Chemometrics and Intelligent Laboratory Systems 174 (2018) 1–14

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

# Hierarchical mixture of linear regressions for multivariate spectroscopic calibration
## An application for NIR calibration

Chenhao Cui *, Tom Fearn

*Department of Statistical Science, University College London, London WC1E6BT, UK*

ARTICLE INFO

*Keywords:*
Near-infrared spectroscopy
Multivariate calibration
Partial least squares regression
Hierarchical mixture of linear regressions
Expectation maximization
Variational inference

ABSTRACT

This paper investigates the use of the hierarchical mixture of linear regressions (HMLR) and variational inference for multivariate spectroscopic calibration. The performance of HMLR is compared to the classical methods: partial least squares regression (PLSR), and PLS embedded locally weighted regression (LWR) on three different NIR datasets, including a publicly accessible one. In these tests, HMLR outperformed the other two benchmark methods. Compared to LWR, HMLR is parametric, which makes it interpretable and easy to use. In addition, HMLR provides a novel calibration scheme to build a two-tier PLS regression model automatically. This is especially useful when the investigated constituent covers a large range.

## 1. Background

Partial least squares regression (PLSR) [1] and principal component regression (PCR) [2] have been applied to NIR calibration for decades. The beauty of these methods is that noisy and mutually dependent NIR spectra are first dimensionally reduced to a few factors, which ensures the following multiple linear regression (MLR) [3] process is properly regularized. However, the limitations of these regression schemes are also obvious: both the dimension reduction and the regression are linear, which means the final derived model is a linear combination of all input variables. This means neither of them can tackle non-linear effects. Lack of non-linearity is critical especially when the target constituent covers a large range. When a single linear model is forced to fit over a wide range, significant biases often appear in the two tails.

Non-linear regression methods were introduced to save prediction models from such heavy biases. Popular examples are support vector machine (SVM) [4] and Gaussian process regression (GPR) [5]. Some recent studies have proven these methods can be used to improve prediction accuracy [6–8]. Very similar to kernel-based regression schemes, local linear regression methods such as locally weighted regression (LWR) were also proposed to give accurate and unbiased prediction across the range [9]. These methods indeed break the limitations of linear regression. However, they are not always practically applicable. First of all, to build such kinds of models, a large data set is required. The methods fail if there are not adequate training instances in the neighborhood of the unknown observation. Secondly, using these methods for out-of-sample prediction requires information of the whole training set, including the spectra (or equivalently, the gram matrix in the kernel methods) and corresponding reference values. The size of the model grows with the size of the training set (i.e., these methods are all non-parametric). Even if the storage of the system is not a limitation, computational cost for making a prediction might be an intolerable issue. Computational cost also makes it extremely difficult to implement these methods in a high-speed on-line system. Finally, these models are complicated for training. Training an LWR, for example, requires prior knowledge of the number of neighbors, distance measurement, and weight function. Even can be tuned by cross-validation, these hyper-parameters are sensitive to the training set variation, which makes them less robust than PLSR and more difficult to scale.

Another good solution is to build a two-tier PLSR model. In this kind of scheme, the training set will be divided into two groups: high or low in target value. Two PLSR models are fitted on each subset separately. This regression method can also remove the bias in the two tails. However, it is unclear how to split the data set, and how to choose the correct component model for making the prediction. Improper segmentation (in the calibration) and selection of model (in the prediction) will affect prediction accuracy, especially for the samples near the segmentation boundary.

The method proposed in this study, hierarchical mixture of linear regressions (HMLR) [10] solves these issues. The calibration method

* Corresponding author.
*E-mail address:* chenhao.cui.14@ucl.ac.uk (C. Cui).

https://doi.org/10.1016/j.chemolab.2017.12.013
Received 21 August 2017; Received in revised form 11 December 2017; Accepted 20 December 2017
Available online 4 January 2018

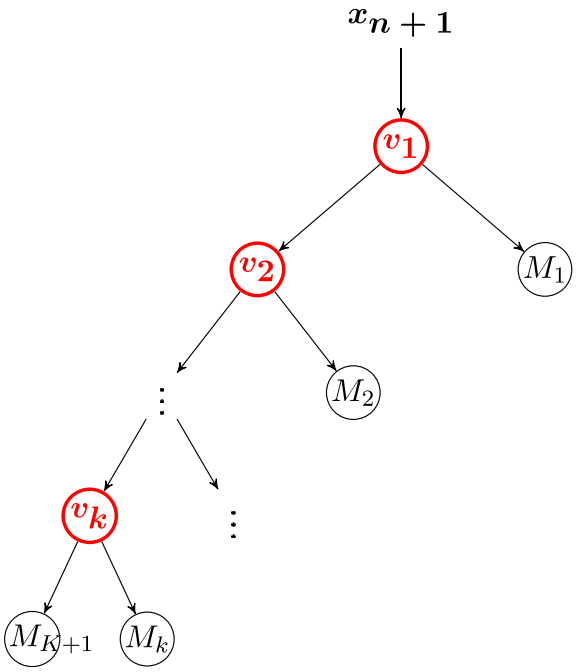


**Fig. 1.** Nested structure for integrating linear regression models. Red nodes: gating nodes, determine weights of component linear models; Black nodes: component models, each end model is an independent linear expert. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

assumes there are two or more underlying linear models behind the dataset. By using the expectation maximisation (EM) and the variational inference, the algorithm will find the most sensible component models automatically through an iterative optimisation process. A set of gating functions will also be trained simultaneously to assign individual observation into the correct component model. In the end, the trained model contains several independent linear models (e.g. PLSR or PCR) for different subsets of the training set and a set of gating functions. Compared to GPR and LWR, it is entirely parametric, which makes it possible to interpret and easy to use. It is also simple for training, only the number of PLS factors and spectral preprocessing methods need to be determined.

