Dirichlet Aggregation: Unsupervised Learning towards an Optimal Metric for Proportional Data

Hua-Yan Wang WANGHY@CIS.PKU.EDU.CN
Hongbin Zha ZHA@CIS.PKU.EDU.CN
State Key Laboratory of Machine Perception, Peking University

Hong Qin QIN@CS.SUNYSB.EDU
Department of Computer Science, State University of New York at Stony Brook

Abstract

Proportional data (normalized histograms) have been frequently occurring in various areas, and they could be mathematically abstracted as points residing in a geometric simplex. A proper distance metric on this simplex is of importance in many applications including classification and information retrieval. In this paper, we develop a novel framework to learn an optimal metric on the simplex. Major features of our approach include: 1) its flexibility to handle correlations among bins/dimensions; 2) widespread applicability without being limited to ad hoc backgrounds; and 3) a "real" global solution in contrast to existing traditional local approaches. The technical essence of our approach is to fit a parametric distribution to the observed empirical data in the simplex. The distribution is parameterized by affinities between simplex vertices, which is learned via maximizing likelihood of observed data. Then, these affinities induce a metric on the simplex, defined as the earth mover's distance equipped with ground distances derived from simplex vertex affinities.

1. Introduction and Motivation

Proportional data (normalized histograms) have been widely used in diverse areas for scientific studies and investigation. Examining types of proportional data, bins of the histograms could be: 1) intrinsically discrete, as in the bag-of-words representation of text (Salton, 1983); or 2) arisen from uniform tessellation or clustering of some continuous space, such as color histogram (Swain, 1991) or bag-of-keypoints representation of images (Csurka, 2004), or 3) mixtures of some basic components from a lower level representation, resulted from methods such as SVD (Deerwester, 1990) and discrete PCA (Buntine, 2004).

From a geometric point of view, normalized histograms could be considered as data points residing in a simplex. In many applications including data classification, visual recognition, and information retrieval, distances between histograms (i.e., documents, images, etc.) are oftentimes necessary measurements. A distance metric on the simplex could be defined in various ways. In this paper, we advocate that the metric should be adapted to observed distribution of data points therein, because the latter one reveals the underlying "true" structure of the simplex. As a mathematical abstraction of this intuition, we propose a novel framework to characterize an optimal metric on the simplex via exploiting affinities/distances between simplex vertices.

An intuitive illustration of our approach is shown in Figure 1, where Histogram 1 originally has roughly the same distance to Histograms 2 and 3 (e.g., measured by L_1 metric). A different metric could be induced by drawing closer two vertices of the 2-simplex (triangle). Under the new metric, Histogram 1 is much closer to Histogram 3 than to Histogram 2.

One could use either bin-by-bin or cross-bin comparison to calculate distance between histograms. The former assumes the bins being independent, such as L_k metric, histogram intersection, or Kullback-Leibler divergence. The latter, such as the earth mover's distance (EMD) (Rubner, 2000), needs inter-bin relationships (a.k.a. ground distances) to be specified. For ex-

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

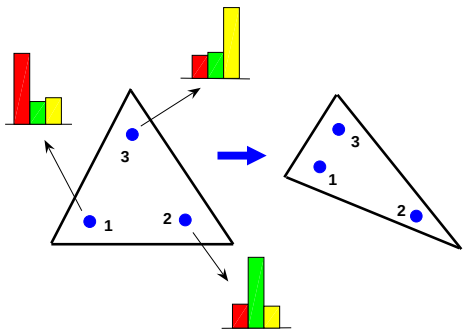

Figure 1. Metric on a 2-simplex induced by distances between simplex vertices.

ample, if the bins correspond to tessellations of some continuous space, neighboring bins are assumed to be more correlated; or if the bins correspond to clusters resulted from vector quantization, distances between bins are set to be the distances between corresponding cluster centers.

However, it is questionable that whether these strategies yield a good metric on the simplex. Bins being independent is an over-simplified assumption in many situations. Besides, Euclidean metric is not optimal in color space and feature space (Wyszecki, 1982; Omer, 2006), in the sense that it does not resemble perceptual distance of human beings; hence, using it to decide inter-bin distances is unjustified. Furthermore, in many situations, we could hardly even quantify relationship between histogram bins. For example, if the bins correspond to English words, although there were endeavors to represent words in some quantitative space (Schütze, 1993), explicitly quantifying their relationships can hardly be objective.

