Chemometrics and Intelligent Laboratory Systems 182 (2018) 166–171

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Accounting for spatial dependency in multivariate spectroscopic data

M. Prakash [a,b,*], J.K. Sarin [a,b], L. Rieppo [a,c], I.O. Afara [a], J. Töyräs [a,b]

[a] Department of Applied Physics, University of Eastern Finland, Kuopio, Finland
[b] Diagnostic Imaging Center, Kuopio University Hospital, Kuopio, Finland
[c] Research Unit of Medical Imaging, Physics and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland

ARTICLE INFO

Keywords:
Linear mixed effects
Articular cartilage
Near infrared (NIR) spectroscopy
Spectroscopic mapping
Principal components
LASSO

ABSTRACT

We examine a hybrid multivariate regression technique to account for the spatial dependency in spectroscopic data due to adjacent measurement locations in the same joint by combining dimension reduction methods and linear mixed effects (LME) modeling. Spatial correlation is a common limitation (assumption of independence) encountered in diagnostic applications involving adjacent measurement locations, such as mapping of tissue properties, and can impede tissue evaluations. Near-infrared spectra were collected from equine joints ($n = 5$) and corresponding biomechanical ($n = 202$), compositional ($n = 530$), and structural ($n = 530$) properties of cartilage tissue were measured. Subsequently, hybrid regression models for estimating tissue properties from the spectral data were developed in combination with principal component analysis (PCA-LME) scores and least absolute shrinkage and selection operator (LASSO-LME). Performance comparison of PCA-LME and principal component regression, and LASSO-LME and LASSO regression was conducted to evaluate the effects of spatial dependency. A systematic improvement in calibration models' correlation coefficients and a decrease in cross validation errors were observed when accounting for spatial dependency. Our results indicate that accounting for spatial dependency using a LME-based approach leads to more accurate prediction models.

1. Introduction

Articular cartilage, a connective tissue covering the ends of bones in a joint, is susceptible to post-traumatic osteoarthritis (PTOA) due to focal injuries caused by sudden excessive impact loading. The injury, although initially localized, often spreads over time, resulting in altered functional performance of the whole joint. Arthroscopic evaluation of tissue properties around the injury site and assessing the spread of the injury could enable optimal surgical intervention, thereby minimizing the risk of PTOA. Currently, in clinical arthroscopies [1], cartilage is assessed visually through an endoscope and by palpating the tissue surface with a metal hook [2]. This method is qualitative, unreliable, and poorly reproducible [3,4], thus necessitating development of novel, quantitative, robust, and reliable methods.

Non-destructive diagnostic tools, such as near-infrared (NIR) spectroscopy, have shown potential for arthroscopic characterization of articular cartilage integrity [5]. NIR spectroscopy is a vibrational spectroscopic technique that has been utilized for spatial assessment of cartilage biomechanical, compositional, and structural properties [6–8]. In these studies, multivariate regression was utilized to relate cartilage NIR spectra with its tissue properties. However, conventional multivariate regression methods, such as partial least squares (PLS), are based on the underlying assumption of independent observations [9], whereas biomedical characterization of tissue integrity, for example in arthroscopy, often involves multiple measurement locations within close proximity in the same joint. This grouping effect of samples introduces spatial dependency and is likely to cause unreliable correlations if unaccounted for in regression modeling [10,11].

Linear mixed effects (LME) regression and its input parameters, namely fixed effects and random effects, can be designed for specific datasets to account for grouping effects. Since only a limited number of regressors (input variables) can be utilized in model creation using LME, adaptation for a large set of variables, such as NIR spectra, requires dimension reduction and/or variable selection methods. Hence, the input variables need to be methodically selected by retaining only the most important ones.

Application of dimension reduction methods, such as principal component analysis (PCA) [12] via PCA score, and variable selection and regularization methods like LASSO (least absolute shrinkage and selection operator) or L1 penalization [13], are effective approaches for

* Corresponding author. Department of Applied Physics, University of Eastern Finland, Kuopio, Finland.
E-mail addresses: mithilesh.prakash@uef.fi (M. Prakash), jaakko.sarin@uef.fi (J.K. Sarin), lassi.rieppo@oulu.fi (L. Rieppo), isaac.afara@uef.fi (I.O. Afara), juha.toyras@uef.fi (J. Töyräs).

https://doi.org/10.1016/j.chemolab.2018.09.010
Received 14 May 2018; Received in revised form 19 September 2018; Accepted 26 September 2018
Available online 27 September 2018

reducing the high dimensionality of the data, such as NIR spectra. PCA finds a set of projections that maximizes the variance in the original dataset; hence, the data structure in the sample space is captured even in the low dimensional subspaces. LASSO [14] is a regularization method ideal for creating sparse models with high statistical accuracy in predictions.

In this study, we propose a novel hybrid technique, which combines dimension reduction methods and LME regression, to account for spatial dependency in analysis of multivariate dataset. This is based on the hypothesis that hybrid regression techniques can effectively model the relationship between cartilage NIR spectra and its properties while accounting for dependencies within the data.

2. Materials and methods

To account for spatial dependency in the dataset, the contributing levels of dependency must first be identified. The levels of dependency are defined by the experimental design and the scope of the application. In our application on NIR-based characterization of cartilage, joint level (measurement locations grouped in one complete joint) and bone level (measurement locations grouped on one bone of a particular joint) were identified (Fig. 1) as the two significant levels of dependency [15]. (Other application specific dependency levels can be accommodated in the design matrix).

