2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Mining NMR spectroscopy using Topic Models

Manuele Bicego, Pietro Lovato, Marco De Bona
Department of Computer Science
University of Verona, Verona, Italy

Flavia Guzzo, Michael Assfalg
Department of Biotechnology
University of Verona, Verona, Italy

***Abstract*—Pattern Recognition techniques have been successfully exploited for the biomedical analysis of NMR spectra. In this context, it is crucial to derive a suitable representation for the data: among others, a successful line of research exploits the Bag of Words representation (called here "Bag of Peaks"). However, despite its success, the Bag of Peaks paradigm has not been fully explored: for example, appropriate probabilistic models (such as topic models) can further distill the information contained in the Bag of Words, allowing for more interpretable and accurate solutions for the task-at-hand. This paper is aimed at filling this gap, by investigating the usefulness of topic models in the analysis of NMR spectra. In particular, we first introduce an unsupervised approach, based on topic models, that performs *soft biclustering* of NMR spectra – this kind of unsupervised analysis being new in the NMR literature. Second, we show that descriptors extracted from topic models can be successfully employed for classification of NMR samples: compared to the original Bag of Words, we prove that our descriptors provide higher accuracies. Finally, we perform an empirical evaluation involving a complex dataset of spectra derived from fruits, and two datasets of medical NMR spectra: our analysis confirms the suitability of such models in the NMR spectra analysis.**

I. Introduction

High-resolution proton nuclear magnetic resonance (NMR) spectroscopy [1] represents an invaluable tool in biomedical research, as it provides data on the metabolic composition of biofluids, and conveys information on their perturbations. These data, produced by standard spectrometers, in the simplest form are available as 1-D traces: depending on several biological factors, distinct peak patterns originate from the mixtures of metabolites in different samples. In recent years, Pattern Recognition techniques have become of paramount importance for the automated analysis of NMR spectra [2], [3], allowing for example the classification of diseased from healthy subjects [4], [5], or assisting experts to gain insights into which metabolites have altered concentrations in the biofluids of diseased subjects (usually by performing peak selection / clustering [6]). Several approaches have been proposed in the past (see for example the reviews [7], [2], [3]), each one characterized by different features like speed, accuracy, interpretability and so on. Among others, a successful line of research exploited the Bag of words representation [8], a famous and well known paradigm firstly introduced in the linguistic scenario and subsequently exploited with success in many other different scenarios. In the NMR scenario, the Bag of Words vector representation has been renamed "Bag of Peaks" [9], and has proved to be very suitable: on one hand, classifiers based on this representation achieved state-of-the-art accuracy performances; on the other hand, the representation can be easily interpreted, allowing interaction with experts, for example to fine tune the dictionary or interpret peaks-words [9], [10].

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

However, not all the potentialities of this paradigm have been explored. In particular, it has been shown in many other contexts (see [11], [12], [13] just to cite a few) that the Bag of Words representation can be largely enriched by exploiting probabilistic models such as topic models [14]. This exploitation, in the NMR scenario, is currently missing, and represents the main goal of this paper. In particular we propose to use probabilistic models for NMR spectrometry mining, in order to further distill the information contained in the Bag of Peaks vector. We will focus on the class of topic models [14], which have been specifically designed to model the Bag of Words representation. Even if they have been firstly introduced in the context of text mining, these models have been extensively exported and tailored for a wide variety of applications in Computer Science (see for example the survey in [14]). Their usage is motivated by their expressiveness and efficiency, by the interpretability of the solution provided [15], and by state of the art results achieved in classification tasks.

In this paper we will demonstrate the usefulness of topic models to analyse NMR spectra in two ways. First, we introduce an unsupervised approach, based on topic models, that performs *soft biclustering* of NMR spectra. Please note that this kind of unsupervised analysis is rather new in the NMR analysis scenario. In fact, typically, in NMR spectra analysis / metabolomics a clustering approach is devised i) for clustering *spectra*, in order to unravel complexities of the dataset ([16], [17], [18], [19], [20], just to cite a few, or the recent [21] and references therein), or ii) for clustering *peaks*, in order to discover metabolites or interesting correlated peaks (see [22] – and all references therein – and [21]). On the contrary, our proposed unsupervised approach based on topic models allows the simultaneous grouping of peaks and samples (thus, biclustering), in order to identify the most relevant peaks for each cluster. Moreover it is a soft approach, i.e. it permits to assign each sample/peak to different groups, this being quantified via a membership function.

