DIRICHLET MIXTURE MATCHING PROJECTION FOR SUPERVISED LINEAR DIMENSIONALITY REDUCTION OF PROPORTIONAL DATA

Walid Masoudimansour

Department of Electrical and Computer Engineering
Concordia University
Montreal, QC, Canada
Email: w_masou@encs.concordia.ca

Nizar Bouguila

Concordia Institute for Information Systems Engineering
Concordia University
Montreal, QC, Canada
Email: nizar.bouguila@concordia.ca

ABSTRACT

An effective novel algorithm to reduce the dimensionality of labeled proportional data is presented which uses an optimal linear projection to project the data into a low-dimensional space. Assuming that each class of the projected data is generated by a mixture of Dirichlet distributions, KL-divergence is used as a dissimilarity measure to maximize the mutual information of projected classes, thus improving separability. Finally, genetic algorithm is used to find such optimal projection. The proposed algorithm is designed as a preprocessing step for binary classification of proportional data, however, it can project multimodal data as well due to use of mixtures and, therefore, can be used for multiclass classification. Experiments show that the proposed technique is effective, and constantly produces better results compared to well-known algorithms from the same category.

Index Terms— Dimensionality reduction, Feature extraction

1. INTRODUCTION

In the past decades due to advancements in technology enormous amounts of data have been collected for different applications. These data typically have tremendous number of features and therefore, their analysis is hindered by several phenomenas such as curse of dimensionality [1] for which dimensionality reduction (DR) is an effective solution. DR methods aim at embedding high-dimensional data in low-dimensional space such that relevant features of the data are preserved. These techniques can be divided into two major categories. Linear DR techniques such as Principal Component Analysis (PCA) [2], Factor Analysis (FA) [3], Fisher Discriminant Analysis (FDA) [4], Local Fisher Discriminant Analysis (LFDA) [5, 6], and Locality Preserving Projection (LPP) [7] use a linear transform to project the data into a lower dimensional space while exploiting second-order statistics of the data. Furthermore, any technique that does not use a linear transform is called a non-linear technique. Examples of such techniques are Kernel PCA (KPCA) [8], Maximum Variance Unfolding (MVU) [9], and Locally Linear Embedding (LLE) [10]. From another perspective, DR methods can be classified to unsupervised and supervised algorithms. Unsupervised DR algorithms reduce data dimensionality without taking advantage of data labels, however, when labeled data are available, the more effective supervised DR algorithms can be used to map the data into a lower dimensional space. Linear supervised DR methods have been proven to be highly effective, however, they may not be able to tackle some problems properly. For example, LPP may fail to handle certain cases of multimodal data [5], and LFDA cannot find the optimum projection due to sparsity of data as it is shown in Section 3. Therefore, to find an effective algorithm that tackles such problems we introduce a novel DR method referred to as Dirichlet Mixture Matching Projection (DMMP), which is specifically devised for proportional data. Proportional data consist of data with non-negative values for which each feature vector sums to one. These data are encountered, for instance, in document and image classification (using visual bag of words model). Considering that the support of Dirichlet distribution is consistent with proportional quantities, it has been used frequently in the literature for modeling such data [11, 12, 13, 14, 15, 16, 17]. DMMP is a linear supervised algorithm and consists of projecting the data using an optimal linear transform and then, matching a mixture of Dirichlet distributions to each class of data separately such that the mutual information of the two densities is maximized. Despite the linearity of the method, due to non-linearity of this process conventional optimization techniques are not efficient to solve this problem. Therefore, genetic algorithm (GA) has been used to find a good candidate for the above transform. The proposed algorithm is highly effective and performs well on multimodal data due to usage of mixture models. The rest of this paper has been organized as follows. In Section 2, the problem is stated and the proposed method is discussed in details. In Section 3, the performance of the algorithm is evaluated using several examples and its computational complexity is discussed. Finally, some concluding remarks are drawn in Section 4.

