Sparse coding Neural Gas for analysis of Nuclear Magnetic Resonance Spectroscopy

Frank-Michael Schleif, Matthias Ongyerth, Thomas Villmann
University Leipzig, Dept. of Medicine, 04103 Leipzig, Germany
{schleif,villmann}@informatik.uni-leipzig.de,{matthias.ongyerth}@medizin.uni-leipzig.de

Abstract

Nuclear Magnetic Resonance Spectroscopy is a technique for the analysis of complex biochemical materials. Thereby the identification of known sub-patterns is important. These measurements require an accurate preprocessing and analysis to meet clinical standards. Here we present a method for an appropriate sparse *encoding of NMR spectral data combined with a fuzzy classification system allowing the identification of sub-patterns including mixtures thereof. The method is evaluated in contrast to an alternative approach using simulated metabolic spectra.* **keywords:** *data analysis, nuclear magnetic resonance, sparse coding*

1 Introduction

Nuclear Magnetic Resonance Spectroscopy (NMR) is a technique for the analysis of complex substances such as cell extracts. One prominent NMR application is metabolite profiling in stem cell biology [3]. NMR spectra are high dimensional functional signals consisting of a multitude of peaks. The peak positions describe the presence of specific chemical compounds in the analyzed material while the area of the peaks are quantitative with respect to the amount of this analyte in the substrate.

To meet clinical standards and high throughput demands a full automatic evaluation is required. Hereby, an accurate preprocessing of the NMR spectra and identification of metabolites in a set of NMR spectra are key issues. High level analyses such as metabolite identification still requires a lot of manual interaction to fit expected patterns against the measured signal. Typically, some assumptions are incorporated to simplify the fitting. The NMR signal consists of multiple peaks of specific shape, thereby the shape is approximated by a Lorentzian or some type of Gaussians, giving a bias in the analyses. Here we present a method which estimates data specific basis functions such that an appropriate compact representation of the NMR spectra becomes possible. This helps to reduce two of the mentioned problems, namely the high dimensionality of the analyzed data and the approximation problem in the fitting approach.

The paper is organized as follows. First we present the used methods, including a short description of the basic preprocessing of the NMR spectra and how the simulated metabolites are obtained. Subsequently we present the method of sparse coding in a specific variant tailored to the analysis of functional data. The introduced approach is applied on sets of simulated pure NMR metabolites as well as of mixtures of such metabolites. We close the paper with a discussion of the obtained results and an outlook on further extensions.

2 Methods

We consider simulated ^1H NMR spectra recorded at 700.15 MHz with $20K$ complex data points. The simulations are done using the gamma-System as given in [13]. It is assumed that each sample is solved in D_2O with an additive of DSS as a standard reference at 4.7 ppm. All spectra were Fourier transformed with 1.5 Hz line broadening. The NMR spectra are phased and baseline corrected, the signal range of the water peak

```
function [vBL] = base_det(vSignal,dDSSPoints)
% Signal size, Window width
ns = size(vSignal,1); w  = dDSSPoints;
% Empty windows whole signal
temp          = zeros(w,ceil(ns/w))+NaN;
% Fill in - over windows, min's per window
temp(1:ns)    = vSignal; [m,h] = min(temp);
g             = h>1 & h<w; % mins, not at borders
% calc minima positions with resp. to x-axis
h             = w*[0:numel(h)-1]+h;
% get valid minima and intensities
m             = m(g);  h = h(g);
% interpolate
vBaseline     = interp1(h,m,1:ns,'pchip');
```

Figure 1. Matlab code for baseline correction by piecewise cubic interpolation using a problem adequat segment width.

is eliminated and all spectra are aligned to the DSS reference. For comparison the data are analyzed by a standard approach employing peak lists, thereby its recommended that the used peaks show a $SNR \geq 5$. While the presented results are given for simulations only, the method is applicable for a wide range of settings in real NMR experiments.

