A Distance-Based Classifier Using Dissimilarity Based on Class Conditional Probability and Within-Class Variation

Kwanyong Lee [1] and Hyeyoung Park [2]

[1.] Department of Computer Science, Korea National Open University, Seoul, 110-791, Korea
[2.] School of Computer Science and Engineering, Kyungpook National University, Daegu, 702-701, Korea
kylee@knou.ac.kr, hypark@knu.ac.kr

Abstract: According to the rapid increase of data, the needs of intelligent data analysis and classification are also increasing. Though there have been developed various methods of classifying given data set into several pre-defined patterns, the distance-based classifier such as nearest neighbor classifier is still one of the most popular methods due to its simplicity and adaptability. However, in order to obtain good performances in practical applications, it is important to choose an appropriate distance measure considering the purpose of task and the distributional properties of data set. In this paper, we propose a new measure of similarity based on two probability densities: the class-conditional probability and the probability of within-class variation. Through statistical estimation of the probability densities using training set, it is possible to obtain an optimized measure for the given data. The efficiency of the proposed measure is confirmed by computational experiments on a few pattern recognition problems using benchmark data sets.

[Lee K, Park H. **A Distance-Based Classifier Using Dissimilarity Based on Class Conditional Probability and Within-Class Variation**. *Life Sci J* 2014;11(7):666-671] (ISSN:1097-8135). http://www.lifesciencesite.com. 95

Keywords: Pattern classification, Distance-based classifier, Nearest neighbor classifier; Distance measure, Class conditional probability, Within-class variation

1. Introduction

Classifying given data set into several pre-defined patterns is one of main tasks in the field of intelligent data analysis. There are diverse applications of intelligent data analysis, such as human identification using various types of bio-signals, text categorization, object detection from images, and so on. For example, the face recognition is to classify given facial images into a number of pre-defined patterns, each of which is corresponding to an individual person. The article categorization is to classify given text inputs into several subject groups such as politics, science, culture, and economics. The pedestrian detection is to find a sub area within a whole image, which includes a pattern of standing human. All these applications need a common technology that is called pattern classification (Duda et al., 2001).

In the studies on pattern classification, there have been developed various classifiers including Bayes classifier, multilayer perceptrons, and support vector machines (Bishop, 2006). Among them, the distance-based classifier, which is also known and K-nearest neighbor classifier, is one of the most popular one because it can be simply implemented and has shown successful performances in many practical applications (Duda et al., 2000).

When an input sample to be classified, which is usually called probe data, is given, a distance-based classifier determines the class label of the probe data based on its distance from each of the registered samples (gallery data). In the case of nearest neighbor classifier, the probe data is assigned to the class in which the nearest gallery data is included. From this procedure, one can easily realize that the performance of this type of classifiers highly depends on the method of measuring distance between probe and gallery data. Thus, it is very important to choose a good distance measure in order to achieve good performance in practical applications.

There are a number of distance functions that have been widely used in various applications. The classical distance functions such as Euclidean distance and cosine distance are the most common ones, and some statistical distances such as Mahalanobis distance have also been used as a better alternative to the fixed distances (Ashby and Ennis, 2007). Though there have been a number of studies on comparing performances of the various distance measures, it is difficult to find a general guideline on choosing a good measure for given applications, because the performances are highly depending on the given data set as well as the purpose of classification tasks.

To solve the problem, it also have been conducted many studies on finding an optimal distance measure through learning (Weinberger et al. 2005; Moghaddam et al., 1999; Lee and Park, 2003; Lee and Park, 2005). Lee and Park (2003) defined the similarity between two samples as the probability that they belong to a same class. In order to obtain an explicit function for calculating the probability, they estimated the probability density function (pdf) of within-class variation, which is defined as the difference between two samples belonging to a same class. Based on the previous works, this paper proposes a new measure of similarity through the combination of the class-conditional probability and the probability density of within-class variations. By additional use of the class-conditional probability, we expect to get more robust classifiers against noisy environment.

