Determination of amino and fatty acid composition of soybeans using near-infrared spectroscopy

Igor Vasylyovych Kovalenko
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/rtd

Part of the Analytical Chemistry Commons, and the Bioresource and Agricultural Engineering Commons

Recommended Citation
http://lib.dr.iastate.edu/rtd/1245

This Dissertation is brought to you for free and open access by Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Determination of amino and fatty acid composition of soybeans using near-infrared spectroscopy

by

Igor Vasylyovych Kovalenko

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Agricultural Engineering (Food and Process Engineering)

Program of Study Committee:
Charles R. Hurburgh, Major Professor
Carl J. Bern
Terri D. Boylston
Thomas J. Brumm
Philip M. Dixon

Iowa State University
Ames, Iowa
2005

Copyright © Igor Vasylyovych Kovalenko, 2005. All rights reserved.
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
This is to certify that the doctoral dissertation of

Igor Vasylyovych Kovalenko

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program
# TABLE OF CONTENTS

ABSTRACT  iv

CHAPTER 1. GENERAL INTRODUCTION  1

CHAPTER 2. NEAR-INFRARED SPECTROSCOPY: DETERMINATION OF AMINO ACID COMPOSITION OF SOYBEANS  20

CHAPTER 3. MEASUREMENT OF SOYBEAN FATTY ACIDS BY NIR SPECTROSCOPY: COMPARISON OF LINEAR AND NONLINEAR CALIBRATION METHODS  49

CHAPTER 4. DIMENSIONALITY REDUCTION OF NIR SPECTRAL DATA USING GLOBAL AND LOCAL IMPLEMENTATIONS OF PCA FOR NEURAL NETWORK CALibrATIONS  67

CHAPTER 5. GENERAL CONCLUSIONS  84

APPENDIX. SAMPLE MATLAB CODE  88

ACKNOWLEDGEMENTS  97
ABSTRACT

Applicability of near-infrared spectroscopy for measurement of amino and fatty acid composition in whole soybeans was the main subject of three research papers included in this dissertation. The effects of type of spectrometer, calibration method, and data preprocessing techniques were also investigated.

Validation of amino acid calibration models resulted in $r^2$ values ranging from 0.04 (tryptophan) to 0.91 (leucine and lysine). Many of the models were usable for research purposes and sample screening, however, no sufficient correlation was found between spectral data and concentrations of cysteine and tryptophan. The variation in predictive ability of equations was determined by how a certain amino acid correlated to reference protein. Comparison of calibration methods demonstrated that (1) performance of partial least squares and support vector machines regressions was significantly better than that of artificial neural networks, and (2) the choice of preferred modeling method was spectrometer-dependent.

Validation of fatty acid calibration equations demonstrated that (1) equations for total saturates had the highest predictive ability ($r^2 = 0.91 - 0.94$) and were usable for quality assurance applications, (2) palmitic acid models ($r^2 = 0.80 - 0.84$) were usable for certain research applications, and (3) equations for stearic ($r^2 = 0.49 - 0.68$), oleic ($r^2 = 0.76 - 0.81$), linoleic ($r^2 = 0.73 - 0.76$), and linolenic ($r^2 = 0.67 - 0.74$) acids could be used for sample screening. The results also showed that support vector machines models produced significantly more accurate predictions than those developed with partial least squares.
regression. Neural network calibrations were not significantly different from the other two methods. Reduction of number of calibration samples reduced predictive ability of all types of equations. However, the rate of performance degradation of support vector machines models was the lowest.

The third study compared applicability of global and local implementations of principal component analysis compression to near-infrared calibration problems solved with the neural networks regression. This was done to better understand how neural networks could be optimized for grain subunit measurements. Two lysine data sets were used for development of control and experimental calibrations. The results demonstrated that local principal component compression could significantly outperform its traditional global counterpart.
CHAPTER 1. GENERAL INTRODUCTION

INTRODUCTION

Near-infrared (NIR) spectroscopy is a rapid, inexpensive, and generally non-destructive analytical technique that has been gaining prominence in product/process quality control applications in agricultural, food, pharmaceutical, and chemical industries. The basis of this technology lays in the fact that NIR electromagnetic energy is absorbed by organic molecules and, by Beer's Law, a sample's compositional information could be extracted from its NIR absorbance spectrum. The process of extracting information about concentration of a specific constituent from spectral data – calibration – is the main step in making NIR spectroscopy work for a specific quantitative or qualitative application. Thus, the objective of this subsection of General Introduction is to introduce readers to fundamentals of calibration of NIR spectrometers and discuss the most important aspects of the steps involved.

The purpose of the spectrometer calibration procedure is to establish a satisfactory mathematical relationship between optical properties (reflectance or transmittance spectral data) of material under investigation and known concentration of one or a few of its constituents (Figure 1.1a). Once this mathematical model is found and validated (tested), it can be applied for prediction of the constituent's concentration in samples with unknown composition from their optical data (Figure 1.1b). An accuracy and precision of predicted concentration values depend on many instrumental (hardware and software) and operator-related factors with the main one being the correctness of the calibration model. Unlike
Figure 1.1. Processes of calibration of NIR spectrometer (a) and prediction of constituent concentration from spectral data (b).
prediction, calibration is usually an expensive and time-consuming process. Therefore, to assure effectiveness of this process, it is very important not only to know the steps involved, but also to understand their implications.

NIR spectrometer calibration process can be divided into four major sequential steps:

1) Data collection
2) Preparation of calibration and validation data sets
3) Development of a calibration model
4) Model validation.

Data Collection

The objective of this step is to collect and organize chemical composition and optical data for all available samples. Composition data are obtained from reference chemical methods, while optical data (reflectance or transmittance values) are obtained by scanning the samples with the NIR spectrometer that requires calibration. The results of both analyses are compiled in the table similar to the one shown in Figure 1.2, where columns $x_1, x_2, \ldots, x_n$ contain spectral data at $n$ wavelengths $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively, and column $y$ holds the reference concentrations of constituent for which the calibration is being developed.

<table>
<thead>
<tr>
<th>Sample #</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$\ldots$</th>
<th>$x_n$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.2. Spectral and chemical composition table for calibration and validation samples.
(Note: The importance of reference analysis quality should not be underestimated, because the accuracy of calibration largely depends on that of the reference analysis. Therefore, every effort should be applied in order to assure that the reference data contains the lowest possible error. Although modern NIR spectrometers are very precise instruments, their potential will not be utilized if the calibration was performed on erroneous reference data.)

**Preparation of Calibration and Validation Data Sets**

As already mentioned, the reliability of the prediction mainly depends on the accuracy of the calibration model. This, in turn, rests on “goodness” of the sample set used for calibration and validation. Therefore, a proper preparation of the sample set is essential for the success of the calibration procedure. Two important aspects of this step include (1) selection of representative samples from the pool of all available data and (2) division of selected samples into calibration and validation subsets.

Because the developed model will be used for prediction of new unknown samples, it is essential for the sample set to include all possible sources of variation of the constituent’s concentration that can be encountered later. For example, in a case of wheat composition analysis, factors contributing to variation of protein concentration include wheat variety, origin of the samples, their moisture content, sample temperature, etc. Therefore, the calibration/validation sample set for the protein analysis must include samples with different origin, moisture content, and so on. Furthermore, these sources of variation have to be represented equally. In other words, the calibration/validation sample set has to uniformly span over the whole range of the expected constituent’s concentrations for every source of variation.
Ideally, the calibration and validation sample subsets should not be interrelated; they have to be assembled independently. However, if the pool of data available for calibration is relatively large (an order of a few hundred samples or more), and it was accumulated over an extended period of time, this requirement can be relaxed. Assuming that this is the case, one can proceed to dividing samples into two subsets. A simple procedure that splits the set into two with the ratio of 3:1 (calibration: validation) can be performed in the following steps:

1) Sorting: list samples from lowest to highest concentration of the constituent of interest
2) Selecting calibration samples: starting from the top of the list, transfer three data points into calibration sample set
3) Selecting validation sample: transfer the next data point into validation sample set
4) Repeating: process the whole list in the manner described in steps (2) and (3)

This simple procedure is applicable to populations of 100 or more samples. For smaller sample sets, it is more reasonable to use all samples for calibration and utilize cross-validation for estimation of prediction error of the model.

**Development of Calibration Model**

Calibration model development is the most important and complicated step of the procedure. Due to its complexity and comparatively large number of available methods, modeling is as much an art as a science. The objective of this step is to analyze the relationship between multiple independent variables \(x_1, x_2, \ldots, x_n\) (reflectance or transmittance at corresponding wavelengths) and dependent variable \(y\) (constituent concentration). In general, the result of this analysis is a multidimensional surface described by equation
\[ \hat{y} = f(w, X), \]  

where \( \hat{y} \) is the estimated value of concentration, \( w \) is a vector of weights, and \( X \) is a vector of predictors. The process of deriving this relationship is usually referred to as multivariate regression. An example of a typical first-order regression equation and its graphical interpretation for a single independent variable \( x_I \) is shown in Figure 1.3. In general, the mathematical problem solved by multivariate regression methods can be described as finding a set of weights \( w \), so that the function \( f(w, X) \) provides the best fit to the set of considered data points. This problem can be solved by two groups of regression algorithms:

1) Linear methods (classical least squares, multiple linear regression, principal component regression, partial least squares), which simplify the task by assuming that the relationship between \( y \) and independent variables \( x_1, x_2, \ldots, x_n \) is of the first order

2) Nonlinear methods (artificial neural networks, locally weighted regression, support vector machines), which do not limit the complexity of the relationship

Figure 1.3. A typical first-order equation for a regression problem with a single independent variable.
(Note: For implementation examples and for more information on these methods and those that are not covered here, the reader is referred to section Recommended Reading.)

**Model Validation**

Model validation is the last step of the calibration procedure. Its purpose is to test the predictive ability of the developed model using validation sample set. This is accomplished by analyzing the concentration values predicted by the model and the actual (reference) chemical data. Four most commonly used statistics in this analysis include:

1) Coefficient of determination \(r^2\), which describes the amount of common variation between NIR and concentration data:

\[
r^2 = \frac{\left( \sum \hat{y}_{y} - \frac{\sum \hat{y} \sum y}{N} \right)^2}{\left( \sum \hat{y}^2 - \frac{(\sum \hat{y})^2}{N} \right) \left( \sum y^2 - \frac{(\sum y)^2}{N} \right)},
\]

where \(N\) is a number of samples in the validation set, \(\hat{y}\) and \(y\) are estimated and reference values of concentration, respectively

2) Bias \(d\), which shows the offset of the predicted values:

\[
d = \frac{\sum (\hat{y} - y)}{N}
\]

3) Standard error of prediction (SEP) corrected for bias, which describes how precise the model will predict constituent's concentration in the future samples:
4) Relative predictive determinant \((RPD)\), which characterizes overall predictive ability of calibration model:

\[
RPD = \frac{SD_y}{SEP},
\]

where \(SD_y\) is a standard deviation of reference data in the validation set.

The calibration model is considered satisfactory if \(r^2\) value is close to 1, \(d\) is close to 0, and \(SEP\) is comparable with the error of the reference method. Alternatively, models predictive ability may be evaluated by \(RPD\) alone. \(RPD\) ranges from 1 to \(+\infty\); the higher the value, the more accurate and precise predictions will be obtained from calibration model.

**Recommended Reading**

The following is a list of books and Internet resources that provide a more in-depth coverage of calibration aspects discussed in this brief review.

**Books**


3) Data Fitting in the Chemical Sciences By the Method of Least Squares by Peter Gans; John Wiley & Sons, 1992, ISBN 0471934127


7) Near-Infrared Applications in Biotechnology by Ramesh Raghavachari; Marcel Dekker, 2001, ISBN 0824700090


**Internet Resources**

1) Council for Near Infrared Spectroscopy (March 15, 2005):
   
   http://www.idrc-chambersburg.org

2) Glossary of NIR Terminology (March 15, 2005):
   

3) Introduction to Analytical Spectroscopy (March 15, 2005):
   

4) Introduction to NIR Technology (March 15, 2005):
   

5) NIR Publications (March 15, 2005):
   
   http://www.nirpublications.com

6) NIR Spectroscopy (March 15, 2005):
   
   http://www.bafz.de/baz99_e/baz_orte/qlb/iqa/spectroscopy/nir_basics.htm

7) Quantitative Analysis Using NIR and Chemometrics (March 15, 2005):
   
   http://www.postech.ac.kr/class/chem441/exp8.htm

8) Selection of a Multivariate Calibration Method (March 15, 2005):
   

9) Spectroscopy Europe: Tony Davies Column (March 15, 2005):
   
   http://www.spectroscopyeurope.com/td_col.html

10) Theory and Principles of NIR Spectroscopy (March 15, 2005):

PROJECT BACKGROUND

Soybeans and, consequently, soybean meal are a main source of plant protein for animal feed formulation. With the development of modern diet balancing methods, increasing nutritional value of soybeans by regulating their amino acid composition (subunits of protein) has gained more attention in plant-breeding community. This, in turn, has called for development of new (or adoption of non-traditional) rapid and cost-effective techniques for amino acid measurement.

Amino acid composition is normally determined using high performance liquid chromatography. This method is neither rapid enough nor inexpensive enough for breeding applications where large numbers of seed samples have to be screened for material with required amino acid profile. As an alternative technique, NIR spectroscopy, which is being successfully utilized for measurement of grain proximates (protein, fat, carbohydrates, ash, and moisture), has been applied to the problem of subunit composition analysis by several researchers with various degrees of success.

Williams et al.\(^1\) reported satisfactory results \((r^2 = 0.66 - 0.96)\) in correlating NIR spectral data of ground wheat and barley to their amino acid concentrations. Wu et al.\(^2\) showed applicability of NIR spectroscopy for amino acid analysis of milled rice. An experiment conducted by Pazdernik et al.\(^3\) demonstrated that the accuracy of NIR screening for amino and fatty acid concentrations in soybeans was improved by grinding seed samples. An extensive research in amino acid profiling of ground grain samples and various feed ingredients was done by Fontaine et al.\(^4,5\) The researchers showed that in regards to soybeans and soybean meal most of the variation of amino acid concentrations \((84 - 98\%)\) could be
explained by NIR spectroscopy. Analysis of all these studies demonstrates that predictive ability of amino acid calibration models is dependent, among other factors, on the type of grain, sample form (whole grain or ground), and type of amino acid. However, little is known on the effect of calibration (regression) method and the type of NIR spectrometer. Therefore, to provide more insight into problem of measurement of amino acid composition of soybeans, the objective of the first research paper presented in this dissertation was to develop NIR calibrations and compare their performance for eighteen amino acids using five models of NIR spectrometers and three regression methods (partial least squares, artificial neural networks, and support vector machines).

Besides altering amino acid profile of soybeans, modification of soy fatty acid composition for improvement of nutritional and/or functional properties of soybean oil is another major objective of plant breeders. Depending on the end-user applications, several directions in soy breeding effort have been taken. Reduction of levels of polyunsaturated fatty acids (particularly linolenic acid) and increase of oleic fatty acid concentration improves oxidative stability of soybean oil during storage and processing. This, in turn, allows avoiding oil hydrogenation process that results in increased concentrations of unhealthy trans-fatty acids. Another example of a breeding strategy is development of soybean varieties with high levels of saturated fatty acids. Soy oils high in palmitic and stearic acids can be important for production of margarine and shortening.

Regardless of the strategy, one of the major elements of a breeding process is identification and keeping track of traits of many seed samples. Therefore, availability of inexpensive and rapid methods for determination of fatty acid composition of seed samples is a key element of success for development of new grain cultivars.
A number of research papers published over the last decade demonstrated applicability of NIR spectroscopy for fatty acid profiling in oilseeds. Validation of calibration models for single rapeseeds reported by Velasco et al.\textsuperscript{11} demonstrated comparatively close relationship between gas-liquid chromatography measurements and those of NIR spectroscopy for oleic ($r^2 = 0.85$) and erucic ($r^2 = 0.88$) fatty acids. However, no reliable correlation was found by these researchers for linoleic ($r^2 = 0.56$) and linolenic ($r^2 = 0.53$) acids. An earlier experiment with bulk rapeseeds conducted by Velasco and Becker\textsuperscript{12} resulted in excellent cross-validation results for oleic, linoleic, linolenic, and erucic acids ($r^2 = 0.95 - 0.98$). In contrast, determination coefficients for saturates such as palmitic, stearic, and eicosenoic acids were not as high: 0.76, 0.62, and 0.69, respectively. Studies by Sato et al.,\textsuperscript{13} Velasco et al.,\textsuperscript{14} Perez-Vich et al.,\textsuperscript{15} and Sato et al.\textsuperscript{16} provide other examples of application of NIR spectroscopy for determination of fatty acid concentrations in oil-bearing crops such as rapeseeds and sunflower seeds.

As far as soybeans are concerned, predictive ability of NIR spectroscopy for analysis of their fatty acid composition is not well documented. Dyer and Feng\textsuperscript{17} reported standard errors of performance of 2.2% for oleic acid and 1.8% for stearic acid calibrations (errors expressed as % of total fatty acids). An experiment conducted by Pazdernik et al.\textsuperscript{3} resulted in models with validation determination coefficients of 0.38 – 0.71 and 0.18 – 0.56 for fatty acids of ground and whole soybean samples, respectively. The objectives of the second research paper presented in this dissertation were to further investigate the applicability of NIR spectroscopy for analysis of fatty acid composition in whole soybeans and determine how performance of different calibration methods compares in this regard.
As mentioned before, one of the regression methods used for calibration of NIR spectrometers for measurement of amino and fatty acid composition in soybeans (first and second research papers of this dissertation) is an artificial neural network (ANN) learning algorithm. ANN, specifically its feedforward backpropagation implementation, has established itself as a strong alternative to traditional linear calibration methods used in NIR spectroscopy. Numerous applications have demonstrated its superiority to techniques that are based on principal component analysis (PCA) and partial least squares (PLS) for solving both regression and classification problems.\textsuperscript{18-23}

The attractiveness of the ANN method, in particular for regression applications, comes from the fact that it is a universal function approximation technique. It performs better than the linear methods when there is a pronounced nonlinearity in the relationship between spectral ($X$) and reference data ($y$), while it can perform as well as the linear methods when the data are linear. However, as any other calibration method, ANN modeling has its shortcomings: (1) it does not extrapolate well (which is characteristic to nonlinear calibration methods in general), therefore constituent concentration of the future samples must be within the concentration range of calibration samples; (2) it is a nondeterministic method in the sense that repeated trainings on the same data set will produce slightly different solutions; (3) the ratio of available training samples to a number of neuron interconnection weights and biases (unknown regression parameters) should be sufficiently large (on the order of tens or hundreds) to effectively employ ANN's generalization capacity. The third paper of this dissertation is focused on this last limitation of ANN calibration technique.

