AIP | Journal of Applied Physics

Noninvasive glucometer model using partial least square regression technique for human blood matrix

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

Citation: Journal of Applied Physics **107**, 104701 (2010); doi: 10.1063/1.3380850
View online: http://dx.doi.org/10.1063/1.3380850
View Table of Contents: http://scitation.aip.org/content/aip/journal/jap/107/10?ver=pdfcov
Published by the AIP Publishing

Noninvasive glucometer model using partial least square regression technique for human blood matrix

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

(Received 9 January 2010; accepted 8 March 2010; published online 21 May 2010)

In this article, we have highlighted the partial least square regression (PLSR) model to predict the glucose level in human blood by considering only five variants. The PLSR model is experimentally validated for the 13 templates samples. The root mean square error analysis of design model and experimental sample is found to be satisfactory with the values of 3.459 and 5.543, respectively. In PLSR templates design is a critical issue for the number of variants participating in the model. Ensemble consisting of five major variants is simulated to replicate the signatures of these constituents in the human blood, i.e., alanine, urea, lactate, glucose, and ascorbate. Multivariate system using PLSR plays important role in understanding chemometrics of such ensemble. The resultant spectra of all these constituents are generated to create templates for the PLSR model. This model has potential scope in designing a near-infrared spectroscopy based noninvasive glucometer.
 [doi:10.1063/1.3380850]

I. INTRODUCTION

In conventional methods of measuring blood glucose concentration, biochemical reagents or enzyme are used to react with blood or plasma to show concentration qualitatively or quantitatively. But these methods have disadvantages such as the wastage of reagent, long measuring period, painful, it might lead to infection and the possible pollution.

Although the noninvasive glucose measurement technique based on near-infrared (NIR) spectroscopy has been an active research area for over twenty years, a reliable monitoring method has not been established yet.[1] The key problem here is that the spectral variations due to glucose concentration are extremely small compared to that from other biological components. In addition, there are also some time-dependent physiological processes, which make the explanation of the model more difficult, especially in the universal calibration. In this paper, we have developed a partial least square regression (PLSR) model to predict the glucose content in blood over a NIR range (2.0–2.5 μm). NIR spectroscopy offers several advantages over mid-infrared spectroscopy in a noninvasive glucose monitoring system, e.g., less background interference due to water absorption as shown in Fig. 1, absorbance due to skin is negligible in that range and penetration depths are greater at the shorter wavelengths, which is necessary for monitoring blood glucose in capillaries and glucose in interstitial fluid and tissue.

Here we have developed a baseline calibration model for the NIR spectral data using PLSR regression technique. PLSR technique which is generally used for analysis in econometrics and social sciences has been found to be excellent method for the determination of concentration of blood analytes such as glucose, ascorbate, lactate, etc. The basic problem is to fit a calibration model to empirical data and use this model to predict certain quantities given a set of test data as input to the calibration model after the training phase. Applying these techniques to predict glucose concentrations from NIR spectroscopic blood serum data, a set of calibration (training) data is formulated from the collected NIR spectroscopic absorption data using the significant portion of the spectra that contains absorption bands associated with glucose.[2] Another set of data is formulated as the reference (target) data, which consists of reference glucose concentration values. In the PLS method, both the absorption data and the concentration data are used at the outset to derive a calibration model.[3] An exhaustive explanation of PLSRwill not be presented, however, an overview of the basic elements of PLS. The details about PLS algorithm used is explained in Sec. II.

II. PARTIAL REGRESSION ANALYSIS FOR MULTIVARIATE SYSTEM

PLSR is an extension of the multiple linear regression models. In its simplest form, a linear model specifies the relationship between a dependent variable "Y'" and a set of predictor variables, the "X's," so that

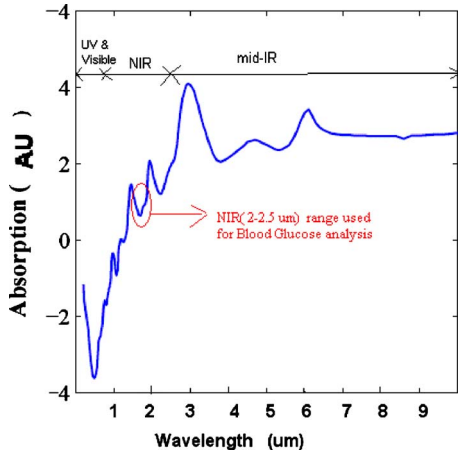

FIG. 1. (Color online) Water Absorbance in AU.