2. PROPOSED METHOD

Consider M samples of proportional data in an N dimensional space and let each sample be represented by column vector $\mathbf{x}_i$ such that $x_{i,j} \geq 0$ and $\sum_{j=1}^{N} x_{i,j} = 1$ where $1 \leq i \leq M,\ 1 \leq j \leq N$ and $x_{i,j}$ is the j-th element of the i-th sample $\mathbf{x}_i$. Let the data be populated in a matrix X of which the columns consist of $\mathbf{x}_i$s. The proposed method projects this corpus of data from the N dimensional space to K dimensional space such that $K < N$. Let this projection be denoted by P, and therefore $Y = PX$ where Y represents the projected data in the K dimensional space populated column-wise. In the rest of this paper, the elements of P will be denoted by $\rho_{r,s}$ where $1 \leq r \leq K$ and $1 \leq s \leq N$. It is worth mentioning that the necessary and sufficient conditions for P such that the projected data remain proportional are [13]

$$\rho_{r,s} \geq 0,\ \sum_{r=1}^{K} \rho_{r,s} = 1, \quad 1 \leq r \leq K, 1 \leq s \leq N \tag{1}$$

To estimate the generating distribution of the projected data for a projection P, we consider the Dirichlet distribution since its support is consistent with proportional data [11]. Assuming a mixture of Dirichlet distributions for projected samples of each of two classes (namely, class 0 and class 1), EM algorithm is used to estimate the distributions. Note that the mixture model allows modeling multimodal data. Let the training samples from classes 0 and 1 be shown by the sets C_0 and C_1 , respectively, such that $|C_0| = M_0$ and $|C_1| = M_1$. Assuming independent samples with identical distributions in each class, the likelihood of the parameter $\boldsymbol{\alpha}^{(\kappa)}, \kappa \in \{0, 1\}$ can be written as

$$L(\boldsymbol{\alpha}^{(\kappa)}|Y) = \prod_{i=1}^{M_\kappa} \left(\sum_{j=1}^{Q} \left(\phi_j^{(\kappa)} \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k,j}^{(\kappa)})}{\prod_{k=1}^{K} \Gamma(\alpha_{k,j}^{(\kappa)})} \prod_{k=1}^{K} (y_{k,i}^{(\kappa)})^{\alpha_{k,j}^{(\kappa)}-1} \right) \right) \tag{2}$$

where $y_{k,i}^{(\kappa)}$ is the k-th element of the i-th projected sample from class κ and the matrix $Y^{(\kappa)}$ consists of all samples of that class as its columns. The parameter $\boldsymbol{\alpha}$ is a matrix with Q columns. Each column of this matrix is a vector of the same size of the vector $\mathbf{y}_k$ with non-negative elements corresponding to one component in the mixture. Moreover, ϕ_js are the priors (mixing weights) where $\sum_{j=1}^{Q} \phi_j = 1$ and superscript (κ) is used to show class-specific parameters. To maximize this log-likelihood, EM algorithm is used. Assuming latent variable $Z_{i,j}$ is equal to 1 if data sample y_i comes from j-th mixture and zero otherwise, an iterative algorithm to find the maximum will consist of the following two steps

Step 1 (Expectation): In this step, the mixing coefficients are calculated as

$$\phi_j^{(\kappa)}(t+1) = \frac{\sum_{i=1}^{M_\kappa} \hat{Z}_{i,j}^{(\kappa)}(t)}{M_\kappa} \tag{3}$$

where t is the iteration number.

