Talanta 107 (2013) 368–375

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Modified locally weighted—Partial least squares regression improving clinical predictions from infrared spectra of human serum samples

David Perez-Guaita [a], Julia Kuligowski [b], Guillermo Quintás [c], Salvador Garrigues [a,*], Miguel de la Guardia [a]

[a] Analytical Chemistry Department, University of Valencia, Edifici Jeroni Muñoz, Burjassot, Valencia 46100, Spain
[b] Division of Neonatology, University Hospital Materno-Infantil La Fe, Bulevar Sur, s/n, Valencia, Spain
[c] Leitat Technological Center, Bio In Vitro Division, Terrassa, Spain

ARTICLE INFO

Article history:
Received 23 October 2012
Received in revised form
16 January 2013
Accepted 19 January 2013
Available online 31 January 2013

Keywords:
Local weighted-partial least squares regression (LW-PLSR)
Human serum analysis
Vibrational spectroscopy
Infrared (IR)
Chemometrics

ABSTRACT

Locally weighted partial least squares regression (LW-PLSR) has been applied to the determination of four clinical parameters in human serum samples (total protein, triglyceride, glucose and urea contents) by Fourier transform infrared (FTIR) spectroscopy. Classical LW-PLSR models were constructed using different spectral regions. For the selection of parameters by LW-PLSR modeling, a multiparametric study was carried out employing the minimum root-mean square error of cross validation (RMSCV) as objective function. In order to overcome the effect of strong matrix interferences on the predictive accuracy of LW-PLSR models, this work focuses on sample selection. Accordingly, a novel strategy for the development of local models is proposed. It was based on the use of: (i) principal component analysis (PCA) performed on an analyte specific spectral region for identifying most similar sample spectra and (ii) partial least squares regression (PLSR) constructed using the whole spectrum. Results found by using this strategy were compared to those provided by PLSR using the same spectral intervals as for LW-PLSR. Prediction errors found by both, classical and modified LW-PLSR improved those obtained by PLSR. Hence, both proposed approaches were useful for the determination of analytes present in a complex matrix as in the case of human serum samples.

1. Introduction

Total protein, triglycerides, glucose and urea contents in blood are commonly determined as components of comprehensive metabolic panels typically included in routine health checkups for the evaluation of conditions such as liver and kidney disease or diabetes, among other metabolic and nutritional disorders. On the other hand, the development of multianalyte reagent-free spectroscopy based methods is an active field of research in analytical and clinical chemistry [1,2] where mid-infrared (IR) spectroscopy shows a number of relevant advantages over other techniques based on enzymatic-colorimetric reactions and enzyme-linked immuno sorbent assay (ELISA) determinations. The broad detection capabilities of IR allow its application to the quantification of a big amount of molecules, from small ions to proteins, without previous derivatization steps thus offering a fast, cheap, reagent-free and compactable alternative to enzymatic-colorimetric methods. However, the wide range of molecules with characteristic absorption bands in the mid-IR region increases the likelihood of interferences arising from other sample constituents. Frequently IR absorption bands of matrix components are strongly overlapped with signals of the analytes of interest in the mid-IR region, which hampers the selection of specific bands suitable for analyte quantification based on univariate regressions. In order to develop direct methods, thus avoiding time consuming sample pre-treatment steps, a common approach is the use of multivariate regression [3–9]. In particular, chemometric modeling of spectra provided by attenuated total reflectance (ATR) was applied for the determination of clinical parameters in human serum samples [10].

Multivariate linear regression models, mostly employing partial least squares regression (PLSR), have been developed for the determination of total protein, triglyceride, glucose and urea contents in serum [11–15], plasma [10] and whole blood [16]. A comparison of the analytical performance of these works can be found in [2] and [1].

The aforementioned PLS methods assumed linear relationships between the IR absorbance and the concentration of the analytes within typical physiological ranges. In spite of that, a number of linearity problems associated with ATR measurements in biological samples have been reported in literature [17] due to the formation of biofilms on the ATR cell [18], the confinement of

* Corresponding author. Tel.: +34 96 354 4838; fax: +34 96 354 4845.
E-mail address: salvador.garrigues@uv.es (S. Garrigues).

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.01.035

analytes inside cells [11] or cell sedimentation [19]. Hence, it can be expected that changes in the penetration depth of the IR beam in ATR systems and the complexity of the serum matrix could affect the linearity of the measurements.

PLS is a multivariate inverse method widely used in chemistry [20]. This method extracts a set of latent variables (LVs) explaining the sources of variation in the *X*-block (i.e., spectra) correlated to an *y*-vector (i.e., analyte concentrations). For accurate predictions of analyte concentrations, it is essential to have access to a set of representative calibration samples that include all sources of variation expected to be present in new unknown samples and, at the same time, their spectral contributions to the overall signal should match. Thus, the selection of appropriate calibration sets might be troublesome due to the complex nature of serum samples. Besides, target analyte concentrations typically show a wide variation hindering the availability of appropriate calibration sets.

