ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Linear Projection based on Noise Variance Estimation - Application to Spectral Data

Amaury Lendasse and Francesco Corona

Helsinki University of Technology - Lab. of Computer and Information Science
P.O. Box 5400 FI-02015 HUT - Finland

Abstract. In this paper, we propose a new methodology to build latent variables that are optimal if a nonlinear model is used afterward. This method is based on Nonparametric Noise Estimation (NNE). NNE is providing an estimate of the variance of the noise between input and output variables. The linear projection that builds latent variables is optimized in order to minimize the NNE. We successfully tested the proposed methodology on a referenced spectral dataset from food industry (Tecator).

1 Introduction

Data from spectrophotometers form vectors with a large number of exploitable variables. Building quantitative models using these variables most often requires using a smaller set of variables than the initial one. Indeed, a too large number of input variables to a model results in a too large number of parameters, leading to overfitting and poor generalization abilities. Partial Least Squares Regression (PLS-R, [1]) has been successfully used to deal with a large number of input variables that are, moreover, near-collinear. In PLS-R, the input variables are linearly projected onto latent structures where the data are assumed to reside. The projection is performed to maximize the non-redundant information needed to build a linear regression model. Specifically, the basic Partial Least Squares formulation combines linear dimension reduction and regression by minimizing correlation between inputs and maximizing covariance with the output.

In practice, PLS-R provides good reference models. However, it is limited when the input-output relationship is nonlinear. In fact, the latent variables that are built by PLS-R are only optimized in order to provide the best input variables if a linear model is assumed. Their optimality is not straightforward when the intrinsic nonlinearities have to be reconstructed using other regression techniques like Least-Squares Support-Vector Machines (LSSVM, [2]) or Multi-Layer Perceptrons (MLP, [3]).

In this paper, we propose a new approach to the construction of latent variables that are optimal if a nonlinear model is used afterward. The method is based on Noise Variance Estimation (NNE, [4]). The NNE provides an estimate of the variance of the noise between input and output variables, or equivalently, an estimate of lowest Mean Squared Error (MSE) that can be achieved by a regression method without overfitting the data. Thus, the NNE is an optimal criterion for either selecting a subset of original inputs or building new latent inputs. In this work, the minimization of the NNE is used to learn a linear projection matrix and build a new set of latent variables with predictive properties.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

The paper is organized as follows. In Section 2, the algorithmic part of the study is briefly overviewed and the approach to building latent variables is outlined in the context of spectroscopic modeling. For brevity, Section 3 is restricted to illustrate only one referenced spectral application from food industry [5].

2 Methodology

For completeness, this section briefly overviews the tools used in the study (namely, the Delta Test, linear projections and an Extended Forward-Backward variable selection) and presents a specific application to spectroscopic modeling.

2.1 Nonparametric Noise Estimation with the Delta Test

Delta Test (DT) is a technique for estimating the variance of the noise or, equivalently, the Mean Square Error (MSE), that can be achieved by a regression model without overfitting; see [4] and references therein. As such, the DT is useful for evaluating the nonlinear correlation between two random variables and can be included in variable selection schemes: the set of inputs minimizing the DT is the one that is to be retained.

Given N input-output pairs: $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, the relationship between $\mathbf{x}_i$ and y_i is modeled as $y_i = f(\mathbf{x}_i) + r_i$ where f is the unknown function to be estimated and r_i is the noise. The Delta Test is a data-derived method to estimating the variance of the noise in such a setting. Denoting by $\mathbf{x}_{NN(\mathbf{x}_i)}$ the first nearest neighbor of point $\mathbf{x}_i$ in the set $\{\mathbf{x}_i\}_{i=1}^N$ and $y_{NN(\mathbf{x}_i)}$ the associated output, the Delta Test, δ, formulates as:

$$\delta = \frac{1}{2N} \sum_{i=1}^{N} ||y_{NN(\mathbf{x}_i)} - y_i||^2. \tag{1}$$

2.2 A Projection based on the Delta Test

Linear projection is a common preprocessing step in both function approximation and classification tasks. When regression is to be performed, the aforementioned PLS-R, as well as other methods like Principal Components Regression (PCR), are two standard approaches based on the idea of combining the original variables by projection. Both methods project the original input variables onto a latent space with reduced dimensionality; in PCR, the projection is constructed in order to keep a maximum of information from the input variables, whereas PLS-R builds new inputs that are also suitable to approximate the output, [1].

This subsection illustrates an efficient strategy to use the Delta Test as a tool to select an optimal linear projection of the original input variables. Being based on the DT, the strategy is mostly suitable when a nonlinear model is used to reconstruct the relationship between the new latent inputs and the output.

