ISSN 1064-5624, Doklady Mathematics, 2014, Vol. 89, No. 3, pp. 301–304. © Pleiades Publishing, Ltd., 2014.
Original Russian Text © K.V. Vorontsov, 2014, published in Doklady Akademii Nauk, 2014, Vol. 456, No. 3, pp. 268–271.

MATHEMATICS

Additive Regularization for Topic Models of Text Collections

K. V. Vorontsov[a, b, c]

Presented by Academician Yu.I. Zhuravlev September 3, 2013

Received September 5, 2013

DOI: 10.1134/S1064562414020185

Topic modeling is an actively developing field in the statistical analysis of texts [1]. A probabilistic topic model identifies the topic of a text collection, describing each topic by a discrete distribution over a set of words and each document by a discrete distribution over a set of topics. Topic models are used for information search, classification, categorization, annotation, and summarizing of texts.

Suppose that we are given three finite sets: a text collection D, a vocabulary W of words, and a set T of topics. It is assumed that the order of words in the documents is of no matter and the collection is a random sample from a discrete distribution $p(d, w, t)$ on $D \times W \times T$. The variables d and w are observable, while t is latent, i.e., the occurrence of each pair (d, w) is associated with some unknown topic t. The text collection is represented by the frequency matrix $F = (\hat{p}_{wd})_{W \times D}$, where $\hat{p}_{wd} = n_{dw}/n_d$ is the frequency estimate of the conditional probability $p(w|d)$, n_{dw} is the number of occurrences of the word w in the document d, and n_d is the length of the document d.

The probabilistic latent semantic analysis (PLSA) model [2] describes the conditional probability of occurrences of words in documents,

$$p(w|d) = \sum_{t \in T} \varphi_{wt}\theta_{td} \tag{1}$$

in terms of the unknown conditional distributions $p(w|t) \equiv \varphi_{wt}$ for each topic $t \in T$ and $p(t|d) \equiv \theta_{td}$ for each document $d \in D$. The problem is reduced to the search for a stochastic matrix decomposition $F = \Phi\Theta$. To find an approximate solution, one maximizes the logarithm of likelihood:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} \to \max_{\Phi, \Theta}. \tag{2}$$

To maximize (2), one uses the EM-algorithm [2, 3], in which two steps are iteratively repeated.

At an E-step, the Bayes formula is used to estimate the conditional distributions of latent topics $p(t|d, w)$ for all words in the documents (d, w):

$$p(t|d, w) = \frac{\varphi_{wt}\theta_{td}}{\sum_{s \in T} \varphi_{ws}\theta_{sd}}. \tag{3}$$

At an M-step, these conditional probabilities are used to calculate the frequency estimates of the desired conditional probabilities:

$$\begin{aligned} \varphi_{wt} &\propto \hat{n}_{wt} = \sum_{d \in D} n_{dw} p(t|d, w), \\ \theta_{td} &\propto \hat{n}_{dt} = \sum_{w \in d} n_{dw} p(t|d, w), \end{aligned} \tag{4}$$

where the proportionality sign $\propto$ means that the expression on the right has to be normalized to obtain a distribution on the left.

The EM-algorithm has been well studied, and its convergence to a local maximum of the likelihood has been proved. Various methods for iteration rearrangement aimed at convergence rate acceleration were described in [4].

The latent Dirichlet allocation (LDA) model [5] introduces an additional probability assumption that the distributions φ_t and θ_d as column vectors of the matrices Φ and Θ are generated by Dirichlet distributions with hyperparameters $\beta = (\beta_w)_{w \in W}$ and $\alpha = (\alpha_t)_{t \in T}$, respectively, which leads to the smoothing of the frequency estimates at an M-step:

$$\varphi_{wt} \propto \hat{n}_{wt} + \beta_w, \quad \theta_{td} \propto \hat{n}_{dt} + \alpha_t. \tag{5}$$

[a] *Dorodnicyn Computing Center, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119333 Russia*
[b] *National Research University "Higher School of Economics," Myasnitskaya 20, Moscow, 101000 Russia*
[c] *Moscow Institute of Physics and Technology (State University), Institutskii per. 9, Dolgoprudnyi, Moscow oblast, 141700 Russia*
e-mail: voron@forecsys.ru

The other differences between EM-like algorithms for the PLSA and LDA models are subsidiary [3]. Moreover, their known modifications can be applied to both models [4].

