# Title page

Comparison and optimization of different multivariate regression techniques based on Near infrared spectroscopy data in equine articular cartilage.

Mithilesh Prakash1, Jaakko K. Sarin1,2, Lassi Rieppo1,3, Isaac O. Afara1,2,4, Juha Töyräs1,2

1 Department of Applied Physics, University of Eastern Finland, Kuopio, Finland

2 Diagnostic Imaging Center, Kuopio University Hospital, Kuopio, Finland

3 Research Unit of Medical Imaging, Physics and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland

4 Department of Electrical and Computer Engineering, Elizade University, Ondo, Nigeria

**An abbreviated title:** Comparison and optimization of NIR data in Equine AC

**Corresponding author:**

Mithilesh Prakash, M.Sc. (Tech.)

Department of Applied Physics

University of Eastern Finland

Kuopio, Finland

Tel: +358 414963056

Fax: +35817162131

Email: [mithilesh.prakash@uef.fi](mailto:mithilesh.prakash@uef.fi)

# Abstract and key terms

The non-destructive near infrared (NIR) spectroscopy has been successfully applied to assess the properties of biological tissues, such as articular cartilage. In general, the spectroscopic data includes absorbance values from broad wavelength region, thus resulting in a large number of contributing factors and complexity amongst variables. Furthermore, this broad spectrum includes information from noisy variables potentially contributing to error during spectroscopic analysis. In this study, several regression techniques such as principal component regression (PCR), partial least squares (PLS), least absolute shrinkage and selection operator (LASSO), ridge regression and least squares support vector machines (LS-SVM) are studied to find the most suitable technique for modeling cartilage properties using pre-existing NIR equine data. Additionally, the most suitable variable selection methods including Monte Carlo uninformative variable elimination (MC-UVE), competitive adaptive reweighted samplings (CARS), variable combination population analysis (VCPA), backwards interval on PLS (biPLS), genetic algorithm (GA) and jack-knife are explored. The comparative analysis indicates PLS technique to be an ideal choice for cartilage NIR research. Furthermore, variable selection methods, especially MC-UVE with PLS enhanced the results. Thus, variable selection based PLS can effectively model cartilage NIR data with functional tissue parameters.

**Key terms:** spectroscopy; variable selection; near infrared; partial least squares (PLS) regression

# Introduction

Articular cartilage (AC) is a specialized type of hyaline cartilage found at the distal ends of bones providing smooth and load bearing interfaces in joints. Degenerative joint conditions such as Osteoarthritis(OA) are generally characterized by disruption of superficial collagen, loss of proteoglycan (PG) and changes in tissue functional properties9,29,14. Non-destructive light-based spectroscopic imaging modalities, such as Optical coherence tomography (OCT)25, Fourier transform infrared (FTIR)28, Near infrared (NIR)31 and Raman7, have been quite effective in cartilage research. Recent cartilage research studies ( Afara et al 2013 6 & 2015 2 , Palukuru et al 201424 & 201620 and Sarin et al 201630) have advocated the use of NIR as a better method due to superior tissue depth penetration. Application of NIR spectroscopy in articular cartilage research has enabled the assessment of the tissue compositional and biomechanical properties in OA 32,5,23 .

Multivariate15 regression techniques are typically used to extract information from the NIR spectral data since univariate approaches are often inefficient due to multi-collinearity in NIR data as suggested by Afara et.al3. Regression methods enable building mathematical models relating NIR spectra data with the functional properties of the sample specimen (e.g. cartilage). The popular multivariate regression techniques utilized in NIR spectroscopy are principal component regression (PCR)22 and partial least squares regression (PLSR)35. PLSR is most commonly used in NIR spectroscopic studies of articular cartilage. However, the regression shrinkage methods, such as ridge regression34 and least absolute shrinkage and selection operator (LASSO), and least square version of support vector machines based regression called LS-SVM10, have not been used in NIR studies of articular cartilage (AC).