In the next section, the conventional distance measures used for pattern classification are briefly described. The previous works on probabilistic learning of distance is also explained in the same section. In Section 3, a novel dissimilarity measure is defined, and the nearest neighbor classifier using the proposed measure is described. Some experimental results on benchmark data sets are given in Section 4, and conclusions are made in Section 5.

2. Distance Measures for Pattern Classification

The most popular measure of distance between two $d-$ dimensional column vector, $\boldsymbol{x}_i = [x_{i1}, \dots, x_{id}]^T$ and $\boldsymbol{x}_j = [x_{j1}, \dots, x_{jd}]^T$, is defined with L_p-norm of the difference vector $\boldsymbol{x}_i - \boldsymbol{x}_j$, which can be written as

$$d_{norm\_p}(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ = L_p(\boldsymbol{x}_i - \boldsymbol{x}_j) = \sqrt[p]{\textstyle\sum_k (x_{ik} - x_{jk})^p} \qquad (1)$$

Note that this is a general form of Euclidean distance and Manhattan distance since L_1 -norm defines Manhattan distance and L_2-norm defines Euclidean distance. We can have a variety of distance functions by just changing the value of p.

Whereas L_p-norm is based on the difference between two vectors, there is another popular measure based on the inner product of two vectors, $\boldsymbol{x}_i^T \boldsymbol{x}_j$. The cosine distance is defined using inner product, which can be written as

$$d_{cos}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \frac{x_i^T x_j}{\sqrt{(x_i^T x_i)(x_j^T x_j)}}. \qquad (2)$$

Since the cosine distance measures the angular difference between two vectors, it is useful when the overlap of two vectors has more important meaning than the absolute subtraction between the corresponding elements of two vectors. Due to these distinct properties of the two measures, they often show significant difference in the classification performance as well. Therefore, in order to achieve good performance, it is very important to choose a proper measure carefully.

On the other hand, the statistical distances use some statistics of given data set to find more appropriate distance measure. Mahalanobis distance uses the covariance matrix $\boldsymbol{\Sigma_x}$ of given data set to define a proper measure for multivariate data with correlations, which is defined as

$$d_{Mah}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j), \qquad (3)$$

As a simpler version, the normalized Euclidean distance uses only diagonal elements of covariance matrix. Although these distances sometimes show improved performances, they have the limitation that they only use the distributional information of the whole data set, and ignore the class-conditional distributions, which are more important in the case of pattern classification.

To solve the problem, metric learning methods try to find an optimal distance metric for the given task and data set by using various machine learning techniques (Xing et al., 2002; Weinberger, et al., 2005). With the same motivation, there also have been probabilistic approaches to define new similarity measures between two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ based on the probability of within-class variation (Moghaddam et al.; 1999; Lee and Park, 2003). In the classification tasks, it is natural to consider that two samples are similar if their membership is same. Therefore, the similarity between two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be written as

$$S(\boldsymbol{x}_i, \boldsymbol{x}_i) = \text{Prob } [y(\boldsymbol{x}_i) = y(\boldsymbol{x}_i)], \qquad (4)$$

where $y(\boldsymbol{x})$ denotes the class label of $\boldsymbol{x}$.

In order to obtain the explicit value of the probability of (4), Lee and Park (2003) proposed the use of new random vector $\boldsymbol{\delta}$, which is the difference between two input samples, and compose the set Ω of all the difference vectors between two samples with the same class label. The set Ω can be obtained from the training set such as

$$\Omega = \{\boldsymbol{\delta}_{ij} \,|\, \boldsymbol{\delta}_{ij} = \boldsymbol{x}_i - \boldsymbol{x}_j, y(\boldsymbol{x}_i) = y(\boldsymbol{x}_i)\}. \qquad (5)$$

Using the set Ω, it is possible to estimate the probability density function $p(\boldsymbol{\delta} \,|\, \Omega)$, which can be considered as the probability of within-class variations.