For N input-output pairs, $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$, a new set of inputs $\mathbf{z}$ is given as:

$$\mathbf{z} = \mathbf{x}\mathbf{P}, \tag{2}$$

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

where $\mathbf{P}$ is the projection matrix. According to Delta Test, the best set of latent variables $\mathbf{z}$ is found as the one that minimizes:

$$\delta = \frac{1}{2N} \sum_{i=1}^{N} ||y_{NN(\mathbf{z}_i)} - y_i||^2, \tag{3}$$

where $y_{NN(\mathbf{z}_i)}$ is now the output for $\mathbf{z}_{NN(\mathbf{z}_i)}$. Thus, we define an optimal $\mathbf{P}$ as:

$$\mathbf{P}^{opt} = \arg\min_{\mathbf{P}} \frac{1}{2N} \sum_{i=1}^{N} ||y_{NN(\mathbf{z}_i)} - y_i||^2, \tag{4}$$

Unfortunately, the optimization for $\mathbf{P}^{opt}$ is difficult because the Delta Test is not differentiable with respect to $\mathbf{P}$; the discontinuity is due to the fact that the Delta Test estimates the variance of the noise based on nearest neighbors.

In order to optimize for $\mathbf{P}^{opt}$, an Extended Forward-Backward optimization technique can be used. Forward-Backward Selection (FBS) is a commonly used strategy for variable selection. The method is fast but there is no guarantee that the optimal set of variables is found [3]. In FBS, each variable can be in two states: "1", meaning that it belongs to the set of selected variables or "0" meaning that it does not and it is temporarily discarded. Given a certain initial state for all variables, the procedure flips the state of each variable at a time and computes a predefined criterion (for example, the Delta Test). The flipping operation that improves performances the most is accepted, and the states are flipped again (excluding the previously accepted change). The process is continued until no improvement is found.

FBS can be extended to any optimization problem for which the importance (or level) of a variable is searched; that is, instead of switching scalars from 0 to 1 or vice versa, by increasing (in case of forward selection) or decreasing (for backward selection) by regular steps $1/h$. In this study, we suggest an application of extended FBS schemes to the problem of optimizing the projection matrix $\mathbf{P}$.

In general, we assume that the initial variables have been normalized, then we assume that the values of $\mathbf{P}$ can be bounded by -1 and 1. In practice, a degree of discretization $h = 10$ is also found to be accurate enough. For a projection onto a 2-dimensional space, the procedure can be summarized as:

1. initialize the first column of $\mathbf{P}$;
2. optimize the first column of $\mathbf{P}$ by FBS from DT in the projected space;
3. initialize the second column of $\mathbf{P}$;
4. optimize the second column of $\mathbf{P}$ by FBS with the first column unchanged.

The data projected onto a 2-dimensional latent space are easily displayed and initially used to investigate their structure in the input space, being the visualization supervised by the output. If visualization is not the main concern, the procedure can be extended to additional columns of $\mathbf{P}$ (e.g., until no significant decrease of the Delta Test is observed) and then used to estimate the output.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

2.3 Methodology for Spectrometric Modeling

The forward backward selection has been previously tested in order to optimize a projection matrix when the criterion is the minimization of the Delta Test. Unfortunately, the method is converging and the results are satisfactory only if the number of variables is small (approximately, 20 variables).

In order to approach such a restriction, we suggest to use the PLS-R as a preprocessing step; thus, allowing a preliminary reduction in the dimensionality of the original problem. The number of latent variables to be retained after performing PLS-R should be a compromise capable to conserve most of the information exploitable by a nonlinear method, but also small enough in order to be able to perform the minimization of the Delta Test. Notice that the number of variables retained from PLS-R is, however, not critical when the choice is conservative; in our experiments, we found that retaining twice the number of latent variables obtained from a cross-validated PLS-R is typically appropriate.

For problems preprocessed by PLS-R, the procedure can be summarized as:

1. build a PLS-R model between the inputs and the output. For a model cross-validated for k_1 latent directions retain as many as $2k_1$;
2. project the data $\mathbf{x}$ onto the space spanned by the first $2k_1$ PLS-R directions:

$$\mathbf{z}_1 = \mathbf{x}\mathbf{P}_1. \tag{5}$$

   Here, $\mathbf{P}_1$ denotes the the projection matrix associated to PLS-R;
3. perform FBS in order to find a second projection matrix $\mathbf{P}_2$ such that the DT between the final set of latent inputs and the output is minimized:

$$\mathbf{P}_2 = \min_{\mathbf{P}} \frac{1}{2N} \sum_{i=1}^{N} ||y_{NN(\mathbf{z}_{1,i})} - y_i||^2, \tag{6}$$

4. project $\mathbf{z}_1$ onto the space spanned by the directions optimized from DT:

$$\mathbf{z}_2 = \mathbf{z}_1\mathbf{P}_2. \tag{7}$$

The linearly projected data $\mathbf{z}_2 = \mathbf{x}\mathbf{P}_1\mathbf{P}_2 = \mathbf{x}\mathbf{P}$ are then used to calibrate any nonlinear model to estimating the output y. In our experiments, the Least-Squares Support-Vector Machine for regression (LSSVM, [2]) is adopted.

