City Research Online

City St George's, University of London

Citation: Mamouei, M. H., Budidha, K., Baishya, N., Qassem, M. & Kyriacou, P. A. (2021). An empirical investigation of deviations from the Beer-Lambert law in optical estimation of lactate. Scientific Reports, 11(1), 13734. doi: 10.1038/s41598-021-92850-4

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/26515/

Link to published version: https://doi.org/10.1038/s41598-021-92850-4

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

scientific reports

Check for updates

OPEN

An empirical investigation of deviations from the Beer–Lambert law in optical estimation of lactate

M. Mamouei[1,2]✉, K. Budidha[2], N. Baishya[2], M. Qassem[2] & P. A. Kyriacou[2]

The linear relationship between optical absorbance and the concentration of analytes—as postulated by the Beer-Lambert law—is one of the fundamental assumptions that much of the optical spectroscopy literature is explicitly or implicitly based upon. The common use of linear regression models such as principal component regression and partial least squares exemplifies how the linearity assumption is upheld in practical applications. However, the literature also establishes that deviations from the Beer-Lambert law can be expected when (a) the light source is far from monochromatic, (b) the concentrations of analytes are very high and (c) the medium is highly scattering. The lack of a quantitative understanding of when such nonlinearities can become predominant, along with the mainstream use of nonlinear machine learning models in different fields, have given rise to the use of methods such as random forests, support vector regression, and neural networks in spectroscopic applications. This raises the question that, given the small number of samples and the high number of variables in many spectroscopic datasets, are nonlinear effects significant enough to justify the additional model complexity? In the present study, we empirically investigate this question in relation to lactate, an important biomarker. Particularly, to analyze the effects of scattering matrices, three datasets were generated by varying the concentration of lactate in phosphate buffer solution, human serum, and sheep blood. Additionally, the fourth dataset pertained to invivo, transcutaneous spectra obtained from healthy volunteers in an exercise study. Linear and nonlinear models were fitted to each dataset and measures of model performance were compared to attest the assumption of linearity. To isolate the effects of high concentrations, the phosphate buffer solution dataset was augmented with six samples with very high concentrations of lactate between (100–600 mmol/L). Subsequently, three partly overlapping datasets were extracted with lactate concentrations varying between 0–11, 0–20 and 0–600 mmol/L. Similarly, the performance of linear and nonlinear models were compared in each dataset. This analysis did not provide any evidence of substantial nonlinearities due high concentrations. However, the results suggest that nonlinearities may be present in scattering media, justifying the use of complex, nonlinear models.

Near Infared (NIR), Mid-Infrared (mid-IR) visible and Ultraviolet (UV) optical spectroscopy provide a low-cost and non-invasive alternative to electrode-based approaches in the characterization of chemical compounds and the quantification of analytes. Such applications necessitate training predictive models on datasets with sufficient variations in the concentration of absorbing species. However, the provision of such datasets is highly time—and resource-demanding, as a result, the number of samples are often small. Moreover, the identification of discriminative optical patterns requires scanning broad ranges of the optical spectrum. Therefore, the acquired optical spectra contain absorbance values at hundreds or thousands of wavelengths. The limited sample sizes, n, along with the large number of wavelengths (variables), p, constitute the main features of most optical spectroscopy datasets; known as large p small n problems[1]. While this poses a challenge for predictive modelling, two conditions make the problem tractable:

[1]Deep Medicine, Nuffield Department of Women's and Reproductive Health, Oxford Martin School, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK. [2]Research Centre for Biomedical Engineering (RCBE), School of Mathematics, Computer Science and Engineering, City, University of London, Northampton Square, London EC1V 0HB, UK. ✉email: mohammad.mamouei@wrh.ox.ac.uk

the presence of redundant variables and the multicollinearity of absorbance values mean the resulting covariance matrices are low rank,
under uniform attenuation conditions, the Beer-Lambert law postulates a linear relationship between the absorbance of monochromatic light and the concentration of absorbing species,

$$A = \log_{10} I_0/I = l \sum_{i=1}^{N} \epsilon_i c_i, \tag{1}$$

where A denotes absorbance, I_o and I are the intensity of beam before and after passing through the absorbing layer, N is the number of absorbing species in the matrix, ϵ_i is the molar decadic extinction coefficient for the i^{th} absorbing species, c_i is the concentration, and l is the path length of light.

These considerations justify the choice of Principal Component Regression (PCR) and Partial Least Squares (PLS) in spectroscopic studies[2,3]. Both of these methods are linear. The former achieves dimensionality reduction by finding the axes of maximal variance in the space of independent variables, $X_{n\times p}$, while the latter does so by finding the axes of maximal covariance between the independent variables and the dependent variable, $Y_{n\times 1}$. In spite of major differences in the interpretation of latent variables in PCR and PLS, they often deliver similar predictive performances[4]. Minor improvements in predictive performance might be expected from PLS, particularly when noise constitutes much of the variance in the independent variables space[5–7].

