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.

# Generalized Matrix Learning Vector Quantizer for the Analysis of Spectral Data

**Petra Schneider**[1], **Frank-Michael Schleif**[2], **Thomas Villmann**[2] and **Michael Biehl**[1]

1- University of Groningen, Mathematics and Computing Science
P.O. Box 407, 9700 AK Groningen - The Netherlands

2- University of Leipzig, Department of Medicine
Semmelweisstrasse 10, 04103 Leipzig - Germany

**Abstract**. The analysis of spectral data constitutes new challenges for machine learning algorithms due to the functional nature of the data. Special attention is paid to the metric used in the analysis. Recently, a prototype based algorithm has been proposed which allows the integration of a full adaptive matrix in the metric. In this contribution we study this approach with respect to band matrices and its use for the analysis of functional spectral data. The method is tested on data taken from food chemistry and satellite image data.

**Keywords:** Learning Vector Quantization, metric adaptation, GMLVQ, band matrices, spectral data.

## 1 Introduction

The analysis of high dimensional functional data is a common task in different fields of natural sciences like medicine and chemistry. Prominent examples are mass spectrometry data (MS) in the field of clinical proteomics, nuclear magnetic resonance spectra (NMR) in the field of chemistry and metabolomics or spectral imagery acquired by satellites to name just a few. Focusing on classification, prototype based classification approaches such as Learning Vector Quantization (LVQ) as proposed by Kohonen [6] or multiple extensions [3, 8] have already proven to be valuable for the analysis of high dimensional data (see [9, 10]). Due to the complexity of the data the use of an appropriate distance measure is of special importance [13] to get an adequate representation of the data. So-called relevance learning techniques [2] extend, e.g. the Euclidean distance, with weight factors for the different dimensions. Together with the prototypes these factors are optimized with respect to the given classification task during training. This allows to scale the axes of the coordinate system of the data space in order to obtain better adaptation towards clusters with axes-parallel ellipsoidal shapes. But this approach ignores correlative effects between different features in general. The recently introduced Generalized Matrix LVQ (GMLVQ) [1, 11] adapts a full matrix of relevance factors in the distance measure. This accounts for pairwise correlations of features and ellipsoidal clusters which are not axes parallel can be obtained. Yet, full adaptive GMLVQ may suffer from the presence of too many adjustable parameters as their number grows quadratically with the dimensions of input. This can lead to instabilities and overfitting. In spectral data the order of the features is not arbitrary and usually local correlations between neighbored dimensions occur. Due to this property, the restriction of GMLVQ to the adaptation of band-limited matrices appears to be natural for the analysis of this kind of data. Hence, the number of free parameters can be reduced without limiting the performance of the algorithm significantly. In this paper we analyze this modification of GMLVQ on two different spectral data sets coming from food chemistry studies and satellite remote sensing.

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 Generalized Matrix LVQ

LVQ aims at parameterizing a classification scheme in terms of prototypes. Assume training data $(\xi_i, y_i) \in \mathbb{R}^N \times \{1, \ldots, C\}$ are given, $N$ denoting the data dimensionality and $C$ the number of different classes. An LVQ network consists of a number of prototypes which are characterized by their location in the weight space $\mathbf{w}_i \in \mathbb{R}^N$ and their class label $c(\mathbf{w}_i) \in \{1, \ldots, C\}$. Classification takes place by a winner takes all scheme. For this purpose, a (possibly parameterized) similarity measure $d^\lambda$ is defined in $\mathbb{R}^N$. Often, the standard Euclidean metric is chosen. A data point $\xi \in \mathbb{R}^N$ is mapped to the class label $c(\xi) = c(\mathbf{w}_i)$ of the prototype $i$ for which $d^\lambda(\mathbf{w}_i, \xi) \leq d^\lambda(\mathbf{w}_j, \xi)$ holds for every $j \neq i$ (breaking ties arbitrarily).

