Dirichlet Component Analysis: Feature Extraction for Compositional Data

Hua-Yan Wang WANGHY@CIS.PKU.EDU.CN
Key Laboratory of Machine Perception (Ministry of Education), Peking University

Qiang Yang QYANG@CSE.UST.HK
Department of Computer Science and Engineering, Hong Kong University of Science and Technology

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

Hongbin Zha ZHA@CIS.PKU.EDU.CN
Key Laboratory of Machine Perception (Ministry of Education), Peking University

Abstract

We consider feature extraction (dimensionality reduction) for compositional data, where the data vectors are constrained to be positive and constant-sum. In real-world problems, the data components (variables) usually have complicated "correlations" while their total number is huge. Such scenario demands feature extraction. That is, we shall de-correlate the components and reduce their dimensionality. Traditional techniques such as the Principle Component Analysis (PCA) are not suitable for these problems due to unique statistical properties and the need to satisfy the constraints in compositional data. This paper presents a novel approach to feature extraction for compositional data. Our method first identifies a family of dimensionality reduction projections that preserve all relevant constraints, and then finds the optimal projection that maximizes the estimated Dirichlet precision on projected data. It reduces the compositional data to a given lower dimensionality while the components in the lower-dimensional space are de-correlated as much as possible. We develop theoretical foundation of our approach, and validate its effectiveness on some synthetic and real-world datasets.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

1. Introduction

Compositional data (positive constant-sum real vectors) are frequently encountered in various scientific disciplines and industrial applications. They quantitatively describe the parts that comprise the entire entity. In geology, scientists investigate relative proportion of different minerals in rocks. In microeconomics, household expenditure in different commodity/service groups is recorded as relative proportion. In information retrieval, documents are usually represented as relative frequencies of words in a prescribed vocabulary. Generally, compositional data are natural representations when the variables (features) are essentially probabilities of complementary and mutually exclusive events. The variables (features) in compositional data are referred to as *components* in this paper.

Feature extraction is often applied in machine learning when the datasets are large and complex. The same is needed for compositional data. The need for feature extraction arises from four aspects. First, prediction performance in classification and regression can benefit from a lower dimensional representation with de-correlated *components* to avoid the curse of dimensionality. Second, feature extraction may improve overall domain understanding, e.g., we could expect the learned *components* to represent latent independent sources from which the data are generated. Third, the computational expense of subsequent data processing can be reduced with a lower dimensionality. Finally, reducing data to two or three dimensions facilitates visualization and further analysis by domain experts.

However, traditional feature extraction techniques are

not suitable for compositional data due to several reasons. First, the traditional measurement of "correlation"[1] implicated by multivariate Gaussian and PCA only captures a linear relationship between two random variables. In contrast, the "curved" nature (Aitchison, 1983) of compositional data and the "spurious correlation" (Pearson, 1896) induced by the constant-sum constraint make it problematic to interpret correlation as merely a linear relationship. We thus need a new concept of "correlation" for compositional data. Second, the positive and constant-sum constraints for compositional data are not considered in most dimensionality reduction techniques, and simply modifying them to accommodate these constraints may induce biases.

PCA is one of the most widely used techniques for feature extraction. Given a target dimension k, PCA identifies an orthogonal projection to a k dimensional subspace that maximizes the estimated Gaussian variance of the projected data. Moreover, the covariance matrix is diagonalized such that the variables are decorrelated. Our approach adapts this framework for compositional data. In particular, we first identify a family of projections that preserve a simplex constraint as substitutes for the orthogonal projections in PCA. Then, we find an optimal projection that minimizes the "Dirichlet correlation" among the projected *components*, as a substitute for maximizing the estimated Gaussian variance in PCA. The Dirichlet correlation among the *components* is defined as the estimated Dirichlet precision on projected data. The *components* are better de-correlated and separated with a smaller Dirichlet correlation. The notion of Dirichlet correlation extends the traditional "linear" interpretation of correlation connoted in the covariance structure of multivariate Gaussian and PCA. Because of our approach's affinity to the Dirichlet distribution, we call it *Dirichlet component analysis (DCA)*.