As a second contribution, we show that descriptors extracted from topic models can be successfully employed for classification of NMR samples: we will prove that the proposed descriptors are better fit for the task than the original Bag of Words representation (on top of which they are built).

Both contributions have been experimentally validated using real datasets, involving a complex dataset of spectra derived from fruits, and two datasets of medical NMR spectra. Obtained results confirm that *i)* topic models can unravel the complexity of the dataset by highlighting different aspects of the analysed data and *ii)* they permit to extract discriminant signatures, able to properly classify NMR spectra.

The rest of the paper is organized as follows: in Sect. II the needed background on Bag of Peaks and Topic Models is introduced; Sect. III then introduces how to exploit topic models for biclustering NMR spectra in a soft fashion. In Sect. IV we will describe how these models can be employed for NMR spectra classification. Finally, Sect. V concludes the paper.

II. Background

In this section we introduce the relevant background on NMR spectra characterization using the Bag of Peaks paradigm [9] and on topic models.

A. Bag of Peaks

The Bag of Peaks representation has been introduced in [9] to characterize NMR spectra on the basis of visible peaks. Briefly, the Bag of Peaks approach works in two steps: dictionary building and spectra characterization.

In the first step (dictionary building), given a training set of NMR spectra, for every trace the most prominent (i.e. highest) peaks are extracted. We can extract a fixed number of such peaks, or all the peaks which are above a given threshold. In all of our experiments, we adopted the first solution. Then we cluster the peak locations into K groups. Different clustering techniques can be used: in [9] the Basic sequential algorithmic scheme (BSAS) approach is used; in our experiments, we employed the Hierarchical Complete Linkage method. The main reason is to avoid fluctuations, due to the random ordering of BSAS (which depends on the initialization, similarly to Kmeans). Finally, the centers of the K clusters represent the words of the dictionary.

In the second step (spectra characterization), given a trace t, a set of most prominent peaks is extracted (typically, using the same strategy used to build the dictionary – i.e. extracting a fixed number of peaks per trace). For every peak p_i, the location and the height (l_i, h_i) are stored. Then, the nearest word (i.e. the nearest peak, in terms of location) in the dictionary is looked for. By making an analogy with the linguistic scenario, where the idea is to count words that appear in a document, we can state that such "word" is present in the trace, with a level of presence (i.e. count) equal to the height h_i. Please note that if two (or more) peaks of the trace are assigned to the same word/peak of the dictionary, the level of presence of such word/peak is the sum of the two (or more) heights. In the end, the trace is characterized by a vector of length K (the size of the dictionary), which in every entry contains the level of presence of a given dictionary word/peak in the trace. Note that such level can be also 0, if the word/peak is not present in the trace.

B. Topic Models

Topic models [14] have been originally introduced in the text analysis community, in order to describe and model a set of documents, represented as Bag of Words vectors. The basic idea underlying these methods is that a document may be characterized by the presence of one or more hidden topics (e.g. sports, finance, politics), each one inducing the presence of some particular words. Here we used the probabilistic Latent Semantic Analysis (pLSA – [23]), one of the first and widely applied models in this family. Given a set of documents, we can learn a pLSA model (via the Expectation Maximization algorithm) in order to obtain two complementary sets of probabilities: $p(z_k|d_t)$ and $p(w_i|z_k)$, where z_k are the different topics, d_t the documents, w_i the words. Intuitively, given a topic z_k, $p(z_k|d_t)$ measures how much the k-th topic is "spoken" in the document d_t, whereas $p(w_i|z_k)$ measures how much the word w_i is used when speaking about the topic z_k.

Due to lack of space, the complete derivation of the pLSA model is not detailed here, and interested readers can refer to [23].

