one observes from stochastic resonance, traditional field sweep, or Fourier transform spectroscopy is identical[5, 6]. When interpreting ω as a random variable, one may think of $p(\omega|\theta)$ as representing the 'payoff' for landing on a particular frequency ω.

In order to gain some insight into how the spectral function can be used to define stiffness, we begin with a simple, analytical model. In "Principles of Nuclear Magnetism"[7], Abragam treats the linewidth problem of multiplet spectra including the effects of quadrupolar relaxation and chemical exchange. For multiplet lines with equal *a priori* probabilities, Abragam derives an expression for the lineshape which may be written in the following (normalized) form

$$p(\omega|\theta) = \frac{1}{\pi(2I+1)} \Re \left[\langle v \,|\, C^{-1}(\omega|\theta) \,|\, v \rangle \right], \tag{2}$$

where $|\, v \rangle$ is a column vector of ones for this problem, and $C^{-1}(\omega|\theta)$ is the inverse of a complex symmetric matrix. The symbol $\Re[\cdot]$ denotes the real part of the expression in square brackets in Equation 2. For the particular case of a spin 1/2 nucleus coupled to a spin $I = 1$ nucleus, Abragam derived the following expression for the matrix C in our notation[7]

$$C(\omega|\theta) = \begin{bmatrix} i(\omega_0 - \omega + J) - 3/5T & 1/5T & 2/5T \\ 1/5T & i(\omega_0 - \omega) - 2/5T & 1/5T \\ 2/5T & 1/5T & i(\omega_0 - \omega - J) - 3/5T \end{bmatrix} \tag{3}$$

A simulated spectrum, including noise, based on Equation 3 is shown in the left panel of Figure 1. The structure of the matrix $C(\omega|\theta)$ in Equation 3 has many features in common with the matrices that arise in the study of complex lineshapes[8, 9], although the details depend on the particular problem at hand[1].

In order to assess the sensitivity of the spectrum to the relevant parameters J and $1/T$, we may compute the covariance matrix of the *relative* changes of the spectrum as the parameters are varied with respect to the spectral lineshape[1, 10]. This is the Fisher information, as defined in Equation 1. As shown earlier[1], this leads naturally to the concept of spectral derivatives. For the simple 3×3 matrix inverse that appears in Equation 2, it is tedious, but straightforward to compute explicitly the necessary parameter derivatives for this case. One finds

$$\frac{\partial p(\omega|\theta)}{\partial \theta^i} = -\frac{1}{\pi(2I+1)} \Re \left[\langle v \,|\, C^{-1}(\omega|\theta) \left(\frac{\partial C(\omega|\theta)}{\partial \theta^i} \right) C^{-1}(\omega|\theta) \,|\, v \rangle \right]. \tag{4}$$

We note that the operator form $\partial C^{-1}/\partial \theta^i = -C^{-1} \partial C/\partial \theta^i C^{-1}$ in Equation 4 is the matrix equivalent of the identity $du^{-1}/dx = -(du/dx)/u^2$.

the Fisher information in the multiplet spectrum may be evaluated numerically by using Equation 4 with Equation 1. For this model, we obtain the following values for the matrix elements of the Fisher information, its eigenvalues and eigenvectors. The

TABLE 1. Left: Matrix elements of the (symmetric) Fisher information derived from the model discussed in the text. Center: Eigenvalues of the Fisher information matrix. Right: Eigenvectors of the Fisher information. The matrix elements are given in the (column) order: J, $1/T$

Fisher Information	Eigenvalues	Eigenvectors
$\begin{bmatrix} 9270 & -224. \\ -224. & 4970 \end{bmatrix}$	$\begin{bmatrix} 4960 & 0 \\ 0 & 9280 \end{bmatrix}$	$\begin{bmatrix} -5.2 \times 10^{-02} & -9.99 \times 10^{-01} \\ -9.99 \times 10^{-01} & 5.2 \times 10^{-02} \end{bmatrix}$

results shown in Table 1 contain useful information that is relevant for the parameter estimation problem. Note that the eigenvectors of the Fisher information identify those linear combinations of J and $1/T$ which are linearly independent. The eigenvalues provide a measure of the relative importance of the eigenvectors. For this example, we see that the eigenvalues are comparable and the eigenvectors do not mix J and $1/T$ significantly. Thus, the Fisher information tells us that for this model, and this parameter range, J and $1/T$ may be optimized separately with a high degree of confidence. The stiffness of the model may be found by forming the product of the eigenvalues of the Fisher information matrix.