The LDA model has become de facto a basis for hundreds of modifications adapted to a wide variety of problems. At the same time, LDA generates two open problems, which are rarely mentioned in the literature.

First, a priori Dirichlet distributions and their generalizations—the Dirichlet and Pitman–Yor processes—have weak linguistic justification and do not model any phenomena in natural languages. Their application is caused only by mathematical convenience, i.e., by the possibility of analytical integration over the parameter space of the model in the case of Bayesian inference.

Second, applications need composite models satisfying a large number of functional requirements [1]. Specifically, scientific search of large collections of publications requires a model that is simultaneously hierarchical, dynamical, n-gram, sparse, robust, multilingual, etc. Bayesian inference becomes too cumbersome when it combines more than two or three requirements in a single model. Such models have not yet been considered in the literature.

Thus, there is necessity of developing new principles of the design of topic models that are free from redundant probabilistic assumptions and simplify the construction of composite models. The proposed theory of additive regularization of topic models (ARTM) solves these problems.

The stochastic matrix decomposition $\Phi\Theta$ is not unique and is determined up to nonsingular transformation: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$. Thus, the construction of a topic model is an ill-posed problem and regularization has to be used for its solution. Instead of Bayesian regularization, we propose using the more general concept of Tikhonov regularization [6].

Assume that, along with likelihood (2), we need to maximize n criteria $R_i(\Phi, \Theta)$, $i = 1, 2, ..., n$, which are called regularizers. To solve the multicriteria optimization problem, we maximize a linear combination of the criteria L and R_i with nonnegative regularization coefficients τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta),$$
$$L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}. \tag{6}$$

As before, this problem can be solved using the EM-algorithm, but with the modified M-step formula:

$$\varphi_{wt} \propto \left(\hat{n}_{wt} + \varphi_{wt}\frac{\partial R(\Phi, \Theta)}{\partial \varphi_{wt}}\right)_+,$$
$$\theta_{td} \propto \left(\hat{n}_{dt} + \theta_{td}\frac{\partial R(\Phi, \Theta)}{\partial \theta_{td}}\right)_+. \tag{7}$$

Adding another regularizer leads to the addition a corresponding term to the M-step formula. Thus, one can construct composite topic models combining many additional requirements, including nonprobabilistic ones.

Below, we give examples of regularizers, some of which are known in the literature (although it is not always obvious that it is a regularizer) and the others are new. The list of regularizers is far from being complete and is rather illustrative.

1. A *smoothing regularizer* formalizes the requirement that the distributions φ_t and θ_d be close to given discrete distributions $\tilde{\beta}$ and $\tilde{\alpha}$ in terms of the Kullback–Leibler divergence:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \tilde{\beta}_w \ln \varphi_{wt}$$
$$+ \alpha_0 \sum_{d \in D} \sum_{t \in T} \tilde{\alpha}_t \ln \theta_{td} \to \max,$$

where β_0 and α_0 are regularization coefficients. Differentiating R immediately yields formulas (5) for an M-step in LDA if we introduce $\beta_w = \beta_0 \tilde{\beta}_w$ and $\alpha_t = \alpha_0 \tilde{\alpha}_t$. Here, we use neither a priori Dirichlet distributions nor Bayesian inference.

In ARTM theory, the Dirichlet distribution loses its central role. This is only one of possible regularizers, which is neither the best nor as universal as is thought. As a basic model, it is more reasonable to use PLSA, which does not have its own regularizers, and to add problem-oriented regularizers.

2. *Sparsing regularizer.* It is natural to assume that each document and each word is related to a small number of topics. Then, among the probabilities φ_{wt} and θ_{td}, many must be zero. This contradicts the LDA model, since the Dirichlet distribution does not admit zero values in the generated vectors.

