Evaluation of Non-linearity in MIR Spectroscopic Data for Compressed Learning

Dixon Vimalajeewa, Donagh Berry, Eric Robson, Chamil Kulatunga
Telecommunications Software and Systems Group, Waterford Institute of Technology, Waterford, Ireland
Teagasc, Animal & Grassland Research and Innovation Centre, Moorepark, Fermoy, Co. Cork, Ireland
Email: dvimalajeewa@tssg.org, Donagh.Berry@teagasc.ie, {erobson, ckulatunga}@tssg.org

Abstract—**Mid-Infrared (MIR) spectroscopy has emerged as the most economically viable technology to determine milk values as well as to identify a set of animal phenotypes related to health, feeding, well-being and environment. However, Fourier transform-MIR spectra incurs a significant amount of redundant data. This creates critical issues such as increased learning complexity while performing Fog and Cloud based data analytics in smart farming. These issues can be resolved through data compression using unsupervisory techniques like PCA, and perform analytics in the compressed-domain i.e. without de-compressing. Compression algorithms should preserve non-linearity of MIRS data (if exists), since emerging advanced learning algorithms can improve their prediction accuracy. This study has investigated the non-linearity between the feature variables in the measurement-domain as well as in two compressed domains using standard Linear PCA and Kernel PCA. Also the non-linearity between the feature variables and the commonly used target milk quality parameters (Protein, Lactose, Fat) has been analyzed. The study evaluates the prediction accuracy using PLS and LS-SVM respectively as linear and non-linear predictive models.**

1. Introduction

Advances in pervasive computation and communication technologies with IoT systems result in rapid adoption of Fog/Edge computing based data analytics to discover near real-time insights in smart farming [1]. The opportunity of collecting and analyzing millions of high-resolution data demands distributed analytics across the resource-constrained Fog devices rather than centralizing raw data. Therefore efficient data storage, communication and processing techniques are vital [2] in Distributed Learning (DL) [6] compared to learning by centralizing data of such applications. This is not only because of scalability, but also due to significant contributions towards energy optimization [3], [12]. Instead of aggregating raw data, DL aggregates rich features from each data source to discover high quality global knowledge. The success of DL depends on the accuracy of knowledge aggregation at the same level where centralized learning could achieve. Therefore, one of the important task in DL is to prepare data in a compressed feature space that enables to maximize information extraction while minimizing computation, communication and storage resource consumption [2], [4].

Pasture-based dairy farming is one of the industries, which has distributed data sources in a large terrain and essentially requires such optimized systems to accelerate current farming strategies [7]. In smart dairy farming, farms are being adopted with the new technologies such as per-animal based milk yield and quality monitoring, sensor-based animal behaviour tracking [5] and robotic milking etc. to improve the quality and efficiency of dairy production. Among them Mid-Infrared Spectroscopic (MIRS) milk quality monitoring and its association analysis with other factors is vital for milk value analysis and for identifying associated phenotypes [8]. To apply DL on these datasets, a *Compressed Learning* (CL) approach (explain in Section 2) is commonly used to extract descriptive features from the raw data. Prior knowledge of the general characteristics of data is essential for a lossy CL approach to retain the precision of learning.

According to the literature [13], [14], [15], the linear/non-linear behaviour of data has a considerable impact on the accuracy of the final learning outcomes. The purpose of most of these studies were very generic because they were based on the fact that non-linear machine learning algorithms have better performances than linear techniques regardless of their complexity and the required computational power. However, linear approaches could achieve the same precision as non-linear techniques with lesser computation. However, recent data analytics, which are capable of doing complex learning with modern computational power, pay attention to employ the most accurate learning approach. Therefore, understanding the original characteristics of the data in particular, non-linearity in CL is vital.

In this study, we investigated the linear and non-linear behaviours of MIRS dataset (Fig. 1) in the context of milk quality predictions. First, pre-processing removed the impact of water absorbances from our dataset. Then non-linearity between the features in measurement-domain as well as in the compressed-domain were investigated for different milk quality parameters. Then the CL approach was used to perform learning from the compressed data, which reduced learning complexity. The impact of non-linearity were taken into account during the data compression based on linear (standard) principal component analysis (LPCA) and Kernel PCA (KPCA) techniques. The learning accuracy of using compressed-domain data was explored with a linear and a non-linear statistical predictive models; partial least square (PLS) and least squares support vector machine (LS-SVM). Section 1 has provided an introduction to the paper with its

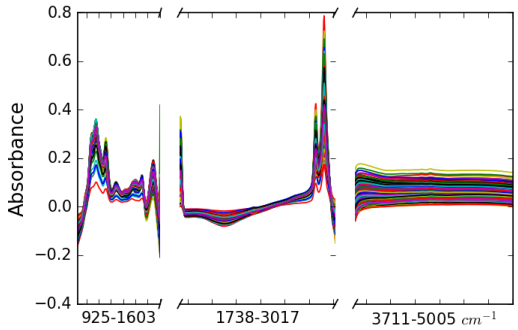

Fig. 1. Water-removed MIR spectra (X) of 712 milk samples in the wave region $925 - 5005cm^{-1}$. Water removal pre-process has reduced the feature-space dimensionality from 1060 wavenumbers to 847.

motivation. The remainder of the paper has been structured as follows. Section 2 discusses the significance of non-linearity in MIRS data and its importance in CL. Section 3 provides the methodologies we used to analyze non-linearity in MIRS milk quality predictions. Section 4 provides the analytical results based on our MIRS data followed by the conclusions in the Section 5.