[a]Electronic mail: jsparab@unigoa.ac.in.

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots \ldots + b_pX_p. \quad (1)$$

In Eq. (1), "b_0" is the intercept and the "b_i" values are the regression coefficients (for variants "X_1" to "Xp") computed from the data sets with known constant concentrations. The multiple linear regression model serves as the basis for a number of multivariate methods such as *discriminant analysis* (DA) (i.e., the prediction of group membership from the levels of continuous predictor variables), *principal components regression* (PCR) (i.e., the prediction of responses on the dependent variables from factors underlying the levels of the predictor variables), and *canonical correlation* (CC) (i.e., the prediction of factors contributing responses of the dependent variables from factors underlying the levels of the predictor variables). These multivariate methods have two important properties in common. These methods impose restrictions such that (1) factors underlying the X and Y variables are extracted from the $\boldsymbol{X'X}$ and $\boldsymbol{Y'Y}$ matrices, respectively, (where $\mathbf{X}'$ is a transpose of $\mathbf{X}$ and $\mathbf{Y}'$ is a transpose of $\mathbf{Y}$), and never from cross-product matrices involving both the $\mathbf{X}$ and $\mathbf{Y}$ variables and (2) the number of prediction functions can never exceed the minimum of the number of $\mathbf{X}$ variables and $\mathbf{Y}$ variables.[4]

PLSR extends multiple linear regressions without imposing the restrictions imposes by DA, PCR, and CC. In PLSR, prediction functions are represented by factors extracted from the $\boldsymbol{Y'XX'Y}$ matrix.

PLSR has been used in various disciplines such as chemistry, economics, medicine, psychology, pharmaceutical, and medical science where predictive linear modeling, especially with a large number of predictors is necessary. PLSR has become a standard tool for modeling linear relations between multivariate measurements in chemometrics.[5]

A. Computational approach for basic PLSR model

As in multiple linear regression, the main purpose of PLSR is to build a linear model, $\boldsymbol{Y=XB+E}$, where $\boldsymbol{Y}$ is an "n" cases by "m" variables response matrix, $\boldsymbol{X}$ is an "n" cases by "p" variables

$$Y = XB + E, \quad (2)$$

predictor matrix, $\boldsymbol{B}$ is a "p" by "m" regression coefficient matrix and $\boldsymbol{E}$ is a noise term for the model which has the same dimensions as that of $\boldsymbol{Y}$. Usually, the variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered by subtracting their means and scaled by dividing by their standard deviations.[6,7]

Both PCR and PLSR produce factor scores as linear combinations of the original predictor variables, so there is no correlation between the factors score variables used in the predictive regression model. Let us consider we have a data set with response variables $\boldsymbol{Y}$ and a large number of predictor variables $\boldsymbol{X}$ and some of which are highly correlated. A regression using factor extraction method for this type of data, generates the factor score matrix $\boldsymbol{T=XW}$, where $\mathbf{W}$ is a weight matrix with "p" by "c" weight.

$$T = XW. \quad (3)$$

For regression technique it can be proved that $\mathbf{B}$ has the form of

$$B = WQ, \quad (4)$$

where $\mathbf{Q}$ is a matrix of regression coefficient for $\mathbf{T}$ *of* "n" *by* "c." Substituting Eqs. (3) and (4) in Eq. (2), we get

$$Y = TQ + E. \quad (5)$$

PCR and PLSR differ in the methods used in extracting factor scores. In short, PCR produces the weight matrix $\boldsymbol{W}$ reflecting the covariance structure between the predictor variables, whereas PLSR produces the weight matrix $\boldsymbol{W}$, reflecting the covariance structure between the predictor and response variables.

One additional matrix which is necessary for a complete description of PLSR procedures is the "p" by "c" factor loading matrix $\boldsymbol{P}$, which gives a factor model

$$X = TP + F, \quad (6)$$

where $\boldsymbol{F}$ is the unexplained part of the $\boldsymbol{X}$ scores.

We can now describe the algorithms for computing PLSR as given below. The PLSR algorithms we have used for prediction are SIMPLE algorithm.