Although the Dirichlet distribution is a natural parametric family on the simplex, its role in modeling compositional data is not well studied. As pointed out in (Aitchison, 1982), the "ultimate independence" property of the Dirichlet family prevents us from directly applying it to model compositional data. Consequently, the use of Dirichlet family in compositional data analysis has been superseded by the *log-ratio* framework (eliminating the constraints by a transformation to $\mathbb{R}^N$) originated from (Aitchison, 1982). For example, the *centered log-ratio* is defined as dividing all *components* by their geometric mean and then applying the *log* function. Although this framework has been very successful, it has certain problems. The *log-ratio* well captures variability in the central area of the simplex, but encounters singularity in peripheral areas. For example, in sparse compositional data (e.g., term frequencies in documents with thousands of terms) the log-ratio is not well defined as most denominators would be zero.

In this paper, we make three main contributions. 1) We identify a rich family of dimensionality reduction transformations for compositional data, as an alternative to existing compositional operators such as *subcomposition*, *amalgamation*, and *partition* (Aitchison, 1982). 2) We exploit the Dirichlet family for compositional data analysis to capture data variability beyond traditional concepts of statistical correlation. 3) We show that the entire framework of DCA is effective and conceptually succinct, and validate its effectiveness on two synthetic datasets and two real-world datasets.

2. Dirichlet Component Analysis

2.1. The Projection Family

Compositional data are positive constant-sum vectors. Without loss of generality, we assume all *components* to sum to one:

$$\mathbf{x} = (x_1, x_2, \cdots, x_N)^T, \quad x_i \geq 0 \text{ for all } i, \quad \sum_{i=1}^{N} x_i = 1 \tag{1}$$

All points satisfying these constraints constitute the $(N-1)$-simplex, denoted as $\mathbb{S}^N$. As low dimensional examples, $\mathbb{S}^3$ is a triangle and $\mathbb{S}^4$ is a tetrahedron.

Given a target dimension K ($K \leq N$), our first aim in dimensionality reduction is to identify a family of projections from $\mathbb{S}^N$ to $\mathbb{S}^K$.

Proposition 1 *For linear projections*

$$\mathbf{y} = \mathbf{R}\,\mathbf{x} \quad \text{where} \qquad \mathbf{R} = (r_{ij})_{K\times N} \tag{2}$$

$\mathbf{y}$ *is in* $\mathbb{S}^K$ *for all* $\mathbf{x}$ *in* $\mathbb{S}^N$ ***if and only if***
1) $r_{ij} \geq 0$ *for all* i, j.
2) $\sum_{i=1}^{K} r_{ij} = 1$, *for* $j = 1, \cdots, N$.

The proof is quite straightforward and we omit it here for brevity.

Such projections could be viewed as *rearranging* mass from the N original *components* to the K new *components*, while the law of conservation of mass is satisfied. Hence we refer to such linear transforms from $\mathbb{S}^N$ to $\mathbb{S}^K$ as *rearrangements*.

Unfortunately, we could have degenerate *rearrangements* when some rows of $\mathbf{R}$ are close to zero and

[1] It is measured by the Pearson's correlation coefficient.

as a result, the corresponding new *component* is almost ignored in the *rearranging* process. Without *a priori* knowledge we should treat the K new *components* equally, which gives rise to the family of *balanced rearrangements*:

Definition 1 (Balanced Rearrangement) *A linear projection $\mathbf{R}\,\mathbf{x} = \mathbf{y}$ is a **balanced rearrangement**, if $\mathbf{R} = (r_{ij})_{K\times N}$ satisfies:*
1) $r_{ij} \geq 0$ for all i, j.
2) $\sum_{i=1}^{K} r_{ij} = 1$, for $j = 1, \cdots, N$.
3) $\sum_{j=1}^{N} r_{ij} = N/K$, for $i = 1, \cdots, K$.

The *balanced* is described by the following proposition, which gives rise to a univariate (symmetric) Dirichlet family, as we will discuss in Section 2.2.

Proposition 2 *If $\mathbf{R}_{K\times N}$ is a balanced rearrangement matrix, $\mathbf{x}$ is a random vector in $\mathbb{S}^N$ satisfying $\mathbf{E}(x_i) = \frac{1}{N}$ for all i, then $\mathbf{y} = \mathbf{R}\,\mathbf{x}$ is a random vector in $\mathbb{S}^K$ and $\mathbf{E}(y_i) = \frac{1}{K}$ for all i.*

The proof is straightforward given the linearity of the expectation operator.

