The efficacy of support vector machines in modelling deviations from the Beer-Lambert law for optical measurement of lactate *

M. Mamouei, *Member, IEEE*[1], K. Budidha, N. Baishya, M. Qassem and P. A. Kyriacou

***Abstract*— Lactate is an important biomarker with a significant diagnostic and prognostic ability in relation to life-threatening conditions and diseases such as sepsis, diabetes, cancer, pulmonary and kidney diseases, to name a few. The gold standard method for the measurement of lactate relies on blood sampling, which due to its invasive nature, limits the ability of clinicians in frequent monitoring of patients' lactate levels. Evidence suggests that the optical measurement of lactate holds promise as an alternative to blood sampling. However, achieving this aim requires better understanding of the optical behavior of lactate. The present study investigates the potential deviations of absorbance from the Beer-Lambert law in high concentrations of lactate. To this end, a number of nonlinear models namely support vector machines with quadratic, cubic and quartic kernels and radial basis function kernel are compared with the linear principal component regression and linear support vector machine. Interestingly, it is shown that even in extremely high concentrations of lactate (600 mmol/L) in a phosphate buffer solution, the linear models surpass the performance of the other models.**

I. Introduction

Lactate is an important fundamental biomarker. It plays an important role in the biochemical processes that lead to the extraction of Adenosine Triphosphate (ATP) from glycogen, namely cellular respiration. ATP is known as the universal energy currency of cells and is the primary energy carrier in all living organisms. Conditions that cause inefficient supply of oxygenated blood to the tissue are some of the common causes of imbalances in lactate levels. Moreover, diseases that affect lactate processing organs such as, the liver, lung and kidney can leave a mark on blood lactate levels. Therefore, not surprisingly abnormal lactate levels have been observed in patients suffering from ischemic stroke, cancer, cardiogenic, septic, obstructive and hypovolemic shock, kidney and lung diseases, to name a few [1, 2, 3, 4]. More broadly, lactate levels above 4 mmol/L are associated with increased likelihood of morbidity and mortality in critically ill patients [5]. Lactate is, therefore, an invaluable biomarker for diagnosis and prognosis of diseases.

In spite of the importance of lactate, its accurate measurement requires blood sampling. The invasive and logistically costly nature of blood sampling, limits the clinicians' ability to frequently monitor lactate levels particularly in intensive care units. The optical measurement of lactate would be a groundbreaking solution to this problem. Previous studies on the subject have shown promising results. In particular, it has been shown that lactate can be accurately quantified in plasma and using the mid-IR region of the optical spectrum; achieving a coefficient of determination, R_v^2, of 0.94 for a validation set and a Root Mean Square Error of Validation (RMSEV) of 0.15 mmol/L [6]. It has also been shown that accurate measurement of lactate in mid-IR region is achievable with highly parsimonious models, i.e. models that that only utilize narrow regions of spectrum and/or small number of wavelengths with $R_{CV}^2 = 0.996$ [7]. The use of the mid-IR region although appropriate for in-vitro applications is of little use in in-vivo applications due to the superficial penetration of mid-IR light into the skin. On the contrary Near InfraRed (NIR) light can effectively reach the microvascular bed of tissue in dermis and hypodermis [8]. The measurement of lactate using NIR spectra has also been successfully reported in whole human blood, $R_{cv}^2 = 0.96$ [9]. While these results lay down a promising foundation for the development of an accurate, optical lactate sensor, and while a non-invasive, continuous lactate sensor would be nothing short of groundbreaking, such a sensor does not yet exist. A better understanding of the optical behavior of lactate, its nonlinearities, interactions and overlaps in optical absorbance of lactate with other molecules, and inter-subject baseline differences are some of the areas that need further investigation. The present study, investigates the significance of absorbance nonlinearity in optical measurement of lactate.

The Beer-Lambert law describes a linear relationship between the absorbance of monochromatic light and the concentration of absorbing species. This is depicted in (1),

$$\log_{10} I_0/I = \epsilon\, cl. \tag{1}$$

where I_o is the intensity of incident beam, I is the intensity of transmitted beam, ϵ is the molar decadic extinction coefficient, c is the concentration of absorbing species and l is the path length of light.

The linearity postulated by the Beer-Lambert law along with the high-dimensional and multicollinear nature of the optical spectra justifies the choice of the widely used Principal Component Regression (PCR) and Partial Least Squares (PLS) in optical spectroscopy. However, high concentrations of analytes, scattering matrices, and non-monochromatic light can lead to non-negligible nonlinearities [10].