B. SIMPLE algorithm

The standard algorithms for computing PLSR components (i.e., factors) is nonlinear iterative partial least-squares and SIMPLE algorithm. We have used the SIMPLE algorithm for the purpose of Multivariate analysis.

For each $h=1,\ldots,c$, where $\boldsymbol{A_0=X'Y}$, $\boldsymbol{M_0=X'X}$, $\boldsymbol{C_0=I}$, and c given

(1) Compute $\boldsymbol{q_h}$, the dominant eigenvector of $\boldsymbol{A_h'A_h}$.
(2) $\boldsymbol{w_h=A_hq_h}$, $\boldsymbol{c_h=w_h'M_hw_h}$, $\boldsymbol{w_h=w_h/sqrt(c_h)}$, and store $\boldsymbol{w_h}$ into $\boldsymbol{W}$ as a column.
(3) $\boldsymbol{p_h=M_hw_h}$ and store $\boldsymbol{p_h}$ into $\boldsymbol{P}$ as a column.
(4) $\boldsymbol{q_h=A_h'w_h}$ and store $\boldsymbol{q_h}$ into $\boldsymbol{Q}$ as a column.
(5) $\boldsymbol{v_h=C_hp_h}$ and $\boldsymbol{v_h=v_h/\|v_h\|}$.
(6) $\boldsymbol{C_{h+1}=C_h-v_hv_h'}$ and $\boldsymbol{M_{h+1}=M_h-p_hp_h'}$.
(7) $\boldsymbol{A_{h+1}=C_hA_h}$.

Where $\mathbf{Q}'$ is transpose of $\mathbf{Q}$. Hence using the $\mathbf{B}$ and $\mathbf{Q}$ matrix obtained from above one can compute the $\mathbf{Y}$ or $\mathbf{X}$ matrix to complete the model.

C. Generalized multivariate model of the system

We have generated required generalized response and predictor matrix using Lorentz oscillator Eq. (7), with respective oscillator strength, width, and central frequency.

$$(n + ik) = \left(\varepsilon_\infty + \sum_j \frac{S_j{\nu_j}^2}{{\nu_j}^2 - \nu^2 - i\nu\Gamma_j}\right)^{1/2}, \quad (7)$$

where "n" is response of spectrum for frequency "v," "ik" is imaginary components of the frequency response, "Sj" is a strengths of the oscillators, "vj" is central frequency, "Γj" is width of oscillator all in wavenumber (cm^{-1}) unit, and "ε_∞" represents the electronic contribution to the complex dielectric constant. From Eq. (7) one can collect only the real part of the frequency components from RHS and generated spectra over the region of interest. The flow chart of the same is given in Fig. 2.

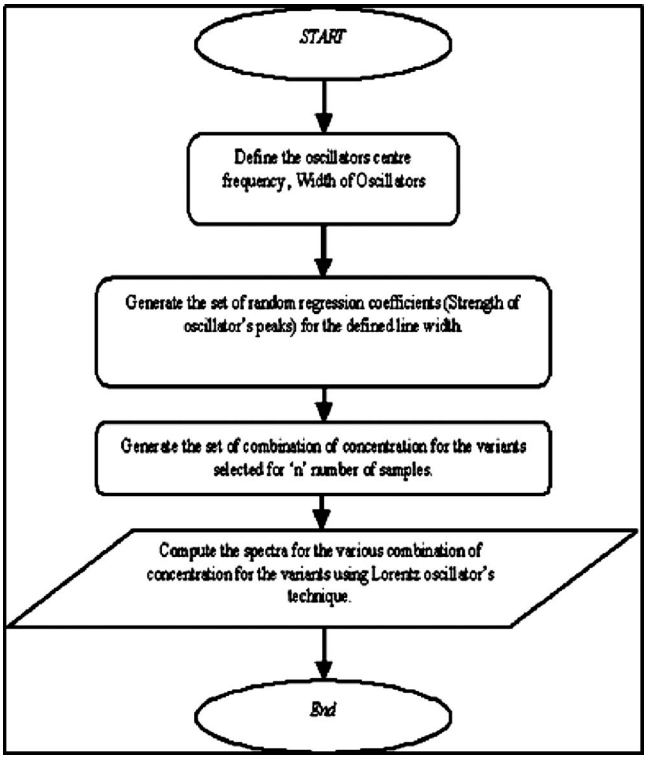

FIG. 2. Flow chart of the simulated spectra generated by Lorentz oscillator model.