If one accepts that the determinant of the Fisher information g is a useful measure of stiffness, then one may compute a distortion energy from the squared difference of the

[1] A suite of `Octave` and `Matlab` scripts is available to study the effects of changes in the relaxation time T and coupling constant J at the Earle group website `http://earlelab.rit.albany.edu/`. We thank Nabin Malakar for valuable assistance in converting the `Octave` scripts to a form usable in `Matlab`.

signal and a model trial function, scaled by the stiffness

$$U = \sum_{\omega \in \{\Omega\}} \frac{1}{2} g \left(S(\omega) - M(\omega|\theta) \right)^2 \tag{5}$$

We repeat that introducing the stiffness is not strictly necessary when analyzing a single magnetic resonance spectrum. The stiffness is significant for simultaneous multifrequency fits as we will sketch in the Applications section. In Equation 5, the signal (plus noise) is represented by $S(\omega)$. The model function, related to $p(\omega|\theta)$, is scaled by a dimensional factor. For example, the signal (plus noise) may be obtained from the output of a lock-in amplifier. In this case, the scale factor for $p(\omega|\theta)$ would have units of μV/Hz, say. We note that the factor of 1/2 is conventional. In the absence of noise, the energy U will achieve a minimum value of zero if the model is a faithful representation of the underlying physics and the chosen parameters are at their optimum values. In the presence of noise, U will achieve a minimum value equal to the total noise energy at the optimum parameter set, when systematic error is not present. Note that the noise in the spectrum can be quantified independently of the details of any model. This is, after all, almost a definition of the noise. Thus, computing the mean square deviation for a fraction of the data, far away from any spectral features of interest, is a practical means for estimating the average noise energy for the entire spectrum.

THERMODYNAMIC CONSIDERATIONS

We seek a distribution function for $U(\omega|\theta)$ where $U = \sum_{\omega \in \{\Omega\}} U(\omega|\theta)$ that is normalized, with expectation value $E_p(U(\omega|\theta)) = U$, and that is maximally uninformative with respect to any other information. A particularly complete and helpful derivation for an analogous system using the technique of Lagrange multipliers may be found elsewhere[11]. Following the maxent prescription of Jaynes[3, 4], this leads to the distribution

$$p(\omega|\theta) = \exp(-\beta U(\omega|\theta))/Z \tag{6}$$

where $Z = \sum_{\omega \in \{\Omega\}} \exp(-\beta U(\omega|\theta))$ in Equation 6. We recognize Z as analogous to the partition function of the canonical distribution function[12, 13]. As is well-known, once a partition function is available, any thermodynamic quantity of interest may be derived from it by taking suitable partial derivatives.

The left panel in Figure 1 shows a simulated spectrum, including noise, from which the 'equilibrium' PDF may be inferred by computing the mean squared deviation in the baseline. The right panel in Figure 1 also shows the PDF appropriate for a model that has an exchange rate that is too slow to correctly model the observed noisy spectrum. As the model parameters approach their optimum values, the regions where the PDF dips towards zero are annealed away. For a noisy spectrum the 'equilibrium' or optimum PDF would have a level 'grassy' appearance. In the absence of noise, the PDF would be a uniform straight line across the spectral range. In non-equilibrium thermodynamics[14, 13, pg. 126], the deviation of the entropy from its equilibrium value is given by the negative of the Kulback-Leibler (KL) divergence, scaled by the Boltzmann constant: $-k_B \sum_{\omega \in \{\Omega\}} p(\omega|\theta) \ln(p(\omega|\theta)/p(\omega|\theta_0)$. Here θ_0 is the optimum parameter set. This