The present study investigates the significance of these nonlinear effects in relation to the prediction of lactate concentrations from NIR spectra. For this purpose PCR and linear Support Vector Machine (SVM) regression are compared with four nonlinear models, namely SVM with

*This research is funded by the Engineering and Physical Sciences Research Council (EPSRC).

[1] All authors are with the Research Centre for Biomedical Engineering (RCBE), School of Mathematics, Computer Science & Engineering, City, University of London, Northampton Square, London, EC1V 0HB, UK (phone: +44 (0) 20 7040 3878; e-mail: mohammad.mamouei@city.ac.uk).

quadratic, cubic, quartic, and Radial Basis Function (RBF) kernels.

II. Materials and methods

A. *The dataset*

A dataset consisting of 57 NIR spectra of different concentrations of lactate in a Phosphate Buffer Solution (PBS) was produced. The dataset contains 31 samples with concentrations of lactate between 0-10 mmol/L (increments of 0.25 mmol/L), 20 samples between 10.5-20 mmol/L (increments of 0.5 mmol/L) and finally, six samples with extremely high concentrations of 100-600 mmol/L (increments of 100 mmol/L). The procedure for the preparation of the solutions is described below.

Analytical grade Sodium L-lactate ($C_3H_5NaO_3$ - 98 + %) and isotonic PBS were acquired in dry form from Thermo Fisher Scientific (*Massachussetts, USA*). A stock solution of 600 mmol/L was prepared by dissolving 67.236 g of Na-lactate powder in 1 L of deionized water (*Deionised Water Company, UK*). A liter of aqueous PBS (1X) was made by dissolving 9.89 g of PBS 10x powder in a liter of deionized water. The lactate stock solution was then serially diluted with PBS to obtain the desired molar concentration of lactate. The concentrations were verified with three independent measurements using the LM5 lactate analyzer (*Analox Instruments Limited, Stourbridge, UK*). The temperature and the pH of the solutions were controlled for and maintained at 24 $°C$ and 6.5 ($\pm$ 0.2) respectively. The pH of the solutions were verified by Orion Star A211 Advanced pH Benchtop Meter (*Thermo Fisher Scientific, Massachusetts, USA*).

The acquisition of the NIR spectra were carried out using the Lambda 1050 dual beam spectrophotometer ($perkin\ Elmer\ Corp, Massuchusetts, USA$). The spectral resolution of 1 nm was chosen. The gain for the indium gallium arsenide (InGaAs) detector, active between 800-1800 nm, was set to 5. For the lead sulfide detector, active between 1800-2600 nm, the gain was 1. The response time of both detectors was set 0.2 seconds. The attenuation in the sample and reference beams were set to 100% and 1% respectively. Background noise baseline correction was performed at 100% transmission / 0% absorbance prior to the acquisition of the spectra.

The solutions were randomly selected (to prevent temporal bias) and transferred into a macro quartz cuvette (λ : 200 nm - 3500 nm) (*Hellma GmbH & Co.KG, Jena, Germany*) with a light path length of 1 mm and placed in the sample compartment of the spectrophotometer. An empty identical cuvette was placed in the reference compartment. Three spectra were obtained for each solution. These three spectra were then averaged to reduce the measurement noise. The samples were randomly chosen to prevent any temporal bias.

Fig. 1. a) demonstrates the raw optical spectra. Fig. 1. b) shows the mean-centered spectra with distinguishable absorption peaks related to lactate. In particular, the peaks between 1660-1780 nm pertain to the first C-H stretching overtone and the peaks between 2230-2230 nm pertain to the combination of C-H stretch with C-H bend [11].

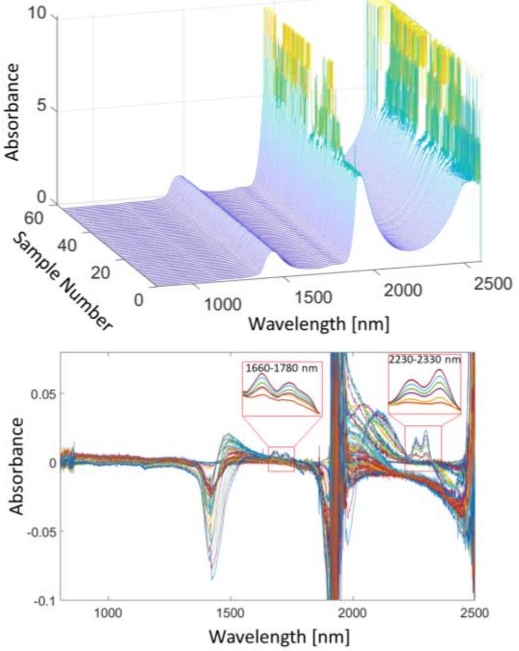

Figure 1. The lactate and PBS dataset a) raw spectra b) the spectrum of PBS is subtracted from the spectra to emphasize variations from the baseline caused by the increasing concentration of lactate in solutions. The small boxes show these peaks for the six high concentrations of lactate between 100-600 mmol/L. These six spectra are smoothed using a Savitzky–Golay filter for visualisation purposes.