Selecting optimal variables for regression models is an essential step as the spectra may contain noisy or irrelevant variables that hinder the analysis. Usually variable selection has been done by restricting the spectral wavelength due to experimental or known restrictions also called manual selection of wavelength, which may result in inconsistent results and is prone to human error. Statistical studies conducted by Xiaobo et al37, Westad et.al36 and Mehmood et.al21 have shown the significance of variable selection methods in multivariate regression techniques. Application of variable selection21 method in regression is based on the principle of either choosing the most contributing variables or eliminating noncontributing variables37. MC-UVE, CARS, VCPA, interval selection methods, GA and jack-knife encompasses different variable selection methods available for NIR research application.

In this study, multiple multivariate regression and variable selection methods are utilized in an attempt to find the most suitable algorithms for cartilage research. We hypothesized the PLS to be optimal regression method and variable selection methods to significantly increase the reliability of the results. To test the hypothesis, calibration models between spectral and reference data were built and evaluated with independent test group

# Materials and methods

1. *Material*

The study was conducted on the pre-existing NIR data30 from equine AC comprised of NIR spectral data and functional properties. Briefly, equine metacarpophalangeal joints (N=5) were gathered and areas of interest (AI, N=44) of intact and damaged cartilage were selected by equine surgeons. The AIs were uniformly divided into grids, thus resulting in a total of 869 measurement points. The spectral data matrix (869 x 2000) included the absorbance information from the wavelength range of 700-1050 nm.

The functional properties including equilibrium, dynamic and instantaneous moduli were obtained via indentation16 testing. Additionally, cartilage thickness was determined with optical coherence tomography12,13. The details of the data acquisition are given in more detail in the paper *Sarin et al* 2016.

1. *Methods*

The NIR spectra data was preprocessed by smoothing and filtering using third degree Savitzky-Golay filter with a 25 nm window to remove the background noise. Furthermore, second derivative was applied on the smoothed and filtered data to remove the offset and linear elements of the baseline and to increase the separation between the absorption peaks.

The dataset was divided into two sets similarly as in the original study by Sarin. et.al 2016. The first set, called the training set, consisting of 799 samples was used for calibration model training, and the second set, called the test set, consisting of 70 samples was used to evaluate the model performance27. Regression models were fine-tuned using the model performance on the test set. After the technique specific optimization, the model parameters were finally optimized for the highest R2 and the lowest root mean square error of prediction (RMSEP) for the test set. The software analysis was done using MATLAB R2014a (8.3.0.532) on a 64-bit computing environment.

*Regression methods*

PCR, PLS, ridge, LASSO and LS-SVM methods are used for multivariate regression comparative analysis. Built-in MATLAB functions *princomp, plsregress, ridge, lasso* and *LS-SVM toolbox* from LS-SVMlab33 v1.8 were used to regress the NIR data with functional parameters respectively.

PCR (*princomp*) was optimized by first building series of models with iteratively increasing number of principal components (PC). The number of PC’s are varied by increasing the number principal component scores and principal component loadings simultaneously until maximum number of components is reached (max =15). Next, the best PCR model with the highest R2 test and the lowest RMSEP was retained and number of components recorded. Similarly, PLSR (*plsregress*) was first optimized by building series of models as function of number of components (n=15) and each model cross validated individually by k-fold operation (k=10) and the best PLSR model with the highest R2 test set and the lowest RMSEP retained. For Ridge regression the vector of the ridge parameters was varied (r = 0 to 1000) in small steps (step size = 0.01) and series of models built as a function of step size. Similar to PCR and PLSR, the best model was retained. In LASSO regression, the initial coefficients are calculated by regularization algorithm (*lasso*) and cross validated by k-fold (k=10). Models series were built as a function of variation in lambda value. Following the previous methods as in PCR, PLR and ridge the best model was retained. For LS-LVM the initialization (*initlssvm*), tuning (*tunelssvm*) and optimization (*trainlssvm*) of the preprocessed is auto handled by the respective functions of the toolbox.

*Variable selection methods*

MC-UVE, CARS, VCPA, biPLS, GA and Jack-knife were used to further enhance the prediction model. The algorithms for MC-UVE and CARS18 were obtained from *Integrated library for PLS and discriminant analysis*19, VCPA40 from *Variable Combination Population Analysis toolbox*, GA algorithm from *PLS-Genetic algorithm toolbox*26 and Jack-knife algorithm11 was coded manually.