While PLS and PCR remain workhorses of quantitative analytics in spectroscopic studies, it is also well-understood that deviations from the linearity assumption can take place when the light source is not monochromatic, the concentration of the analytes are high, and the medium is highly scattering. Mayerhöfer et al.[8] showed errors that arise from the Beer-Lambert law can exceed an order of magnitude compared to the exact solution of the Maxwell equations. Tolbin et al.[9] derived analytical expressions for the critical concentration and the extinction coefficient beyond which deviations from the Beer-Lambert law become significant. Firstly, the expectation of nonlinearities that are challenging to quantify apriori, and secondly, the prevalence and success of nonlinear Machine Learning (ML) models in different fields have paved the way for their application in spectroscopic studies. For instance[10], used Artificial Neural Networks (ANN) for classification of drug strength from NIR spectra. Santana et al.[11] compared the application of discriminant PLS and Random Forest (RF) on classification of adulterated oil and spice samples from Fourier Transform Infrared (FT-IR) and NIR spectra. They reported that RF delivers a superior performance. Mekonnen et al.[12] compared the performance of methods such as PLS, Support Vector Regression (SVR), ensemble decision trees, and ANNs on the estimation of the concentration of glucose in aqueous solutions. This comparison was conducted on an NIR dataset comprising of 47 concentrations of glucose. SVR, ANN, and a variant of ensemble decision trees obtained better performances. Balabin and Lomakinab[13] compared the performance of PLS, SVR, and ANN on NIR spectra obtained from different petrochemical matrices. It was shown that ANN and SVR deliver comparable performances and both offer more accurate predictions than PLS; the authors concluded that SVR can provide a robust alternative to ANN. Similar investigations of linear and nonlinear regression models have been reported in the literature for the estimation of soil carbon content, sugar content of orange, active substance content of tablets, moisture, fat and protein content of meat, and finally protein content of wheat in[14–16]. The PLS model is often found to deliver poorer performance compared to nonlinear models.

The present study focuses on lactate. The association of lactate with one of the most fundamental processes in the body, namely cellular respiration, makes it an important biomarker akin to glucose. Therefore, not surprisingly, clinical literature underlines the diagnostic and prognostic value of lactate in relation to numerous life-threatening conditions and diseases, such as sepsis, diabetes, cancer, pulmonary and kidney diseases[17–20]. Lactate has also been referred to as an important indicator of the risk of morbidity and mortality in critically ill patients[21]. Currently, the gold standard in the measurement of lactate requires blood sampling. This limits the ability of intensivists to frequently monitor patients' lactate levels; in spite of the calls for its routine measurements in patients with sepsis[22]. These considerations have given rise to the pursuit of non-invasive and continuous alternatives to intermittent blood sampling for lactate measurement.

Petibois et al.[23] used the mid-IR region of the optical spectrum to estimate the concentration of lactate in plasma and reported a coefficient of determination, R^2_T, of 0.94 in the test set and a Root Mean Square Error of Prediction (RMSEP) of 0.15 mmol/L. Lafrance et al.[24] showed the potential of NIR spectra in the estimation of lactate concentration in blood, reporting a coefficient of determination of 0.96 with cross-validation,R^2_{CV}. Mamouei et al.[25] applied a number of variable selection methods to the mid-IR spectra of lactate and showed that highly accurate estimates, $R^2_{CV} = 0.996$, can be achieved with models that only use a small subset of wavelengths. Budidha et al.[26] conducted a comprehensive comparison of the different regions of the optical spectrum, namely ultraviolet/visible, NIR, and mid-IR, for the measurement of lactate and highlighted the merits of mid-IR for in-vitro applications and NIR for transcutaneous applications.

In this study we adopt an empirical approach to investigate potential deviations from the Beer-Lambert law that arise from high concentrations of lactate and scattering matrices. To this end, we compare the performance of linear models, namely PCR, PLS, and linear SVR, with nonlinear models, specifically, SVR with quadratic, cubic, quartic, and Radial Basis Function (RBS) kernels. To isolate the effects of high concentrations, this comparison is performed on three partially overlapping datasets comprising of NIR spectra of lactate in PBS with concentrations in the range of 0–10 mmol/L, 0–20 mmol/L and 0–600 mmol/L. To investigate the effects of scattering matrices, the comparison is extended to incrementally more scattering matrices; from PBS to serum, whole blood, and in vivo transcutaneous spectra.

Model	RMSECV [mmol/L]					
	PBS			Human Serum	Sheep Blood	Invivo
	0–11 mmol/L	0–20 mmol/L	0–600 mmol/L	7.7–15 mmol/L	4.8–13.8 mmol/L	1.1–11.7 mmol/L
PLS	0.88	0.87	1.61	1.11	1.58	1.03
PCR	1.61	0.88	1.65	1.35	1.66	1.12
SVR-linear	**1.60**	**0.87**	**1.53**	1.23	1.65	**1.08**
SVR-quadratic	1.95	1.21	12.87	1.22	1.70	1.21
SVR-cubic	1.68	1.37	15.20	**1.15**	1.60	1.37
SVR-quartic	1.86	1.61	8.53	1.28	1.66	1.42
SVR-RBF	1.77	2.18	38.25	1.32	**1.54**	1.30

Table 1. The comparison of pls, pcr and svr models with diefferent kernel functions in different matrices.

Results

The performance of seven linear and nonlinear models are compared to investigate the extent of nonlinearities caused by high concentrations and scattering matrices. The hypothesis is that if these factors introduce significant nonlinearities, nonlinear models are expected to perform better compared to linear models.

The performance of models is evaluated using cross-validation, with test sets of size three in each iteration. These test sets are randomly selected with uniform distribution and without replacement. This is referred to as the model evaluation cross-validation loop. For SVR models, hyperparameter optimization (to find the values of C, ϵ , and the kernel scale) is performed within each fold with another five-fold cross-validation routine and Bayesian optimizer. This (hyperparameter optimization) cross-validation is nested inside the model evaluation cross-validation loop to ensure that the prediction results are representative of the external predictive performance while minimizing the risk of hyperparameter misspecification. Firstly, the model evaluation cross-validation ensures that the predictive performance of each model is tested across all samples. The alternative approach of using a single test set, given the small sample size, can be susceptible to outliers and the selection of unrepresentative tests sets, therefore, it can lead to unreliable measures of predictive performance. Secondly, the nested hyperparameter tuning cross-validation reduces the risk of hyperparameter misspecification. The alternative approach of using a randomly selected validation set would result in less samples in the training set. Also, similarly, a single validation set might not be representative of the spectra and, therefore, lead to hyperparameter misspecification. After the completion of the model evaluation cross-validation loop, Root Mean Square Error of Cross-Validation (RMSECV) and R^2_{CV} are calculated. Note the R^2_{CV} presented here is the coefficient of determination pertaining to the cross-validation routine and is different from the R^2 of calibration. This assesses the goodness of fit between the predicted values of the holdout spectra in the cross-validation routine and their reference values. Since in the model evaluation cross-validation routine, each spectrum is held out and predicted exactly once, the R^2_{CV} demonstrates the predictive performance across all spectra.

