Learning document representations using subspace multinomial model

Santosh Kesiraju[1,2], *Lukáš Burget*[1], *Igor Szőke*[1], *Jan "Honza" Černocký*[1]

[1] Brno University of Technology, Speech@FIT and
IT4I Center of Excellence, Brno, Czech Republic
[2] International Institute of Information Technology - Hyderabad, India

qkesiraju@stud.fit.vutbr.cz, {burget, szoke, cernocky}@fit.vutbr.cz

Abstract

Subspace multinomial model (SMM) is a log-linear model and can be used for learning low dimensional continuous representation for discrete data. SMM and its variants have been used for speaker verification based on prosodic features and phonotactic language recognition. In this paper, we propose a new variant of SMM that introduces sparsity and call the resulting model as ℓ_1 SMM. We show that ℓ_1 SMM can be used for learning document representations that are helpful in topic identification or classification and clustering tasks. Our experiments in document classification show that SMM achieves comparable results to models such as latent Dirichlet allocation and sparse topical coding, while having a useful property that the resulting document vectors are Gaussian distributed.

Index Terms: Document representation, subspace modelling, topic identification, latent topic discovery

1. Introduction

Learning document representations that elicit the underlying semantics (or topics) is essential for tasks such as document classification, topic identification, query based document retrieval. In this paper, we propose to use subspace multinomial model (SMM) [1], for obtaining a compact and continuous representation for the documents, which are further used for document classification and clustering tasks. By using evaluation metrics like classification accuracy and normalized mutual information (NMI), we show that the obtained representations can be useful for the aforementioned tasks. SMM was first proposed for speaker verification based on prosodic features [1]. It was later used for phonotactic language recognition (LRE), where discrete n-gram counts of an utterance are assumed to be generated by a specific multinomial (or n-gram) distribution, and the parameters of these distributions are modelled using subspace techniques [2].

Document clustering or latent topic discovery has been studied for more than two decades. In most of the approaches, the first step is to represent a document in the form of vector, where every element corresponds to the weighted frequency of a word in the vocabulary as occurred in a document. Word selection can be done to decide the size of the vocabulary, which will essentially eliminate the stop words and other irrelevant words that do not significantly contribute to the semantics of the document [3]. This allows us to obtain a fixed dimensional representation for every document, which further enables to perform clustering or train classifiers. Even with careful word selection, the document vectors are usually sparse as many words are not shared across all the documents. To overcome this, latent variables are introduced, which are in a much lower dimension than the actual dimension of the document vectors. In the latent space, it is possible to compute similarities between documents even if they do not share common words. The latent space also allows us to project both the documents and words into the same space. Popular approaches include latent semantic analysis (LSA) [4], non-negative matrix factorization (NMF) [5], probabilistic LSA (PLSA) [6], and latent Dirichlet allocation (LDA) [7].

LSA and NMF use matrix factorization to obtain the latent space while minimizing the reconstruction error. PLSA and LDA come under probabilistic topic models (PTM), and are seen as generative models for documents [8]; where every document is modelled as a distribution over topics (latent variables) and every topic is modelled as a distribution over vocabulary.

It was argued that sparsity in the semantic space is desired in text modelling [9, 10], and NLP taks [11]. Sparsity inducing topic model was proposed in [10], where the authors show that it is possible to obtain sparse representation for document codes and word codes. In [11], sparse log-linear models were used to learn a small and useful feature space for dialogue-act classification.

In the proposed SMM, we obtain a low dimensional continuous representation for every document while the objective function is based on maximizing the likelihood of the observed data, which is similar to that of PTM. In PTM, the latent vectors have probabilistic interpretation (points on simplex), whereas in SMM the latent vectors are Gaussian-like distributed [2], that helps in clustering and training classifiers. Similar idea was proposed in [12], where subspace technique was used to obtain compact representation of multiple topic spaces learned from LDA. The technique used in [12] is different from SMM, as the former was based on subspace Gaussian mixture models which is proposed for modelling continuous data, where as the SMM acts directly on the observed word counts (discrete data).

2. Subspace multinomial model

Let D be the number of document in the collection, with d representing the document index and V be the total number of words in the vocabulary, with i representing the word index. Every word in a document can be seen as an independent event (bag-of-words) generated by a document specific multinomial distribution, and let $\mathbf{c}_d$ denote the vector of word occurrences in document d. We now describe the basic subspace multinomial model [1].