2. Non-linearity and CL in MIRS

The main objective of the traditional data compression techniques is to reduce data storage and communication requirements as much as possible while minimizing information losses. These compression techniques do not contribute much to reduce learning complexity as de-compression was performed to a similar complexity prior to the learning process. The challenge of performing efficient data analytics with higher dimensions with highly redundant data remains unchanged. CL concept can be used to overcome this issue in Fog/Edge computing and in big data analytics. In CL, the original data (measurement-domain) is compressed while preserving the original learning accuracy. De-compression can be postponed until only if it is necessary. Thus, CL significantly reduces learning complexity. The data reduction techniques such as PCA and Wavelet Transformation (WT) has been used for CL [21].

Fig. 2 illustrates the CL process with unsupervised PCA compression where the measurement-domain and the compressed-domain data can either be in the same or different processing entities. Suppose matrix $Y_{n\times p}$ contains data for p response variables and need to build a regression model for those in Y (e.g. Lactose, Protein, Fat milk quality parameters) using $X_{n\times m}$ with m feature variables and n data samples. In order to preserve the original information and improve the learning performance in the compressed-domain, (l, G, P) from the compression should represent the original characteristics of X as much as possible.

In general, CL can be performed either in a single processing entity or in many geo-distributed processing entities. In a single processing scenario, both compression and learning can be supervisory since the compression unit is aware of what the compressed data is used for. Therefore, an optimal compression can be performed and continue to the learning process. In distributed scenario, compression and learning may

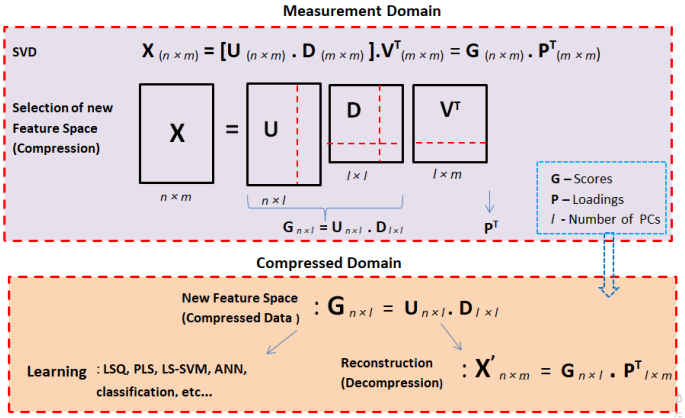

Fig. 2. The matrices of SVD used in classical PCA for deriving our compressed-domain data, which later be used in CL to predict MQPs.

be performed independently at two different locations. Unsupervised compression and a supervised learning should be employed since the compression unit may not know what will be the learning purpose. Therefore the compression entity will not aware of the most specific and relevant information required for the learning. It may neglect generally the least significant information according to the properties of the compression algorithm (e.g. variants in PCA). In order to achieve a robust analytic outcome by extracting the most accurate information, a proper understanding of X with Y is important because it helps to form well-represented compressed data and then perform learning in CL [24].

Performing a comprehensive pre-analysis with careful attention at all possible characteristics such as non-linearity, redundancy (including co-linearity), scaling and normalization of data helps to understand the data before applying CL. Therefore, such analysis overcomes the most decisive challenge in CL to select suitable compression and learning techniques based on the underlying behaviour of data. Understanding non-linearity between the feature variables and with the response variables can make a significant impact on the accuracy of CL. If a linear compression technique is used on the dataset without knowing that data has non-linear behaviours, compression may loose non-linear property of the original data. The information losses can be minimized by first understanding the behaviour though a pre-analysis and using a non-linear compression to preserve both linear and non-linear characteristics.

Most of the past studies in MIR spectrometry have followed the centralized analytics. *Y. M. Chen et al.* [13] studied non-invasive determination of sorghum species with different dimensionality reduction techniques and non-linear predictive models. Their study proved that the concern about non-linearity of MIRS sorghum data contributed for dimensionality reduction as well as improving the robustness of learning outcomes. The authors in [14] also considered non-linear associations between melamine content and MIRS spectra of dairy products (liquid milk, milk powder and infant formula). The generalization performance of linear and non-linear dimensionality reduction with a non-linear learning technique (SVM) has been studied by *L. J.*

Cao et al. [15]. They explored non-linear dimensionality reduction methods (KPCA and ICA) to capture higher order information of the input signal than linear methods (PCA). As a result, they were able to improve the generalization performance of their predictive models. In this study, we looked at non-linearity of MIRS data used in DL scenarios using CL.

3. Evaluation Methodologies

In this study, we first analyzed linear and non-linear associations between the measurement-domain variables in X and then between each compressed-domain significant feature variables (G) for three selected target variables Y (Protein, Fat and Lactose). The linear/non-linear correlation coefficients, PCA reconstruction error and non-linearity rate (NLR) measures were used with unsupervised CL (only needed X). Partial residual plots (PRP) and *Durbin-Wotsan* (DW) test were used with supervised approach (needed both X and Y) to describe the impact of non-linearity using LPCA and KPCA. PLS and LS-SVM learning approaches were used to examine the quality of compression based on non-linearity of the compressed data.

3.1. Linear/Non-linearity Evaluation Measures

Correlation Coefficients: There are different types of correlation measures such as *Pearson's correlation* (*cor*) and *Maximal correlation* (*mcor*), which are used for different purposes. *cor* captures only the linear correlation between random variables (generally called as the correlation coefficient), which is a statistical measure used to quantify association between random variables $X_i, X_j \in \mathbb{R}$,

$$cor(X_i, X_j) = \frac{cov(X_i, X_j)}{\sqrt{var(X_i)}\sqrt{var(X_j)}} \tag{1}$$

$cor(X_i, X_j) = 0$ does not mean that there is no association because *cor* cannot detect if there is a non-linear association. *mcor* enables measuring non-linear correlations by transforming the data, where associations are not detectable in the original data space and is defined as;

$$mcor(X_i, X_j) = max_{f,g} cor(f(X_i), g(X_j)) \geq 0 \tag{2}$$

where $f, g \in \mathbb{R} \rightarrow \mathbb{R}$ are two functions selected so that they maximize the correlation of X_i and X_j. If there are non-linear associations, $mcor \geq cor$ and otherwise $mcor = |cor|$. The Alternating Conditional Expectation (ACE) algorithm was used to compute *mcor* in our evaluations. In this study, *cor* and *mcor* measures [16] were used to recognize linear and non-linear associations in our MIRS data X.