Due to these difficulties, our approach abandons any over-simplified assumption or background-specific information, and learns inter-bin distances directly from examples of histograms. This gives rise to special features of our approach: not being restricted to any ad hoc background condition, and being easily applicable to a variety of areas with a similar data representation.

The necessity of adapting distance metric to data point distribution in a space has long been recognized by researchers. A variety of linear/non-linear dimensionality reduction techniques could be viewed as implicitly addressing this issue. Metric on the learned low-dimensional manifold is actually "adapted" to data point distribution in the original high-dimensional space. When distribution of data points exhibits sophisticated non-linear patterns, traditional techniques compromise to characterize global structure of data via accumulating information locally, either by preserving geodesic distance (Tenenbaum, 2000)[1] or fitting local linear structures (Roweis, 2000).

As a major contribution of our framework, in contrast to "think globally, fit locally" (Roweis, 2000), we propose to "think globally, fit globally". Fitting globally inevitably confront us with the difficulty of characterizing sophisticated patterns of data point distribution in a unified parametric form, which is generally intractable in N-dimensional space. However, by normalization, data points are mapped onto a $(N-1)$-simplex, on which we can devise a parametric distribution that indeed exhibits such flexibility.

The distribution bridges "aggregation" patterns of data points and "aggregation" patterns of simplex vertices. Specifically, it is a Dirichlet mixture in a restricted form, parameterized by affinities between simplex vertices. These affinities are learned via maximizing likelihood of observed data. Then, an induced metric on the simplex is characterized by the earth mover's distance (EMD) (Rubner, 2000), equipped with ground distances derived from those learned simplex vertex affinities. Optimality of the induced metric is validated on a variety of data representations.

1.1. Related Work

Lebanon (2003) addressed the metric learning problem on a simplex. Their approach chooses an optimal metric from a parametric family that maximizes the inverse volume of the observation (equivalent to maximum likelihood from a statistical perspective). And the parametric family is constructed as pull-back metrics of the Fisher information metric on the simplex, under a parametric family of transformations from the simplex to itself. The transformations are essentially independent scalings of individual vertices/dimensions, hence their approach does not directly handle correlations among dimensions.

Omer (2006) also formalized the idea of adapting distance metric to data point distribution, but in an explicit form. They defined the "bottleneck affinity" between two pixel features, based on feature point density along the straight line connecting them. This approach essentially judges whether the two points reside in a same "cluster", in which case they should be drawn closer. However, their method has a "narrow" sight (restricted to a straight line) in contrast to the

[1] Although Isomap is regarded as a "global" method, it accumulates information locally to characterize geodesic distance.

integrated global vision of our solution.

A broad class of techniques (Blei, 2003; Buntine, 2004; Deerwester, 1990; Gehler, 2006; Hofmann, 2001; Marlin, 2004; Welling, 2004) provided us with a hierarchical representation where a low-dimensional high-level simplex[2] is embedded (possibly in a probabilistic manner) in the original simplex[3]. Thus our framework could be built on the former one, where redundancy of the representation is largely reduced.

Blei (2006) used the logistic normal distribution as a component of their correlated topic model derived from LDA (Blei, 2003). The logistic normal distribution has a covariance structure among simplex vertices, which is similar to our model.

The distribution we devised could be viewed as a Dirichlet mixture in a restricted form. Dirichlet mixture model was addressed by researchers in a variety of areas including bioscience (Sjolander, 1996), text modeling (Yamamoto, 2005), and image analysis (Bouguila, 2006). However, none of these work addressed the relationship between simplex vertex affinities and Dirichlet mixture parameters, which is fully exploited in this paper.

2. Mathematical Preliminaries

A normalized histogram

$$\mathbf{d} = (d_1, d_2, \cdots, d_N), \quad d_i \geq 0 \quad \sum_{i=1}^{N} d_i = 1 \qquad (1)$$

resides in a $(N-1)$-simplex, which is a $N-1$ dimensional area in N dimensional space. Each vertex of the $(N-1)$-simplex represents a histogram with value 1 in one bin and 0 in the others.