2.1. Basic SMM

Let ϕ_{di} be the parameters (word probabilities) of the document specific multinomial distribution, and the corresponding log likelihood of the document can be written as

$$\log(P(\mathbf{c}_d \mid \boldsymbol{\phi}_d)) = \sum_{i=1}^{V} c_{di} \ \log(\phi_{di}) \,, \tag{1}$$

where $\sum_i \phi_{di} = 1, \phi_{di} \geq 0$ and c_{di} is the count of word i in document d. The parameter ϕ_{di} of the multinomial distribution, which belongs to the exponential family, can be re-parameterized using the *natural parameters* (η) [13] as

$$\phi_{di} = \frac{\exp(\eta_{di})}{\sum_i \exp(\eta_{di})} \,, \tag{2}$$

which is also known as the $\mathrm{softmax}$ function. The subspace model assumes that these *natural parameters* live in a much smaller space, and can be expressed as

$$\boldsymbol{\eta}_d = \mathbf{m} + \mathbf{T}\,\mathbf{w}_d \,. \tag{3}$$

Here $\mathbf{w}_d \in \mathbb{R}^K$ is the document specific latent vector, also known as the i-vector, $\mathbf{T} \in \mathbb{R}^{V \times K}$ is known as the total variability matrix (bases matrix) which spans a linear subspace in log-probability domain, and $\mathbf{m} \in \mathbb{R}^V$ can be seen as a vector of offset or bias.

The model parameters, $\mathbf{w}$ is initialized to zeros, $\mathbf{T}$ with small random values, and $\mathbf{m}$ with log of probabilities of words as estimated from the training set (this can be seen as an average distribution over the entire training set). The parameters $\mathbf{w}$ and $\mathbf{T}$ are updated iteratively and alternately ($\mathbf{m}$ is not updated in our experiments) by using Newton-Raphson like update steps that maximizes the joint log-likelihood of all the documents,

$$\mathcal{L} = \sum_{d=1}^{D} \ \sum_{i=1}^{V} c_{di} \log(\phi_{di}) \,. \tag{4}$$

The update equations take the following form [1]:

$$\mathbf{w}_d^{\text{new}} = \mathbf{w}_d + \mathbf{H}_d^{-1} \ \nabla \mathbf{w}_d \,, \tag{5}$$

$$\mathbf{t}_i^{\text{new}} = \mathbf{t}_i \ \ + \mathbf{H}_i^{-1} \ \nabla \mathbf{t}_i \,. \tag{6}$$

Here $\mathbf{t}_i$ is the i$^{\text{th}}$ row in $\mathbf{T}$. $\nabla \mathbf{w}_d$ and $\nabla \mathbf{t}_i$ are the gradients with respect to the objective function in Eq. (4). Since the parameters (rows in $\mathbf{T}$ and every document i-vector) are updated independently, the corresponding $\mathbf{H}$ matrices ($\mathbf{H}_i$ and $\mathbf{H}_d$) are much smaller and faster to compute [14], as compared to the conventional full Hessian matrix in Newton-Raphson optimization [13].

2.2. Limitations

In the document collection, the most frequently occurring words are the stop words which do not have the ability to semantically discriminate the documents. So, when using full vocabulary including the stop words, the number of parameters in the model increases, which could lead to over-fitting. To over come this, the model can be regularized. A variant of SMM (subspace n-gram model) was proposed for language recognition in [2], where the authors used ℓ_2 regularized model, which could be interpreted as MAP point estimates of the parameters with Gaussian prior. It was observed in [2] that, the i-vectors ($\mathbf{w}$) exhibit Gaussian-like distribution across various dimensions, and the rows in $\mathbf{T}$ exhibit Laplace-like distribution (which does not comply with Gaussian prior assumption). It was also suggested in [2], that ℓ_1 regularization could be applied for $\mathbf{T}$, which could be interpreted as MAP point estimate with Laplace prior. Following this, we propose to regularize $\mathbf{T}$ with ℓ_1 and $\mathbf{w}$ with ℓ_2, and call the resulting model as ℓ_1 SMM.

2.3. ℓ_1 SMM

The objective function from Eq. (4) becomes,

$$\mathcal{L} = \sum_{d=1}^{D} \ \sum_{i=1}^{V} \Big(c_{di} \log(\phi_{di}) - \gamma \|\mathbf{t}_i\|_1 - \frac{\lambda}{2} \|\mathbf{w}_d\|_2 \Big) \tag{7}$$

Here γ and λ are the regularization coefficients for $\mathbf{t}$ and $\mathbf{w}$ respectively. It is essential to regularize both $\mathbf{t}$ and $\mathbf{w}$. Otherwise, restricting the magnitude of one parameter will be compensated by dynamic range increase in the other, during the iterative update steps (Eq. (5) and (6)).

