AIP | Review of Scientific Instruments

Note: Multivariate system spectroscopic model using Lorentz oscillators and partial least squares regression analysis

R. S. Gad, J. S. Parab, and G. M. Naik

Citation: Review of Scientific Instruments **81**, 116103 (2010); doi: 10.1063/1.3499359

View online: http://dx.doi.org/10.1063/1.3499359

View Table of Contents: http://scitation.aip.org/content/aip/journal/rsi/81/11?ver=pdfcov

Published by the AIP Publishing

A WHOLE NEW LOOK FOR *PHYSICS TODAY* MAGAZINE COMING JANUARY 2016
1948
1964
1982
1985
1992
1997
2006
2012
2016

Note: Multivariate system spectroscopic model using Lorentz oscillators and partial least squares regression analysis

R. S. Gad,[a)] J. S. Parab, and G. M. Naik
Department of Physics, Electronic Section, Goa University, Goa 403206, India

(Received 25 June 2010; accepted 20 September 2010; published online 30 November 2010)

Multivariate system spectroscopic model plays important role in understanding chemometrics of ensemble under study. Here in this manuscript we discuss various approaches of modeling of spectroscopic system and demonstrate how Lorentz oscillator can be used to model any general spectroscopic system. Chemometric studies require customized templates design for the corresponding variants participating in ensemble, which generates the characteristic matrix of the ensemble under study. The typical biological system that resembles human blood tissue consisting of five major constituents i.e., alanine, urea, lactate, glucose, ascorbate; has been tested on the model. The model was validated using three approaches, namely, root mean square error (RMSE) analysis in the range of ±5% confidence interval, clerk gird error plot, and RMSE versus percent noise level study. Also the model was tested across various template sizes (consisting of samples ranging from 10 up to 1000) to ascertain the validity of partial least squares regression. The model has potential in understanding the chemometrics of proteomics pathways. © *2010 American Institute of Physics.* [doi:10.1063/1.3499359]

Most of spectroscopic methods are differentiated as either atomic (or molecular) based on whether or not they apply to atoms (or molecules). Along with that distinction, they can be classified on the nature of their interaction. Absorption spectroscopy is well expressed by Lambert–Bouguer law $I=I_0 10^{-\varepsilon cl}$ where ε is molar absorption coefficient, c a concentration, and l the thickness of sample. Another attractive technique is the reflectance measurement where the reflectivity of an absorbing media in air is given by $R=(n-1)^2+k^2/(n+1)^2+k^2$ where k is extinction coefficient of the sample and n is its refractive index. However, the mirror type reflection is difficult for detection and is generally not used in sensor-based instrumentation; hence most often diffuse reflectance is used in sensor-based instrumentation.[1] The probing light is diffusely reflected and passed through the sensitive analytes material. Hence the reflected light decreases with the increase in the absorption coefficient. One widely used model for diffuse light spectroscopy is that of Kubelka–Munk[2] and is expressed as $F(R)=[(1-R)^2/2R]=K/S$ where R is reflected light, S is the scattering coefficient, and K is the absorption coefficient.

Most of the spectroscopic techniques work with some type of indicator X, which is changing following interaction with the analyte A as $[X]+[A]\leftrightarrow[X-A]$ which can be solved as $[A]\propto[X]/[X-A]$. There are not many ratiometric indicators. However, any intensity measurements can be converted to ratiometric measurements if mixtures of two luminophores are used. Most of these models are valid under certain limited conditions,[3] as light entering the medium must be monochromatic and perfectly collimated, and the medium itself must be purely and uniformly absorbing. Therefore, certain errors will arise when applying the law to practical spectroscopic measurements (for example, even lasers are not perfectly monochromatic).

Understanding optical imaging of biological tissue is a real challenge due to the predominance of light scattering as incident photons propagate into the tissue.[4] There are many constituents in biological tissues which absorb light radiation, collectively known as tissue chromophores, each of which has its own unique spectrum. As expressed in Eq. (1) the total extinction coefficient k of a mixture of compounds is equal to the sum of their individual extinction coefficients, weighted by their relative concentrations. Hence one can approximate a biological tissue as a homogeneous mixture of compounds; the overall light absorption in tissue at a given wavelength depends upon the type and concentration of chromophores present thereby giving rise to multivariate system.[5] Thus in a ensemble containing a mixture of n absorbing constituents, the total absorption is the sum of the individual extinction coefficients multiplied by the distance l as given in Eq. (1).

$$\varepsilon cl = kl = (\varepsilon_1 c_1 + \varepsilon_2 c_2 + \varepsilon_3 c_3 + \ \ldots\ \ldots\ . + \varepsilon_n c_n)l. \qquad (1)$$

System identification is defined as problem of identifying nonlinear model structure which involves diverse characteristics as linearity, degree of nonlinearity, model structure, performance, and model validation, which has to be considered. Measures of long-term prediction error, for example, have been used in combination with genetic algorithms for on-line parameter identification,[6–9] as an alternative to recursive least-squares methods. More recently, Koza[10] have used a tree-structure representation, which is based on an input-output model, to represent these systems. Also a nonlinear difference equation model known as nonlinear autoregressive moving average with exogenous input (NARMAX) model was introduced by Billings and Chen.[11] Wide class of nonlinear systems has been represented using NARMAX which is given in Eq. (2).