The simplest way to increase the ratio of training samples to a number of regression parameters is to increase the number of training samples. However, with NIR data where the
number of ANN inputs (wavelengths) alone can reach hundreds or even thousands, this task may not be feasible due to economic considerations or to the lack of samples. Another way to increase the ratio is to reduce the dimensionality of ANN input space by compressing $X$ data. PCA technique is often employed for this purpose.$^{24-27}$

PCA is a classical linear dimensionality reduction method, which transforms original correlated variables (wavelengths in our case) into a set of new uncorrelated variables or principal components (PCs). The main idea is to first determine orthogonal directions of highest variance of uncompressed data and then project the data into a new coordinate system. Resultant PCs are linear combinations of original variables, which are ordered in such a manner that several first ones capture most of the variation of the original data and the last ones retain supposedly unimportant nonlinearity and noise. Therefore, only several first PCs should be able to approximate original high-dimensional data, and the higher the collinearity in the data, the fewer PCs are needed. (However, it has been demonstrated by Yeung and Ruzzo$^{28}$ on a classification problem with microarray gene expression data that first $n$ PCs do not necessarily contain the most important information for problem solution, and there exist subsets of $n$ disjointed PCs that can produce better results.)

It should be pointed out that despite PCA’s popularity and the fact that it is the optimal transformation method when the relationship between variables of original data is for the most part linear, several other dimensionality reduction methods as applied to data from various fields of science and engineering have been shown to outperform traditional PCA. These include nonlinear variants of PCA for chemical engineering and structural dynamics applications,$^{29,30}$ wavelet-based technique in handwritten numerals recognition,$^{31}$ and methods based on data clustering in genetic studies.$^{32,33}$
An attractive data compression method that combines two multivariate data analysis techniques, namely clustering and PCA, has been described by Archer and Leen\textsuperscript{34} and Kerschen and Golinval.\textsuperscript{30} This approach, also known as local PCA, overcomes PCA's global linearity by performing dimensionality reduction task in two steps: division of the data space into clusters and local compression of each cluster using PCA. To our best knowledge, this technique has not been applied to NIR spectral data, therefore, the objective of the third paper of this dissertation was to analyze applicability of local PCA method to NIR calibration problems solved with ANN regression and compare it to traditional PCA data compression.

**GENERAL OBJECTIVE**

Over the last few decades, NIR spectroscopy has established itself as a reliable method for proximate analysis of grain. Today, the limits of this technology are being pushed further into analysis of subunits of protein and fat. To contribute to knowledge about capabilities of NIR spectroscopy, the general objective of this project was to apply and compare performance of various multivariate data analysis techniques to measurement of amino and fatty composition in soybeans.

**DISSERTATION ORGANIZATION**

This dissertation is written in the alternative format. The General Introduction section is followed by chapters containing manuscripts of three research papers: (1) Near-infrared spectroscopy: Determination of amino acid composition of soybeans, (2) Measurement of soybean fatty acids by NIR spectroscopy: Comparison of linear and nonlinear calibration
methods, and (3) Dimensionality reduction of NIR spectral data using global and local implementations of PCA for neural network calibrations. These are followed by General Conclusions, Appendix, and Acknowledgements. Appendix contains samples of custom MATLAB code used in this study. The papers were formatted for submission to the Journal of NIR Spectroscopy.

REFERENCES


CHAPTER 2. NEAR-INFRARED SPECTROSCOPY: DETERMINATION OF AMINO ACID COMPOSITION OF SOYBEANS

A paper to be submitted to Journal of Near Infrared Spectroscopy

Igor V. Kovalenko*, Glen R. Rippke*, Charles R. Hurburgh*a,b

*aDepartment of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA

bDepartment of Food Science and Human Nutrition, Iowa State University, Ames, IA

ABSTRACT

With the development of modern diet balancing methods, increasing nutritional value of soybeans by regulating their amino acid composition has gained more attention in the plant-breeding community. This, in turn, has called for development of new rapid and cost-effective techniques for amino acid measurement. Near-infrared (NIR) spectroscopy has been applied to this problem by several researchers with various degrees of success. To provide more insight into problem of measurement of amino acid composition in whole soybeans, the objective of this experiment was to develop NIR calibrations and compare their performance for eighteen amino acids using partial least squares (PLS), artificial neural networks (ANN), and support vector machines (SVM) regression methods and five models of NIR spectrometers. Validation of computed prediction models resulted in coefficients of determination $r^2$ ranging from 0.04 (tryptophan) to 0.91 (leucine and lysine). Most of the developed models were usable for research purposes and sample screening, however, no sufficient correlation was found between spectral data and concentrations of such important amino acids as cysteine and tryptophan. It was established that the variation in NIR models’
predictive ability was determined by how a certain amino acid correlated to reference protein. Therefore, future research should attempt to break this correlation by introducing calibration samples with non-typical amino acid profiles. Comparison of calibration methods demonstrated that (1) performance of PLS and LS-SVM was significantly better than that of ANN, and (2) choice of preferred modeling method was spectrometer-dependent.

INTRODUCTION

Soybeans and, consequently, soybean meal are a main source of plant protein for animal feed formulation. With the development of modern diet balancing methods, increasing nutritional value of soybeans by regulating their amino acid composition has gained more attention in the plant-breeding community. This, in turn, has called for development of new (or adoption of non-traditional) rapid and cost-effective techniques for amino acid measurement.

Amino acid composition is normally determined using high performance liquid chromatography. This method is neither rapid enough nor inexpensive enough for breeding applications where large numbers of seed samples have to be screened for material with required amino acid profile. As an alternative technique, which is being successfully utilized for measurement of grain proximates (protein, fat, carbohydrates, ash and moisture), near-infrared (NIR) spectroscopy has been applied to the problem of protein subunit composition analysis by several researchers with various degrees of success.

Williams et al.\(^1\) reported satisfactory results \((r^2 = 0.66 - 0.96)\) in correlating NIR spectral data of ground wheat and barley to their amino acid concentrations. Wu et al.\(^2\) showed
applicability of NIR spectroscopy for amino acid analysis of milled rice. An experiment conducted by Pazdernik et al. demonstrated that the accuracy of NIR screening for amino and fatty acid concentrations in soybeans may be improved by grinding seed samples. An extensive research in amino acid profiling of ground grain samples and various feed ingredients was done by Fontaine et al. The researchers showed that in regards to soybeans and soybean meal most of the variation of amino acid concentrations (84 – 98 %) could be explained by NIR spectroscopy. Analysis of all these studies demonstrates that predictive ability of amino acid calibration models is dependent, among other factors, on the type of grain, sample form (whole grain or ground), and type of amino acid. However, little is known on the effect of calibration (regression) method and the type of NIR spectrometer. Therefore, to provide more insight into problem of measurement of amino acid composition of soybeans, the objective of the present experiment was to develop NIR calibrations and compare their predictive ability for eighteen amino acids using three regression methods (main principles of which are summarized in the Theory section) and five models of NIR spectrometers (characteristics are provided in Materials and Methods).

THEORY

Several linear and nonlinear function approximation methods exist for solving spectrometer calibration problems. Three of them, partial least squares (industry standard), artificial neural networks, and support vector machines, were used in this work and are briefly discussed below.
Partial Least Squares (PLS)

The principle behind the PLS algorithm is to extract the important information from variation of both optical (X) and reference chemical composition (y) data and compress it in a set of new independent latent variables. The prediction equation becomes

\[ \hat{y} = f(w, l) = w_0 + w_1 l_1 + w_2 l_2 + w_{(p-1)} l_{(p-1)} + w_p l_p, \]

(2.1)

where \( \hat{y} \) is predicted concentration, \( w \) is a vector of weights (regression coefficients), \( l \) is a vector of new independent variables, and \( p \) is the number of latent variables. The elements of \( l \) are defined as successive linear combinations of those original variables (wavelengths) that have the greatest covariance with optical data. The optimal number of latent variables is usually found by cross-validation.

Artificial Neural Nets (ANN)

The ANN modeling technique was inspired by attempts to imitate biological neural systems capable of learning from examples. A neural network is a set of interconnected neurons. In the case of spectrometer calibrations, this network of neurons establishes a relationship between optical properties of the material and its chemical composition, provided from a set of examples, then uses it for future predictions. A trained network is a function described by number of hidden layers, number of neurons at each layer (with their transfer functions), and a set of weights (including bias terms) assigned to links connecting the neurons. For example, the equation for a neural network with \( D \) inputs, \( K \) neurons in one hidden layer, and transfer (activation) function \( \sigma \) in both output and hidden layers (Figure 2.1) takes the form:
Figure 2.1. Graphical representation of an artificial neural network with $D$ inputs, $K$ neurons in one hidden layer, and one output neuron.

$$\hat{y} = \sigma_2 \left( \sum_{j=1}^{K} \left( \sum_{i=1}^{D} w_{ij} x_i + b_j \right) v_j + b_0 \right), \quad (2.2)$$

where $X_i$ is $i$th input variable, $w_{ij}$ is the weight of the connection from $i$th input to $j$th neuron of the hidden layer (number of $w$-weights is equal to $D$ for each hidden layer neuron); $v_j$ is the weight of the connection from $j$th neuron of the hidden layer to output neuron (number of $v$-weights is equal to $K$); $b_j$ is bias of $j$th neuron of the hidden layer; $b_0$ is bias of the output neuron; $\sigma_1$ and $\sigma_2$ are functions defined, for example, as
The main limiting factor of this function approximation method is a sufficient number of training samples, because weights and biases are determined by trial and error optimization. The more complicated the network is, the more examples it needs during training process to perform adequately during prediction. When ANN is used with NIR data where the number of input variables (wavelengths) is usually large (on the order of tens, hundreds, or even thousands) and the number of training samples is limited, it is a good idea to reduce the number of dimensions of the input space using, for example, principal component analysis (PCA). The optimal number of new inputs and number of neurons in hidden layer(s), as in case with PLS, is found by minimizing cross-validation standard error.

For more details on the ANN method refer to Haykin, Cherkassky and Mulier, Williams and Norris, and Næs et al.

**Support Vector Machines (SVM)**

The SVM method is based on principles of statistical learning theory developed by Vapnik and was intended for solving classification problems. Later, this technique was adapted for linear and nonlinear function estimation.

In the SVM regression approach, data from original input space is transformed using a mapping function \( \phi(x) \) into a high dimensional feature space where linear regression is performed. This problem is formulated as constrained quadratic optimization in high-dimensional space. The solution of this problem using the Least Squares SVM regression (LS-SVM) algorithm implemented by Suykens et al. is given by the model

\[
\begin{equation}
\sigma_1(X) = \sigma_2(X) = \frac{1}{1 + \exp(-X)}.
\end{equation}
\]
\[
\hat{y} = \sum_{k=1}^{N} a_k K(x, x_k) + b,
\]

where vector \(x\) represents new sample, \(x_k\) is \(k\)th training sample, \(a_k\) is Lagrangian multiplier for \(k\)th training sample, \(b\) is bias term, \(N\) is number of training samples, \(K(x, x_k)\) is a kernel function defined as

\[
K(x, x_k) = \varphi(x)^T \varphi(x_k)
\]

In this way, SVM model contains information about relevance of each training sample for calculation of \(\hat{y}\) and makes its predictions based on relative comparison of new (unknown) sample spectra to the spectra of \(k\) training samples. SVM training is computationally intensive if \(k\) is large.

More information on SVM may be found in Vapnik et al., Smola and Scholkopf, Cherkassky and Mulier, Suykens et al., Cogdill and Dardenne.

**MATERIALS AND METHODS**

**Raw Data**

A calibration set of 526 soybean samples from 1997 – 2001 crop and a test set of 147 samples (various lines from all regions of the U.S.) from 2002 crop were used for model development and testing. NIR spectra of the whole soybeans were obtained using five NIR spectrometers: FOSS Infratec 1241 Grain Analyzer (FOSS, www.foss.dk), DICKEY-john
OmegAnalyzerG (DICKEY-john Corporation, www.dickey-john.com), Perten DA 7200 (Perten Instruments AB, www.perten.com), Bruker Optics/Cognis QTA (Bruker Optics Inc., www.brukeroptics.com and Cognis Corporation, www.cognis.com), and ASD LabSpec Pro (Analytical Spectral Devices Inc., www.asdi.com). Specifications of the instruments are provided in Table 2.1. To provide a reference for visualization of Figure 2.2 illustrates NIR scans of the same soybean sample obtained with the five spectrometers. (Note: Overlapping spectra from Perten and ASD instruments, as well as parallel data from Bruker Optics/Cognis QTA, suggest possibility for development of common calibration database.)

The primary 18 amino acids were considered in this study: alanine (ALA), arginine (ARG), aspartic acid (ASP), cysteine (CYS), glutamic acid (GLU), glycine (GLY), histidine (HIS), isoleucine (ISO), leucine (LEU), lysine (LYS), methionine (MET), phenylalanine (PHE), proline (PRO), serine (SER), threonine (THR), tryptophan (TRY), tyrosine (TYR), and valine (VAL). Their concentrations were determined at Experiment Station Chemical

Table 2.1. Specifications of the five NIR spectrometers used in the experiment.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>FOSS Infratec 1241 Grain Analyzer</th>
<th>DICKEY-john OmegAnalyzerG 7200</th>
<th>Perten DA 7200</th>
<th>Bruker Optics/Cognis QTA</th>
<th>ASD LabSpec Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>Scanning monochromator, Si detector</td>
<td>Scanning monochromator, Si detector</td>
<td>InGaAs photodiode array</td>
<td>FT-NIR, RT-PbS detector</td>
<td>Si and InGaAs photodiode arrays</td>
</tr>
<tr>
<td>Mode</td>
<td>Transmittance 850 - 1048 nm</td>
<td>Transmittance 730 - 1100 nm</td>
<td>Reflectance 950 - 1650 nm</td>
<td>Reflectance 12000 - 4000 cm²</td>
<td>Reflectance 350 - 2500 nm</td>
</tr>
<tr>
<td>Spectral range</td>
<td>7 nm</td>
<td>(not available)</td>
<td>3.125 nm/diode</td>
<td>2 - 256 cm⁻¹</td>
<td>3 nm, 10 nm</td>
</tr>
<tr>
<td>Spectral resolution</td>
<td>2.0 nm</td>
<td>0.5 nm</td>
<td>5.0 nm</td>
<td>7.7 cm⁻¹</td>
<td>1.4 nm, 2.0 nm</td>
</tr>
<tr>
<td>Number of data points</td>
<td>100</td>
<td>741</td>
<td>141</td>
<td>1037</td>
<td>2151</td>
</tr>
</tbody>
</table>

* Spectral resolution of 16 cm⁻¹ was recommended by the manufacturer for this study.
Figure 2.2. NIR scans of the same whole soybean sample obtained with five spectrometers. (DICKEY-john OmegAnalyzerG spectrum is baseline shift corrected.)
Laboratories, University of Missouri, using official method AOAC 982.30 E (a, b, c) Ch. 45.3.05.\textsuperscript{15} Statistics of reference amino acid concentrations in calibration and test samples are given in Table 2.2. Only those samples that had obviously erroneous spectra and/or concentration values were considered gross outliers and were excluded from calibration and test sets.

**Data Preprocessing**

Two methods for reduction of the light scatter effect, multiplicative scatter correction (MSC) and differentiation using Savitzky-Golay algorithm,\textsuperscript{16} were considered as primary pretreatments for spectral data. The preference was given to differentiation due to its superior effect on performance of all calibrations, regardless of regression method and spectrometer. An interesting observation was made on the effect of combining the two pretreatments for one of the spectrometers. Although not normally done in practice, performing MSC with subsequent differentiation improved prediction accuracy of ASD LabSpec Pro calibrations up to 9\% compared to differentiation alone.
Optimal combination of Savitzky-Golay algorithm parameters - window size, polynomial order, and derivative order (1\textsuperscript{st} or 2\textsuperscript{nd}) - was established based only on predictive ability of PLS calibrations. Performing search of optimal parameter sets for all three regression methods was not feasible due to tremendous amount of computation time required. (The drawback here is that this could have given some advantage to PLS over the other calibration methods.) Best preliminary results were obtained with 2\textsuperscript{nd} derivative for all spectra except for DICKEY-john data which required only 1\textsuperscript{st} order derivation. This was most likely due to the fact that raw spectra from this instrument had already been corrected for baseline shift by the instrument software.

In addition to differentiation, spectral data from all instruments were normalized to have zero mean and unity standard deviation. For more details on data transformations for each spectrometer refer to Table 2.3.

Table 2.3. Transformations (in sequential steps) applied to absorbance data from five spectrometers.