As a result of the random sampling of the training and test sets within both cross-validation loops and due to the stochastic nature of the Bayesian optimizer, different hyperparameters and models may be found in different runs, leading to different results (RMSECV and R^2_{CV}). Therefore, the process (i.e. the main model evaluation cross-validation) is repeated 10 times; R^2_{CV} values are visualized with boxplots and the lowest RMSECV among all ten runs is reported separately in Table 1. The presentation of results as boxplots helps capture the stochastic nature of the results as well as the convergence properties of the optimization in SVR models; a wider spread implies poorer convergence and model misspecification. Capturing this aspect is particularly important, as it is a direct cost of using more complex models compared to PLS and PCR. The inclusion of these considerations in the present study distinguishes the results presented here from our previous preliminary work on the PBS spectra[27].

In the interpretation of the results, it is worth noting that the PLS model is fitted directly on the spectra while the PCR and all SVR models are fitted to the PCs that present a loss of 0.01% in variance in the PBS, serum, and blood datasets, and a loss of 20% in the invivo dataset. Therefore, a major difference between the performance of PLS and PCR models can imply that a) the variance not captured by the PCs are important for predictions b) the ratio between the number of PCs and the number of observations is large, leading to poorer regression. Hence, the assessment of linear and nonlinear effects should be primarily based on the comparison of how well the PCR and the SVR with linear kernel perform relative to the nonlinear models. A fairer comparison may take the hyperparameter tuning requirements of the SVR models into account, restricting pairwise comparisons to SVR with linear kernel to SVR with nonlinear kernels.

Nonlinearities due to high concentration. To investigate the effect of high concentrations of lactate, first the PBS dataset is analyzed where scattering due to compounds other than lactate is minimal. To this end, three partially overlapping datasets with concentrations of lactate ranging between 0–11, 0–20, and 0–600 mmol/L are used to compare the performance of linear and nonlinear regression methods. The choice of the ranges of lactate levels was primarily motivated by computational considerations. The approach described above for hyperparameter tuning and model evaluation, while delivers more reliable results, is very computationally demanding. Performing this for SVR models takes multiple days for each dataset on a high-end workstation with a 10-core Intel® Xeon® Silver 4114 processor and 32 GB RAM . The preferable approach would be to incrementally increase the range of the dataset—by adding one or multiple samples at the time—and analyzing

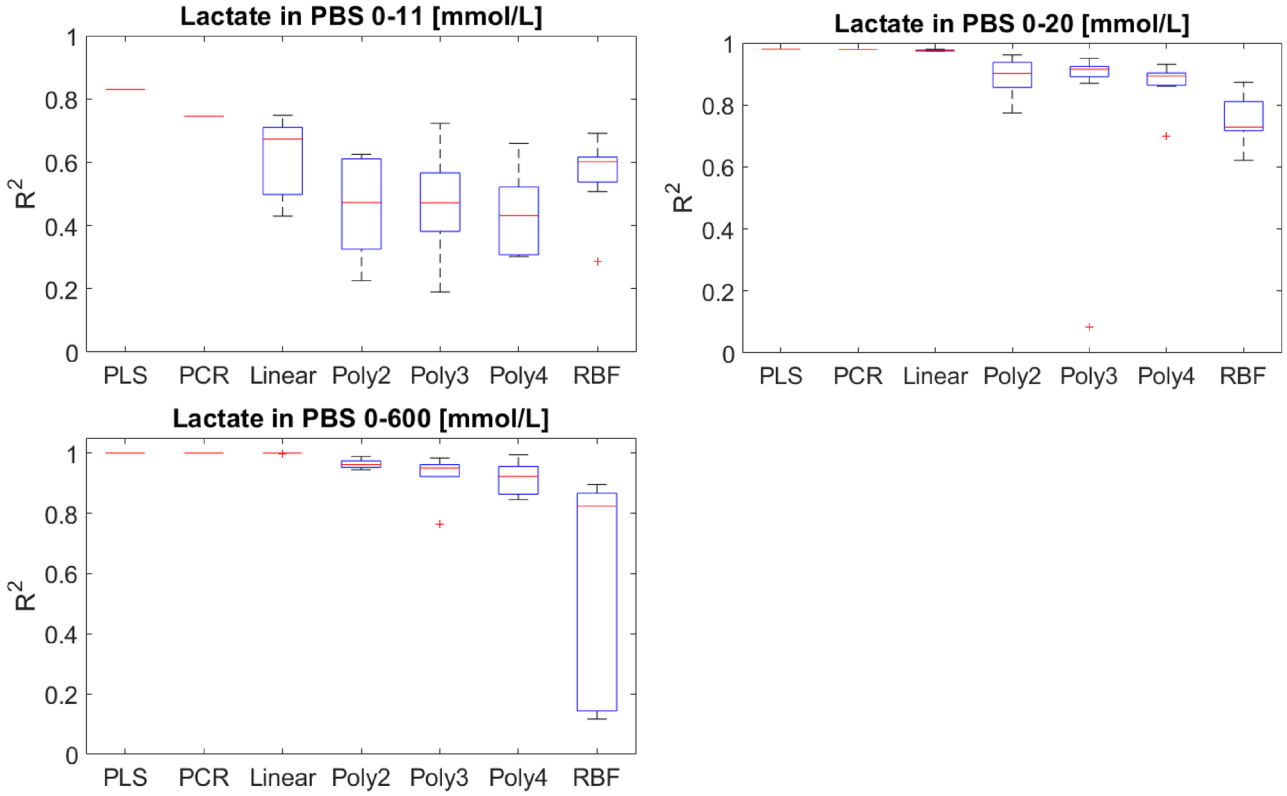

Figure 1. The comparison of the performance of linear and nonlinear models is datasets with low, medium, and high ranges of lactate concentrations. The boxplots summarise the results pertaining to ten runs of cross-validation for each model. After each run, an estimate of the coefficient of determination, R^2_{CV}, is obtained.