III. MULTIVARIATE MODEL OF THE HUMAN WHOLE BLOOD TISSUE FOR GLUCOSE

Methods of noninvasive optical diagnosis involves various issues which is described above. The absorption spectrum depends on the type of predominant absorption center and water content of tissue. Absolute value of absorption coefficients for typical tissue lie in the range of 10^{-2} to 10^{14} cm^{-1}.[8–11] Glucose is one of the most important carbohydrate nutrient sources and is fundamental to almost all biological processes. Quantification of glucose concentration is important in monitoring and analysis of control and regulation of cell culture, and diagnosis and control of human disease including diabetes. The NIR region of the optical spectrum extend from 700–2500 nm which is used for quantitative measurement of organic functional groups, specially C—H, O—H, N—H, and C═O. The absorption bands in the NIR are composed primarily of overtones and combination bands of stretching and vibrational modes of organic molecules. Major seven constituents interfering with glucose in the region of 2000–2500 nm are serum alanine, serum urea, bovine serum albumin (BSA), high density lipoprotein (HDL) lactate, glucose, ascorbate, and triacetin, interfere with the whole blood glucose to generate a complex signature in the spectrum region.[22] Also from the literature it is found that tissue temperature and skin complexion due to pigment variation over the globe population has influence on the transmission characteristics. There are many other factor influencing the signature of a spectrum but the analysis becomes complex as the physiology of the body is very sensitive and dynamic over the catabolic processes involved. It has been decided to model this complex ensemble over various concentrations of five chemical constituents and two physical components namely skin complexion and the body temperature.[12–21]

TABLE I. Various oscillators used in the Lorentz expression. [CF: Centre frequency; OW Oscillator width; OS Oscillator strength.]

Variants	Oscillator number	1	2	3	4	5
Alanine	CF	2130	2242	2295	2320	2530
	OW	150	25	30	30	50
	SO	230	30	20	51	40
Urea	CF	2080	2150	2200	2240	2550
	OW	40	50	50	140	250
	SO	1	3	10	16	12
HDL lactate	CF	2050	2150	2198	2258	2350
	OW	150	100	80	40	80
	SO	1.5	10	10	10	12
Glucose	CF	2100	2150	2250	2320	2480
	OW	100	100	60	60	100
	SO	1.4	0.6	0.5	0.5	0.6
Ascorbate	CF	2125	2160	2280	2340	2490
	OW	70	100	60	100	50
	SO	48	90	30	40	40

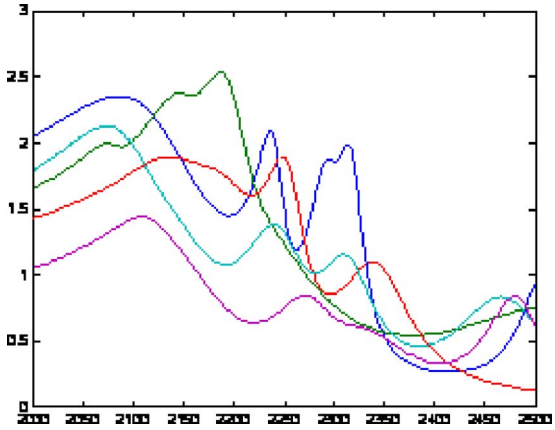

FIG. 3. (Color online) Signature of five major components simulated using Lorentz oscillators.

The whole blood spectrum model for various concentrations of the five variants has been generated. The model considered the following values of concentration for the human blood chemometrics system well within the pathological range. The concentrations of the five constituents were C1 =alanine (1–21 gm/dl), C2=serum urea (7–18 g/dl), C3 =HDL lactate (4.5–14.4 mgm/dl), C4=glucose (70–110 mg/dl), C5=ascorbate (0.4–1.5 gm/dl), and other two physical parameters were C6=temperature (25–40 °C) and C7 =skin complexion (0.2–0.4). Oscillator's parameters as shown in Table I were used to generate the signature of the seven chromospheres as shown in Fig. 3.