Step 2 (Maximization): To maximize the log-likelihood, the following set of equations for which the solution is $\alpha_{m,n}^{(\kappa)}(t+1)$ must be solved for $1 \leq m \leq K$ and $1 \leq n \leq Q$

$$\sum_{i=1}^{M_\kappa} \hat{Z}_{m,n}(t) \frac{\partial}{\partial \alpha_{m,n}^{(\kappa)}} \log \Bigg(\sum_{j=1}^{Q} \phi_j^{(\kappa)} \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k,j}^{(\kappa)})}{\prod_{k=1}^{K} \Gamma(\alpha_{k,j}^{(\kappa)})} \times \prod_{k=1}^{K} (y_{k,i}^{(\kappa)})^{\alpha_{k,j}^{(\kappa)}-1} \Bigg) = 0 \tag{4}$$

The above set of equations can be organized as a system of non-linear equations, and therefore, solved using Newton-Raphson (NR) method. In this case, the set of equations consists of the following

$$\psi(\alpha_{j,s}^{(\kappa)}(t+1)) - \psi(\sum_{k=1}^{K} \alpha_{j,k}^{(\kappa)}(t+1)) = \frac{\sum_{i=1}^{M_\kappa} Z_{i,j}^{(\kappa)}(t) \log y_{i,s}}{\sum_{i=1}^{M_\kappa} Z_{i,j}^{(\kappa)}(t)} \tag{5}$$

where $1 \leq s \leq K, 1 \leq j \leq Q$ and $\psi(.)$ is the Digamma function. To solve this system of equations, NR method requires inversion of the Jacobian of the system. Note that since the equations are decoupled with respect to n the resulting Jacobian is block diagonal, and therefore, inverting such a Jacobian can be carried out block by block. Assuming cubic complexity for the inversion operator, this results in $\mathcal{O}(K^3)$ complexity instead of $\mathcal{O}(Q^3K^3)$ which is a considerable improvement.

In the next step, we use KL-divergence [18] as a measure of distance between the class distributions. Since calculating the KL-divergence between two Dirichlet mixtures is not straightforward, following a method analogous to the one introduced in [19], one can find an approximation for this value. Assume $f(x) = \sum_{i=1}^{Q_f} w_i f_i(x)$ and $g(x) = \sum_{j=1}^{Q_g} u_j g_j(x)$ are two Dirichlet mixtures such that $\sum_{i=1}^{Q_f} w_i = 1$ and $\sum_{j=1}^{Q_g} u_j = 1$ where Q_f and Q_g are the number of components in each mixture. Starting with the definition of KL-divergence we obtain

$$KL(f(x), g(x)) = \int f(x) \log f(x) \mathrm{d}x - \int f(x) \log g(x) \mathrm{d}x \tag{6}$$

Consider the first term of the above equation. Using Jensen's inequality, a lower bound can be calculated as

$$\begin{aligned} &\int f(x) \log f(x) \mathrm{d}x = \sum_{i=1}^{Q_f} \int w_i f_i(x) \log \sum_{j=1}^{Q_f} w_j f_j(x) \mathrm{d}x = \\ &\sum_{i=1}^{Q_f} \int w_i f_i(x) \log \sum_{j=1}^{Q_f} \zeta_{i,j} \frac{w_j f_j(x)}{\zeta_{i,j}} \mathrm{d}x \geq \\ &\sum_{i=1}^{Q_f} \int w_i f_i(x) \sum_{j=1}^{Q_f} \zeta_{i,j} \log \left(\frac{w_j f_j(x)}{\zeta_{i,j}} \right) \mathrm{d}x \end{aligned} \tag{7}$$

where $\zeta_{i,j}$ are to be determined such that the lower bound of Eq.7 is maximized. Note that, Jensen's inequality imposes that $\sum_{j=1}^{Q_f} \zeta_{i,j} = 1$ for $1 \leq i \leq Q_f$. Defining $\bar{J}$ as the following cost function

$$\bar{J} = \sum_{i=1}^{Q_f} \int w_i f_i(x) \sum_{j=1}^{Q_f} \zeta_{i,j} \log\left(\frac{w_j f_j(x)}{\zeta_{i,j}}\right) \mathrm{d}x \quad (8)$$
$$+ \sum_{i=1}^{Q_f} \left(\lambda_i \sum_{j=1}^{Q_f} \zeta_{i,j} - 1\right) + \sum_{i=1}^{Q_f} \mu_i(-\zeta_{i,j})$$

and considering the KKT conditions [20], one can solve for ζs as

$$\zeta_{i,j} = \frac{w_j e^{H(f_i,f_j)}}{\sum_{k=1}^{Q_f} w_k e^{H(f_i,f_k)}} \quad (9)$$

where

$$H(f_i, f_j) = \int f_i(x) \log f_j(x) \mathrm{d}x \quad (10)$$

Substituting $\zeta_{i,j}$ in the last term of Eq.7 we obtain

$$\int f(x) \log f(x) \mathrm{d}x \geq \sum_{i=1}^{Q_f} w_i \log\left(\sum_{k=1}^{Q_f} w_k e^{H(f_i,f_k)}\right) \quad (11)$$