MC-UVE (*mcuvepls*) variable selection was optimized by first calculating the reliability index of all the wavelengths and then determining the optimal threshold for reliability index by finding the maximum correlation with the training set. CARS (*carspls*) and VCPA (*vcpa*) did not require additional input and the respective functions auto handled the optimization. In interval selection method the algorithm is optimized by elimination of the non-informative intervals as the Leardi method17, the 3 intervals (interval = 5, 17, 19) with the lowest RMSECV are retained. In GA algorithm the effective number of evaluation (n1= 182) and number of variables (n2 = ~100) was first evaluated using the *gaplsopt* function in the toolbox and the main function *gaplssp* was invoked to perform the GA algorithm. In Jack-knife the student’s t-statistics is used to determine the variable selection by selecting variables with values less than threshold (t = 0.05).

*Statistics and model comparison*

The calibration models developed were analyzed on 5 key parameters, root mean square error of calibration (RMSEC), R2 training set, root mean square error of prediction (RMSEP), R2 testing set and error percentage in test dataset. Additionally, for PLSR, PCR and variable based PLS the number of components are recorded.

# Results

The regression models were built and optimized for all the measured functional properties, namely cartilage thickness, instantaneous modulus, equilibrium modulus and dynamic modulus. PLSR technique was found to the best with the highest R2 test and the lowest RMSEP and error percentage compared to other regression methods (table 1).

**Table 1: Comparison of different regression methods across different tissue parameters. Data is arranged in the descending order of the R2 Test as highlighted in bold. Number of components for PLSR and PCR is indicated by C.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Cartilage Thickness (mm)** | | | |  |
| **Regression method** | **Train Set** | | **Test Set** | | **Error %** | |
| **R2 Train** | **RMSEC** | **R2 Test** | **RMSEP** |
| **PLSR (C = 5)** | 70.28 | 0.13 | **75.57** | 0.11 | 5.94 | |
| **RIDGE** | 73.02 | 0.12 | **74.09** | 0.11 | 6.17 | |
| **LASSO** | 72.90 | 0.12 | **68.63** | 0.12 | 6.90 | |
| **LSSVM** | 77.55 | 0.11 | **67.87** | 0.13 | 6.89 | |
| **PCR (C= 13)** | 60.44 | 0.15 | **67.38** | 0.13 | 7.02 | |
|  | **Instantaneous Modulus (MPa)** | | | |  |
| **PLSR (C = 6)** | 41.82 | 2.63 | **57.16** | 2.31 | 9.04 | |
| **RIDGE** | 41.82 | 2.63 | **54.08** | 2.39 | 9.76 | |
| **LASSO** | 40.82 | 2.65 | **52.67** | 2.42 | 9.79 | |
| **PCR (C = 10)** | 29.60 | 2.89 | **51.50** | 2.45 | 9.76 | |
| **LSSVM** | 98.97 | 0.35 | **43.78** | 2.64 | 11.41 | |
|  | **Equilibrium Modulus (MPa)** | | | |  |
| **PLSR (C = 5)** | 65.95 | 0.87 | 66.84 | 0.97 | 15.73 | |
| **LASSO** | 82.03 | 0.63 | 60.51 | 1.06 | 17.36 | |
| **RIDGE** | 75.88 | 0.73 | 54.54 | 1.17 | 20.39 | |
| **LSSVM** | 60.28 | 0.93 | 49.27 | 1.20 | 20.66 | |
| **PCR (C = 15)** | 24.74 | 1.29 | 35.18 | 1.36 | 21.63 | |
|  | **Dynamic Modulus (MPa)** | | | |  |
| **LASSO** | 69.17 | 3.44 | **66.35** | 3.56 | 13.85 | |
| **PLSR (C = 2)** | 37.27 | 4.91 | **64.88** | 3.64 | 13.06 | |
| **RIDGE** | 63.46 | 3.75 | **61.31** | 3.83 | 15.44 | |
| **PCR (C = 6)** | 27.03 | 5.30 | **60.35** | 3.87 | 14.55 | |
| **LSSVM** | 72.83 | 3.23 | **59.08** | 3.93 | 15.43 | |

However, in case of dynamic modulus property, LASSO was found to have the highest R2 for train set, but it also had a higher error percentage than PLSR.