The space of balanced rearrangement projections from $\mathbb{S}^N$ to $\mathbb{S}^K$ is a $NK - N - K + 1$ dimensional vector space, which is closed with respect to the operator of weighted average. This property is useful in developing the optimization algorithm in Section 3:

Proposition 3 *If $\mathbf{R_1}$ and $\mathbf{R_2}$ are balanced rearrangement matrices, α and β are positive real numbers, then $(\alpha\mathbf{R_1} + \beta\mathbf{R_2})/(\alpha+\beta)$ is a balanced rearrangement matrix.*

This is easy to validate from the definition of balanced rearrangement.

A noticeable property of balanced rearrangements is the "shrinking effects" stated as follows:

Proposition 4 *Let $\min(\mathbf{x})$ be the minimum component of $\mathbf{x}$. $\mathbf{R}_{K\times N}$ is a balanced rearrangement matrix with $K \leq N$, then $\min(\mathbf{Rx}) \geq \min(\mathbf{x})$ for all $\mathbf{x}$ in $\mathbb{S}^N$.*

The proof is obvious as long as we notice that each component of $\mathbf{Rx}$ is N/K times a weighted average of the components of $\mathbf{x}$, where equality holds only in some trivial cases. For example, $\mathbf{R}$ is the identify matrix or $\mathbf{x} = (1/N, 1/N, \cdots, 1/N)$.

Intuitively, Proposition 4 states that the balanced rearrangements always make data points "shrink" toward the central area of the simplex, which is undesirable because it diminishes variabilities of data[2]. To solve this problem, we induce the *regularization* operator for compositional data. As shown in Figure 1, we impose on the data points a parallel move along the direction $x_1 = x_2 = \cdots = x_N$, and then project the data points back to the simplex by radial projection:

Definition 2 (Regularization) *Given a compositional dataset $\mathcal{X} = \{\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^M\}$, a **regularization** on the dataset is denoted as: $\widetilde{\mathcal{X}} = \{\widetilde{\mathbf{x}^1}, \widetilde{\mathbf{x}^2}, \cdots, \widetilde{\mathbf{x}^M}\}$, where $\widetilde{\mathbf{x}^i} = \frac{1}{\sum_{j=1}^{N}(x_j^i - \delta)}(x_1^i - \delta, x_2^i - \delta, \cdots, x_N^i - \delta,)$ for $i = 1, 2, \cdots, M$, and the **regularization factor** $\delta = \min(\min(\mathbf{x^1}), \min(\mathbf{x^2}), \cdots, \min(\mathbf{x^M}))$.*

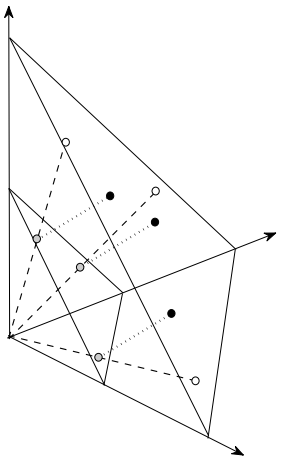

Figure 1. Regularization of compositional data points (black) is performed by parallel projection to the gray points, then radial projection to the white points.

The *regularization* operator can be viewed as a "scaling", which preserves Euclidean geometrical properties such as distance (allowing a constant scaling factor) and angle. Intuitively it "expands" the data points and compensates for the "shrinking effect" of balanced rearrangements. Its usefulness will be illustrated in a toy example in Section 2.3.1.

2.2. Dirichlet Correlation

The Dirichlet distribution (3) is conjugate prior of the multinomial, which is quite natural for compositional data arisen from independent *components*.

$$\mathrm{Dir}(\mathbf{x} \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{N}\alpha_i)}{\prod_{i=1}^{N}\Gamma(\alpha_i)}\prod_{i=1}^{N} x_i^{\alpha_i - 1} \tag{3}$$

Parameters $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)$ could be summarized by the Dirichlet precision $\sum_{i=1}^{N}\alpha_i$ and the Dirichlet

[2]Actually, as we will show, it also increases the Dirichlet correlation among components, which is undesirable.