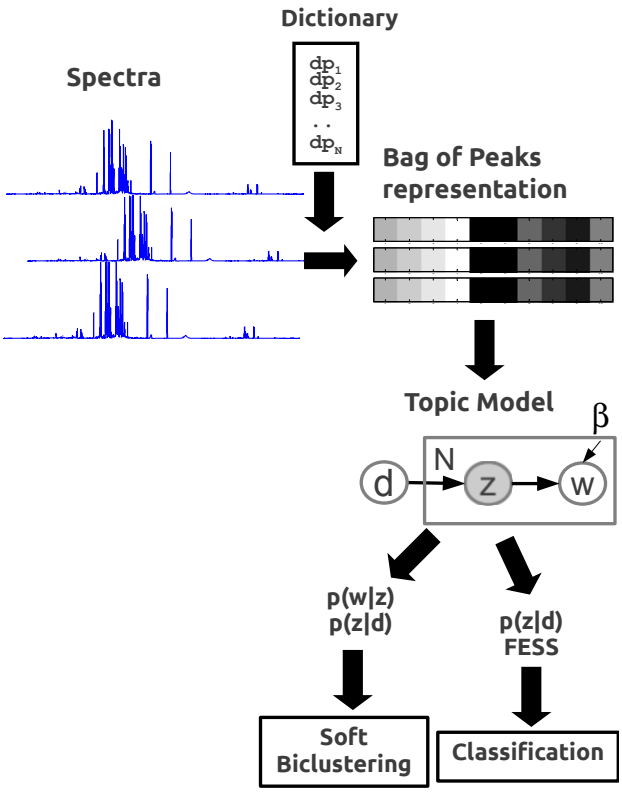

Fig. 1. Scheme of the proposed approach.

III. Topic Models for soft biclustering of NMR spectra

Topic models are built on top of Bag of Words representations [8]. Clearly the Bag of Peaks represents a "Bag of words" representation: in this case a parallelism is established between the linguistic and the NMR scenarios. Reasonably,

it seems appropriate to establish the correspondence between traces/documents and peaks/words. Actually every spectrum (document) is characterized by the different presence of different peaks (the words). To learn a pLSA, documents should be represented with an occurrence matrix (count matrix), where each entry counts the number of times a given word occurs in a given document. This is similar to what the Bag of Peaks representation provides: the level of presence of a given word/peak i in a trace t can be reasonably intended as the count value for that word and document[1]. Given this parallelism, we can learn a pLSA to model the input dataset.

It is important to note that pLSA has been seen recently as a biclustering algorithm [24], able to characterize groups of documents and words (i.e. biclusters). The usage of biclustering algorithms in the NMR analysis scenario is rather new. Typically, clustering algorithms in NMR spectra analysis / metabolomics are employed for clustering *spectra* to unravel complexities of the dataset ([16], [17], [18], [19], [20], just to cite a few, or the recent [21] and references therein) or for clustering *peaks*, in order to discover metabolites or interesting correlated peaks (see [22] – and all references therein – and [21]). pLSA may allow to simultaneously cluster both peaks and samples, therefore permitting a more deep understanding of the information contained in the data.

Following [24], we can say that every topic represents a bicluster, active in different samples and involving the presence of different words. The bicluster memberships are expressed via probabilities, thus permitting a more fine analysis of the spectra. However, there are situations where we need to know which spectra belong to a given bicluster (hard assignment). Even if some strategies to face this problem have been investigated in [24], the problem remains almost open: here we introduce a novel method, based on a statistically sound randomization test, which permits us to get a p-value for a word w_i belonging to a topic z_k (the same reasoning can be applied for a topic z_k "spoken" by document d_t). The main idea is to compare the obtained probability $p(w_i|z_k)$ with random generated ones $p_1(w_i|\tilde{z}), \ldots, p_R(w_i|\tilde{z})$, counting how many times w_i is assigned to topic $\tilde{z}$ with equal or higher probability than the obtained $p(w_i|z_k)$. In other words, we are measuring the chance of obtaining a value $p(w_i|\tilde{z})$ equal to or more extreme than what was actually observed: and this represents exactly the definition of p-value. More specifically, we first notice that the distribution $p(w|z_k)$ is multinomial (by construction of the pLSA model [23]); thus, the random values of $p(w|\tilde{z})$ are sampled from a Dirichlet distribution with uniform parameters, simulating the process of data generation that the pLSA is modeling. In our experiments, the p-value is obtained after 100,000 randomization tests. For more details on how to use pLSA for biclustering please refer to [24].