Addressing the abovementioned concerns, LW-PLSR can be seen as a suitable strategy to overcome a lack of linearity of the relationship between signals and analyte concentrations and to facilitate the selection of proper calibration sets. Local regression approximations are based on the assumption that the use of specific calibration equations for each new sample to be analyzed, using small calibration sets tailored to the unknown sample from a large library of samples, improves prediction accuracy [21]. As described by Pérez-Marin et al. 'local regression, combines the advantages of global calibration in using one database to cover a large product domain, with the accuracy obtainable with specific calibrations' [22]. In other words, for each unknown sample to be predicted, local regression models are based on an initial selection of a reduced set of calibration spectra providing similar features, in order to develop a specific calibration model. Locally weighted regression was first proposed by Cleveland et al. for the estimation of nonlinear regression surfaces when little information about the surface was available [23]. Naes et al. found a nonlinear relationship between sample composition and principal component scores [24,25]. To correct this non-linearity, the same authors applied LW-PLSR, where only the closest samples characterized by a minimum distance in the scores space were employed for local model calculation. Briefly, the approach consisted in four steps: (i) development of a PCA model; (ii) computation of the Euclidean distance between the query and the calibration samples in the scores space, (iii) selection of the nearest neighbors to the query and (iv) calculation of a Principal Component Regression (PCR) or PLSR employing the selected samples. This method has been criticized because it only takes into account "spectral aspects" of the response variable (*X*-block) and ignores the "chemical information" of the concentration vector *y* [26]. Therefore, numerous approaches have been proposed as alternatives for the selection of samples for local models, as for example the similarity estimated from the output of a global method [26] or the Euclidean distance in the *X*-block [27,28]. In addition, methods using the correlation between variables [29] and most recently methods where all samples are stored in the calibration set and weighted depending on their distance to the query [30,31] have been proposed.

The aim of this work is to evaluate the advantages and drawbacks of the application of LW-PLSR to the direct determination of total protein, triglycerides, glucose and urea contents in human serum employing ATR-FTIR spectra. For this propose spectra from 1400 serum samples acquired during a previous study [15] dealing with the determination of the influence of the origin of serum samples on PLS model performance, were used. For the present study, samples from all origins were randomly divided into calibration and validation sets for PLS and LW-PLSR model calculation and comparison of their respective prediction capabilities. In the case of serum samples it must be remarked that the problem of using the PC space to find similarities between samples with similar concentrations of the analyte under study is that results may be influenced by matrix effects. Because of that, a simple approach based on the use of an analyte specific spectral region and a multi-parametric method for the selection of the optimum number of PCs was developed. In addition, a simple, modified LW-PLSR approach is proposed in order to optimize obtained results based on the use of an analyte specific spectral region for PCA clustering to overcome difficulties arising due to matrix effects in complex human serum samples.

2. Materials and methods

2.1. Sample collection and data acquisition

A total of 1400 samples obtained from the Hospital Dr. Peset Alexandre (Valencia, Spain) were analyzed. Reference concentrations of total protein, triglycerides, glucose and urea contents in human serum of the samples included in the study, were obtained through the use of an Abbott Architect c16000 auto-analyzer (Libertyville, IL, USA) as described elsewhere [15] in the clinical laboratory of the hospital, providing precisions $\leq 5\%$. Detailed information about the samples can be found in a previous work [15]. Table 1 summarizes the main descriptive statistical parameters of the sample reference data used throughout this study.

FTIR spectra were acquired on a Bruker Tensor 27 (Bremen, Germany) spectrophotometer equipped with a deuterium triglycine sulfate detector and an ATR DuraSampleIR accessory with a nine reflection diamond/ZnSe Dura disc from Smiths detection Inc. (Warrintong, UK). Aliquots of 150 μL of each serum sample were deposited on the ATR crystal and covered using an N-BK7 PCV lens to avoid sample evaporation. A total of 100 accumulated scans in the 600–4000 cm^{-1} range at a resolution of 4 cm^{-1} were averaged to increase the signal to noise ratio using a background of the empty ATR cell obtained under the same instrumental conditions and ATR correction was applied to the resulting mean spectrum. A blank spectrum of water was subtracted to each serum spectrum. All serum samples were measured by triplicate and averaged. The contribution of water vapor to the final spectrum was subtracted to the average spectrum of each sample. Further details on samples and spectra acquisition are available in [15].

2.2. Software and spectral preprocessing

Data analysis was run under Matlab 7.7.0 from Mathworks (Natick, USA, 2004). PLS Toolbox 6.2 from Eigenvector Research Inc. (Wenatchee, WA, USA) was used for building of PLS and classical LW-PLSR models and in-house written MATLAB functions were employed for modified LW-PLSR.

Table 1
Descriptive statistical parameters of the reference values of samples used throughout this study.

Analyte	Calibration set				Validation set			
	N	Mean conc.	SD	Interval	*N*	Mean conc.	SD	Interval
Proteins	332	6.5	0.7	4.1–9.3	331	6.5	0.7	4.1–8.6
Urea	584	48	36	15–249	583	48	36	15–242
Glucose	592	106	40	32–490	591	106	38	35–424
Triglycerides	510	127	84	51–1280	509	126	72	51–598

Note, *N*: number of samples and SD: standard deviation. All values are in mg/dL except for proteins which are in g/dL.

For each analyte samples were sorted according to their concentration and odd and even indexed samples were used, respectively, to construct calibration and validation sets, thus ensuring that both datasets cover the whole range of concentrations (see Table 1). For all models, first derivative spectra were used and data sets were mean centered as a pre-processing step.

2.3. PLS model calculation

PLS models for each component under study were calculated using spectral regions where the analyte of interest showed characteristic absorbance bands ($PLSR_a$). PLSR scaling parameters used throughout this study were exclusively selected by using the calibration data set. The number of factors (i.e., latent variables, LV) providing the lowest root mean square error of cross validation (RMSECV), calculated using venetian blinds (10 splits) was selected. The maximum number of LV assayed was 12. Further information on the theoretical background of PLS can be founded in [20].