Learning aims at determining weight locations for the prototypes such that the given training data are mapped to their corresponding class labels. A very flexible learning approach has been introduced in [4]. It is derived as a minimization of the cost function

$$\sum_i \Phi\left(\frac{d_J^\lambda - d_K^\lambda}{d_J^\lambda + d_K^\lambda}\right) \tag{1}$$

where $\Phi$ is a monotonic function, e.g. the identity or the logistic function, $d_J^\lambda = d^\lambda(\mathbf{w}_J, \xi_i)$ is the distance of data point $\xi_i$ from the closest prototype $\mathbf{w}_J$ with the same class label $y_i$, and $d_K^\lambda = d^\lambda(\mathbf{w}_K, \xi_i)$ is the distance from the closest prototype $\mathbf{w}_K$ with a different class label than $y_i$. Taking derivatives with respect to the prototypes and metric parameters yields gradient based adaptation rules. The choice of the similarity measure as standard Euclidean metric yields GLVQ [8]. The squared *weighted* Euclidean metric $d^\lambda(\mathbf{w}, \xi) = \sum_i \lambda_i (w_i - \xi_i)^2$ where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$ constitutes a powerful alternative, GRLVQ [5], particularly suitable for high dimensional data with input dimensions of different (but a priori unknown) relevance. In GMLVQ, a full matrix which can account for pairwise correlations of the dimensions, is used. The metric has the form

$$d^\Lambda(\mathbf{w}, \xi) = (\xi - \mathbf{w})^T \Lambda (\xi - \mathbf{w})$$

where $\Lambda$ is an $N \times N$ matrix. The above similarity measure only corresponds to a meaningful distance if $\Lambda$ is positive (semi-) definite. We can achieve this by substituting $\Lambda = \Omega \Omega^T$. Without loss of generality we consider only symmetric $\Lambda$. We can furthermore assume that $\Omega$ itself is symmetric as the (unique) symmetric square root of $\Lambda = \Omega^2$ always exists. To obtain the adaptation formulas we need to compute the derivatives of (1) with respect to $\mathbf{w}$ and $\Omega$. We get the updates

$$\begin{aligned}
\Delta \mathbf{w}_J &= +\epsilon_1 \cdot \phi'(\mu(\xi)) \cdot \mu^+(\xi) \cdot \Omega\Omega \cdot (\xi - \mathbf{w}_J) \\
\Delta \mathbf{w}_K &= -\epsilon_1 \cdot \phi'(\mu(\xi)) \cdot \mu^-(\xi) \cdot \Omega\Omega \cdot (\xi - \mathbf{w}_K) \\
\Delta \Omega_{lm} &= -\epsilon_2 \cdot \phi'(\mu(\xi)) \cdot \\
&\quad \Big( \mu^+(\xi) \cdot \Big( [\Omega(\xi - \mathbf{w}_J)]_m (\xi_l - w_{J,l}) + [\Omega(\xi - \mathbf{w}_J)]_l (\xi_m - w_{J,m}) \Big) \\
&\quad - \mu^-(\xi) \cdot \Big( [\Omega(\xi - \mathbf{w}_K)]_m (\xi_l - w_{K,l}) + [\Omega(\xi - \mathbf{w}_K)]_l (\xi_m - w_{K,m}) \Big) \Big)
\end{aligned}$$

for the prototypes and matrix elements $\Omega_{lm}$ with $\mu(\xi) = (d_J^\Lambda - d_K^\Lambda)/(d_J^\Lambda + d_K^\Lambda)$, $\mu^+(\xi) = 2 \cdot d_K^\Lambda / (d_J^\Lambda + d_K^\Lambda)^2$, and $\mu^-(\xi) = 2 \cdot d_J^\Lambda / (d_J^\Lambda + d_K^\Lambda)^2$. (See [1] for the

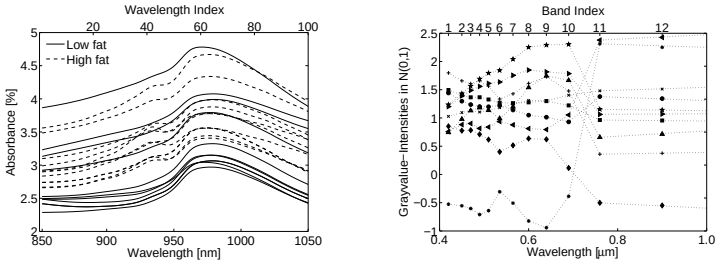


**Fig. 1:** Left: Multiple spectra of both classes taken from the Tecator data set. Right: Average spectra of all classes of the satellite data set. The class labels are indicated by different point symbols. ● Alfalfa, ▼ Soil, ★ Corn, ■ Oats, ♦ Red Clover, ◄ Rye, ► Soybeans, ▲ Water, * Wheat1, + Wheat2