A. Experimental Results

We tested the proposed biclustering scheme on a complex dataset involving 12 different varieties of cherry, collected in North-East Italy. The list of the samples, together with some other metadata, is reported in table I. A detailed description of orchard area, sampling plan and analytical procedures is reported in [25]. In few words, the dataset includes NMR-based metabolomics data of fruits from three early ripening and nine late ripening sweet cherry cultivars. The fruits were also assessed for the following quality parameters: hardness, degrees Brix (depending on the sugars accumulated within the fruits) and pH (depending on organic acids accumulated within the fruits).

TABLE I
DETAILS OF THE FRUITS DATASET.

Sample	Cultivar	Ripening	Brix	pH	Hardness
1	sandra	early	13.12	3.897	57.757
2	early bigi	early	12.70	3.790	44.367
3	francese	early	13.24	3.792	48.146
4	milanese	late	15.82	3.530	79.770
5	durone rosso	late	16.62	3.660	57.632
6	bella italia	late	16.91	3.610	55.485
7	sandra tardiva	late	18.02	3.600	55.670
8	van	late	17.20	3.740	62.850
9	giorgia	late	15.93	3.650	67.400
10	ferrovia	late	17.78	3.693	60.543
11	kordia	late	17.57	3.697	61.260
12	regina	late	20.05	3.720	60.670

On this dataset we applied the proposed scheme: we extracted 20 peaks from each trace, and we built the Bag of Peaks representation with a dictionary of length 30. On top of this we trained a pLSA model, using 5 topics and stopping the learning after likelihood convergence. In our approach the pLSA model has been trained via a variational EM algorithm, starting from a clever initialization. In particular, it is known that choosing a good initialization for word-topic probabilities is crucial for a proper learning [26]. Typically, this is done at random, with the risk of solution convergence to poor local minima. In order to overcome this issue, following [27] we cluster words into Z groups (where Z represents the number of topics) using the complete link algorithm, which performs an agglomerative clustering. Then, we initialize the topic-word probabilities so that each topic has high probability of generating the words inside its cluster, and low probability of generating words outside the cluster.

The result of the application of our approach is reported in Fig. 2. In particular, for every topic we plot the distribution $p(z|d)$ among all samples; in bold we highlight samples for which the p-value is over 0.01. All the potentialities of the proposed scheme can be deduced if we intersect these plots with the metadata presented in Tab. I: in particular we can observe how the different topics are able to capture different aspects of the proposed dataset. Topic 5 completely characterizes the early ripening cultivars: all significant samples belong to that family; at the same time, it also characterizes samples with the highest pH. In topic 2 all significance samples have high hardness; finally almost all the significant samples of topic 3 have a large Brix.

Different information can also be evinced from the com-

[1]A similar parallelism has been established for gene expression data [24], [12].