2.4. LW-PLSR optimization and model calculation

During the development of the LW-PLSR approach the method employed for selecting the samples used in each calibration model and the regression method for predicting the analyte should be chosen in advance. In this work, the selection of the nearest neighbors based on the Euclidean distance measured in the PCA space, and PLSR were used for sample selection and model calculation, respectively. The number of principal components (F) used for the calculation of the Euclidean distances ranged between 1 and 8, the sample size of the local calibration models (N) varied between 20 and 240 in steps of 10, and a maximum number of 12 LVs was employed for local PLSR models. The optimal combination of the aforementioned three parameters was selected from results obtained by a multi-parametric approach using the RMSECV as response function. Accordingly, the combination of the values of the three parameters providing the lowest RMSECV calculated as described above, was selected and used for the evaluation of the PLSR predictive performance using the external validation set. Three LW-PLSR models namely $LW\text{-}PLSR_a$, $LW\text{-}PLSR_w$ and $LW\text{-}PLSR_m$, employing different spectral regions (see Section 3.1) for the selection of samples used for calibration and PLSR modeling, were calculated for each compound under study: $LW\text{-}PLSR_w$ models were constructed using the fingerprint region (900–1750 cm^{-1}) for sample selection and model calibration; for $LW\text{-}PLSR_a$ models, analyte characteristic spectral regions were employed for sample selection and model calibration; in $LW\text{-}PLSR_m$ the selection of the samples used for calibration was carried out employing analyte specific regions and PLSR was performed using the whole spectral range between 900 and 1750 cm^{-1}. A detailed explanation of the mathematical background of LW-PLSR can be found in [21–31].

3. Results and discussion

3.1. ATR-FTIR spectra of samples and analytes

For selecting analyte specific regions, spectra of aqueous standard solutions of glucose and urea were acquired. Furthermore spectra of albumin and tryolein standards were recorded (see Fig. 1) as representative compounds of the protein and triglyceride fraction in human sera, respectively. Representative spectra of all standard solutions are shown in Fig. 1. Accordingly, the selected intervals were the C–O stretching region for glucose (900–120 cm^{-1}), amide bands for proteins (1380–1690 cm^{-1}), C=0 and C–N stretching regions (1683–1560 and 1495–1405 cm^{-1}, respectively) for urea and C=0, C–H_2 and C–O stretching regions (1708–1760, 1420–1480 and 1128–1199 cm^{-1}, respectively) for triglycerides. Furthermore Fig. 1 shows the average spectrum of the calibration set, where a strong contribution of the protein signal overlaps with absorption bands of other compounds in the regions of Amide I–IV (1380–1690 cm^{-1}).

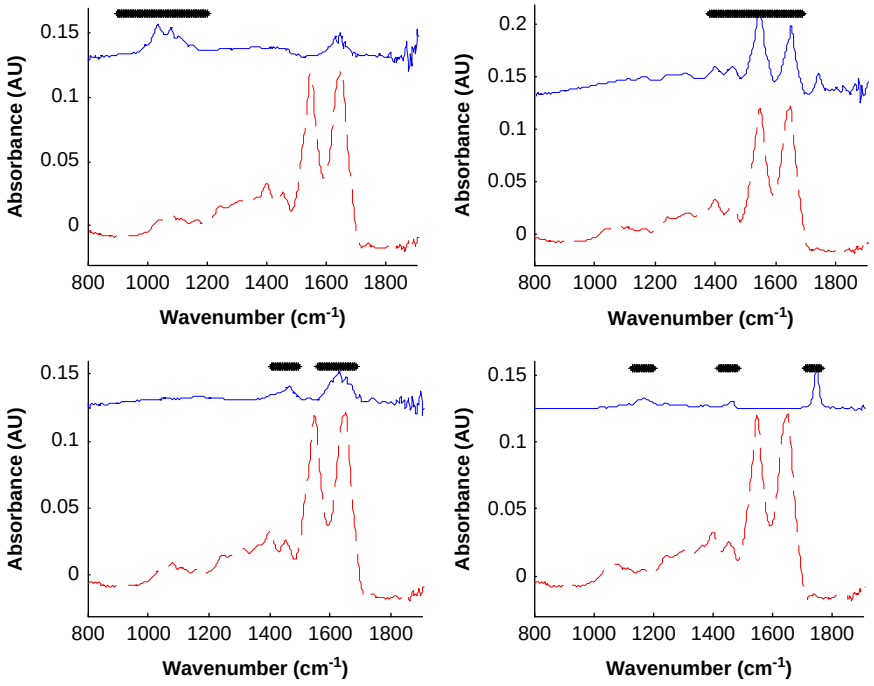

Fig. 1. Fingerprint region of infrared ATR spectra. Average spectra of the calibration set (red dotted line) and standard solutions (solid blue line) of 3 g/dL albumin (1), 100 mg/dL glucose (2), 15 mg/dL urea (3) and spectrum obtained from drying 2 μL of a 0.02% solution of tryolein in hexane (4) The blacks lines corresponds to intervals used for the modeling of each analyte. *Note*: Spectra were shifted along the *y*-axis to clearly show their bands. Additional information on the standards can be found in [15]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Sample analysis by PLSR