The Lorentz model is so flexible that just by varying the

TABLE II. RMSE analysis for glucose.

Sample	RMSE	+5% CI RMSE	−5% CI RMSE
Theoretical (15 sample)	3.459	2.534	5.448
Experimental (13 sample)	5.543	3.981	9.120

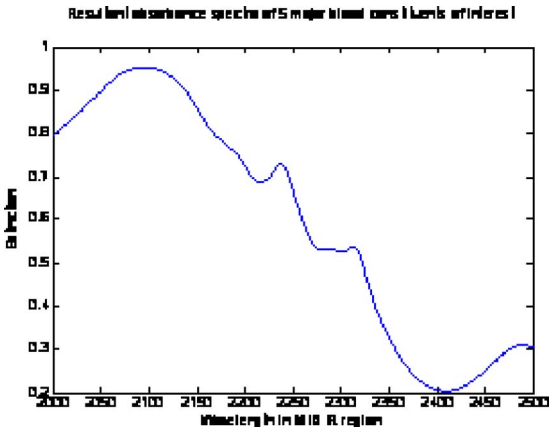

FIG. 4. (Color online) Resultant normalized signature of various components simulated using Lorentz oscillator.

strength, line width, and natural frequency any practical spectrum can be generated with highly nonlinear behavior. The unknown spectrum within the confidence interval of the calibrated spectra can be generated adopting same principle that of calibration spectral as shown in flow chart (Fig. 2). Model was validated with the PLSR software (PARLES software).[23] A Root mean square error (RMSE) analysis was performed over the prediction of the variants concentrations as shown in Table II. The resultant spectra of all blood constituents will have the form shown in Fig. 4. Samples template where generated for different combinations of blood constituents as shown in Fig. 5. We had taken five different blood constituents like glucose, alanine, ascorbate, urea, and lactate in different proportions and thirteen spectra's in the range 2.0–2.5 μm of were generated with the help of Schimatzu fourier transform infrared spectrophotometer (Fig. 6).

IV. RESULTS AND DISCUSSIONS

Major challenges involved in calibrating the noninvasive glucometer are weak signal in the transmission and reflectance mode with the interferences from approximately 118 constituents from the blood, varied skin complexion over the

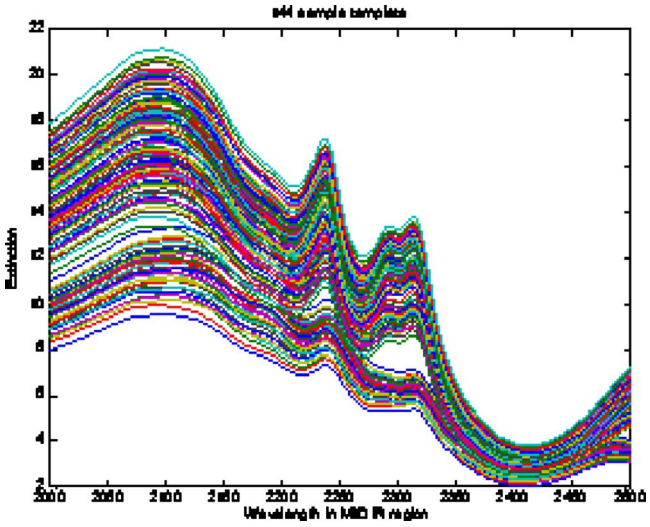

FIG. 5. (Color online) 1024 samples template for the PLSR model.

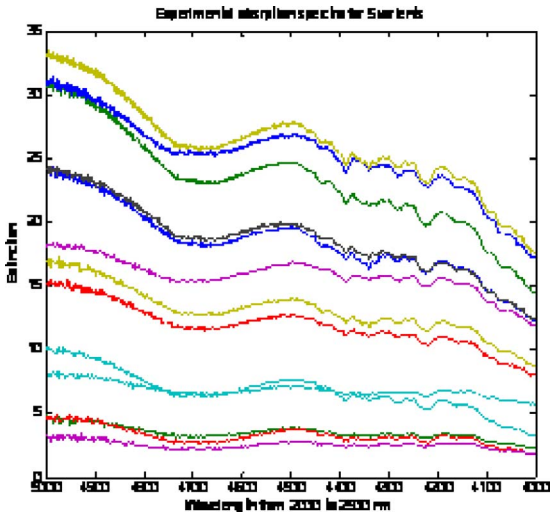

FIG. 6. (Color online) 13 solutions experimental template in nonhygroscopic mode for the PLSR model.

globe, temperature compensation, physiology dependent calibration, dependence of wavelength, and width of the absorber on the local chemical environment surrounding the bonds.

