Analytica Chimica Acta 813 (2014) 1–14

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Non-linear calibration models for near infrared spectroscopy

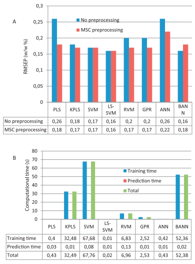
Wangdong Ni[a,b,*], Lars Nørgaard[a,b], Morten Mørup[c]

[a] FOSS Analytical A/S, Foss Allé 1, DK-3400 Hillerød, Denmark
[b] Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
[c] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Matematiktorvet, Building 321, DK-2800 Kgs. Lyngby, Denmark

HIGHLIGHTS

- We present a comprehensive comparative study of nonlinear calibration techniques.
- The methods are connected in terms of traditional calibration by ridge regression.
- Three real-life near infrared (NIR) data sets are used for comparison of the methods.
- Different practical aspects of the methods are discussed for spectral analysis.

GRAPHICAL ABSTRACT



	PLS	KPLS	SVM	LS-SVM	RVM	GPR	ANN	BAN N
No preprocessing	0,26	0,18	0,17	0,16	0,2	0,2	0,26	0,16
MSC preprocessing	0,18	0,17	0,17	0,16	0,17	0,17	0,22	0,18

	PLS	KPLS	SVM	LS-SVM	RVM	GPR	ANN	BANN
Training time	0,4	32,48	67,68	0,01	6,83	2,52	0,42	52,36
Prediction time	0,03	0,01	0,08	0,01	0,13	0,01	0,01	0,02
Total	0,43	32,49	67,76	0,02	6,96	2,53	0,43	52,38

ARTICLE INFO

Article history:
Received 9 October 2013
Received in revised form
26 November 2013
Accepted 2 December 2013
Available online 9 December 2013

Keywords:
NIR
Chemometrics
Nonlinear calibrations
GPR
BANN
LS-SVM

ABSTRACT

Different calibration techniques are available for spectroscopic applications that show nonlinear behavior. This comprehensive comparative study presents a comparison of different nonlinear calibration techniques: kernel PLS (KPLS), support vector machines (SVM), least-squares SVM (LS-SVM), relevance vector machines (RVM), Gaussian process regression (GPR), artificial neural network (ANN), and Bayesian ANN (BANN). In this comparison, partial least squares (PLS) regression is used as a linear benchmark, while the relationship of the methods is considered in terms of traditional calibration by ridge regression (RR). The performance of the different methods is demonstrated by their practical applications using three real-life near infrared (NIR) data sets. Different aspects of the various approaches including computational time, model interpretability, potential over-fitting using the non-linear models on linear problems, robustness to small or medium sample sets, and robustness to pre-processing, are discussed. The results suggest that GPR and BANN are powerful and promising methods for handling linear as well as nonlinear systems, even when the data sets are moderately small. The LS-SVM is also attractive due to its good predictive performance for both linear and nonlinear calibrations.

1. Introduction

Near infrared spectroscopy (NIR) is a powerful analytical tool in many different fields due to its non-invasive, fast, and informative characteristics. It has become one of the most widely applied instrumental methods to determine critical attributes closely related to product quality in areas and industries as diverse as process analytical technology (PAT) [1,2], food [3], pharmaceuticals [4], agricultural [5], and petrochemical [6]. NIR data usually include hundreds or thousands of wavelengths that contain chemical, physical, and biological information of the analyzed material. Multivariate calibration methods, such as partial least squares (PLS) regression [7,8] and principal component regression (PCR) [9], relate the spectral data to specific variables, based on a linearity assumption [10] which implies that the NIR spectra are linearly related to the concentrations. While PLS and PCR have become

[*] Corresponding author at: Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark. Tel.: +45 35 33 32 39; fax: +45 35 33 32 45.
E-mail addresses: wni@life.ku.dk, antony2006ster@gmail.com (W. Ni).

0003-2670/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.aca.2013.12.002

popular due to their ease of use, fast computation, good predictive performance and easy interpretable representations – the linearity assumption is not always valid, and thus PLS or PCR may give non-optimal results when the spectra exhibit nonlinearities [11,12].

In order to improve performance, PLS and PCR can be extended to handle nonlinearity by the use of polynomial functions [13], or nonlinear kernels (KPLS) [14,15]. As a more powerful and flexible non-linear regression tool, artificial neural network (ANN) has become a popular method in many areas [16–18], and ANN has been widely applied in the chemometrics community for spectroscopic calibration since the 1990s [16,17,19]. Its broad application to spectroscopic data analysis is however still limited due to disadvantages e.g. issues with local minima solutions where initial parameter settings affect the final performance and the enlarged risk of over-fitting leading to a requirement for very large sample sizes.

An alternative method, SVM [20–22] for regression, has been applied to develop spectral calibrations. It allows modeling certain nonlinear relationships in the spectral space by introducing nonlinear kernels. Opposed to standard ANN methods it has a so-called convex objective function admitting globally optimal solutions that are most often unique [20,21]. However, computational issues arise as the model has no closed form solution but forms a quadratic programming problem and model parameters need to be optimized which generally includes a grid based cross-validation procedure [20,23]. The computational difficulty can be overcome by its variant, least squares-SVM (LS-SVM) [20,23,24] where the quadratic problem reduces to solving a set of linear equations allowing a fast and easy implementation with less parameters to estimate [23,24]. SVM and LS-SVM have been successfully applied to spectroscopic calibration development in several cases and they are increasingly regarded as an alternative to ANN [20,24]. A few methods derived from the Bayesian framework, have been developed to address the challenges in SVM-based models; Tipping [25] proposed a Bayesian method for SVM denoted relevance vector machine (RVM), which has been introduced by Hernández et al. [26] for spectroscopic calibration.

