AN ABSTRACT OF THE THESIS OF

Madan Kumar Thangavelu for the degree of Master of Science in Computer Science presented on December 04, 2009.

Title: On Error Bounds for Linear Feature Extraction.

Abstract approved: _____

Raviv Raich

Linear transformation for dimension reduction is a well established problem in the field of machine learning. Due to the numerous observability of parameters and data, processing of the data in its raw form is computationally complex and difficult to visualize. Dimension reduction by means of feature extraction offers a strong preprocessing step to reduce the complexity of the data. In applications dealing with classification of high dimensional data, the goal of a feature extraction step is to achieve a classification accuracy close to that achieved by utilizing the complete high dimensional data. In search for better classification with reduced complexity, numerous dimension reduction methods have been proposed that directly or indirectly aim at minimizing the classification error.

This thesis proposes a novel set of bounds on the probability of classification error for the dimension reduced data. A criteria called the Chernoff union bound is developed which acts as the upper bound on the bayes classification error in the transformed subspace. The bounds offer a closed-form solution to our problem under various data model assumptions. We demonstrate its applicability in feature extraction for parametric and non-parametric data model assumptions. A detailed

numerical study has been presented comparing the performance with many state-of-the-art methods demonstrating the competitiveness and validity of the proposed criteria. .

On Error Bounds for Linear Feature Extraction

by

Madan Kumar Thangavelu

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented December 04, 2009
Commencement June 2010

Master of Science thesis of Madan Kumar Thangavelu presented on December 04, 2009

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Madan Kumar Thangavelu, Author

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

The advancement in data collection and storage techniques in the past decade has led to the acquisition of large amount of high dimensional data. Research areas involving engineering, biology, economics, astronomy, and auditing acquire numerous high dimensional data as part of their work. One of the principle challenges is to identify relevant parts of information from the high dimensional data. Most statistical learning tools, as simple as a classifier are difficult to learn when the data is high dimensional. The idea behind the field of dimension reduction[1] (DR) remains that, even though each data is described in terms of a large number of variables (features), the most relevant information would be concentrated within a very few features. The most relevant features extracted are enough to describe the original data in its entirety. Any utilization of this lower dimensional reduced feature data should offer similar results as the original high dimensional data.

Mathematically, during data collection step we obtain a high dimensional vector $z \in \mathbb{R}^m, z = [z_1, z_2, \ldots, z_m]^T$ and our goal is to obtain a new lower dimensional vector $x \in \mathbb{R}^k, x = [x_1, x_2, \ldots, x_k]^T$ where, $k \ll m$. It is important to note that the performance of the dimension reduction step is task dependent.

---

[1]Dimension reduction has acquired various names over the past decades. Dimensions are analogous to the terms like "feature", "variable", "dimensionality" and "attribute" while the term reduction is replaced by two possible approaches namely "extraction" and "selection".

1.2. Task-oriented dimension reduction

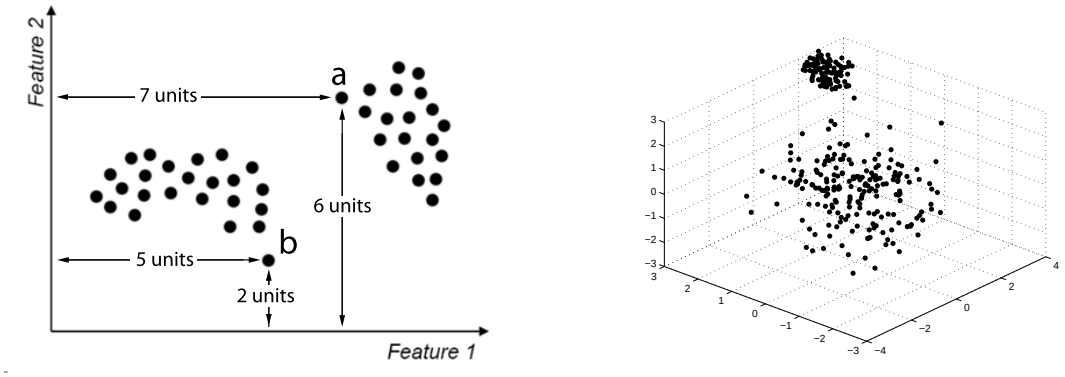
Dimension reduction is adopted for a variety of reasons. One of the biggest task in hand is to extract relevant information from this ever growing abstract high dimensional data. In contrast to simple traditional data collection and statistical analysis, the problem under consideration involves numerous observations where each observation is associated with large number of variables. DR is broadly grouped as a tool for visualization, compression, and noise reduction.

1.2.1 Visualization

In biological and medical research areas the requirement of visualization is of great importance. Data collected by measurements, and experiments can be used to derive important underlying conclusions. A good visualization tool can identify a relation two parameters in a measurement, e.g., A plot between radiation exposure vs the cases of cancer can clearly bring out relevant information. Often, more than one parameter (e.g., radiation, smoking) may correlate with a particular phenomenon (e.g., presence of cancer in a patient). Such situations form an ideal setup for applying DR for visualization. In the scenario just discussed, DR methods offer a plot in 2D after neglecting the numerous other irrelevant measurements (features) from the original data. Visualization is often used as a tool to analyze clustering properties, identify parameter relationships and developing a sense of neighborhood or proximity among data points. To illustrate the concept we consider a toy example in further discussions.

When a multidimensional data is in 2-dimensional or in 3-dimensional space, the visualization of data is possible. Consider two 2-dimensional vectors $a = [7, 6]$

and $b = [5, 2]$. The numeric representation of the vector is independent of the feature associated with the value, i.e, the two values of features in 'a' could represent any physical attribute like [Height, Width], [Length, Angle] or [Concentration 1, Concentration 2]. Such observation can be visualized in a 2D space using a scatter plot as shown in Fig. 1.1(a).



(a) Representation of 2D vectors (b) Representation of 3D vectors

FIGURE 1.1: Visualization of vectors.

In cases where the number of features m for a data is 3 the data is visualized in a 3-dimensional space shown in Fig. 1.1(b). When $m > 3$ visualization of the data in a m dimensional space is not possible. In such cases DR offers a tool for visualization of the data. To illustrate the idea, we consider an image of a hand written digit of dimensions 16 x 16 pixels. One such image is represented as a vector of dimension 256 (16×16). Visualization tools at such high dimensions are unavailable and hence a dimension reduction to a 2-Dimensional or a 3-Dimensional space is desired. DR to 2D provides a 2D vector corresponding to each 256D high dimensional vector which can now be visualized using a scatter plot as mention previously. A 2-dimensional representation of hand written number is shown in Fig. 1.2(a).

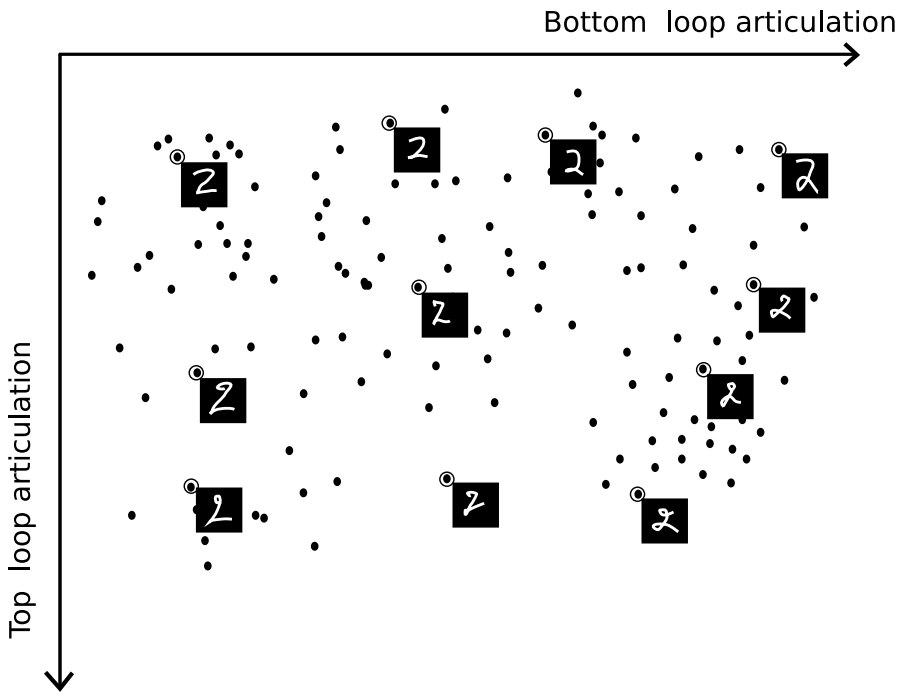


FIGURE 1.2: A 2D representation of hand written images.

1.2.2 Data compression

In applications such as hyper-spectral imaging the image data are captured over a wide range of electromagnetic spectrum. Unlike the standard RGB based image which is represented as a 3D cube of data, a hyper-spectral image constitutes of a m dimensional cube where m is large. Storage and processing of such images are expensive. Since some of the information in these images is redundant, we are only interested in parts of the image that contain interesting underlying information. Dimension reduction helps to identify such features thus reducing the cost of storage and easier processing of the dimension reduced data.

1.2.3 Noise removal

In applications involving sensor networks noise is associated with the raw data which is obtained. Dimension reduction helps in identifying an underlying model in which the data is generated and can be used to reduce noise features considerably. PCA [1] is a well known method that is utilized for noise removal.