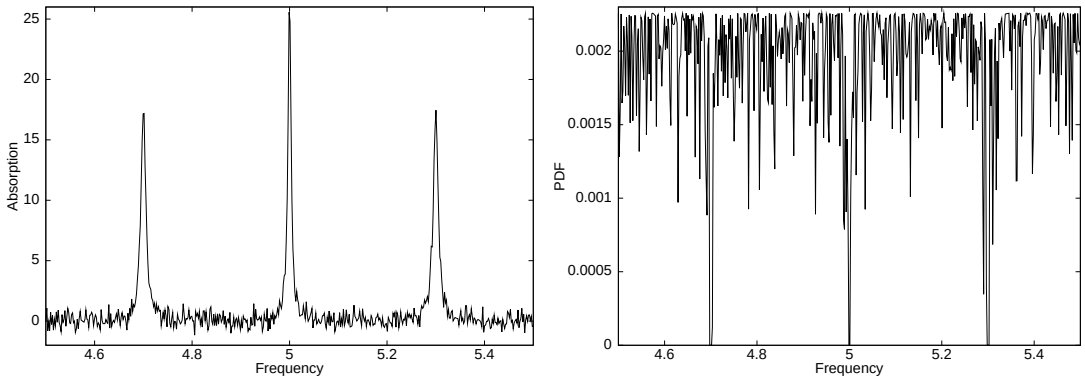

FIGURE 1. Left: Multiplet Exchange Spectrum including noise. Right: PDF computed from a model with an exchange rate that is too slow to fit the spectrum in the left panel.

observation suggests to us that the parameter optimization problem may be profitably viewed as a problem in transport theory where model parameters corresponding to generalized coordinates are varied in order to minimize the KL divergence, or maximize the entropy. For the example shown here, the KL divergence is 0.02 nats for the model with the 'wrong' exchange rate.

A physical picture emerges that given a prescription for defining a partition function one may define forces conjugate to the model parameters. Parameter optimization is then a process of allowing the system to do work on the constraints in order to minimize the Helmholtz free energy and maximize the entropy. A parameter optimization procedure such as nested sampling[15] may be cited as an example of how these notions are implemented in practice. In fact, one of the motivations for developing the approach given here was to gain insights into nested sampling procedure. Now consider the case that we repeat the experiment N times in order to improve the signal to noise, for example. We assume that the repetitions are all independent and that permuting the order of the repetitions would have no effect on the outcome. At a particular observation point $i : i \in \{\Omega\}$, one then has $p_i(\theta) \rightarrow \Pi_{j=1}^{N} p_i(R_{i,j}|\theta)$. Here we introduce new notation $R_{i,j}$ corresponding to the jth measured residual at the observation frequency i. Note that $p_i(\theta) \equiv p(i|\theta)$ in our previous notation. Note that if permutations of the repetitions are truly equivalent, then to assign all permutations to the same equivalence class it is necessary to divide p_i by $N!$. In order to preserve the normalization condition on p_i, it is then necessary to divide the partition function by $N!$ as well. The steps outlined here are parallel to the ones conventionally used in the derivation of the Sackur-Tetrode equation[11, 12]. We shall assume that we may replace the jth residual $R_{i,j}$ by its average $\bar{R}_i \equiv (1/N)\sum_{j=1}^{N} R_{i,j}$. This is, in some sense, a mean field approximation. When this is valid, the modified partition function becomes

$$\bar{Z} \rightarrow \frac{1}{N!}\left(\sum_{i=1}^{m} \exp\left(-\beta \frac{1}{2}\left(\bar{S}_i - M_i(\theta)\right)^2\right)\right)^N . \tag{7}$$

This choice of $\bar{Z}$ in Equation 7 has an interpretation in terms of the ideal gas law, as we will see below.

In order to provide a plausibility argument for the 'mean field' approximation, note that $p_i \propto \exp(-\beta \bar{R}_i^2/2)$ in that approximation, which is of the form $\exp(-(x-\mu)^2/\sigma^2)$. After N measurements, we substitute $p_i \rightarrow (p_i)^N$ up to an overall factor which is equivalent to replacing σ^2 with σ^2/N. This implies that after N independent measurements the standard deviation $\sqrt{\sigma^2} \rightarrow \sqrt{\sigma^2/N}$ is reduced by a factor of $\sqrt{N}$, the standard result for averaging N independent, normally distributed measurements. At least in this case, the 'mean field' approximation is a useful tool for modeling multiple observations.