The sparser the distribution, the lower its entropy. The maximal entropy is possessed by the uniform distribution. For this reason, we use a regularizer maximizing the divergence between the uniform distribution and the desired ones:

$$R(\Phi, \Theta)$$
$$= -\beta \sum_{t \in T} \sum_{w \in W} \ln \varphi_{wt} - \alpha \sum_{d \in D} \sum_{t \in T} \ln \theta_{td} \to \max.$$

As a result, we obtain an M-step formula that differs from a smoothing regularizer in the sign of the parameter and leads to sparsity:

$$\varphi_{wt} \propto (\hat{n}_{wt} - \beta)_+,$$
$$\theta_{td} \propto (\hat{n}_{dt} - \alpha)_+.$$

3. *Regularizer for semi-supervised learning.* To improve the interpretability of a topic model, experts can define training data. Suppose that it is known that some of the documents $d \in D_0$ concern the topics $T_d \subset T$ and some of the topics $t \in T_0$ are related to the words $W_t \subset W$. Let φ^0_{wt} be a uniform distribution on W_t and θ^0_{td} be a uniform distribution on T_d. Consider the regularizer

$$R(\Phi, \Theta)$$
$$= \tau_1 \sum_{t \in T_0} \sum_{w \in W_t} \varphi^0_{wt} \ln \varphi_{wt} + \tau_2 \sum_{d \in D_0} \sum_{t \in T_d} \theta^0_{td} \ln \theta_{td} \to \max.$$

According to (7), the M-step formulas become

$$\theta_{td} \propto \hat{n}_{dt} + \tau_1 \theta^0_{td}, \quad d \in D_0;$$
$$\varphi_{wt} \propto \hat{n}_{wt} + \tau_2 \varphi^0_{wt}, \quad t \in T_0.$$

This is also a smoothing regularizer, but, in contrast to LDA, it is applied only to those θ_{td} and φ_{wt} for which there are training data.

4. *Covariance regularizer for topics.* It is believed that a high dissimilarity between the topics improves the interpretability of a model [7]. A regularizer minimizing the covariances between the column vectors φ_t,

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \varphi_{wt} \varphi_{ws} \to \max,$$

leads to the M-step formula

$$\varphi_{wt} \propto \left(\hat{n}_{wt} - \tau \varphi_{wt} \sum_{s \in T \setminus t} \varphi_{ws} \right)_+.$$

The meaning of this formula is that the conditional probabilities $\varphi_{wt} = p(w|t)$ are gradually decreased for those words w that have higher values of the probability φ_{ws} in other topics. In the course of iterations of the EM-algorithm, for each word, the probabilities of more significant topics become increasingly higher, while the probabilities of less significant topics decrease and can vanish. Thus, this regularizer is also sparsing. Moreover, it has an additional useful property of grouping stop words in separate topics [7].

5. *Covariance regularizer for documents.* Sometimes there is additional information on the relations between documents of similar topics. For example, these can be documents belonging to the same category or placed placed into the same folder by users of a digital library or referring to each other. Suppose that $G = \langle D, E \rangle$ is a given directed graph and the graph edges $(d, c) \in E$ mean that the topics of the document c are close to the topics of the document d. This assumption is formalized by the regularizer

$$R(\Theta) = \tau \sum_{(d, c) \in E} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc} \to \max,$$

where n_{dc} is a weight of the edge (d, c), for example, the number of references to c in d. In [8] a similar model (LDA–JS) was proposed, in which the maximization of the covariance is replaced by the minimization of the Jensen–Shannon divergence between θ_d and θ_c. According to (7), the M-step formula for θ_{td} becomes

$$\theta_{td} \propto \hat{n}_{dt} + \tau \theta_{td} \sum_{c:\,(d, c) \in E} n_{dc} \theta_{tc}.$$

Thus, in the course of iterations, the conditional distributions $\theta_{td} = p(t|d)$ approach the distributions θ_{tc} of documents connected to d.

6. *Maximization of coherence.* A topic is called coherent if the words occurring most often in it frequently occur nearby in the documents. The average coherence of the topics is considered a good measure of the interpretability of a topic model. Let C_{uv} be an estimate of the joint occurrence frequency of words $(u, v) \in Q \subset W^2$. The following M-step formula for the Gibbs sampling algorithm with justification via the generalized Pólya urn model was proposed in [9]:

$$\varphi_{wt} \propto \hat{n}_{wt} + \tau \sum_{u \in W \setminus w} C_{uw} \hat{n}_{ut}.$$

It is easy to show that this formula also follows from the regularizer

$$R(\Phi) = \tau \sum_{t \in T} \sum_{(u, v) \in Q} C_{uv} \hat{n}_{ut} \ln \varphi_{vt} \to \max,$$

which minimizes the sum of the divergences between each distribution φ_{vt} and its empirical estimate over all words occurring with v.