Although the following two concepts are quite straightforward, we give formal definitions due to their significance in articulating properties of our model.

Definition 1 (Sub-simplex) *$\widehat{V}$ is a sub-simplex of a $(N-1)$-simplex arisen from $V \subseteq \{1, 2, \cdots, N\}$, if*

$$\widehat{V} = \{\boldsymbol{d} \,|\, d_i \geq 0\,,\ \sum_{i=1}^{N} d_i = 1\,,\ d_i = 0 \text{ for } i \notin V\} \qquad (2)$$

Definition 2 (Generalized principal sub-matrix) *For a symmetric matrix Λ of order N, its generalized principle sub-matrix is obtained by eliminating i-th row and i-th column, for all $i \notin V$, where $V \subseteq \{1, 2, \cdots, N\}$.*

[2]It is referred as "topic-simplex" in (Blei, 2003)
[3]It is referred as "word-simplex" in (Blei, 2003)

3. Dirichlet Aggregation

3.1. Vertex Affinities of Simplex

A commonly used distribution on a simplex is the Dirichlet distribution,

$$\mathbf{P}(\mathbf{d}|\alpha) = \frac{\Gamma(\sum_{i=1}^{N} \alpha_i)}{\prod_{i=1}^{N} \Gamma(\alpha_i)} \prod_{i=1}^{N} d_i^{\alpha_i - 1} \qquad (3)$$

It has N parameters $\alpha_1, \alpha_2, \cdots, \alpha_N$, each associated with one vertex of the simplex. A property of Dirichlet distribution is that

$$\mathbf{E}(d_i|\alpha) = \frac{\alpha_i}{\sum_{i=1}^{N} \alpha_i} \qquad (4)$$

which means that each α_i controls "aggregation" of mass near the corresponding vertex (see Figure 2).

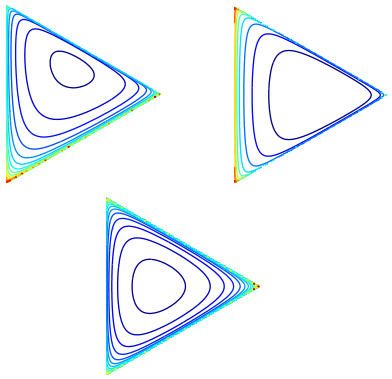

Figure 2. Contour lines of the Dirichlet distribution on a 2-simplex (triangle). Outer area has higher probability density. Top left: $\alpha_1 = 0.7$, $\alpha_2 = \alpha_3 = 0.1$. Top right: $\alpha_1 = \alpha_2 = 0.7$, $\alpha_3 = 0.1$. Bottom: $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$.

A crucial limitation of Dirichlet distribution is easily revealed. Consider affinities between 4 English words: "economy", "market", "geography", and "terrain", as in Figure 3. Intuitively, "economy" and "market" are somewhat close to each other; so do "geography" and "terrain"; yet the two couples are relatively far apart. Suppose we have collected a number of documents containing some of these words, thus each document could be represented as a point (histogram) in the 3-simplex (tetrahedron). Mass in such a 3-simplex should be intuitively predicted to be concentrated near the sub-simplex (edge) connecting "economy" and "market", and also the sub-simplex connecting "geography" and "terrain". This phenomenon, in principle, can not be modeled by a Dirichlet distribution. Because all words are considered to be equally important in such a setting, 4 parameters of the Dirichlet distribution need

to be all identical, which yields a undesirable trivial symmetric distribution.

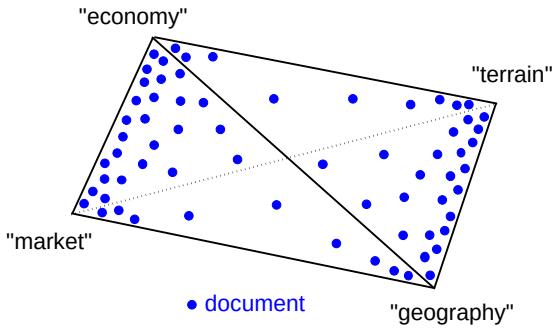

Figure 3. A toy example of 4 English words leads to an expected distribution of documents.