The previous works (Lee and Park 2003; Lee and Park 2012) used the estimated pdf $p(\boldsymbol{\delta} \,|\, \Omega)$

to measure the similarity between two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ by calculating $p(\boldsymbol{x}_i - \boldsymbol{x}_j \mid \Omega)$. Furthermore, if we assume that the within-class variation $\boldsymbol{\delta}$ is subject to Gaussian distribution, we can get an explicit function of $p(\boldsymbol{x}_i - \boldsymbol{x}_j \mid \Omega)$ such as

$$p(\boldsymbol{x}_i - \boldsymbol{x}_j \mid \Omega) = \frac{1}{Z} exp\left\{-\frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{x}_j - \mu_\delta\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{x}_j - \mu_\delta\right)\right\} \quad (6)$$

where Z is the normalization factor; the parameters μ_δ and $\boldsymbol{\Sigma}_{\boldsymbol{\delta}}$ are the mean and covariance of random variable $\boldsymbol{\delta}$, which can be estimated by using the set Ω. This probability density function can be used as a similarity measure between two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, because high probability density implies strong likelihood that two samples belong to the same class. Based on the probability density, we can define a simpler function for measuring dissimilarity such as

$$d_{with}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\boldsymbol{x}_i - \boldsymbol{x}_j - \mu_\delta\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{x}_j - \mu_\delta\right). \quad (7)$$

Based on this measure, we propose a new measure that has more desirable properties, as we shall describe in the following sections.

3. Proposed Measure and Classifier

Although the probabilistic measure (7) defined by using within-class variation has shown successful performances in various pattern recognition problems (Moghaddam et al., 1999; Lee and Park 2003), instability and performance degradation have also been reported in some cases (Lee and Park, 2012). Since the measure is based on the estimated probability density $p(\boldsymbol{\delta} \mid \Omega)$, it is obvious that the inaccurate estimation of $p(\boldsymbol{\delta} \mid \Omega)$ leads to low classification performance. When the number of training data is not sufficient or the dimensionality of estimation parameter (μ_δ and $\boldsymbol{\Sigma}_{\boldsymbol{\delta}}$) is large, it is difficult to get good estimation of the parameters. In these situations, we cannot expect to get good classification performance using the measure $d_{with}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$. In addition, the measure $d_{with}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ only takes the within-class variations into account, and the distributional relationship among classes is ignored, which may cause the loss of important information for discriminating among all the classes. This drawback would be emphasized when the number of classes is large.

Considering these problems of the previous measure, we propose a new dissimilarity measure by combining two probabilities: the probability density of within-class variations $p(\boldsymbol{\delta} \mid \Omega)$ and the class-conditional probability density $p(\boldsymbol{x}|C_i)$. The class-conditional probability density $p(\boldsymbol{x}|C_i)$ implies the likelihood that an input feature $\boldsymbol{x}$ is observed in the ith class C_i. The estimation of $p(\boldsymbol{x}|C_i)$ can be done by using the subset of whole training set that can be defined as $C_i = \{\boldsymbol{x}|\ y(\boldsymbol{x}) = i\}$. Through the additional use of this class-conditional probability, we can utilize the information on the distribution of classes and thus can expect to increase the discriminative ability of the measure.

Using the two probability densities, we can define a new measure of similarity between a training data $\boldsymbol{x}_i$ and the newly given input $\boldsymbol{x}_{new}$, which can be written as

$$S(\boldsymbol{x}_i, \boldsymbol{x}_{new}) = p(\boldsymbol{x}_i - \boldsymbol{x}_{new} \mid \Omega)\, p(\boldsymbol{x}_{new}|C_{y(x_i)}). \quad (8)$$

This first factor in right hand of equation (8) implies the likelihood that $\boldsymbol{x}_i$ and $\boldsymbol{x}_{new}$ belong to a same class. Similarly, the second factor implies the likelihood that $\boldsymbol{x}_{new}$ is observed from the class $C_{y(x_i)}$, in which $\boldsymbol{x}_i$ is included. Consequently, the similarity measure gives large values when the class labels of $\boldsymbol{x}_i$ and $\boldsymbol{x}_{new}$ are same.