Using a similar method, one can find the lower bound of the second term of Eq.6 as the following

$$\int f(x) \log g(x) \mathrm{d}x \geq \sum_{i=1}^{Q_f} w_i \log\left(\sum_{k=1}^{Q_g} u_k e^{H(f_i,g_k)}\right) \quad (12)$$

and therefore, the approximate value of the KL-divergence of two Dirichlet mixtures simplifies to

$$KL_L(f(x), g(x)) = \sum_{i=1}^{Q_f} w_i \log \frac{\sum_{k=1}^{Q_f} w_k e^{-KL(f_i,f_k)}}{\sum_{k=1}^{Q_g} u_k e^{-KL(f_i,g_k)}} \quad (13)$$

where we have used the fact that

$$KL(f_i, f_j) = H(f_i, f_i) - H(f_i, f_j) \quad (14)$$

Finally, the KL-divergence of two Dirichlet distribution with parameters $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ is

$$KL(f_i, f_j) = E\left\{\log \frac{f_i(x)}{f_j(x)}\right\} = \quad (15)$$
$$E\left\{\log\left(\frac{\Gamma(\sum_{k=1}^K \alpha_{i,k}) \prod_{k=1}^K \Gamma(\alpha_{j,k})}{\Gamma(\sum_{k=1}^K \alpha_{j,k}) \prod_{k=1}^K \Gamma(\alpha_{i,k})} \prod_{k=1}^K y_k^{\alpha_{i,k}-\alpha_{j,k}}\right)\right\}$$
$$= \log\left(\frac{\Gamma(\sum_{k=1}^K \alpha_{i,k}) \prod_{k=1}^K \Gamma(\alpha_{j,k})}{\Gamma(\sum_{k=1}^K \alpha_{j,k}) \prod_{k=1}^K \Gamma(\alpha_{i,k})}\right) + E\left\{\log \prod_{k=1}^K y_k^{\alpha_{i,k}-\alpha_{j,k}}\right\}$$

where $E\{.\}$ denotes the expected value with respect to $f_i(x)$. The last term in the above equation can be simplified to

$$E\left\{\log \prod_{k=1}^K y_k^{\alpha_{i,k}-\alpha_{j,k}}\right\} = \sum_{k=1}^K (\alpha_{i,k} - \alpha_{j,k}) E_{f_i(x)}\{\log y_k\} =$$
$$\sum_{k=1}^K (\alpha_{i,k} - \alpha_{j,k}) \left(\psi(\alpha_{i,k}) - \psi(\sum_{k=1}^K \alpha_{i,k})\right) \quad (16)$$

Substituting this term in Eq.15 yields

$$KL(f_i, f_j) = \log\Gamma(\sum_{k=1}^K \alpha_{i,k}) - \log\prod_{k=1}^K \Gamma(\alpha_{i,k}) - \log\Gamma(\sum_{k=1}^K \alpha_{j,k}) \quad (17)$$
$$+ \log\prod_{k=1}^K \Gamma(\alpha_{j,k}) + \sum_{k=1}^K (\alpha_{i,k} - \alpha_{j,k}) \left(\psi(\alpha_{i,k}) - \psi(\sum_{k=1}^K \alpha_{i,k})\right)$$