In real applications, we may have to deal with even higher dimensional representations and the simplex vertices could exhibit sophisticated aggregation patterns. In order to achieve sufficient flexibility, we explicitly address affinities between vertices of the simplex. Consider the following matrix:

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1N} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{N1} & \lambda_{N2} & \dots & \lambda_{NN} \end{pmatrix} \tag{5}$$

with $\lambda_{ij} \in [0, 1]$ indicating affinity between vertex i and j. Naturally we add two restrictions to these quantities:

$$\lambda_{ij} = 1 \quad \text{for} \quad i = j \tag{6}$$

$$\lambda_{ij} = \lambda_{ji}, \quad \text{for} \quad i, j = 1, 2, \cdots, N \tag{7}$$

We define a distribution on the simplex parameterized by Λ,

$$\mathbf{P}(\mathbf{d}|\Lambda) = \sum_{i=1}^{N} \frac{\Gamma(\sum_{j=1}^{N} \lambda_{ij})}{N \prod_{j=1}^{N} \Gamma(\lambda_{ij})} \prod_{j=1}^{N} d_j^{\lambda_{ij}-1} \tag{8}$$

To see its relationship to the Dirichlet distribution, consider the following restriction instead of (6) and (7).

$$\lambda_{1i} = \lambda_{2i} = \cdots = \lambda_{Ni} = \alpha_i, \tag{9}$$

Obviously, (8) degenerates to the Dirichlet distribution (3) under the restriction (9).

To exploit properties of distribution (8), we first consider two extreme cases:

Extreme Case 1:

$$\lambda_{ij} = 0, \quad \text{for} \quad i \neq j \tag{10}$$

In this case, distribution (8) degenerates to a discrete distribution on simplex vertices, which means, bins of the histograms are exclusive. Histograms reside right at a simplex vertex with probability 1.

Extreme Case 2:

$$\lambda_{ij} = 1, \quad \text{for all } i, j \tag{11}$$

In this case, (8) indicates a uniform distribution on the simplex, which also has a straightforward explanation. All the histogram bins are intrinsically identical, and any mixture ratio is equally preferred.

With the definitions in Section 2, it is obvious that each sub-simplex is in correspondence with a generalized principle sub-matrix of the parameter matrix (5). Both are coming from a certain sub-set V of simplex vertices. This connection provides us with a principled way to link observed data points with affinities between simplex vertices. When observed data points "aggregate" near some sub-simplex, increasing of affinities between relevant simplex vertices is directly reflected on the corresponding generalized principal sub-matrix of the parameter matrix.

To visualize the phenomenon of Dirichlet aggregation. We use a low dimensional illustration resembling the 4-word toy example we have discussed. Let the parameter matrix

$$\Lambda = \begin{pmatrix} 1 & 0.7 & 0.1 & 0.1 \\ 0.7 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.6 \\ 0.1 & 0.1 & 0.6 & 1 \end{pmatrix} \tag{12}$$

We sampled 1000 points from such a distribution (see Figure 4). Two generalized principal sub-matrices of the parameter matrix corresponding to sub-simplex "1-2" and sub-simplex "3-4" exhibit high values. In accord with "aggregation" of the simplex vertices, data points in the 3-simplex "aggregate" near the sub-simplex "1-2" and the sub-simplex "3-4".

3.2. Induced Metric

As we have stated, our framework aims to defining a metric on the simplex, adapted to distribution of data points therein. This intuition becomes straightforward in the low-dimensional case in Figure 4. The underlying "true" distance between a pair of data points residing near sub-simplex "1-2" should be smaller than that of a pair residing near sub-simplex "2-3" with the

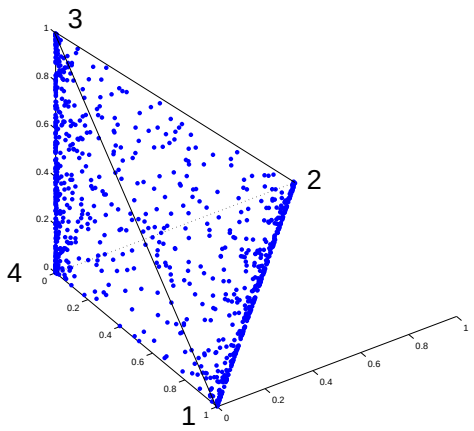

Figure 4. 1000 points sampled from the distribution (8) with parameter matrix (12). Note that it resembles the imaginary distribution in Figure 3.