NLR: NLR is a quantitative measure for the degree of non-linearity in data. Most of the measures of non-linearity are based on the residuals from linear and non-linear regression fittings. The residual difference between two fittings gives an idea about the non-linearity. According to [11], NLR can be defined assuming that non-linear techniques fit perfectly to the data (i.e. non-linear fitting residual error is nearly zero).

$$\begin{aligned} NLR &= \frac{1}{n\sigma}\sum_{i=1}^{n}(||L_i - X_i||^2 - ||H_i - X_i||^2) \\ &\simeq \frac{1}{n\sigma}\sum_{i=1}^{n}||L_i - X_i||^2 \end{aligned} \tag{3}$$

where n number of data points (X_i), L_i and H_i are supporting points of linear and non-linear regression fittings, respectively. $\sigma = \frac{1}{n}\sum_{i=1}^{n}||X_i - \mu||^2$ is the variance of data X and μ is the mean of X. The Equation (3) indicates the amount of residuals from linear fitting. Higher NLR will result in higher non-linearity and vise-versa. Suppose the linear fitting is LPCA, then NLR can be derived as follows.

$$NLR = 1 - \frac{\sum_{i=1}^{l}\lambda_i}{\sum_{i=1}^{m}\lambda_i} \tag{4}$$

where λ is the i^{th} eigenvalue computed form the covariance matrix of X and l is the selected dimension for $l = 1, 2, \cdots, m$. The proof of this formula can be found in [11].

PRP: A plot obtained using Least Squares Regression (LSR) fitting can be used to understand the usefulness of the LSR model parameters and their unknown functional forms (e.g. non-linearity). According to [23], partial residuals (component+residuals) are the residuals of a LSR model fitting added to the mis-specified part of the model. PRPs are the plots of partial residuals against the mis-specified part. Suppose a LSR model in the form;

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + f(x_k) + \epsilon \tag{5}$$

where $f(x_k), k = 1, \cdots, m$ is an unknown function to be identified (mis-specified part), β_i's are the LSR model parameters of the predictor variable x_i's for $i = 1, \cdots, m$ (β_0 is intercept) and ϵ is the random error. PRP of f gives an graphical overview regarding the effect of f to y when the effect of all other x_is are controlled. This concept was used to check the relationship of each feature to their corresponding response.

DW Test: This statistical test is used as a measure of auto-correlation (ρ) of residuals from a LSR fitting to check whether there is a correlation between the successive residuals. Since, residual ρ indicates the goodness of LSR fit, this can be used as a technique to identify the relationship (linear/non-linear) of response variables to its feature variables. The null hypothesis states $H_0 : \rho = 0$ and alternative hypothesis states $H_1 : \rho > 0$. The test statistic d is computed by,

$$d = \frac{\sum_{i=1}^{n-1}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{6}$$

where e and n reflect a residual and the number of samples. If $d < d_L$, H_0 is accepted (residuals are uncorrelated and normality exists in the model). H_0 is

rejected, if $d > d_U$, which reflects that there exists a correlation in residuals and linearity in the model. The test is inconclusive, if $d_L < d < d_U$. The d_L and d_U are lower and upper critical values for the test [19].

3.2. Data Compression and Regression Methods

PCA is used for dimensionality reduction, visualizations, compression (with loss), de-noising (removing small variance in the data) and whitening (de-correlation so that features have unit covariance). PCA is a variance-based statistical dimensionality reduction technique. It draws a low dimensional space and represents each data point by its projection along the orthogonal directions, which represents maximal variance of the data. The low dimensional space is called compressed-domain feature space and the projections along the directions are called principal components (PCs). We used LPCA and KPCA [10] in our evaluations of CL.

LPCA: Fig. 2 shows the process of LPCA with singular value decomposition (SVD) within the measurement-domain entity. Given the mean-centered data set ($X_{n\times m}$), SVD decomposes X in the form $X = UDV^T$, where U, V are respectively upper and lower triangular matrices where $U^TU = I = V^TV$ (I is an identity matrix). D is a diagonal matrix in which elements follow the condition $d_{11} \leq d_{22} \leq \cdots \leq d_{mm}$ (eigen values of the co-variance matrix of X). Then the score matrix $G = UD$ (compressed feature space) and the loading matrix $P = V$ are derived. Finally, the data is transformed into its compressed-domain by selecting the scores of the significant l ($< m$) PCs, which minimizes the reconstruction error. LPCA is based on the assumption that correlations are linear.