1.3. Application areas

Techniques of DR are widely used in a broad set of fields. It is extensively used in the field genetics and microbiology [2]. A well known application in this are is to model the relationship between phenotype and gene expression from a micro-array data in order to classify samples (e.g., classification of tumors). The audio and speech processing [3] industry applies DR in order to remove noise, genre classification and modeling user preferences. High dimensional data are obtained by using the 'Bag of words' model in document analysis and classification [4]. Document clustering, visualization and classification applications rely on dimension reduction techniques. In the recent years a new field of recommendation systems [5] is developing at a fast pace. It deals with product recommendation for users based on ratings obtained by other users in the system. The number of products and users are large and dimension reduction offers a great tool to identify the underlying hidden models within the data. The recent contest organized by Netflix offered a prize money of one million dollars for developing a good movie recommendation system to users. The best performing algorithms in the contest were ones that utilized the DR method of SVD to identify principle features to model the user and movie parame-

ters. Hyper-spectral imaging [6] industry extensively use DR to reduce the load on storage and processing of high dimensional spectral images obtained across a wide electromagnetic spectrum. Applications in this area involve identifying object on the earth surface form images taken from space. Face recognition [7] is a common application in the image processing area. With current imaging techniques, high dimensional (usually in mega-pixels) data is obtained. In an application for face recognition, a good classifier is difficult to learn at high dimensions (mega-pixels) and thus a good generalization and computational advantage can be gained using DR methods. The reason for such widespread application of DR is the ability to represent data in each of these fields in the form of numerical vectors which are standard input for all DR methods. After the DR step the obtained results can be translated back to the original parameters to develop an inference about the data. Hence DR remains a widely applicable method to reduce unwanted complexity in processing high dimensional data.

1.4. Dimension reduction for classification

Consider two distributions of data belonging to class $c1$ and $c2$ as show in Fig. 1.3(a). D denotes the decision boundary between the two distributions. Given new data $x = [x_1, x_2]$, the goal of classification is to assign a class label to x based on some criterion. Classification is an important property that is required in many applications of DR.

Dimension reduction as a preprocessing step in classification refers to the evaluation criterion of DR being the classification error. Consider 2-dimensional data as shown in Fig. 1.4(a) We consider the case of DR to a 1-dimensional line. Fig-

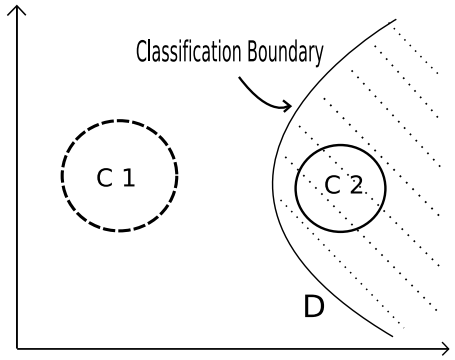
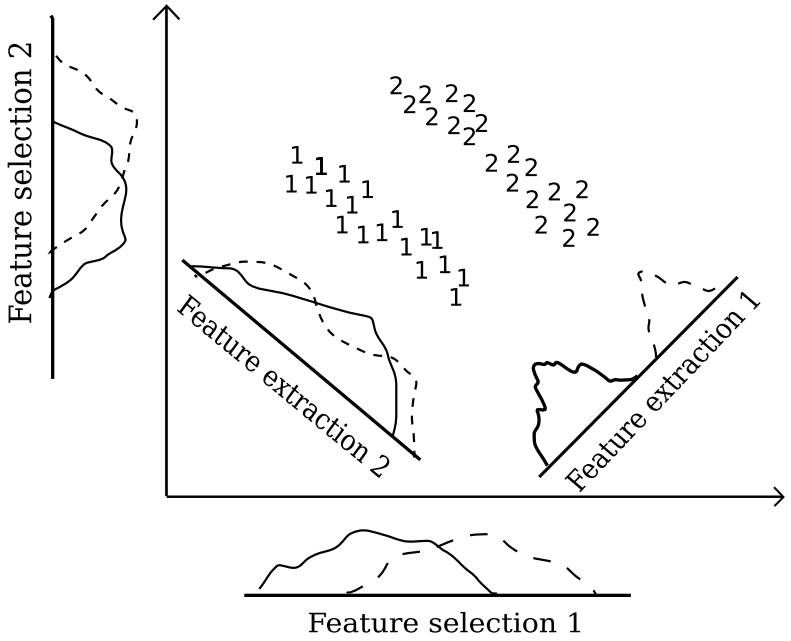


FIGURE 1.3: Classification of data.



(a) Projection plane vs class separation

FIGURE 1.4: A 2D to 1D Dimension reduction.

ure 1.4(a) shows four different projections to a 1-dimension line. In projections along the x-axis, y-axis and along the principle component of the data, the two classes are overlapping. Overlapping classes produce a higher classification error probability and hence are not desirable. The 1-dimension plane that best fits this scenario is illustrated by the vector along 45° tagged as 'Feature extraction 1'. On

this projection the two classes of data are clearly separable and hence produces least classification error. Goal of DR for classification is to obtain one such projection where the classification error is minimum.

1.5. Feature selection vs Feature extraction

Two main approaches to DR are feature selection and feature extraction. In feature selection a subset of the original dimensions are selected. Feature selection is computationally tedious, as the process is the selection of the best combination among the original m dimensions. The total number of choices of such combinations is given by $\binom{m}{1} + \binom{m}{2} + \ldots + \binom{m}{m} = 2^m - 1$ which is computationally intractable for very high dimensions. In Fig. 1.4(a) the x-axis and the y-axis are the two possible 1D feature selections. The drawback of feature selection in this scenario is that either 1D projection will not offer good classification.

Feature extraction offers the flexibility of constructing new lower dimensional feature space form existing high dimensional variables. It is different from feature selection in the sense that the new lower dimensional features are not a subset of the original features but a linear or non-linear combination of the original features. The most standard algorithm for feature extraction is the principal component analysis (PCA). In PCA the data is factorized into basis vectors (eigen vectors) and mixing coefficients (eigen values), the data is then reconstructed using the most relevant basis vectors called the principal components.

Feature extraction is computationally faster as the problem is not combinatorial as in the case of feature selection, also it offers better DR as illustrated in Fig. 1.4(a). In Fig. 1.4(a) Feature selection is restricted to two vectors along the

x-axis and the y-axis, on the other hand we are able to utilize any vector across all angles from 0° - 360° for feature extraction. It can be noted that the hypothesis space of feature selection is a subset of the hypothesis space of feature extraction. Such flexibility allows us to identify the line tagged 'Feature extraction 1' in Fig. 1.4(a) as a dimension of projection on which the classes are clearly separable.

This thesis deals with methods that are classified under the feature extraction methods for DR.

1.6. Data Models

Modeling the data allows us to develop machine learning algorithms that take advantage of the statistical properties of the data. Numerous DR methods have been developed assuming a particular data model. Among the most common data models assumed are Gaussian distribution, Gaussian mixture distribution, and the non parametric distribution based on kernel density estimates. It is best to adopt a data model that best describes the data for better DR or classification results. In the course of this thesis we will develop a generalized framework that will be applicable for all the aforementioned data models. We will now go over a brief review of the data models.

1.6.1 Gaussian distribution

In DR the original data is high dimensional and by virtue of central limit theorem its low-dimensional projection can be assumed Gaussian. Data that follow a Gaussian distribution is characterized by its mean and covariance. In the context of multivariate distributions of dimension m , the data is completely parametrized

by two parameters namely mean (μ) and covariance (Σ). The mean is characterized by a vector of size $1 \times m$ and the covariance is described by a matrix of dimension $m \times m$. Numerous DR methods adopt this data model due to the availability of analytic expressions and mathematical tools to handle Gaussian expressions. The probability density function of a Gaussian distribution is give by

$$p(x) = \frac{1}{\sqrt{|2\pi\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}. \tag{1.1}$$

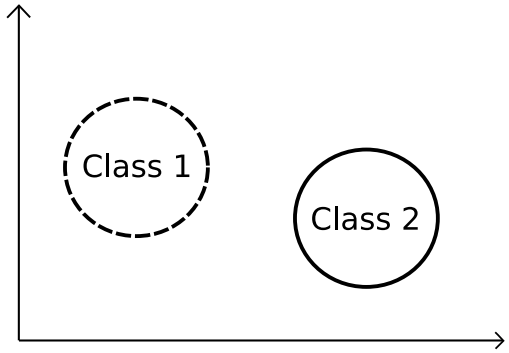


FIGURE 1.5: Gaussian distribution of data.

1.6.2 Gaussian Mixture Model

A Gaussian mixture model (GMM) is a probabilistic model that describes the density function for the data in terms of multiple Gaussian distributions and mixing factors. GMM have the flexibility to represent complicated data while preserving the analytic advantages of Gaussian distribution. A number of cases occur in DR when there is a need for GMM, e.g, in character recognition problem, consider the two versions of the letter 'r' as shown in Fig. 1.6(a) A 2D visualization of this data demonstrates that the class 'r' clusters into two groups. A Gaussian

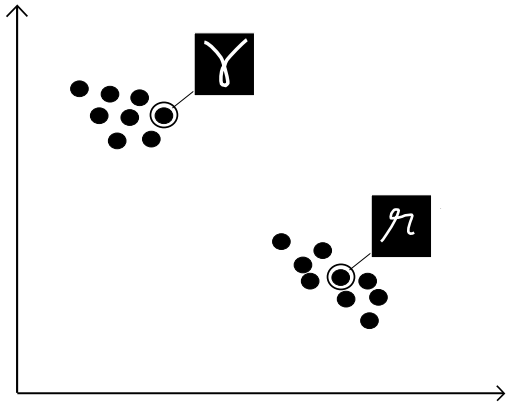


FIGURE 1.6: Gaussian mixture of letter 'r'.