same Euclidean distance, because the former essentially reside in a same "cluster". In other words, the simplex needs to be warped such that the sub-simplex "1-2" and "3-4" are shrinking, resembling the simplex in Figure 3. In higher dimensions, the warping could exhibit diverse sophisticated patterns, which are well captured because of the flexibility of distribution (8).

Realization of the warping is achieved by EMD with ground distance between the i-th and j-th bin set to be $-\log \lambda_{ij}$, where λ_{ij} is the learned affinity between corresponding simplex vertices. EMD between two histograms is the overall work of transporting mass in order to make them identical. Equivalently, we can perceive it as the overall work of moving a point (histogram) to another in the simplex (the path corresponds to intermediate states of the transportation process). Hence, in Figure 4, EMD with learned ground distances actually warps the simplex, such that moving a point near the edge "1-2" needs less effort than moving a point near the edge "2-3". Furthermore, for normalized histograms, with the same overall mass 1, Rubner (2000) proved that EMD is a true metric. In our framework, it is a true metric induced by affinities between simplex vertices. For brevity, readers are referred to (Rubner, 2000) for more details of EMD.

Computing EMD involves solving a linear programming problem, with $O(N^2)$ time complexity, which makes it difficult to directly apply our approach to efficiency demanding and large scale applications. We leave this issue to further research.

3.3. Parameter Estimation

Without enforcing the restrictions (6) and (7), distribution (8) could be viewed as a Dirichlet mixture model with N equally weighted components:

$$\mathbf{P}(\mathbf{d}|\Lambda) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{P}_i(\mathbf{d}|\Lambda) \tag{13}$$

where each $\mathbf{P}_i$ is a Dirichlet distribution.

Estimating parameters of a Dirichlet distribution was thoroughly discussed by Minka (2003). We employed a modified Newton iteration in our EM-like algorithm.

During the E-step, we compute the probability density at each $\mathbf{d}_j$, $j = 1, \cdots, M$, under all $\mathbf{P}_i$, based on the current estimation of Λ. Let $w_{ij} = \mathbf{P}_i(\mathbf{d}_j|\Lambda)$. And for each i, we normalize w_{ij} such that $\sum_{j=1}^{M} w_{ij} = 1$.

During the M-step, parameters of each $\mathbf{P}_i$ are estimated using $\mathbf{d}_j$ $(j = 1, \cdots, M)$ re-weighted by w_{ij}.

Newton iteration is modified as follows. To estimate parameters of $\mathbf{P}_i$: 1) whenever averaging $\mathbf{d}_j$ or $\log \mathbf{d}_j$, re-weight them by w_{ij}; 2) during initialization, set $\lambda_{ii} = 1$ to meet the restriction we have imposed; 3) when calculating $\mathbf{H}^{-1}\mathbf{g}$ (inverted Hessian times gradient), set $(\mathbf{H}^{-1}\mathbf{g})_i = 0$, such that λ_{ii} is not changed during Newton iteration. For brevity, readers are referred to (Minka, 2003) for more details of estimating Dirichlet parameters. After each M-step, we add a regularization step (set both to their average) to force $\lambda_{ij} = \lambda_{ji}$.

Initially, we set $\lambda_{ij} = 1$ for $i = j$, and $0 < \lambda_{ij} \ll 1$ (this value has so little effect on the results, according to our experiments) for $i \neq j$. This could be viewed as placing a kernel at each vertex of the simplex.

It is noticeable that the restriction $\lambda_{ij} \leq 1$ is not guaranteed in our algorithm. Actually it rarely exceeds 1 in practice. And when λ_{ij} exceeds 1, vertex i and j are indeed extremely correlated (actually they were almost identical "topics" in our experiments). In these situations, we simply treat λ_{ij} as 1 for subsequent procedures. Actually, the ground distance between relevant vertices would be 0, i.e. they are merged and the dimensionality of the simplex is reduced by 1.