**Table 2: Variable methods comparison for different tissue parameters. the data is presented in the descending order of R2 test set**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Cartilage Thickness (mm)** | | | | |  |
| **Variable Selection method** | **No. Of PLS Components** | **Model** | | **TEST** | | **Error Percentage** | |
| **R2** | **RMSEC** | **R2** | **RMSEP** |
| **MC-UVE-PLS** | 4 | 70.53 | 0.14 | **76.20** | 0.11 | 5.89 | |
| **JK-PLS** | 1 | 59.65 | 0.15 | **74.05** | 0.11 | 6.14 | |
| **GA-PLS** | 6 | 70.65 | 0.15 | **70.76** | 0.12 | 6.47 | |
| **BIPLS** | 12 | 59.05 | 0.15 | **70.05** | 0.12 | 6.59 | |
| **VCPA** | 3 | 63.15 | 0.15 | **66.13** | 0.13 | 7.02 | |
| **CARS** | 4 | 70.58 | 0.14 | **62.02** | 0.14 | 7.65 | |
|  | **Instantaneous Modulus (MPa)** | | | | |  |
| **CARS** | 5 | 43.39 | 2.79 | **56.72** | 2.32 | 8.99 | |
| **MC-UVE-PLS** | 4 | 39.19 | 2.86 | **56.66** | 2.32 | 9.46 | |
| **JK-PLS** | 7 | 32.40 | 2.98 | **53.21** | 2.41 | 9.41 | |
| **GA-PLS** | 2 | 30.97 | 2.88 | **51.56** | 2.45 | 9.86 | |
| **VCPA** | 3 | 33.41 | 2.84 | **51.19** | 2.46 | 9.48 | |
| **BIPLS** | 3 | 20.16 | 3.08 | **43.55** | 2.65 | 9.98 | |
|  | **Equilibrium Modulus (MPa)** | | | | |  |
| **MC-UVE-PLS** | 3 | 59.26 | 1.20 | **62.98** | 1.03 | 17.25 | |
| **GA-PLS** | 10 | 61.41 | 1.31 | **60.47** | 1.06 | 17.60 | |
| **VCPA** | 5 | 48.47 | 1.13 | **58.35** | 1.09 | 17.58 | |
| **JK-PLS** | 8 | 41.43 | 1.34 | **52.87** | 1.16 | 19.14 | |
| **CARS** | 2 | 44.49 | 1.26 | **45.37** | 1.25 | 19.82 | |
| **BIPLS** | 5 | 36.84 | 1.18 | **41.65** | 1.29 | 20.34 | |
|  | **Dynamic Modulus (MPa)** | | | | |  |
| **MC-UVE-PLS** | 3 | 63.53 | 4.80 | **72.31** | 3.23 | 12.93 | |
| **GA-PLS** | 3 | 50.88 | 4.83 | **69.75** | 3.38 | 12.29 | |
| **VCPA** | 2 | 43.06 | 4.86 | **67.61** | 3.50 | 13.76 | |
| **JK-PLS** | 5 | 43.64 | 5.27 | **67.47** | 3.50 | 13.49 | |
| **CARS** | 3 | 63.13 | 4.45 | **67.05** | 3.53 | 13.20 | |
| **BIPLS** | 7 | 38.73 | 4.85 | **55.98** | 4.07 | 14.87 | |

PLS models optimized by using variable selection methods showed much better model performance (table 2). MC-UVE-PLS was found to improve the PLS performance parameters the most.

# Discussion

In this study, for the first time a comparative analysis (table 1) of multivariate regression methods in NIR spectroscopy on cartilage data is presented. First, comparison of different optimized multivariate regression techniques, namely PLS, PCR, LASSO, ridge and LS-SVM, using key statistical parameters was conducted. The model performances were measured on all important tissue properties4. PLS regression was found to be the best regression method based on the results of the first study.

Second, the effect of variable selection methods on the performance of PLS regression models was evaluated (table 2). The wavelength region of input NIR spectra was limited to 700-1050 nm range to follow the procedure of the original study by Sarin et.al 2016. As a result, the variable selection methods had a relatively narrow spectrum of variables. Nonetheless, the results show the applicability of variable selection methods in regression analysis, which is consistent with the findings from Abrahamsson .et.al1 research on pharmaceutical application.