mean $(\alpha_1, \alpha_2, \cdots, \alpha_N)/\sum_{i=1}^{N} \alpha_i$. The Dirichlet mean actually encodes the expectation of each *component*:

$$\mathbf{E}_{\mathrm{Dir}}(x_i) = \alpha_i / \sum_{j=1}^{N} \alpha_j \tag{4}$$

Without domain knowledge, we assume the *components* in original data to be equally important. According to Proposition 2, the feature extraction process should not "prefer" any new *component*. We therefore adopt a uniform Dirichlet mean:

$$\mathrm{Dir}(\mathbf{x} \mid \alpha_0) = \frac{\Gamma(N\alpha_0)}{\Gamma(\alpha_0)^N} \prod_{i=1}^{N} x_i^{\alpha_0 - 1} \tag{5}$$

The traditional concept of "correlation" (Pearson product-moment correlation coefficient) encodes linear relationships between *components* (variables). With strong linear relationships, some *components* are redundant and the total amount of information declines. In information theory, the amount of information is measured by "uncertainty" of a distribution. The Gaussian distribution with larger variances is more "uncertain", thus is preferred in PCA. For the Dirichlet distribution (5), a smaller α_0 indicates higher "uncertainty" (amount of information) and less "correlation" among the *components* (see Figure 2), which coincides with the traditional statistical interpretation of "correlation". Hence we define correlation for compositional data in terms of α_0:

Definition 3 (Dirichlet Correlation) *Given i.i.d. compositional data set $\mathcal{X} = \{\mathbf{x^1}, \mathbf{x^2}, \cdots, \mathbf{x^M}\}$ arisen from (5), the **Dirichlet correlation** among the components with respect to $\mathcal{X}$ is defined as the maximum likelihood estimation of α_0.*

Note that α_0 is the overall (not pairwise) "correlation" among *all* components.

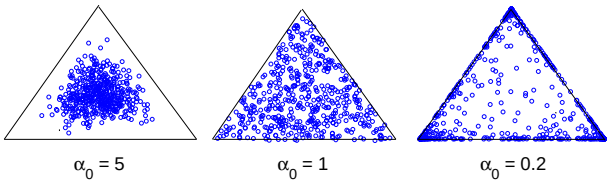

Figure 2. Points sampled from the univariate Dirichlet distribution (5) on $\mathbb{S}^3$ with different α_0.

The intuitive interpretation of the Dirichlet correlation is shown in Figure 2: 1) when $\alpha_0 > 1$, the distribution is bump-shaped, where the *components* are highly correlated and are likely to mix together in samples; 2) when $\alpha_0 = 1$, the distribution is uniform, and any proportion of mixture is equally preferred; 3) when $\alpha_0 < 1$, the distribution is valley-shaped with peaks at simplex vertices, and the *components* are better de-correlated such that the components present themselves as more purified elements in the data samples.

With the specially designed transform family and correlation measure for compositional data, we define Dirichlet component analysis (DCA) as follows:

Definition 4 (Dirichlet Component Analysis) *Given i.i.d. compositional data set $\mathcal{X} = \{\mathbf{x^1}, \mathbf{x^2}, \cdots, \mathbf{x^M}\}$ with N components, and the target dimension K, **Dirichlet component analysis** (DCA) applies a balanced rearrangement $\bar{\mathbf{R}}_{K\times N}$ and a regularization on $\mathcal{X}$ to minimize the Dirichlet correlation among the resulted K components:*

$$\bar{\mathbf{R}} = \operatorname*{argmin}_{\mathbf{R}} \operatorname*{argmax}_{\alpha_0} \mathrm{Dir}(\widetilde{\mathbf{R}(\mathcal{X})} \mid \alpha_0) \tag{6}$$

where $\widetilde{\mathbf{R}(\mathcal{X})}$ denotes that we first apply balanced rearrangement $\mathbf{R}$ to $\mathcal{X}$, and then apply a *regularization* according to Definition 2. The *i.i.d.* assumption is for factorization of the joint likelihood. The optimization problem will be discussed in Section 3.

2.3. Illustrative Examples

2.3.1. Example 1: Composition of Rocks in Geology

In this example, suppose that some rock samples are collected in a geological study in an attempt to analyze their composition. Original representation of each rock sample is a point in $\mathbb{S}^3$ (see Figure 3 left) indicating relative proportion of 3 minerals. The data points demonstrate three peaks that correspond to three substances that have fixed compositions in terms of the minerals. These peaks are formed because the formation of different substances depends on certain geological factors that vary from site to site. Hence a particular substance tends to dominate rock samples collected from some particular site. The substances had been decomposed by the chemical tests on the rocks, so that we only observe proportions of minerals.