the trends in model performance, however, due to the high computational time this was not feasible. Instead, we allocated the spectra to the three aforementioned sets to present significant increase in the range of lactate levels. Since the model evaluation cross-validation loop sampled test sets of threes, minor adjustments were made to keep the number of spectra in the sets divisible by three. The third dataset was formed by adding six spectra pertaining to lactate concentrations of 100, 200, 300, …, and 600 mmol/L to the data. These values are far outside the physiological range of lactate, but they were included to assess if extremely high concentrations of lactate cause nonlinearities. Figure 1. Depicts the results.

In all datasets—specifically in the datasets with medium and high concentrations of lactate—the linear models obtain the best performances. Therefore, this analysis does not provide any evidence of the presence of significant nonlinearities due to high concentrations of lactate.

Nonlinearities due to scattering matrices. Figure 2, summarizes the performance of models in increasingly more scattering matrices, namely PBS, human serum, sheep blood and in transcutaneous spectra acquired from the participants of an exercise study.

While the results do not show a strong upward trend with increasing model nonlinearity, both in the serum dataset and the blood dataset interesting patterns emerge. Unlike PBS, in serum and blood, whether the best observed R^2_{CV} in ten runs (top end of the top whisker) or the median R^2_{CV} (the red bar) are considered, the trend is either almost flat or upward. In serum, the SVR model with cubic polynomial kernel shows a better performance than the linear SVR and PCR models when the top whisker is considered. This effect is much more pronounced in the blood dataset. The SVR model with RBF kernel obtains the best R^2_{CV}, in terms of both the median and the best observed case. This is despite the fact that in small datasets, the complexity and the requirement of hyperparameter tuning for the SVR models puts them at a disadvantage compared to simpler PCR and PLS. This suggests that the nonlinearities must be substantial to compensate for and exceed the expected marginal loss of accuracy due to additional model complexity. The noticeably high nonlinearity observed in sheep blood relative to human serum, could be related to the differences between the two mediums. The composition of blood and the distribution of lactate between red blood cells and plasma are different in the two mediums, potentially contributing to differential scattering properties and optical absorbance profiles.

Surprisingly, the same pattern does not emerge in the invivo dataset. One possible explanation is that proxies of lactate concentration might have been detected rather than lactate itself. In the exercise study it was expected that other absorbing species, such as oxyhaemoglobin and deoxyhaemoglobin, can show highly correlated variations with lactate. The observations that RMSECV is lower and R^2_{CV} is higher in this dataset relative to serum and blood, supports this possibility. Especially because in this dataset, a portable spectrophotometer with a much lower resolution and a shorter wavelength range was used, the number of observations was smaller, and baseline