derivation of these formulas.) Thereby, the learning rate for the metric can be chosen independently of that for the prototypes. Note that the update preserves the symmetry of $\Omega$. After each update, $\Omega$ is normalized to prevent the algorithm from degeneration. We set $\sum_i \Lambda_{ii} = \sum_{i,j} \Omega_{ij}^2 = 1$ which fixes the sum of diagonal elements and, thus, the sum of eigenvalues of $\Lambda$.
Band-limited GMLVQ can be achieved by limiting the number of non-zero adjacent diagonals in $\Omega$ or $\Lambda$, respectively. If $k$ adjacent diagonals above and below the main diagonal are considered in $\Omega$ the respective bandwidth including the main diagonal is given as $n = 2 \cdot k + 1$ in $\Lambda$. We refer to this as GMLVQ-$n$. This restriction leads to a focus on locally correlated frequency bands in spectral data. The parameter $n$ should be in correspondence to the correlation range in the spectra, which is problem specific. Note that GMLVQ-1 corresponds to GRLVQ.

## 3 Data

We test the approach using the Tecator benchmark data set which is available at http://lib.stat.cmu.edu/datasets/tecator. It contains 215 infrared absorption spectra of meat samples. Each spectrum has been measured at 100 wavelengths ranging from $850\,nm$ to $1050\,nm$. The classification task consists in the prediction of the binary fat content (low/high) of the probes. Figure 1 (left) visualizes several example spectra of both classes. Apart from a tendency towards dints around channel $41$ for high fat content, a substantial overlap of the classes is evident.
Furthermore, the algorithm is applied to the Flightline C1 data set obtained form [7] which contains 12-band multispectral gray-value data. It was taken by an M7-scanner over Tippecanoe Country, Indiana and discriminates between 10 different classes (Alfalfa, Soil, Corn, Oats, Red Clover, Rye, Soybeans, Water, Wheat1, Wheat2). The spectral window $0.4\,\mu m$ to $1.0\,\mu m$ covers the visible - near-infrared range. The visible range mainly judges leaf pigments (chlorophyll) and the infrared range is mostly responsible for cell structures (spongy-mesophyll cells). The data set is splitted into 11451 samples for training and 70549 samples for testing. Figure 1 (right) depicts the mean spectra of the different classes. A more detailed description of the data is available in [7].

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.