[a)]Author to whom correspondence should be addressed. Electronic mail: rsgad@unigoa.ac.in.

0034-6748/2010/81(11)/116103/3/$30.00

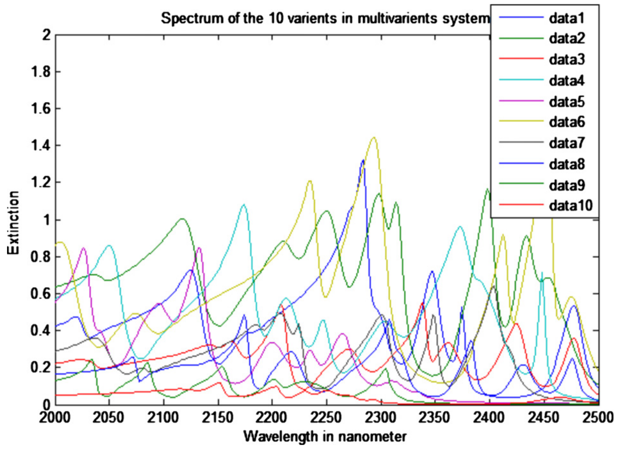

FIG. 1. (Color online) Spectra's of various concentrations for variants (data "1–10").

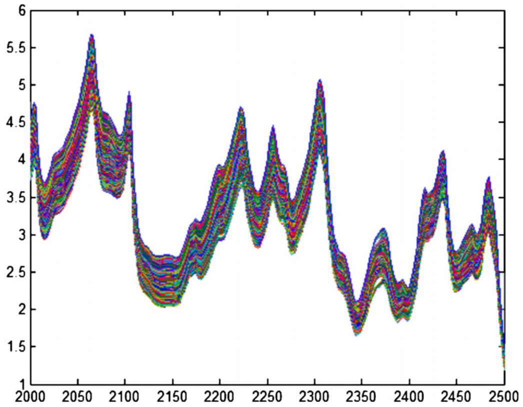

FIG. 2. (Color online) Template spectra for various strengths, central frequencies, and width for ten oscillators.

$$y(t) = F[y(t-1), \ldots \ldots, y(t-n_y), u(t-1), \ldots \ldots, u(t - n_u), e(t-1), \ldots \ldots, e(t-n_e)] + e(t), \qquad (2)$$

where $u(t)$, $y(t)$, and $e(t)$ represent input, output, and noise at time t, respectively. In practice, the nonlinear function F of Eq. (2) can be approximated,[12] for example, by a polynomial expansion of a given degree. Doing so, one obtains the representation as given in Eq. (3), where $n=n_y+n_u+n_e$, all b represent scalar coefficients, and all $x(t)$ represent lagged terms in u, y, or e.

$$y(t) = b_0 + \sum_{i_1=1}^{i=n} b_{i_1} x_{i_1}(t) + \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} b_{i_1 i_2} x_{i_1}(t) x_{i_2}(t) + \ldots \ldots + \sum_{i=1}^{n} \cdots \sum_{i_l=i_{l-1}}^{n} b_{i_1 \ldots\ldots i_l} x_{i_1}(t) \ldots x_{i_1}(t) + e(t). \qquad (3)$$

Since significance of the higher order terms is negligible, Eq. (3) can be simplified as Eq. (4). Equation (4) clearly belongs to linear regression model.

$$y(t) = \sum_{i=1}^{M} b_i x_i(t) + \xi(t); \quad t = 1, \ldots \ldots N, \qquad (4)$$

Modeling experiments are different in nature then optimizing experiments, since quantitative parameters are required from the experiments and also a good knowledge of the system. Under such circumstances it might be correct to say that the model is within 95% likelihood. In order to do this it is first necessary to propose a model. For a three factor experiment a quadratic model is given in Eq. (5).