KPCA: When data has complex non-linear associations, which is more realistic in practical datasets such as MIRS data, KPCA like non-linear feature extraction methods have to be used for data compression in CL [10], [15], [20]. It has been proved in many studies that non-linear methods perform well in dimensionality reduction by capturing global characteristics in data [15]. In KPCA, the original data matrix $X_{n\times m} \in \mathbb{R}^m$ is mapped into a new higher dimensional space (feature space) $\mathbb{F}^M$ by a non-linear function ϕ such that,

$$\phi\colon \mathbb{R}^m \rightarrow \mathbb{F}^M \tag{7}$$

For a certain selection of ϕ, $\mathbb{F}^M$ has arbitrarily large dimension and then LPCA is performed using "*kernel trick*". According to the *Mercer*'s theorem, non-linear mapping function ϕ and the kernel function K are associated by the equation $K(x^i, x^j) = \phi(x^i)^T \cdot \phi(x^j)$. Given the kernel function K, the normalized kernel matrix $\hat{K}_{m\times m}$ of the data $X_{n\times m}$ is computed as follows.

$$\hat{K} = K - 2I_{1/n}K + I_{1/n}KI_{1/n} \tag{8}$$

where $I_{1/n}$ is a matrix with all elements $1/n$. Then LPCA is applied on $\hat{K}$ in the feature space, which is equivalent to non-linear PCA in the original data domain. There are different types of kernel functions such as Gaussian (radial basis - RBF) and polynomial [15] where the RBF kernel; $K(x, x_i) = exp(-||x - x_i||^2/\sigma^2)$ was used with $\sigma^2 = 0.25 \times m \times mean(var(X))$ in our MIRS data compression.

PLS: The ordinary LSR derives a relationship between X and Y with the assumption that X variables are uncorrelated. However, since some data such as MIRS data violates this assumption, PLS builds regression models by considering correlations of variables in X itself as well as between X and Y. Therefore PLS is considered as a bilinear modelling method in which X data is projected into a feature space (or latent variables, LVs) and then simplify relationship between X and Y to predict Y selecting least number of LVs via cross-validation. First, decompose both X and Y as the decomposition was performed in LPCA;

$$X = TP^T + H \qquad Y = RQ^T + L \tag{9}$$

where, T and R are the score matrices and P and Q are the loading matrices. H and L are respectively the error matrices, which come from the process of PLS regression of X and Y. Then, LSR is applied for scores T and R such that $R = WT + e$, where W and e are respectively the weight matrix (to be estimated) and the error term, which fits a LSR model for X and Y [9].

LS-SVM: As same as the process explained in KPCA, when data has complex non-linear associations, linear models cannot capture them properly. Therefore, LS-SVM is used to form a regression model in the feature space $\{\phi(x^i)\}_{i=1}^n$. The regression model in LS-SVM is given by,

$$y(x) = W^T\phi(x) + b \tag{10}$$

where $W \in \mathbb{R}^n$ is the weight vector and b is the bias. LS-SVM is an optimized algorithm based on the standard SVM [18]. The optimization problem is formulated as follows.

$$minJ(W, e) = \frac{1}{2}W^TW + \frac{1}{2}\gamma\sum_{i=1}^{n} e_i^2$$

where γ is the regularization parameter and e_i is the random error. The *Lagrange multiplier* method is used to solve the optimization task in the LS-SVM algorithm.

$$L(W, b, e, \alpha) = J(W, e) - \sum_{i=1}^{n} \alpha_i\{W^T\phi(x_i) + b + e_k - y_k\} \tag{11}$$

where α_i is Lagrange multipliers. The above Equation (11) is solved by partial differentiation with respect to each variable. Then estimation function of y can be obtained as,

$$y(x) = \sum_{i=1}^{n} \alpha_k K(x, x_k) + b; \quad i, j = 1, 2, 3, \cdots, n \tag{12}$$

where K is the kernel function. The selection of the parameter values γ and σ (RBF kernel parameter) is important. This is because γ improves the generalization performance of the model and σ controls the regression error and also reflects the sensitivity of LS-SVM model due to noise in input variables [17]. Thus, large γ and σ reflect respectively more non-linear model and global properties. There are different techniques to set parameter values in LS-SVM model such as cross-validation, grid search, Bayesian optimizer [22].

4. Evaluation Results

4.1. MIR Spectroscopic Milk Quality Data

The data used in this paper has been obtained from Teagasc research dairy farm at Moorepark, Ireland where MIR spectra was collected (in 35 days starting from August 2013 and ending in August 2014) using 605 different dairy cattle. The composition of milk was determined using FOSS MilkScan prediction equations using FT-MIR technology. The input data matrix contained the spectra of 712 different milk samples in the wavenumber region $925 - 5005cm^{-1}$ with a resolution of $3.853cm^{-1}$. When the wavenumbers were rounded to the nearest integer, a given spectrum contained 1060 transmittance data points. Therefore, the original MIRS spectra used (called the gold standard) to apply compression algorithms was a (712×1060) dimensional matrix.

Since spectral values were given in transmittance, we converted them to absorbance by taking log_{10} of the reciprocal of given transmittance values. According to the impact of water absorption in MIRS at $25°C$, two corresponding wave regions were removed as $1607 - 1734cm^{-1}$ and $3021 - 3707cm^{-1}$. This reduced our spectra to 847 wavenumbers, which we used as the input data matrix X (Fig. 1) in our analysis. In addition, the percentages of the selected MQPs corresponding to each sample were stored in a matrix (Y). Among them three most commonly used MQPs; Lactose, Protein and Fat were taken into the evaluations. Then our data compression and regression model calibration/validation were applied on this gold standard data. *R*-software was used for non-linearity analysis and *MATLAB* was used for PLS and LS-SVM model building and evaluations.