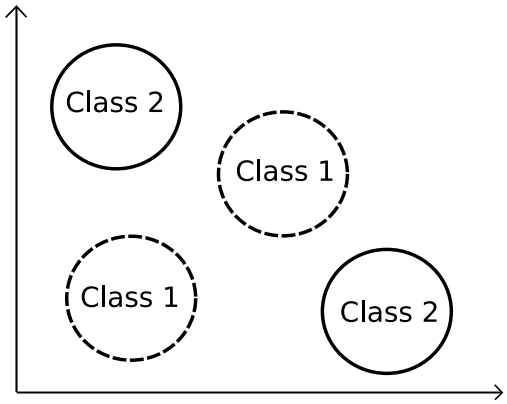


FIGURE 1.7: Gaussian mixture model.

assumption of the data clearly misrepresents the data and in this case a GMM is inevitable. When data cannot be described accurately using a Gaussian distribution, a GMM comes in handy. A Gaussian mixture is completely described by a number of Gaussian distributions ϕ (parametrized by μ and Σ)and their mixing factor α such that $\sum_{k=1}^{k_i} \alpha_{ik} = 1$ and $\alpha \geq 0$. The probability density function of a Gaussian mixture is given as,

$$p(x) = \sum_{k=1}^{k} \alpha_k \phi_k(x). \tag{1.2}$$

1.6.3 Non-parametric model

The above data models assumed parametrization for the data. In situations where the true data does not follow a distribution close to Gaussian or GMM then, it can be advantageous to represent the data accurately using non-parametric model based on kernel density estimates (KDE). The advantage of KDE is that we can define data models of arbitrary shape. KDE is associated with a parameter called kernel which is the distribution assumed around each point of data. Under this parametrization, the probability density function of a random variable x is given by

$$p(x) = \frac{1}{N} \sum_{k=1}^{N} I(y_k = i) K(x - x_k; \sigma) \tag{1.3}$$

where N is the number of instances in the data, indicator I allows summing only points belonging to the same class and for our particular setup we assume $K(x - x_k; \sigma)$ to be the Gaussian kernel given by,

$$K(x - x_k; \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \ \exp^{-\frac{\|x - x_k\|^2}{2\sigma^2}} \tag{1.4}$$

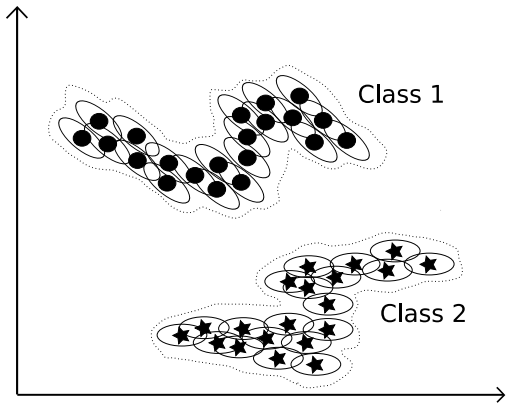


FIGURE 1.8: Non-Parametric Model

1.7. DR methods and data models

Numerous dimension reduction methods have been developed in literature with various driving criterion aimed at obtaining the best noise reduction, clustering, classification, geometry properties in the lower dimension. Also, these methods assume one of the three data parametrization discussed above. In subsequent chapters we will layout the review of methods grouping them based on criterions and the data model that they are applicable to. The advantages and disadvantages of a criterion and data model over each other will be discussed over the course of this thesis.

1.8. Organization of the thesis

This thesis discusses the design of DR algorithms that preserve the classification accuracy of the data.

Chapter 2 provides a thorough literature review of the area of DR. Our discussion will go through the classical methods of PCA and LDA and incline towards methods that follow various criterions for DR. We will finally categorize DR techniques based on the data model that they are applicable and the criterion that they utilize.

Chapter 3 details the mathematical formulation of DR. In our specific case, the choice of the evaluation criterion is the classification accuracy/ classification error and hence we will describe how the problem of DR can be related to in terms of the probability of error.

Chapter 4 develops a bound on the probability of error for multiclass classifica-

tion of the data. The two class bound on the probability of error is well established via the Chernoff bound. In this chapter we demonstrate how the multiclass classification error is bounded in terms of pairwise two class error bounds.

Chapter 5 investigates the applicability of the error bound developed in Chapter 4 to different data models. In our case, we analyze how the bound is applicable when the data is modeled as Gaussian distributions, Gaussian mixtures. We finally develop the bound in the non-parametric cases when no data model is assumed and the class distribution is developed using kernel density estimates (KDE).

Chapter 6 provides a method to optimize the cost function developed in Chapter 5. Our particular case will be based on a gradient descent algorithm. The details of the technique and the methods adopted to speed up are discussed.

Chapter 7 offers the numerical study based on the three DR methods that were developed based on the error bounds on probability and a detailed comparison to various other state-of-the-art DR methods along with the classical methods of PCA and LDA.

Chapter 8 summarizes the offerings of this thesis and the prospectus for future work in this area.

2. LITERATURE REVIEW

2.1. A review of DR techniques

With growing data acquisition and storage rates, dimension reduction (DR) methods take an important place in data processing and analysis. One of the most discussed application of DR is classification. As a preprocessing step, DR allows for a low dimensional data classification to alleviate the curse of dimensionality. Additionally, the use of DR enables visualization as well as a computational advantage. With a reduction in the number of features that represent the data, some discriminatory information may be lost in favor of computational advantage and generalization.

2.2. Broad categorization

We consider categorizing DR into two broad approaches. In *feature selection*, an informative subset of the available features is of interest. For a survey on methods for feature selection we refer the read to [8]. In *feature extraction*, a lower dimensional embedding of the data is of interest. In other words, a new set of lower dimensional features are constructed from the original feature vectors. For a survey on feature extraction methods we refer the reader to [9].

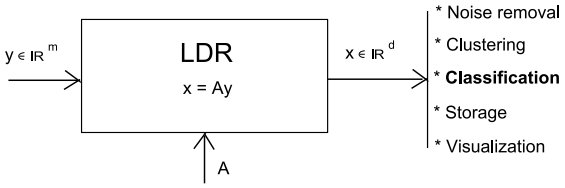
2.3. Linear and non-linear DR

Feature extraction has been addressed using both linear and non-linear methods. Until recently, nonlinear methods primarily concentrate on preserving the high-dimensional geometrical structure of feature vectors in the lower dimensional embedding [10, 11, 12]. Manifold learning deals with non-linear DR for exploring the intrinsic dimension of the data and its geometrical and topological structure. Some efforts towards nonlinear dimension reduction for classification have been made (see e.g., [13, 14, 15]). Generally, the computation of a nonlinear transformation from a high-dimensional data space to a low-dimensional space requires the computation of a transformation parametrized by a number of parameters proportional to the data size and hence are considered computationally complex when the data size is large. Moreover, the applications of the methods to test data often requires the entire training set imposing a memory constraint.

2.4. Supervised vs unsupervised LDR methods

Linear dimension reduction (LDR) offers computational efficiency relative to their nonlinear counterparts in return for some performance compromise. LDR methods can be separated into supervised and unsupervised methods. In unsupervised LDR methods, the label of the data is either unavailable (or simply ignored) and hence techniques focus on the structure of the data. Principle component analysis (PCA) [1] implicitly assumes that the data is generated by Gaussian distribution. Using an eigen decomposition of data sample covariance matrix, data is projected onto the eigenvectors corresponding to the largest eigenvalues. Due to

the simplicity and low computational complexity, PCA is the most popular method for linear dimension reduction. A kernel based nonlinear generalization of PCA was proposed in [16]. Independent component analysis (ICA) [17] considers LDR to produce independent features removing the Gaussian assumption used by PCA. Supervised linear feature extraction is one of the most popular areas of dimension reduction. Supervised methods take into consideration the class label during the dimension reduction process and hence offer more control in applications that deal with classification and pattern recognition.



(a) Block diagram of LDR

FIGURE 2.1: Dimension reduction as a preprocessing step for various other applications like noise removal, clustering, classification, storage and visualization.

2.5. DR for classification

In recent years, DR methods have been developed with the further utilization of the data in mind as shown in Fig. 2.1. For noise reduction, one may consider PCA whereas for data visualization, nonlinear manifold learning may be advantageous. In information retrieval [18], DR is used to preserve the information in the data. Similarly, classification can benefit from a classification-optimized DR.

2.6. Information criterion for DR

A natural criterion for performing optimal DR for classification is the classification error probability. Linear discriminant analysis (LDA) [19] is one of the most well-known supervised linear feature extraction. It is shown to be optimal in the two-class case when the class-conditional distributions are homoscedastic Gaussian probability density functions. LDA has many variants and a popular version uses a generalized eigendecomposition to maximize the between class covariance matrix and within class covariance matrix ratio. Ultimately the minimization of the probability of classification error should guide DR for classification. However, since an analytic expression for the probability of error cannot be computed in closed-form for most scenarios, alternatives have been considered. In [20, 21, 22], the principle of minimum classification error (MCE) is applied using a differentiable approximation to the sample error probability trading-off fidelity to the original criterion for applicability. Another alternative to minimum error probability involves the maximization of probabilistic distance measures and probabilistic dependence measures [23] between different classes. A popular approach is based on the maximization of mutual information between the features and the class labels [24, 25, 26]. The motivation to these approaches is based on the connection between classification error probability and mutual information [27]. A number of variation on the theme and approximation to mutual information have been proposed. In [28], a quadratic divergence measure is used instead of the Shannon's original mutual information criterion and is called the quadratic mutual information (QMI) criterion. The formulation of QMI allows it to be applied to non-parametric dimension reduction. In [29], approximation of the entropy of a Gaussian mixture is used to compute the

mutual information for the Gaussian case. Other probabilistic distance measures that are used for DR are Kullback-Liebler divergence, Bhattacharyya distance [30] and Chernoff distance [31]. These divergences offer a tight surrogate for binary classification. However, heuristic extensions to multiclass often result in sub-optimal performance. The methods in [32, 33, 34] offer such extensions for multiclass linear dimension reduction. Another manifestation of the principle of minimum probability of classification error is presented in [35]. Neighborhood component analysis (NCA) considers minimization of the probability of error for a specific classifier namely k-nearest neighbor (KNN).

2.7. Data model based categorization

LDR methods can also be categorized based on the statistical data model considered. PCA [1], [32], [36] assume Gaussian class-conditional distributions and allow heteroscedasticity of the class covariances. LDA [19] on the other hand is optimal under the assumption of Gaussian class-conditional distributions with homoscedastic class covariances, though LDA is applied equally often to heteroscedastic case. It should be noted that all methods that allow heteroscedasticity are applicable for homoscedastic data as a special case. Methods like MMI [24] have been demonstrated to work well with Gaussian mixture models (GMM). Methods that assume no model for the data are categorized as non-parametric and includes, NCA [35], IDA [29], QMI [37] . Table 2.7. lists the above mentioned methods in the form of a table listing various methods based on the model parameter assumed and the approach of dimension reduction.