Finally, substituting the above value in Eq.13 one can find an approximation for the KL-divergence of two Dirichlet mixtures. It is worth mentioning that KL-divergence is not symmetric, and therefore, the following distance measure is used as an alternative

$$KL_s(f(x), g(x)) = KL_L(f(x), g(x)) + KL_L(g(x), f(x)) \quad (18)$$

The above symmetric KL-divergence is a measure of dissimilarity (distance) of the two data classes. To maximize the distance one can use this measure along with an optimization algorithm which results in a projection that separates the data effectively. Despite the linearity of the projection, and considering the non-linearity embedded in the process of calculating the optimum transform, heuristic search algorithms are proper tools to solve this problem. GA is one of the most well-known of such heuristic search tools and it has been used effectively in the proposed technique. The search is performed in the space of all matrices that conform to Eq.1. Starting with an initial population, the data is projected using each member (matrix P). Then, a Dirichlet mixture is matched to each class in the new low-dimensional space. Note that the mixture facilitates the matching of multimodal classes. In the next step, the approximate value of the symmetric KL-divergence between the two classes is used as a fitness function for each member of the population, and the next generation of population is generated based on this fitness function. In the final step, the fittest member of the current population is chosen as the optimum projection.

3. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of DMMP algorithm using real data. To demonstrate the effectiveness of the proposed technique, it is compared against four supervised linear methods: LFDA [5], SOLPP [21], SLPP [7], and LSDA [22]. We have used two datasets for this matter. Note that, in all examples, SVM is used to classify the data, and each experiment is repeated 4 times while 5-fold cross validation is used to find the average and standard deviation of classification accuracy. For instance, in Example 1, the first test yields average accuracy of $\%95.07$ and standard deviation of 0.01 for DMMP.

Example 1: In this example, we use the `20-Newsgroups` dataset to build the bag of words model while stop words and low-frequency words are ignored. Two tests have been performed for unimodal and multimodal data and the dimensionality is reduced to 3. Tables 1 and 2 show the resulting

classification accuracies and, as it can be seen, DMMP outperforms the rest of the algorithms. Also note that LFDA could not solve the eigenvalue problem in this experiment, and thus, produced no results. Furthermore, SOLPP produces unacceptable results in some cases which is due to sparsity of the data.

Table 1: Classification accuracy (%) of 20-Newsgroups dataset for unimodal data, 1500 original features and target dimension 3.

Classes	SOLPP	SLPP	LSDA	DMMP
MS-Windows vs. Hockey	70.12±0.14	84.15±0.02	83.66±0.02	**95.07±0.01**
Electronics vs. Guns	52.20±0.04	70.16±0.03	69.90±0.03	**84.72±0.02**
Baseball vs. Politics	51.89±0.04	74.12±0.03	73.38±0.03	**87.19±0.02**
Autos vs. Med	53.78±0.04	71.91±0.03	71.01±0.03	**86.84±0.02**
Crypt vs. Space	62.34±0.11	75.59±0.03	75.01±0.03	**88.56±0.02**
Graphics vs. Hardware	54.60±0.03	63.71±0.04	63.93±0.04	**82.88±0.02**
Autos vs. Electronics	53.55±0.02	71.28±0.02	71.20±0.03	**79.08±0.01**

Table 2: Classification accuracy (%) of 20-Newsgroups dataset for multimodal data, 2000 original features and target dimension 3.

Classes	SOLPP	SLPP	LSDA	DMMP
Hardware vs. Graphics + Christian	50.70±0.03	79.70±0.02	78.89±0.02	**84.45±0.02**
MS-Windows vs. Hardware + Politics	54.40±0.06	69.49±0.02	68.75±0.02	**82.37±0.02**
Electronics vs. IBM + Motorcycles	58.98±0.08	77.17±0.02	76.49±0.02	**79.74±0.01**
Religion vs. Crypt + Space	55.57±0.13	71.08±0.02	69.81±0.02	**86.26±0.01**
Guns vs. Mideast + IBM	59.77±0.11	81.79±0.02	80.82±0.02	**82.61±0.02**
Forsale vs. Atheism + Christian	58.48±0.06	83.70±0.02	83.25±0.02	**94.83±0.01**
Politics vs. Windows X + Forsale	53.77±0.02	79.54±0.02	78.84±0.02	**90.02±0.02**