4.2. Non-linearity in Measurement-domain Data

To emphasize that there are linear and non-linear correlations in X, the *cor* and *mcor* were computed for all every pairs of wavenumbers in X. The results are shown in Fig. 3 with their absolute differences. According to the variations of color intensity, there are high and low variations respectively in the regions 925-$3025cm^{-1}$ and 3025-$5005cm^{-1}$. The figure shows both linear and non-linear correlations in the region 925-$3025cm^{-1}$ due to strong *mcor*s values. Even though $mcor \geq cor$ among the wavenumbers in both regions 925-$3025cm^{-1}$ and 3025-$5005cm^{-1}$, the region 925-$3025cm^{-1}$ shows a higher variation. Both correlations seem to be similar (linear/no correlation) among the

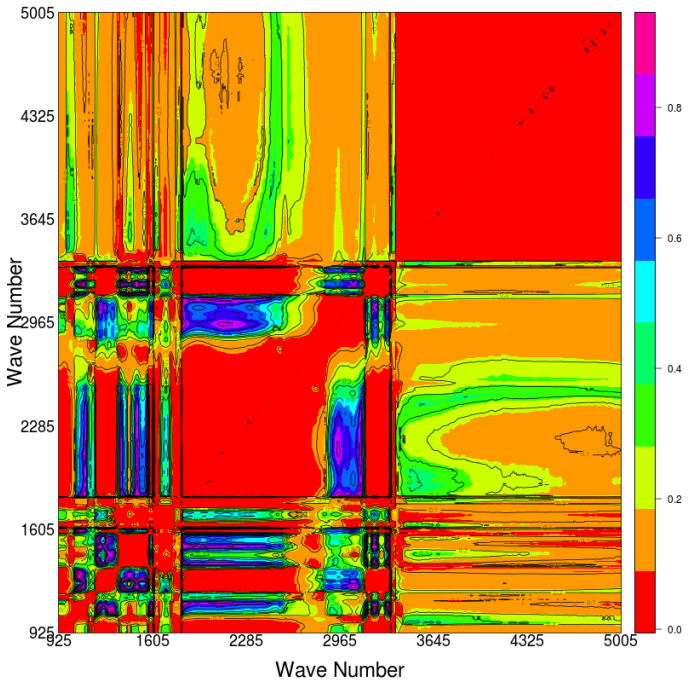

Fig. 3. The absolute difference of Maximal (*mcor*) and Pearson's (*cor*) correlation coefficients ($|mcor - cor|$) between feature variables in X.

wavenumbers in the region 3025-$5005cm^{-1}$ and some in 952-$3025cm^{-1}$ since there are no much color variations (red regions). Within the region of 925-$3025cm^{-1}$, at some points the correlation difference is even greater than 0.6. Maximum of 0.96 was observed between the wavenumbers $1387cm^{-1}$ and $1152cm^{-1}$ (the strongest *mcor*).

In general, there are both linear and non-linear associations in X and in particular, more non-linear correlations exist among the wavenumbers up to the $3025cm^{-1}$. The correlations between the wavenumbers above $3025cm^{-1}$ are not much stronger in terms of non-linearity.

To explore the importance of the linear/non-linear correlations of the feature variables in X, PLS regression was applied and the regression coefficients (βs) were derived for each MQP. Then the correlations of the wavenumber at significant βs (e.g. $\beta \geq 3\sigma_\beta$) with the other features were computed for each MQP. Fig. 4 shows the βs of each MQP and the absolute correlations (*mcor*s and *cor*s) of wevenumbers at the highest significant β (Lactose - $1745cm^{-1}$, Fat - $1734cm^{-1}$ and Protein - $1541cm^{-1}$) with other coefficients.

In each plot, $mcor \geq cor$ for all βs and most of the correlations are high and fluctuated sharply within the region 925-$3012cm^{-1}$ compared to the correlations of the βs above the wavenumber $3710cm^{-1}$. The plot for Protein are clearly non-linear because the differences between *mcor* and *cor* are higher for many βs. Even though there is no much non-linearity in the plots for Lactose and Fat compared to Protein, correlations in the region 2730-$2817cm^{-1}$ show a clear non-linearity. Thus these plots reveal that the correlations associated with the most significant β are non-linear for Protein compared to the correlations associated with the most