Estimating the parameters of any ℓ_1 regularized function is not trivial, as it introduces discontinuities at points where the function is crossing the axis. To address this, several optimization techniques were proposed [15, 16]. One of such techniques, called as orthant-wise learning is explored in our work, as it could be translated in a straightforward way to our optimization scheme (Eq. (6)).

Orthant is a region in the n-dimensional space where the sign of the variables does not change. It is equivalent to quadrant in 2D and octant in 3D. The important property of any ℓ_1 regularized function is that it is differentiable over any given orthant. In general, for any ℓ_1 regularized convex objective function, if the initial point is in the same orthant as the minimum, then the simple Newton-Raphson updates will lead to the minimum. In cases where we need to cross the orthant to find the minimum, orthant-wise learning can be adopted [16].

2.4. Parameter estimation using orthant-wise learning

The gradient of $\mathbf{t}_i$ with respect to the function in Eq. (7) is given by

$$\nabla \mathbf{t}_i = \sum_{d=1}^{D} \Big(c_{di} - \phi_{di}^{\text{old}} \sum_{i=1}^{V} c_{di} \Big) \mathbf{w}_d^T - \ \gamma \ \mathrm{sign}(\mathbf{t}_i) \,. \tag{8}$$

Here sign is the element-wise sign operation on the vector $\mathbf{t}_i$. At co-ordinates where the objective function is not differentiable (i.e., when any of the co-ordinates k in $\mathbf{t}_i \in \mathbb{R}^K$ equals to 0), we compute the pseudo-gradient $\tilde{\nabla} \mathbf{t}_i$.

$$\tilde{\nabla} t_{ik} \triangleq \begin{cases} \nabla t_{ik} + \gamma, & t_{ik} = 0, \ \nabla t_{ik} < -\gamma \\ \nabla t_{ik} - \gamma, & t_{ik} = 0, \ \nabla t_{ik} > \gamma \\ \quad 0, & t_{ik} = 0, \ |\nabla t_{ik}| \leq \gamma \\ \ \ \nabla t_{ik}, & |t_{ik}| > 0 \ . \end{cases} \tag{9}$$

Otherwise, $\tilde{\nabla} \mathbf{t}_i = \nabla \mathbf{t}_i$. For the updates following Newton-Raphson like method, we need to ensure two things: (i) We need to find the ascent direction $\mathbf{d}$, which leads us into the correct orthant, and, (ii) a step in the ascent direction should not cross the point of non-differentiability. In general, the search direction $\mathbf{d} \in \mathbb{R}^K$ will be of the form,

$$\mathbf{d}_i \triangleq \mathbf{H}_i^{-1} \tilde{\nabla} \mathbf{t}_i \tag{10}$$

To ensure that the new updates ($\mathbf{t}_i^{\text{new}}$) are along ascent direction ($\mathbf{d}_i \tilde{\nabla} \mathbf{t}_i > 0$), the co-ordinates in the search direction $\mathbf{d}_i$ are set to zero, if the sign does not match with the co-ordinates in the steepest ascent $\tilde{\nabla}\mathbf{t}_i$. This sign projection is denoted by $\mathcal{P}_{\mathcal{S}}$:

$$\mathcal{P}_{\mathcal{S}}(\mathbf{d})_i \triangleq \begin{cases} d_{ik}, & \text{if} \quad d_{ik}(\tilde{\nabla} t_{ik}) > 0\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{11}$$

To ensure that the step does not cross the point of non differentiability, we apply the following orthant projection:

$$\mathcal{P}_{\mathcal{O}}(\mathbf{t}+\mathbf{d})_i \triangleq \begin{cases} 0 & \text{if} \quad t_{ik}(t_{ik}+d_{ik}) < 0\,, \\ t_{ik}+d_{ik} & \text{otherwise}\,. \end{cases} \tag{12}$$

This orthant projection will set the co-ordinates in $\mathbf{t}_i^{\text{new}}$ to zero, if they differ in sign with $\mathbf{t}_i$. Finally, the update for $\mathbf{t}_i$ is given as follows:

$$\mathbf{t}_i^{\text{new}} = \mathcal{P}_{\mathcal{O}}[\mathbf{t}_i + \, \mathcal{P}_{\mathcal{S}}[\mathbf{H}_i^{-1}\tilde{\nabla}\mathbf{t}_i]]\,. \tag{13}$$