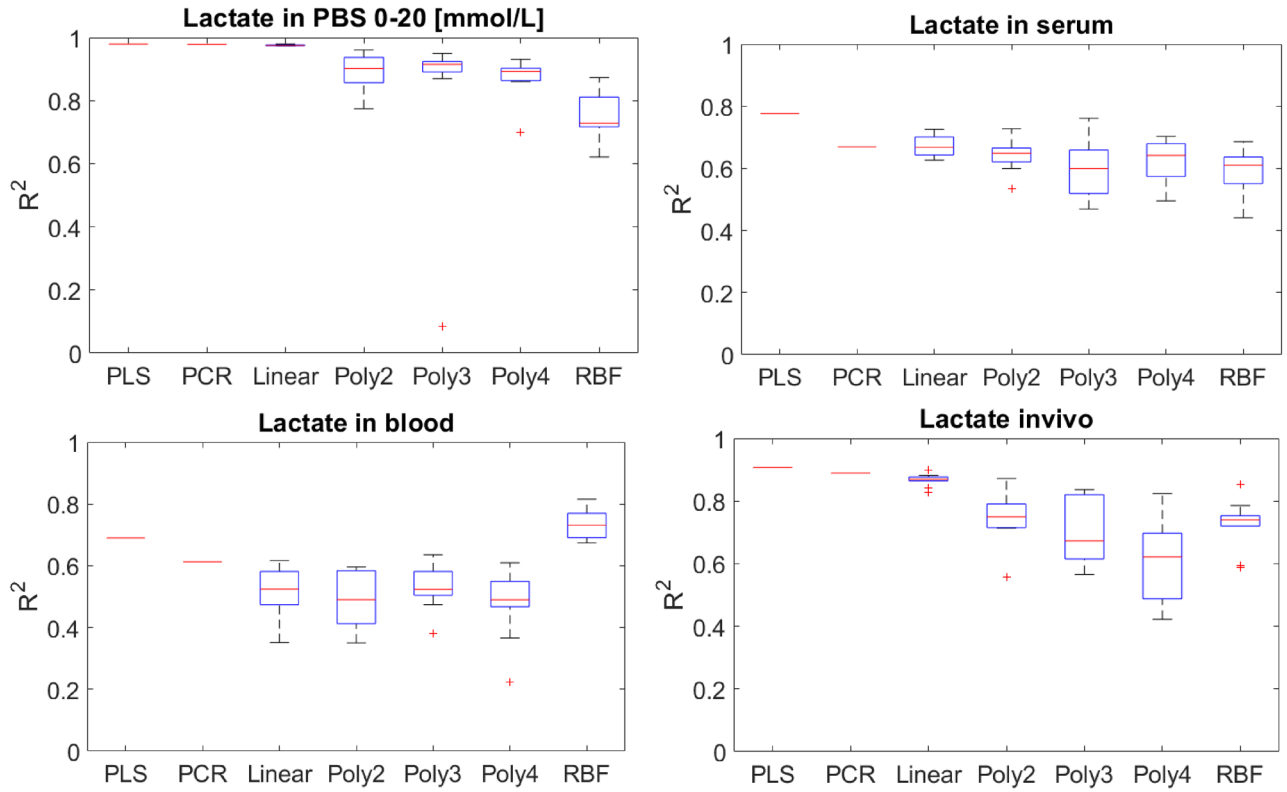

Figure 2. The comparison of the performance of linear and nonlinear models in increasingly more scattering matrices. The boxplots summarise the results pertaining to ten runs of cross-validation for each model. After each run, an estimate of the coefficient of determination, R^2_{CV}, is obtained.

differences in optical properties of the four participants are expected to be far greater. Table 1. Shows the best RMSECV obtained in ten different runs of the cross-validations routine.

Discussion and conclusion

This study focused on the optical behavior of lactate; an important and fundamental biomarker that, if measured routinely and accurately, can shed light on numerous diseases and health problems. Previous studies have underlined the potential of optical estimation of lactate as a noninvasive, inexpensive and continuous alternative to blood sampling. However, accurate and reliable optical measurements are still not within reach. A better understanding of the interactions of light and this biomarker is necessary to recognize the merits, limitations, and practical issues of optical sensing. Since previous attempts in the optical estimation of lactate have used linear models, we investigated potential deviations from the linearity assumption postulated by the Beer-Lambert law. To this end, a series of experiments were conducted to analyze potential nonlinear effects that can arise due to high concentrations of lactate and scattering matrices. The assumption was that if nonlinear effects become substantial, nonlinear models will deliver better accuracies than linear models.

Seven linear and nonlinear models were compared in datasets with concentrations of lactate ranging between 0–11 mmol/L, 0–20 mmol/L, 0–600 mmol/L. For this investigation a minimally scattering matrix (PBS) was used to ensure potential deviations are mainly due to high concentrations. This analysis did not provide any evidence of significant nonlinearities. A similar comparison in incrementally more scattering matrices, namely human serum, sheep blood, and transcutaneous spectra, showed some merits to the use of nonlinear models. Both in serum and blood, nonlinear models obtained better performances than PCR and SVR with linear kernel.

In summary, (a) the results confirm the potential of optical measurements of lactate, both invitro and invivo, albeit the latter is likely to have been indirect. Therefore, more studies, with more participants, and in different scenarios is necessary to assess the feasibility of indirect optical sensing of lactate. (b) It was shown that concentrations of lactate, even far beyond the biological range, do not present substantially nonlinear absorbance. (c) Nonlinear models showed merits in direct measurement of lactate when the medium is scattering, i.e. serum and blood.

Finally, the authors would like to emphasize that the RMSECV presented in Table 1 is not representative of the best accuracies that can be achieved. Previous studies have shown that major improvements may be observed when nonlinear baseline corrections are used and redundant wavelengths are excluded. In the present study, since the primary objective was the analysis of nonlinear absorbance, the aforementioned topics were not covered.

Methods

The datasets. All datasets were produced after obtaining approval by the Senate Research Ethics Committee (SREC) at City, University of London (SREC 17–18 05 6ii 27.06.2018) and all methods were carried out in accordance with the relevant guidelines and regulations.

PBS. The dataset consists of 57 NIR spectra of different concentrations of lactate in a Phosphate Buffer Solution (PBS). The procedure for the preparation of the solutions and acquisition of the spectra is detailed in[27]. The dataset contains 31 samples with concentrations of lactate between 0–10 mmol/L (increments of 0.25 mmol/L), 21 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). Spectra were acquired using the Lambda 1050 dual-beam UV/Vis/NIR spectrophotometer (Perkin Elmer Corp, Massachusetts, USA), with a spectral resolution of 1 nm. The light source used in the spectrophotometer was a halogen tungsten lamp. Indium gallium arsenide (InGaAs) and the lead sulfide detectors (PbS) were used to detect the transmitted NIR light in the range between 800 and 2600 nm. Baseline correction was performed on the spectrophotometer prior to the acquisition of a spectra, at 100% Transmission / 0% absorbance to remove background noise. Once, background correction was performed, 300 µl of each sample was transferred to in a macro quartz cuvette (Hellma GmbH & Co.KG, Jena, Germany) with a path length of 1 mm to acquire the three NIR spectra of each sample in the desired wavelength range. The three spectra were then averaged, and the resulting spectrum of each sample was considered for further analysis.

Serum. The dataset consists of 36 NIR spectra of lactate in human blood serum. Mixed pool human serum collected from healthy volunteers was purchased from TCS Biosciences Ltd., (Buckingham, UK). The base lactate of the purchased serum was 7.7 mmol/L. Thirty-five serum samples of 29 mL were then serially diluted with 1 mL of stock solutions containing varying concentration of lactate in PBS. The concentration of the serum samples was measured before the acquisition of spectra using the ABL 825 Flex (Radiometer UK Limited, Crawley, West Sussex, UK). The concentration of lactate within the samples ranged between 7.7 and 15 mmol/L with an average increment of 0.20 mmol/L.