Suppose that we are able to collect data of sufficiently high signal to noise ratio that all of the exponentials in $\bar{Z}$ approach unity. Then we may approximate $\bar{Z} \approx (1/N!)\,(V/v_0)^N$. Here, (V/v_0) is the number of observations in $\{\Omega\}$ per experiment and v_0 is the minimum resolvable frequency increment in the experiment. The entropy for this partition function may be shown to be

$$S = k_B N \left[\ln\left(\frac{V}{N} \frac{1}{v_0} \right) + 1 \right]. \tag{8}$$

Equation 8 is the Sackur-Tetrode equation appropriate for this system[11, 12]. Note that it is extensive in the number of experiments N. The pressure in this system may be computed from the Helmholtz free energy as follows $p \equiv -\partial A/\partial V = \rho/\beta$ where $A = -\ln(\bar{Z})/\beta$. One may put this expression for p into the ideal gas form by identifying $\rho = N/V$, $\beta = 1/k_B T$, where T is a pseudo temperature and k_B is Boltzmann's constant. With these substitutions, we find $pV = Nk_B T$, the ideal gas law as claimed.

As the model coordinates are varied the free energy will change according to the following prescription $dA = \Theta_j d\theta^j$. Using the definition of Z in Equation 7 one finds that

$$\Theta_j = -N \frac{\sum_{i=1}^m g\,(S_i - M_i(\theta))\,\frac{\partial M_i(\theta)}{\partial \theta^j} \exp(-\beta g (S_i - M_i(\theta)^2/2)}{\sum_{i=1}^m \exp(-\beta g (S_i - M_i(\theta)^2/2)} \equiv -N\langle g(S-M)M_j \rangle, \tag{9}$$

where $M_j \equiv \partial M/\partial \theta^j$ on the right hand side of Equation 9. Note that the generalized forces scale the changes in the free energy due to coordinate changes. If the model is not particularly sensitive to θ^j over a particular range, then Θ_j will be small and the contribution to dA from $\Theta_j d\theta^j$ will be small compared to those terms in dA for which the generalized forces Θ_j are significant. These coordinate changes correspond to the system doing work on itself in order to minimize the free energy. When this has been achieved, the sum of square residuals is a minimum and the entropy is a maximum.

APPLICATIONS

In this section, we will provide some suggestions for applications of the thermodynamic analogy that we have presented here. This is currently work in progress. One of the more intriguing possibilities suggested by the definition of a partition function for a spectrum is that one can define its heat capacity from $C_V \equiv \beta^2(\partial^2 \ln Z/\partial \beta^2)$. This is relevant for the problem of parameter optimization when spectra from several spectral bands are available. Consider the following classic chemical physics problem from Reif[16]: Two

substances with different heat capacities C_A and C_B at temperature T_A and T_B are brought into contact. What is the final temperature? Answer: $T_f = (C_A T_A + C_B T_B)/(C_A + C_B)$ when C_A and C_B are independent of temperature. Extensions to more systems in contact are obvious. This problem may be solved by noting that the partition function for a composite system may be written in the following form

$$Z = \prod_j \frac{1}{N_j!} \left(\frac{V_j}{v_0} \zeta_j \right)^{N_j}, \tag{10}$$

where ζ_j measures departures from ideal gas behavior. With our definitions

$$\zeta_j = \frac{1}{m_j} \sum_{i=1}^{m_j} \exp\left(-\beta_j \frac{1}{2} g_j \left(\bar{S}_i - M_i(\theta) \right)^2 \right). \tag{11}$$

Here, $m_j = V_j/v_0$ the number of observed frequencies in the j^{th} spectrum. In the absence of systematic error and if the model is a faithful representation of the underlying physics $\zeta_j \to 1$ at the optimum parameter set. Starting from the composite partition function of Equation 10, we suggest the following algorithm for simultaneous multifrequency fits

1. Estimate starting parameters $\{\theta\}$
2. Infer β_j's and g_j's
3. Vary $\{N_j\}$ to make β_j's all equal (defines an isothermal ensemble)
4. Vary $\{\theta\}$ according to a search algorithm which maximizes the entropy defined by the composite partition function.
5. Update β, N_j's and g_j's so that they are consistent with $\{\theta\}$.
6. When the stopping criterion of the search algorithm is reached, estimate parameter uncertainties by computing the change in Z as $\{\theta\}$ is varied about its optimum values.