Subsequently, models were developed for relating the predictors ($\boldsymbol{X}$) to the response variables ($\boldsymbol{y}$) while accounting for the identified dependency levels (grouping effects). The adaptation of LME can be written in the equation form:

$$\boldsymbol{y_i} = \boldsymbol{X\beta} + \boldsymbol{Zu}_1 + \boldsymbol{Mu}_2 + \boldsymbol{\varepsilon}, \quad (1)$$

where $\boldsymbol{y_i}$ is an N(number of observations)-by-1 response vector of reference values for the $\boldsymbol{i}$th tissue property, $\boldsymbol{X}$ is an N-by-P (dimension reduced NIR spectra) matrix of fixed effect regressors, $\boldsymbol{\beta}$ is a P-by-1 vector of fixed effects coefficients, $\boldsymbol{Z}$ is an N-by-Q (grouping count) random effects design matrix, $\boldsymbol{M}$ is an N-by-1 vector of additional random effects vector, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are the mixed effects coefficients of sizes Q-by-1 and 1-by-1 respectively, and $\boldsymbol{\varepsilon}$ is an N-by-1 vector representing the observation error. Restricted maximum likelihood method was employed for estimating LME [16].

2.1. Equine cartilage dataset

In this study, we utilized NIR spectral and tissue reference data from equine cartilage measured in earlier studies [17,18]. In summary, metacarpophalangeal joints ($n = 5$) were acquired from a slaughterhouse, and specific areas of interest (AI, $n = 44$) with cartilage lesions of varying severity were selected by a veterinary surgeon. Subsequently, a 15 × 15 mm grid consisting of 25 measurement locations was marked on each AI with a felt-tip pen (Fig. 1). The measurement locations (d=2mm) were equally spaced (*interdistance* = 2.5 mm), and locations with highly eroded cartilage were excluded, yielding a total of 869 measurement points. NIR spectral measurements and thickness values were acquired on each of the 869 measurements; however, biomechanical measurements were performed only on 202 locations and compositional analysis on 530 locations due to limitations set by sample preservation and geometry, respectively. NIR spectra were matched with corresponding tissue property based on location during regression analysis.

2.2. NIR spectral measurements

The NIR spectroscopy instrumentation consisted of a halogen light source (wavelength range: 360–2500 nm, power 5 W, optical power: 239 μW in a $d_{fiber} = 600$ μm, Avantes BW, Apeldoorn, Netherlands), and a spectrometer (wavelength range: 200–1160 nm, Avantes BW, Apeldoorn, Netherlands). A customized fiber optic probe ($d = 5$ mm) consisting of seven fibers ($d_{fiber} = 600$ μm) within the central window ($d = 2$ mm), the six outer fibers for transmitting, and the central one for collecting the reflectance spectrum, was utilized. Prior to sample measurements, dark and reference spectra were acquired. Dark spectrum was acquired with the spectrometer light source switched off in order to collect background noise. With the light source switched on, reference spectrum was acquired from a reflectance standard (Spectralon, SRS-99, Labsphere Inc., North Sutton, USA). The absorbance values of each sample spectra were scaled as per Beer-Lambert's law using the dark and reference spectra. In addition, signal acquisition time was optimized to maximize the signal to noise ratio. The average of three spectral measurements that each consisted of eight co-added spectral scans ($t_{eight\ scans} = 720$ ms) was calculated. To preprocess the spectra (700–1050 nm), Savitzky-Golay estimates of the second derivative using 41 points (or 25 nm) and a third-order polynomial for the smoothing were computed. This preprocessing not only removes baseline offset and dominant linear terms but also enhances subtle absorption peaks.

2.3. Cartilage thickness and biomechanical measurements

Cartilage thickness at all NIRS measurement locations was determined using optical coherence tomography (OCT) via the Ilumien PCI Optimization System, (St. Jude Medical, St. Paul, MN, USA) at an operating wavelength of 1305 ± 55 nm, axial resolution <20 μm, and lateral resolution 25–60 μm. The samples were fully immersed in phosphate-buffered saline (PBS) during the measurements.

Biomechanical indentation measurements were performed at 202 locations using a customized material testing device consisting of a load cell (Sensotec, Columbus, OH, USA) with force resolution of 5 mN, an actuator (PM1A1798-1 A, Newport, Irvine, CA, USA) with displacement resolution of 0.1 μm (PM500-1 A, Newport, Irvine, CA, USA), and a plane-ended cylindrical indenter ($d = 0.53$ mm). Equilibrium modulus (E_{eq}) and dynamic modulus (E_{dyn}) were calculated using an indentation protocol detailed in Korhonen et al. [19] and Sarin et al. [17].

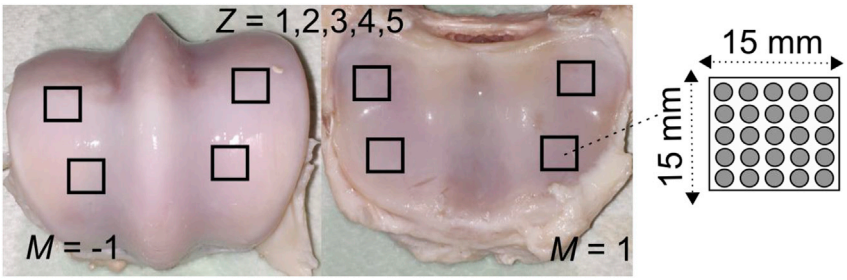

Fig. 1. Areas of interest (AI) marked (black squares, not to scale) on the articulating surfaces of equine metacarpophalangeal joint. In this study, grouping information is on two dependency levels, i.e. joint level and bone level, which is held in $\boldsymbol{Z}$ (sample count × 5) and $\boldsymbol{M}$ (sample count × 1) design matrices.