Once the samples were prepared and the concentration measured, the spectra of each sample was acquired again in transmission mode using the Lambda 1050 dual-beam UV/Vis/NIR spectrophotometer (Perkin Elmer Corp, Massachusetts, USA). The acquisition procedure of serum spectra was similar to that of the PBS spectra, as detailed above.

Sheep blood. The dataset consists of 36 spectra of lactate in sheep blood acquired using the Lambda 1050 dual-beam UV/Vis/NIR spectrophotometer (Perkin Elmer Corp, Massachusetts, USA) equipped with 100 mm InGaAs Integrating Sphere. All the spectra were acquired in reflectance mode using a 300 µl sample in the range between 900 and 2500 nm. The procedure for the preparation of the solutions and acquisition of the spectra is detailed in[28]. The concentration of lactate is within the range of 4.8–13.8 mmol/L and the average increment is around 0.25 mmol/L. The spectral resolution and acquisition setup used to acquire blood spectra were similar to that of PBS.

Invivo. The dataset consists of 27 reflectance spectra obtained from four healthy participants during maximal effort cycling on a spinning bike (WattBike Pro, Wattbike Ltd, Nottingham, UK). The participants were 22–31 years old and gave informed written consents before commencing the experiment. Since blood lactate is known to increase with physical effort, the experiment was designed to induce changes in participants' blood lactate levels over the course of the exercise study. The participants were asked to cycle for as long as they can at a fixed peddle speed and magnetic resistance, the air resistance was increased after every minute until volitional exhaustion of the volunteer or upon reaching 90% of predicted maximal heart rate (derived using the equation max heartrate = 220 − age in years). A 45 s rest was allowed after every minute of cycling. During the rest period, optical spectra were acquired from the right index of the volunteer and a drop of blood was drawn from the left index finger using a sterile lancet. The capillary blood was collected on a finger stick and the lactate level in blood was measured using the portable Lactate Pro 2 analyzer. The concentration of lactate measured was within the range of 1.1–11.7 mmol/L and the average increment was 0.41 mmol/L. As intended, overall, the participants' lactate levels followed an upward trend, however, significant variations in patterns of increase and maximum lactate levels were observed amongst the participants. For instance, for one of the participants, who is a semi-professional cyclist, the highest lactate level was around 4 mmol/L after 20 min of exercise, while another volunteer reached exhaustion and a lactate of around 8 mmol/L only after 5 min.

Invivo spectra in the 900–1700 nm range were acquired using the NIRQUEST 512–1.9 NIR Spectrophotometer (Ocean Optics Inc., Florida, USA). A reflectance optical fiber probe (600 um fibers) was used to transmit and detect NIR light reflected from the finger of health volunteers. A small slit of 25 µm was chosen to improve optical resolution and stop the detector from saturating. While it would be preferable to use the same spectrophotometer across all datasets, NIRQUEST was selected for the invivo study due to its high sampling rate. This allowed us to collect a sufficiently large number of spectra during the 45-s rest periods and average them to minimize the motion artifacts. The same could not be achieved with the Lambda 1050 spectrophotometer that has a much lower sampling rate.

Figure 3. depicts the raw spectra in each dataset.

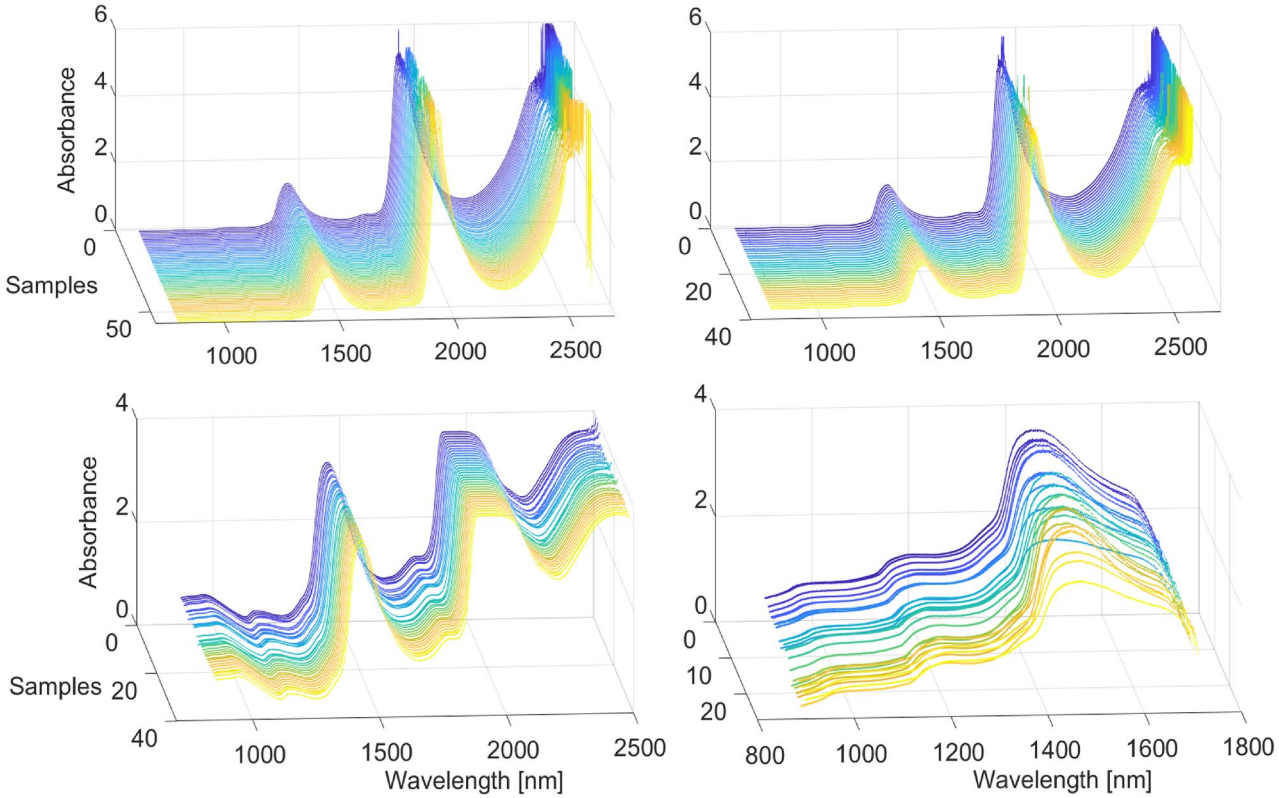

Figure 3. Raw absorbance of spectra of lactate in (**a**) PBS (**b**) human serum (**c**) sheep blood (**d**) invivo human tissue.