To obtain an explicit form for calculating the similarity, we simply use the Gaussian model with unit covariance matrix for class-conditional probability density. Then the simplified function for measuring dissimilarity can be given as

$$d_{prop}(\boldsymbol{x}_i, \boldsymbol{x}_{new}) = (\boldsymbol{x}_i - \boldsymbol{x}_{new} - \mu_\delta)^T \boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_{new} - \mu_\delta). + \left(\boldsymbol{x}_{new} - \mu_{y(x_i)}\right)^T \left(\boldsymbol{x}_{new} - \mu_{y(x_i)}\right). \quad (9)$$

where the parameter $\mu_{y(x_i)}$ is the mean vector of class $C_{y(x_i)}$, which can be estimated by the sample mean of the subset $C_i = \{\boldsymbol{x}|\ y(\boldsymbol{x}) = i\}$.

In order to classify a newly given probe data $\boldsymbol{x}_{new}$ using the nearest neighbor classifier, we calculate the value of dissimilarity $d_{prop}(\boldsymbol{x}_i : \boldsymbol{x}_{new})$ for all $\boldsymbol{x}_i$ in the previously registered gallery (training) set D, and find the nearest neighbor $\boldsymbol{x}_{nn}$, which can be written as

$$\boldsymbol{x}_{nn} = \mathrm{argmin}_{x_i \in D}\{\, d_{prop}(\boldsymbol{x}_i,\ \boldsymbol{x}_{new})\}. \quad (10)$$

Then we can assign $\boldsymbol{x}_{new}$ to the same class to which the nearest neighbor $\boldsymbol{x}_{nn}$belongs, such as

$$y(\boldsymbol{x}_{new}) = y(\boldsymbol{x}_{nn}). \quad (11)$$

Although we use the Gaussian model for estimating $p(\boldsymbol{\delta} \mid \Omega)$ and $p(\boldsymbol{x}|C_i)$ in this paper, other various models can also be applied.

4. Experimental Results

In order to confirm the efficiency of the proposed dissimilarity measure and the related classifier, we conducted a number of experiments on two benchmark data: FERET face database (available at http://www.itl.nist.gov/iad/humanid/feret/feret_master.html) and PICS face database (available at http://pics.psych.stir.ac.uk/). The examples of the two databases are shown in Figure 1.

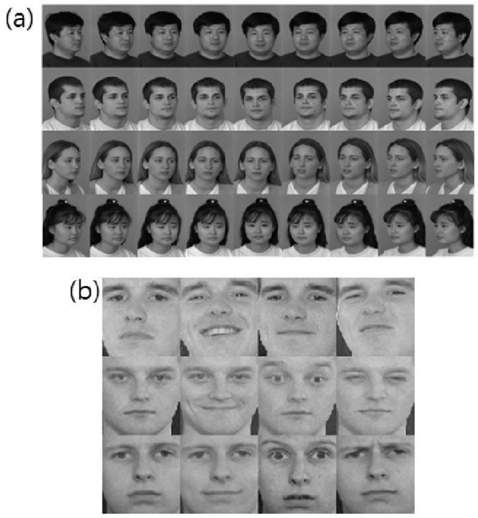

Figure 1. Samples of facial images used in the experiments: (a) FERET database, (b) PICS database

FERET database is composed of 450 facial images with nine different poses from 50 subjects. Using the FERET database, we conducted two types of classification tasks: face recognition and pose recognition. For face recognition, the left (+60^{o}), right (-60^{o}), and frontal images of each subject were used as training (gallery) set, and the remaining 300 images were used for testing (probe data). For pose recognition, images from 25 different subjects were used for training, and the remaining 225 images from the other 25 subjects were used for testing.

The PICS database is composed of 276 images from 69 persons; four images with different expressions were taken from each person. For the PICS database, we also conducted two types of classification tasks: face recognition and expression recognition. For face recognition, we used three images from each person for training, and the remaining one image with a neutral expression was used for testing. For expression recognition, 20 images per each facial expression were used for training and the remaining 49 images were used for testing.

Instead of using raw input images, we apply the principal component analysis (PCA) to original images so as to obtain low dimensional features (Martinez and Kak, 2001). The dimensionality of the feature vectors obtained through PCA was chosen to give the best performance for each distance measure and task. We compared the performance of the proposed measure (d_{prop}) with those of four other conventional measures: Euclidean distance (d_{norm2}), the cosine distance (d_{cos}), Mahalanobis distance (d_{Mah}), and the probabilistic dissimilarity based on within-class variations (d_{with}). Figure 2 shows the classification rates of the measures in the four classification tasks.