TABLE 2.1: A list of DR methods placed based on the underlying principle and the data model applicable.

	Distance	Mutual Information	Approximation of error	Bound on error
Gaussian-Homoscedastic	[19]	—-	—-	—-
Gaussian-Heteroscedastic	[36, 32]	[29]	[20]	[38]
Gaussian mixture models	—-	[24]	—-	[39]
Kernel density estimates	[35]	[37]	—-	CUB-KDE

2.8. Issues with existing methods

Methods based on approximation of probability of error suffer form the lack of control over the approximation accuracy. On the other hand, methods based on probabilistic distance measures adopt a heuristic approach for multiclass classification which can lead to significant performance degradation. Empirical evaluation of mutual information leads to undesirably high computational complexity. Most methods seem to have an advantage when the data follows the data model assumed and suffer a degradation in performance as the data deviates from the assumed model.

A probability of error bound approach for DR has been proposed in [40] [41]. The bound explored in [40] is based on Fisher distance and suffers from optimization due to the non-smooth nature of the maximization of a *min* criterion [38]. The closest method to our approach is suggested in [41] where the error probability of the multiclass classification is bounded by a Union Bhattacharya bound. This bound [41] is applicable only to Gaussian data model and generalization to other data models are unavailable.

2.9. Our approach

We present an LDR method that is based on the minimization of error probability. We introduce a bound on the probability of error for multiclass classification. A closed-form expression for the bound is provided for the case where the classes are: i) Gaussian, ii) Gaussian mixture model and iii) follow a Gaussian kernel density estimate. The novel bound introduced addresses three issues: i) multiclass classification ii) smooth surrogate for probability of error and iii) variability in distribution. We present an LDR method based on the bound and introduce a gradient descent implementation of it. We demonstrate the superiority of the proposed LDR method compared to other state-of-the-art LDR methods via numerical analysis.

3. PROBLEM FORMULATION

3.1. Linear transformation

We start by introducing the problem of LDR for multiclass classification. For the convenience of the reader, we introduce the notations used in this thesis in table 3.1. An observation $z \in \mathbb{R}^m$ is drawn from one of n classes, the set of classes $\mathcal{C} = \{1, 2, \ldots, n\}$. Each class in $\mathcal{C}$ has a prior probability $\pi_1, \pi_2, \cdots, \pi_n$, respectively. We denote the class conditional distributions as $p_1(z), p_2(z), \cdots, p_n(z)$ and restrict the $p_i(x)$'s to probability density functions (PDFs) such that $\int p_i(z)dz = 1$ and $p_i(z) \geq 0$. A linear dimension reduction (LDR) of an m dimensional observation $z \in \mathbb{R}^m$ is defined by a function $g(\cdot) : z \mapsto x \in \mathbb{R}^d$ where $d \leq m$ (and sometimes $d \ll m$). For LDR, the transformation is given by

$$x = g(z) = Az, \tag{3.1}$$

where A is the $d \times m$ linear transformation matrix. We denote $x \in \mathbb{R}^d$ as the low dimensional projection of the observation z. A linear transformation as described above also parametrizes the PDF of the low dimensional projected data x. Corresponding to the PDFs of the data in the original space $p_1(z), p_2(z), \cdots, p_n(z)$, we denote the PDFs for the low dimensional projected data by $p_1(x; A), p_2(x; A), \cdots, p_n(x; A)$ for each class respectively.

TABLE 3.1: A list of symbols and their description.

Symbol	Description
n	Number of classes
m	Original dimension of data
d	Reduced dimension of data
z	Original data, $z \in \mathbb{R}^m$
x	Dimension reduced data, $x \in \mathbb{R}^d$
A	Transformation matrix $d \times m$
$h(\cdot)$	Classifier
π_i	Prior probability of class i
$P_e(\cdot)$	Probability of error
Σ_i	Covariance of class i
$\delta\mu_{ij}$	Difference in mean between class i and j

3.2. Linear transformation for multiclass classification

To analyze the classification error rate associated with a linear transformation, we start by defining a classifier $h(\cdot) : x \in \mathbb{R}^d \mapsto h(x) \in \mathcal{C}$. A classifier is a function that maps an observation x onto one of the n classes in $\mathcal{C} = \{1, 2, \ldots, n\}$. Since the classifier in this setup is applied to the lower dimensional projections of the data, the associate performance of the classifier is dependent upon the specific projects through A. In this setup, we can define the probability of error in classification as,

$$\begin{aligned} P_e(A) &= \sum_{i=1}^{n} \pi_i P(h(x) \neq i|i) \\ &= \sum_{i=1}^{n} \pi_i \int I(h(x) \neq i) p_i(x; A) dx, \end{aligned} \tag{3.2}$$

where the indicator function $I(\cdot)$ becomes 1 when its argument is true and 0 otherwise. The optimal classifier $h^*(x)$ that minimizes the probability of error in (3.2) is

known as the Bayes classifier [42] and is given by

$$h^*(x) = \arg\max_i \pi_i p_i(x; A). \tag{3.3}$$

The error probability $P_e^*(A)$ associated with Bayes classifier $h^*(x)$ can be obtained by substituting (3.3) into (3.2),

$$P_e^*(A) \;=\; \sum_{i=1}^{n} \int I\left(\max_{j\neq i} \frac{\pi_j p_j(x; A)}{\pi_i p_i(x; A)} > 1\right) \pi_i p_i(x; A) dx. \tag{3.4}$$

Replacing the error event $\{\max_{i\neq j} \frac{\pi_j p_j(x;A)}{\pi_i p_i(x;A)} > 1\}$ in (3.4) by $\bigcup_{i\neq j} B_{ij}$ where, $B_{ij} = \{\frac{\pi_j p_j(x;A)}{\pi_i p_i(x;A)} > 1\}$, we rewrite the error probability for the optimal Bayes classifier as

$$P_e^*(A) = \sum_{i=1}^{n} \int I\Big(\bigcup_{i\neq j} \big(\frac{\pi_j p_j(x; A)}{\pi_i p_i(x; A)} > 1\big)\Big) \pi_i p_i(x; A) dx. \tag{3.5}$$

From (3.5) it is evident that the probability of error $P_e^*(A)$ depends on the linear transformation matrix A through the PDF of the reduced dimensional data $p_i(x; A)$ and therefore by controlling the value of A we can control the value $P_e^*(A)$. Our objective is the minimization of probability of error $P_e^*(A)$ w.r.t the transformation matrix.The optimal transformation matrix A^* is formally defined as:

$$A^* = \arg\min_A P_e^*(A). \tag{3.6}$$

The probability of error for classification which is based on observations from the reduced dimension space is given by (3.5) in terms of the PDF of the data. However, the true class-conditional probability distributions of the data or the class priors are often unavailable. Instead, only samples are available as $\{(z_k, y_k)\}$, where y_k

is the class label associated with the k^{th} sample and z_k is the high-dimensional feature vector associated with the same sample. In this framework, we pursue both parametric and non-parametric approaches to represent $p_i(x)$ so that we can apply its form to evaluate (3.5).

3.2.1 Difficulty in evaluating error probability

Although some ambiguities are resolved by specifying a model for $p_i(x)$, some difficulties remain. The indicator in (3.5) yields a set in terms of x over which the integral is carried out. In general, the integral is not available in closed-form for even simple scenarios such as the case with heteroscedastic Gaussian $p_i(x)$s (i.e., with different covariance structures). Moreover, the union within the indicator in (3.5) yields a set defined in parts further reducing the possibilities of a closed-form expression for the RHS of (3.5). To address these issues we consider an upper bound approach on the probability of error. By minimizing a closed-form upper bound on the probability of error, we aim to produce a method for LDR that is consistent with the underlying goal of classification.

3.2.2 Previous solutions

Due to the issues associated with the evaluation of (3.5) and its optimization w.r.t. A, various alternatives have been adopted through direct and indirect surrogates for the probability of error. We proceed by describing the various approaches adopted in literature that follow the approach of optimization of surrogates for the error probability. In minimum classification error (MCE) methods, a cost function that is an approximation of the probability of error is developed. In [20], a differentiable cost function is proposed to replacing the indicator with a smooth approximation using an exponential or a sigmoid function. This approach has been

discussed to be inefficient in cases where the data size and dimension are high [43]. Another widely adopted statistical measure used in feature extraction is based on the maximization of the mutual information (MI) between the class labels and the features. In [27], Hellman and Raviv provide a bound on the probability of classification error in terms of the MI, suggesting that the minimization of classification error probability is coupled with the maximization of the MI. In developing an expression for the MI for DR, a wide range of data models have been deployed, e.g., Gaussian models [24], Gaussian mixture models [44], and non-parametric kernel density estimates [37]. However, for either data model assumption a closed-form expression is unavailable and all MI methods referenced adopt a sample estimate approximation. The effects of the sample-based approximation on the accuracy of the methods remains unclear. Another approach in LDR, considers divergence measures between class-conditional PDFs. The divergence measures adopted in the literature for LDR include the Kullback-Leibler divergence [45], Chernoff distance [36] (which is directly related to Renyi-α divergence), and the Bhattacharya distances (which is directly related to Hellinger distance). The hope with these approaches is that by maximizing the distance between the PDFs of two classes the associated probability of error will be minimized. While this connection has strong statistical foundations (see [33]), the extension to the multiclass case is non-trivial. The main drawback in the current implementation divergence based LDR is the way in which the pairwise distances are combined to develop the cost function for the multiclass case.

For convenience we will represent the probability density function in the lower dimensional plane denoted by $p_i(x;A), p_2(x;A), \cdots, p_n(x;A)$ as $p_i(x), p_2(x), \cdots, p_n(x)$ in the remainder of the thesis. It can be noted that the parameter $p_i(x)$ is dependent of the linear transformation matrix A, while the parametrization on the true

dimension of the data denoted by $p_i(z)$ is independent of the transformation matrix.