Recent cartilage research studies have favored projection regression methods such has PLSR and PCR due to ease of implementation. However, variable selection methods have not been applied earlier. PLSR modelled (table 1) the optical response of the tissue for articular cartilage thickness better (10% to 30% higher) compared to other functional tissue properties. This is attributed to direct relationship tissue thickness6 due to path length effect to the distance the light travels6. The regression comparison highlighted some limitations of sophisticated regression techniques in modelling cartilage data, as LS-LVM and PCR seemed to suffer from overfitting and underfitting, respectively. The results of the second study show that the variable selection improves the model performance and enhances the results of PLSR, and hence MC-UVE-PLS is quite favorable in cartilage application.

The prediction models built with variable selection methods has smaller number of components when compared to just PLSR in case of cartilage thickness (difference = 1), instantaneous modulus (difference = 1 in CARS and difference = 2 in MC-UVE-PLS), Equilibrium modulus (difference = 2) and same number of components in dynamic modulus (n =3). Hence, variable selection methods promote simpler models less prone to overfitting. Also MC-UVE-PLS models had decreased RMSEP (1% - 11%) compared to PLS models without variable selection, and was seen to produce consistently good results.

The results of the variable selection methods when benchmarked against the results of PLS and manual selection of wavelength conducted on same equine data conducted by Sarin et al. 2016, promotes the usage of variable selection methods for cartilage thickness and dynamic modulus property of the cartilage tissue. MC-UVE-PLS prediction models displayed about 7-8 % improvement in the R2 test set and showed 15% lower RMSEP in case of thickness measurement and dynamic modulus, respectively, compared to standard PLSR models. On the other hand, in cases of instantaneous and equilibrium moduli, the variable selection methods showed no improvement in R2 or RMSEP over the standard PLS models. This may be due to limited spectral range used in the study. However, it should be noted that the models were simpler in terms of PLS components than the standard PLS models, which reduces the possibility of overfitting.

The results of the study in comparison to original research indicate that manual variable selection of wavelengths does not always result in optimal prediction model, as it requires more understanding of the underlying measurement process and a priori knowledge of significance of spectral regions. In contrast, in case of variable selection methods, the models are more robust, and they do not require thorough understanding of the measurement procedure and constraints due to experimental setup. Furthermore, they seem to improve the model performance in terms of R2 and RMSEP values and/or reduce the required number of PLS components, which results in more stable regression models. MC-UVE-PLS gave consistently good results, and may be considered as the best variable selection method for this data set .

The GA-PLS in NIR applications has been studied in horticultural studies38, food engineering39 and fuel analytical8 studies as the optimal variable selection methods. However, in the current cartilage study, MC-UVE-PLS surpassed GA-PLS and VCPA, which is in contrast with the results of Yun et.al study with VCPA40. Limitations imposed on the wavelength bandwidth might have reduced the performance of the VCPA, which in theory is the better variable selection method. Likewise, the narrow wavelength bandwidth probably limits the performance of all other variable selection methods as there are less relevant variables to work with compared to the use of full NIR range. The study recommends the usage of PLS technique as the best regression tool in cartilage study and results can be further optimized by MC-UVE-PLS.

# Acknowledgments

This study was funded by the Academy of Finland (project 267551, University of Eastern Finland) and Kuopio University Hospital (VTR projects 5041750 and 5041744, PY210 Clinical Neurophysiology).

# Contributions

# Prakash, M: implementation of the analysis, interpretation and was the main writer of the manuscript.

**Sarin, J.K.:** was involved in equine *ex vivo* study, data acquisition, manuscript supervision.

**Rieppo, L:** Conception of the exploratory idea, data interpretation and manuscript supervision.

**Afara I.O.:** supervision of statistical analyses and manuscript development.

**Töyräs J.:** contributed in the study conception and interpretation of data

All authors contributed in the preparation and approval of the final submitted manuscript.

# Conflict of Interest

The authors have no conflicts of interest in the execution of this study and preparation of the manuscript.