Example 2: In a different application, `Food-101` dataset which contains shots from 101 food types is used to demonstrate the efficacy of the proposed algorithm in image classification. Again, the test is performed for unimodal and multimodal classes after extracting image features using SIFT and constructing a visual bag of words using a dictionary size of 750 while the dimensionality is reduced to 3. The results of this test are shown in Tables 3 and 4. In this test, similar to the previous one, DMMP constantly yields better classification accuracy than other algorithms.

Table 3: Classification accuracy (%) of Food-101 dataset for unimodal data and target dimension 3.

Classes	LFDA	SOLPP	SLPP	LSDA	DMMP
Steak vs. Sashimi	75.74±0.02	52.09±0.04	75.71±0.02	75.76±0.02	**81.85±0.02**
Pizza vs. Hamburger	76.85±0.02	52.04±0.03	76.79±0.02	76.64±0.02	**82.22±0.02**
Macarons vs. Frozen Yogurt	77.45±0.02	51.51±0.03	77.37±0.02	77.21±0.02	**82.86±0.02**
Hot Dog vs. Nachos	76.32±0.02	52.51±0.03	75.94±0.02	75.26±0.02	**80.33±0.01**
Caesar Salad vs. Poutine	75.17±0.02	51.00±0.03	74.96±0.02	75.01±0.02	**81.36±0.01**
Fried Rice vs. Baklava	82.98±0.02	56.75±0.07	82.65±0.02	82.63±0.02	**88.31±0.01**

Performance and Computational Complexity: While most linear supervised DR methods use a generalized eigenvalue problem, DMMP uses a novel different approach to find an optimum projection to the lower dimension space. Note that the effect of original dimensionality of the data on this method is minor since, first, the data is projected into the desired low-dimensional space, and the rest of the algorithm is performed on K dimensional data where $K \ll N$, while the other algorithms involve inverting matrices with large sizes. Also, the major time consuming part of the algorithm is calculation of the fitness function for each member. This can be implemented efficiently using GPUs and parallel computing methods since calculation of the fitness of each member is independent of others. Furthermore, when dealing with extremely high-dimensional data, most algorithms that rely on solving the generalized eigenvalue problem fail to provide a reliable solution due to sparsity of the data and singularity of the involved matrices, while DMMP does not suffer from such problem since it projects the data into a lower dimensional space first, resulting in a non-sparse data matrix. Moreover, in high dimensions, the running time of DMMP is comparable to that of the rest of the algorithms since they involve inverting very large size matrices which is $\mathcal{O}(N^3)$. Finally, some of the algorithms used for comparison need extra parameters to be set by the user which makes them less effective than DMMP.

Table 4: Classification accuracy (%) of Food-101 dataset for multimodal data and target dimension 3.

Classes	LFDA	SOLPP	SLPP	LSDA	DMMP
Macarons vs. Frozen Yogurt + Chicken Wings	83.34±0.01	58.18±0.03	83.18±0.01	83.03±0.01	**85.36±0.01**
Crab Cakes vs. Lasagna + Oysters	69.10±0.03	56.25±0.02	69.88±0.02	69.45±0.02	**73.10±0.01**
Risotto vs. Creme Brulee + Apple Pie	79.87±0.02	55.99±0.02	78.96±0.02	78.33±0.02	**81.85±0.01**
Pho vs. Ice Cream + Hummus	87.74±0.01	56.78±0.03	87.44±0.01	86.92±0.01	**89.50±0.01**
Caprese Salad vs. Takoyaki + Ravioli	59.24±0.05	56.66±0.03	74.76±0.02	74.43±0.02	**77.46±0.01**
Pad Thai vs. Carrot Cake + Greek Salad	84.37±0.02	55.45±0.04	84.40±0.01	84.05±0.01	**85.54±0.01**
Macarons vs. Nachos + Panna Cotta	61.28±0.1	57.99±0.04	81.76±0.02	81.70±0.02	**83.30±0.01**