7. *Maximization of likelihood in classification problems.* Suppose that there is additional information on the classification of documents, and it is assumed that the documents of the same class usually have similar topics. As classes, one can use categories, authors, publication years, quoting or quoted authors or documents, and users (readers) of documents. Specific models have been developed for all these cases [1]. Assume that each document d is associated with a collection of elements C_d from a finite set of class labels C. The problem is to identify the relations between the classes and topics, to improve the quality of the topic model with the help of additional information on classifications, and to construct a classification algorithm for new documents. One of the best classification topic models is the Dependency LDA [10], which determines the distribution $p(c|d)$ over classes for each document in terms of the distribution over classes for each topic $\psi_{ct} = p(c|t)$ and the distribution over topics for each document $\theta_{td} = p(t|d)$ by analogy with the basic topic model (1):

$$p(c|d) = \sum_{t \in T} \psi_{ct} \theta_{td}, \tag{8}$$

where the new unknown is the matrix $\Psi = (\psi_{ct})_{C \times T}$. In [10] a rather cumbersome derivation of the Gibbs sampling algorithm is given within the framework of the Bayesian approach. However, the same result can be achieved with the help of a regularizer minimizing the divergence between the classification model $p(c|d)$ and the empirical class frequency in documents m_{dc}:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td} \to \max, \quad (9)$$

where the regularization coefficient τ brings together the word frequencies n_{dw} and the class frequencies m_{dc}.

When we use a linear combination of regularizers R_i, the problem arises of choosing a coefficient vector $\tau = (\tau_i)_{i=1}^n$. A similar problem is effectively solved in ElasticNet while combining L_1- and L_2-regularizations in regression and classification problems [11]. In topic modeling problems, the variety of regularizers is wider and they affect each other in a nontrivial manner. Preliminary experiments have shown that some regularizers can worsen convergence if they are included too early or too abruptly. Therefore, the regularization coefficients should be increased gradually, so that they follow a certain trajectory. The construction of such trajectories in topic modeling problems is as yet an open problem.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 11-07-00480, 14-07-00847, 14-07-00908) and by the Branch of Mathematics of the Russian Academy of Sciences (the program "Algebraic and Combinatorial Methods of Mathematical Cybernetics and Information Systems of New Generation").

REFERENCES

1. A. Daud, J. Li, L. Zhou, and F. Muhammad, Frontiers Comput. Sci. China **4**, 280–301 (2010).
2. T. Hofmann, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, New York, 1999), pp. 50–57.
3. A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, in *Proceedings of International Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, 2009).
4. K. V. Vorontsov and A. A. Potapenko, Komp'yut. Issled. Model. **4**, 693–706 (2012).
5. D. M. Blei, A. Y. Ng, and M. I. Jordan, J. Machine Learning Res. **3**, 993–1022 (2003).
6. A. N. Tikhonov and V. Ya. Arsenin, *Solutions of Ill-Posed Problems* (Halsted, New York, 1977; Nauka, Moscow, 1986).
7. Y. Tan and Z. Ou, in *Proceedings of the 7th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (Taiwan, 2010), pp. 224–228.
8. L. Dietz, S. Bickel, and T. Scheffer, in *Proceedings of the 24th International Conference on Machine Learning ICML'07* (ACM, New York, 2007), pp. 233–240.
9. D. Mimno, H. M. Wallach, E. Talley, M. Leenders, A. McCallum, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP'11* (Assoc. Comput. Linguistics, Stroudsburg, PA, 2011), pp. 262–272.
10. T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, Machine Learning **88** (1–2), 157–208 (2012).
11. J. H. Friedman, T. Hastie, and R. Tibshirani, J. Stat. Software **33** (1), 1–22 (2010).

Translated by I. Ruzanova