Given the target dimension of three, DCA obtains a new representation of the rock samples (see Figure 3 right). The learned new components correspond to three underlying substances in the rock samples. Three peaks are found near the vertices of the simplex, which indicates that the new *components* are "decorrelated" in the sense that the samples tend to be

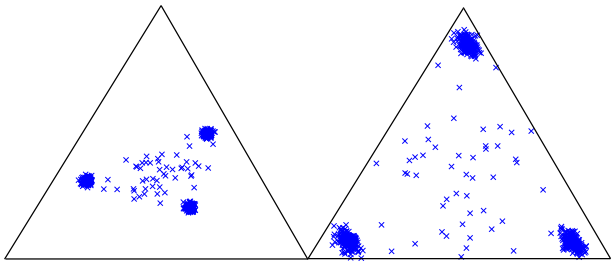

Figure 3. Left: synthetic data, composition of rock samples (small 'x') in terms of the old *components*. Right: representation in terms of new *components* (right) obtained by optimization algorithm discussed in Section 3.

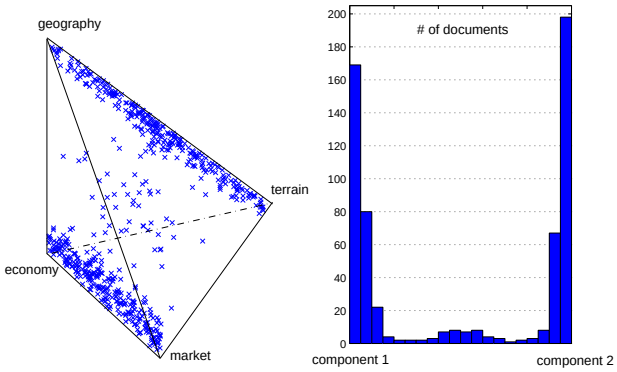

Figure 4. Left: Term frequencies of four words on $\mathbb{S}^4$, each small 'x' denotes a document. The data are synthetic. Right: new representation obtained by the optimization algorithm discussed in Section 3. The histogram illustrates the distribution of documents on $\mathbb{S}^2$.

explained by individual *components* instead of linear combinations of multiple *components*. This effect of de-correlation could be interpreted analogous to PCA. In PCA, we diagonalize the covariance matrix in order that variance in data is separately "explained" by individual variables rather than linear combinations of multiple variables. In our representation, we reveal more information about the rocks' substances because the individual components are easier to explain in further statistical analysis. In contrast, we note that PCA cannot be used to solve this rock analysis problem because by its nature this problem cannot be resolved through an orthogonal transformation.

2.3.2. Example 2: Term Frequencies in Document Retrieval

We consider a simplified bag-of-words model for document retrieval, where relative frequency values of four terms are counted in a set of documents. Each document is represented as a point in $\mathbb{S}^4$ (see Figure 4 left).

Predictably, many documents would mention both "economy" and "market" a lot, and many documents would mention both "terrain" and "geography" a lot, which gives rise to two ridge-shaped modes, corresponding to two underlying classes in these documents (one concerns economical issues, and the other discusses geological issues). Reducing the dimensionality is very likely to boost the prediction performance in classification tasks because it helps avoid overfitting (the curse of dimensionality), especially in more sophisticated high-dimensional document datasets.

Given the target dimension of two, DCA identifies two latent *components* (see Figure 4 right). The projection actually merges two pairs of semantically close *components*, and the resulting representation best preserves the information that distinguishes the two classes. Note that as an unsupervised approach, DCA cannot see any class label—all it does is minimizing the Dirichlet correlation. Although applying PCA to this toy case may have similar effects, our approach greatly outperforms PCA in higher dimensional cases, because it is specially designed for compositional data (as shown on a real-world dataset in Section 4.2).

3. Optimization

The optimization problem of DCA as defined in (6) lacks an explicit analytical loss function. Moreover, the *regularization* operator adds to the difficulty in identifying gradients or judging convexity in the parameter space.

Maximum likelihood estimation of Dirichlet precision can be carried out efficiently (Minka, 2003). The solution space is closed with respect to weighted average (Proposition 3), which motivates us to use the *genetic algorithm* (Goldberg, 1988), in which the weighted average serves as the *crossover* operator[3]. Although *genetic algorithm* is generally inefficient, it is still tractable with additional acceleration tricks. Nevertheless, genetic algorithm is just one of many choices in the optimization of DCA.