4. ERROR PROBABILITY BOUND

Due to the limitations of the aforementioned methods in obtaining a closed-form surrogate for error probability (3.5) that address both fidelity to the error probability case and the issues associated with the multiclass case, we pursue the approach of bounding the probability of error. In the following, we develop a series of bounds on the error probability $P^*(A)$ (3.5).

4.1. Union bound for multiclass classification

We start with a bound that offers a surrogate to the multiclass probability of error in terms of the pairwise error probabilities [46]. In a multiclass classification problem, the union of error events in (3.5) is difficult to evaluate. Applying the union bound $I(\cup_j A_j) \leq \sum_j I(A_j)$ to (3.5) with $A_j = \{\frac{\pi_j p_j(x)}{\pi_i p_i(x)} > 1\}$, we bound the probability of error associated with Bayes classifier (3.3) by

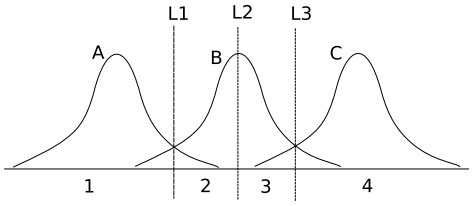
$$P_e^* \leq \sum_{i=1}^{n} \sum_{j \neq i} \int I\left(\frac{\pi_j p_j(x)}{\pi_i p_i(x)} > 1\right) \pi_i p_i(x) dx. \tag{4.1}$$

Expressing the multiclass error in terms of pairwise classification error, (see appendix 2)

$$P_e(p_1, p_2, \cdots, p_L, \pi_1, \pi_2, \cdots, \pi_L) \leq \sum_{i=1}^{n} \sum_{j>i} (\pi_i + \pi_j) P_e(p_i, p_j, \tilde{\pi}_i, \tilde{\pi}_j). \tag{4.2}$$

where, $\tilde{\pi}_i = \frac{\pi_i}{\pi_i+\pi_j}$, $\tilde{\pi}_j = \frac{\pi_j}{\pi_i+\pi_j}$, and $P_e(p_i, p_j, \tilde{\pi}_i, \tilde{\pi}_j)$ can be obtain by evaluating (3.5) for the two class case as

$$P_e(p_i, p_j, \tilde{\pi}_i, \tilde{\pi}_j) = \int I\Big(\frac{\tilde{\pi}_j p_j(x)}{\tilde{\pi}_i p_i(x)} > 1\Big)\tilde{\pi}_i p_i(x)dx + \int I\Big(\frac{\tilde{\pi}_i p_i(x)}{\tilde{\pi}_j p_j(x)} > 1\Big)\tilde{\pi}_j p_j(x)dx. \quad (4.3)$$



(a)

FIGURE 4.1: Probability distributions of three equiprobable classes A,B,C when projected on a one-dimensional plane. L1,L2,L3 denote the classifier boundaries when only a pair of the classes are considered for classification at any instant. L1,L2 and L3 mark the classifier boundaries between classes , A-B, A-C and B-C respectively.

The implication of the union bound is that error event could be counted multiple times on the RHS of (4.2), resulting in a gap between the predicted error probability denoted by the bound (4.2) and the true error probability for the problem (3.4). While in some cases the gap could constitute a significant portion of the original error, in other the gap may be negligible resulting in a tight bound. To illustrate this point, we consider the following example. Our goal is bring out the difference between the true error probability in this setup and the bound on the error probability developed in equation (4.2). Consider a three class classification problem in one-dimension with PDFs as illustrated in Fig. (4.1). Consider the case in which the classes are equiprobable. The classes are marked as A, B, and C. The overlap in the picture between the PDFs of the different classes illustrates the source of misclassification error. We mark the misclassification regions by circled numbers (e.g., ① when x was produced from $p_B(x)$ by $p_A(x) > p_B(x)$). Thus error event

$p(h(x) = A|B)$ can be written as $p(①|B)$. Furthermore, the line L1 denotes the decision boundary between class A and B, line L2 denotes the decision boundary between class A and C, and line L3 denotes the boundary between class B and C. Under this setup, the inequality in (4.2) is given by

$$\begin{aligned} P_e(A,B,C;P(A),P(B),P(C)) \leq (P(A)+P(B))P_e(A,B;\tfrac{P(A)}{P(A)+P(B)},\tfrac{P(B)}{P(A)+P(B)}) \\ +(P(A)+P(C))P_e(A,C;\tfrac{P(A)}{P(A)+P(C)},\tfrac{P(C)}{P(A)+P(C)}) \\ +(P(B)+P(C))P_e(B,C;\tfrac{P(B)}{P(B)+P(C)},\tfrac{P(C)}{P(B)+P(C)}) \quad (4.4) \end{aligned}$$

Since the class are equiprobable, i.e., $P(A) = P(B) = P(C) = \frac{1}{3}$, (4.4) can be written as

$$P_e(A,B,C;\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}) \leq \tfrac{2}{3}P_e(A,B;\tfrac{1}{2},\tfrac{1}{2}) + \tfrac{2}{3}P_e(A,C;\tfrac{1}{2},\tfrac{1}{2}) + \tfrac{2}{3}P_e(B,C;\tfrac{1}{2},\tfrac{1}{2}). \quad (4.5)$$

To assess the quality of the bound, we proceed by computing both sides starting with the LHS of (4.5)

$$P_e(A,B,C;\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}) = p(②\cup③\cup④|A)\tfrac{1}{3} + p(①\cup④|B)\tfrac{1}{3} + p(①\cup②\cup③|C)\tfrac{1}{3} \quad (4.6)$$

Using the fact that the ⓘs are disjoint, we obtain

$$\begin{aligned} P_e^* = \frac{1}{3}\Big(&p(②|A)+p(③|A)+p(④|A)+p(①|B) \\ &+p(④|B)+p(①|C)+p(②|C)+p(③|C)\Big). \end{aligned} \quad (4.7)$$

Similarly, we proceed by computing the three terms in the bound on the RHS of

(4.5)

$$P_e(A,B;\frac{1}{2},\frac{1}{2}) = \frac{1}{2}\Big(p(\textcircled{1}|B) + p(\textcircled{2}\cup\textcircled{3}\cup\textcircled{4}|A)\Big)$$
$$P_e(A,C;\frac{1}{2},\frac{1}{2}) = \frac{1}{2}\Big(p(\textcircled{3}\cup\textcircled{4}|A) + p(\textcircled{1}\cup\textcircled{2}|C)\Big)$$
$$P_e(B,C;\frac{1}{2},\frac{1}{2}) = \frac{1}{2}\Big(p(\textcircled{4}|B) + p(\textcircled{1}\cup\textcircled{2}\cup\textcircled{3}|C)\Big)$$

Hence the RHS of (4.5) is given by

$$\frac{1}{3}\Big(p(\textcircled{1}|B) + p(\textcircled{2}\cup\textcircled{3}\cup\textcircled{4}|A) + p(\textcircled{3}\cup\textcircled{4}|A) \tag{4.8}$$
$$+p(\textcircled{1}\cup\textcircled{2}|c) + p(\textcircled{4}|B) + p(\textcircled{1}\cup\textcircled{2}\cup\textcircled{3}|C)\Big). \tag{4.9}$$

Using the fact that the $\textcircled{i}$s are disjoint and collecting similar terms, we obtain

$$\frac{1}{3}\Big(p(\textcircled{2}|A) + 2p(\textcircled{3}|A) + 2p(\textcircled{4}|A) + p(\textcircled{1}|B)$$
$$+p(\textcircled{4}|B) + 2p(\textcircled{1}|C) + 2p(\textcircled{2}|C) + p(\textcircled{3}|C)\Big). \tag{4.10}$$

A comparison of the true error probability as given in (4.7) and its bound in (4.10) reveals that the gap between the two terms is $\frac{1}{3}(p(\textcircled{3}|A) + p(\textcircled{4}|A) + p(\textcircled{1}|C) + p(\textcircled{2}|C))$ or $\frac{1}{3}(p(\textcircled{3}\cup\textcircled{4}|A) + p(\textcircled{1}\cup\textcircled{2}|C))$. This gap is a result of counting some of the error events twice. However, in our example the event $p(\textcircled{3}\cup\textcircled{4}|A)$ is negligible compared to $p(\textcircled{2}|A)$ and similarly $p(\textcircled{1}\cup\textcircled{2}|C)$ is negligible compared to $p(\textcircled{3}|C)$. This can be observed by examining that the tail distribution for each class as it decreases further from the class center. In Fig. 4.1 we see that the area under the distribution of class C is negligible in the region $\textcircled{2}$ and $\textcircled{3}$. By using the pairwise evaluation of the multiclass error probability, we effectively eliminated the need to

evaluate union over the class pairs which is difficult to evaluate and replaced it with a sum over errors between the class pairs resulting in a bound on the error probability that remains close to the true error probability. This approach allows us to shift our attention to bounding the pairwise error probabilities.