Most of the models of noninvasive glucose instrumentation have difficulty in satisfying the Clarke error grid[24] shown in Fig. 7. This could be due to poor multivariate model. The fine-tuning of the same multivariate model is required by incorporating the various parameters influencing glucose. With proper understanding of the dynamics of the physiology of the human body and the catabolic processes triggered by signal for the alternative path depending on the situation therein the cell, a closer estimate of glucose is possible.

In general, there are three basic types of absorption processes as follows: (1) electronic, (2) vibrational, and (3) rotational. Electronic transition occurs in both atoms and molecules, whereas vibrational and rotational transitions occur only in molecules. PLSR model is more difficult to build because of variation in skin pigments, body temperature, overlapping absorption spectra's of other blood analytes, etc. It may be noted that even a slight change in temperature, the

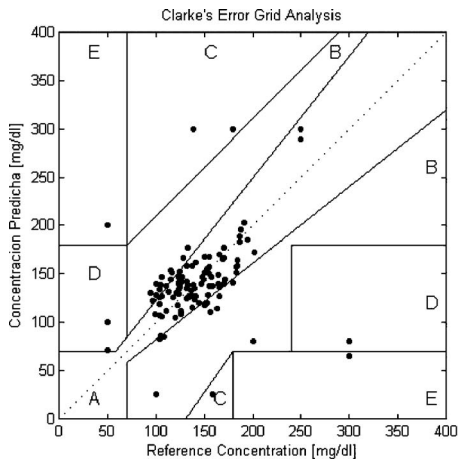

FIG. 7. Clarke error grid.

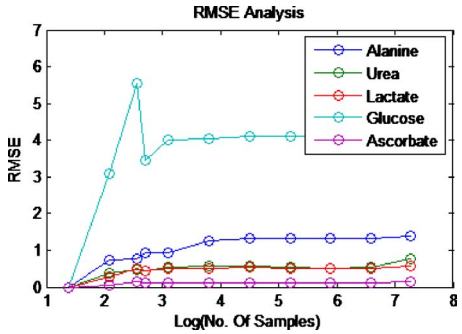

FIG. 8. (Color online) RMSE analysis of five variants over 1024–4 samples.

absorption of the background water spectrum will shift, severely impacting measurement of the glucose signal.[25–29]

The physiological ranges of glucose values seen in the normal human body range is from 80–110 mg/dl and should ideally remain around 100 mg/dl (5.5 mM). Required accuracy of a useful glucometer is 10 mg/dl (0.55 mM). For the most identifiable NIR glucose peak at approximately 2.27 μm. Considering transmission measurement made through 1 mm of body tissue. The background absorption due to water will be about 1.6 and that due to glucose will be about 1.26×10^{-4}. Further for the accuracy requirement, we must be able to discriminate an absorption change of about 1.26×10^{-4} on a background of 1.6. For this reason, high-order multivariate models that incorporate analysis must use entire spectra to extract NIR glucose information.[30,31]

Calibration of the multivariate model and validation of the result by incorporating the samples obtained from different patients population, i.e., different ages, sex type, ethnic and racial origins, blood group, cultural variation, daily diet habits, skin complexion corrections, etc. The model then should be calibrated to individual user.

The above discussion opens the scope for extending the low order multivariate model to high-order. The generalized model using Lorentz oscillators can play major role in validating the model for the calibration purpose and satisfying the Clarke error grid.

V. CONCLUSIONS

The designed glucometer chemometrics model has around $\pm5\%$ error (refer Table II), this error can be minimized if the intraconstituents and interconstituents chemistry is known by means of various pathways triggered by the ambient conditions in the process of catabolism of glucose.

The RMSE analysis of design model and experimental sample is found to be satisfactory with the values of 3.459 and 5.543, respectively, as shown in Fig. 8. This model has potential scope in designing a NIR spectroscopy based non-invasive glucometer. This model can be used to predict the concentration of any blood constituents.