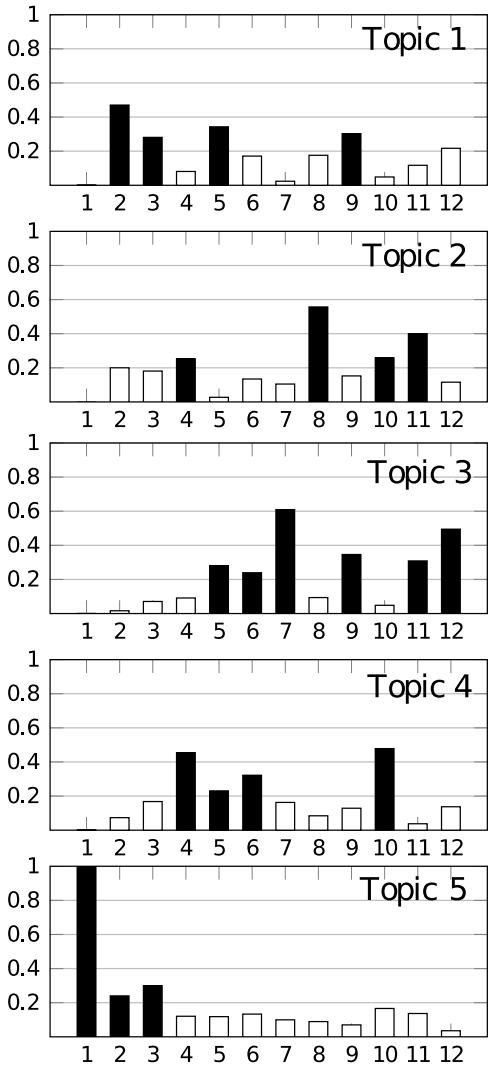

Fig. 2. Relative topic frequencies for distinct samples of the fruit data collection.

plementary analysis of the $p(w|z)$. In particular, considering the dictionary and analysing the most important "words" for the different topics we can observe different things. Actually, the words correspond to a spectral region that shows intense signals from carbohydrates (glucose, fructose and xylose) and free amino acids (aspartate, valine, threonine, alanine, asparagine). Due to the fact that carbohydrates display an elevated number of peaks, their signals are represented in several words, but each word may correspond to accumulated contributions from distinct molecules. Thus, while some words spoken in topic 3 can be attributed exclusively to glucose and xylose, a word spoken in topic 5 has contributions from asparagine and fructose, the majority of words cannot be explained in terms of single or couples of metabolites, but may be associated with several molecules. It is indeed an advantage of the NMR method that individual metabolite peaks need not to be separated while still contributing to the observed spectral profile.

IV. Topic Models for NMR spectra classification

In this section we want to show the potentialities of topic models for classification of NMR spectra. Even if topic models have been introduced for clustering purposes, different strategies has been proposed in recent years to employ them in classification tasks. The main strategy is to employ a hybrid generative-discriminative classification scheme [28], [29], a class of approaches which merge the best characteristics of generative and discriminative paradigms: once learned from data, a generative model (good in describing the problem) is exploited to project every object of the problem in a feature space (often called generative embedding space), in which a discriminative classifier can be used. Here we learned a single pLSA model on the whole training set of spectra (different possibilities are still available – see [30]), performing inference to get distributions on the testing set. Then we employed two schemes to get signatures: the first, called *Topic Proportions*, simply encodes an object by using the probabilities $p(z|d)$ – this represents the first and most used method [11]. As a second scheme we encode the recent Free Energy Score Space (FESS) approach [31]: in few words, – interested readers are referred to [31] – the FESS vector is able to capture how well each object of problem fits the different parts of the generative model, modelled via the variational free energy (which represents a lower bound of the negative log-likelihood). It has been shown in [31] that such representation is highly informative for classification, permitting to reach state-of-the-art results in different bioinformatics and computer vision problems.

A. Experimental Results

Here we tested the classification strategy using two sets of NMR spectra. The first, recently used in [32], derives from a study investigating the urine metabolome of patients affected by immunoglobulin A nephropathy, the most common form of primary glomerulonephritis worldwide. The NMR spectral fingerprints of 24 patients were compared to those of 68 healthy matched controls to verify the occurrence of a urine metabolic signature of the disease. In the paper the classification task is solved by using a Principal Component Analysis plus Correlation Analysis approach. The second dataset concerns the detection of diabetes in children [9]. In particular, the study involved 35 Sardinian under 10-year-old children. The goal was to classify the NMR traces derived from their urine samples in two classes (children having or not Type I diabetes). The NMR traces were characterized by using the above described Bag of Peaks representation, and then classified using different standard classifiers (nearest neighbor, Support Vector Machines and so on).

To be comparable with the two papers we used the same experimental protocol (10 fold cross validation with 100 repetitions for [32] and Leave One Out cross validation for [9]). For [32] we used the same classifier in the representation space (the 5-Nearest Neighbor), whereas for [9] we choose the 1-Nearest Neighbor classifier, since it showed the averaged best performances in such paper. Classification performances

have been assessed by measuring classification accuracies. For what concerns pLSA and Bag of Peaks we performed a thorough evaluation, by varying the number of extracted peaks, the dimension of the dictionary and the number of topics. The pLSA model has been trained as described in the previous section, namely via a variational EM algorithm initialized via clustering.