Recently, Gaussian process regression (GPR) was introduced as an alternative approach to ANN from a non-parametric Bayesian learning perspective [11]. GPR was initially developed by O'Hagan [27], and a significantly increasing interest is observed in different areas, such as machine learning [28–30] and dynamic process modeling [16,31–33]. The concept of Gaussian processes (GP) emerged from the area of neural networks. It has been shown that Bayesian neural networks (BANN) converge to a GP when an infinite number of hidden units are used and through assigning Gaussian priors over the weight space of neural networks [33–37]. In other words, BANN with infinite number of hidden neurons could be formulated as a GP which is defined by a mean and covariance function. However, as pointed by Williams [38,44], only a specific transfer function used in neural networks can be presented as a Gaussian process with a specific covariance function. BANN requires in general less training samples than ANN as it invokes prior distributions on the model parameters that are tuned to address issues of over-fitting [33]. In order to control for over-fitting in ANN [37] an 'early stopping' strategy can be employed to stop training by an independent stop set. In the BANN model parameters are estimated by Bayesian inference without using a separate stop data set or cross-validation strategy as for regular ANNs as pointed out by Neal [33] and MacKay [34]. Instead, the Bayesian approach to parameter estimation makes use of the model evidence [35] to address issues of over-fitting. A few empirical comparative studies [11,12,16,31,32] have confirmed the efficient predictive performance of GPR including its application to spectroscopic calibrations.

The nonlinear calibration techniques have been compared in several studies for spectroscopic data analysis. Chen et al. [11] compared GPR with quadratic PLS (QPLS) and ANN and demonstrated that GPR performed best for the modeling of nonlinear spectroscopic data sets. In Hernandez's study [26], RVM, ANN, SVM, and LS-SVM were compared to demonstrate that the RVM resulted in sparse solution where only about 20% or even less training objects were retained for three spectroscopic calibrations. Comparison of GPR, LS-SVM, and ANN was performed by Wang et al. [12] demonstrating good predictive performance of GPR for the modeling of three NIR data sets. A performance comparison of PLS, LS-SVM, and ANN for large NIR data sets is given in the work of Fernández Pierna et al. [24], where the LS-SVM model achieved the best performance. Direct comparison between KPLS and SVM was conducted by Czekaj et al. [39] and the results showed that KPLS and SVM can generate similar predictive performance. Balabin and Lomakina [20] presented the comparison of a polynomial PLS (poly-PLS), ANN, SVM, and LS-SVM. SVM based methods achieved a comparative accuracy in predictive performance to ANN and they were recommended for real applications due to their much higher robustness.

The main goal of the present work, compared to previous studies, is to comprehensively compare nonlinear methods for spectroscopic calibrations with PLS as the linear benchmark and ridge regression (RR) [40,41] as a theoretical connection between these calibration approaches. Properties/evaluation criteria include computational time, parametric vs. non-parametric methods, and performance across different data set sizes. The presented methods are related and highlighted in Table 1 for a compact outline of nonlinear methods. The methods to be compared are: KPLS, SVM, LS-SVM, ANN, RVM, BANN, and GPR. The methods will be compared with respect to performance on three selected publicly available NIR data sets reflecting both non-linear and linear relations between the spectra and the dependent variable. Furthermore, discussions on the computation time, model interpretability, potential over-fitting using non-linear models on linear problems, robustness to small sample sets, and pre-processing will be included when relevant.

2. Introduction of calibration models

Usually, N observations of the training set $D = \{\mathbf{x}_n, y_n\}$, $n = 1, \ldots, N$, where $\mathbf{x}_n$ has M variables or predictors, are used for spectroscopic calibration. The relation between the spectral matrix $\mathbf{X}$ and reference vector $\mathbf{y}$ is generally described by linear regression as the problem of finding the regression vector $\mathbf{b}$, as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1}$$

where $\mathbf{b} = (b_1, \ldots, b_M)^{\mathrm{T}}$ and $\mathbf{e} = (e_1, \ldots, e_N)^{\mathrm{T}}$ is the residuals. However, the linearity is not always valid and as a result some calibration problems are nonlinear. One way to address nonlinearity in the data is to extend Eq. (1) to a linear combination of the response of a set of nonlinear basis functions as follows [25,26]:

$$y_n = \sum_{j=1}^{M} \phi_j(\mathbf{x}_n) b_j + e = \phi(\mathbf{x}_n)\mathbf{b} + e, \quad \mathbf{y} = \Phi(\mathbf{X})\mathbf{b} + \mathbf{e} \tag{2}$$

where $\phi_j(\mathbf{x}_n)$ is the response of the jth basis function to input $\mathbf{x}_n$. Suppose $\phi(\mathbf{x}_n) = [\phi_1(\mathbf{x}_n), \ldots, \phi_M(\mathbf{x}_n)]$ is a row vector containing the response of all basis functions to input $\mathbf{x}_n$, $\phi_j = [\phi_j(\mathbf{x}_1), \ldots, \phi_j(\mathbf{x}_n)]^T$, is a column vector containing the response of basis function $\phi_j(\mathbf{x})$ to all training inputs, and Φ is an $N \times M$ matrix whose jth column is vector ϕ_j and whose nth row is vector $\phi(\mathbf{x}_n)$. M, herein, is just the number of basis functions, which does not necessarily equal the number of variables in the spectra or data set, M. In the following, it will be assumed that the residuals contain independent zero mean Gaussian noise with variance σ^2, $e_n \sim G(0, \sigma^2)$. As a result, derived from the idea of Bayesian learning [11,25,42], given the training