4. CONCLUSIONS

A novel and effective algorithm of dimensionality reduction has been introduced for labeled proportional data that tackles some problems of the existing methods. The data are assumed to be from two different classes, however, since the algorithm is able to process multimodal data, it can be used for multiclass data as well. The proposed method is a linear algorithm and finds an optimal projection to project the data such that the mutual information of projected classes is maximized. The data are projected to the lower dimensional space using a transform matrix. Then, a Dirichlet mixture is estimated as the generating distribution for each class separately, and the KL-divergence between the mixtures is also approximated using a maximized lower bound. Finally, a genetic algorithm uses this KL-divergence as fitness value to search for the best candidate for the projection. Several experiments and comparisons show that the algorithm is highly effective and, when used as preprocessing step for classification, constantly produces high classification rates compared to existing well-known linear supervised DR methods.

5. REFERENCES

[1] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, pp. 1–33, 2000.

[2] I. T. Jolliffe, "Principal Component Analysis," *Encyclopedia of Statistics in Behavioral Science*, vol. 30, no. 3, pp. 487, 2002.

[3] D. J. Bartholomew, "The foundations of factor analysis," *Biometrika*, vol. 71, no. 2, pp. 221–232, 1984.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[5] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.

[6] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local Fisher discriminant analysis for dimensionality reduction," *Journal of Machine Learning*, vol. 78, no. 1-2, pp. 35–61, 2010.

[7] X. He and P. Niyogi, "Locality preserving projections," *Neural information processing systems*, vol. 16, pp. 153, 2004.

[8] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[9] K. Q. Weinberger and L. K. Saul, "An introduction to nonlinear dimensionality reduction by maximum variance unfolding," *AAAI*, pp. 1683–1686, 2006.

[10] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding.," *Science*, vol. 290, no. 5500, pp. 2323–6, 2000.

[11] N. Bouguila and D. Ziou, "A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 107–122, 2010.

[12] W. Masoudimansour and N. Bouguila, "Dimensionality reduction of proportional data through data separation using dirichlet distribution," *Image Analysis and Recognition*, vol. 9164, pp. 141–149, 2015.

[13] H. Y. Wang, Q. Yang, H. Qin, and H. Zha, "Dirichlet component analysis: feature extraction for compositional data," in *25th international conference on Machine learning*, 2008, pp. 1128–1135.

[14] N. Bouguila, "Hybrid Generative/Discriminative Approaches for Proportional Data Modeling and Classification," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 12, pp. 2184–2202, 2012.

[15] E. Epaillard and N. Bouguila, "Proportional data modeling with hidden Markov models based on generalized Dirichlet and Beta-Liouville mixtures applied to anomaly detection in public areas," *Pattern Recognition*, vol. 55, pp. 125–136, jul 2016.

[16] E. Epaillard and N. Bouguila, *Hidden Markov models based on generalized dirichlet mixtures for proportional data modeling*, vol. 8774, 2014.

[17] W. Fan and N. Bouguila, "Expectation propagation learning of a Dirichlet process mixture of Beta-Liouville distributions for proportional data clustering," *Engineering Applications of Artificial Intelligence*, vol. 43, pp. 1–14, aug 2015.

[18] S. Kullback, *Information theory and statistics*, Dover Publications, 1997.

[19] J. R. Hershey and P. A. Olsen, "Approximating the Kullback-Leibler divergence between Gaussian mixture models," *Acoustics, Speech and Signal Processing*, vol. 4, no. 6, pp. 317–320, 2007.

[20] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Second Berkeley Symposium on Mathematical Statistics and Probability*, 1951, pp. 481–492.

[21] W. K. Wong and H. T. Zhao, "Supervised optimal locality preserving projection," *Pattern Recognition*, vol. 45, no. 1, pp. 186–197, 2012.

[22] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," *20th International Joint Conference on Artifical Intelligence*, pp. 708–713, 2007.