ACKNOWLEDGMENTS

The authors would like to acknowledge financial assistance provided by Indian Council of Medical Research (ICMR, New Delhi).

[1]R. Liu, B. Deng, W. Chen, and K. Xu, Opt. Quantum Electron. **37**, 1305 (2005).
[2]F. M. Hamy, G. M. Cohenz, I. Kostanicy, and B. R. Goochx, *Physiological Measurement* (IOP, London, 1995), Vol. 37, pp. 1–20.
[3]M. Cope, P. van der Zee, M. Essenpreis, S. R. Arridge, and D. T. Delpy, Proc. SPIE **1431**, 251 (1991).
[4]H. Martens and T. Naes, *Multivariate Calibration*, 2nd ed. (Wiley, New York, 1991).
[5]S. De Jong, Chemom. Intell. Lab. Syst. **18**, 251 (1993).
[6]P. Geladi and B. R. Kowalski, Anal. Chim. Acta **185**, 1 (1986).
[7]K. Faber and B. R. Kowalski, J. Chemom. **11**, 181 (1998).
[8]M. P. Fuller, G. L. Ritter, and C. S. Draper, Appl. Spectrosc. **42**, 217 (1988).
[9]B. Chance, Annu. Rev. Biophys. Biophys. Chem. **20**, 1 (1991).
[10]*Handbook of Optical Biomedical Diagnostics*, edited by V. V. Tuchin (SPIE, Bellingham, WA, 2002), Vol. PM107.
[11]M. J. Goetz, Jr., G. L. Cote, W. E. March, R. Erckens, and M. Motamedi, IEEE Trans. Biomed. Eng. **42**, 728 (1995).
[12]H. A. MacKenzie, H. S. Ashton, S. Spiers, Y. Shen, S. S. Freeborn, J. Hannigan, J. Lindberg, and P. Rae, Clin. Chem. **45**, 1587 (1999).
[13]Y. C. Shen, E. H. Linfield, A. G. Davies, P. F. Taday, D. D. Arnone, and T. S. Elsey, J. Biol. Phys. **29**, 129 (2003).
[14]Y. C. Shen, A. G. Davies, E. H. Linfield, T. S. Elsey, P. F. Taday, and D. D. Arnone, Phys. Med. Biol. **48**, 2023 (2003).
[15]M. A. Arnold, L. Liu, and J. T. Olesberg, J. Diabetes Sci. Technol. **1**, 454 (2007).
[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]J. Chen, M. A. Arnold, and G. W. Small, Anal. Chem. **76**, 5405 (2004).
[19]M. A. Arnold, J. Am. Chem. Soc. **126**, 14678 (2004).
[20]S. Pan, H. Chung, M. A. Arnold, and G. W. Small, Anal. Chem. **68**, 1124 (1996).
[21]K. Schugerl, J. Biotechnol. **85**, 149 (2001).
[22]M. A. Arnold, J. J. Burmeister, and G. W. Small, Anal. Chem. **70**, 1773 (1998).
[23]R. A. Viscarra Rossel, Chemom. Intell. Lab. Syst. **90**, 72 (2008).
[24]H. Chung, M. A. Arnold, M. Rhil, and D. W. Mruhammer, Appl. Spectrosc. **50**, 270 (1996).
[25]A. M. K. Nilsson, G. W. Lucassen, and W. Verkruysee, Photochem. Photobiol. **65**, 366 (1997).
[26]I. Gabriely, R. Wozniak, M. Mevorach, J. Kaplan, Y. Aharon, and H. Shamoon, Diabetes Care **22**, 2026 (1999).
[27]M. R. Robinson, R. P. Eaton, D. M. Haaland, G. W. Koepp, E. V. Thomas, B. R. Stallard, and P. L. Robinson, Clin. Chem. **38**, 1618 (1992).
[28]J. S. Maier, S. A. Walker, S. Fantini, M. A. Franceschini, and E. Gratton, Opt. Lett. **19**, 2062 (1994).
[29]E. Angelopoulou, J. Electron. Imaging **11**, 5 (2001).
[30]G. L. McClure, *Computerized Quantitative Infrared Analysis* (American Society for Testing and Materials, Philadelphia, PA, 1987)
[31]H. Martens and T. Naes, *Multivariate Calibration* (Wiley, New York, 1989).