# References

1. Abrahamsson, C., J. Johansson, A. Sparén, and F. Lindgren. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tabletsdoi:10.1016/S0169-7439(03)00064-9

2. Afara, I. O. *et al.* Optical absorption spectra of human articular cartilage correlate with biomechanical properties, histological score and biochemical composition. *Physiol. Meas.* 36:1913–1928, 2015.

3. Afara, I. O., M. Hauta-Kasari, J. S. Jurvelin, A. Oloyede, and J. Töyräs. Optical absorption spectra of human articular cartilage correlate with biomechanical properties, histological score and biochemical composition. *Physiol. Meas.* 36:1913–1928, 2015.

4. Afara, I., I. Prasadam, R. Crawford, Y. Xiao, and a. Oloyede. Non-destructive evaluation of articular cartilage defects using near-infrared (NIR) spectroscopy in osteoarthritic rat models and its direct relation to Mankin score. *Osteoarthr. Cartil.* 20:1367–1373, 2012.

5. Afara, I., S. Singh, and a. Oloyede. Application of near infrared (NIR) spectroscopy for determining the thickness of articular cartilage. *Med. Eng. Phys.* 35:88–95, 2013.

6. Afara, I., S. Singh, and A. Oloyede. Application of near infrared (NIR) spectroscopy for determining the thickness of articular cartilage. *Med. Eng. Phys.* 35:88–95, 2013.

7. Bonifacio, A. *et al.* Chemical imaging of articular cartilage sections with Raman mapping, employing uni- and multi-variate methods for data analysis. *Analyst* 135:3193, 2010.

8. Breitkreitz, M. C. *et al.* Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration. *Analyst* 128:1204–1207, 2003.

9. Buckwalter, J. A., and H. J. Mankin. Articular cartilage: degeneration and osteoarthritis, repair, regeneration, and transplantation. *Instr. Course Lect.* 47:487–504, 1998.

10. Chauchard, F., R. Cogdill, S. Roussel, J. M. Roger, and V. Bellon-Maurel. Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes. *Chemom. Intell. Lab. Syst.* 71:141–150, 2004.

11. Efron, B. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, 1982.doi:10.1137/1.9781611970319

12. Fard, A. M., P. Vacas-Jacques, E. Hamidi, H. Wang, R. W. Carruth, J. A. Gardecki, and G. J. Tearney. Optical coherence tomography--near infrared spectroscopy system and catheter for intravascular imaging. *Opt. Express* 21:30849–58, 2013.

13. Fercher, A. F. *et al.* Optical coherence tomography - principles and applications. *Reports Prog. Phys.* 66:239–303, 2003.

14. Guilak, F., A. Ratcliffe, N. Lane, M. P. Rosenwasser, and V. C. Mow. Mechanical and biochemical changes in the superficial zone of articular cartilage in canine experimental osteoarthritis. *J. Orthop. Res.* 12:474–484, 1994.

15. Kim H. Esbensen, Dominique Guyot, Frank Westad, L. P. H. Multivariate Data Analysis: In Practice : an Introduction to Multivariate Data Analysis and Experimental Design. 2002, 598 pp.

16. Korhonen, R. ., M. . Laasanen, J. Töyräs, J. Rieppo, J. Hirvonen, H. . Helminen, and J. . Jurvelin. Comparison of the equilibrium response of articular cartilage in unconfined compression, confined compression and indentation. *J. Biomech.* 35:903–909, 2002.

17. Leardi, R., and L. Nørgaard. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *J. Chemom.* 18:486–497, 2004.

18. Li, H., Y. Liang, Q. Xu, and D. Cao. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* 648:77–84, 2009.

19. Li H.-D., Xu Q.-S., L. Y.-Z. An Integrated Library for Partial Least Squares Regression and Discriminant Analysis. *PeerJ Prepr. 2e190v* , 2014.

20. McGoverin, C. M., A. Hanifi, U. P. Palukuru, F. Yousefi, P. B. M. Glenn, M. Shockley, R. G. Spencer, and N. Pleshko. Non-destructive Assessment of Engineered Cartilage Composition by Near Infrared Spectroscopy. *Ann. Biomed. Eng.* , 2016.doi:10.1007/s10439-015-1536-8

21. Mehmood, T., K. H. Liland, L. Snipen, and S. Sæbø. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* 118:62–69, 2012.