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Spectral data preprocessing</th>
</tr>
</thead>
</table>
| FOSS Infratec 1241 Grain Analyzer | 1) 2\textsuperscript{nd} derivative (5, 3)\textsuperscript{a}  
2) Normalization                                                   |
| DICKEY-john OmegAnalyzerG   | 1) 1\textsuperscript{st} derivative (17, 2)  
2) Normalization                                                      |
| Perten DA 7200              | 1) 2\textsuperscript{nd} derivative (5, 3)  
2) Normalization                                                      |
| Bruker Optics/Cognis QTA    | 1) Delete noisy data points on the range of 12000 – 11533 cm\textsuperscript{-1} (833 – 867 nm)  
2) Smooth (37, 2) noisy spectra on the range of 11533 – 910 cm\textsuperscript{-1} (867 – 1122 nm)  
3) 2\textsuperscript{nd} derivative (25, 3)  
4) Normalization                                                      |
| ASD LabSpec Pro             | 1) Delete noisy data points on the range of 350 – 440 nm  
2) Use every other wavelength for subsequent steps  
3) MSC  
4) 2\textsuperscript{nd} derivative (9, 3)  
5) Normalization                                                      |

\textsuperscript{a} Parentheses contain window size and polynomial order for Savitzky-Golay differentiation and smoothing algorithm.
Modeling

**PLS**

PLS_Toolbox 3.0 (Eigenvector Research Inc., www.eigenvector.com) for MATLAB (The MathWorks Inc., www.mathworks.com) was used for PLS modeling. The number of latent variables was selected using 5-block cross-validation on the training set.

**ANN**

MATLAB/Neural Network Toolbox (The MathWorks Inc., www.mathworks.com) was used for development of ANN calibration models. Feedforward backpropagation networks were trained on 80% of the available training patterns. The other 20% of the training set were utilized as an early stopping set to prevent over-fitting during training process. Input dimensionality was reduced by PCA. The best number of network inputs (principal components) and number of neurons in one hidden layer was determined by 5-block cross-validation on the training set. A tangent sigmoid function and linear function were used as activation functions of hidden layer neurons and an output neuron, respectively.

**LS-SVM**

LS-SVM lab 1.5 toolbox for MATLAB developed by Suykens et al.\(^{12}\) was utilized for this part of the experiment. The radial basis function (RBF)

\[
K(x, x_k) = \exp(-\|x-x_k\|^2/\sigma^2),
\]

where \(\sigma^2\) is the RBF bandwidth, was used as a kernel function.\(^{12}\) The best pair of complexity regularization parameter (required for model training) and RBF bandwidth for every amino acid was determined by 5-block cross-validation on the training set.
Model Validation

An independent test set of 147 samples was applied to all calibration models and following parameters characterizing their predictive ability were computed: (1) coefficient of determination, $r^2$, (2) standard error of prediction corrected for bias, $SEP$, (3) bias or mean difference between NIR-predicted and reference concentrations, $d$, and (4) relative predictive determinant, $RPD$. Definitions of these parameters can be found in Williams and Norris\textsuperscript{8} and AACC method 39-00.\textsuperscript{17}

RESULTS AND DISCUSSION

Overall Observations

The results of the experiment in terms of test statistics of calibration models (PLS, ANN, and LS-SVM) for all five spectrometers are provided in Tables 2.4 – 2.8. Coefficients of determination $r^2$ ranged from 0.04 (TRY) to 0.91 (LEU and LYS) and values of $RPD$ extended from 0.98 (TRY) to 3.29 (LEU). By Williams and Norris,\textsuperscript{8} four groups of amino acid calibration models could be distinguished based on $r^2$:

1) $r^2 = 0.83...0.90$: models “usable with caution for most applications”
   
   ARG, ASP, GLY, HIS, LEU, LYS, PHE, TYR

2) $r^2 = 0.66...0.81$: models usable for sample screening
   
   ALA, GLU, ISO, PRO, THR, VAL

3) $r^2 = 0.50...0.64$: models usable for rough sample screening
   
   MET, SER

4) $r^2 = 0.00...0.49$: unusable calibration models
   
   CYS, TRY
Table 2.4. FOSS Infratec 1241 Grain Analyzer: test statistics of the three types of calibration models.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Test statistic</th>
<th>PLS</th>
<th>LS-SVM</th>
<th>ANN</th>
<th>Amino Acid</th>
<th>Test statistic</th>
<th>PLS</th>
<th>LS-SVM</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>Model par.</td>
<td>3</td>
<td>67/3934</td>
<td>9/3/1</td>
<td>LYS</td>
<td>Model par.</td>
<td>5</td>
<td>273/5806</td>
<td>11/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.80</td>
<td>0.79</td>
<td>0.82</td>
<td>$r^2$</td>
<td>0.89</td>
<td>0.89</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>SEP</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.06</td>
<td>0.09</td>
<td>0.06</td>
<td>$d$</td>
<td>-0.10</td>
<td>0.14</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.22</td>
<td>2.18</td>
<td>2.33</td>
<td>RPD</td>
<td>2.89</td>
<td>2.86</td>
<td>2.88</td>
<td></td>
</tr>
<tr>
<td>ARG</td>
<td>Model par.</td>
<td>3</td>
<td>45/2141</td>
<td>3/8/1</td>
<td>MET</td>
<td>Model par.</td>
<td>3</td>
<td>145/48713</td>
<td>8/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.85</td>
<td>0.87</td>
<td>0.85</td>
<td>$r^2$</td>
<td>0.56</td>
<td>0.58</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.15</td>
<td>0.14</td>
<td>0.15</td>
<td>SEP</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.05</td>
<td>0.21</td>
<td>0.07</td>
<td>$d$</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.60</td>
<td>2.70</td>
<td>2.59</td>
<td>RPD</td>
<td>1.50</td>
<td>1.54</td>
<td>1.44</td>
<td></td>
</tr>
<tr>
<td>ASP</td>
<td>Model par.</td>
<td>5</td>
<td>80/4799</td>
<td>12/4/1</td>
<td>PHE</td>
<td>Model par.</td>
<td>5</td>
<td>59/6248</td>
<td>10/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.88</td>
<td>0.87</td>
<td>0.87</td>
<td>$r^2$</td>
<td>0.87</td>
<td>0.88</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.16</td>
<td>0.17</td>
<td>0.17</td>
<td>SEP</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.07</td>
<td>0.10</td>
<td>0.02</td>
<td>$d$</td>
<td>-0.06</td>
<td>0.08</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.88</td>
<td>2.81</td>
<td>2.82</td>
<td>RPD</td>
<td>2.82</td>
<td>2.82</td>
<td>2.80</td>
<td></td>
</tr>
<tr>
<td>CYS</td>
<td>Model par.</td>
<td>4</td>
<td>48/7638</td>
<td>8/4/1</td>
<td>PRO</td>
<td>Model par.</td>
<td>6</td>
<td>628/20415</td>
<td>12/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.36</td>
<td>0.36</td>
<td>0.33</td>
<td>$r^2$</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>SEP</td>
<td>0.09</td>
<td>0.10</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.01</td>
<td>-0.01</td>
<td>-0.02</td>
<td>$d$</td>
<td>-0.04</td>
<td>0.07</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>1.25</td>
<td>1.24</td>
<td>1.17</td>
<td>RPD</td>
<td>2.28</td>
<td>2.27</td>
<td>2.30</td>
<td></td>
</tr>
<tr>
<td>GLU</td>
<td>Model par.</td>
<td>4</td>
<td>58/5519</td>
<td>11/6/1</td>
<td>SER</td>
<td>Model par.</td>
<td>4</td>
<td>51/17770</td>
<td>4/8/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.72</td>
<td>0.72</td>
<td>0.71</td>
<td>$r^2$</td>
<td>0.53</td>
<td>0.52</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.52</td>
<td>0.52</td>
<td>0.54</td>
<td>SEP</td>
<td>0.17</td>
<td>0.17</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.19</td>
<td>0.38</td>
<td>0.10</td>
<td>$d$</td>
<td>-0.06</td>
<td>0.05</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>1.87</td>
<td>1.89</td>
<td>1.81</td>
<td>RPD</td>
<td>1.44</td>
<td>1.43</td>
<td>1.34</td>
<td></td>
</tr>
<tr>
<td>GLY</td>
<td>Model par.</td>
<td>4</td>
<td>23/4143</td>
<td>12/4/1</td>
<td>THR</td>
<td>Model par.</td>
<td>4</td>
<td>129/36867</td>
<td>12/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.86</td>
<td>0.85</td>
<td>0.84</td>
<td>$r^2$</td>
<td>0.76</td>
<td>0.76</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>SEP</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>$d$</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.65</td>
<td>2.62</td>
<td>2.53</td>
<td>RPD</td>
<td>2.01</td>
<td>1.98</td>
<td>1.94</td>
<td></td>
</tr>
<tr>
<td>HIS</td>
<td>Model par.</td>
<td>3</td>
<td>602/78469</td>
<td>10/4/1</td>
<td>TRY</td>
<td>Model par.</td>
<td>2</td>
<td>375/68599</td>
<td>12/5/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.87</td>
<td>0.89</td>
<td>0.88</td>
<td>$r^2$</td>
<td>0.07</td>
<td>0.05</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.02</td>
<td>0.03</td>
<td>0.01</td>
<td>$d$</td>
<td>0.05</td>
<td>-0.04</td>
<td>-0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.59</td>
<td>2.70</td>
<td>2.66</td>
<td>RPD</td>
<td>1.02</td>
<td>0.98</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>ISO</td>
<td>Model par.</td>
<td>4</td>
<td>812/2.4E+5</td>
<td>11/3/1</td>
<td>TYR</td>
<td>Model par.</td>
<td>4</td>
<td>115/6670</td>
<td>11/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.82</td>
<td>0.81</td>
<td>0.80</td>
<td>$r^2$</td>
<td>0.89</td>
<td>0.87</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
<td>SEP</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.08</td>
<td>0.09</td>
<td>0.07</td>
<td>$d$</td>
<td>-0.04</td>
<td>0.05</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.32</td>
<td>2.26</td>
<td>2.24</td>
<td>RPD</td>
<td>2.95</td>
<td>2.80</td>
<td>2.71</td>
<td></td>
</tr>
<tr>
<td>LEU</td>
<td>Model par.</td>
<td>6</td>
<td>207/7556</td>
<td>11/5/1</td>
<td>VAL</td>
<td>Model par.</td>
<td>4</td>
<td>50/10463</td>
<td>11/5/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>$r^2$</td>
<td>0.78</td>
<td>0.75</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.13</td>
<td>0.17</td>
<td>0.07</td>
<td>$d$</td>
<td>-0.06</td>
<td>0.08</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>3.21</td>
<td>3.29</td>
<td>3.12</td>
<td>RPD</td>
<td>2.07</td>
<td>1.97</td>
<td>2.05</td>
<td></td>
</tr>
</tbody>
</table>

*a Model parameters provide number of latent variables for PLS, number of inputs and neurons in hidden and output layers for ANN, and RBF bandwidth/complexity regularization parameter for LS-SVM.
Table 2.5. DICKEY-john OmegAnalyzerG: test statistics of the three types of calibration models.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Test statistic</th>
<th>PLS</th>
<th>LS-SVM</th>
<th>ANN</th>
<th>Amino Acid</th>
<th>Test statistic</th>
<th>PLS</th>
<th>LS-SVM</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>Model par.</td>
<td>9</td>
<td>234/37935</td>
<td>12/4/1</td>
<td>LYS</td>
<td>Model par.</td>
<td>7</td>
<td>2.5E+3/6.2E+4</td>
<td>9/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.80</td>
<td>0.80</td>
<td>0.81</td>
<td>$r^2$</td>
<td>0.86</td>
<td>0.89</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>SEP</td>
<td>0.08</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>$d$</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.25</td>
<td>2.24</td>
<td>2.26</td>
<td>RPD</td>
<td>2.68</td>
<td>2.96</td>
<td>2.79</td>
<td></td>
</tr>
<tr>
<td>ARG</td>
<td>Model par.</td>
<td>12</td>
<td>3.5E+4/2.5E+5</td>
<td>13/5/1</td>
<td>MET</td>
<td>Model par.</td>
<td>9</td>
<td>83/9143</td>
<td>8/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.77</td>
<td>0.87</td>
<td>0.85</td>
<td>$r^2$</td>
<td>0.54</td>
<td>0.48</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.19</td>
<td>0.14</td>
<td>0.15</td>
<td>SEP</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.05</td>
<td>0.04</td>
<td>0.02</td>
<td>$d$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.05</td>
<td>2.74</td>
<td>2.58</td>
<td>RPD</td>
<td>1.44</td>
<td>1.36</td>
<td>1.31</td>
<td></td>
</tr>
<tr>
<td>ASP</td>
<td>Model par.</td>
<td>7</td>
<td>5.7E+3/9.0E+4</td>
<td>16/3/1</td>
<td>PHE</td>
<td>Model par.</td>
<td>12</td>
<td>4.1E+3/7.2E+4</td>
<td>9/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.86</td>
<td>0.88</td>
<td>0.87</td>
<td>$r^2$</td>
<td>0.87</td>
<td>0.88</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.18</td>
<td>0.16</td>
<td>0.17</td>
<td>SEP</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.07</td>
<td>-0.01</td>
<td>-0.06</td>
<td>$d$</td>
<td>0.02</td>
<td>0.02</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.62</td>
<td>2.93</td>
<td>2.73</td>
<td>RPD</td>
<td>2.79</td>
<td>2.85</td>
<td>2.69</td>
<td></td>
</tr>
<tr>
<td>CYS</td>
<td>Model par.</td>
<td>6</td>
<td>3.8E+4/1.2E+5</td>
<td>7/4/1</td>
<td>PRO</td>
<td>Model par.</td>
<td>7</td>
<td>827/2.0E+5</td>
<td>8/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.32</td>
<td>0.18</td>
<td>0.29</td>
<td>$r^2$</td>
<td>0.79</td>
<td>0.77</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>SEP</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.02</td>
<td>-0.04</td>
<td>-0.03</td>
<td>$d$</td>
<td>-0.03</td>
<td>-0.02</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>1.19</td>
<td>0.98</td>
<td>1.14</td>
<td>RPD</td>
<td>2.14</td>
<td>2.07</td>
<td>2.05</td>
<td></td>
</tr>
<tr>
<td>GLU</td>
<td>Model par.</td>
<td>7</td>
<td>857/17681</td>
<td>9/5/1</td>
<td>SER</td>
<td>Model par.</td>
<td>7</td>
<td>1.8E+3/1.4E+5</td>
<td>7/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.68</td>
<td>0.71</td>
<td>0.67</td>
<td>$r^2$</td>
<td>0.46</td>
<td>0.46</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.56</td>
<td>0.54</td>
<td>0.57</td>
<td>SEP</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.02</td>
<td>0.14</td>
<td>-0.02</td>
<td>$d$</td>
<td>0.03</td>
<td>0.04</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>1.76</td>
<td>1.84</td>
<td>1.73</td>
<td>RPD</td>
<td>1.37</td>
<td>1.36</td>
<td>1.32</td>
<td></td>
</tr>
<tr>
<td>GLY</td>
<td>Model par.</td>
<td>8</td>
<td>3.4E+3/6.5E+4</td>
<td>11/3/1</td>
<td>THR</td>
<td>Model par.</td>
<td>8</td>
<td>604/1.3E+5</td>
<td>11/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.84</td>
<td>0.82</td>
<td>0.83</td>
<td>$r^2$</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>SEP</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>$d$</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.48</td>
<td>2.37</td>
<td>2.41</td>
<td>RPD</td>
<td>1.88</td>
<td>1.89</td>
<td>1.85</td>
<td></td>
</tr>
<tr>
<td>HIS</td>
<td>Model par.</td>
<td>7</td>
<td>1.0E+3/2.1E+4</td>
<td>9/5/1</td>
<td>TRY</td>
<td>Model par.</td>
<td>9</td>
<td>356/9.1E+3</td>
<td>15/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.87</td>
<td>0.89</td>
<td>0.86</td>
<td>$r^2$</td>
<td>0.08</td>
<td>0.07</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.01</td>
<td>$d$</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.69</td>
<td>2.93</td>
<td>2.59</td>
<td>RPD</td>
<td>1.01</td>
<td>0.99</td>
<td>1.03</td>
<td></td>
</tr>
<tr>
<td>ISO</td>
<td>Model par.</td>
<td>7</td>
<td>4.3E+3/7.9E+4</td>
<td>9/5/1</td>
<td>TYR</td>
<td>Model par.</td>
<td>10</td>
<td>1.2E+3/6.9E+4</td>
<td>13/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.80</td>
<td>0.79</td>
<td>0.78</td>
<td>$r^2$</td>
<td>0.86</td>
<td>0.87</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>SEP</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>$d$</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.24</td>
<td>2.16</td>
<td>2.14</td>
<td>RPD</td>
<td>2.70</td>
<td>2.79</td>
<td>2.67</td>
<td></td>
</tr>
<tr>
<td>LEU</td>
<td>Model par.</td>
<td>7</td>
<td>504/27731</td>
<td>9/3/1</td>
<td>VAL</td>
<td>Model par.</td>
<td>5</td>
<td>2.0E+3/2.8E+5</td>
<td>7/5/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.89</td>
<td>0.91</td>
<td>0.89</td>
<td>$r^2$</td>
<td>0.77</td>
<td>0.79</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.02</td>
<td>0.04</td>
<td>0.03</td>
<td>$d$</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>3.02</td>
<td>3.23</td>
<td>3.03</td>
<td>RPD</td>
<td>2.02</td>
<td>2.15</td>
<td>2.08</td>
<td></td>
</tr>
</tbody>
</table>