4.2. Chernoff upper bound on error probability

Although the bound in (4.2) reduces the search for closed-form surrogates for the probability of error to a search for surrogates for the pairwise error probability, it is still not computable in closed-form. A closed-form evaluation of the probability of error between two classes denoted by $P_e^*(p_i, p_j, \tilde{\pi}_i, \tilde{\pi}_j)$ is not trivial as it involves the integration over a non-differentiable indicator function as in (4.3). To bound $P_e^*(p_i, p_j, \tilde{\pi}_i, \tilde{\pi}_j)$, we consider the application of Chernoff bound. Specifically, We bound the indicator function in (4.3) using $I(a \geq 1) \leq a^s$ where $a \geq 0$ and $0 \leq s \leq 1$. The resulting bound is given by

$$\begin{aligned} P_e^*(p_i, p_j, \tilde{\pi}_i, \tilde{\pi}_j) &\leq \int \left(\frac{\tilde{\pi}_j p_j(x)}{\tilde{\pi}_i p_i(x)}\right)^s \tilde{\pi}_i p_i(x) dx + \int \left(\frac{\tilde{\pi}_i p_i(x)}{\tilde{\pi}_j p_j(x)}\right)^{s'} \tilde{\pi}_j p_j(x) dx \\ &= \tilde{\pi}_j^s \tilde{\pi}_i^{1-s} \int p_j^s(x) p_i^{1-s} dx + \tilde{\pi}_i^{s'} \tilde{\pi}_j^{1-s'} \int p_i^{s'}(x) p_j^{1-s'} dx. \quad (4.11) \end{aligned}$$

Note that the second term on the RHS on (4.11) can be made identical to the first term by setting $s' = 1 - s$. Moreover, since the bound holds for any $0 < s, s' < 1$, it is common to consider a tighter form of the bound by minimizing over the RHS,

yielding

$$P_e^*(p_i, p_j, \tilde{\pi}_i, \tilde{\pi}_j) \leq \min_s 2\tilde{\pi}_j^s \tilde{\pi}_i^{1-s} \int p_j^s(x) p_i^{1-s} dx. \tag{4.12}$$

The Chernoff distance between two PDFs $p_i(x)$ and $p_j(x)$ is given by [42]

$$d_{ch}(p_i(x), p_j(x); s) = -\log\Big(\int p_j^s(x) p_i^{1-s}(x) dx\Big). \tag{4.13}$$

Therefore, the two-class bound can be written in terms of the Chernoff distance as

$$P_e(p_i, p_j, \tilde{\pi}_i, \tilde{\pi}_j) \leq \min_s 2\tilde{\pi}_j^s \tilde{\pi}_i^{1-s} e^{-d_{ch}(p_i(x), p_j(x); s)}. \tag{4.14}$$

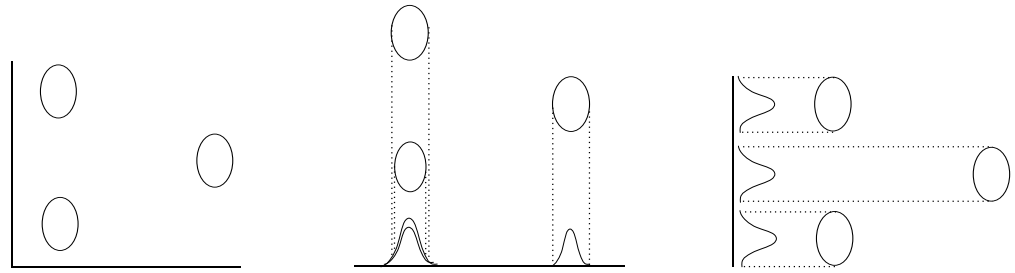
Note that the RHS is exponentially decaying in the Chernoff distance between the two PDFs. Substituting the two-class bound (4.14) into (4.2), we obtain a multi-class extension of the Chernoff bound

$$P_e^* \leq \sum_{i=1}^{n} \sum_{j \neq i} \min_s \pi_j^s \pi_i^{1-s} e^{-d_{ch}(p_i(x), p_j(x); s)}. \tag{4.15}$$

Since the upper bound to the Bayes error probability for a multiclass classification problem in (4.15) was obtained by application of two bounds (i.e., union and Chernoff). We refer to this bound hence on as the Chernoff union bound (CUB). A special case of CUB when $s = \frac{1}{2}$ yields the union Bhattacharya bound which has been demonstrated for DR in [41].

4.3. Sum of distances vs bound on error probability

The bound in (4.15) referred to as CUB considers a weighted sum of exponents of the negative of the pairwise Chernoff distance between classes. This result is significantly different from that of the sum of pairwise distances approach [36] or other possible weighted distance combinations [47, 48]. Since we consider the application of the bound for DR, we proceed with a toy example that provides an intuitive explanation to the advantage of the approach suggested by this form over the sum of distances approach. To demonstrate this, we will consider a three-class toy dataset with classes $\mathcal{C} = 1, 2, 3$ in two-dimensional, i.e., $p_i(x) : \mathbb{R}^2 \rightarrow \mathbb{R}$. Assume that the $p_i(x)$s are of a Gaussian distribution. The specific configuration of the three class is illustrated in Fig. 4.2(a).



(a) Original data in 2D space (b) Projection on 1D x-axis (c) Projection on 1D y-axis

FIGURE 4.2: A 2D to 1D projection that suggests that a projection with a maximum pairwise class distance may not result in the right projection plane for classification purposes.

A goal of an LDR method is to find a linear projection into a one-dimensional space to produce minimum classification error. The algorithm described in [36] extends the maximization of the Chernoff distance approach to the multiclass case by directly summing up the between class Chernoff distances. This criterion is dominated by the between class distances which are the largest. Fig. 4.2(b) illustrates

that by maximizing such criterion, large between class distances are maximized at the expense of smaller between class distances resulting in an overlap of the PDF of two classes far from the PDF of a third class. By overlapping the two class, they become essentially indiscriminable leading to a large classification error. Counter to the sum of distances approach, the proposed bound (4.15) is dominated by the small between class distance driving close classes further apart. In turn, the contribution of the leading sources of error (i.e., the close classes) is reduced leading to a small error probability. Figure 4.2(c) illustrates that while no two class are as apart as in Fig. 4.2(b) the minimum distance between close classes is kept fairly large. In the Gaussian homoscedastic case the Chernoff distance is propositional to Fisher's discriminant and hence our approach is immediately relevant to combining Fisher's discriminant for the multiclass case. While the importance of properly combining pairwise between class discriminants was previously raised in [47], their offered heuristic solution suggested was that of a weighted sum of discriminants approach. Our DR approach following the bound (4.15) is strongly motivated by the probability of error.

5. EXPLICIT ERROR BOUNDS FOR DIFFERENT DATA MODELS

In this section, we provide explicit error bounds for three widely used data models. We consider the case where the class-conditional PDFs follow i) the Gaussian distribution (for both the homoscedastic and the heteroscedastic case, ii) a Gaussian mixture model, and iii) a non parametric kernel density estimate (KDE).

5.1. Chernoff union bound for Gaussian class-conditional distributions

One of the most widely used distributions in data modeling is the Gaussian distribution. The Gaussian model trades-off representation accuracy for tractability. In the case of LDR, the Gaussian model can be motivate by the central limit theorem since the lower-dimensional representation is obtained by linear combination of a high-dimensional feature vector. We consider the case where the class conditional PDF is a Gaussian. Hence, $p_i(x) = \mathcal{N}(x|\mu_i, \Sigma_i)$, where

$$\mathcal{N}(x|\mu_i, \Sigma_i) = \frac{1}{\sqrt{|2\pi\Sigma_i|}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}, \tag{5.1}$$

where $|\cdot|$ applied to a square matrix denotes the determinant of the matrix. Following this model the pairwise Chernoff distance $d_{ch}(p_i, p_j; s)$ between $p_i(x)$ and $p_j(x)$

is given in closed-form [42, p. 98] as

$$d_{ch}(p_i, p_j; s) = \frac{s(1-s)}{2}\Delta\mu_{ij}^T\Sigma_{ij}^{-1}\Delta\mu_{ij} + \frac{1}{2}\log\left(\frac{|\Sigma_{ij}|}{|\Sigma_j|^s|\Sigma_i|^{1-s}}\right), \tag{5.2}$$

where $\Delta\mu_{ij} = (\mu_j - \mu_i)$ and $\Sigma_{ij} = s\Sigma_j + (1-s)\Sigma_i$. By substituting $d_{ch}(p_i, p_j; s)$ as in (5.2) in (4.15), we obtain the bound for the probability of error for the Gaussian class-conditional model as

$$P_e^* \leq \sum_{i=1}^{n}\sum_{j\neq i}\left(\frac{\pi_j|\Sigma_j|}{|\Sigma_{ij}|}\right)^{s_{ij}}\left(\frac{\pi_i|\Sigma_i|}{|\Sigma_{ij}|}\right)^{1-s_{ij}} e^{-\frac{s_{ij}(1-s_{ij})}{2}\Delta\mu_{ij}^T\Sigma_{ij}^{-1}\Delta\mu_{ij}}. \tag{5.3}$$

5.1.1 The homoscedastic case

The bound in (5.3) is applicable for any class covariance structure and is referred to as the heteroscadastic Gaussian case. A special case is the case in which each class distribution is model as Gaussian with identical covariance structure, i.e., $\Sigma_i = \Sigma$. This scenario is referred to in the literature as the homoscedastic case. Application of this property suggest that $\Sigma_{ij} = \Sigma_i = \Sigma_j = \Sigma$ and hence the $\log(\cdot)$ term in (5.2) vanishes. The corresponding pairwise Chernoff distance between classes is

$$d_{ch}(p_i, p_j; s) = \frac{s(1-s)}{2}\Delta\mu_{ij}^T\Sigma^{-1}\Delta\mu_{ij}.$$

Note that in this case the optimal choice for s maximizing the distance and hence tightening the bound is $s = \frac{1}{2}$. This lead to a simple form of the bound for the

Gaussian homoscedastic case given by

$$P_e^* \leq \sum_{i=1}^{n} \sum_{j \neq i} \sqrt{\pi_i \pi_j} e^{-\frac{1}{8} \Delta\mu_{ij}^T \Sigma^{-1} \Delta\mu_{ij}}. \tag{5.4}$$

Note that the argument of the exponent is negative of Fisher's discriminant. This bound suggests a sum of negative exponential approach for extending Fisher's discriminant analysis to the multiclass case.

The Gaussian model and even more so the homoscedastic Gaussian model may not accurately represent the data due to their simplified structure. We are interested in extending our approach to the Gaussian mixture model.

5.2. Chernoff union bound for GMM class-conditional distributions

The generality of the Gaussian mixture model (GMM) makes it an attractive alternative to the Gaussian model. Gaussian mixture models can describe fairly complex non-Gaussian distributions quite accurately with adequate number of mixture components. Unfortunately, a closed-form expression for the Chernoff distance between two GMMs is unavailable. Mutual information based LDR methods that assume that the data follows the GMM use sample estimates in evaluating the MI. A direct computation in high-dimensions is computationally costly and Monte-Carlo evaluation [49, 50] may introduce estimation errors. We tackled the aforementioned issues using the closed-form bound for the Gaussian case. Next, we present an extension of the bound to the GMM case.