2.1 Preprocessing of NMR spectra

Phase correction is a necessary step [4] made by use of the approach given in [2] followed by a Fourier transformation. Next, the known region of the water peak is removed and the DSS peak is identified, located at 4.7 ppm. The DSS is used as a reference to align multiple spectra and to specify the typical peak shape including a typical width and height of a peak. Due to underground artifacts in the signal a baseline correction has to be applied. This involves two steps: baseline estimation and subtraction [1, 17]. In general baseline correction algorithms define a sequence of simple e.g. polynomial functions connected by some support points estimating the underground of the signal. Thereby the critical point is the number and position of the support points as well as the chosen functions constructing the baseline. If some additional information is available about the specific nature of the data, the baseline correction can be made problem specific. Here we use information about the DSS signal width to define the width between two supporting points and obtain an effective and fast algorithm depicted in Figure 1. The baseline is subtracted from the signal and negative values are truncated to zero. Following the basic preprocessing an optional encoding step can be applied, to reduce the dimensionality of the data. In general the identification of patterns in NMR data is realized by a difference matching approach. This requires complicated fitting procedures, sensitive to model assumptions such as the peak shape in the signal. For difference matching an additional encoding is omitted and the processing takes place on the whole dimensionality of the data. Taking this approach in an automatic analysis one obtains concentrations for each pattern, thereby it is not always obvious if small values are due to a missing pattern, wrongly identified, or due to small concentrations of a truly present pattern. An alternative is a principal component analysis (PCA) based representation as suggested in [14]. However, if the data space is non-linear, PCA may be suboptimal. Further the application of the PCA on an extremely high dimensional input space may give inappropriate results due to an increasing influence of noise [15]. While the PCA does not take much knowledge into account a more technology aware approach is available by a peak based analysis. Thereby the high dimensional NMR spectra are reduced to its peak lists (line spectra). These peak lists are obtained by application of a peak picking algorithm on the spectra. This algorithm identifies the peaks including attributes such as start/end/center position, width, area, maximal intensity and context information e.g. if the peak belongs to a set of peaks (doublet, triplet, ...) or is a single peak. Here we focus on the spectral peak detection and tracking approach given in [11]. Thereby a rough peak picking is done considering negative and positive slopes of the spectrum. Adapted to NMR we use the peak structure information provided by the DSS signal. The negative magnitude threshold is fixed to 0 and the positive magnitude threshold is defined as the minimal peak height assumed as 10% of the DSS height. Subsequently the identified peaks are further screened using a second analysis step. Thereby for each peak the start and end positions are determined which is done in accordance to the

DSS characteristics. Each peak is limited to a minimal peak width assumed as 20% of the DSS width (DSSw) and a maximal width of 1.5× DSSw. Each peak list of a spectrum constitutes its line spectrum and can be considered as a dense representation of the original signal. While the peak picking approach is promising, e.g. an analysis of noise free data is possible, there are also some problems. First the peak picking becomes complicated in case of stronger noise, partially resolved peaks and overlapping signals, second the quantification of the identified patterns based on the peak lists may be biased due to slightly incorrect start/end positions of the peaks and hiding effects due to overlapping peaks. As a positive point the peak shape becomes less crucial and one gets a natural measure about the safety of a pattern match, considering the number of matched peaks in the pattern. In conclusion, the difference approach always gives a match and concentration value, sensitive to the fit, the PCA method may potentially lead to incorrect results due to the huge number of dimensions or failed constraints, the peak picking on the other hand is promising to achieve most of the requirements but is also sensitive with respect to noise and signals with poor resolution. To overcome some of these problems an alternative is proposed which can be seen as a compromise between all the different approaches.

2.2 Sparse coding for functional data

Sparse coding refers back to the work given in [10], which has shown that images in the primary visual cortex of mammals are encoded by sparse representations of the image data such that a set of sparse codes is obtained. Here we focus on a special variant of sparse coding applied on sets of functional spectral data. This encoding can be seen as a natural compact representation of the data balancing dimensionality reduction and information preservation.