An initial evaluation of the predictive capabilities of PLSR models for the direct quantification of total protein, triglycerides, glucose and urea contents in human sera employing preprocessed ATR-FTIR spectra was performed. Based on the spectral information provided by reference spectra shown in Fig. 1, four spectral regions were selected (see Tables 2–5). Subsequently PLSR models were built using the calibration data set and results obtained were used as reference values for the evaluation of the performance of locally weighted regression models. Tables 2–5 summarize the main figures of merit for glucose, urea, total protein and triglycerides quantification. Obtained root mean square errors of prediction (RMSEP) for glucose, urea, total protein and triglycerides were 25, 13, 250 and 45 mg/L, respectively, in the same range as both, cross validation (CV) figures of merit (see Tables 2–5) and previously reported average RMSEP values of 23, 14, 254 and 51 mg/dL [15]. As indicated in a previous section, CV was performed for each model and the number of LVs that provided the lowest RMSECV was selected. Although this procedure might lead to model over-fitting, it was employed to be consistent with the procedure for the selection of the number of PLS LVs used for locally weighted regressions. Nonetheless, the performance of PLSR models was evaluated by using an external validation set.

3.3. Sample analysis by LW-PLSR

3.3.1. Selection of the LW-PLSR model parameters

The effect of the three parameters under study (*N*, *F* and LVs) on the predictive capabilities of LW-PLSR models estimated by CV is shown in Fig. 2. This plot evidences that LW-PLSR strongly affected the outcome of the analysis. Results found using the three LW-PLSR approaches (LW-PLSR$_w$, LW-PLSR$_a$ and LW-PLSR$_m$) for the determination of the four considered analytes confirmed

Table 2
Parameters and results obtained for multivariate determination of glucose in sera.

Approach	Region for sample selection (cm^{-1})	Region for regression (cm^{-1})	LV	*F*	*N*	RMSEC (mg/dL)	RMSECV (mg/dL)	RMSEP (mg/dL)
PLS$_a$	–	Glucose	12	–	–	23.5	25.1	24.7
LW-PLSR$_a$	Glucose	Glucose	7	7	240	10.3	23.4	22.2
LW-PLSR$_w$	Whole spectrum	Whole spectrum	12	2	240	11.2	19.3	19.7
LW-PLSR$_m$	Glucose	Whole spectrum	11	4	210	13.7	16.8	17.1

Note: Glucose region corresponds to 990–1200 cm^{-1} and whole spectrum to 900−1750 cm^{-1}.
LV: latent variables; *F*: number of factors; *N*: number of samples; RMSEC: root mean square error of calibration; RMSECV: root mean square error of cross validation; RMSEP: root mean square error of prediction.

Table 3
Parameters and results obtained for the multivariate determination of urea in sera.

Approach	Region for sample selection (cm^{-1})	Region for regression (cm^{-1})	LV	*F*	*N*	RMSEC (mg/dL)	RMSECV (mg/dL)	RMSEP (mg/dL)
PLS$_a$	–	Urea	12	–	–	12.2	13.0	13.4
LW-PLSR$_a$	Urea	Urea	6	5	240	7.9	11.4	11.6
LW-PLSR$_w$	Whole spectrum	Whole spectrum	10	2	240	6.7	10.7	11.3
LW-PLSR$_m$	Urea	Whole spectrum	10	7	230	7.1	9.3	9.5

Note: Urea region corresponds to 1683–1560 and 1495–1405 cm^{-1} and whole spectrum to 900–1750 cm^{-1}.
LV: latent variables; *F*: number of factors; *N*: number of samples; RMSEC: root mean square error of calibration; RMSECV: root mean square error of cross validation; RMSEP: root mean square error of prediction.

Table 4
Parameters and results obtained for the multivariate determination of proteins.

Approach	Region for sample selection (cm^{-1})	Region for regression (cm^{-1})	LV	*F*	*N*	RMSEC (g/dL)	RMSECV (g/dL)	RMSEP (g/dL)
PLS$_a$	–	Protein	6	–	–	0.210	0.221	0.253
LW-PLSR$_a$	Protein	Protein	3	120	3	0.183	0.212	0.243
LW-PLSR$_w$	Whole spectrum	Whole spectrum	3	130	3	0.180	0.215	0.240
LW-PLSR$_m$	Protein	Whole spectrum	3	100	3	0.206	0.216	0.240

Note: Protein region corresponds to 1380–1690 cm^{-1} and whole spectrum to 900–1750 cm^{-1}.
LV: latent variables; *F*: number of factors; *N*: number of samples; RMSEC: root mean square error of calibration; RMSECV: root mean square error of cross validation; RMSEP: root mean square error of prediction.

Table 5
Parameters and results obtained for the multivariate determination of triglycerides.

Approach	Region for sample selection (cm^{-1})	Region for regression (cm^{-1})	LV	*F*	*N*	RMSEC (mg/dL)	RMSECV (mg/dL)	RMSEP (mg/dL)
PLS$_a$	–	Triglycerides	12	–	–	42.7	48.7	45.0
LW-PLSR$_a$	Triglycerides	Triglycerides	8	5	240	24.1	42.8	43.2
LW-PLSR$_w$	Whole spectrum	Whole spectrum	10	6	240	14.0	45.5	50.6
L W-PLSR$_m$	Triglycerides	Whole spectrum	12	7	200	24.5	39.7	40.1

Note: Triglycerides region corresponds to 1708–1760, 1420–1480 and 1128–1199 cm^{-1} and whole spectrum to 900–1750 cm^{-1}.
LV: latent variables; *F*: number of factors; *N*: number of samples; RMSEC: root mean square error of calibration; RMSECV: root mean square error of cross validation; RMSEP: root mean square error of prediction.