4. Experimental Results

4.1. Reuters-21578 Corpus

We tested our model on the widely used corpus Reuters-21578[4]. All documents must undergo pre-

[4]The corpus is from http://ai-nlp.info.uniroma2.it/moschitti/corpora.htm.

processing steps including removing stop words and rare words. After pre-processing, the corpus has 22556 unique words, 12897 documents, and 90 categories (plus one category labelled "unknown"). One document could exhibit multiple category labels. Each document is represented by a count histogram of words.

The dimensionality of the representation is then reduced by LDA (Blei, 2003), after which each document is represented by a histogram of topic proportions. Note that our model is equally applicable on representations obtained by other methods (Buntine, 2004; Deerwester, 1990; Gehler, 2006; Hofmann, 2001; Marlin, 2004; Welling, 2004).

For a "query" document[5], we rank all other documents in the corpus by their distances to the query. A document is marked as "correct" iff it has at least one same category label with the query. P-R plots averaged over the whole corpus are shown in Figure 5.

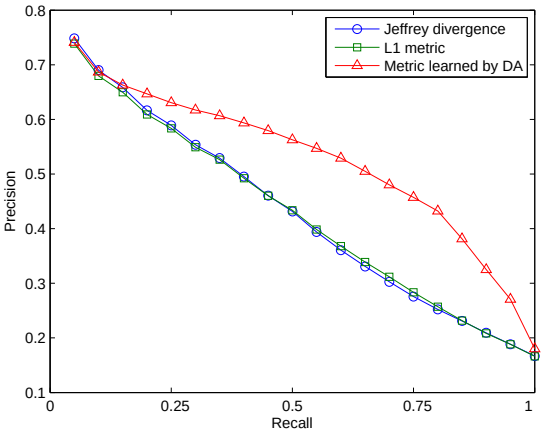

Figure 5. P-R curves on Reuters database. 50-topic representation obtained by LDA. "DA" appearing in legend stands for Dirichlet aggregation. Note that the first two curves are almost overlapping with each other.

We make use of Jeffrey divergence[6] (Puzicha, 1997) and L_1 metric for comparison purpose. In addition, we try other methods such as cosine similarity and L_2 metric. They generally perform no better than these two methods. Moreover, it can be proved that EMD with a constant ground distance is equivalent to L_1 metric (allowing a constant scaling factor, which equals to the ground distance). Therefore, we can assert that the significant improvement was brought by our learned inter-vertex affinities, but not EMD.

[5]Documents with the category label "unknown" are not used as query.

[6]A modified symmetric and more robust version of Kullback-Leibler divergence.

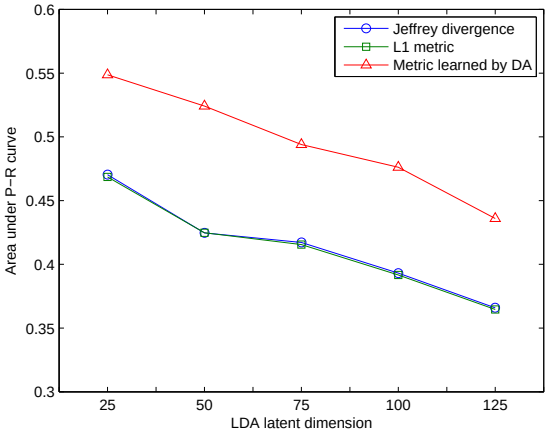

Figure 6. Area under P-R curve vs. representation dimension on Reuters database.

To validate performance of Dirichlet aggregation under different representation dimensions, we work on a considerable range of LDA latent dimensions (number of "topics"), and we have observed a steadily superior performance of Dirichlet aggregation. (see Figure 6).

4.2. The Caltech4 Database

The Caltech4 database is a subset of the Caltech101 database (Li Fei-Fei, 2004). It consists of 2233 images, 4 categories: faces, motorcycles, airplanes, and leopards, with considerable intra-class variations. All images are converted to gray scale. We first compute 128-dimensional SIFT (Lowe, 2004) descriptors on detected keypoints of all images. Vector quantization (VQ) is then applied to all these descriptors, after which every image is represented by a count histogram of visual-words (cluster centers).