*aModel parameters provide number of latent variables for PLS, number of inputs and neurons in hidden and output layers for ANN, and RBF bandwidth/complexity regularization parameter for LS-SVM.
Table 2.6. Perten DA 7200: test statistics of the three types of calibration models.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Test statistic</th>
<th>PLS</th>
<th>LS-SVM</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>Model par.</td>
<td>4</td>
<td>200442835</td>
<td>11/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.80</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.16</td>
<td>2.11</td>
<td>2.13</td>
</tr>
<tr>
<td>ARG</td>
<td>Model par.</td>
<td>4</td>
<td>6152986</td>
<td>11/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.88</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.14</td>
<td>0.17</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.05</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.89</td>
<td>2.27</td>
<td>2.61</td>
</tr>
<tr>
<td>ASP</td>
<td>Model par.</td>
<td>6</td>
<td>9142.6E+4</td>
<td>13/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.89</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.93</td>
<td>2.82</td>
<td>2.81</td>
</tr>
<tr>
<td>CYS</td>
<td>Model par.</td>
<td>6</td>
<td>3719.0E+3</td>
<td>11/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.40</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>1.25</td>
<td>1.17</td>
<td>1.22</td>
</tr>
<tr>
<td>GLU</td>
<td>Model par.</td>
<td>4</td>
<td>6371.3E+4</td>
<td>9/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.71</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.55</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.16</td>
<td>0.39</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>1.84</td>
<td>1.83</td>
<td>1.84</td>
</tr>
<tr>
<td>GLY</td>
<td>Model par.</td>
<td>4</td>
<td>4015.5E+3</td>
<td>9/7/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.86</td>
<td>0.86</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.03</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.65</td>
<td>2.56</td>
<td>2.27</td>
</tr>
<tr>
<td>HIS</td>
<td>Model par.</td>
<td>4</td>
<td>2483.8E+3</td>
<td>11/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.88</td>
<td>0.87</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.87</td>
<td>2.73</td>
<td>2.51</td>
</tr>
<tr>
<td>ISO</td>
<td>Model par.</td>
<td>5</td>
<td>2.7E+3/3.5E+4</td>
<td>9/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.81</td>
<td>0.78</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.04</td>
<td>-0.03</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.28</td>
<td>2.10</td>
<td>2.00</td>
</tr>
<tr>
<td>LEU</td>
<td>Model par.</td>
<td>6</td>
<td>3.8E+3/4.2E+4</td>
<td>9/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.91</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.09</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>3.24</td>
<td>3.14</td>
<td>2.78</td>
</tr>
<tr>
<td>LYS</td>
<td>Model par.</td>
<td>6</td>
<td>2675.8E+3</td>
<td>9/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.86</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.69</td>
<td>2.56</td>
<td>2.60</td>
</tr>
<tr>
<td>ARG</td>
<td>Model par.</td>
<td>4</td>
<td>6152986</td>
<td>11/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.88</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.14</td>
<td>0.17</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.05</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.89</td>
<td>2.27</td>
<td>2.61</td>
</tr>
<tr>
<td>ASP</td>
<td>Model par.</td>
<td>6</td>
<td>9142.6E+4</td>
<td>13/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.89</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.93</td>
<td>2.82</td>
<td>2.81</td>
</tr>
<tr>
<td>CYS</td>
<td>Model par.</td>
<td>6</td>
<td>3719.0E+3</td>
<td>11/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.40</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>1.25</td>
<td>1.17</td>
<td>1.22</td>
</tr>
<tr>
<td>GLU</td>
<td>Model par.</td>
<td>4</td>
<td>6371.3E+4</td>
<td>9/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.71</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.55</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.16</td>
<td>0.39</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>1.84</td>
<td>1.83</td>
<td>1.84</td>
</tr>
<tr>
<td>GLY</td>
<td>Model par.</td>
<td>4</td>
<td>4015.5E+3</td>
<td>9/7/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.86</td>
<td>0.86</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.03</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.65</td>
<td>2.56</td>
<td>2.27</td>
</tr>
<tr>
<td>HIS</td>
<td>Model par.</td>
<td>4</td>
<td>2483.8E+3</td>
<td>11/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.88</td>
<td>0.87</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.87</td>
<td>2.73</td>
<td>2.51</td>
</tr>
<tr>
<td>ISO</td>
<td>Model par.</td>
<td>5</td>
<td>2.7E+3/3.5E+4</td>
<td>9/3/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.81</td>
<td>0.78</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.04</td>
<td>-0.03</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.28</td>
<td>2.10</td>
<td>2.00</td>
</tr>
<tr>
<td>LEU</td>
<td>Model par.</td>
<td>6</td>
<td>3.8E+3/4.2E+4</td>
<td>9/4/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.91</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.09</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>3.24</td>
<td>3.14</td>
<td>2.78</td>
</tr>
<tr>
<td>VAL</td>
<td>Model par.</td>
<td>3</td>
<td>2186932</td>
<td>8/5/1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.78</td>
<td>0.75</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.01</td>
<td>-0.02</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>RPD</td>
<td>2.10</td>
<td>1.84</td>
<td>1.91</td>
</tr>
</tbody>
</table>

*a Model parameters provide number of latent variables for PLS, number of inputs and neurons in hidden and output layers for ANN, and RBF bandwidth/complexity regularization parameter for LS-SVM.
Table 2.7. Bruker Optics/Cognis QTA: test statistics of the three types of calibration models.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Test statistic</th>
<th>PLS</th>
<th>LS-SVM</th>
<th>ANN</th>
<th>Amino Acid</th>
<th>Test statistic</th>
<th>PLS</th>
<th>LS-SVM</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>Model par.</td>
<td>5</td>
<td>309/181/927</td>
<td>10/3/1</td>
<td>LYS</td>
<td>Model par.</td>
<td>5</td>
<td>457/8.3E+4</td>
<td>12/4/1</td>
</tr>
<tr>
<td>r^2</td>
<td>0.79</td>
<td>0.79</td>
<td>0.78</td>
<td></td>
<td>r^2</td>
<td>0.85</td>
<td>0.85</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.04</td>
<td>-0.05</td>
<td>-0.04</td>
<td></td>
<td>d</td>
<td>-0.04</td>
<td>-0.05</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.16</td>
<td>2.14</td>
<td>2.15</td>
<td></td>
<td>RPD</td>
<td>2.61</td>
<td>2.60</td>
<td>2.43</td>
<td></td>
</tr>
<tr>
<td>ARG</td>
<td>Model par.</td>
<td>5</td>
<td>15/2629</td>
<td>10/3/1</td>
<td>MET</td>
<td>Model par.</td>
<td>5</td>
<td>62/1.0E+4</td>
<td>11/4/1</td>
</tr>
<tr>
<td>r^2</td>
<td>0.86</td>
<td>0.84</td>
<td>0.84</td>
<td></td>
<td>r^2</td>
<td>0.55</td>
<td>0.48</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.15</td>
<td>0.16</td>
<td>0.16</td>
<td></td>
<td>SEP</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td></td>
<td>d</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.70</td>
<td>2.53</td>
<td>2.46</td>
<td></td>
<td>RPD</td>
<td>1.45</td>
<td>1.34</td>
<td>1.40</td>
<td></td>
</tr>
<tr>
<td>ASP</td>
<td>Model par.</td>
<td>7</td>
<td>102/38427</td>
<td>12/4/1</td>
<td>PHE</td>
<td>Model par.</td>
<td>11</td>
<td>171/3.9E+4</td>
<td>12/3/1</td>
</tr>
<tr>
<td>r^2</td>
<td>0.86</td>
<td>0.86</td>
<td>0.83</td>
<td></td>
<td>r^2</td>
<td>0.85</td>
<td>0.88</td>
<td>0.8604</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.18</td>
<td>0.18</td>
<td>0.20</td>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.07</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0.02</td>
<td>0.02</td>
<td>0.07</td>
<td></td>
<td>d</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.71</td>
<td>2.68</td>
<td>2.41</td>
<td></td>
<td>RPD</td>
<td>2.62</td>
<td>2.90</td>
<td>2.60</td>
<td></td>
</tr>
<tr>
<td>CYS</td>
<td>Model par.</td>
<td>4</td>
<td>72/13.5E+4</td>
<td>11/5/1</td>
<td>PRO</td>
<td>Model par.</td>
<td>5</td>
<td>194/6.4E+4</td>
<td>10/4/1</td>
</tr>
<tr>
<td>r^2</td>
<td>0.39</td>
<td>0.33</td>
<td>0.30</td>
<td></td>
<td>r^2</td>
<td>0.79</td>
<td>0.78</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.04</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
<td>SEP</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td></td>
<td>d</td>
<td>0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>1.25</td>
<td>1.19</td>
<td>1.18</td>
<td></td>
<td>RPD</td>
<td>2.17</td>
<td>2.15</td>
<td>2.05</td>
<td></td>
</tr>
<tr>
<td>GLU</td>
<td>Model par.</td>
<td>5</td>
<td>2.7E+3/2.7E+5</td>
<td>12/5/1</td>
<td>SER</td>
<td>Model par.</td>
<td>4</td>
<td>29/2.4E+4</td>
<td>7/5/1</td>
</tr>
<tr>
<td>r^2</td>
<td>0.70</td>
<td>0.68</td>
<td>0.68</td>
<td></td>
<td>r^2</td>
<td>0.52</td>
<td>0.46</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.56</td>
<td>0.58</td>
<td>0.58</td>
<td></td>
<td>SEP</td>
<td>0.17</td>
<td>0.18</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.01</td>
<td>-0.09</td>
<td>0.08</td>
<td></td>
<td>d</td>
<td>-0.03</td>
<td>-0.05</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>1.82</td>
<td>1.77</td>
<td>1.76</td>
<td></td>
<td>RPD</td>
<td>1.43</td>
<td>1.36</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>GLY</td>
<td>Model par.</td>
<td>7</td>
<td>519/1.2E+5</td>
<td>13/3/1</td>
<td>THR</td>
<td>Model par.</td>
<td>7</td>
<td>642/4.4E+4</td>
<td>12/3/1</td>
</tr>
<tr>
<td>r^2</td>
<td>0.84</td>
<td>0.83</td>
<td>0.80</td>
<td></td>
<td>r^2</td>
<td>0.72</td>
<td>0.70</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td></td>
<td>SEP</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.01</td>
<td></td>
<td>d</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.49</td>
<td>2.41</td>
<td>2.23</td>
<td></td>
<td>RPD</td>
<td>1.87</td>
<td>1.82</td>
<td>1.70</td>
<td></td>
</tr>
<tr>
<td>HIS</td>
<td>Model par.</td>
<td>7</td>
<td>58/2.3E+4</td>
<td>15/3/1</td>
<td>TRY</td>
<td>Model par.</td>
<td>3</td>
<td>7.3E+5/553</td>
<td>4/6/1</td>
</tr>
<tr>
<td>r^2</td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td></td>
<td>r^2</td>
<td>0.14</td>
<td>0.04</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td></td>
<td>SEP</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.01</td>
<td></td>
<td>d</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.57</td>
<td>2.55</td>
<td>2.35</td>
<td></td>
<td>RPD</td>
<td>1.07</td>
<td>0.98</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>ISO</td>
<td>Model par.</td>
<td>7</td>
<td>809/6.6E+4</td>
<td>14/5/1</td>
<td>TYR</td>
<td>Model par.</td>
<td>6</td>
<td>100/1.7E+4</td>
<td>12/3/1</td>
</tr>
<tr>
<td>r^2</td>
<td>0.80</td>
<td>0.80</td>
<td>0.74</td>
<td></td>
<td>r^2</td>
<td>0.87</td>
<td>0.88</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.09</td>
<td></td>
<td>SEP</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.02</td>
<td></td>
<td>d</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.22</td>
<td>2.21</td>
<td>1.96</td>
<td></td>
<td>RPD</td>
<td>2.77</td>
<td>2.85</td>
<td>2.57</td>
<td></td>
</tr>
<tr>
<td>LEU</td>
<td>Model par.</td>
<td>8</td>
<td>222/4.4E+4</td>
<td>13/3/1</td>
<td>VAL</td>
<td>Model par.</td>
<td>3</td>
<td>143/236924</td>
<td>13/4/1</td>
</tr>
<tr>
<td>r^2</td>
<td>0.89</td>
<td>0.89</td>
<td>0.86</td>
<td></td>
<td>r^2</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.02</td>
<td></td>
<td>d</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.97</td>
<td>2.98</td>
<td>2.63</td>
<td></td>
<td>RPD</td>
<td>1.97</td>
<td>2.04</td>
<td>2.03</td>
<td></td>
</tr>
</tbody>
</table>

^1 Model parameters provide number of latent variables for PLS, number of inputs and neurons in hidden and output layers for ANN, and RBF bandwidth/complexity regularization parameter for LS-SVM.
Table 2.8. ASD LabSpec Pro: test statistics of the three types of calibration models.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Test statistic&lt;sup&gt;a&lt;/sup&gt;</th>
<th>PLS</th>
<th>LS-SVM</th>
<th>ANN</th>
<th>Amino Acid</th>
<th>Test statistic&lt;sup&gt;a&lt;/sup&gt;</th>
<th>PLS</th>
<th>LS-SVM</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>Model par.</td>
<td>3</td>
<td>3.0E+3/2.7E+3</td>
<td>12/3/1</td>
<td>LYS</td>
<td>Model par.</td>
<td>4</td>
<td>135/2.9E+5</td>
<td>21/3/1</td>
</tr>
<tr>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.79</td>
<td>0.72</td>
<td>0.75</td>
<td></td>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.87</td>
<td>0.88</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.06</td>
<td>0.07</td>
<td>0.06</td>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.07</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.04</td>
<td>-0.05</td>
<td>-0.04</td>
<td></td>
<td>d</td>
<td>-0.04</td>
<td>-0.03</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.17</td>
<td>1.90</td>
<td>1.98</td>
<td></td>
<td>RPD</td>
<td>2.81</td>
<td>2.85</td>
<td>2.32</td>
<td></td>
</tr>
<tr>
<td>ARG</td>
<td>Model par.</td>
<td>4</td>
<td>3.8E+3/2.0E+3</td>
<td>8/7/1</td>
<td>MET</td>
<td>Model par.</td>
<td>2</td>
<td>309/1.7E+3</td>
<td>13/3/1</td>
</tr>
<tr>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.87</td>
<td>0.80</td>
<td>0.82</td>
<td></td>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.57</td>
<td>0.51</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.14</td>
<td>0.18</td>
<td>0.17</td>
<td></td>
<td>SEP</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.01</td>
<td></td>
<td>d</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.78</td>
<td>2.21</td>
<td>2.35</td>
<td></td>
<td>RPD</td>
<td>1.49</td>
<td>1.42</td>
<td>1.35</td>
<td></td>
</tr>
<tr>
<td>ASP</td>
<td>Model par.</td>
<td>6</td>
<td>1.5E+4/4.8E+3</td>
<td>18/4/1</td>
<td>PHE</td>
<td>Model par.</td>
<td>6</td>
<td>316/2.2E+5</td>
<td>12/4/1</td>
</tr>
<tr>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.86</td>
<td>0.84</td>
<td>0.84</td>
<td></td>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.87</td>
<td>0.87</td>
<td>0.815</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
<td></td>
<td>SEP</td>
<td>0.07</td>
<td>0.07</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0.13</td>
<td>0.06</td>
<td>0.09</td>
<td></td>
<td>d</td>
<td>0.02</td>
<td>0.02</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.58</td>
<td>2.50</td>
<td>2.51</td>
<td></td>
<td>RPD</td>
<td>2.79</td>
<td>2.74</td>
<td>2.30</td>
<td></td>
</tr>
<tr>
<td>CYS</td>
<td>Model par.</td>
<td>3</td>
<td>5.9E+3/1.1E+3</td>
<td>9/4/1</td>
<td>PRO</td>
<td>Model par.</td>
<td>4</td>
<td>53.5/6.4E+4</td>
<td>13/5/1</td>
</tr>
<tr>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.30</td>
<td>0.22</td>
<td>0.31</td>
<td></td>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.80</td>
<td>0.79</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
<td>SEP</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td></td>
<td>d</td>
<td>0.02</td>
<td>0.03</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>1.11</td>
<td>1.12</td>
<td>1.13</td>
<td></td>
<td>RPD</td>
<td>2.21</td>
<td>2.19</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>GLU</td>
<td>Model par.</td>
<td>4</td>
<td>57/4.4E+3</td>
<td>16/4/1</td>
<td>SER</td>
<td>Model par.</td>
<td>3</td>
<td>2.2E+5/1.5E+3</td>
<td>9/5/1</td>
</tr>
<tr>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.72</td>
<td>0.69</td>
<td>0.68</td>
<td></td>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.50</td>
<td>0.43</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.55</td>
<td>0.58</td>
<td>0.59</td>
<td></td>
<td>SEP</td>
<td>0.17</td>
<td>0.19</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.05</td>
<td>-0.07</td>
<td>-0.01</td>
<td></td>
<td>d</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>1.86</td>
<td>1.78</td>
<td>1.74</td>
<td></td>
<td>RPD</td>
<td>1.42</td>
<td>1.32</td>
<td>1.35</td>
<td></td>
</tr>
<tr>
<td>GLY</td>
<td>Model par.</td>
<td>4</td>
<td>40/5.3E+4</td>
<td>17/3/1</td>
<td>THR</td>
<td>Model par.</td>
<td>3</td>
<td>1.4E+6/1.8E+3</td>
<td>9/3/1</td>
</tr>
<tr>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.84</td>
<td>0.82</td>
<td>0.82</td>
<td></td>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.73</td>
<td>0.67</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td></td>
<td>SEP</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
<td>d</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.46</td>
<td>2.34</td>
<td>2.30</td>
<td></td>
<td>RPD</td>
<td>1.90</td>
<td>1.72</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td>HIS</td>
<td>Model par.</td>
<td>4</td>
<td>85/1.3E+5</td>
<td>21/3/1</td>
<td>TRY</td>
<td>Model par.</td>
<td>2</td>
<td>2.8E+3/1.0E+3</td>
<td>4/5/1</td>
</tr>
<tr>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.86</td>
<td>0.87</td>
<td>0.81</td>
<td></td>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.13</td>
<td>0.06</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td></td>
<td>SEP</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
<td>d</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.62</td>
<td>2.71</td>
<td>2.25</td>
<td></td>
<td>RPD</td>
<td>1.06</td>
<td>1.02</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>ISO</td>
<td>Model par.</td>
<td>3</td>
<td>3.1E+5/2.3E+3</td>
<td>8/3/1</td>
<td>TYR</td>
<td>Model par.</td>
<td>4</td>
<td>46/7.1E+4</td>
<td>10/4/1</td>
</tr>
<tr>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.76</td>
<td>0.68</td>
<td>0.72</td>
<td></td>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.85</td>
<td>0.86</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.09</td>
<td>0.10</td>
<td>0.09</td>
<td></td>
<td>SEP</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.03</td>
<td>-0.04</td>
<td>-0.05</td>
<td></td>
<td>d</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>2.02</td>
<td>1.76</td>
<td>1.86</td>
<td></td>
<td>RPD</td>
<td>2.56</td>
<td>2.64</td>
<td>2.30</td>
<td></td>
</tr>
<tr>
<td>LEU</td>
<td>Model par.</td>
<td>6</td>
<td>26/2.8E+5</td>
<td>21/3/1</td>
<td>VAL</td>
<td>Model par.</td>
<td>3</td>
<td>99/6/1713</td>
<td>8/5/1</td>
</tr>
<tr>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.90</td>
<td>0.90</td>
<td>0.86</td>
<td></td>
<td>r&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.75</td>
<td>0.64</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.09</td>
<td>0.10</td>
<td>0.12</td>
<td></td>
<td>SEP</td>
<td>0.08</td>
<td>0.10</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.03</td>
<td></td>
<td>d</td>
<td>-0.01</td>
<td>0.00</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>3.11</td>
<td>3.07</td>
<td>2.52</td>
<td></td>
<td>RPD</td>
<td>1.98</td>
<td>1.67</td>
<td>1.84</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Model parameters provide number of latent variables for PLS, number of inputs and neurons in hidden and output layers for ANN, and RBF bandwidth/complexity regularization parameter for LS-SVM.
Figure 2.3. Relationship between $RPD$ and $r^2$. Each data point represents one of 270 calibration models developed in this study. Mathematically, the relationship can be expressed as $RPD = 1/(1-r^2)^{0.5}$ (from personal communication with David B. Funk). 