We start with required mathematical notations. As with the classification

setup in Section 3., the class-conditional PDF for the ith class is given by $p_i(x)$. In the GMM representation for class i, we assume k_i Gaussian mixture components $\phi_{i1}(x), \phi_{i2}(x), \ldots, \phi_{ik_i}(x)$ with the mixing coefficients of $\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}$ such that $\sum_{k=1}^{k_i} \alpha_{ik} = 1$ and $\alpha_{ik} \geq 0$. The ith class-conditional PDF for the GMM is given by

$$p_i(x) = \sum_{k=1}^{k_i} \alpha_{ik} \phi_{ik}(x). \tag{5.5}$$

Substituting the expression for $p_i(x)$ from (5.5) into (4.13), we obtain

$$e^{-d_{ch}(p_i,p_j;s)} = \int \Big(\sum_{l=1}^{k_j} \alpha_{jl} \phi_{jl}(x) \Big)^s \Big(\sum_{k=1}^{k_i} \alpha_{ik} \phi_{ik}(x) \Big)^{(1-s)} dx. \tag{5.6}$$

A closed-form expression for the integral in (5.6) is unavailable and hence we resort to an application of an additional bound. Consider the following inequality (see appendix 1)

$$\Big(\sum |x_i| \Big)^t \leq \sum |x_i|^t, \quad 0 < t < 1. \tag{5.7}$$

Applying the inequality in (5.7) to (5.6), we obtain a bound for $e^{-d_{ch}(p_i,p_j;s)}$

$$\begin{aligned} e^{-d_{ch}(p_i,p_j;s)} &\leq \int \sum_{k=1}^{k_j} \left(\alpha_{jk} \phi_{jk}(x) \right)^s \sum_{l=1}^{k_i} \left(\alpha_{il} \phi_{il}(x) \right)^{(1-s)} dx \\ &= \sum_{k=1}^{k_j} \sum_{l=1}^{k_i} \alpha_{jk}^s \alpha_{il}^{1-s} \int \phi_{jk}(x)^s \phi_{il}(x)^{(1-s)} dx. \end{aligned} \tag{5.8}$$

Application of the bound in (5.8) to (4.15) yields

$$P_e^* \leq \sum_{i=1}^{n} \sum_{j \neq i} \min_s \pi_i^s \pi_j^{1-s} \sum_{k=1}^{k_i} \sum_{l=1}^{k_j} \alpha_{ik}^s \alpha_{jl}^{1-s} \cdot \int \phi_{ik}(x)^s \phi_{jl}(x)^{(1-s)} dx. \tag{5.9}$$

Rearranging, we obtain the bound on the error probability as,

$$P_e^* \leq \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{k_i} \sum_{l=1}^{k_j} \min_s (\pi_i \alpha_{ik})^s (\pi_j \alpha_{jl})^{1-s} \cdot \exp(-d_{ch}(\phi_{ik}, \phi_{jl}, s)). \tag{5.10}$$

By the definition of the GMM, $\phi_{ik}(x)$ and $\phi_{jl}(x)$ are Gaussian PDFs and hence $\phi_{ik}(x) = \mathcal{N}(x, \mu_{ik}, \Sigma_{ik})$. The Chernoff distance between the mixture components is given by,

$$d_{ch}(\phi_{ik}, p_j jl; s) = \frac{s(1-s)}{2} \Delta\mu_{kl}^{ij\,T} \Sigma_{kl}^{ij\,-1} \Delta\mu_{kl}^{ij} + \frac{1}{2} \log \left(\frac{|\Sigma_{kl}^{ij}|}{|\Sigma_{lj}|^s |\Sigma_{ik}|^{1-s}} \right), \tag{5.11}$$

where $\Delta\mu_{kl}^{ij} = (\mu_{jl} - \mu_{ik})$ and $\Sigma_{kl}^{ij} = s\Sigma_{lj} + (1-s)\Sigma_{ik}$. In the previous section, we had already mentioned the availability of closed-form expression for Chernoff distances between Gaussian PDFs (5.2).

5.3. Chernoff union bound for non-parametric class-conditional distributions

One important factor that can greatly affect the performance of a GMM based LDR is the uncertainty in estimating the GMM parameters and the number of mixture components that must be assumed or estimated through model order selection. This directly affects the performance of the proposed LDR algorithm. As an alternative, we consider the use kernel density estimation [51], i.e., a non-parametric approach to DR. Under the KDE model, each class-conditional distribution $p_i(x)$ is constructed by adding kernel functions centered about the data points whose label is i:

$$p_i(x) = \frac{1}{N_i} \sum_{k=1}^{N} I(y_k = i) K(x - x_k; \sigma), \tag{5.12}$$

where $N_i = \sum_k I(y_k = i)$ is the number of points drawn from class i, and σ is a bandwidth parameter typically used in balancing bias and variance associated with the KDE. To fit to the GMM framework used earlier, we consider a kernel function $K(\cdot; \sigma) : \mathbb{R}^d \to \mathbb{R}$ of the form

$$K(x; \sigma) = \frac{1}{(\sqrt{2\pi\sigma^2})^d} \; e^{-\frac{\|x\|^2}{2\sigma^2}}. \tag{5.13}$$

Since $p_i(x)$ is given a convex combination of Gaussian PDFs, it can be viewed as a special case of the GMM where the number of mixtures in $p_i(x)$ is N, the mixture components are $\phi_{ik} = \mathcal{N}(x; x_k, \sigma^2)$, the mixing coefficients are $\alpha_{ik} = \frac{1}{N_i} I(y_k = i)$ and the prior probability is $\pi_i = \frac{N_i}{N}$. In such a case, the expression in (5.14) naturally offers an upper bound on the probability of error. Substituting the aforementioned

values for ϕ_{ik}, α_{ik}, and π_i into (5.10), we obtain

$$P_e^* \leq \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{N} \sum_{l=1}^{N} \min_s \left(\frac{N_i}{N} \frac{1}{N_i} I(y_k = i) \right)^{1-s} \left(\frac{N_j}{N} \frac{1}{N_j} I(y_l = j) \right)^{s} \cdot \exp(-d_{ch}(\phi_{ik}, \phi_{jl}, s)). \quad (5.14)$$

Since $\phi_{ik} = \mathcal{N}(x; x_k, \sigma^2)$ and $\phi_{jl} = \mathcal{N}(x; x_l, \sigma^2)$, we have

$$d_{ch}(\phi_{ik}, \phi_{jl}; s) = \frac{s(1-s)}{2} \frac{\|x_k - x_l\|^2}{\sigma^2} \quad (5.15)$$

Substituting (5.15) into (5.14) and simplifying, yields

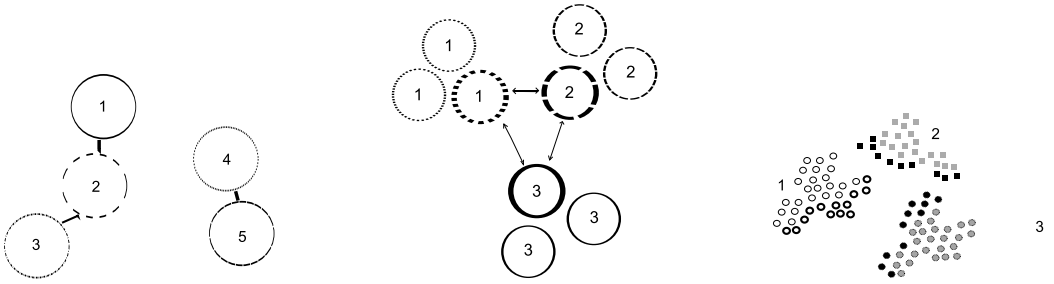
$$P_e^* \leq \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{N} \sum_{l=1}^{N} \min_s \frac{1}{N} I(y_k = i) I(y_l = j) e^{\frac{s(1-s)}{2} \frac{\|x_k - x_l\|^2}{\sigma^2}}. \quad (5.16)$$

Since $\sum_i \sum_{j \neq i} I(y_k = i) I(y_l = j) = \sum_i I(y_k = i) \sum_{j \neq i} I(y_l = j) = \sum_i I(y_k = i) I(y_l \neq i) = I(y_k \neq y_l)$ and since the minimum on the RHS is achieved at $s = \frac{1}{2}$, we can write (5.16) as

$$P_e^* \leq \frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} I(y_k \neq y_l) e^{-\frac{\|x_k - x_l\|^2}{8\sigma^2}}. \quad (5.17)$$

The bound derived in (5.17) can be interpreted in the following way. The dominant terms on the RHS are those exponentials for which $\|x_k - x_l\|^2/(8\sigma^2)$ is small. This suggests that the cases when the two points x_k and x_l are close, they contribute most to the probability of error. In LDR, the x_is depend on the linear transformation matrix A. To minimize the the bound for the minimum error probability, we must keep the closest pairs of points far to keep the bound to a minimum. An illustration

of the above is given in Fig. (5.1(c)). We present a two-dimensional projection of a three-dimensional data in which the points that contribute to most probability of error are marked in dark. The LDR method based on our approach will thus be controlled by these anchor points that lie in the boundary of the classes.



(a) Gaussian model (b) Gaussian mixture model (c) Kernel density estimate

FIGURE 5.1: An illustration marking the physical counterpart of the terms in the CUB bounds that increase the probability of error for the three data model. A dark line between distributions indicate that the distance between them introduces most of the classification error.