2.2.1 Minimum sparse coding

We suppose that N functional data $\mathbf{f}_k$ are available, each containing D values, i.e. $\mathbf{f}_k \in \mathbb{R}^D$ and $\|\mathbf{f}_k\| = 1$. A set of M, maybe overcomplete and/or not necessarily orthogonal, basis function vectors $\phi_j \in \mathbb{R}^D$ should be used for representation of the data in form of linear combination:

$$\mathbf{f}_k = \sum_j \alpha_{j,k} \cdot \phi_j + \xi_k \tag{1}$$

with $\xi_k \in \mathbb{R}^D$ being the reconstruction error vector and $\alpha_{j,k}$ are the weighting coefficients with $\alpha_{j,k} \in [0,1]$, $\sum_j \alpha_{j,k} = 1$ and $\alpha_k = (\alpha_{1,k}, \ldots, \alpha_{M,k})$. We define a cost function E_k for $\mathbf{f}_k$ as

$$E_k = \|\xi_k\|^2 - \lambda \cdot S_k \tag{2}$$

which has to be minimized. It contains a regularization term S_k. This term judges the sparseness of the representation. It can defined as

$$S_k = \sum_j g\left(\alpha_{j,k}\right) \tag{3}$$

whereby $g(x)$ is a non-linear function like $\exp\left(-x^2\right)$, $\log\left(\frac{1}{1+x^2}\right)$, etc.. Another choice would be to take

$$S_k = H\left(\alpha_k\right) \tag{4}$$

being the entropy of the vector α_k. We remark that minimum sparseness is achieved iff $\alpha_{j,k} = 1$ for exactly one arbitrary j and zero elsewhere. Using this minimum scenario, optimization is reduced to minimization of the description errors $\|\xi_k\|^2$ or equivalently to the optimization of the basis functions ϕ_j. The span for a set of data vectors, consists of vectors ϕ_j chosen as principal components. Minimum principal component analysis requires at least the determination of the first principal component. Taking into account higher components improves the approximation. However, as mentioned above, if the data space is non-linear, principal component analysis (PCA) may be suboptimal. In NMR spectroscopy, one possible way to overcome this problem is to split the data space into continuous patches, building homogenous subsets on these patches and to cary out a PCA on each subset taking only the first principal component. The respective approach to determine the principal component is a combination of adaptive PCA

(Oja-learning, [9]) and prototype-based vector quantization (neural gas [8]) called sparse coding neural gas.

2.2.2 Sparse coding neural gas (SCNG)

We now briefly describe the basic variant of SCNG according to [7]. In SCNG N prototypes $\mathbf{W} = \{\mathbf{w}_i\}$ approximate the first principal component $\mathbf{p}_i$ of the subsets Ω_i. A functional data vector $\mathbf{f}_k$ belongs to Ω_i iff its correlation to $\mathbf{p}_i$ defined by the inner product $O(\mathbf{w}_i, \mathbf{f}_k) = \langle \mathbf{w}_i, \mathbf{f}_k \rangle$ is maximum:

$$\Omega_i = \left\{ \mathbf{f}_k | i = \underset{j}{\operatorname{argmax}} \langle \mathbf{p}_j, \mathbf{f}_k \rangle \right\} \quad (5)$$

The approximations $\mathbf{w}_i$ can be obtained adaptively by Oja-learning starting with random vectors $\mathbf{w}_i$ for time $t = 0$ with $\|\mathbf{w}_i\| = 1$. Let P be the the probability density in Ω_i. Then, for each time step t a data vector $\mathbf{f}_k \in \Omega_i$ is selected according to P and the prototype $\mathbf{w}_i$ is updated by

$$\triangle \mathbf{w}_i = \varepsilon_t O(\mathbf{w}_i, \mathbf{f}_k)(\mathbf{f}_k - O(\mathbf{w}_i, \mathbf{f}_k)\mathbf{w}_i) \quad (6)$$

with $\varepsilon_t > 0$, $\varepsilon_t \underset{t\to\infty}{\to} 0$, $\sum_t \varepsilon_t = \infty$ and $\sum_t \varepsilon_t^2 < \infty$ which is a converging stochastic process [6]. The final limit of the process is $\mathbf{w}_i = \mathbf{p}_i$ [9].