3 Experimental

The Tecator dataset is a referenced problem in spectroscopy [5] for predicting the fat content of 215 observations of minced meat samples from Near-Infrared (NIR) absorption spectra. The input spectra consists of 100 near-collinear variables corresponding to the absorbance of the meat sample measured in the correspondence of 100 light wavelengths ranging from 850 to 1050 nanometers. The

output fat content ranges from 0.9 to 49.1 percent and it is measured in laboratory. Based on the dataset guidelines, the first 172 observations are used as a learning set $\mathcal{L}$ and the remaining 43 are used as a test set $\mathcal{T}$ to assess the final regression model. The input spectra are illustrated in the left panel of Figure 1.

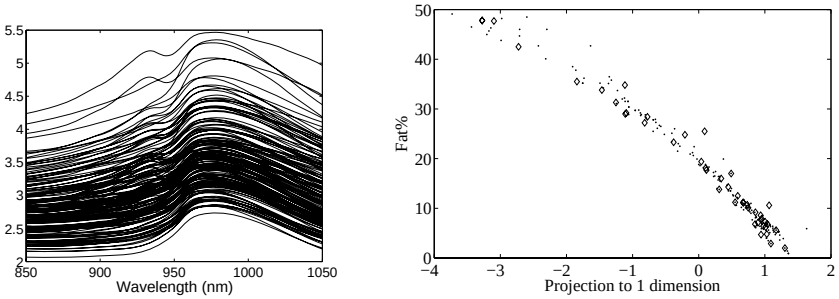

Fig. 1: A selection of input spectra (left) and the fat content with respect to the one-dimensional projection (right). In the right panel, dots ($\cdot$) correspond to the projection of the learning set and the diamonds ($\diamond$) to the testing set.

In the case of Tecator data, the number of latent variables that is found optimal by cross-validation for a PLS-R model equals 4. According to the methodology presented in Subsection 2.3, we have projected the initial input variables onto the space corresponding to the first 8 latent directions. Then, the FBS is used to optimize the 1D projection; the left panel in Figure 1 depicts the fat content against the 1D input projection. From the figure, it can be noticed that the learning and the testing set show an identical behaviour, being projected to the same region. Such a behavior probably indicates that no overfitting in the projection process has occured. It is also interesting to notice that the relationship between the output and the input is not totally linear.

These observations show the qualitative advantage of projecting onto a very low-dimensional space. The projection matrix $\mathbf{P_2}$ from the optimization is:

$$\mathbf{P}_2 = (0.1 \quad -1 \quad -0.5 \quad -0.5 \quad 0 \quad 0.1 \quad 0 \quad 0)^\top, \tag{8}$$

where only the second, third and fourth latent variables appear to be important for the problem. It is worthwhile noticing that the reduced importance of the first variable was already observed for this dataset. Additionally, the result corroborates the number of latent variables cross-validated for the PLS-R.

In conclusion, LSSVM models between the projection variables and the output are build. The prediction results (Table 1) are compared to other methods reported in literature using the testing set and 2 different measures: the MSE and the Normalized Mean Square Error (NMSE). The obtained accuracies clearly indicate the advantages of coupling optimized projections to a nonlinear model. Specifically, using LSSVM reduces the MSE by a factor of 3 to 5 when compared

to any linear model. Moreover, the application is computationally sustainable; for the presented case study, the optimization is performed in 1 minute and the calibration of the LSSVM in 2 minutes using a desktop computer.

Method	**N. of variables**	MSE_T	$NMSE_T$
PLS-R	100 (8)	4.45	0.0274
Projection to 1 dim. + LS-SVM	1 (8)	1.35	0.0083
Projection to 2 dim. + LS-SVM	2 (8)	0.85	0.0052

Table 1: Results for the Tecator. Number of latent variables given in parenthesis.

4 Conclusions

This study presented an alternative approach to build a nonlinear model in the context of spectroscopic modeling. The methodology is based on the optimization of a linear projection that reduces the number of latent variables. The optimality criterion is an estimate of the variance of the noise using the Delta Test. The resulting projection matrix is suitable for any nonlinear model to be used afterward. Based on our results, the main advantages of the suggested methodology can be summarized as: 1) accurate results in the presence of nonlinearities; 2) reduced computational burden and 3) projection on 1-2D spaces where the data can be be visualized, outliers detected and the degree of nonlinearity assessed. As a concluding remark, we point out that, for linear problems, the approach can be used to validate the number of latent variables. It is our future goal to improve the strategy used for the optimization.

References

[1] Wold H. Partial least squares. In *Encyclopedia of Statistical Sciences*, volume 6, pages 581–591. Wiley, New York, 1985.

[2] Suykens J.A.K., Van Gestel T., De Brabanter J., De Moor B., and Vanderwalle J. *Least-Squares Support-Vector Machines*. World Scientific, Singapore, 2002.

[3] Haykin S. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, New York, second edition, 1999.

[4] Jones A. New tools in non-linear modeling and prediction. *Computational Management Science*, 1:109–149, 2004.

[5] Rossi F., Lendasse A., Francois D., Wertz V., and Verleysen M. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems*, 80(2):215–226, 2006.