Because of correlation between $r^2$ and $RPD$ (Figure 2.3), the same classification of models could be derived from $RPD$ values.

It is worth mentioning, that $r^2$ values of this experiment were generally higher than those previously reported by Pazdernik et al. for both whole-seed and ground-seed soybean samples. This could most likely be attributed to a much larger calibration set used in this study (526 samples vs. 90 samples) and form of expression of amino acid concentrations (% of total sample weight vs. % of crude protein). Validation statistics of NIR calibrations for LEU, LYS, MET, and THR in ground soybeans reported by Fontaine et al. were superior to our results, which supports the theory that grinding soybean samples improves predictive
ability of NIR spectroscopy for amino acid or protein measurement. Unfortunately, in this case NIR spectroscopy loses its non-invasive property.

The attempts to explain the variation in models' predictive ability by correlating $r^2$ or $RPD$ of a specific regression method to such properties of amino acids as average reference concentration in soybeans, relative variation of concentration (range divided by average), molecular weight, solubility in water, and isoelectric point did not result in any reliable relationship. However, when NIR $RPD$ values were compared to those of linear regression of reference amino acids to reference protein, it became apparent that variation in NIR models' predictive ability was determined by how a certain amino acid was correlated to protein (for reference, Table 2.2 provides determination coefficients describing correlation of a particular amino acid with protein). Figure 2.4 illustrates this concept. If amino acid concentration can

![Figure 2.4. Scatter plot of $RPD$ values of linear protein regression models vs. those of the best NIR calibration models.](image)
be predicted accurately from known value of protein concentration, it can be accurately estimated using NIR spectroscopy. If the correlation is poor, as the case with CYS and TRY, NIR predictions will be equally inaccurate. A similar observation about correlation between soy amino acid contents predicted by NIR and linear protein regression, although based on bias values, was made by Fontaine et al.\textsuperscript{4} This implies that NIR spectroscopy measures amino acid concentration in whole soybeans indirectly by deriving it from the total amount of nitrogen-containing molecules. Analysis of regression vectors of PLS calibration models supports this statement. Most of the regression curves in Figure 2.5 follow the same pattern, which indicates that, for the most part, calibrations predict protein. Those regression vectors that fall out of the general pattern represent low-$R_{PD}$ calibration models such as TRY, CYS, and SER. Therefore, the biggest challenge that is faced in NIR measurement of amino acids

![Figure 2.5](image)

Figure 2.5. Regression vectors of 18 amino acid calibration models (PLS regression) developed for FOSS Infratec 1241 Grain Analyzer. Most of the curves follow the same pattern, which indicates that, for the most part, calibrations predict protein.
in soybeans - and possibly in other legumes and cereal grains - is to break the correlation between amino acid and protein concentrations. Future research should attempt to address this issue by introducing calibration samples (possibly artificially created) with unusual amino acid profiles.

Comparison of Calibration Methods and Spectrometers

Overall performance of three calibration methods and five spectrometers was evaluated and compared based on RPD coefficient, which is viewed as a standardized parameter of model's predictive ability. The effects of regression method and type of spectrometer on RPD was tested using analysis of variance and least squares fit of the form

\[ RPD = AA + M + S + M*S + \text{Error}, \]  

(2.7)

where AA is amino acid factor, M is method factor, S is spectrometer factor, and M*S is method-spectrometer interaction. The analysis indicated that all factors used in the model had significant effect (p < 0.0001) on RPD. Due to a large number of samples, M*S interaction factor could be ignored for all practical purposes. As far as calibration methods are concerned, performance of PLS and LS-SVM regressions was significantly better (\( \alpha = 0.05 \)) than that of ANN (Figure 2.6). ANN's inferior performance could most likely be explained by (1) an insufficient size of training set for this method, and/or (2) use of PCA for dimensionality reduction of the input space, which discards information on nonlinearity that is contained in high-order principal components. Comparison of overall performance of spectrometers demonstrated significant advantage (\( \alpha = 0.05 \)) of FOSS Infratec 1241 Grain Analyzer (Figure 2.7), and this is in spite of its shortest optical range and smallest number of spectral data points among all tested instruments.
Figure 2.6. Mean RPD values (based on 18 amino acids x 5 spectrometers) of calibration models for three regression methods. Error bars indicate +/- three standard errors. Means with the same letter are not significantly different (α = 0.05) by Tukey HSD test.

Figure 2.7. Mean RPD values (based on 18 amino acids x 3 regression methods) of calibration models for five spectrometers. Error bars indicate +/- three standard errors. Means with the same letter are not significantly different (α = 0.05) by Tukey HSD test.
In order to determine whether the same calibration priority pattern, PLS − LS-SVM − ANN, applied to all spectrometers, levels of method-spectrometer interaction factor were analyzed (Figure 2.8). Results showed that the choice of a preferred calibration method as well as the advantage of one calibration method over the other were spectrometer-dependent. While Perten DA 7200 and ASD LabSpec Pro had largest differences in the mean RPD values of calibration methods, FOSS Infratec 1241 Grain Analyzer showed nearly identical performance of all methods. DICKEY-john OmegAnalyzerG, unlike the other spectrometers, demonstrated an advantage of LS-SVM method over PLS.

Since RPD variations among calibration methods were not the same for all instruments, calibration models of the best-performing methods (PLS for FOSS Infratec 1241 Grain Analyzer, Perten DA 7200, Bruker Optics/Cognis QTA, and ASD LabSpec Pro; and LS-
SVM for DICKEY-john OmegAnalyzerG) were used to further compare performance of the spectrometers. Analysis of least squares fit of the form

\[ RPD = AA + S + Error \]  

(2.8)

demonstrated significance of amino acid (as expected), and spectrometer factors \((p < 0.0001\) and \(p < 0.0029\), respectively). Figure 2.9 shows that the difference between spectrometers became less distinct. Amino acid predictive ability of Perten DA 7200, which had the highest mean \( RPD \), was comparable to that of FOSS Infratec 1241 Grain Analyzer and DICKEY-
john OmegAnalyzerG, but significantly better than Bruker Optics/Cognis QTA and ASD LabSpec Pro.

An interesting observation was made by analyzing bias on the test set (Figure 2.10). While average bias of all amino acid predictions from four spectrometers approached zero, all of the FOSS Infratec PLS calibration models except for the two unusable calibrations, CYS and TRY, had a negative bias, indicating that this spectrometer tended to overpredict amino acid concentrations. A completely opposite pattern was observed with this spectrometer in combination with LS-SVM and ANN calibration methods: all of the predictions except for CYS and TRY had a positive bias (figure not shown), indicating that with nonlinear calibrations the spectrometer underpredicted amino acid concentrations. This phenomenon could not be explained, since the calibration and test sets were nearly identical for all spectrometers.

CONCLUSIONS

Calibration models for determination of amino acid concentration in whole soybeans were developed using five NIR spectrometers and three regression methods. Models were characterized by various degrees of accuracy, and most of them were usable for research purposes and sample screening. Unfortunately, no correlation could be established between spectral data and concentrations of such important amino acids as CYS and TRY. The variation in NIR models' predictive ability was determined by the degree to which a certain amino acid correlated to crude protein. Therefore, future research should attempt to
Figure 2.10. Test set bias values of best-performing calibration methods (PLS for FOSS Infratec 1241 Grain Analyzer, Perten DA 7200, Bruker Optics/Cognis QTA, and ASD LabSpec Pro; and LS-SVM for DICKEY-john OmegAnalyzerG) for five tested spectrometers.
break this correlation by introducing calibration samples with non-typical amino acid profiles.

Comparison of calibration methods demonstrated that (1) performance of PLS and LS-SVM was significantly better than that of ANN, and (2) choice of preferred modeling method was spectrometer-dependent. Comparison of instruments showed some advantage of FOSS Infratec 1241 Grain Analyzer and Perten DA 7200.

REFERENCES


18. Personal communication with David B. Funk (Szent Istvan University, Budapest, Hungary) during 12th International Diffuse Reflectance Conference (IDRC), Chambersburg, PA (2004).
CHAPTER 3. MEASUREMENT OF SOYBEAN FATTY ACIDS BY NIR SPECTROSCOPY: COMPARISON OF LINEAR AND NONLINEAR CALIBRATION METHODS

A paper to be submitted to Journal of Near Infrared Spectroscopy

Igor V. Kovalenko\textsuperscript{a}, Glen R. Rippke\textsuperscript{a}, Charles R. Hurburgh\textsuperscript{a,b}

\textsuperscript{a} Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA
\textsuperscript{b} Department of Food Science and Human Nutrition, Iowa State University, Ames, IA

ABSTRACT

Improvement of nutritional and/or functional properties of soybean oil by modification of soy fatty acid composition is one of the objectives of plant breeders. The key element of success for development of new cultivars is availability of inexpensive and rapid method for measurement of fatty acids in seed samples. A number of research papers published over the last decade demonstrated applicability of near-infrared (NIR) spectroscopy for fatty acid profiling in oilseeds. Therefore, the objectives of this study were to investigate the applicability of NIR spectroscopy for analysis of fatty acid composition in whole soybeans and determine how performance of different calibration methods compares in this regard. Calibration equations were developed using partial least squares (PLS), artificial neural networks (ANN), and support vector machines (SVM) regression methods. Validation results demonstrated that (1) equations for total saturates had the highest predictive ability ($r^2 = 0.91 - 0.94$) and were usable for quality assurance applications, (2) palmitic acid models ($r^2 = 0.80 - 0.84$) were usable for certain research applications, and (3) equations for stearic ($r^2 = 0.49 - 0.68$), oleic ($r^2 = 0.76 - 0.81$), linoleic ($r^2 = 0.73 - 0.76$), and linolenic ($r^2 = 0.67 - 0.74$) acids
could be used for sample screening. The results also showed that SVM models produced significantly more accurate predictions than those developed with PLS regression. ANN calibrations were not significantly different from the other two methods. Reduction of number of calibration samples reduced predictive ability of all types of NIR equations, however the rate of performance degradation of SVM models was the lowest.

**INTRODUCTION**

Improvement of nutritional and/or functional properties of soybean oil by modification of soy fatty acid composition is one of the objectives of plant breeders. Depending on the end-user applications, several directions in soy breeding effort have been taken. Reduction of levels of polyunsaturated fatty acids (particularly linolenic acid) and increase of oleic fatty acid concentration improves oxidative stability of soybean oil during storage and processing. This, in turn, allows avoiding oil hydrogenation process that results in increased concentrations of unhealthy trans-fatty acids.\(^1\)\(^-\)\(^3\) Another example of a breeding strategy is development of soybean varieties with high levels of saturated fatty acids. Soy oils high in palmitic and stearic acids can be important for production of margarine and shortening.\(^4\)\(^-\)\(^5\)

Regardless of the strategy, one of the major elements of a breeding process is identification and tracking of traits of many seed samples. Therefore, availability of inexpensive and rapid methods for determination of fatty acid composition of seed samples is a key element of success for development of new grain cultivars.

A number of research papers published over the last decade demonstrated applicability of near-infrared (NIR) spectroscopy for fatty acid profiling in oilseeds. Validation of calibration models for single rapeseeds reported by Velasco *et al.*\(^6\) demonstrated comparatively close
relationship between gas-liquid chromatography measurements and those of NIR spectroscopy for oleic ($r^2 = 0.85$) and erucic ($r^2 = 0.88$) fatty acids. However, no reliable correlation was found by these researchers for linoleic ($r^2 = 0.56$) and linolenic ($r^2 = 0.53$) acids. An earlier experiment with bulk rapeseeds conducted by Velasco and Becker\textsuperscript{7} resulted in excellent cross-validation results for oleic, linoleic, linolenic, and erucic acids ($r^2 = 0.95 - 0.98$). In contrast, determination coefficients for saturates such as palmitic, stearic, and eicosenoic acids were not as high: 0.76, 0.62, and 0.69, respectively. Studies by Sato \textit{et al.},	extsuperscript{8,9} Velasco \textit{et al.},	extsuperscript{10} and Perez-Vich \textit{et al.} \textsuperscript{11} provide other examples of application of NIR spectroscopy for determination of fatty acid concentrations in oil-bearing crops such as rapeseeds and sunflower seeds.

In soybeans, predictive ability of NIR spectroscopy for fatty acid analysis is not well documented. Dyer and Feng\textsuperscript{12} reported standard errors of performance of 2.2 % for oleic acid and 1.8 % (in % of total fatty acids) for stearic acid calibrations. An experiment conducted by Pazdernik \textit{et al.} \textsuperscript{13} resulted in models with validation determination coefficients of 0.38 – 0.71 and 0.18 – 0.56 for fatty acids of ground and whole soybean samples, respectively. The objectives of this study were to (1) further investigate the applicability of NIR spectroscopy for analysis of fatty acid composition in whole soybeans, and (2) determine how performance of three linear and nonlinear calibration methods compares in this regard.

\textbf{MATERIALS AND METHODS}

\textbf{Raw Data}

A pool of approximately 1,400 soybean samples (U.S. crop of 1991, 1993 – 1998, and 2003) with fatty acid profiles was used in this study. Whole soy samples were scanned on
three FOSS spectrometers with the same spectral characteristics, Infratec Grain Analyzers
1225, 1229, and 1241 (FOSS Group, www.foss.dk), and a common calibration database
consisting of 4,144 scans was created. The Infratec was used because (1) the majority of
units in commercial trade are Infratecs, and (2) this instrument was as effective as any other
unit in the previous amino acid study. Concentrations of total saturates (palmitic plus stearic),
palmitic (C16:0), stearic (C18:0), oleic (C18:1), linoleic (C18:2), and linolenic (C18:3) acids
(in % of total fatty acids) were determined in Department of Agronomy at Iowa State
University using gas chromatography method described by Hammond.\textsuperscript{14}

Selection of Samples and Data Preprocessing

To reduce computational load by removing redundant information from the plane of
reference values, the original calibration database was resampled and six new approximately
uniformly distributed (based on fatty acid concentrations) data subsets were created for
further analysis. In addition, samples with abnormally low or high variation of 2\textsuperscript{nd} derivative
of NIR signal were considered spectral outliers (Figure 3.1) and excluded from further
calculations. Reference data statistics for the six final calibration/validation subsets are
provided in Table 3.1. 75 % of the samples from each subset were used for calibration and
the other 25 % were reserved for model validation.

NIR spectra were corrected for scatter effects by estimating their 2\textsuperscript{nd} derivative using
Savitzky-Golay algorithm\textsuperscript{15} (5-point window and 3\textsuperscript{rd}-order polynomial). In addition,
samples' spectral and reference data (rows) were normalized to have zero mean and unity
standard deviation.
Figure 3.1. Detection of spectral outliers by analyzing variations of 2nd derivative of log(1/T).

Table 3.1. Reference data statistics and a total number of soybean samples used in this study for calibration and model validation (concentrations are in % of total fatty acids)

<table>
<thead>
<tr>
<th>Fatty acid</th>
<th>Number of samples(^b)</th>
<th>Mean, %</th>
<th>Range, %</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total saturates (0.003)(^a)</td>
<td>721</td>
<td>16.4</td>
<td>5.3 - 37.3</td>
<td>7.46</td>
</tr>
<tr>
<td>Palmitic C16:0 (0.001)</td>
<td>616</td>
<td>11.0</td>
<td>2.8 - 32.3</td>
<td>6.94</td>
</tr>
<tr>
<td>Stearic C18:0 (0.01)</td>
<td>619</td>
<td>4.7</td>
<td>2.0 - 8.4</td>
<td>1.34</td>
</tr>
<tr>
<td>Oleic C18:1 (0.15)</td>
<td>771</td>
<td>27.7</td>
<td>11.8 - 51.0</td>
<td>8.71</td>
</tr>
<tr>
<td>Linoleic C18:2 (0.03)</td>
<td>758</td>
<td>51.4</td>
<td>32.4 - 69.4</td>
<td>7.27</td>
</tr>
<tr>
<td>Linolenic C18:3 (0.33)</td>
<td>976</td>
<td>6.5</td>
<td>1.2 - 13.3</td>
<td>3.01</td>
</tr>
</tbody>
</table>

\(^a\) Parentheses contain determination coefficient describing correlation of a fatty acid (in % of total fatty acids) with total oil content.