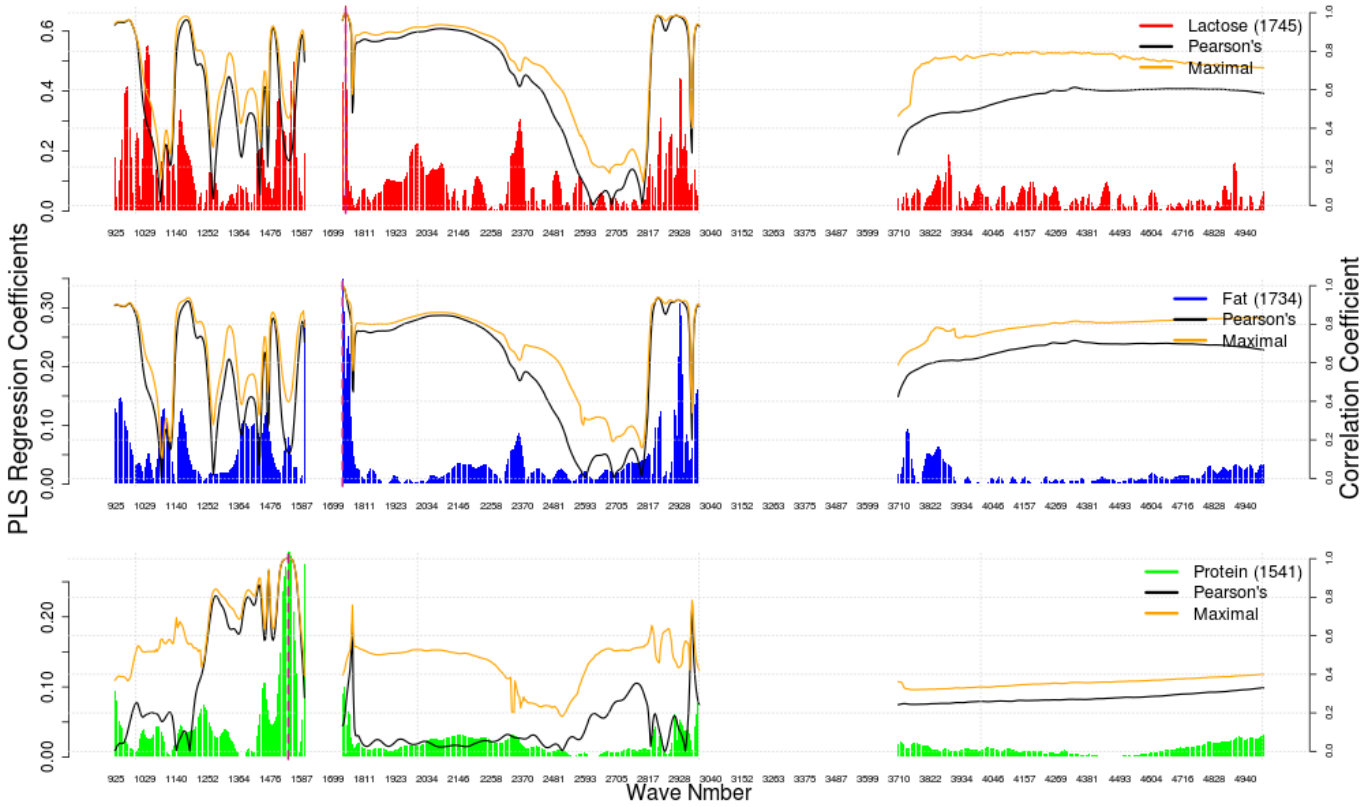

Fig. 4. PLS regression coefficients and the absolute values of *mcor* and *cor* correlations for Lactose, Fat and Protein. The shown correlation coefficients are between the wavenumber at the most significant regression coefficient with the rest of the wavenumbers.

significant βs of Lactose and Fat.

With regard to βs of each MQP, the highest significant βs lie in the regions where both *mcor* and *cor* are more or less similar (linear correlation) in each plot. There are many coefficients those lie where the correlations are non-linear. For instance, correlations of βs of Protein in the regions 925-1250cm^{-1} and 2034-2370cm^{-1} contain most of the significant coefficients. If data is compressed with LPCA, then the wavenumbers in these regions will likely to be removed due to lesser *cor* correlations. As a consequence, a high information loss can happen. Non-linear KPCA may be able to capture those non-linear as well as linear associations. Therefore, KPCA compressed data may retain more characteristics from the original data than LPCA. Therefore, it is important to understand the non-linearity as a prior knowledge before applying CL.

4.3. Performance of Linear/Non-linear PCA

After analyzing the behaviour of correlations in the dataset X, its impact on the PCA compression was investigated. The amount of information captured by LPCA and KPCA algorithms were considered by computing REs for the first 100 PCs. The results are shown in Fig. 5 (left). According to the figure, REs of KPCA is less than that of LPCA. It turns out that KPCA incurs lesser REs with a lower number of PCs than in the LPCA. According to KPCA, this is due to existence of non-linearity in X. For instance, REs of LPCA and KPCA with 20 PCs are respectively 5.9×10^{-4} and 8.6×10^{-7}. Therefore, LPCA needs more PCs to achieve

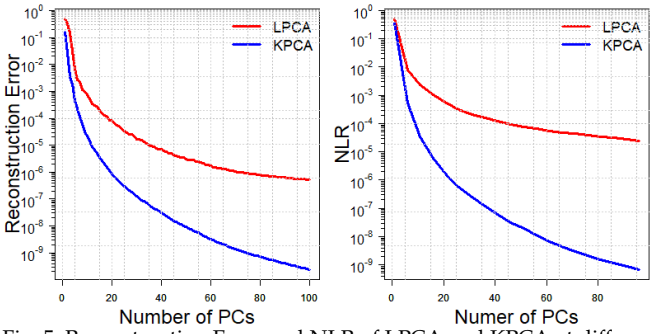

Fig. 5. Reconstruction Error and NLR of LPCA and KPCA at different numbers of selected PCs.

the same RE where KPCA can achieve with a lesser number of PCs. It confirms that the nature of associations among the variables in data directly affect the compression.

To further investigate the existence of non-linearity in X and its impact on compression, the degree of NLR was computed with LPCA and KPCA at different number of PCs. Fig. 5 (right) shows the variation of NLR which represents the linear fitting residual error. LPCA incurs a higher degree of NLR than KPCA, which means that non-linear fitting produces lower residual errors than linear fitting, which confirms the outcomes of REs. For instance, NLR with 20 PCs is 2.99×10^{-4} of LPCA, which is twice higher that in KPCA (5.59×10^{-8}). It confirms that there is a non-linearity between feature variables in X. KPCA captures non-linearity better than LPCA. Further, the degree of

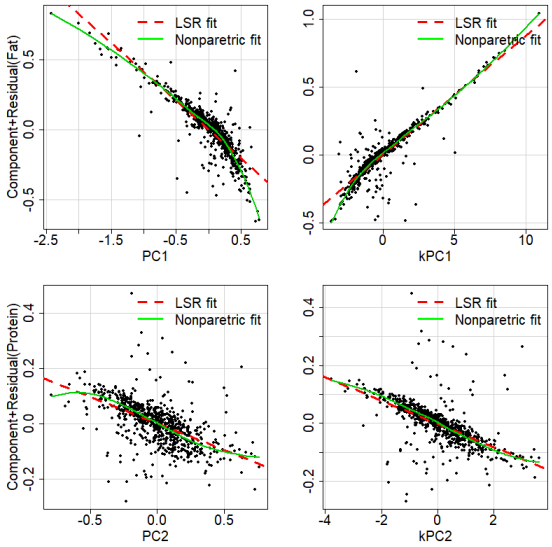

Fig. 6. PRPs for PC1 (Fat) and PC2 (Protein) to detect non-linearity in MIRS data with the first 10 PCs of LPCA (left) and KPCA (right).