Yet, the subsets Ω_i are initially unknown but requires the knowledge about their first principal components $\mathbf{p}_i$ according to (5). This problem is solved in analogy to the original neural gas in vector quantization [8]. For a randomly selected functional data vector $\mathbf{f}_k$ (according P) for each prototype the correlation $O(\mathbf{w}_i, \mathbf{f}_k)$ is determined and the rank r_i is computed according to

$$r_i(\mathbf{f}_k, \mathbf{W}) = N - \sum_{j=1}^{N} \theta(O(\mathbf{w}_i, \mathbf{f}_k) - O(\mathbf{w}_j, \mathbf{f}_k)) \quad (7)$$

counting the number of pointers $\mathbf{w}_j$ for which the relation $O(\mathbf{w}_i, \mathbf{f}_k) < O(\mathbf{w}_j, \mathbf{f}_k)$ is valid [8]. $\theta(x)$ is the Heaviside-function. All prototypes are updated according to

$$\triangle \mathbf{w}_i = \varepsilon_t h_\sigma(\mathbf{v}, \mathbf{W}, i) O(\mathbf{w}_i, \mathbf{f}_k)(\mathbf{f}_k - O(\mathbf{w}_i, \mathbf{f}_k)\mathbf{w}_i) \quad (8)$$

with

$$h_{\sigma_t}(\mathbf{f}_k, \mathbf{W}, i) = \exp\left(-\frac{r_i(\mathbf{f}_k, \mathbf{W})}{\sigma_t}\right) \quad (9)$$

is the so-called neighborhood function with neighborhood range $\sigma_t > 0$. Thus, the update strength of each prototype is correlated with its matching ability. Further, the temporary data subset $\Omega_i(t)$ for a given prototype is

$$\Omega_i(t) = \left\{ \mathbf{f}_k | i = \underset{j}{\operatorname{argmax}} \langle \mathbf{w}_j, \mathbf{f}_k \rangle \right\} \quad (10)$$

For $t \to \infty$ the range is decreased as $\sigma_t \to 0$ and, hence, only the best matching prototype is updated in (8) in the limit. Then, in the equilibrium of the stochastic process (8) one has $\Omega_i(t) \to \Omega_i$ for a certain subset configuration which is related to the data space shape and the density P [16]. Further, one gets $\mathbf{w}_i = \mathbf{p}_i$ in the limit. Both results are in complete analogy to usual neural gas, because the maximum over inner products is mathematically equivalent to the minimum of the Euclidean distance between the vectors [5],[8].

2.2.3 Classification with Fuzzy Labeled Self Organizing Map

The sparse coded spectra have been fed into a special variant of a self organizing map, called Fuzzy Labeled Self Organizing Map (FL-SOM) as given in [12]. We do not detail FL-SOM here but mention that it generates a classifier and a topological mapping of the data. The parameters of the FL-SOM are: map size 5×10, final neighborhood range 0.75 and with remaining parameters as in [12]. The map has been trained upto convergence as specified in [12]. To obtain the sparse coding on NMR data, the spectra were splitted into 90 so called patches, which are fragments of the NMR signal (see [7]), with a width of 200 points, motivated by the DSS width. For the SCNG algorithm 30 prototypes have been used, determined by a grid search over different values. We would like to mention that the number of prototypes did not significantly influence the results but should be chosen in accordance to

	A	G	Y	S	12	13	14	23	24	34
A	.9					.1				
G		.8								
Y			1					.2		
S				0.9					.2	
12					1					
13						.9				
14							1			
23								.8		
24									.8	
34										1

Table 1. Classification of metabolites using peak lists. Simulated metabolites are almost perfectly recovered (A,G,Y,S) whereas for the unknown mixtures some miss identifications can be observed.

	A	G	Y	S	12	13	14	23	24	34
A	.8									
G		1			1				.4	1
Y			1							
S				.8						
12										
13	.2					1	.7			
14							.3			
23				.2				1		
24									.6	
34										

Table 2. Classification of metabolites using sparse coding. Pure metabolites are almost perfectly recovered (A,G,Y,S) whereas for the unknown mixture data stronger miss identifications are observed.

the diversity of the substructures expected in the overall dataset. The sparse coding got a dimensionality reduction by a factor of ≈ 10.