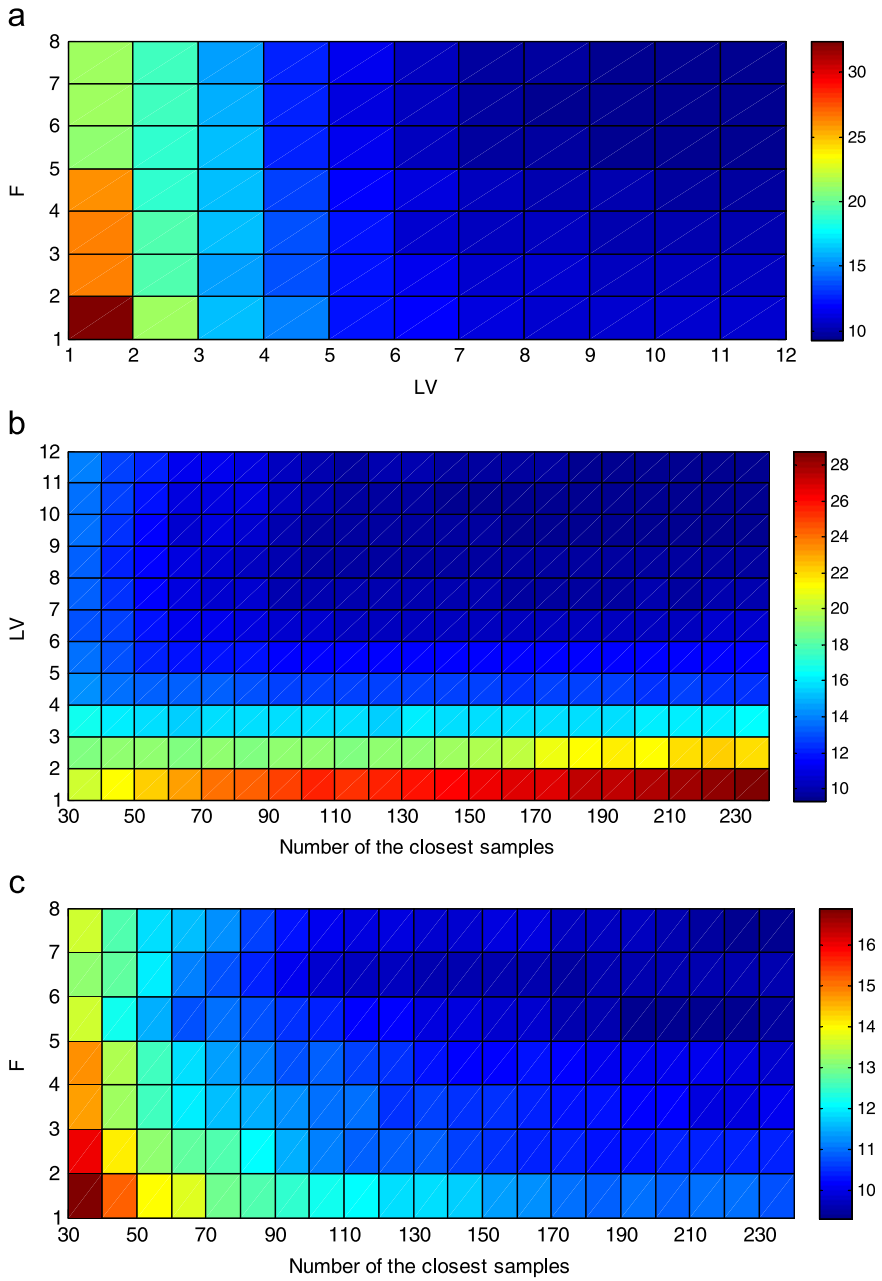

Fig. 2. Effect of the number of factors chosen for measuring the distance (F), the number of samples in each sub-model (N) and the number of latent variables (LVs) on the RMSECV for LW-PLSR modeling of urea using the analyte specific spectral region. F vs. LVs (for $N=240$) (a) LVs vs. N (for $F=6$) (b) and F vs. N (for LVs$=5$) (c).

the existence of local minima in the response surfaces. Thus, the evaluation of a set of combinations of LW-PLSR parameters simplifies the selection process of the optimal values. Despagne et al. [28] established that, for the prevention of over-dimensioned models, the complexity of the local models cannot be higher than the complexity of the global model, and additionally, RMSECV values for the local models must be lower than those of the global models. As shown in Tables 2–5, results obtained in this study, evidence that both conditions are accomplished for all evaluated LW-PLSR models.

3.3.2. *Comparison of different LW-PLSR approaches*

The use of local regression models improved both, RMSECV and RMSEP figures for the four considered analytes, as summarized in Tables 2–5 LW-PLSR$_a$ provided RMSEP values between 4 and 14% lower than those obtained by PLSR$_a$. A similar trend was observed for LW-PLSR$_w$ where the improvement was slightly higher and the RMSEP values decreased between 5 and 22% with respect to those obtained for PLSR (with the exception of triglycerides determination).