NLR decreases with increasing number of PCs. This shows that extraction of higher dimensional feature space from the original data has lower degree of non-linearity.

4.4. Non-linearity between the Features and Target Variables

DW test was conducted to investigate the non-linearity between feature variables in X and the response variables based on LSR modeling. The scores of the first 10 PCs derived from LPCA and KPCA were used to compute LSR residuals to evaluate DW test statistics for each MQP. The test statistics and corresponding critical values are given in the Table 1. The results reveal that Protein and Fat predictions have non-linear behaviors and Lactose has a linear relationship in the MIRS dataset X.

To make a visual interpretation of non-linearity in Fat and Protein predictions, which was evidenced by DW test, their PRPs were drawn using the scores of the first 10 PCs. To get an idea about non-linearity of Fat and Protein, only the PRP of PC1 and PC2 are shown in Fig. 6 (LPCA-left and KPCA-right). The divergence of the non-parametric fit from the fitted LSR line indicates a degree of non-linearity and the type of non-linear relationship. The PRPs from the compressed-domain of LPCA represent higher divergence from the LSR fit than those using compressed-domain of KPCA. This confirms that Protein and Fat predictions have non-linear relationships to X.

Table 1: DW TEST for the SELECTED MQPs WITH FIRST 10PCs DERIVED FROM LPCA AND KPCA ($d_L = 1.8498$ and $d_U = 1.9019$).

MQP	DW test statistic (d)		Decision (linear/non-linear)
	LCPA	KPCA	
Lactose	1.6559	1.6391	$d < d_L$ linear
Fat	1.9680	1.9505	$d > d_U$ non-linear
Protein	1.9805	1.9779	$d > d_U$ non-linear

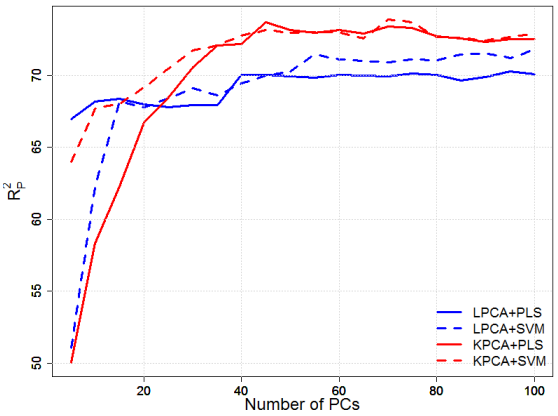

Fig. 7. Prediction accuracy of PLS and LS-SVM based on LPCA and KPCA compressed data for Protein.

4.5. Learning Accuracy of PLS and LS-SVM

To study the learning performance of the regression models based on LPCA and KPCA compressed data, the learning accuracy was computed from PLS and LS-SVM. First, the dataset was divided into two subsets as calibration (80% of the samples) and validation (the remaining samples) by using *Kernard-stone* sampling method. Calibrations and validations were performed for the first 100 PCs and computed the validation R_P^2 (coefficient of determination) as the learning accuracy. The number of latent variables (LVs) in PLS was selected employing *10-fold* cross-validation. We used the *Bayesian* optimization approach to select LS-SVM model parameters; γ and σ.

Fig. 7 shows the learning accuracy computed from PLS and LS-SVM models for Protein with different PCs derived from LPCA and KPCA. Almost the same maximum learning accuracy of nearly 74% was achieved with KPCA compressed data using first 45 and 70 PCs respectively from PLS and LS-SVM. The learning accuracy was higher with LPCA compressed data than in KPCA only for lower number of PCs (around $\leq$ 20 PCs). The number of PCs where the maximum learning accuracy was achieved is higher with KPCA than LPCA. In general, comparing all the values, it turns out that non-linear compression and learning has improved the leaning accuracy although the feature space is higher compared to the linear approach.

The same procedure was repeated for Fat and Lactose and the results are shown in Table 2. The observed highest learning accuracies from the original and LPCA/KPCA compressed data with the corresponding PLS and LS-SVM model parameters (including Protein). The number of PCs of with those accuracies were observed are also given. All the learning outcomes show that CL gives higher leaning accuracy than that was obtained from the original data. Further, the performance of CL from KPCA compression is better than that of LPCA except Lactose. The learning performances of LS-SVM models are always higher than PLS models regardless of the compression technique. The results show that the learning accuracy from LPCA compressed data is higher than KPCA. This turns out that there is a linear relationship between MIRS data for Lactose prediction.

Table 2: COMPRESSED-DOMAIN LEARNING ACCURACY (R^2) of PLS AND LS-SVM PREDICTIONS for LACTOSE,FAT AND PROTEIN

MQP	Original Data		LPCA				KPCA			
	PLS	LS-SVM	PLS		LS-SVM		PLS		LS-SVM	
	R^2_p(LVs)	$R^2_p(\sigma,\gamma)$	$\#PCs$	R^2_p	$\#PCs$	R^2_p (σ,γ)	$\#PCs$	R^2_p	$\#PCs$	R^2_p (σ,γ)
Lactose	83.51 (12)	87.59 (26.57, 58.48)	55	83.49	45	88.03 (12.25, 66.91)	95	78.13	95	79.32 (2.28,12.18)
Fat	88.04 (5)	89.35 (35.66, 69.56)	65	88.20	45	88.82 (17.56, 50.5)	55	88.17	75	88.89 (14.2, 63.5)
Protein	70.3 (15)	72.2 $(5.01, 7.4 \times 10^4)$	40	70.41	55	71.63 $(42.35, 1.58 \times 10^3)$	45	73.67	70	73.9 (39.66,64.47)

LS-SVM model parameters confirm a non-linearity for Protein and Fat predictions and a linearity for Lactose within the MIRS dataset. This is because the highest γ values were observed for Protein with both LPCA and KPCA. γ for Fat was higher than Lactose. Further, these values verify the results given in Fig. 7 and Table 1. The behaviour of σ values was same as γ, which means Protein predictions have more global behaviour than Fat and Lactose.