\(^b\) Different number of samples was required to create a uniformly distributed set for each fatty acid.
Calibration Procedures: Theory

One linear (partial least squares) and two nonlinear (artificial neural networks and support vector machines) regression methods were used in this study for development of calibration models. Their brief description is provided below.

**Partial Least Squares (PLS)**

The principle behind the PLS on algorithm is to extract the important information from variation of both optical ($X$) and reference chemical composition ($y$) data and compress it in a set of new independent latent variables. The prediction equation becomes

$$y = \hat{y} = W_0 + W_1 f_1 + W_2 f_2 + \ldots + W_p f_p,$$  \hspace{1cm} (3.1)

where $\hat{y}$ is predicted concentration, $W$ is a vector of weights (regression coefficients), $f$ is a vector of new independent variables, and $p$ is the number of latent variables. The elements of $f$ are defined as successive linear combinations of those original variables (wavelengths) that have the greatest covariance with optical data. The optimal number of latent variables is usually found by cross-validation.

**Artificial Neural Networks (ANN)**

The ANN modeling technique was inspired by attempts to imitate biological neural systems that are capable of learning on examples. A neural network is a set of interconnected neurons. In the case of spectrometer calibrations, this network of neurons establishes a relationship between optical properties of the material and its chemical composition, provided from a set of examples, then uses it for future predictions. A trained network is a function described by number of hidden layers, number of neurons at each layer (with their
transfer functions), and a set of weights (including bias terms) assigned to links connecting the neurons. For example, the equation for a neural network with \( D \) inputs, \( K \) neurons in one hidden layer, and transfer (activation) function \( \sigma \) in both output and hidden layers takes form:

\[
\hat{y} = \sigma \left[ \sum_{j=1}^{K} \sigma \left( \sum_{i=1}^{D} w_{ij} X_i + b_j \right) \nu_j + b_0 \right],
\]

(3.2)

where \( X_i \) is \( i \)th input variable, \( w_{ij} \) is the weight of the connection from \( i \)th input to \( j \)th neuron of the hidden layer (number of \( w \)-weights is equal to \( D \) for each hidden layer neuron); \( \nu_j \) is the weight of the connection from \( j \)th neuron of the hidden layer to output neuron (number of \( \nu \)-weights is equal to \( K \)); \( b_j \) is bias of \( j \)th neuron of the hidden layer; \( b_0 \) is bias of the output neuron; \( \sigma \) and \( \sigma \) are functions defined, for example, as

\[
\sigma_1(X) = \frac{1}{1 + \exp(-X)}.
\]

(3.3)

The main limiting factor of this function approximation method is a sufficient number of training samples. More complicated networks require more training examples to perform adequately during prediction. When ANN is used with NIR spectral data where the number of input variables (wavelengths) is usually large (on the order of tens, hundreds, or even thousands) and the number of training samples is limited, it is practical to reduce the number of dimensions of the input space. The optimal number of new inputs and number of neurons
in hidden layer(s), as in case with PLS, is found by minimizing cross-validation standard error.

For more details on the ANN method refer to Haykin,\textsuperscript{16} Cherkassky and Mulier,\textsuperscript{17} Borggaard,\textsuperscript{18} and Næs \textit{et al.}\textsuperscript{19}

\textit{Support Vector Machines (SVM)}

The SVM method is based on principles of statistical learning theory developed by Vapnik\textsuperscript{20} and was intended for solving classification problems. Later, this technique was adapted for linear and nonlinear function estimation.\textsuperscript{21}

In the SVM regression approach, data from original input space is transformed using a mapping function $\phi(x)$ into a high dimensional feature space where linear regression is performed. This problem is formulated as constrained quadratic optimization in high-dimensional space. The solution of this problem using the Least Squares SVM regression (LS-SVM) algorithm implemented by Suykens \textit{et al.}\textsuperscript{22} is given by the model

\begin{equation}
\hat{y} = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b, \quad (3.4)
\end{equation}

where vector $x$ represents new sample, $x_k$ is $k$th training sample, $\alpha_k$ is Lagrangian multiplier for $k$th training sample, $b$ is bias term, $N$ is number of training samples, $K(x, x_k)$ is a kernel function defined as

\begin{equation}
K(x, x_k) = \phi(x) \cdot \phi(x_k), \quad (3.5)
\end{equation}
In this way, SVM model contains information about relevance of each training sample for calculation of $\hat{y}$ and makes its predictions based on relative comparison of new (unknown) sample spectra to the spectra of $k$ training samples. SVM training is computationally intensive if $k$ is large.

More information on SVM may be found in Vapnik et al., Suykens et al., Smola and Scholkopf, Cherkassky and Mulier, and Cogdill and Dardenne.

**Calibration Procedures: Application**

**PLS**

PLS_Toolbox 3.0 (Eigenvector Research Inc., www.eigenvector.com) for MATLAB (The MathWorks Inc., www.mathworks.com) was used for PLS modeling. The number of latent variables was selected using 5-block cross-validation on the training set.

**ANN**

MATLAB/Neural Network Toolbox (The MathWorks Inc., www.mathworks.com) was used for development of ANN calibration models. Feedforward backpropagation networks were trained on 80% of the calibration samples available for each fatty acid. The other 20% of the calibration samples were utilized as an early stopping set to prevent over-fitting during training process. Input dimensionality was reduced from 100 to 25 by taking every fourth wavelength of the NIR spectra. (Note: preliminary study demonstrated that this simple resampling resulted in ANN calibrations superior to those developed on data compressed with principal component analysis.) The best number of neurons in one hidden layer was determined by 5-block cross-validation on the training set. A tangent sigmoid function and linear function were used as activation functions of hidden layer neurons and an output neuron, respectively.
**LS-SVM**

LS-SVMLab1.5 toolbox for MATLAB developed by Suykens et al.\(^{22}\) was utilized for this part of the experiment. Radial basis function (RBF)

\[
K(x, x_k) = \exp(-||x-x_k||^2/\sigma^2),
\]

where \( \sigma^2 \) is the RBF bandwidth, was used as a kernel function. The best pair of complexity regularization parameter and RBF bandwidth for every fatty acid calibration model was determined by 5-block cross-validation on the training set.

**Comparison of Calibration Methods**

Parts of the data sets (25% of total number of samples) described in Table 3.1 that were not used for calibration were applied to corresponding calibration models and following parameters characterizing models’ predictive ability were computed: coefficient of determination \( r^2 \), standard error of prediction corrected for bias SEP, bias or mean difference between NIR-predicted and reference concentrations \( d \), and relative predictive determinant \( RPD \). Definitions of these parameters can be found in Williams and Norris.\(^{25}\)

To establish significance of calibration method factor, \( RPD \) coefficients, which characterize overall predictive ability of calibrations, were compared using analysis of variance. In addition, the effect of calibration set size on performance of the regression methods was studied. In this part of the experiment, calibration sets for saturated and linoleic fatty acids were reduced to smaller data sets that ranged from 50% to 5% of the original size. Models for the two constituents, three regression methods, and all reduced calibration sets were developed and then tested using constant validation sets.
RESULTS AND DISCUSSION

Overall Results

Validation results of calibration models developed with PLS, ANN, and LS-SVM regression methods for six fatty acids in whole soybeans are shown in Table 3.2. In terms of determination coefficients, predictive ability of models ranged from 0.49 – 0.68 (stearic acid) and 0.91 – 0.94 (total saturates). Based on guidelines for interpretation of $r^2$ coefficients outlined by Williams and Norris,$^25$ NIR calibration equations for saturated fatty acids were usable for quality assurance applications, while those for palmitic acid were “usable with caution for most applications, including research”. Models for the other four fatty acids had lower predictive power, however, they could still be utilized for sample screening, which is an important routine task in seed breeding programs. It is important to note, that predictive ability of NIR calibration equations was not dependent on correlation between total oil and fatty acid concentration (refer to Table 3.1 for determination coefficients describing relationship of individual fatty acids with total oil content). For example, validation $r^2$ values of NIR calibration models (PLS, ANN, and LS-SVM) for saturates were high (0.91 – 0.94), while correlation between oil content and saturates was practically zero (0.003). This suggests that NIR spectroscopy and calibration methods used in this study could be utilizing information from individual fatty acid absorption bands, not from wider total fat absorption bands.

Most of the variation (79 %) among $r^2$ coefficients of the fatty acid prediction equations could be explained by standard deviation of reference data in calibration sets (graph is not shown). Thus, introduction of larger number of samples with extremely low and high values...
Table 3.2. Validation statistics of PLS, ANN, and LS-SVM calibration equations developed for estimation of fatty acid composition in whole soybeans.

<table>
<thead>
<tr>
<th>Fatty acid</th>
<th>Test statistic</th>
<th>PLS model</th>
<th>ANN model</th>
<th>LS-SVM model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturates</td>
<td>$r^2$</td>
<td>0.91</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>(C16:0+C18:0)</td>
<td>$SEP$</td>
<td>2.23</td>
<td>2.13</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.01</td>
<td>-0.31</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>$RPD$</td>
<td>3.3</td>
<td>3.5</td>
<td>4.2</td>
</tr>
<tr>
<td>Palmitic</td>
<td>$r^2$</td>
<td>0.80</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>(C16:0)</td>
<td>$SEP$</td>
<td>3.16</td>
<td>2.79</td>
<td>2.94</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.33</td>
<td>-0.44</td>
<td>-0.39</td>
</tr>
<tr>
<td></td>
<td>$RPD$</td>
<td>2.2</td>
<td>2.5</td>
<td>2.4</td>
</tr>
<tr>
<td>Stearic</td>
<td>$r^2$</td>
<td>0.49</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>(C18:0)</td>
<td>$SEP$</td>
<td>0.97</td>
<td>0.82</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.08</td>
<td>-0.01</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td>$RPD$</td>
<td>1.4</td>
<td>1.7</td>
<td>1.8</td>
</tr>
<tr>
<td>Oleic</td>
<td>$r^2$</td>
<td>0.76</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>(C18:1)</td>
<td>$SEP$</td>
<td>4.27</td>
<td>3.93</td>
<td>3.88</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.63</td>
<td>-0.63</td>
<td>-0.28</td>
</tr>
<tr>
<td></td>
<td>$RPD$</td>
<td>2.1</td>
<td>2.2</td>
<td>2.3</td>
</tr>
<tr>
<td>Linoleic</td>
<td>$r^2$</td>
<td>0.73</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>(C18:2)</td>
<td>$SEP$</td>
<td>3.77</td>
<td>3.67</td>
<td>3.56</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.18</td>
<td>0.19</td>
<td>2.27</td>
</tr>
<tr>
<td></td>
<td>$RPD$</td>
<td>1.9</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Linolenic</td>
<td>$r^2$</td>
<td>0.67</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>(C18:3)</td>
<td>$SEP$</td>
<td>1.74</td>
<td>1.56</td>
<td>1.53</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.09</td>
<td>-0.02</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td>$RPD$</td>
<td>1.7</td>
<td>1.9</td>
<td>2.0</td>
</tr>
</tbody>
</table>
of stearic and linolenic acids into corresponding calibration data sets may improve predictive ability of NIR spectroscopy for these constituents.

A previous study by Pazdernik et al.\textsuperscript{13} on the applicability of NIR spectroscopy for determination of fatty acid composition in soybeans resulted in validation $r^2$ values of 0.38 – 0.71 for models developed on ground-seed samples and 0.18 – 0.56 for those of the whole-seed samples. The $r^2$ coefficients of our experiment were higher than both sets of results reported by Pazdernik et al., which suggests that satisfactory accuracy of NIR predictions may be achieved without grinding the seed samples.

**Comparison of Calibration Methods**

Validation \textit{RPD} values for eighteen calibration models are illustrated in Figure 3.2. Visual analysis of the bar chart suggested superior performance of nonlinear regression methods. In order to confirm this, ANOVA modeling of the form

$$RPD = FA + M + \text{Error}, \quad (3.7)$$

where \( FA \) is fatty acid factor, \( M \) is calibration method factor, was performed. (Note: as could be seen in Figure 3.2, \( RPDs \) for palmitic acid calibrations had a method priority pattern different from the rest of fatty acid groups, suggesting possible \( FA \times M \) interaction; nonetheless, this factor was ignored since inclusion it into ANOVA model would drive degrees of freedom for \( \text{Error} \) to zero.) Also, mean values of \( M \) factor were compared using Tukey test at \( \alpha = 0.05 \). The results of statistical analysis demonstrated that both \( FA \) and \( M \) factors had significant effect on \( RPD \) coefficient (\( p < 0.0001 \) and \( p = 0.017 \), respectively). Mean \( RPD \) of LS-SVM equations (mean = 2.43, standard error = 0.0638), was significantly
better than that of PLS equations (mean = 2.11, standard error = 0.0638). However, mean $RDP$ of ANN calibrations (mean = 2.30, standard error = 0.0638) was not significantly different from the other two methods.

To further compare performance of regression methods, calibration models for saturated and linoleic fatty acids were developed using reduced calibration data sets as described in the last paragraph of Materials and Methods section. The results – $RDP$ coefficient as a function of calibration set size – are shown in Figure 3.3. As expected, predictive ability of calibration equations dropped as the number of calibration (training) samples decreased, regardless of the regression method or type of predicted constituent. However, a rate of performance degradation was dependent on calibration method and constituent. Out of three types of
Figure 3.3. Validation RPD of PLS, LS-SVM, and PLS equations as a function of calibration set size for saturated and linoleic fatty acids in soybeans.

equations, ANN models demonstrated the highest rate of performance degradation for both saturated and linoleic fatty acids. Judging from their low RPD values, these equations became unusable when they were developed using calibration sets of fewer than 90 samples for saturates and 150 samples for linoleic acids. LS-SVM calibrations for linoleic acid demonstrated behavior similar to their PLS counterparts; however, LS-SVM equations for saturates displayed the best tolerance to reduction of number of calibration samples. Another important observation about saturates models was that variation of RPDs of LS-SVM equations developed on calibration sets of fewer than 100 – 150 samples was higher than that of PLS models. This suggests a strong sensitivity of this nonlinear regression method to outliers and/or unusual samples in small calibration sets.
CONCLUSIONS

NIR calibration models for determination of fatty acid concentrations in whole soybeans were developed using PLS, ANN, and LS-SVM regression methods. Validation results demonstrated that (1) equations for total saturates had the highest predictive ability and were usable for quality assurance applications, (2) palmitic acid models were usable for certain research applications, and (3) equations for stearic, oleic, linoleic, and linolenic acids could be used for sample screening. The results also showed that LS-SVM models produced significantly more accurate predictions than those developed with PLS regression. ANN calibrations were not significantly different from the other two methods. Reduction of number of calibration samples reduced predictive ability of all types of NIR equations, however the rate of performance degradation of LS-SVM models was the lowest.

REFERENCES


CHAPTER 4. DIMENSIONALITY REDUCTION OF NIR SPECTRAL DATA USING GLOBAL AND LOCAL IMPLEMENTATIONS OF PCA FOR NEURAL NETWORK CALIBRATIONS

A paper to be submitted to *Journal of Near Infrared Spectroscopy*

Igor V. Kovalenko\(^a\), Glen R. Rippke\(^a\), Charles R. Hurburgh\(^ab\)

\(^a\) Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA
\(^b\) Department of Food Science and Human Nutrition, Iowa State University, Ames, IA

**ABSTRACT**

Artificial neural network (ANN) learning algorithm has established itself as a strong alternative to traditional linear calibration methods used in near-infrared (NIR) spectroscopy. One of the limitations of this method comes from the fact that its generalization capacity could be effectively employed only when the ratio of available training samples to a number of neuron interconnection weights and biases (unknown regression parameters) is sufficiently large. Traditionally, this ratio is increased by reducing dimensionality of ANN input space by compressing $X$ data using principal component analysis (PCA). However, several other dimensionality reduction methods have been shown to outperform it. An attractive data compression method that combines two multivariate data analysis techniques, namely clustering and PCA, has been described in the literature. This approach, known as local PCA, overcomes PCA's global linearity by performing dimensionality reduction task in two steps: division of the data space into clusters and local compression of each cluster using PCA. Therefore, the objective of this study was to compare applicability of global and local implementations of PCA compression to NIR calibration problems solved with ANN
regression. In this experiment, two data sets were used for development of control (based on PCA) and experimental (based on local PCA) ANN calibrations. Predictive ability of two types of models was compared for both data sets. The results demonstrated that local PCA could significantly outperform traditional global PCA compression. However, the choice of preferred dimensionality reduction method was case-dependent. In addition, the study showed that performance of local PCA-based calibrations degraded rapidly as compression rate increased, while global PCA allowed achieving higher compression at minimal cost of prediction accuracy.