To get a first glance of the results we reported in table II the classification accuracies obtained by using the best parametrization for every variant of the method on the two datasets, comparing the results with the method proposed in [32] and in [9]. Even if many other methods for NMR spectra classification are present in the literature [2], [3], here we focus on the techniques employed on these datasets, taking baseline results from the papers in which the two datasets have been introduced [32], [9]. For what concerns the results on the first

TABLE II
COMPARATIVE RESULTS: (A) RESULTS ON DATASET OF [32]; (B) RESULTS ON DATASET IN [9].

Method	Accuracy
Bag of Peaks	89.45%
PLSA (Topics Prop.)	92.86%
PLSA (FESS)	92.53%
PCA + Canonical Analysis [32]	88.00%

(a)

Method	Accuracy
Bag of Peaks	93.75%
PLSA (Topics Prop.)	100.00%
PLSA (FESS)	96.875%
PCA (99.9% variance) [9]	84.00%
PCA (Scree Test) [9]	87.00%
Multidimensional Scaling [9]	84.00%

(b)

dataset [32] (table II(a)), we can observe the improvements in the accuracies obtained when employing topic models with respect to the classic Bag of Peaks. It is important to note that the proposed scheme also significantly outperforms the method proposed in [32], suggesting that Bag of Peaks and topic models can be a good alternative to standard schemes. Please note that the accuracy can be increased even more if we adopt a more complex classifier: for example, if we use a linear SVM[2], the accuracies raise to 92.09%, 93.41% and 96.70% for Bag of Peaks, PLSA (Topics Prop.) and PLSA (FESS), respectively. In this case it is also clearer the advantage got when using the FESS classification scheme. Similar observations can be derived from the results of the analysis on the dataset of [9], shown in II(b). Also here it is evident the gain obtained when using the topic models, as well as the improvement over competing standard schemes.

As a second analysis we investigated the relation between Bag of Peaks and topic models when varying the parameters (i.e. number of peaks and dictionary size). Such analysis is reported in Fig. 3: the plots in the left column refer to the number of peaks, whereas those in the right column are related to the size of the dictionary. The first row the analysis is for the dataset in [32], whereas the second is for the dataset of [9]. In such curves we display the best classification accuracy of the Bag of Peaks and the two topic models variants when using a particular value of the analysed parameter. At a first glance, we can note that the methods based on topic models consistently outperform those based on Bag of Peaks for almost all values of the parameters, thus confirming the observations reported above. For what concerns the number of extracted peaks, it seems that the best behavior is obtained when using a number of peaks in the middle of the range. Probably, too few peaks do not contain enough information to characterize the traces, whereas with too many peaks also noisy information is extracted. On the other side, it is more difficult to find a trend in the curves related to the dictionary size. However, it seems that the Bag of Peaks tends to deteriorate when increasing the size of the dictionary. This seems reasonable, since when increasing the dimension of the dictionary we increase the dimensionality of the space where the classifier is trained – thus being more prone to have problems linked to the curse of dimensionality. This is not true for topic models, where the dimensionality of the space does not depend on the size of the dictionary, but on the number of topics.

[2]The parameter C has been set by performing a crossvalidation analysis on the training set.

V. CONCLUSION

In this paper we investigated the usefulness of topic models in the context of NMR spectroscopy. In particular we presented an unsupervised method, based on such models, able to perform soft biclustering of NMR spectra – this kind of unsupervised analysis being new in the NMR literature. Second, we investigated the usefulness of descriptors extracted from topic models for classification of NMR samples. The quantitative and empirical evaluation, involving a complex dataset of spectra derived from fruits, and two datasets of medical NMR spectra, confirms the suitability of such models in the NMR spectra analysis.

ACKNOWLEDGMENTS

M. Bicego and P. Lovato were partially supported by the University of Verona through the program "Bando di Ateneo per la Ricerca di Base 2015".