5. Conclusions

First we investigated non-linear behaviours between the wavenumbers (features) of the MIRS data. Our investigation has shown that there is a considerable non-linearity exists and should be captured by the compression algorithms. Then we have compressed the original data using both LPCA/KPCA and investigated non-linearity between the compressed-domain feature variables and with three selected response variables (Fat, Protein and Lactose). According to this analysis, we conclude that Fat and Protein predictions show non-linear behaviours, which we need to capture in compressed learning. Finally we applied PLS and LS-SVM regression models on the two compressed-domain data to show that there is an improvement in accuracies using non-linear predictions. Therefore, we conclude that use of a linear unsupervised compression technique has negative impacts on the prediction accuracy of different MQPs. Use of non-linear compression techniques such as KPCA at the compression entity is highly desirable in compressed learning approach. Otherwise, the advantages of using complex non-linear predictive models will not be useful in MIRS based milk quality analytics.

Acknowledgment

This research was supported by the Science Foundation Ireland (SFI) through the project "PrecisionDairy" (ID: 13/1A/1977).

References

[1] S. Wolfert, L. Ge, C. Verdouw, M. J. Bogaard, Big Data in Smart Farming - A Review, Elsevier J. Agricultural Systems, vol. 153, May 2017.
[2] B. H. Park, H. Kargupta, Distributed Data Mining: Algorithms, Systems and Applications, Data Mining Handbook (Editor: Nong Ye), 2002.
[3] F. Jalali, et al., Fog Computing may help to save Energy in Cloud Computing, IEEE Journal on Selected Areas in Communications, May 2016.
[4] H. Zheng, S. R. Kulkarni, H. W. Poor, Dimensionally Distributed Learning Models and Algorithms, IEEE International Conference on Information Fusion, July 2008.
[5] K. Bhargava, S. Ivanov, C. Kulatunga and W. Donnelly, "Fog-enabled WSN system for animal behavior analysis in precision dairy", IEEE International Conference on Computing, Networking and Communications (ICNC), Jan. 2017.
[6] M. Chen, et al., On the Computation Offloading at Ad Hoc Cloudlet: Architecture and Service Modes, IEEE Communications Magazine, vol. 53 (6), June 2015.
[7] C. Kulatunga, L. Shalloo, W. Donnelly, E. Robson, S. Ivanov, Opportunistic Wireless Networking for Smart Dairy Farming, IEEE IT Professional Magazine, vol. 19 (2), March 2017.
[8] G. Visentin, et al., Prediction of Bovine Milk Technological Traits from Mid-Infrared Spectroscopy Analysis in Dairy Cows, J. Dairy Science, vol. 98, September 2015.
[9] P. H. Garthwaite, An Interpretation of Partial Least Squares, J. American Statistical Association, vol. 89, March 1994.
[10] L. J. P. ven der Maaten, In Introduction to Dimensionality Reduction using MatLab, MICC report, July 2007.
[11] W. Huang, H. Yin, Linear and Nonlinear Dimensionality Reduction for Face Recognition, IEEE International Conference on Image Processing (ICIP), 2009.
[12] Mu Li, et al., Scaling Distributed Learning with the Parameter Server, USENIX Operating Systems Design and Implementation (OSDI), 2014.
[13] Y. M. Chen, et al., Combination of the Manifold Dimensionality Reduction Methods with Least Square Support Vector Machines for Classifying Species of Sorghum, Scientific Reports, vol. 6, Jan 2016.
[14] R. M. Balabin, S. V. Smirnov, Melamine Detection by MIRS and NIRS: A Quick and Sensitive Method for Dairy Products Analysis Including Liquid Milk, Infant Formula, and Milk Powder, J. Talanta, 2011.
[15] L.J. Cao, et al., A comparison of PCA, KPCA, and ICA for Dimensionality Reduction in Support Vector Machine, J. Neurocomputing, 2003.
[16] H. V. Nguyen et. al., Multivariate Maximal Correlation Analysis, International Conference on Machine Learning, 2014.
[17] Yan-de Liu, et al., On-Line Predicting Soluble Solids Contents of Intact Pears Combination with Wavelet Transform and Support Vector Regression, IEEE International Conference on Natural Computation, 2010
[18] T. V. Gestel, et al., Benchmarking Least Squares Support Vector Machine Classiers, J. of Machine Learning, vol. 54, 2004.
[19] J. V. Anand, S. Titus, Regression based Analysis of Effective Hydrocast in Underwater Environment, IEEE Region 10 Conference on TENCON, 2014.
[20] G. A. Licciardi, F. D. Frate, Pixel Unmixing in Hyperspectral Data by Means of Neural Networks, IEEE Tran. on Geoscience and Remote Sensing, vol. 49, 2011.
[21] R. Calrebank, S. Jafarpor, R. Schapier, Compressed Learning: Universal Sparse Dimensionality Reduction and Learning in the Measurement Domain, 2009.
[22] K. Pelckmans, et al.,LS-SVMlab: A MatLab/C Toolbox for Least Squares Support Vector Machines, (http://www.esat.kuleuven.be/sista/lssvmlab)
[23] J. W. McKean, Exploring Data Sets using Partial Residual Plots based on Robust Fits, IMS Lecture Notes-Monograph Series, vol. 31, 1997.
[24] Y. Bengio, A. Courville, P. Vincent, Representative Learning: A Review and New Perspectives, IEEE Tran. on Pattern Analysis and Machine Intelligence, vol. 35, Aug. 2013.