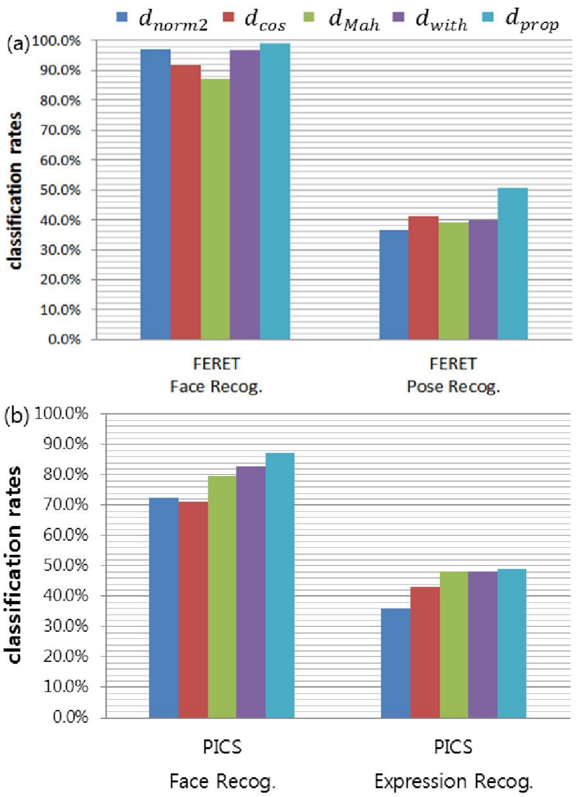

Figure 2. Classification performances of the distance measures in the four tasks: (a) face and pose recognition on FERET database, (b) face and expression recognition on PICS database.

From the figure, we can see that the proposed measure gives the best performance for all the four tasks. We can also see that performances of Euclidean distance (d_{norm2}) and the cosine distance (d_{cos}) are dependent on the database and tasks; d_{norm2} gives relatively good performance for face recognition on FERET data but is worst for pose recognition on the same data; d_{cos} shows relatively better performance for pose and expression recognition. On the other hand, the probabilistic learning approaches (d_{with} and d_{prop}) show stable performances for all the tasks.