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_1 x_1 + b_{11} x_1^2 + b_{22} x_2^2 + b_{33} x_3^2 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3. \qquad (5)$$

One sees three cross product terms in Eq. (5), which corresponds to interaction of variants. Further for five factors experiment quadratic model has ten cross products interact terms. The explosive growth of the number of possible monomial terms with the degree of nonlinearity and the order (or maximum lag) assumed for the model implies that even relatively small values of degree (say ≤ 3) and lag (say nu, $ny \leq 5$, and $ne = 0$) would result in a model too complex to be useful in practice $\left[\binom{5+5+3}{3} = 286 \text{ terms}\right]$. Normally regression methods are used to determine the coefficient b from the observed response y at different level of x. The form of model is important and dictates the minimum number of experiments required. If the model consists of too few terms, then the experiments might not properly describe the system. If the model consists of too many terms, many more experiments than necessary would be performed. The situation can be ameliorated if the total set contains a wide range of terms but, it has been found that, if the number of terms is increased, the identification process can become unnecessarily time consuming. The genetic programming approach is applied to the identification of nonlinear system polynomial model and provides a trade-off between the complexity and the performance of the models. Even though the genetic programming approach is also restricted by the number of nodes permissible in a tree, the search space is still extremely large and its variable size and dynamic representation provides diversification in the population.[8] Hence many times the system is described by using Gaussian and Lorentz oscillators.

Beyond modeling glasses and other disordered materials, the Gaussian can be applied to many different types of materials. The primary advantage of the Gaussian is that it rapidly approaches zero beyond $En \pm Br$. This extremely useful characteristic makes the Gaussian an all-purpose oscillator that can model materials which are transparent over a limited portion of the measured spectral range. This behavior is quite different from the Lorentz oscillator (LO), where it decreases slowly.[13] The LO is a useful tool for modeling complex optical properties.[14] Using the quantum mechanical form of the LO as expressed in Eq. (6), and allowing for the construction of models involving multiple oscillators, one can build models and determine complex optical properties. The spectra of

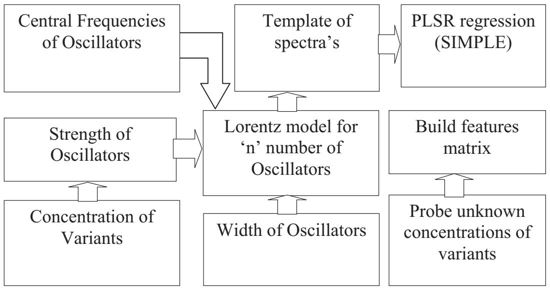
Central Frequencies of Oscillators
Template of spectra's
PLSR regression (SIMPLE)
Strength of Oscillators
Lorentz model for 'n' number of Oscillators
Build features matrix
Concentration of Variants
Width of Oscillators
Probe unknown concentrations of variants

FIG. 3. Block diagram of the spectroscopic model.

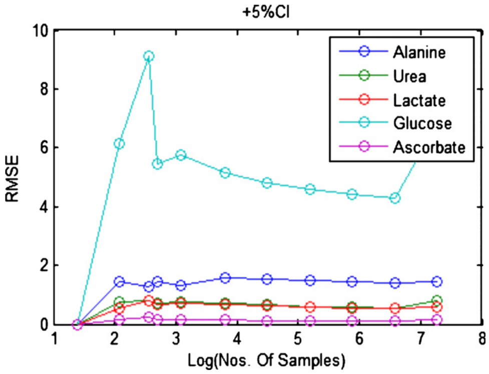

FIG. 4. (Color online) RMSE analysis of five variants over 1000–4 sample (±5% CI).

ten variant complex properties generated are shown in Fig. 1.

$$y = (n + i\,.\,k) = \left(\varepsilon_0 + \sum_j \frac{Sj\upsilon j2}{\upsilon_j^2 - \upsilon^2 - i\upsilon\Gamma j} \right)^{1/2} \qquad (6)$$

Where n is frequency of generated spectrum; $i.k$ is imaginary components of the frequency; Sj is strength of the oscillators; υj is central frequency; Γj is width of oscillator all in wavenumber (cm^{-1}) unit, and ε_0 represents the electronic contribution to the complex dielectric constant. Applying Eq. (1) for these variants one can generate templates of samples for range of concentrations as indicated in Fig. 2.

The four sets of templates (as shown in Fig. 2) generated using spectroscopic model (as shown in Fig. 3), for ten oscillators were studied. Here the concentrations of variants were within the confidence interval, which are further multiplied with the strength of oscillators to give extinction coefficients at particular wavelength of absorption. The templates of spectra generated are solved through the SIMPLE algorithm of partial least squares regression to obtain the features matrix, which is characteristics matrix of the template under study.

The root mean square error (RMSE) analysis was performed for the five templates in and about ±5% confidence interval (CI) for 10–1000 samples template and the results were satisfactory (i.e., within the one unit). This model was confirmed for template consisting of 13 experimental samples.[15] Human blood tissue was selected because the chemometrics of this tissue is very important in design of complex application such as noninvasive glucometer.[16–18] These analyses were performed using PARLES (Ref. 19) software. In RMSE analysis the glitch in the graph (Fig. 4) at the 2.5 (corresponding to "log 13" at natural base) is due to an experimental samples error.