Preprocessing of spectra. In the PBS and serum datasets, wavelengths between 1900–1980 nm and 2450–2600 nm show high levels of noise and are hence removed. This noise is caused by the oversaturation of the lead sulfide detector in the transmittance mode due to water absorption peaks. The PBS spectra were processed using Multiplicative Scattering Correction (MSC) and a Savitzky-Golay (SG) filter with the window length of 135, second order polynomial and second order derivative. The serum spectra, were processed with MSC and SG filter with window length of 151, third order polynomial and second order derivative. The blood spectra were acquired in reflectance mode and subsequently transformed to absorbance. Therefore, no noise was observed in the aforementioned regions, however, for consistency, they were removed. The transcutaneous spectra were processed with SG filter, polynomial order of three, window length of seven and derivative order of three.

Dimensionality reduction. Principal Component Analysis (PCA) is applied to all datasets to reduce the dimensionality of the spectra prior to model fitting. For the in-vitro sets the number of components is selected such that 99.99% of the variance is explained by the PCs. This led to the selection of 12, 14, 13 PCs in the PBS datasets with low, medium, and high ranges of concentrations respectively, 16 PCs in the serum dataset, and 22 PCs in the blood dataset. For the tissue dataset, given the noisy nature of the data, this criterion led to a large number of PCs and, consequently, overfitting. The explained variance of 80% was found to produce good results across all models and was therefore selected. This led to 8 PCs.

For the PLS model the number of Latent Variables (LVs) was selected as the point where the Predicted Residual Error Sum of Squares (PRESS) plateaus. This criterion led to 11, 9 and 8 LVs for the PBS datasets with low, medium, and high ranges of concentrations respectively, 10 LVs for the serum dataset, 12 LVs for the blood dataset, and 6 LVs for the invivo dataset.

Linear and nonlinear models. The linear models used in this study are of PCR, PLS, and SVR,

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

In SVR the objective is to find the flattest line while the prediction error shows minimal deviation beyond ϵ. Therefore, given the training set $(x_i, y_i), i \in \{1, 2, \ldots, n\}$, w^* is defined as

$$w^* = \min_{\arg w} \left\{ \frac{1}{2} w^2 + C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*) \right\}$$
$$\text{s.t.} \begin{cases} y_i - w.x_i - b \leq +\zeta_i \\ w.x_i + b - y_i \leq +\zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0 \end{cases} \quad (3)$$

where $\| \cdot \|$ is the Euclidean norm, and ζ_i, ζ_i^* are slack variables that absorb excess errors when a solution is not possible that guarantees errors restricted to ϵ boundaries. With an ϵ-insensetive loss, ζ_i is defined as,

$$\zeta_i = \begin{cases} 0 & if \ |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon & otherwise \end{cases}, \quad (4)$$

where C is the capacity control parameter that determines the tradeoff between higher loss values and higher norm w^2 (less flat plane).

The incorporation of nonlinearities can be achieved by using nonlinear transformations, $\phi(x)$, to map the explanatory variables into new hyperdimensinoal spaces. For instance, a quadratic polynomial augmentation of a two-dimensional feature space $(x_1,x_2) \subset \mathbb{R}^2$ may include all polynomial terms of degree two $(x_1,x_2, x_1.x_2, x_1^2,x_2^2) \subset \mathbb{R}^5$. In high-dimensional data the explicit use of such nonlinear transformations, $\phi(x)$, can be intractable. However the use of the "kernel trick" provides a computational effective way to achieve this. Specifically, solving the optimization above necessitates the calculation of $\phi(x)^T.\phi(x')$, these computationally demanding transformations can be avoided by finding the equivalent kernel, $K(x, x') = \phi(x) \cdot \phi(x')$. Therefore, the four main parameters that need to be selected are the scaling of features, the kernel function, $K(x, x')$, the loss function, ϵ, and the capacity control parameter, C[29,30].

In the present sudy, different kernel functions are compared, namely polynomial kernels,

$$K_{poly}(x, x') = \left(1 + x^T.x'\right)^p \quad p \in \{2, 3, 4\}, \quad (5)$$

and Radial Basis Function (RBF),

$$K_{RBF}(x, x') = e^{-(x-x')^T \cdot (x-x')}. \quad (6)$$

The value of C, ϵ, and the kernel scale are optimized.