INTRODUCTION

Artificial neural network (ANN) learning algorithm, specifically its feedforward backpropagation implementation, has established itself as a strong alternative to traditional linear calibration methods used in near-infrared (NIR) spectroscopy. Numerous applications have demonstrated its superiority to techniques that are based on principal component analysis (PCA) and partial least squares (PLS) for solving both regression and classification problems.\textsuperscript{1-6}

The attractiveness of the ANN method, in particular for regression applications, comes from the fact that it is a universal function approximation technique. It performs better than the linear methods when there is a pronounced nonlinearity in the relationship between spectral (\(X\)) and reference data (\(y\)), while it can perform as well as the linear methods when the data are linear. However, as any other calibration method, ANN modeling has its shortcomings: (1) it does not extrapolate well (which is characteristic to nonlinear calibration
methods in general), therefore constituent concentration of the future samples must be within
the concentration range of calibration samples; (2) it is a nondeterministic method in the
sense that repeated trainings on the same data set will produce slightly different solutions; (3)
the ratio of available training samples to a number of neuron interconnection weights and
biases (unknown regression parameters) should be sufficiently large (on the order of tens or
hundreds) to effectively employ ANN's generalization capacity. The present paper is focused
on this last limitation.

The simplest way to increase the ratio of training samples to a number of regression
parameters is to increase the number of training samples. However, with NIR data where the
number of ANN inputs (wavelengths) alone can reach hundreds or even thousands, this task
may not be feasible due to economic considerations or to the lack of samples. Another way to
increase the ratio is to reduce the dimensionality of ANN input space by compressing \( X \) data.
PCA technique is often employed for this purpose.\(^7\)-\(^10\)

PCA is a classical linear dimensionality reduction method, which transforms original
correlated variables (wavelengths in our case) into a set of new uncorrelated variables or
principal components (PCs). The main idea is to first determine orthogonal directions of
highest variance of uncompressed data and then project the data into a new coordinate
system. Resultant PCs are linear combinations of original variables, which are ordered in
such a manner that several first ones capture most of the variation of the original data and the
last ones retain supposedly unimportant nonlinearity and noise. Therefore, only several first
PCs should be able to approximate original high-dimensional data, and the higher the
collinearity in the data, the fewer PCs are needed. However, Yeung and Ruzzo\(^11\)
demonstrated on a classification problem with microarray gene expression data that first \( n \)
PCs do not necessarily contain the most important information for problem solution, and there may exist subsets of \( n \) disjointed PCs that can produce better results.

Despite PCA’s popularity and the fact that it is the optimal transformation method when the relationship between variables of original data is for the most part linear, several other dimensionality reduction methods outperform traditional PCA. These include nonlinear variants of PCA for chemical engineering and structural dynamics applications, wavelet-based technique in handwritten numerals recognition, and methods based on data clustering in genetic studies.

An attractive data compression method that combines two multivariate data analysis techniques, namely clustering and PCA, has been described by Archer and Leen and Kerschen and Golinval. This approach, also known as local PCA, overcomes PCA’s global linearity by performing dimensionality reduction task in two steps: partitioning of variables into clusters and local compression of each cluster using PCA. To our best knowledge, this technique has not been applied to NIR spectral data, therefore, the objective of this study was to analyze its applicability to NIR calibration problems solved with ANN regression and compare it to traditional PCA data compression.

MATERIALS AND METHODS

Data and Instrumentation

Data compression methods, PCA and local PCA, were tested on two sets of NIR spectral data. Scans of whole soybean samples were collected at room temperature (22 °C) with two spectrometers: Perten DA 7200 (Perten Instruments AB, www.perten.com), and ASD
LabSpec Pro (Analytical Spectral Devices Inc., www.asdi.com). Instruments specifications are provided in Table 4.1. For reference, typical NIR spectra of a whole soybean sample are shown in Figure 4.1. (Note: visible region of ASD LabSpec Pro spectra was not employed in calibration development and, therefore, is not shown in the figure.) Lysine concentrations of soybean samples determined at Experiment Station Chemical Laboratories of University of Missouri by official method AOAC 982.30 E (a, b, c) Ch. 45.3.05\(^{18}\) were used as reference data for development and validation of ANN calibration models. (Note: our previous work showed that lysine NIR predictions are essentially protein predictions.) Lysine concentrations in % of total weight on dry basis ranged from 2.15 % to 3.28 % and were approximately normally distributed (mean = 2.69 %, standard deviation = 0.2). 474 soybean samples from 1997 – 2001 crop and 130 samples from 2002 crop were used as calibration and validation sets, respectively.

Table 4.1. Specifications of two NIR spectrometers (Perten DA 7200 and ASD LabSpec Pro) used in this study.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Spectrometer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Perten DA 7200</td>
</tr>
<tr>
<td>Technology</td>
<td>One InGaAs photodiode array detector</td>
</tr>
<tr>
<td>Mode</td>
<td>Reflectance</td>
</tr>
<tr>
<td>Spectral range</td>
<td>950 – 1650 nm</td>
</tr>
<tr>
<td>Spectral resolution</td>
<td>12 nm</td>
</tr>
<tr>
<td>Sampling interval</td>
<td>5 nm</td>
</tr>
<tr>
<td>Number of data points per spectrum</td>
<td>141</td>
</tr>
</tbody>
</table>

\(^{a}\) Only NIR region (700 – 2500 nm) of ASD LabSpec Pro scans was used in this study.

\(^{b}\) 1801 data points on the range of 700 – 2500 nm.
Procedure

Data preprocessing, development of calibration models, their validation, and statistical analysis of the results were done in MATLAB programming environment using Neural Network, Fuzzy Logic, and Statistics toolboxes (The MathWorks Inc., www.mathworks.com). Experimental procedure was divided into following steps:

1) Preprocess calibration and validation spectra using multiplicative scatter correction

2) Compress wavelengths space of calibration data using PCA (control treatment) or local PCA (experimental treatment) into a new c-dimensional space ($c = 10, 12, 13$, ...)
15, 17, 20, 25, 34, 50, 100):

**Control treatment**

a) Normalize spectral data (zero mean and unity standard deviation)

b) Perform PCA on X-variables (wavelengths) and use first c PCs as ANN inputs

or

**Experimental treatment**

a) Partition wavelengths into c groups using fuzzy c-means (FCM) clustering algorithm. FCM is an unsupervised classification algorithm that allows each wavelength to belong to all clusters with a different degree of membership ranging from zero to one. Grades of membership are calculated in iterative optimization process during which the following objective function is minimized with respect to a membership matrix $\mathbf{U}$ and a matrix $\mathbf{V}$ of c cluster centers:

$$J(U,V) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2,$$

where $u_{ij}$ is a degree of membership for $i$th cluster center and $j$th data point $(0 \leq u \leq 1)$, $m$ is a weighting exponent $(1 \leq m < \infty)$; and $d_{ij}$ is the Euclidian distance between $i$th cluster center and $j$th data point. Note: for a detailed description of FCM clustering, refer to Dunn, Bezdek, and Theodoridis and Koutroumbas.

b) Normalize spectral data
c) Perform PCA for each cluster and replace each group of wavelengths with its first
PC for ANN training

3) Perform training of the feedforward backpropagation ANN as follows:

a) Create an untrained network with \( c \) inputs, two neurons (tangent sigmoid transfer
functions) in one hidden layer, and one output neuron (linear transfer function).
(Note: preliminary studies showed that increasing the number of hidden layers or
number of neurons in a hidden layer did not improve ANN’s predictive ability in
this experiment)

b) Train network on 80 % of the available training samples and use the other 20 % as
an early stopping set to prevent over-fitting

c) Repeat steps 3a and 3b thirty times

d) Out of thirty trained networks, select one whose mean square error (MSE) of
cross-validation is the closest to the median MSE. (Note: this step reduced
variability of performance parameters among replicated calibration models.)

4) Perform appropriate transformations on validation spectra using parameters obtained
during transformation of calibration set (PCA transformation matrices, clustering
information, means, and standard deviations)

5) Test predictive ability of ANN calibration model from step 3d with validation data.
Compute and record coefficient of determination \( r^2 \), bias or mean difference between
NIR-predicted and reference concentrations \( d \), standard error of prediction corrected
for bias \( SEP \), and relative predictive determinant \( RPD \). (Note: \( RPD \) is a ratio of
standard deviation of reference data in the validation set to \( SEP \)). Definitions of these
parameters can be found in Williams and Norris\(^\text{22}\)
Because of nondeterministic nature of backpropagation training algorithm and, consequently, expected variability in validation results of repeatedly trained calibration models, the above procedure was repeated five times for each data set, compression level, and compression method combination, and values of $r^2$, $d$, $SEP$, and $RPD$ were recorded.

**RESULTS AND DISCUSSION**

**Data Set 1: Perten DA 7200 Spectra and Lysine Concentrations**

The effect of dimensionality reduction method and rate of compression on predictive ability of ANN calibration model for the first data set is illustrated by Figure 4.2. On this

![Graph showing RPD coefficient as a function of number of new variables obtained through compression of spectral data using PCA and local PCA methods.]

Figure 4.2. Data set 1: $RPD$ coefficient of ANN calibration models as a function of number of new variables obtained as a result of compression of spectral data with PCA and local PCA methods.
graph, $RPD$ coefficient is plotted as a function of number of new variables that were used as spectral data. Each number of new variables (x-axis values of 10, 12, 13, 15, 17, 20, 25, 34, 50, 100) corresponds to five $RPD$ replicates for each compression method (refer to the last paragraph of Materials and Methods). $RPD$ coefficients ranged from 1.89 to 2.92 and from 1.80 to 2.76 for PCA and local PCA, respectively. On the range of 34 to 100 new inputs, both dimensionality reduction methods demonstrated comparable performance, however, as compression rate increased, performance of calibrations based on local PCA degraded much faster. ANNs trained and tested on PCA-treated data tolerated higher compression rates better than those based on local PCA compression. Two observations supported this conclusion: (1) $RPDs$ of PCA-based models with 10, 12, 13, 15, 17 inputs had smaller variations than their local PCA counterparts, and (2) control treatment models demonstrated consistent performance on the range of 10 – 17 inputs. The latter observation suggests that, for this data set, PCA compression could achieve higher compression rates at minimal cost of prediction accuracy.

Both control and experimental calibration models reached their peaks of prediction accuracy when original 141 dimensions of spectral data were reduced to 20 new variables, which corresponded to compression rate of 7.05. Since this level of compression was optimal and, therefore, was of the most interest, validation parameters of ANN calibration models with 20 inputs were used for further comparison of dimensionality reduction methods. Table 4.2 provides averages and differences of these validation results for the two methods. The differences between values of $r^2$ (1.1 %), $SEP$ (-6.2 %), and $RPD$ (6.0 %) were statistically significant at a level of 0.05 demonstrating advantage of PCA compression. However,
Table 4.2. Data set 1: Validation results of control and experimental ANN calibration models with 20 inputs. (Note: reduction of dimensionality of spectral data from 141 to 20 showed to be optimal for both PCA and local PCA compression methods.)

<table>
<thead>
<tr>
<th>ANN model validation parameter</th>
<th>PCA compression&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Local PCA compression</th>
<th>Difference&lt;sup&gt;b&lt;/sup&gt;, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r^2$</td>
<td>0.88</td>
<td>0.87</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.007)</td>
<td></td>
</tr>
<tr>
<td>$d$</td>
<td>-0.026</td>
<td>0.007</td>
<td>127.8</td>
</tr>
<tr>
<td></td>
<td>(0.0045)</td>
<td>(0.0070)</td>
<td></td>
</tr>
<tr>
<td>$SEP$</td>
<td>0.074</td>
<td>0.079</td>
<td>-6.2</td>
</tr>
<tr>
<td></td>
<td>(0.0014)</td>
<td>(0.0018)</td>
<td></td>
</tr>
<tr>
<td>$RPD$</td>
<td>2.85</td>
<td>2.68</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>(0.054)</td>
<td>(0.062)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Parentheses contain standard deviation.

<sup>b</sup> All differences are significant at $\alpha = 0.05$.

Analysis of average bias values showed that control calibrations tended to overpredict lysine concentrations, while bias of experimental models was not significantly different from zero.

Even though the overall performance of local PCA compression in terms of $RPD$ was inferior to the global PCA method, the results from the first step of this procedure, wavelength clustering, are worth some discussion. Like analysis of PC loadings in PCA, analysis of clusters of similar wavelengths may provide some useful information about between-samples variability in NIR data for a particular set of calibration samples. Figure 4.3 depicts a cluster map of 20 groups of wavelengths for the first data set. On the map, the clusters were ordered on y-axis by their size, from the largest (#1) to the smallest (#20).

The map showed that cluster size ranged from 1 to 20 variables (wavelengths). Partitioning of wavelengths into large clusters such as #1, 2, and 3 (Figure 4.3) indicated that there was a large number of variables within each of these clusters that contained redundant...
Figure 4.3. Data set 1: Cluster map of 20 groups of wavelengths. Clusters are ordered on y-axis by their size: #1 is the largest, #20 is the smallest.

information. On the other hand, small clusters such as #18, 19, and 20 were composed of a very few "unique" variables that had a strong influence on calibration model.

Data Set 2: ASD LabSpec Pro Spectra and Lysine Concentrations

RPD coefficient as a function of number of new variables (ANN inputs) for the second data set is illustrated in Figure 4.4. For the considered in this study range of new variables, RPD values ranged from 1.55 to 2.22 and from 1.38 to 2.71 for PCA and local PCA, respectively. Several similarities and differences could be pointed out as the graphs in Figures 4.2 and 4.4 were compared. As with the first data set, PCA-treated calibration models demonstrated graceful performance degradation with the decrease of the number of new
variables (increase of compression rate). In addition, $RPD$s of the control models had smaller variations than their local PCA counterparts. The major change in the behavior of $RPD$ of the second data set calibrations was at the point of optimal level of compression when original 1801 dimensions of ASD LabSpec Pro spectra were reduced to 50 new variables (compression rate of 36.02). Here, control treatment models demonstrated significantly better performance (Table 4.3). It is worth mentioning that the advantage of one compression method over the other was much higher than in the case with the data set 1 (refer to percent difference columns of Tables 4.2 and 4.3). Bias values of both groups of calibrations were
Table 4.3. Data set 2: Validation results of control and experimental ANN calibration models with 20 inputs. (Note: reduction of dimensionality of spectral data from 141 to 20 showed to be optimal for both PCA and local PCA compression methods.)

<table>
<thead>
<tr>
<th>ANN model validation parameter</th>
<th>PCA compression(^a)</th>
<th>Local PCA compression</th>
<th>Difference,(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r^2)</td>
<td>0.78</td>
<td>0.85</td>
<td>-9.3</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.018)</td>
<td></td>
</tr>
<tr>
<td>(d)</td>
<td>-0.034</td>
<td>-0.036</td>
<td>-6.3</td>
</tr>
<tr>
<td></td>
<td>(0.0122)</td>
<td>(0.0042)</td>
<td></td>
</tr>
<tr>
<td>(SEP)</td>
<td>0.100</td>
<td>0.082</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>(0.0049)</td>
<td>(0.0050)</td>
<td></td>
</tr>
<tr>
<td>(RPD)</td>
<td>2.11</td>
<td>2.57</td>
<td>-22.1</td>
</tr>
<tr>
<td></td>
<td>(0.097)</td>
<td>(0.148)</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Parentheses contain standard deviation.

\(^b\) All differences, except for \(d\) value of -6.3, are significant at \(\alpha = 0.05\).

not significantly different from each other, and their negative signs indicated some overprediction of lysine concentrations.

A cluster map of 50 groups of wavelengths for the second data set spectra is shown in Figure 4.5. Cluster size ranged from 6 to 98 wavelengths. Distribution of cluster size was closer to uniform than in the case with data set 1. This indicates a more equal representation of various spectral regions by a set of new variables. Interestingly, a large number of clusters covered up to five disjointed regions of the spectrum (clusters such as #8, 9, 14, and 21), which suggests that even widely separated wavelengths carried redundant spectral information.
CONCLUSIONS

The study demonstrated that local PCA could be a viable alternative to traditional global PCA method for dimensionality reduction of NIR spectral data to use in ANN calibrations. For one spectrometer, this technique significantly outperformed traditional PCA compression. However, local PCA did not always result in ANN calibrations with an optimal predictive ability, and the choice of the dimensionality reduction method should be case-dependent. In addition, both examples used in the experiment showed that performance of
local PCA-based calibrations degraded rapidly as the compression rate increased, while
global PCA allowed achieving higher compression at minimal cost of prediction accuracy.

REFERENCES


CHAPTER 5. GENERAL CONCLUSIONS

GENERAL DISCUSSION

NIR spectroscopy is being successfully utilized for measurement of protein and oil content in grain. However, analysis of grain on a subunit level (amino and fatty acids) is a new application for this method. The results and conclusions drawn from the three research papers included in this dissertation provide some insight on the challenges related to this application.

In the first paper, calibration models for determination of amino acid concentration in whole soybeans were developed using five NIR spectrometers and three regression methods. The study resulted in models characterized by various degrees of accuracy, most of which were usable for research purposes and sample screening. Unfortunately, no sufficient correlation could be established between spectral data and concentrations of such important amino acids as cysteine and tryptophan. The variation in NIR models’ predictive ability was determined by the degree to which a certain amino acid correlated to crude protein. In other words, NIR instrumentation and calibration methods used in this study appeared to measure soybean amino acid concentrations indirectly by deriving them from predicted protein. Comparison of calibration methods demonstrated that performance of PLS and LS-SVM was significantly better than that of ANN. Comparison of spectrometers showed some advantage of monochromator-based FOSS Infratec 1241 Grain Analyzer and diode array Pertem DA 7200. (Note: the only feature that sets these two instruments apart from the other three is a lower number of data points in their spectra – refer to Table 2.1.)
The second paper focused on comparison of NIR calibration models for determination of fatty acid concentrations in whole soybeans that were developed for FOSS Infratec spectrometers using three regression methods. Validation results demonstrated that (1) equations for total saturates had the highest predictive ability and were usable for quality assurance applications, (2) palmitic acid models were usable for certain research applications, and (3) equations for stearic, oleic, linoleic, and linolenic acids could be used for sample screening. The results also showed that LS-SVM models produced significantly more accurate predictions than those developed with PLS regression. ANN calibrations were not significantly different from the other two methods. Reduction of number of calibration samples reduced predictive ability of all types of NIR equations, however the rate of performance degradation of LS-SVM models was the lowest.