B. *Preprocessing of spectra*

Two of the water absorption peaks, specifically wavelengths between 1900-1967 nm and 2450-2600 nm are affected by the over saturation of the lead sulfide detector and are removed. Subsequently, the spectra were processed using Multiplicative Scattering Correction (MSC) and a Savitzky-Golay filter with the window length of 135, second order polynomial and second order derivative.

C. *Dimensionality reduction*

Typically, in optical spectroscopy the number of features are significantly greater than the number of samples. This is known as the "large p, small n problem". However many wavelengths are collinear and many are redundant. Principal Component Analysis (PCA) helps effectively deal with the former issue by projecting the p-dimensional data with correlated axes onto a new orthogonal, c-dimensional space. In NIR spectra due to the high degree of multicolinearity often $c \ll p$.

In the lactate dataset, using 13 Principal Components (PCs) 99.99% of the variance in the preprocessed spectra can be explained. All subsequent operations and regressions are carried out on these 13 components. Figs 2.a and 2.b show the spectra (after pre-processing) and the spectra reconstructed from the PCs, respectively. The residuals, the portion of the spectra that are not explained by the 13 PCs, are shown in Fig 2.c).

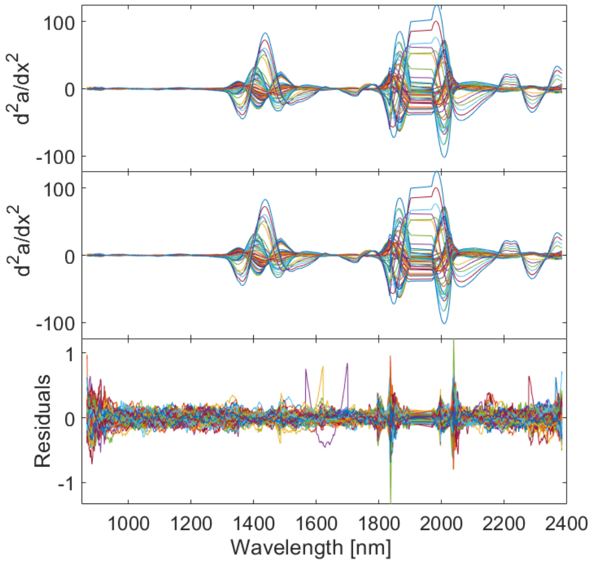

Figure2. a) The NIR spectra pertaining to the 57 lactate and PBS solutions. The water absorption peaks have been removed, and the spectra are processed with a savitzky-golay second derivative filter with polynomial order of two, and window length of 135. b) The spectra reconstructed from the first 13 principal components. These principal components explain 99.99% of the variance of the input data. c) The portions of the spectra that are not explained by the first 13 principal components.

D. *Linear and nonlinear models*

PLS and PCR are linear models that are commonly used in optical spectroscopy. In PCR, first PCA is employed to find the axes of maximal variation in the input space. Subsequently, the application of the multiple linear regression on PCs as regressors yields the PCR model. In PLS, however, the axes of the new hyperspace are chosen to maximize the correlation between the regressors and the regressand. While some favor one model over the other for theoretical reasons, in practice both models perform very similarly. For instance, in the lactate dataset, a PLS model with six components obtains the same Root Mean Square Error Cross-Validation (RMSECV) as a PCR model with 13 PCs. In the present study, in order to achieve a better comparison between the models and to isolate the importance of nonlinearity, PCR is selected and the same 13 PCs are used in all linear and nonlinear models.

Support vector machines are powerful and computationally efficient methods that are widely used in classification and regression tasks. The use of the "kernel trick" provides a simple, elegant, and computationally efficient way of incorporating nonlinearity in SVM models without the explicit definition of transfer functions or explicit transformation of the data to new hyperdimensional spaces which could be computationally intractable.