The algorithm is formalized in Algorithm 1, where "BR" is abbreviation for balanced rearrangement matrix; "DC" is abbreviation for Dirichlet correlation; "MAX" is the maximum number of iterations allowed; "SIZE" is the size of population. The *fitness* score is

[3]We do not use the *mutation* operator in our algorithm.

Algorithm 1 Genetic Algorithm for DCA

Input: dataset $\mathcal{X} \subset \mathbb{S}^N$, target dimension K
Initialize population of BR, denoted as P_0.
for $iter = 0$ **to** MAX $- 1$ **do**
  for $j = 0$ **to** SIZE $- 1$ **do**
    Apply BR_j in P_{iter} to $\mathcal{X}$
    Apply the regularization operator
    Estimate DC for transformed data
  end for
  Find minimum DC in P_{iter}
  if converged **then**
    break
  end if
  Put the BR with minimum DC into P_{iter+1}
  Compute *fitness* score for all BR
  Reduce SIZE
  for $j = 1$ **to** SIZE $- 1$ **do**
    Sample two BR from P_{iter}, probability proportional to their *fitness* scores
    Put their average, weighted by *fitness* scores, into P_{iter+1},
  end for
end for

computed as:

$$fitness = -\log(\min(\frac{DC}{\text{median } DC}, 1)), \qquad (7)$$

where "median DC" is the median Dirichlet correlation in current population. This is a key trick to accelerate the algorithm, because it prunes half of the population by assigning zero *fitness* scores. The pruning is based on the intuitive observation that: 1) the *regularization* factor is a continuous function of the BR matrix given the dataset $\mathcal{X}$; 2) the Dirichlet precision is a continuous function of the *regularization* factor and BR matrix. Hence the target function is approximately continuous and smooth in the solution space. Retaining 50% good candidate solutions in each generation is sufficient. Since the total diversity of the population diminishes, the population size could be reduced accordingly in each iteration.

4. Experimental Results

4.1. The Llobregat River Basin Hydrogeochemistry

We investigate the hydrogeochemistry dataset from the Llobregat River Basin (northeast Spain)[4] with DCA. This dataset had been studied in (Tolosana-Delgado, 2005) using factor analysis under the log-ratio framework, with which they obtained interpretable latent factors. Applying our approach on this dataset yields even more interesting results.

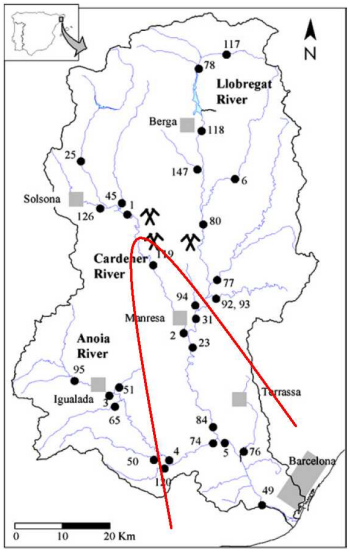

Figure 5. Sampling sites in the Llobregat River Basin, classified as "upstream" and "downstream" by the red line.

The dataset consists of 485 samples, each being a 14 dimensional compositional vector representing the concentrations of major ions (e.g. H^+, Na^+, NH_4^+, Cl^-, HCO_3^-, etc.) in the water samples. These samples are collected monthly over a certain period of time from 31 sites in the Llobregat River Basin. We classify the sites into two categories: *upstream* and *downstream*, separated by the red line (see figure 5). The 485 water samples are also classified into two categories according to the site from which they had been collected.

DCA is applied on this dataset with a target dimension of three to facilitate visualization. Visualization of high-dimensional data is crucial in disciplines such as geology, chemistry, etc., because it facilitates further analysis by domain experts.

Interestingly, although there is no location information in this dataset (locations are known from labels unseen for DCA), the two categories are well separated in the latent representation (see Figure 6). This underlying pattern is attributable to various geological and anthropogenic factors thoroughly described in (Tolosana-Delgado, 2005), which we omit here for brevity. These new patterns that are discovered by DCA was not reported in (Tolosana-Delgado, 2005), a fact highlighting the power of DCA in knowledge discovery.