Incidentally, the results from amino and fatty acid papers show the difference between the two types of NIR analysis. Unlike the amino acid measurement problem, where subunit concentrations were derived indirectly from total predicted protein, fatty acid NIR calibrations appeared to be utilizing information from individual fatty acid absorption bands, not from wider total fat absorption bands.

As far as calibration techniques are concerned, results of both studies were quite favorable for LS-SVM. Moreover, fatty acid study clearly demonstrated significant advantage of this non-linear technique over PLS regression. However, the shortcoming of LS-SVM calibration was that it became computationally intensive when the number of training samples was large.

The third study was focused more on methodology of NIR spectroscopy, rather than on application, and demonstrated how selection of a proper preprocessing method may improve
overall predictive ability of calibrations. The experiment explored issues of PCA-based
dimensionality reduction of NIR spectral data for development of ANN calibrations
predicting lysine concentration in whole soybeans. Results demonstrated that local PCA
compression technique could significantly outperform its traditional global counterpart.
However, local PCA did not always result in ANN calibrations with an optimal predictive
ability, and the choice of the dimensionality reduction method should be case-dependent. (In
particular, application of local PCA compression to spectral data from ASD LabSpec Pro in
the amino acid study could have improved predictive ability of this instrument’s ANN
models.) In both examples, performance of local PCA-based calibrations degraded rapidly as
compression rate increased, while global PCA allowed higher compression at minimal cost
of prediction accuracy.

RECOMMENDATIONS FOR FUTURE RESEARCH

Following are several recommendations to researchers that will continue studying
applicability of NIR spectroscopy for measurement of amino and fatty acid composition (or
subunits in general) of soybeans:

1) The biggest challenge that is faced in NIR measurement of amino acids in soybeans is
to break correlation between subunits and main unit concentrations. Future research
should attempt to address this issue by introducing calibration samples (possibly
artificially created) with non-typical (for this commodity) amino acid profiles
2) The applicability of FT-NIR spectrometers for amino acid measurement should be further investigated. Specifically, attention should be paid to performance of this type of instruments with spectral resolution set to values of 2, 4, and 8 cm\(^{-1}\).

3) Published literature provides evidence that grinding seed samples improves performance of NIR spectroscopy in measurement of amino and fatty acid profiles. Therefore, it may be worth repeating first two experiments presented in this dissertation – probably in somewhat reduced scale – using ground soybean samples.

4) Finally, in regards to methodology of spectral data compression for ANN calibration. Dimensionality reduction methods, such as local PCA, PLS, and local PLS should be applied to amino and fatty acid data sets used in this study to find out whether they may improve accuracy of ANN prediction equations.
APPENDIX. SAMPLE MATLAB CODE

% MATLAB script performing crossvalidation for determination of the % optimal number of latent variables for PLS regression

% Begin script

%% Load
fa='linolenic'; % user input
datafile=[fa '.mat'];
load(datafile);
data=dataMatrix; % user input

%% Prepare inputs and targets;
trainData=data([1:4:end 2:4:end 4:4:end],:);
testData=data(3:4:end,:);
inputTrain=trainData(1:end,2:end);
targetTrain=trainData(1:end,1);
inputTest=testData(:,2:end);
targetTest = testData(:,1);

figure; hist(data(:,1),40);
title((sprintf('%s histogram',fa)));
title((sprintf('Histogram: %s (all data) ', fa)));
xlabel('Concentration, %1')
ylabel('Frequency');
clear trainData;

%% Derivative;
inputTrain=savgol(inputTrain,5,3,2);

%% Normalize inputs and targets;
[inputTrainNorm,meanInputTrainNorm,stdInputTrainNorm]=auto(inputTrain);
[targetTrainNorm,meanTargetTrainNorm,stdTargetTrainNorm]=...
=auto(targetTrain);

%% Crossval for pls;
disp(sprintf('Constituent: %s',fa));
[press,cumpress,rmsecv,rmsec,cvpred]=...
crossval(inputTrainNorm,targetTrainNorm,'sim',{'con' 5},25);
title((sprintf('PLS crossvalidation: %s',fa)));

% End script

% MATLAB script performing PLS regression and model validation

% Begin script
%% Load
fa='linolenic'; % user input
nLV=16; % user input
datafile=[fa '.mat'];
load(datafile);
data=dataMatrix; % user input

%% Prepare inputs and targets;
trainData=data([1:4:end 2:4:end 4:4:end],:);
testData=data([3:4:end,:]);

inputTrain=trainData(1:end,2:end);
targetTrain=trainData(1:end,1);

inputTest=testData(:,2:end);
targetTest=testData(:,1);

clear trainData testData;

%% Derivative;
inputTrain=savgol(inputTrain,5,3,2);
inputTest=savgol(inputTest,5,3,2);

%% Normalize inputs and targets;
[inputTrainNorm,meanInputTrainNorm,stdInputTrainNorm]=auto(inputTrain);
inputTestNorm=scale(inputTest,meanInputTrainNorm,stdInputTrainNorm);
[targetTrainNorm,meanTargetTrainNorm,stdTargetTrainNorm]=
auto(targetTrain);

%% PLS;
plsOptions=pls ('options ');
plsOptions.display='off';
plsOptions.plots='none';

plsModel=pls(inputTrainNorm,targetTrainNorm,nLV,plsOptions);

plspredictedNorm=pls(inputTestNorm,plsModel,plsOptions);
plspredicted=rescale(plspredictedNorm.pred{2},
meanTargetTrainNorm,StdTargetTrainNorm);

[plsslope,plsintercept,plsl]=postreg(plspredicted',targetTest');
plsbias = sum(targetTest-plspredicted)./length(targetTest-... plspredicted);
plssep = std(plspredicted-plspredicted);
plsrpd = std(targetTest)/plssep;

PLS.model=plsModel;
PLS.constituent=fa;
PLS.numOfLVs=nLV;
PLS.residuals=targetTest-plspredicted;
PLS.r=plsr;
PLS.r^2=plsr^2;
PLS.SEP=plssep;
PLS.slope=plsslope;
PLS.intercept=plsintercept;
PLS.bias=plsBias;
PLS.RPD=plsRPD;

close all;

%% Display results
disp(sprintf('
'));
disp(sprintf('Constituent: %s',fa));
disp('==========================================');

plsMod=PLS;

figure; plot(plsMod.residuals,'bx');
title((sprintf('Residuals of PLS model: %s',fa)));
xlabel('Sample (from min to max ref. concentration)');
ylabel('Prediction error');
hold on;
plot([1:length(plsMod.residuals)],0,'r-');
hold off;

plsPredictedNorm=pls(inputTestNorm,PLS.model,plsOptions);
plsPredicted=rescale(plsPredictedNorm.pred{2},meanTargetTrainNorm,...
stdTargetTrainNorm);
figure; postreg(plsPredicted',targetTest');
title((sprintf('PLS model: %s',fa)));
xlabel('Actual');
ylabel('Predicted');

disp('==========================================');
disp(sprintf('Training set size: %g',length(targetTrain)));

% End script

% MATLAB script performing LS-SVM regression and model validation

% Begin script

%% Load;
fa='linolenic'; % user input
gamSig2Range=[100 10; 1000000 100000]; % user input
datafile=[fa '.mat'];
load(datafile);
data=dataU; % user input

%% Prepare inputs and targets;
trainData=data([1:4:end 2:4:end 4:4:end,:]);
testData=data(3:4:end,:);

inputTrain=trainData(1:end,2:end);
targetTrain=trainData(1:end,1);
inputTest=testData(:,2:end);
targetTest=testData(:,1);
clear data trainData testData;

%% Derivative;
inputTrain=savgol(inputTrain,5,3,2);
inputTest=savgol(inputTest,5,3,2);

%% Normalize inputs and targets;
[inputTrainNorm,meanInputTrainNorm,stdInputTrainNorm]=auto(inputTrain);
[inputTestNorm=scale(inputTest,meanInputTrainNorm,stdInputTrainNorm);

[targetTrainNorm,meanTargetTrainNorm,StdTargetTrainNorm]=...
  auto(targetTrain);

%% LS-SVM regression
lssvmModel=initlssvm(inputTrainNorm,targetTrain,'f',1,0.1,...
    'RBF_kernel','original');
lssvmModel=tunelssvm(lssvmModel,gamSig2Range,'gridsearch',{},...
    'crossvalidate',{inputTrainNorm,targetTrain,5,'mse',...
    'mean','original'});
lssvmModel=trainlssvm(lssvmModel);
lssvmPred=simlssvm(lssvmModel,inputTestNorm);

figure;
[lssvmSlope,lssvmIntercept,lssvmR]=postreg(lssvmPred',targetTest');
title((sprintf('%s, LSSVM model',fa)));
xlabel('Actual');
ylabel('Predicted');

lssvmBias = sum(targetTest-lssvmPred)./length(targetTest-lssvmPred);
lssvmSEP = std(targetTest-lssvmPred);
lssvmRPD = std(targetTest)/lssvmSEP;
LSSVM.mode1=lssvmModel;
LSSVM.constituent=fa;
LSSVM.residuals=targetTest-lssvmPred;
LSSVM.r=lssvmR;
LSSVM.rSq=lssvmR^2;
LSSVM.SEP=lssvmSEP;
LSSVM.slope=lssvmSlope;
LSSVM.intercept=lssvmIntercept;
LSSVM.bias=lssvmBias;
LSSVM.RPD=lssvmRPD;
LSSVM.testData.pedlssvmPred;
LSSVM.testData.actual=targetTest;

disp(sprintf('
'));
disp(sprintf('Constituent: %s',fa));
disp('-----------------------------------------------');

lssvmMod=LSSVM
% figure; postreg(LSSVM.testData.ped,LSSVM.testData.actual');
% title((sprintf('%s, LSSVM model',LSSVM.constituent)));
```matlab
figure; plot(LSSVM.residuals,'bx');
title((sprintf('Residuals of LS-SVM model: %s',fa)));
xlabel('Sample (from min to max ref. concentration)');
ylabel('Prediction error');
hold on;
plot([1:length(LSSVM.residuals)],0,'r-');
hold off;

disp('==========================================');
disp(sprintf('Training set size: %g',length(targetTrain)));

% MATLAB script performing ANN regression and model validation
% The script uses custom ANN training function annfatrain.m

% Begin script
fa='linolenic'; % user input
numOfNeurons=3; % user input

for count=1:20
    [theNet]=annfatrain(fa,numOfNeurons);
    allNets(count)=theNet;
end

close all;
diff=(allNets.RPD)-median(allNets.RPD).
index=find(diff==min(diff-median(allNets.RPD).
index=index(1);
finalNet=allNets(index)

disp(sprintf('
'));
allRPDs=allNets.RPD

pred=sim(finalNet.net » finalNet.testData.inputTestNorm);
figure; postreg(pred,finalNet.testData.targetTest);
title((sprintf('ANN model: %s',fa)));
xlabel('Actual');
ylabel('Predicted');

figure; plot(finalNet.residuals,'bx');
title((sprintf('Residuals of ANN model: %s',fa)));
xlabel('Sample (from min to max ref. concentration)');
ylabel('Prediction error');
hold on;
plot([1:length(finalNet.residuals)],0,'r-');
hold off;

% End script
```
% MATLAB ANN training function annfatrain.m used in the script above

% Begin function
function [finalNet]=annfatrain(fa,numOfNeurons);
%ANNPATRAIN function
% Example
%
% [finalNet]=annfatrain('linolenic',3);

%% Read all train and test data;
datafile=[fa '.mat'];
load(datafile);
data=dataMatrix;

allTrainData=data([1:4:end 2:4:end 4:4:end],:);
temp=allTrainData(:,1);
allTrainData(:,1)=[];
allTrainData=[allTrainData,temp];
allTrainData=allTrainData';

allTrainData=allTrainData(:,1:end);

allTestData=data(3:4:end,:);
temp=allTestData(:,1);
allTestData(:,1)=[];
allTestData=[allTestData,temp];
allTestData=allTestData';

%% Prepare inputs and targets;
inputTrain=allTrainData(1:4:end-1,[1:10:end,3:10:end,4:10:end,...
6:10:end,8:10:end,10:10:end]);
targetTrain=allTrainData(end,[1:10:end,3:10:end,4:10:end,...
6:10:end,8:10:end,10:10:end]);
inputVal=allTrainData(1:4:end-1,[2:10:end 5:10:end 7:10:end 9:10:end]);
targetVal=allTrainData(end,[2:10:end 5:10:end 7:10:end 9:10:end]);
inputTest=allTestData(1:4:end-1,:);
targetTest=allTestData(end,:);
clear allTrainData allTestData ;

%% Derivative;
inputTrain=inputTrain';
inputVal=inputVal';
inputTest=inputTest';

inputTrain=sav Gol(inputTrain,5,3,2);
inputVal=savGol(inputVal,5,3,2);
inputTest=savGol(inputTest,5,3,2);

inputTrain=inputTrain';
inputVal=inputVal';
inputTest=inputTest';
% Normalize inputs;
[inputTrainNorm,meanInputTrainNorm,stdInputTrainNorm]=
    prestd(inputTrain);
inputValNorm = trastd(inputVal,meanInputTrainNorm,stdInputTrainNorm);
inputTestNorm = trastd(inputTest,meanInputTrainNorm,stdInputTrainNorm);
% Validation data and initialization
val.P=inputValNorm;
val.T=targetVal;
valMse=1000;
numOfInputs=size(inputTest);
numOfInputs=numOfInputs(1);
% ANN training;
disp(sprintf('Reinitialization: %g',reinit));
for reinit=1:10;
    disp(sprintf('Neurons in a hidden layer: %g',numOfNeurons));
    disp(sprintf('Number of inputs: %g',numOfInputs));
    disp(sprintf('Constituent: %s',fa));
    net=newff(minmax(inputTrainNorm),[numOfNeurons 1],
        {'tansig' 'purelin'},'trainlm');
    [net,trRec]=train(net,inputTrainNorm,targetTrain,[],[],val);
    predictedFromTest = sim(net,inputTestNorm);
    [fSlope,fIntercept,fR] = postreg(predictedFromTest,targetTest);
    fBias = sum(targetTest-predictedFromTest)/length(targetTest-
        predictedFromTest);
    fSEP = std(targetTest-predictedFromTest);
    fRPD = std(targetTest)/fSEP;
    if (trRec.vperf(end)<valMse)
        valMse=trRec.vperf(end)
        finalNet.net=net;
        finalNet. constituents =fa;
        finalNet.residuals=targetTest-predictedFromTest;
        finalNet.testData.inputTestNorm=inputTestNorm;
        finalNet.testData.targetTest=targetTest;
        finalNet.numOfInputs=numOfInputs;
        finalNet.numOfHLNeurons=numOfNeurons;
        finalNet.r=fR;
        finalNet.rSq=fR^2;
        finalNet.SEP=fSEP;
        finalNet.slope=fSlope;
        finalNet.intercept=fIntercept;
        finalNet.bias=fBias;
        finalNet.RPD=fRPD;
        finalNet.means=meanInputTrainNorm;
        finalNet.stds=stdInputTrainNorm;
    end;
end;
disp(sprintf('n));
disp(sprintf('Constituent: %s',fa));
disp('================================================:');
finalNet;
disp(sprintf('n));
disp('================================================:');
finalNet=finalNet;
close all;
pred=sim(finalNet.net,finalNet.testData.inputTestNorm);
figure; postreg(pred,finalNet.testData.targetTest);

% End function

% MATLAB script that modifies original calibration data set to the one
% with approximately uniform distribution of reference concentration
% values (Input data matrix should be sorted by reference concentration)

% Begin script

% Samples (rows) in the input data matrix should be sorted by reference
% concentration from lowest to highest; first column - concentration

%% Load
load linolenic; % user input
data=dataMatrix; % user input

%% Plot original distr. & define size of bin
num=200; % number of bins; user input
conc=data(:,1);
figure; hist(conc,40);
title('Original distribution');
minc=min(conc);
maxc=max(conc);
bin=(maxc-minc)/num;

%% Initialize bins
for j=1:num
    bins(j).samps=[];
end;

%% Assign samples to bins
for i=1:length(conc)
    b=ceil((conc(i,:)-minc)/bin);
    if b==0
        b=1;
    end;
    bins(b).samps=[bins(b).samps; data(i,:)];
end;

%% Resample bins
clear j;
for j=1:num
    numSamps=size(bins(j).samps);
    numSamps=numSamps(1);
    if numSamps>10
        step=floor(numSamps/10);
        bins(j).samps=bins(j).samps(1:step:end,:);
    end;
end;

% New uniform data set
dataU=[];
clear j;
for j=1:num
    dataU=[dataU;bins(j).samps];
end;

% Plot new distribution
figure; hist(dataU(:,1),40);
title('Modified distribution');

% End script
I am appreciative and thankful to many people who helped me during my doctoral work. I would especially like to thank my major professor, Dr. Charles Hurburgh, for his technical support, readiness to discuss my ideas, and continuous stimulation of my analytical thinking. I wish to thank my committee members, Dr. Carl Bern, Dr. Terri Boylston, Dr. Thomas Brumm, and Dr. Philip Dixon, who helped me in building my experiments and interpreting the results. My gratitude is due to my M.S. major professor, Dr. Jenni Briggs, who set me on the path to becoming a researcher. I am also very grateful to Dr. Eric Bartlett and Dr. Julie Dickerson for introducing me into the fascinated field of pattern recognition and computer learning. I extend many thanks to all of my former and current colleagues from Grain Quality Laboratory at Iowa State University. Finally, I would like to thank my wife, Oksana, for her immense moral support during the years of my graduate studies.

This research was funded by the United Soybean Board and carried out in Grain Quality Lab at Iowa State University.