## 4 Experiments and Results

We apply the proposed modification of GMLVQ with different bandwidth settings and compare the classification performance to known results taken form [7, 12]. The following statements hold for both data sets: the samples are normalized to zero mean and unit variance in each feature and we employ one prototype per class respectively. The learning rates are continuously reduced in the course of training. The initial values are set to $\epsilon_1 = 5 \cdot 10^{-3}, \epsilon_2 = 5 \cdot 10^{-4}$ (Tecator data) and $\epsilon_1 = 0.01, \epsilon_2 = 10^{-4}$ (satellite data). We use the same learning rate schedule as in [1] with $c = 10^{-4}$. Pretraining with simple GLVQ is mandatory in all experiments. Because of the small number of samples contained in the Tecator data set, the spectra are partitioned randomly into $4/5$ samples for training and $1/5$ patterns for testing averaged in a 5-fold cross validation. Figure 2 and Table 1 summarize the obtained classification accuracies. We observe that adapting only a small number of bands in the relevance matrix is sufficient to clearly improve the classification performance compared to GRLVQ. The learning curves for the satellite data depict a clear gap between the accuracies obtained with the settings $k = 0$ and $k = 1$. Furthermore, we can conclude that the number of bands needed to obtain the performance of GMLVQ is rather small compared to the dimension of the input data. We find the bandwidths of $n = 21$ for the Tecator data set and $n = 5$ for the satellite data set to be sufficient to achieve this performance. Adding further bands yields no significant benefit for the classification any more. Accordingly, correlations between features far apart in the sequence of all features can be ignored without constraining the performance of the GMLVQ-algorithm significantly. Thus the number of free parameters can be reduced, efficiently. It is interesting to note, however, that the learning curves indicate that the system needs more training epochs to reach the same performance for small $n$ in case of high input dimensionality (Tecator data).
Figure 3 visualizes the relevance matrices obtained by GMLVQ and GMLVQ-21 on the Tecator data set. The diagonal elements of the full matrix reflect that the region around feature 41 is ranked highest. The most relevant range captures the indices 30 to 50. This conforms to the visual impression given in Figure 1 and the property of local correlations in spectral data. Accordingly, off-diagonal elements which are significantly different from zero are detected only in the neighborhood of the main diagonal around index 41. Consequently, the restricted band matrix in GMLVQ-21 is adequate to represent the meaningful structure in the data appropriately and to achieve the classification performance of GMLVQ.
Our findings for the satellite data are in good agreement with prior results published in [7]. The accuracies on the test data given in [7] are slightly better ($\approx 5\% - 7\%$) than our best results. However, the complex feature selection scheme used in [7] implies a lot more computation effort to achieve this performance. Further it should be mentioned that this feature selection is not an inherent part of the classifier method. The detailed steps in [7], p. 179 suggest that a potential bias with respect to the feature selection could have effected these results. Therefore the method independent experimental settings given here and in [7] are slightly different, complicating direct comparisons. Considering the results including the expert knowledge in [7] about relevant spectral bands for vegetation discrimination, interesting findings can be made. In particular both visible and infrared frequencies contribute to the identification. A bandwidth of 5 in the given data set comprises at least parts of the visible and near-infrared spectrum. Therefore the respective correlations are taken into account. Smaller bandwidths lead to a loss of this correlation information, whereas larger ones, above 5, give no significant information gain. For this optimum 5-band case the experiment was repeated using 5

prototypes per class. We achieved a prediction of 86.4% which is comparable to the result given in [7] with 91% using only 4 features. This selection was done such that the features are almost independent, but covering visible and infrared frequencies. The main diagonal elements of matrix $\Lambda$ (relevance profile), reflect that red and infrared frequencies are especially relevant for the classification. This underlies the above mentioned features of chloro- and mesophyll level for vegetation discrimination.

| Tecator | | Satellite | |
|---|---|---|---|
| Algorithm | Prediction | Algorithm | Prediction |
| GMLVQ-1 | 66.7% | GMLVQ-1 | 78.5% |
| GMLVQ-3 | 66.7% | GMLVQ-3 | 82.9% |
| GMLVQ-11 | 84.3% | GMLVQ-5 | 86.2% |
| GMLVQ-21 | 97.1% | GMLVQ-7 | 86.3% |
| GMLVQ-31 | 97.1% | GMLVQ-9 | 86.6% |
| GMLVQ-41 | 96.7% | GMLVQ-11 | 86.4% |
| GMLVQ-F | 95.7% | GMLVQ-F | 86.6% |
| SVM-RBF | 68.9% | SVM-RBF | 70.7% |
| SVM-Lin | 73.3% | SVM-Lin | 85.3% |
| C-GRLVQ | 97% | C-GRLVQ | n.a. |

**Table 1:** Classification accuracies achieved on the Tecator- and the satellite data set using different bandwidth settings for GMLVQ-$n$ (1 to F-full) in comparison to correlation based GRLVQ (C-GRLVQ) with 20 prototypes [12] and two types of a SVM (Lin-linear, RBF-radial basis function kernel) obtained using Yale (http://yale.cs.uni-dortmund.de)

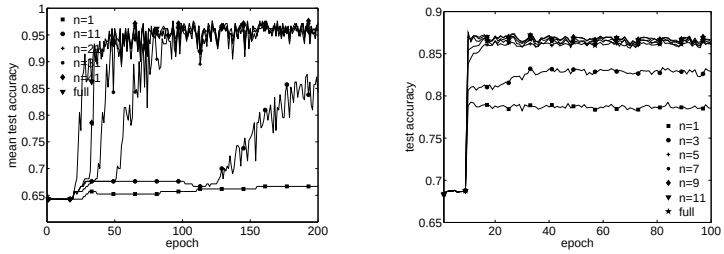


**Fig. 2:** Prediction accuracies as a function of training time for both data sets using matrices of different bandwidths. Left: Tecator data set (matrix adaptation starts after 20 epochs). Right: Satellite data set (matrix adaptation starts after 10 epochs).