REFERENCES

[1] R. R. Ernst, G. Bodenhausen, A. Wokaun *et al.*, *Principles of nuclear magnetic resonance in one and two dimensions*. Clarendon Press Oxford, 1987.

[2] A. Smolinska, L. Blanchet, L. Buydens, and S. Wijmenga, "Nmr and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review," *Analytica Chimica Acta*, vol. 750, pp. 82–97, 2012.

[3] A. Alonso, S. Marsal, and A. Julià, "Analytical methods in untargeted metabolomics: State of the art in 2015," *Frontiers in Bioengineering and Biotechnology*, vol. 3, p. 23, 2015.

[4] S. Mahadevan, S. L. Shah, T. J. Marrie, and C. M. Slupsky, "Analysis of metabolomic data using support vector machines," *Analytical Chemistry*, vol. 80, no. 19, pp. 7562–7570, 2008.

[5] S.-F. Chen, H. Gu, M.-Y. Tu, Y.-P. Zhou, and Y.-F. Cui, "Robust variable selection based on bagging classification tree for support vector machine in metabonomic data analysis," *Journal of Chemometrics*, 2017.

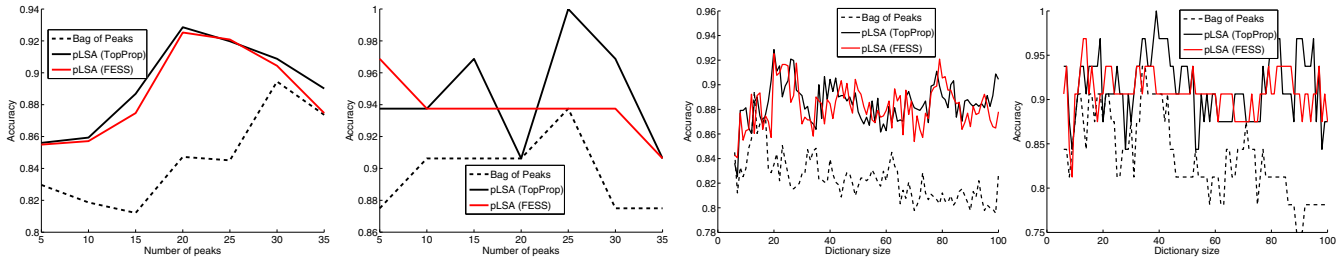

Fig. 3. Comparison between Bag of Peaks and Topic Models when varying the number of extracted peaks (columns 1-2) and the size of the Dictionary (columns 3-4). Columns 1, 3 display results for spectra in [32], whereas Columns 2, 4 are for the spectra in [9].

[6] Y. Cheng, X. Gao, and F. Liang, "Bayesian peak picking for nmr spectra," *Genomics, proteomics & bioinformatics*, vol. 12, no. 1, pp. 39–47, 2014.

[7] J. C. Lindon, E. Holmes, and J. Nicholson, "Pattern recognition methods and applications in biomedical magnetic resonance." *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 39, pp. 1–40, 07 2001.

[8] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[9] G. Brelstaff, M. Bicego, N. Culeddu, and M. Chessa, "Bag of peaks: interpretation of nmr spectrometry," *Bioinformatics*, vol. 25, no. 2, pp. 258–264, 2009.

[10] S. Stanciu, D. Tranca, G. Stanciu, R. Hristu, and J. Bueno, "Perspectives on combining nonlinear laser scanning microscopy and bag-of-features data classification strategies for automated disease diagnostics," *Optical and Quantum Electronics*, vol. 48, p. 320, 2016.

[11] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Proc. of European Conf. on Computer Vision*, vol. 3954. Springer, 2006, pp. 517–530.

[12] M. Bicego, P. Lovato, A. Perina, M. Fasoli, M. Delledonne, M. Pezzotti, A. Polverari, and V. Murino, "Investigating topic models' capabilities in expression microarray data classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 6, pp. 1831–1836, 2012.

[13] S. Kim, S. Sundaram, P. G. Georgiou, and S. Narayanan, "Audio Scene Understanding using Topic Models," in *Proceedings of the Neural Information Processing Systems (NIPS) Workshop*, 2009.