We test our model on two different representations: 1) 2000 visual-words are obtained by VQ and then we apply LDA, after which each image is represented by a histogram of "visual-topics"; 2) 100 visual-words are obtained by VQ, and each image is represented by a histogram of visual-words. Note that the latter is consistent with case 2 discussed in the beginning of this paper. For the former representation, we also try several different LDA latent dimensions. Dirichlet aggregation performs well in all these situations. (see Figures 7, 8, 9).

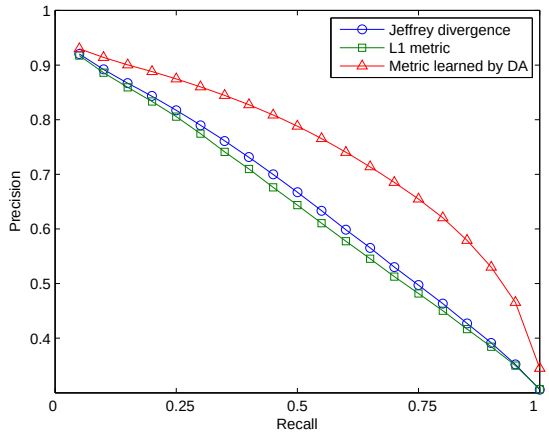

Figure 7. P-R curves on Caltech4 database. 40-topic representation obtained by LDA.

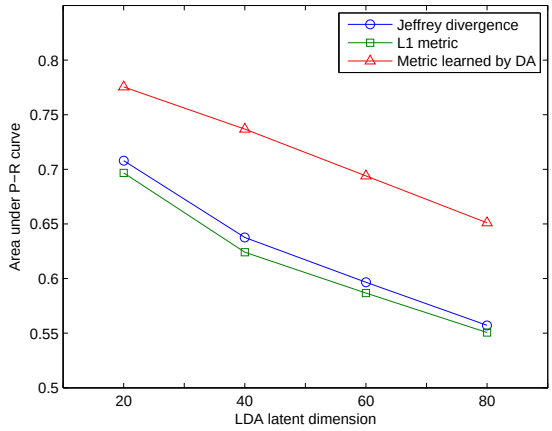

Figure 8. Area under P-R curve vs. representation dimension on Caltech4 database.

5. Conclusion and Discussion

In this paper, we have proposed a novel framework, Dirichlet aggregation, to learn an optimal metric on a simplex, by exploiting "aggregation" patterns of data points and "aggregation" patterns of simplex vertices, which are too complicated to be simply modeled as deterministic clusters. The induced metric is a true metric, and yields a considerable improvement for retrieval performance. Our approach performs steadily well on both text and images, and on a wide range of representation dimensions, which indicates that there could be other potential applications of Dirichlet aggregation in various areas, as long as we have a basic representation of normalized histograms.

In our experiments, the P-R curves of our learned metric are superior mainly for intermediate recalls, yet not so much for high recalls[7](see Figures 5, 7). However, an exception is observed in Figure 9, in which P-R curve of our learned metric outperforms others even for high recalls.

These phenomena could be explained as follows: although the overall "shape" of the simplex is warped in an optimal way, "micro-structure" in "peripheral" areas (near low-dimensional sub-simplexes) of the simplex is largely preserved. A crucial difference between visual-word-histogram representation (Figure 9) and "topic"-histogram representation (Figure 5, 7) is that: in the latter case, a Dirichlet prior had been imposed on the topic-simplex by LDA, hence points (histograms) mainly reside in "peripheral" areas of the simplex (i.e., near edges, triangles, etc.), where most bins have a value close to zero. High recalls of the P-R curve is determined by points that reside "very" close to each other. If two histograms are "indeed" very similar, especially when their mass are concentrated in two or three bins, their distance would be very small no matter how we measure it. Nonetheless, difference for intermediate recalls (see Figures 5, 7, 9) indicates that points with moderate distances are always rearranged in an optimal way by our learned metric. In the case of Figure 9, histograms distribute more uniformly in the simplex, and superiority of the metric learned by Dirichlet aggregation is even significant for histograms of high similarity.