## 2. Methodology

### 2.1. Dimension reduction

For all calibration methods, the target is to train a regression model from a set of $N$ training instances $\{\boldsymbol{x_n}, y_n\}$, where $\boldsymbol{x_n}$ is the original NIR spectrum and $y_n$ is the lab measurement of the target constituent. Since NIR spectra are mostly high dimensional and mutually dependent, dimension reduction is applied to the original space. Partial least squares (PLS) and principal component analysis (PCA) are the most popular data shrinkage methods [11,12]. In this research PLS with NIPALS decomposition was used to transform original space $\boldsymbol{x_n}$ into a selection of latent variables $\phi_n$ [13]. PLSR is a result of direct MLR of $\phi_n$ on $y_n$. HMLR and the other benchmark method LWR were also learned on the transformed training set $\{\phi_n, y_n\}$.

### 2.2. Hierarchical mixture of linear regression

#### 2.2.1. Overview of the graphical model

The structure of HMLR can be illustrated by Fig. 1. Red nodes in the graph represent a set of sequential gating functions, and black nodes are component linear regression models. When the new observation $\boldsymbol{x}_{n+1}$ is obtained, it flows through the decision tree, each gating node $\boldsymbol{v_i}$ can decide whether the corresponding component model $M_i$ should be used to predict the observation.

There are two different prediction strategies: hard split and soft mix. Hard split means one and only one end model will be used to make the prediction. In the soft mix scheme, gating nodes are probabilistic. Outputs of gating nodes are not binary decisions, but a set of probabilities of using the corresponding component models. In the end, predictions from all component models will be weighted averaged out by the likelihood. In this study, soft mix of the component models was used to make the prediction.

The critical part is how to learn the gating functions $\boldsymbol{V}$ and component linear models $\boldsymbol{M}$ from a given training set. For a total number of $N$ training instances $\{\phi_n, y_n\}$, assume there are $K+1$ underlying linear regression functions $\boldsymbol{w_i}, i = 1, 2, \ldots, K+1$. Gating functions $\boldsymbol{V}$ indicate the correct component model (or the weight of each component model) for input $\phi_n$. A labeling variable $\boldsymbol{z_n}$ is assigned to each of the training instances. In this study, $\boldsymbol{z_n}$ is binary in training (i.e., either 0 or 1 depending on whether the corresponding end model is used). $\boldsymbol{z_n}$ is a vector of length $K+1$, corresponding to $K+1$ component models. The goal of the whole training process is to find a set of parameters of $\{\boldsymbol{V}, \boldsymbol{W}\}$ that can explain the training set. It is helpful to put appropriate regularization on these parameters to avoid over-fit.

Here Bayesian directed acyclic graph (DAG) [14] can be introduced to describe dependency of different parameters and variables in the calibration process. The DAG is shown in Fig. 2. The rectangular box represents $N$ training instances. The connection from the black dot $X_n$ to $\phi_n$ is deterministic transformation, which is a PLS data shrinkage process. $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ are parameters of the regularization on coefficients $\{\boldsymbol{W}, \boldsymbol{V}\}$. $\boldsymbol{\tau}$ is a vector of length of $K+1$, indicates the precision of reference for each Bayesian linear regression.

#### 2.2.2. Loss function: complete data log-likelihood

As introduced, the goal of training is to find a set of parameters that can explain the training set. The complete data log-likelihood (CDLL) is introduced as the loss function. According to the graphical model in Fig. 2, CDLL can be described by the following equation:

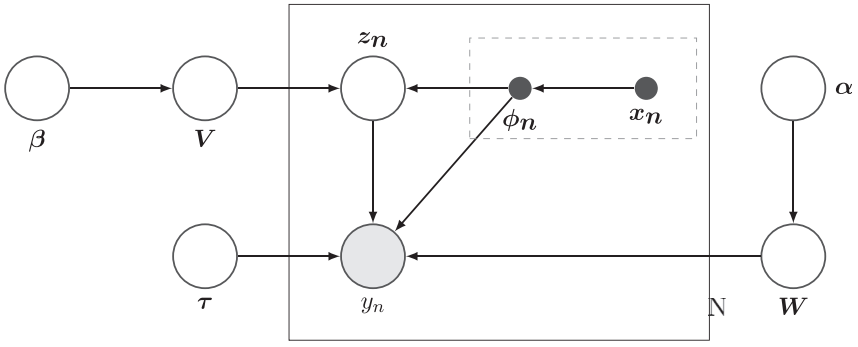


**Fig. 2.** Directed graphical model for mixture of linear experts.