With n observations in the training set (x_i, y_i), $i \in \{1,2,\dots,n\}$, in the linear case (2),

$$f(x) = w.x + b, \tag{2}$$

and with an ϵ-insensitive loss function (3),

$$L_\epsilon(y) = \begin{cases} 0 & if\ |y - f(x)| \le \epsilon \\ |y - f(x)| - \epsilon & otherwise \end{cases}, \tag{3}$$

the aim is to find the flattest line with minimal deviation outside the boundaries set by ϵ. This is demonstrated by (4),

$$\min_w \{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*)\}$$
$$\text{s.t.} \begin{cases} y_i - w.x_i - b \le \epsilon + \zeta_i \\ w.x_i + b - y_i \le \epsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \ge 0 \end{cases} \tag{4}$$

where w is the weight vector, $\|.\|$ is the Euclidean norm, ζ_i, ζ_i^* are slack variables that take up the excess amount for observations that fall outside the ϵ boundaries; these slack variables are added to make the equation feasible. C is the capacity control parameter and similar to a regularization parameter, helps determine the tradeoff between deviations larger than ϵ and flatness of $f(x)$. In the nonlinear case, the input data are mapped onto a new hyperdimensional space with a nonlinear transformation $\phi(x)$. Solving the optimization problem requires the calculation of the dot product in the feature space $\phi(x).\phi(x')$. However, this can be avoided by finding the equivalent kernel in the input space,

$$K(x, x') = \phi(x).\phi(x')\ . \tag{5}$$

In the application of these models, the three main parameters that need to be selected are the kernel function, $K(x, x')$, the loss function $L(y)$, and the capacity control parameter, C [12].

In NIR spectroscopy applications, SVM has been shown to effectively model nonlinearities and outperform PLS. For instance, SVM regression has been shown to outperform PLS in quantification of caffeine from spectra of tea [13]. It is also shown to outperform PLS in quantification of brix and pol from sugarcane spectra [14]. A similar comparison between PLS, SVM, and ANN on 14 datasets and a variety of regressands concluded that SVM and Artificial Neural Network (ANN) outperform PLS [15].

In the present study, different kernel functions are examined, namely linear, quadratic, cubic, quartic and radial basis function. Finally, linear ϵ-insensitive loss function is used. The values of ϵ, the capacity control parameter, C, and the kernel scale are optimized using 5-fold cross-validation. This 5-fold cross validation is performed internally within each iteration of the leave-one-out cross-validation to avoid data leakage.

III. Results

In order to assess the impact of increasing nonlinearity in the absorbance behavior of lactate, the performance of the models are separately evaluated in three sets, (a). The set containing concentrations of lactate between 0-10 mmol/L (31 spectra), (b). a set containing the concentrations between 0-20 mmol/L, (51 spectra), and finally (c). a set containing all spectra with concentrations of lactate ranging between 0-600 mmol/L (57) spectra. The hypothesis is that if nonlinearities become significant in high concentrations of lactate, nonlinear models are expected to produce better fits in the training stage and as a result deliver better prediction results.

Table I, summarizes the results of leave-one-sample-out cross-validation. For SVM models the hyperparameters, namely the loss-insensitive margin, ϵ, the capacity control parameter, C, and the kernel scale, have been optimized within each fold and based on the RMSECV obtained from 5-fold cross-validation routine.

TABLE I. COMPARISON OF PCR AND SVM REGRESSION WITH DIFFERENT KERNEL FUNCTIONS

Model	**RMSECV [mmol/L]**		
	0-10 mmol/L	***0-20 mmol/L***	***0-600 mmol/L***
PCR	**1.19**	**1.02**	**1.64**
SVM-linear	1.52	1.42	1.86
SVM-quadratic	1.83	3.42	15.12
SVM-cubic	1.68	3.65	19.02
SVM-quartic	1.93	3.28	22.96
SVM-RBF	1.35	3.93	114.18

IV. DISCUSSION AND CONCLUSION

Table I shows that PCR consistently outperforms all models. Interestingly the performance of the nonlinear models noticeably deteriorates in the set that includes very high concentrations of lactate. Clearly, the investigated nonlinear models fail to generalize well. Although high concentrations of lactate are present in one of the datasets, the results suggest that the nonlinear effects have remained marginal in PBS. In other words, given the limited number of samples used in the study, the flexibility of the nonlinear models is outweighed by their increasing susceptibility to overfitting.

In the introduction section, it was mentioned that another contributory factor to the nonlinearity of absorbance lies in the scattering properties of solutions. The use of PBS in this study, which is a low scattering solution, minimizes this factor. This question will be examined further in our future work by using high scattering solutions and mediums, namely serum, whole blood and in transcutaneous measurements.