3 Experiments and Results

Here we compare the peak picking encoding and sparse coding for a set of simulated metabolite spectra. We consider four types of metabolites, relevant in metabolic studies of the stem cell: Alanine (Ala), Glutamine (Gln), Gycine (Gly) and Serine (Ser), simulated at 39 different linear increased concentration levels $(1-39)$. Hence we obtain 156 spectra simulated using the prior mentioned NMR system parameters. Additionally we generated mixtures of these metabolites by combining two metabolites up to all combinations, with 39 concentration levels. This gives 6×39 mixture spectra, which are not used in subsequently training steps but used for external validations. All spectra are processed as mentioned above and either encoded to peak lists or alternatively encoded by sparse coding. The results for the peak based approach are collected in Table 1 and the sparse coding in 2 (Alanine - A, Guanine G, Glycine - Y, Serine - S). Thereby the peak lists of the patterns are directly matched against the peak lists of the measurement using a tolerance of 0.005 ppm.

For the peak lists we observe a very good recognition as well as prediction accuracy. In average the recognition (on the 4 training classes) is 91% and on the unknown 6 mixture classes $\approx 90\%$. It should be noted that the fractions in a column of Table 1 do not necessary accumulate to $1.0 = 100\%$ because, the peak based identification is not forced to identify one of the metabolites in each analysis[1].

The sparse coded data have been analyzed using the FL-SOM and the obtained map (which was topological preserving, topographic error < 0.05). The model has been trained with 4 classes of metabolites. The FL-SOM model generates a fuzzy-labeling of the underlying prototypes and hence is also able to give assignments to more than one class. Using a majority vote scheme to classify the data the training data have been learned with 100% accuracy. But we also wanted to determine the 6 new mixture classes. To do this we defined prototypical labels for each class such as $\{1,0,0,0\}$ for class 1-Alanine and $\{0.5,0.5,0,0\}$ for a mixture class of Alanine and Glycine. Spectra where assigned to the closest prototype and labeled by the label which was closest to the labeling of the data point using Euclidean distance. For example let a data point $\mathbf{v}$ have a fuzzy label $\{1,0,0,0\}$ which assigns it to class 1 or Alanine. Let further be some prototype $\mathbf{w}$ be the winner (closest prototype) for this data point with a fuzzy label of $\{0.6,0.4,0,0\}$. Then two classifications are possible. Using majority vote the prototype label becomes $\{1,0,0,0\}$ and hence the data point is assigned to class 1 - Alanine, which is correct. Using the alternative scheme the fuzzy label $\{0.60.40,0\}$ is closer to $\{0.5,0.5,0,0\}$ then to $\{1,0,0,0\}$ and hence the prototype is labeled as a mixture of alanine and glutamine, consequently the data point is assigned to the 1/2 or Alanine/Glutamine class leading to a (in this case) wrong classification because the data point was labeled as Alanine. Using this scheme and considering the receptive fields

[1] At least 75% of the peaks had to match to count the classification

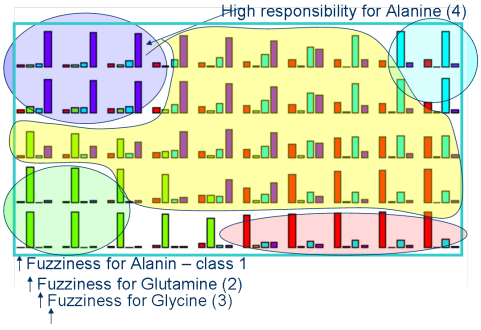

Figure 2. FL-SOM (bar plot) for the 4 metabolite classes. The map is given as a 5×10 grid and for each grid 4 bars are depicted indicating the fuzzy values for the respective classes. As shown in the picture (indicated by manually added closed regions - e.g. ellipsoids in the corners) the map contains receptive field with high responsibility for a single class, but there is also a region in the map responsible for data points which are topologically located between different classes - in our case metabolite mixtures.

of the prototypes of the FL-SOM the picture is a bit different as shown in Table 2. In average the recognition (on the 4 training classes) becomes 87% and on the unknown 6 mixture classes we obtain $\approx 50\%$. However, it should be noted that the used FL-SOM classifier did never see the mixture classes during training but also learned prototypes which are located between different classes in a topology preserved manner. The error for the mixtures Ala/Gln and Gly/Ser are caused due to the fact that no prototype were learned on the map representing these mixtures as depicted in Figure 2.