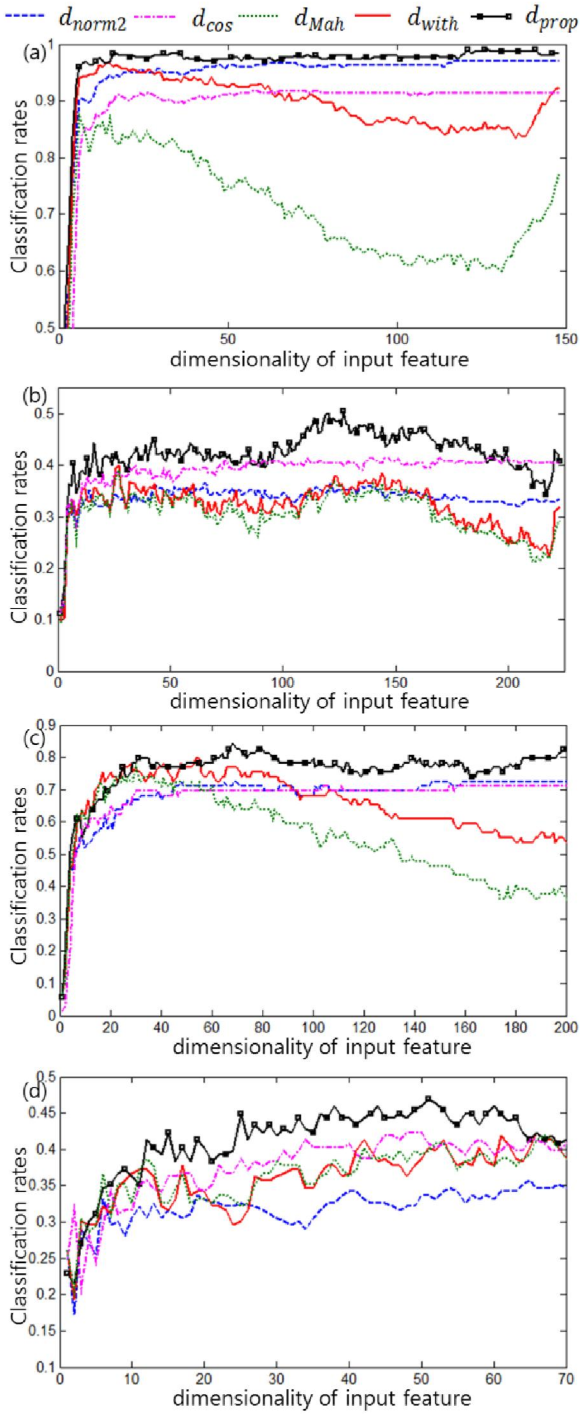

Figure 3. Change of performances depending on the dimensionality of input feature: (a) face recognition on FERET data, (b) pose recognition on FERET data, (c) face recognition on PICS data, and (d) expression recognition on PICS data.

Since we use the features obtained from PCA, we also need to check performance sensitivity to the change of the dimensionality of feature vector. Figure 3 shows the change of classification rates according to the increase of the dimensionality of PCA features. Whereas d_{norm2} and d_{cos} show stable performances in the change of dimensionality, d_{Mah} and d_{with} show performance degradation according to the increase of dimensionality in Figure 3 (a) and (c). Because these two measures have parameters to be estimated through learning with training data, the increase of dimensionality can cause inaccurate estimation, and thus leads to low classification performances. Nevertheless, we can see that the proposed measure can achieve stability in the performances.

5. Conclusions

In order to improve the performance of the distance-based classifier, we proposed a probabilistic learning method for obtaining an appropriate measure of dissimilarity. By combining class conditional probability and the probability of within-class variations, we achieved reliable performances on various pattern classification problems. Compared to the fixed arithmetic distances, the proposed measure can achieve good performance by obtaining appropriate measures through learning. Compared to the Mahalanobis distance and the probabilistic distance based on within-class variation, the proposed method can improve classification rates and stability by adding class-conditional probability. Though we use Gaussian model for estimating the probability densities, it would be interesting to use more sophisticated models. In addition, it is also necessary to investigate the efficiency of the proposed measure in diverse applications with various types of input signals.

Acknowledgements:

This research was supported by Kyungpook National University Research Fund, 2012. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2013061831).

Corresponding Author:
Prof. Hyeyoung Park
School of Computer Science and Engineering
Kyungpook National University.
Sangyuk-dong, Buk-gu, Daegu, 702-701, Korea
E-mail: hypark@knu.ac.kr

References

1. Duda, R.O., Hart, P. E., and Stork, D. G., Pattern Classification (2ed.). Wiley, 2001.
2. Bishop, C.M., Pattern Regognition and Machine Learning, MIT Press, 2006.
3. Ashby, F. G. and Ennis, D. M., Similarity measures, Scholarpedia, 2(12), 4116, 2007.
4. Weinberger, K. Q., Blitzer, J., and Saul, L. K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. Adv. in Neural Inf. Proc. Sys. (NIPS17), 2005.
5. Moghaddam, B., Jebara, T., Pentland, A. Bayesian modeling of facial similarity. NIPS11, 1999.
6. Lee, K., Park, H., A new similarity measure based on intraclass statistics for biometric systems. ETRI J. 25 (5), 401–.406, 2003.
7. K. Lee and H. Park, Probabilistic learning of similarity measures for tensor PCA, Pattern Recognition Letters 33, 1364-1372, 2012.
8. Xing, E., Ng, A., Jordan, M., and Russell, S. Distance metric learning with application to clustering with side-information. Advances in Neural Information Processing Systems. MIT Press, 2002.
9. Martinez, A., Kak, A.: PCA versus LDA. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(2), 228–33, 2001.

3/10/2014