Further the "RMSE" versus "percent noise level" study for each variant over 5 and 300 template samples (Fig. 5), indicate that the RMSE is within 0.2 and increases further to 0.5 for noise level of 50% and above. It is found that the model behavior was uncertain for noise level above 50%. The four variants concentrations (c1, c2, c4, and c6 as indicated in italics in Fig. 5) selected were higher in the magnitude (almost first and second orders) hence their RMSE values were dominating the curve as indicated in Fig. 5, this indicates that model has capability in predicting overlapped signatures across the order of magnitude.

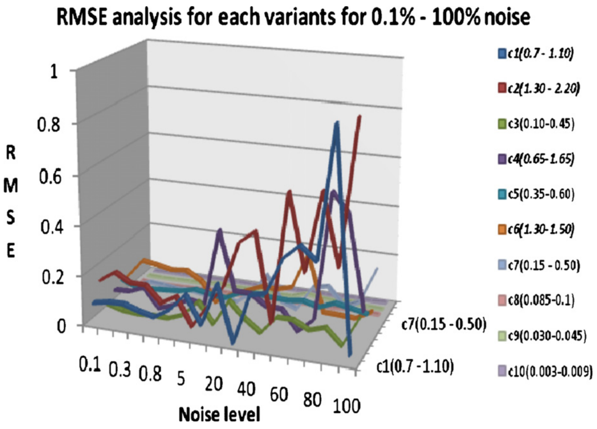

FIG. 5. (Color online) RMSE analysis for each variant over noise level of 0.1% to 100% for four samples template.

Authors would like to acknowledge financial assistance from Indian Council of Medical Research (ICMR, New Delhi) and University Grant Commission (UGC, New Delhi). We are grateful to Dr. Raphael Viscarra Rossel, soil scientist (Pedometrics), CSIRO, Land and Water; Canberra ACT 2601 Australia for providing PARLES software.

[1] Y. Kostov and G. Rao, Rev. Sci. Instrum. **71**, 4361 (2000).
[2] A. J. Gurthrie, R. Narayanaswamy, and D. A. Russel, Analyst (Cambridge, U.K.) **118**, 457 (1988).
[3] M. Cope, P. van der Zee, M. Essenpreis, S. R. Arridge, and D. T. Delpy, *Data Analysis Methods for Near Infrared Spectroscopy of Tissue: Problems in Determining the Relative Cytochrome Concentration* (SPIE, Bellingham, 1991), Vol. 1431.
[4] J. C. Smith, J.-P. Lambert, F. Elisma, and D. Figeys, Anal. Chem. **79**, 4325 (2007).
[5] H. Martens and T. Næs, *Multivariate Calibration*, 2nd ed. (Wiley, New York, 1991).
[6] K. Kristinsson and G. A. Dumont, IEEE Trans. Syst. Man Cybern. **22**, 1033 (1992).
[7] H. Oakley, *Advances in Genetic Programming*, Two Scientific Applications of Genetic Programming: Stack Filters and Linear Equation Fitting to Chaotic Data, edited by K. E. Kinnear, Jr. (MIT Press, Cambridge, 1994), pp. 369–389.
[8] C. M. Fonseca and P. J. Fleming, Evol. Comput. **3**, 1 (1995).
[9] C. M. Fonseca and P. J. Fleming, Proceedings of the 13th World Congress of IFAC, San Francisco 1996, pp. 187–192.
[10] J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT Press, Cambridge, 1992).
[11] S. A. Billings and S. Chen, Int. J. Control **50**, 1897 (1989).
[12] I. J. Leontaris and S. A. Billings, Int. J. Control **41**, 311 (1985).
[13] D. De Sousa Meneses, M. Malki, and P. Echegut, J. Non-Cryst. Solids **352**, 769 (2006).
[14] F. Wooten, *Optical Properties of Solids* (Academic, New York, 1972).
[15] J. S. Parab, R. S. Gad, and G. M. Naik, J. Appl. Phys. **107**, 104701 (2010).
[16] M. Ren and M. A. Arnold, Anal. Bioanal. Chem. **387**, 879 (2007).
[17] A. K. Amerov, G. W. Small, and M. A. Arnold, Proc. SPIE **6007**, 180 (2005).
[18] S. Pan, H. Chung, M. A. Arnold, and G. W. Small, Anal. Chem. **68**, 1124 (1996).
[19] R. A. Viscarra Rossel, Chemom. Intell. Lab. Syst. **90**, 72 (2008).