4 Conclusions

We presented a method for the sparse coded representation of functional data applied in NMR spectroscopy and compared it to an alternative peak based approach. All approaches were able to recognize the plain metabolite spectra at different concentration levels. For the analysis of mixtures the peak picking approach performed better but this result is potentially biased because the simulated data always show a perfect peak shape. For the SCNG approach, we found promising results, the metabolic information encoded in the spectra could be preserved and a significant data reduction by a factor of 10 was achieved. The SCNG provided a sufficient and accurate data reduction such that the FL-SOM classifier method could be used in a topology preserved manner. The SCNG encoding also allows the application of other data analysis methods, such as different classifiers or statistical tests, which need a compact data representation. The SCNG generated a compact and discriminative encoding. Future directions of improvement will focus on a better combination of sparse coded data and the FL-SOM, the additional integration of NMR specific knowledge and an advanced determination of the patches. In a next step all methods will be analyzed on the basis of real NMR metabolite and NMR cell extract measurements. [2].

References

[1] David Chang, Cory D. Banack, and Sirish L. Shah. Robust baseline correction algorithm for signal dense nmr spectra. *Journal of Magnetic Resonance*, 187(2):288–292, 2007.

[2] Li Chen, Zhiqiang Weng an Laiyoong Goh, and Marc Garland. An efficient algorithm for automatic phase correction of nmr spectra based on entropy minimization. *Journal of Magnetic Resonance*, 158(1-2):164–168, 2002.

[3] M. Cross, R. Alt, and D. Niederwieser. The case for a metabolic stem cell niche. *Cells Tissues Organs*, in press, 2008.

[4] S. W. Homans. *A dictionary of concepts in NMR*. Clarendon Press, Oxford, 1992.

[5] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[6] H.J. Kushner and D.S. Clark. *Stochastic Appproximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York, 1978.

[7] K. Labusch, E. Barth, and T. Martinetz. Learning data representations with sparse coding neural gas. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks ESANN*, page in press. d-side publications, 2008.

[8] Thomas M. Martinetz, Stanislav G. Berkovich, and Klaus J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[9] E. Oja. Neural networks, principle components and suspaces. *International Journal of Neural Systems*, 1:61–68, 1989.

[10] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Letters to Nature*, 381:607–609, 1996.

[11] T.H. Park. *Towards Automatic Musical Instrument Timbre Recognition*. PhD thesis, Princeton University, 2004.

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

[13] S.A. Smith, T.O. Levante, B.H. Meier, and R.R. Ernst. Computer simulations in magnetic resonance. an object oriented programming approach. *J. Magn. Reson.*, 106a:75–105, 1994.

[14] Lucksanaporn Tarachiwin, Koichi Ute, Akio Kobayashi, and Eiichiro Fukusaki. 1h nmr based metabolic profiling in the evaluation of japanese green tea quality. *J. Agric. Food Chem.*, 55(23):9330–9336, 2007.

[15] M. Verleysen and D. François. The curse of dimensionality in data mining and time series prediction. In J. Cabestany, A. Prieto, and F. S. Hernández, editors, *Computational Intelligence and Bioinspired Systems, Proceedings of the 8th International Work-Conference on Artificial Neural Networks 2005 (IWANN), Barcelona.*

[16] T. Villmann and J.-C. Claussen. Magnification control in self-organizing maps and neural gas. *Neural Computation*, 18(2):446–469, 2006.

[17] B. Williams, S. Cornett, B. Dawant, A. Crecelius, B. Bodenheimer, and R. Caprioli. An algorithm for baseline correction of maldi mass spectra. In *Proceedings of the 43rd annual Southeast regional conference - Volume 1*, pages 137–142, Kennesaw, Georgia, 2005. ACM.

[2] ACKNOWLEDGMENT: We are grateful to Thomas Riemer IZKF, Leipzig University