[4]This dataset "Hydrochem.txt" is available online at: http://rss.acs.unt.edu/Rdoc/library/compositions/data/

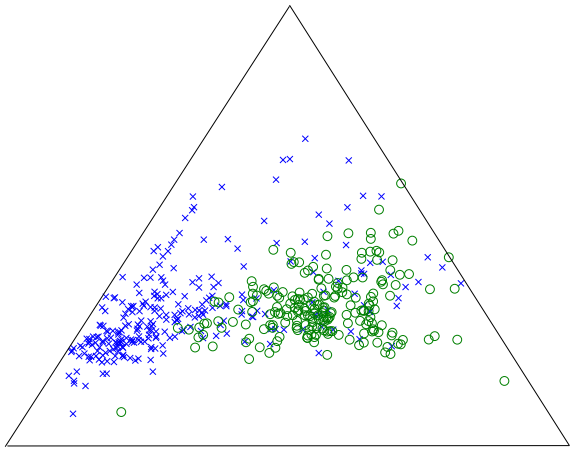

Figure 6. Latent representation of the hydrogeochemistry dataset on $\mathbb{S}^3$, learned by DCA. The two classes are well separated, see text for explanation.

4.2. Twenty newsgroups dataset

Using the 20 newsgroup data set [5], we consider a classification task of the "alt" class (798 documents) versus the "misc" class (965 documents). We show that our approach avoids overfitting and improves the prediction accuracy when we train the classifiers with a very small number of training examples, in which case the problem of overfitting could be the severest. In the preprocessing step, the "stop words" and scarce words with less than 10 total occurrences are removed. Thus the dataset we used consists of 1763 documents, where each document is represented by a 2711 dimensional sparse vector of relative word frequency values, which satisfy the constant-sum constraint.

The dimensionality is reduced to K with DCA, PCA, and LDA (latent Dirichlet allocation) (Blei, 2003), respectively. We then used a linear SVM to classify these low dimensional representations as well as the original high dimensional data for comparison. Performance results on the test test dataset are plotted with a varying number of training samples (see Figure 7) for target dimensions of $K = 10$, 20, and 50. Different choices of training data may affect the prediction performance, especially in our case where the size of training set is very small. So the performance results in our experiments are averaged over 500 different random choices of the training set. The advantage of DCA in improving prediction is clear when comparing to other techniques with the same target dimensionality, especially with very small training sets.

[5]The dataset is available at http://people.csail.mit.edu/ jrennie/20Newsgroups/

Although the highly specialized technique for bag-of-words data (LDA) beats our approach in some cases, these cases are extreme (the target dimension is 10 and the number of training samples exceeds 30). The results also justify the applicability of DCA on sparse compositional data, for which the traditional *log-ratio* framework (Aitchison, 1982) is not applicable.

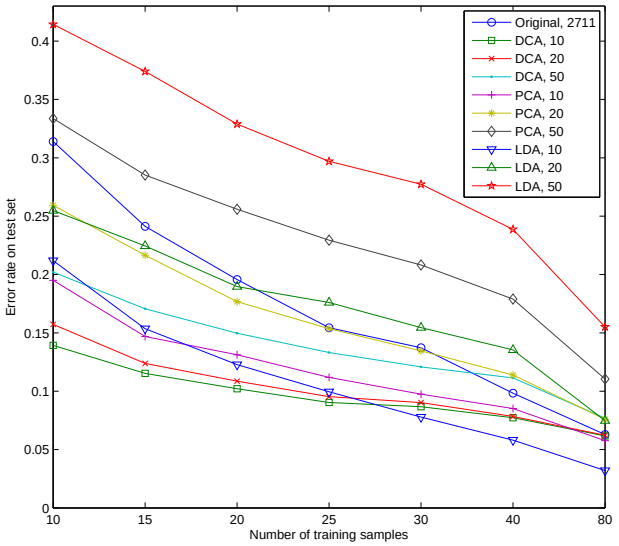

Figure 7. The "alt" versus "misc" classification performances using linear SVM. Representations with different dimensionality obtained by different methods are compared. E.g., "LDA, 10" indicates 10 dimensional representation obtained by latent Dirichlet allocation.

To make the optimization step tractable for high-dimensional data, we employed several variations in implementation. In order to reduce the solution space, we used a restricted family of transformations rather than general rearrangements. The restricted family is "amalgamations" (binary rearrangement matrices) introduced in (Aitchison, 1982), and the "balanced" requirement is not imposed. This is inspired from the toy example in Section 2.3, for which the optimal rearrangement matrix is actually an amalgamation.