We expect that this algorithm will be useful for the common situation where increased spectral resolution at high frequencies, parameterized by a larger g value, is partially offset by reduced signal to noise. This is commonly the case for simultaneous multifrequency fits of ESR spectra[17]. In fact, this work was partially motivated by the need to account for frequency-dependent parameter sensitivities in the presence of frequency-dependent noise. We also note that the canonical distribution defined by our procedure imposes a geometry on parameter space due to the normalization constraint and the constraint on the mean squared residual. In our view, the following observation, loosely paraphrased from Murray and Rice, is apt[18]: *Depicting any coordinate system in a Cartesian way implies a Cartesian geometry, but few people take that geometry seriously. Once you differentiate vector fields or compute Taylor series expansions in the usual way, you have taken that geometry very seriously even if you don't realize it.* Based on preliminary work[1], we expect that the constraints imposed by the PDF will lead to improved estimates of errors and we are actively exploring this possibility for simultaneous multifrequency fits. In particular, we expect that curvature corrections to the Hessian used for updating parameter searches will be especially important[19]. The

work of Sepulchre and co-workers will also be relevant to algorithm development that takes account of the intrinsic geometry defined by the PDF[20].

Finally, although the examples given here have been limited to applications in magnetic resonance, the approach is generic and should be applicable to any problem for which a spectral function and its derivatives are available. Space constraints have limited us to only presenting an overview of current work, but a more complete account with applications is in preparation[2].

ACKNOWLEDGMENTS

KAE thanks Kevin H. Knuth and Ariel Caticha for many useful discussions. This work was partially supported by a Faculty Research Awards Program grant from the University at Albany. The ACERT center at Cornell University is thanked for the use of its computational resources.

REFERENCES

1. K. A. Earle, L. Mainali, I. Dev Sahu, and D. J. Schneider, *J. Magn. Reson.* **37**, 865–880 (2010).
2. K. A. Earle, L. Mainali, I. D. Sahu, and K. H. Knuth, Estimating Parameter Sensitivity from Spectral Amplitudes, *in preparation* (2010).
3. E. T. Jaynes, *Physical Review* **106**, 620–630 (1957).
4. E. T. Jaynes, *Physical Review* **108**, 171–190 (1957).
5. B. Blümich, *Progress in NMR Spectroscopy* **19**, 331–417 (1987).
6. M. Fuhs, T. Prisner, and K. Möbius, *J. Magn. Reson.* **149**, 67–73 (2001).
7. A. Abragam, *Principles of Nuclear Magnetism*, Oxford Science Publications, 1982.
8. D. J. Schneider, and J. H. Freed, "Calculating Slow Motional Magnetic Resonance Spectra: A User's Guide," in *Spin Labeling: Theory and Applications, Vol. III*, edited by L. J. Berliner, and J. Reuben, Plenum, 1989, vol. 8 of *Biological Magnetic Resonance*, pp. 1–76.
9. D. J. Schneider, and J. H. Freed, "Spin Relaxation and Motional Dynamics," in *Lasers, Molecules and Methods*, edited by R. E. W. J. O. Hirschfelder, and R. D. Coalson, Wiley, 1989, vol. 73 of *Adv. Chem. Phys.*, chap. 10.
10. S. Amari, and H. Nagaoka, *Methods of Information Geometry*, vol. 191 of *Translations of Mathematical Monographs*, American Mathematical Society, 2000.
11. A. Ben-Naim, *A Farewell to Entropy: Statistical Thermodynamics Based on Information*, World Scientific, 2008.
12. R. K. Pathria, *Statistical Mechanics*, Pergammon Press, 1972.
13. R. F. Streater, *Statistical Dynamics: A Stochastic Approach to Nonequilibrium Thermodynamics*, Imperial College Press, 2009, second edn.
14. S. R. de Groot, and P. Mazur, *Non-Equilibrium Thermodynamics*, North-Holland, 1962, reprint edition of 1963 edn.
15. D. S. Sivia, and J. Skilling, *Data Analysis: A Bayesian Tutorial*, Oxford University Press, 2006, second edn.
16. F. Reif, *Foundations of Statistical and Thermal Physics*, McGraw-Hill, 1965.
17. Z. Zhang, M. R. Fleissner, D. S. Tipikin, Z. Liang, J. K. Moscicki, K. A. Earle, W. L. Hubbell, and J. H. Freed, *J. Phys. Chem. B* **114**, 5503–21 (2010).
18. M. K. Murray, and J. W. Rice, *Differential Geometry and Statistics*, Chapman and Hall/CRC, 1993.
19. H. Shima, *The Geometry of Hessian Structures*, World Scientific, 2007.
20. P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008.