Here $\mathbf{H}_i \in \mathbb{R}^{K\times K}$ is computed as follows:

$$\mathbf{H}_i = -\left(\sum_{d=1}^{D} \max\left(c_{di}, \phi_{di}^{\text{old}} \sum_{i=1}^{V} c_{di}\right)\right) \mathbf{w}_d \mathbf{w}_d^T\,. \tag{14}$$

The updates for $\mathbf{w}_d$ are according to Eq. (5), with the following gradient:

$$\nabla \mathbf{w}_d = \sum_{i=1}^{V} \mathbf{t}_i^T (c_{di} - \, \phi_{di}^{\text{old}} \sum_{i=1}^{V} c_{di}) - \lambda \mathbf{w}_d\,. \tag{15}$$

The $\mathbf{H}_d \in \mathbb{R}^{K\times K}$ for updating $\mathbf{w}_d$ is given as follows:

$$\mathbf{H}_d = -\sum_{i=1}^{V} \mathbf{t}_i^T \mathbf{t}_i \, \max\left(c_{di}, \, \phi_{di}^{\text{old}} \sum_{i=1}^{V} c_{di}\right) \, - \lambda \mathbf{I}\,. \tag{16}$$

More details on estimating the $\mathbf{H}$ matrices are given in [14].

If the updates of $\mathbf{T}$ and $\mathbf{w}$ fail to increase the objective function in Eq. (7), we keep backtracking by halving the update step. Typically the model converges after 15 to 20 iterations. Once the model is trained, the i-vectors $\mathbf{w}_d$ for every document d are extracted by keeping the $\mathbf{T}$ fixed and using updates in Eq. (5) that maximize the objective function. The i-vectors are extracted for both the training and test datasets and takes 3 to 5 iterations to converge.

3. Experiments

The experiments were conducted on the 20 newsgroups dataset as it is well-studied with several benchmarking baseline systems. We have used the `20-news-bydate` version as used in [10], which contains 18775 documents in 20 categories, with a total vocabulary of 61188 words. The training set consists of 11269 documents with 53975 unique words and the test set consists of 7505 documents.

3.1. Document classification

Since the document vectors (i-vectors) exhibit Gaussian-like distribution, we have used linear Gaussian classifier, where every class has a specific mean and the covariance matrix is shared [13]. The classification accuracy on the test set for ℓ_1 and ℓ_2 SMM for various values of γ (regularization coefficient of $\mathbf{T}$) and i-vector dimensions are shown in Fig. 1. For the purpose of illustration, we have fixed the value of λ (regularization coefficient of i-vectors) at 10^{-4}. We also give a comparison with LDA, Discriminative LDA (DiscLDA) [17], sparse topical coding (STC) and max-margin supervised STC (MedSTC) [10] in Table 1, along with the corresponding latent variable dimension for which the classification accuracy is reported to be maximum [10, 17]. Detailed comparison of STC and its variants along with various other models can be found in [10]. It is important to note that DiscLDA and MedSTC which achieve better classification results are supervised models i.e., topic label information is incorporated while obtaining the latent vector representation; whereas their counterparts, LDA and STC are completely unsupervised models like SMM.

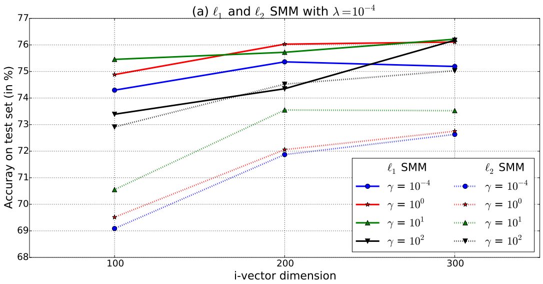

Figure 1: Classification accuracy of ℓ_1 and ℓ_2 SMM for various values of γ, and i-vector dimensions.

Table 1: Comparison of classification accuracy (in %) across various models for the best latent variable dimension (K).