The improved prediction capability using the whole region instead of using the analyte specific region is difficult to explain. Fig. 3a shows the variable importance projection (VIP) scores of a PLS model built for glucose prediction using the whole spectrum. Although the most important variables are localized in the region of the specific C–O stretching band of glucose between 900 and 1100 cm^{-1}, an additional strong influence of other regions can be observed. This fact could be due to a correlation of glucose concentration values with other compounds such as proteins. However, the coefficient of correlation between glucose concentration and protein or urea values were found to be below 0.01 in both cases. Another justification could be the utilization of these

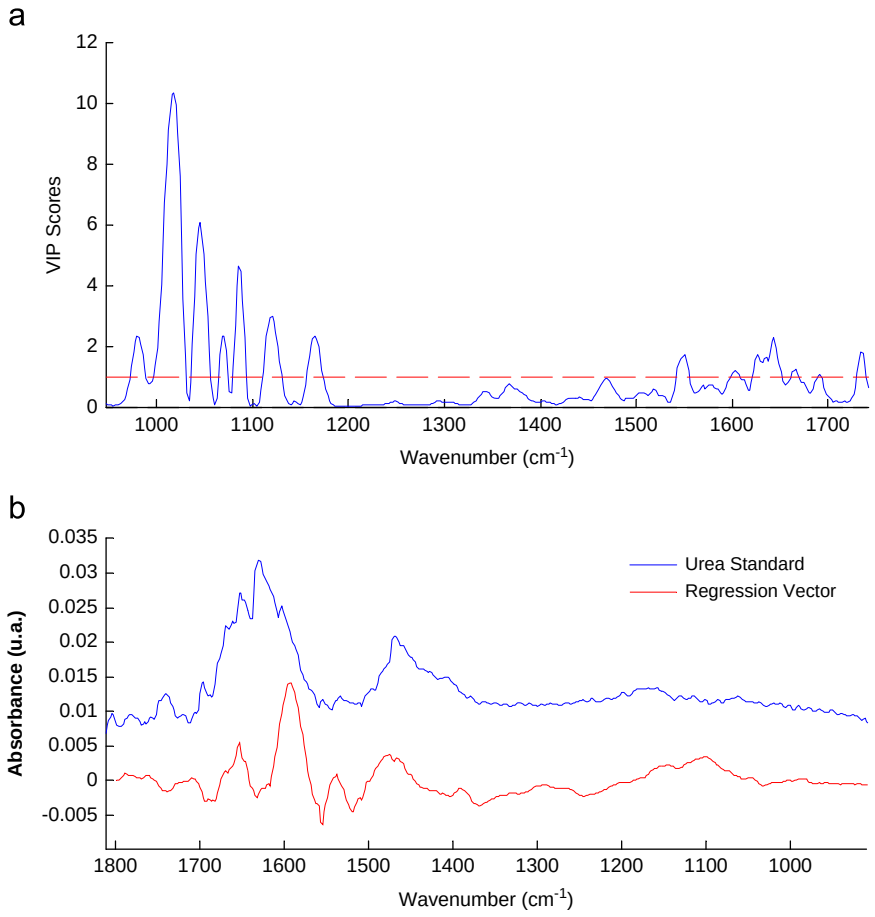

Fig. 3. Variable importance projection (VIP) scores of the LW-PLSR$_w$ model calculated for the determination of glucose using the whole spectrum (a) and comparison of the regression vector of the LW-PLSR$_w$ model obtained for the determination of urea (only applying mean centering in the preprocessing step) using the whole spectrum. *Note*: For an easier interpretation (b) also shows a spectrum of a 15 mg/dL standard solution of urea.

regions for compensating the matrix effect. Fig. 3b evidences negative values observed for the regression vector of urea models in regions where this analyte does not show any absorbance, especially in the region of the amide bands I and II, where the protein matrix shows a strong signal. Thus, this behavior could be associated with the compensation of the spectral contributions of other sample components.

3.4. *Results for the determination of glucose, urea, proteins and triglycerides in human serum obtained by LW-PLSR$_m$*

Fig. 4a evidences the good correlation of the first two PCA scores with protein concentration, which is the major component of serum. In contrast, when the glucose concentration is related to the PCA projection (see Fig. 4b) the aforementioned correlation cannot be observed. This was expected, since protein concentration is at least 10 times higher than that of any other analyte in human sera, thus showing a vast contribution to the total spectrum (see Fig. 1). Due to this distribution in the scores space, it seems difficult that LW-PLSR could properly select samples for building local models for minor components and it is not clear that LW-PLSR could compensate spectral interferences caused by the matrix. Likewise, it could explain the problems for obtaining local models integrated by samples with similar concentration values to the query in multi-component matrices described elsewhere [26].

In contrast, when PCA is built using the glucose specific spectral region, as shown in Fig. 4c, a correlation was appreciated between depicted scores and glucose concentrations, being samples with similar levels of glucose close to each other. This fact is also evidenced in Table 6 where correlations between y values of samples used for building local models and the query are shown after the selection of the 20 closest neighbors using Euclidean distance in the scores space. Coefficients of determination as well as standard deviations were strongly affected by the number of PCA factors selected and the region used for PC decomposition. If samples were selected according to their proximity in the y value, best sub-models were obtained when standard deviations of the y values of the selected samples were low and their mean close to the query concentration. In the case of proteins the use of low F values produced better results and, since proteins are the major compound of the matrix, the chosen region does not seem to affect the results. Regarding models for glucose prediction, the subset selection is much better when high F values were selected and only the glucose specific region was used. In summary, data show that in this case, PCA in the specific region of the analyte together with a careful selection of the PCs employed for the calculation of the distance between samples and query could help to perform an appropriate selection of similar samples to build local models, not only in the X-block space, but also in the concentration vector y.

So, from results obtained, a two-steps strategy (LW-PLSR$_m$) was selected involving: (i) the calculation of a PCA using analyte specific regions for the selection of appropriate calibration subsets, and (ii) the calculation of PLSR models using the whole spectrum.