Received: 30 January 2021; Accepted: 14 June 2021
Published online: 02 July 2021

References

1. Johnstone, I. M. & Titterington, D. M. Statistical challenges of high-dimensional data. *Phil. Trans. R. Soc. A* **367**, 2009 (1906).
2. Abdi, H. Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Comput. Stat.* **2**(1), 97–106 (2010).
3. Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Comput. Stat.* **2**(4), 433–459 (2010).
4. Wentzell, P. D. & Montoto, L. V. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemom. Intell. Lab. Syst.* **65**(2), 257–279 (2002).
5. Yaroshchyk, P., Death, D. L. & Spencer, S. J. Comparison of principal components regression, partial least squares regression, multi-block partial least squares regression, and serial partial least squares regression algorithms for the analysis of Fe in iron ore using LIBS. *J. Anal. At. Spectrom.* **27**(1), 92–98 (2012).
6. Hemmateenejad, B., Akhond, M. & Samari, F. A comparative study between PCR and PLS in simultaneous spectrophotometric determination of diphenylamine, aniline, and phenol: Effect of wavelength selection. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **67**(3–4), 958–965 (2007).
7. Khajehsharifi, H., Pourbasheer, E., Tavallali, H., Sarvi, S. & Sadeghi, M. The comparison of partial least squares and principal component regression in simultaneous spectrophotometric determination of ascorbic acid, dopamine and uric acid in real samples. *Arab. J. Chem.* **10**(2), 3451–3458 (2017).
8. Mayerhöfer, T. G., Mutschke, H. & Popp, J. Employing theories far beyond their limits—the case of the (Boguer-) Beer–Lambert Law. *ChemPhysChem* **17**(13), 1948–1955 (2016).
9. Tolbin, A. Y., Pushkarev, V. E. & Tomilova, L. G. A mathematical analysis of deviations from linearity of Beer's law. *Chem. Phys. Lett.* **706**, 520–525 (2018).
10. Wua, W. *et al.* Artificial neural networks in classification of NIR spectral data: design of the training set. *Chemom. Intell. Lab. Syst.* **33**(1), 35–46 (1996).
11. Santana, F. B., Neto, W. B. & Poppi, R. J. Random forest as one-class classifier and infrared spectroscopy for food adulteration detection. *Food Chem.* **293**, 323–332 (2019).
12. Mekonnen, B. K., Yang, W., Hsieh, T. H., Liaw, S. K. & Yang, F. L. Accurate prediction of glucose concentration and identification of major contributing features from hardly distinguishable near-infrared spectroscopy. *Biomed. Signal Process. Control* **59**(1), 101923 (2020).
13. Balabin, R. M. & Lomakinab, E. I. 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* **136**(8), 1703–1712 (2011).
14. Ding, J., Yang, A., Wang, J., Sagan, V. & Yu, D. Machine-learning-based quantitative estimation of soil organic carbon content by VIS/NIR spectroscopy. *PeerJ* **6**, e5714 (2018).
15. Ni, W., Nørgaard, L. & Mørup, M. Non-linear calibration models for near infrared spectroscopy. *Anal. Chim. Acta* **813**, 1–14 (2014).
16. Liu, Y., Sun, X., Zhou, J., Zhang, H. & Yang, C. Linear and nonlinear multivariate regressions for determination sugar content of intact Gannan navel orange by Vis–NIR diffuse reflectance spectroscopy. *Math. Comput. Model.* **51**(11–12), 1438–1443 (2010).
17. Bellomo, R. Bench-to-bedside review: Lactate and the kidney. *Crit. Care* **6**(4), 322–326 (2002).

18. De Backer, D., Creteur, J., Zhang, H., Vincent, J.-L. & Norrenberg, M. Lactate production by the lungs in acute lung injury. *Am. J. Respir. Crit. Care Med.* **156**(4), 1099–1102 (1997).
19. Matz, O. *et al.* Lactate as a diagnostic marker in transient loss of consciousness. *Seizure* **40**, 71–75 (2016).
20. Wu, Y. *et al.* Lactate, a neglected factor for diabetes and cancer interaction. *Mediators Inflamm.* **2016**, 12 (2016).
21. Bakker, J., Nijste, M. W. & Jansen, T. C. Clinical use of lactate monitoring in critically ill patients. *Ann. Intensive Care* **3**, 1–8 (2013).
22. Levy, M., Evans, L. E. & Rhodes, A. The surviving sepsis campaign bundle: 2018 update. *Crit. Care Med.* **46**(6), 997–1000 (2018).
23. Petibois, C., Cazorla, G. & Cassaigne, A. Plasma protein contents determined by Fourier-transform infrared spectrometry. *Clin. Chem.* **47**(4), 730–738 (2001).
24. Lafrance, D., Lands, L. C. & Burns, D. H. Measurement of lactate in whole human blood with near-infrared transmission spectroscopy. *Talanta* **60**(4), 635–641 (2003).
25. Mamouei, M., Budidha, K., Baishya, N., Qassem, M. & Kyriacou, P. Comparison of wavelength selection methods for in-vitro estimation of lactate: a new unconstrained, genetic algorithm-based wavelength selection. *Sci. Rep.* **10**, 16905 (2020).
26. Budidha, K. *et al.* Identification and quantitative determination of lactate using optical spectroscopy—towards a noninvasive tool for early recognition of sepsis. *Sensors* **20**(18), 5402 (2020).
27. Mamouei, M., Budidha, K., Baishya, N., Qassem, M. & Kyriacou, P. A. The efficacy of support vector machines in modelling deviations from the Beer-Lambert law for optical measurement of lactate. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada (2020).
28. Baishya, N. *et al.* Investigations into the effects of pH on quantitative measurements of lactate in biological media using ATR-FTIR spectroscopy. *Molecules* **25**(16), 3695 (2020).
29. Vapnik, V., Golowich, S. E. & Smola, A. J. Support vector method for function approximation, regression estimation and signal processing. In *Advances in Neural Information Processing Systems 9*, Denver, Colorado, US (1996).
30. Devos, O., Ruckebusch, C., Durand, A., Duponchel, L. & Huvenne, J.-P. Support vector machines (SVM) in near infrared (NIR) spectroscopy: focus on parameters optimization and model interpretation. *Chemom. Intell. Lab. Syst.* **96**(1), 27–33 (2009).

Author contributions

M.M. conceived and performed the analysis, K.B. and N.B. carried out the experiments, M.Q. and P.K. provided expertise and feedback. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021