Model	LDA	Disc LDA	STC	Med STC	ℓ_2 SMM	ℓ_1 SMM	
%	75.0	80.0	74.0	81.0	74.95	**75.46**	**78.84**
K	110	110	90	100	100	100	1000

3.2. Document clustering

Clustering is not implicit in SMM as in NMF [5] and the latent vector (i-vector) dimension doesn't necessarily correspond to the number of latent topics like PTM. Since, the i-vectors exhibit Gaussian-like distribution, clustering techniques like k-means or Gaussian mixture models can be used. In our experiments, we used k-means to obtain the hard clusters. The clustering was performed on the entire dataset (training + test), while keeping the subspace trained only on the training set (to maintain consistency with the classification experiments). The resulting clusters are evaluated using NMI [3], and the average (over 5 runs) scores are shown in Table 2. Here the number of clusters in k-means are fixed to 20 (same as the actual number of classes).

In [18], a model which is a mixture of LDAs (multi-grain clustering topic model, MGCTM) was proposed that integrates topic modelling with document clustering. In Table 3, we give the comparison of the proposed SMM with other techniques as reported in [18]. Here the hyper-parameters (including latent vector dimension) are tuned to achieve best clustering performance, and document specific parameters of MGCTM are initialized using LDA. It can be seen that the proposed SMM per-

Table 2: *Comparison of average NMI scores of ℓ_1 and ℓ_2 SMM for various values of γ, $\lambda = 10^{-4}$, ivector-dimension$=$ 100 and no. of clusters $=$ 20.*

γ	10^{-4}	10^{0}	10^{1}	10^{2}	10^{3}
ℓ_2 SMM	0.50	0.49	0.49	0.50	0.52
ℓ_1 SMM	0.56	0.57	0.58	0.58	0.45
Sparsity (%)	0.2	3.6	22.0	53.5	78.8

Table 3: *Comparison of average NMI scores of other systems with ℓ_1 and ℓ_2 SMM for $\gamma = 10^1$, $\lambda = 10^{-4}$, ivector-dimension$=$ 100 and no. of clusters $=$ 20.*

Method	ℓ_2 SMM	ℓ_1 SMM	LDA	NMF	MGCTM
NMI	**0.52**	**0.58**	0.48	0.36	0.61

forms better at clustering and classification at the same time with the same model (i.e., with the same model parameters including i-vector dimension). More experiments on various datasets with analysis on STC, LDA, MGCTM and SMM are left to future work.

4. Discussion

In our experiments, we have observed that the classification accuracy of SMM increases with the increasing dimensionality of the latent variable (i-vector) which is not the case with STC or PTM [10]. Further, we achieved **78.84** classification accuracy on the test set for ℓ_1 SMM with i-vector dimension=1000.

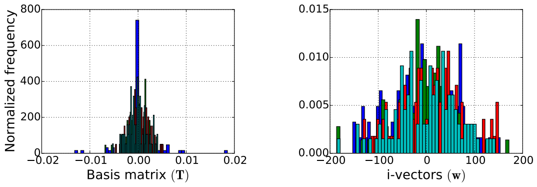

Figure 2: *Histograms showing the distribution of values in rows of bases matrix $\mathbf{T}$, and i-vectors.*

Fig. 2 shows an example histogram of 5 randomly selected rows in the bases $\mathbf{T}$ with $\gamma = 10.0$, and the i-vectors with $\lambda = 10^{-4}$ and $K = 100$, after the training phase. The distribution of the bases matrix in Fig. 2 suggests that Laplace prior is suitable for $\mathbf{T}$. The Laplace prior enforces sparsity in the bases matrix ($\mathbf{T}$), which suggests that feature (word) selection is implicit. Table 2 gives the percentage of sparsity in $\mathbf{T}$ for ℓ_1 SMM for various values of γ, along with the NMI scores. It can be observed that ℓ_1 SMM performs better than ℓ_2 SMM across various values of γ, reinforcing the suitability of Laplace prior over the bases matrix $\mathbf{T}$.

In an unsupervised scenario, it is possible to obtain a set of words for each cluster that represent or discriminate it from other clusters. One way is to subtract the global mean from the cluster mean of i-vectors ($\mathbf{w}_d$) and project the resulting vector on to the bases matrix $\mathbf{T}$ and find the indices of large positive values. These indices corresponds to the words for which the probabilities significantly increase as compared to the average distribution over words for the given cluster. Table 4 shows an example of words from all the 20 clusters obtained using k-means for ℓ_1 SMM with $\lambda = 10^{-4}, \gamma = 10^1$ and i-vector dimension (K) = 100.