[7]This phenomenon is also significant in all other plots we are unable to show due to page limitation

Acknowledgments

We thank all reviewers for their helpful comments on our work. We also thank Yi Liu for many thoughtful discussions.

References

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(2003), 993-1022.

Blei, D. M. and Lafferty, J. D. (2006) Correlated Topic Models. In *NIPS*, volume 18.

Bouguila, N. and Ziou, D. (2006) Unsupervised Selection of a Finite Dirichlet Mixture Model: An MML-Based Approach. IEEE *Transactions on Knowledge and Data Engineering*. Vol.18, No.8.

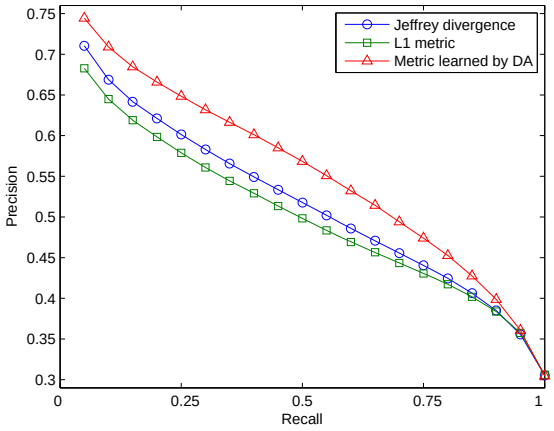

Figure 9. P-R curves on Caltech4 database. Visual-word histogram representation with 100 visual-words obtained by vector quantization.

Buntine, W. and Jakulin, A. (2004) Applying Discrete PCA in Data Analysis In *Proceedings of the 20th UAI.*

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., Bray, C. (2004) Visual Categorization with Bags of Keypoints. In *the 8th ECCV, Workshop on Statistical Learning in Computer Vision.*

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6), 391-407.

Gehler, P. V., Holub, A. D., Welling, M. (2006) The Rate Adapting Possion Model for Information Retrieval and Object Recognition. In *Proceedings of the 23rd ICML.*

Hofmann, T. (2001) Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42, 177-196.

Lebanon, G. (2003) Learning Riemannian Metrics. In *Proceedings of the 19th UAI.*

Li Fei-Fei, Fergus, R., Perona, P. (2004) Learning Generative Visual Models from Few Training Examples: an Incremental Bayesian Approach Tested on 101 Object Categories. IEEE. *CVPR, Workshop on Generative-Model Based Vision.*

Lowe, D. G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91-110.

Marlin, B. and Zemel, R. (2004) The Multiple Multiplicative Factor Model for Collaborative Filtering. In *Proceedings of the 21st ICML.*

Minka, T. P. (2003) Estimating a Dirichlet Distribution. Technical Report, Microsoft Research.

Omer, I. and Werman, M. (2006) Image Specific Feature Similarities. In *Proceedings of the 9th ECCV.*

Puzicha, J., Hofmann, T., Buhmann, J. (1997) Nonparametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval. In *Proceedings of the IEEE. CVPR*, pp.267-272.

Roweis, S. and Saul, L. (2000) Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science.* 290: 2323-2326

Rubner, Y., Tomasi, C., Guibas, L. J. (2000) The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2), 99-121.

Salton, G. and McGill, M. editors (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill.

Schütze, H. (1993) Word Space. In *NIPS*, volume 5.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., Haussler, D. (1996) Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology. *CABIOS*, 12(4): 327-345

Swain, M. J. and Ballard, D. H. (1991) Color Indexing. *International Journal of Computer Vision*, 7(1), 11-32.

Tenenbaum, J., Silva, V.d., Langford, J. (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science.* 290: 2319-2323

Welling, M., Rosen-Zvi, M., Hinton, G. (2004) Exponential Family Harmoniums with an Application to Information Retrieval. In *NIPS*, volume 16.

Wyszecki, G. and Stiles, W. S. (1982) *Color Science: Concepts and Methods, Quantitative Data and Formulae.* John Wiley and Sons: New York, NY.

Yamamoto, M. and Sadamitsu, K. (2005) Dirichlet Mixtures in Text Modeling. Technical report. CS-TR-05-1. University of Tsukuba.