Using this strategy, significant improvements in accuracy were obtained, as summarized in Tables 2–5: RMSEPs obtained for glucose, urea, proteins and triglycerides were 30, 29, 5 and 11% lower than those obtained for PLSR$_a$.

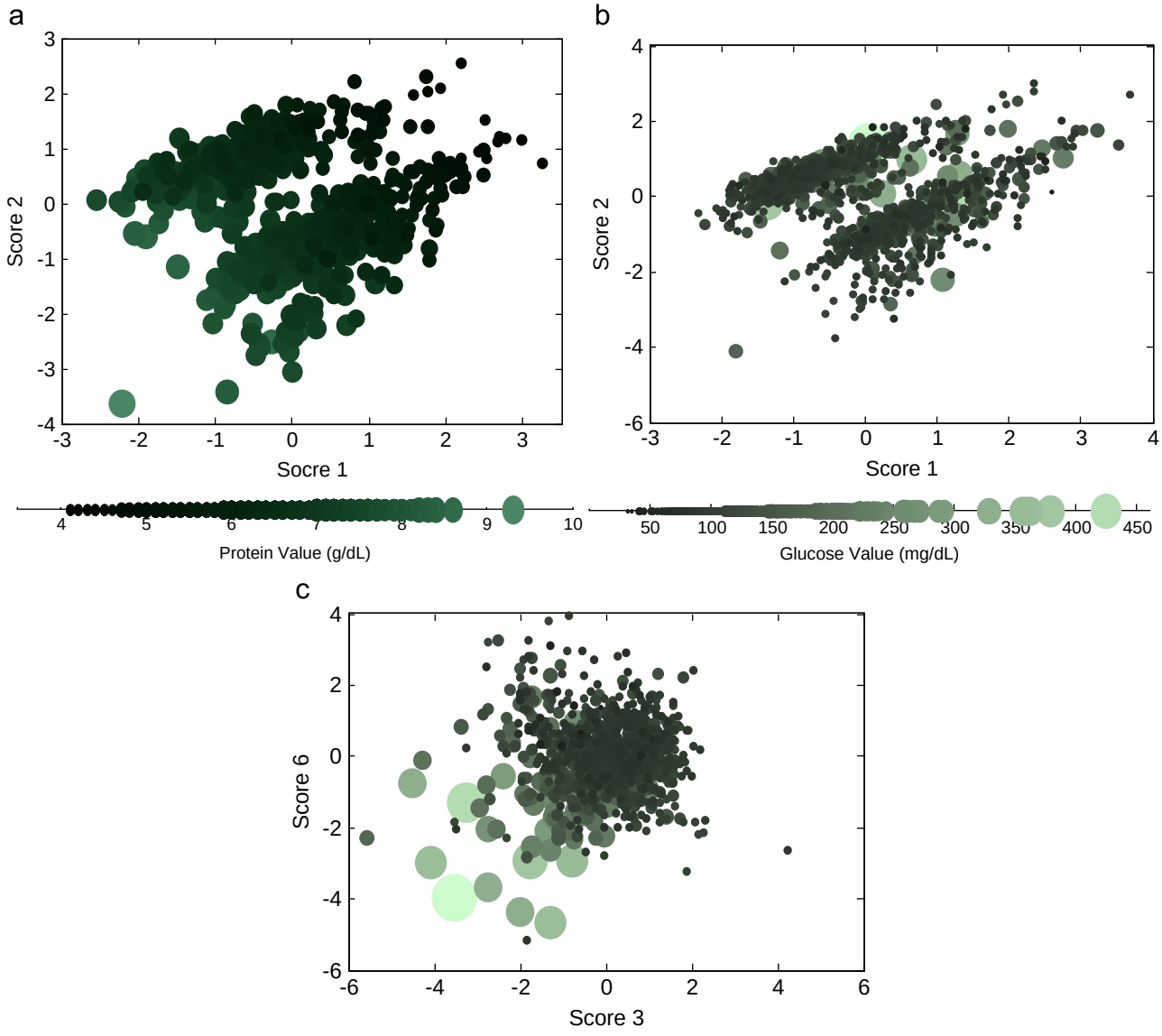

Fig. 4. PCA scores plot of PC1 vs. PC2 obtained using the whole spectrum and concentration values of proteins (a) and glucose (b) and relationship between PCA scores obtained employing the analyte specific region of glucose and the glucose concentration values (c).

Table 6
Relation between the *y* values of the selected samples and the query for PCA treatment of proteins and glucose concentrations in sera.

PCA (*N*)	Whole region		Specific region	
	R^2	SD	R^2	SD
Proteins (1–2)	0.934	0.345[b]	0.933	0.3477[b]
Proteins (1–8)	0.913	0.539[b]	0.917	0.5465[b]
Glucose (1–2)	0.052	36.33[a]	0.614	23.77[a]
Glucose (1–8)	0.211	33.61[a]	0.856	20.49[a]

Note. R^2: correlation coefficient between the mean value of the 20 closest neighbors of a sample in the scores space and the sample value. SD: mean of the standard deviation of the 20 closest samples selected for each sub-model. *N*: number of factors used for the PCA.

[a] mg/dL.
[b] g/dL.