Table 4: *Top 5 significant words from all 20 clusters.*

acceleration	preferably	scotia	autoexec
suspension	architecture	sluggo	windows
wagon	databases	compuserv	exe
tires	publisher	nursery	icons
chevy	blvd	pruden	ini
xlib	sale	waco	sacred
widget	packaging	atf	worship
parameter	obo	fbi	christianity
openwindows	shipping	convicted	atheist
xview	cod	koresh	prophet
physicians	income	murders	privacy
patients	socialism	criminals	encryption
therapy	abortion	firearm	denning
infection	welfare	handguns	clipper
diagnosed	cramer	criminal	crypto
rbi	israeli	compute	spacecraft
dodgers	lebanon	algorithms	lunar
hitters	occupied	polygon	moon
pitcher	palestinians	shareware	exploration
pitching	palestinian	surfaces	orbit
resistor	hockey	nubus	zx
amplifier	potvin	quadra	bikes
resistors	leafs	meg	motorcycle
volt	nhl	slots	riding
voltage	playoff	adapter	bike

5. Conclusions and future work

In this paper, we have proposed a new variant of subspace multinomial model called ℓ_1 SMM its application to topic identification and document clustering. We have shown that it is possible to introduce sparsity in the semantic space (bases matrix), while retaining the useful property of the document vectors to be Gaussian distributed. Having such a distribution, helped in using simple classifiers and clustering techniques, rather than relying on sophisticated models for each of the tasks.

By applying optimization techniques, we have shown how the ℓ_1 SMM could be trained. There are many optimization techniques for ℓ_1 regularized objective functions that could be explored [16]. Our future work involves, exploring discriminative SMM and fully Bayesian modelling of SMM.

6. Acknowledgements

This work was supported by the U.S. DARPA LORELEI contract No. HR0011-15-C-0115. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The work was also supported by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602". Santosh Kesiraju is partly funded by TCS Research Fellowship program, India.

7. References

[1] M. Kockmann, L. Burget *et al.*, “Prosodic speaker verification using subspace multinomial models with intersession compensation,” in *INTERSPEECH, ISCA*, September 2010, pp. 1061–1064.

[2] M. Soufifar, L. Burget, O. Plchot *et al.*, “Regularized Subspace n-Gram Model for Phonotactic iVector Extraction,” in *INTERSPEECH, ISCA*, August 2013, pp. 74–78.

[3] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[4] S. Deerwester, S. T. Dumais *et al.*, “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[5] W. Xu, X. Liu, and Y. Gong, “Document Clustering Based on Non-negative Matrix Factorization,” in *SIGIR*. New York, USA: ACM, 2003, pp. 267–273.

[6] T. Hofmann, “Probabilistic Latent Semantic Indexing,” in *SIGIR*. New York, USA: ACM, 1999, pp. 50–57.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *JMLR*, vol. 3, pp. 993–1022, March 2003.

[8] M. Steyvers and T. Griffiths, “Probabilistic Topic Models,” *Handbook of Latent Semantic Analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[9] M. V. S. Shashanka, B. Raj, and P. Smaragdis, “Sparse Overcomplete Latent Variable Decomposition of Counts Data,” in *NIPS*, December 2007, pp. 1313–1320.

[10] J. Zhu and E. P. Xing, “Sparse Topical Coding,” in *Proceedings of the 27th Conference on UAI*, July 2011, pp. 831–838.

[11] Y. Chen, W. Y. Wang, and A. I. Rudnicky, “An empirical investigation of sparse log-linear models for improved dialogue act classification,” in *IEEE ICASSP*, May 2013, pp. 8317–8321.

[12] M. Morchid, M. Bouallegue *et al.*, “An I-vector Based Approach to Compact Multi-Granularity Topic Spaces Representation of Textual Documents,” in *EMNLP*, October 2014, pp. 443–454.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[14] D. Povey, L. Burget *et al.*, “The Subspace Gaussian Mixture model-A Structured Model for Speech Recognition,” *Comput. Speech Lang.*, vol. 25, no. 2, pp. 404–439, Apr. 2011.

[15] G. Andrew and J. Gao, “Scalable Training of L1-Regularized Log-Linear Models,” in *ICML*. New York, USA: ACM, 2007, pp. 33–40.

[16] M. Schmidt, “Graphical Model Structure Learning with ℓ_1 Regularization,” Ph.D. dissertation, The University of British Columbia, August 2010.

[17] S. Lacoste-Julien, F. Sha, and M. I. Jordan, “DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification,” in *NIPS*, 2009, pp. 897–904.

[18] P. Xie and E. P. Xing, “Integrating Document Clustering and Topic Modeling,” in *UAI*, August 2013.