5. Related Work

Feature extraction for de-correlating and reducing variables date back to K. Pearson's original idea (Pearson, 1901) on PCA. There have been a large body of research papers in the statistics and machine learning literature that address this issue, including ICA (Hyvärinen, 2001), kernel PCA (Schölkopf, 1998), etc. Directed and undirected graphical models (Blei, 2003; Welling, 2004) have also been exploited to handle this

problem, where they treat the target variables as latent nodes in the graph. Besides, the manifold assumption motivates a family of non-linear methods (Tenenbaum, 2000; Roweis, 2000), in which they use coordinates on the manifold to encode original high dimensional data.

Statistical analysis of compositional data has received a lot of concern since J. Aitchison's seminal work (Aitchison, 1982). The author proposed to transform from $\mathbb{S}^N$ to $\mathbb{R}^{N+1}$ by *log-ratio* functions, and transplanted PCA to $\mathbb{S}^N$ under the log-ratio framework (Aitchison, 1983). Our approach is an alternative PCA-like technique on $\mathbb{S}^N$, which focuses on different statistical properties (Dirichlet correlation) of data. Moreover, the *log-ratio* is not well-defined for sparse compositional data. In contrast, our approach do not have this problem. Algebraic-geometric structures (Pawlowsky-Glahn, 2001) on the simplex had been investigated, which facilitate analysis of relationship among compositional data points. Unsupervised metric learning for compositional data had been addressed in the machine learning literature (Lebanon, 2003; Wang, H.-Y., 2007).

6. Discussion

A major unresolved issue in the DCA framework is the theoretical implication of the *regularization* operator (see Figure 1), which is not compatible with the popular *log-ratio* framework, because it does not preserve the ratio between different *components*. Nevertheless, the *regularization* operator preserves Euclidean geometrical properties such as distance (allowing a constant scaling factor) and angle. Although these properties are not emphasized in the log-ratio framework, they are nonetheless meaningful as long as classification or regression tasks are concerned.

Acknowledgement

This work was supported in part by NKBRPC No. 2004CB318000, NHTRDP 863 Grant No. 2006AA01Z302, and No. 2007AA01Z336. Qiang Yang thanks Hong Kong CERG grants 621307 and and CAG grant HKBU1/05C.

References

Aitchison, J. (1982). "The statistical analysis of compositional data", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol.44, No.2, pp.139-177

Aitchison, J. (1983). "Principal component analysis for compositional data", *Biometrika*, Vol.70, No.1, pp.57-65

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

Goldberg, D.E. and Holland, J.H. (1988) "Genetic Algorithms and Machine Learning", *Machine Learning* Volume 3, Numbers 2-3.

Hyvärinen, A., Karhunen, J., Oja, E. (2001). "*Independent Component Analysis*", John Wiley & Sons.

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

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

Pawlowsky-Glahn, V. and Egozcue, J.J. (2001). "Geometric approach to statistical analysis on the simplex", *Stochastic Environ. Res. Risk Assess. (SERRA)*, v.15, no.5, pp.384-398.

Pearson, K. (1896). "On a form of spurious correlation which may arise when indices are used in the measurements of organs", *Proceedings of the Royal Society of London*, 60, pp.489-502

Pearson, K. (1901). "On lines and planes of closest fit to systems of points in space", *Philosophical Magazine*, 2, (6), pp.559-572

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

Schölkopf, B., Smola, A., Müller, K.R. (1998). "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation*, Vol 10, p.1299-1319,

Tenenbaum, J., Silva, V., Langford, J. (2000). "A global geometric framework for nonlinear dimensionality reduction", *Science*, 290, pp.2319-2323

Tolosana-Delgado, R., Otero, N., Pawlowsky-Glahn, V. and Soler, A. (2005). "Latent Compositional Factors in the Llobregat River Basin (Spain) Hydrogeochemistry", *Mathematical Geology*, Vol.37, No.7, pp.681-702

Wang, H.-Y., Zha, H., Qin, H. (2007). "Dirichlet aggregation: unsupervised learning towards an optimal metric for proportional data", In *Proceedings of the 24th ICML*

Welling, M., Rosen-Zvi, M., Hinton, G. (2004). "Exponential family harmoniums with an application to information retrieval", In *NIPS*, volume 16