[14] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[15] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.

[16] G. Pierens, M. Palframan, C. Tranter, A. Carroll, and R. Quinn, "A robust clustering approach for nmr spectra of natural product extracts," *Magn Reson Chem.*, vol. 43, no. 5, pp. 359–365, 2005.

[17] M. Cuperlovic-Culf, N. Belacel, A. Culf, I. Chute, R. Ouellette, I. Burton, T. Karakach, and J. Walter, "Nmr metabolic analysis of samples using fuzzy k-means clustering," *Magn Reson Chem*, vol. 47 Suppl 1, pp. S96–104, 2009.

[18] E. Beckonert, M. Bollard, T. Ebbels, H. Keun, and H. A. et al, "Nmr-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches," *Analytica Chimica Acta*, p. 315, 2003.

[19] J. Hageman, R. V. D. Berg, J. Westerhuis, H. Hoefsloot, and A. Smilde, "Bagged k-means clustering of metabolome data," *Critical reviews in analytical chemistry*, vol. 36, p. 211220, 2006.

[20] X. Li, X. Lu, J. Tian, P. Gao, and e. a. H. Kong H, "Application of fuzzy c-means clustering in data analysis of metabolomics," *Analytical Chemistry*, vol. 81, pp. 4468–4475, 2009.

[21] X. Zou, E. Holmes, J. Nicholson, and R. Loo, "Statistical homogeneous cluster spectroscopy (shocsy): An optimized statistical approach for clustering of 1h nmr spectral data to reduce interference and enhance robust biomarkers selection," *Analytical Chemistry*, vol. 86, no. 11, p. 53085315, 2014.

[22] S. Robinette, K. Veselkov, E. Bohus, M. Coen, H. Keun, T. Ebbels, O. Beckonert, E. Holmes, J. Lindon, and J. Nicholson, "Cluster analysis statistical spectroscopy using nuclear magnetic resonance generated metabolic data sets from perturbed biological systems," *Analytical Chemistry*, vol. 81, no. 16, pp. 6581–6589, 2009.

[23] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1–2, pp. 177–196, 2001.

[24] M. Bicego, P. Lovato, A. Ferrarini, and M. Delledonne, "Biclustering of expression microarray data with topic models," in *2010 20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 2728–2731.

[25] M. Commisso, M. Bianconi, F. Di Carlo, S. Poletti, A. Bulgarini, F. Munari, S. Negri, M. Stocchero, S. Ceoldo, L. Avesani, M. Assfalg, G. Zoccatelli, and F. Guzzo, "Multi-approach metabolomics analysis and artificial simplified phytocomplexes reveal cultivar-dependent synergy between polyphenols and ascorbic acid in fruits of the sweet cherry (prunus avium l.)," *PLoS ONE*, vol. 12, no. 7, p. e0180889, 2017.

[26] P. Lovato, M. Bicego, V. Murino, and A. Perina, "Robust initialization for learning latent dirichlet allocation," in *Proc. Int. Workshop on Similarity-Based Pattern Analysis and Recognition (SIMBAD2015)*, 2015, pp. 117–132.

[27] P. Lovato, A. Giorgetti, and M. Bicego, "A multimodal approach for protein remote homology detection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 12, no. 5, pp. 1193–1198, 2015.

[28] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems*, 1999, pp. 487–493.

[29] J. Lasserre, C. Bishop, and T. Minka, "Principled hybrids of generative and discriminative models," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 87 – 94.

[30] M. Bicego, M. Cristani, V. Murino, E. Pekalska, and R. Duin, "Clustering-based construction of hidden Markov models for generative kernels," in *Proc. Int. Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2009, pp. 466–479.

[31] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic, "Free energy score spaces: Using generative information in discriminative classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1249–1262, 2012.

[32] L. Del Coco, M. Assfalg, M. DOnofrio, F. Sallustio, F. Pesce, F. Fanizzi, and F. Schena, "A proton nuclear magnetic resonance-based metabolomic approach in iga nephropathy urinary profiles," *Metabolomics*, vol. 9, pp. 740–751, 2013.