22. Næs, T., and H. Martens. Principal component regression in NIR analysis: Viewpoints, background details and selection of components. *J. Chemom.* 2:155–167, 1988.

23. Padalkar, M. V., R. G. Spencer, and N. Pleshko. Near infrared spectroscopic evaluation of water in hyaline cartilage. *Ann. Biomed. Eng.* 41:2426–2436, 2013.

24. Palukuru, U. P., C. M. McGoverin, and N. Pleshko. Assessment of hyaline cartilage matrix composition using near infrared spectroscopy. *Matrix Biol.* 38:3–11, 2014.

25. Puhakka, P. H., N. C. R. te Moller, I. O. Afara, J. T. A. Mäkelä, V. Tiitu, R. K. Korhonen, H. Brommer, T. Virén, J. S. Jurvelin, and J. Töyräs. Estimation of articular cartilage properties using multivariate analysis of optical coherence tomography signal. *Osteoarthr. Cartil.* 23:2206–2213, 2015.

26. R. Leardi and A. Lupiáñez. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst.* 41:95–207, 1998.

27. Reunanen, J. Overfitting in Making Comparisons Between Variable Selection Methods. *J. Mach. Learn. Res.* 3:1371–1382, 2003.

28. Rieppo, L., J. Rieppo, J. S. Jurvelin, and S. Saarakkala. Fourier transform infrared spectroscopic imaging and multivariate regression for prediction of proteoglycan content of articular cartilage. *PLoS One* 7:, 2012.

29. Saarakkala, S., P. Julkunen, P. Kiviranta, J. Mäkitalo, J. S. Jurvelin, and R. K. Korhonen. Depth-wise progression of osteoarthritis in human articular cartilage: investigation of composition, structure and biomechanics. *Osteoarthr. Cartil.* 18:73–81, 2010.

30. Sarin, J. K., M. Amissah, H. Brommer, D. Argüelles, J. Töyräs, and I. O. Afara. Near Infrared Spectroscopic Mapping of Functional Properties of Equine Articular Cartilage. *Ann. Biomed. Eng.* 1–11, 2016.doi:10.1007/s10439-016-1659-6

31. Spahn, G. *et al.* Near-infrared (NIR) spectroscopy. A new method for arthroscopic evaluation of low grade degenerated cartilage lesions. Results of a pilot study. *BMC Musculoskelet. Disord.* 8:47, 2007.

32. Spahn, G., H. Plettenberg, H. Nagel, E. Kahl, H. M. Klinger, T. Mückley, M. Günther, G. O. Hofmann, and J. A. Mollenhauer. Evaluation of cartilage defects with near-infrared spectroscopy (NIR): an ex vivo study. *Med. Eng. Phys.* 30:285–92, 2008.

33. Suykens, J. A. K., T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. Least Squares Support Vector Machines. WORLD SCIENTIFIC, 2002.doi:10.1142/5089

34. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc.* 58:267–88, 1996.

35. Trygg, J., and S. Wold. PLS regression on wavelet compressed NIR spectra. *Chemom. Intell. Lab. Syst.* 42:209–220, 1998.

36. Westad, F., and H. Martens. Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression. *J. Near Infrared Spectrosc.* 8:117, 2000.

37. Xiaobo, Z., Z. Jiewen, M. J. W. Povey, M. Holmes, and M. Hanpin. Variables selection methods in near-infrared spectroscopy. *Anal. Chim. Acta* 667:14–32, 2010.

38. Xiaobo, Z., Z. Jiewen, H. Xingyi, and L. Yanxiao. Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of “Fuji” apple based on different PLS models. *Chemom. Intell. Lab. Syst.* 87:43–51, 2007.

39. Ying, Y., and Y. Liu. Nondestructive measurement of internal quality in pear using genetic algorithms and FT-NIR spectroscopy. *J. Food Eng.* 84:206–213, 2008.

40. Yun, Y.-H., W.-T. Wang, B.-C. Deng, G.-B. Lai, X. Liu, D.-B. Ren, Y.-Z. Liang, W. Fan, and Q.-S. Xu. Using variable combination population analysis for variable selection in multivariate calibration. *Anal. Chim. Acta* 862:14–23, 2015.