In conclusion, while the results do not reject the significance of nonlinearity, they emphasize the importance of choosing the simplest possible model. The successful application of nonlinear machine learning models that require fine-tuning of tens and hundreds of parameters, such as bagged and boosted trees models or artificial neural networks, may entice their application in areas such as optical spectroscopy. However, the large number of variables (wavelengths) in these application, in combination with the logistical difficulty of sample preparation and the acquisition of spectra, means simple, linear models, namely PLS and PCR remain viable choices.

References

[1] R. Bellomo, "Bench-to-bedside review: Lactate and the kidney," *Critical care,* vol. 6, no. 4, p. 322–326, 2002.

[2] D. De Backer, J. Creteur, H. Zhang, J.-L. Vincent and M. Norrenberg, "Lactate Production by the Lungs in Acute Lung Injury," *American Journal of Respiratory and Critical Care Medicine,* vol. 156, no. 4, pp. 1099-1102, 1997.

[3] O. Matz, C. Zdebik, S. Zechbauer, L. Bündgens, J. Litmathe, K. Willmes, J. Schulz and M. Dafotakis, "Lactate as a diagnostic marker in transient loss of consciousness," *Seizure,* vol. 40, pp. 71-75, 2016.

[4] Y. Wu, Y. Dong, M. Atefi, Y. Liu, Y. Elshimali and J. V. Vadgama, "Lactate, a Neglected Factor for Diabetes and Cancer Interaction," *Mediators of Inflammation,* vol. 2016, p. 12, 2016.

[5] J. Bakker, M. W. Nijste and T. C. Jansen, "Clinical use of lactate monitoring in critically ill patients," *Annals of Intensive Care,* 2013.

[6] C. Petibois, G. Cazorla and A. Cassaigne, "Plasma Protein Contents Determined by Fourier-Transform Infrared Spectrometry," *Clinical Chemistry,* vol. 47, no. 4, pp. 730-738, 2001.

[7] M. Mamouei, M. Qassem, K. Budidha, N. Baishya, P. Vadgama and P. Kyriacou, "Comparison of a Genetic Algorithm Variable Selection and Interval Partial Least Squares for quantitative analysis of lactate in PBS," in *IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019.

[8] T. A. Henderson and L. D. Morries, "Near-infrared photonic energy penetration: can infrared phototherapy effectively reach the human brain?," *Neuropsychiatric Disease and Treatment,* vol. 21, no. 11, pp. 2191-208, 2015.

[9] D. Lafrance, L. C. Lands and D. H. Burns, "Measurement of lactate in whole human blood with near-infrared transmission spectroscopy," *Talanta,* vol. 60, no. 4, pp. 635-641, 2003.

[10] A. Liu, G. Li, Z. Fu, Y. Guan and L. Lin, "Non-linearity correction in NIR absorption spectra by grouping modeling according to the content of analyte," *Scientific Reports,* vol. 8, p. 8564, 2018.

[11] A. Davies, "An Introduction to near Infrared Spectroscopy," *NIR News,* vol. 16, no. 7, pp. 9-11, Nov. 2005.

[12] V. Vapnik, S. E. Golowich and A. J. Smola, "Support Vector Method for Function Approximation, Regression Estimation and Signal Processing," in *Advances in Neural Information Processing Systems 9*, Denver, Colorado, US, 1996.

[13] S. Chanda , A. K. Hazarika, N. Choudhury , S. A. Islam, R. Manna, S. Sabhapondit, T. Bipan and R. Bandyopadhyay, "Support vector machine regression on selected wavelength regions for quantitative analysis of caffeine in tea leaves by near infrared spectroscopy," *Journal of Chemometrics,* vol. 33, no. 10, p. e3172, 2019.

[14] R. Tange, M. A. Rasmussen, E. Tairac and R. Broa, "Application of support vector regression for simultaneous modelling of near infrared spectra from multiple process steps," *Journal of Near Infrared Spectroscopy,* vol. 23, no. 2, pp. 75-84, 2015.

[15] R. M. Balabin and E. I. Lomakinab, "Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data," *Analyst,* vol. 136, no. 8, pp. 1703-1712, 2011.

[16] O. Devos, C. Ruckebusch, A. Durand, L. Duponchel and J.-P. Huvenne, "Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation," *Chemometrics and Intelligent Laboratory Systems,* vol. 96, no. 1, pp. 27-33, 2009.