4. Conclusions and outlook

In this work, the application of LW-PLSR to the analysis of clinical parameters in serum was evaluated for major and minor analytes. The comparison of RMSEP values obtained for classical LW-PLSR and PLSR evidenced a slight improvement of the errors when local regression was applied. Besides, after the identification of the problems of PCA for the selection of samples for local models when the whole spectrum was used, an alternative two-steps strategy ($LW\text{-}PLSR_m$) was selected based on the calculation of a PCA model using analyte specific regions for the selection of appropriate calibration subsets, and a subsequent calculation of PLSR models using the whole spectrum. By using this approach, RMSEP errors found were 5–30% lower than those obtained by PLSR. Although the optimization of LW-PLSR parameters can be computing intensive, this study evidences that LW-PLSR provides improved predictive capabilities for the determination of clinical parameters based on direct FTIR measurements of untreated serum samples.

Acknowledgements

Authors gratefully acknowledge the financial support of the Ministerio de Educación y Ciencia (***Project CTQ2011-25743***) and

the Generalitat Valenciana (**Project *PROMETEO 2010-055***). DPG acknowledges the "**V Segles**" grant provided by the University of Valencia to carry out this study. JK acknowledges her personal grant (Sara Borrell CD12/00667) from the *Instituto Carlos III* (Ministry of Economy and Competitiveness).Authors are grateful to Dr. Josep Ventura (Hospital Dr. Peset, Valencia) for providing samples and reference values used in this study.

References

[1] J. Wang, M. Sowa, H.H. Mantsch, A. Bittner, H.M. Heise, Trac-Trend. Anal. Chem. 15 (1996) 286–296.
[2] D. Rohleder, G. Kocherscheidt, K. Gerber, W. Kiefer, W. Kohler, J. Mocks, W. Petrich, J. Biomed. Opt. 10 (2005).
[3] D. Naumann, AIP Conf. Proc. 430 (1998) 96–109.
[4] L.Q. Wang, B. Mizaikoff, Anal. Bioanal. Chem. 391 (2008) 1641–1654.
[5] G. Deleris, C. Petibois, Vib. Spectrosc. 32 (2003) 129–136.
[6] Y.Z. Li, R. Chen, L. Liu, S.Y. Feng, B.H. Huang, P. Soc. Photo. -Opr. Ins. 5630 (2005) 229–234.
[7] W. Petrich, Appl. Spectrosc. Rev. 36 (2001) 181–237.
[8] H.J. Gulley-Stahl, S.B. Bledsoe, A.P. Evan, A.J. Sommer, Appl. Spectrosc. 64 (2010) 15–22.
[9] H.H. Mantsch, L.P. Choo-Smith, R.A. Shaw, Vib. Spectrosc. 30 (2002) 31–41.
[10] G. Janatsch, J.D. Kruse-Jarres, R. Marbach, H.M. Heise, Anal. Chem. 61 (1989) 2016–2023.
[11] D. Perez-Guaita, J. Ventura-Gayete, C. Pérez-Rambla, M. Sancho-Andreu, S. Garrigues, M. de la Guardia, Anal. Bioanal. Chem. 404 (2012) 649–656.
[12] K.Z. Liu, R.A. Shaw, A. Man, T.C. Dembinski, H.H. Mantsch, Clin. Chem. 48 (2002) 499–506.
[13] E. Diessel, P. Kamphaus, K. Grothe, R. Kurte, U. Damm, H.M. Heise, Appl. Spectrosc. 59 (2005) 442–451.
[14] K.H. Hazen, M.A. Arnold, G.W. Small, Anal. Chim. Acta. 371 (1998) 255–267.
[15] D. Perez-Guaita, J. Ventura-Gayete, C. Pérez-Rambla, M. Sancho-Andreu, S. Garrigues, M. de la Guardia, Microchem. J. 106 (2013) 202–211.
[16] G. Hosafci, O. Klein, G. Oremek, W. Mantele, Anal. Bioanal. Chem. 387 (2007) 1815–1822.
[17] M. Mecozzi, E. Pietrantonio, M. Amici, G. Romanelli, Analyst 126 (2001) 144–146.
[18] G. Mazarevica, J. Diewok, J.R. Baena, E. Rosenberg, B. Lendl, Appl. Spectrosc. 58 (2004) 804–810.
[19] M. Meinke, G. Mueller, H. Albrecht, C. Antoniou, H. Richter, J. Lademann, J. Biomed. Opt. 13 (2008) 014021.
[20] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[21] P. Berzaghi, J.S. Shenk, M.O. Westerhaus, J. Near Infrared Spectosc. 8 (2000) 1–9.
[22] D. Pérez-Marín, A. Garrido-Varo, J.E. Guerrero, Talanta 72 (2007) 28–42.
[23] W. Cleveland, S. Delvin, J. Am. Stat. Assoc. 83 (1988) 569–910.
[24] T. Isaksson, T. Naes, Appl. Spectrosc. 42 (1988) 1273–1284.
[25] T. Naes, T. Isaksson, B. Kowalski, Anal. Chem. 62 (1990) 664–673.
[26] Z. Wang, T. Isaksson, B.R. Kowalski, Anal. Chem. 66 (1994) 249–260.
[27] V. Centner, D.L. Massart, Anal. Chem. 70 (1998) 4206–4211.
[28] F. Despagne, D.L. Massart, P. Chabot, Anal. Chem. 72 (2000) 1657–1665.
[29] K. Fujiwara, M. Kano, S. Hasebe, Chemometr. Intell. Lab. 101 (2010) 130–138.
[30] S. Kim, M. Kano, H. Nakagawa, S. Hasebe, Int. J. Pharm. 421 (2011) 269–274.
[31] H. Nakagawa, T. Tajima, M. Kano, S. Kim, S. Hasebe, T. Suzuki, H. Nakagami, Anal. Chem. 84 (2012) 3820–3826.