## 5 Conclusion

In this article band-limited GMLVQ has been investigated for classification of spectral data. For both considered data sets we observe an overall improvement in prediction, compared to simple GRLVQ. The improvement is the same as achieved by unrestricted GMLVQ. However band-limitation can be applied successfully without significant information loss. The obtained optimum bandwidths can be discussed in the light of spectra properties of the underlying problems. Thus band-limiting can be used to reduce the number of adjustable parameters of standard GMLVQ to improve the stability. These findings may carry over to other kinds of spectral data such as mass spectra

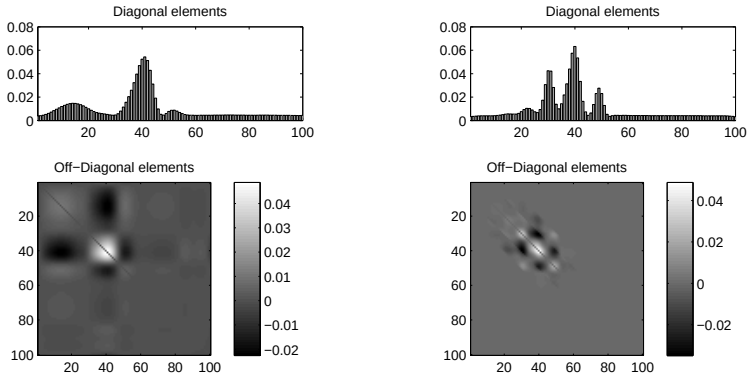


**Fig. 3:** Visualization of the relevance matrices obtained by GMLVQ (left) and GMLVQ-21 (right) on the Tecator data set. The diagonals are set to zero in the off-diagonal-presentations.

(MS) or Ion Mobility Spectroscopy (IMS) which is an important analysis technique in chemistry and the field of security. In our experiments we considered only the effect of symmetric correlations between neighbored features using different bandwidths. This maybe not always optimal with respect to the underlying data characteristics. In future work also more generic correlation scenarios will be analyzed.

## References

[1] M. Biehl, B. Hammer, and P. Schneider. Matrix learning in learning vector quantization. Technical report, Department of Informatics, Clausthal University of Technology, 2006.

[2] T. Bojer, B. Hammer, D. Schunk, and K. Tluk von Toschanowitz. Relevance determination in learning vector quantization. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks*, pages 271–276, 2001.

[3] B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, February 2005.

[4] B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, 2005.

[5] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[6] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, second edition, 1997.

[7] D. Landgrebe. *Signal Theory Methods in Multispectral remote sensing*. Wiley, New Jersey, 2003.

[8] A. Sato and K. Yamada. Generalized learning vector quantization. In M. C. Mozer D. S. Touretzky and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9, Cambridge, MA, USA, 1996. MIT Press.

[9] F.-M. Schleif, T. Elssner, M. Kostrzewa, T. Villmann, and B. Hammer. Analysis and visualization of proteomic data by fuzzy labeled self organizing maps. In *International Symposium on Computer-Based Medical Systems*, pages 919–924. IEEE press, 2006.

[10] F.-M. Schleif, T. Villmann, and B. Hammer. Prototype based fuzzy classification in clinical proteomics. *International Journal of Approximate Reasoning on Approximate reasoning and Machine learning for Bioinformatics*, 47(1):4–16, 2008.

[11] P. Schneider, M. Biehl, and B. Hammer. Relevance matrices in lvq. In M. Verleysen, editor, *European Symposium on Artifiacal Neural Networks*, pages 37–42, Bruges, Belgium, April 2007.

[12] M. Strickert, W. Weschke N. Sreenivasulu, T. Villmann, and B. Hammer. Generalized relevance LVQ (GRLVQ) with correlation measures for gene expression analysis. *Neurocomputing*, 69(7-9):651–659, 2006.

[13] T. Villmann, F.-M. Schleif, and B. Hammer. Comparison of relevance learning vector quantization with other metric adaptive classification methods. *Neural Networks*, 19:610–622, 2006.