5.4. summary

In this chapter, we presented the generalized CUB bound for the error probability and applied it to three data models, i.e., Gaussian, GMM, and KDE. All the CUB bounds follow a similar flavor of summation over exponent of negative distances. The relation between exponent of negative distances to probability of error surfaces when we look at the terms that increase the probability of error, which are the distances that are small. These terms that dominate our cost functions are what we would like to describe as the region of interest. To visualize the region of interest when using CUB algorithm, a lower dimensional projection of datasets is illustrated in fig. (5.1). For a particular lower dimensional transformation A, the

CUB algorithms is illustrated for cases when the data is parametrized as Gaussian, GMM and Kernel density estimates. The class pairs that adds most weight to the CUB learning process is marked with a dark connection between the distributions. In the Gaussian case (fig 5.1(a)) we that the class pairs $1-2, 2-3, 4-5$ in the lower dimensional plane are the dominant factors which will influence CUB. In the case of GMM (Fig. 5.1(b)) the distances between the mixture components $1, 2$ and 3 will guide the CUB LDR. In non-parametric CUB the data points that are closer to the other class act as support points that help in finding a good projection plane, it is represented in Fig. 5.1(c) with support points marked darker than the rest of the data points. We see that CUB bounds identify the region of interest that most intuitively contribute to the classification error, Next we demonstrate how the bound can be used for LDR.

6. ERROR BOUND MINIMIZATION ALGORITHMS

6.1. Cost function

In this section, we present a criterion for LDR based on the CUB presented in the previous section and provide an gradient based algorithm for the minimization of the criterion. Under the correct data model, the minimization of the bound as a function of the linear transformation matrix A provide upper bound guarantees on the probability of error. The bound derived in the GMM case is the most general. Both the bound for the Gaussian case and the KDE case (see previous section) can be obtained through a special choice of the parameters of the GMM. For example, when number of the mixture components in each class is 1, the bound for the GMM case reduces to the bound for the Gaussian case. Therefore, we proceed with the bound for the GMM as the LDR objective function

$$J(A) = \sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k}\sum_{l}(\pi_i\alpha_{ik})^s(\pi_j\alpha_{jl})^{1-s} \exp(-d_{ch}(\phi_{ik}(\cdot,A),\phi_{jl}(\cdot,A);s)), \tag{6.1}$$

where $\phi_{ik}(\cdot,A) = \mathcal{N}(\cdot|A\mu_{ik}^z, A\Sigma_{ik}^z A^T)$ and μ_{ik}^z and Σ_{ik}^z are the mean and covariance of the kth component of the ith class in the original space $\mathbb{R}^m$, respectively. It follows that,

$$d_{ch}(\phi_{ik}(\cdot,A),\phi_{jl}(\cdot,A);s)) = \frac{s(1-s)}{2}\Delta\mu_{kl}^{ij\,T}A^T\Sigma_{kl}^{ij\,-1}A\Delta\mu_{kl}^{ij} + \frac{1}{2}\log\left(\frac{|A\Sigma_{kl}^{ij\,z}A^T|}{|A\Sigma_{ik}^z A^T|^s|A\Sigma_{jl}^z A^T|^{1-s}}\right), \tag{6.2}$$

where, $\Sigma_{kl}^{ij} = A\Sigma_{kl}^{ij^z}A^T$, $\Sigma_{kl}^{ij^z} = (s\Sigma_{ik}^z + (1-s)\Sigma_{jl}^z)$, $\Delta\mu_{kl}^{ij} = A\Delta\mu_{kl}^{ij^z}$ and $\Delta\mu_{kl}^{ij^z} = \mu_{jl}^z - \mu_{ik}^z$.

Without loss of generality, we propose the following constraint minimization

$$\min_{A\in\mathbb{R}^{d\times n}} J(A) \text{ subject to } AA^T = I_d. \tag{6.3}$$

Suppose that a full rank A does not satisfy the constraint, we can always find $B = GA$, where G is $d \times d$ invertible matrix that satisfies that constraint since $BB^T = GAA^TG^T$ can be made I. Note however that the cost function remains the same if we replace A with $B = GA$. Transforming the features with a one-to-one transformation implies that the features can then be recovered and hence we expect the classification performance to remain the same. Hence the solution of the constraint problem is identical to the solution of the unconstrained problem.

6.2. Gradient Descent with Unitary Constraints

In our specific setting, A is orthogonal such that $AA^T = I_d$, where I_d is an identity matrix of dimensions $d \times d$. Orthogonality restriction helps to avoid uncontrolled scaling of the transformation matrix. A standard technique to optimize a cost function such as J(A) is the gradient based optimization. To minimize the cost function (6.1) w.r.t. A, we adopt the gradient descent method. The update

Algorithm 1: GMM-CUB LDR Algorithm

Input : Set of $\{\mu_i^z, \Sigma_i^z, threshold\lambda\}$
Output: Matrix $A \in \mathbb{R}^{m \times d}$

begin
  $A^{(0)} \leftarrow R$ and $(m, d) : AA^T = I_d$
  repeat
    for $i \leftarrow 1$ **to** n **do**
      for $j \leftarrow i + 1$ **to** n **do**
        for $k \leftarrow 1$ **to** n_k **do**
          for $l \leftarrow 1$ **to** n_l **do**
            evaluate $d_{ch}(\phi_{ik}, \phi_{lj}; s)$
            evaluate $\nabla_A d_{ch}(\phi_{ik}, \phi_{lj}; s)$
          end
        end
      end
    end
    evaluate $J(A)$ and $\nabla_A J(A)$
    $A^{(t+1)} = A^{(t)} - \epsilon \nabla_A \tilde{J}(A^{(t)})$
  until $|J(A)^t - J(A)^{(t-1)}| < \lambda$;
end

iteration is given by,

$$A^{(t+1)} = A^{(t)} - \epsilon \nabla_A \tilde{J}(A^{(t)}), \tag{6.4}$$

$$\nabla_A \tilde{J}(A) = \nabla_A J(A) - \frac{1}{2}\left(\nabla_A J(A) A^T + A \nabla_A J(A)^T\right). \tag{6.5}$$

The variable ϵ is the step size of the gradient descent and $\tilde{J}(A^{(t)})$ is the gradient of our cost function. The use of $\nabla_A \tilde{J}(A)$ instead of $\nabla_A J(A)$ is to incorporate orthogonality constraints to our transformation matrix A [52]. In our simulations, we consider a line backtracking method to determine a value for ϵ that guarantees a reduction in the cost function. Since our cost function in (6.1) is continuous and

differentiable, we can obtain its gradient as

$$\nabla_A J(A) = -\sum_i^n \sum_{j\neq i}^n \sum_k \sum_l (\pi_i \alpha_{ik})^s (\pi_j \alpha_{jl})^{1-s} \\ e^{-d_{ch}(\phi_{ik},\phi_{jl},s,A)} \nabla_A d_{ch}(\phi_{ik}, \phi_{jl}, s, A). \quad (6.6)$$

The gradient of $d_{ch}(\phi_{ik}, \phi_{jl}, s, A)$, the Chernoff distance between two Gaussian mixture components has been discussed in eq. (19) of [38] or in [33] and is given by

$$\nabla_A d_{ch}(\phi_{ik}, \phi_{jl}, s, A) = s(1-s)(\Sigma_{kl}^{ij\,-1} \Delta\mu_{kl}^{ij} \Delta\mu_{kl}^{ij^z T} - \Sigma_{kl}^{ij\,-1} \Delta\mu_{kl}^{ij} \Delta\mu_{kl}^{ij\,T} \Sigma_{kl}^{ij\,-1} A \Sigma_{kl}^{ij^z}) \quad (6.7)$$

$$+(\Sigma_{kl}^{ij\,-1} A \Sigma_{kl}^{ij^z} - s\Sigma_{ik}{}^{-1} A \Sigma_{ik}^z - (1-s)\Sigma_{jl}{}^{-1} A \Sigma_{jl}^z) \quad (6.8)$$

A summary of our algorithm is provided in Algorithm 1. Convergence of the algorithm to local minima has been tackled by repeated random initializations.

7. NUMERICAL STUDY

7.1. Toy example

In this section, we consider a comparison of the non-parametric version of our algorithm to other methods that are based on the non-parametric assumption. Our comparison includes two DR methods: i) neighborhood component analysis (NCA) [35] and ii) quadratic mutual information (QMI) [28]. NCA is based on the maximization of probability that the first nearest neighbor belongs to the same class for every data point. The idea is based on a smooth extension to the hard notion of a neighborhood. QMI follows the approach of maximum mutual information while replacing the difficult to compute Shannon's MI with a quadratic mutual information. All of the algorithms considered consider an optimization of an objective function over the space of linear transformation.

To gain some insight regarding the cost function used in our proposed method and the two aforementioned algorithms, we consider the following toy example. The dataset created consists of three Gaussian classes $\mathcal{C} = \{1, 2, 3\}$ defined over a two-dimensional Euclidean space (see Fig. 7.1(a)). The mean and covariance of each class is denoted by μ_1, μ_2, and μ_3 and Σ_1, Σ_2, and Σ_3 respectively. We consider identical diagonal covariance matrices, i.e., $\Sigma_1 = \Sigma_2 = \Sigma_3$. The variance along the x-axis is 1 and the variance along the y-axis is 49. The classes are centered at $\mu_1 = [0, 26]^T$, $\mu_2 = [0, -26]^T$ and $\mu_3 = [-20, 0]^T$. Introducing the class mean in this manner places the classes 1 and 2 at a large distance of 52 vertically on top of each other and the class 3 at a relatively smaller distance of 20 to the left of classes 1 and 2. The bandwidth selection process is unsupervised, the bandwidth

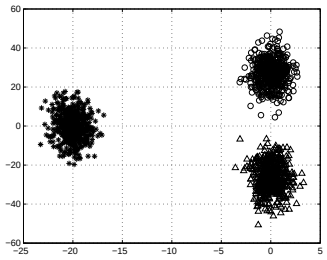
will be biased towards a larger variance vertically (to accommodate classes 1 and 2). Under this condition we evaluate the aforementioned methods. The classes are clearly separable on 1D projections along the angles of $10^\circ - 18^\circ$, $163^\circ - 169^\circ$ e.g., a projection along 13° produces a clear class separation as shown in Fig. 7.1(b). For the toy example considered, the cost function used by QMI can be found in closed-form

$$J_{QMI}(\theta) = \frac{1}{\sqrt{4\pi\sigma_d}} + \sum_i^3 \sum_j^3 \frac{1}{\sqrt{4\pi\sigma_d^2}} \exp(-(\mu_{ij}^x \cos(\theta) + \mu_{ij}^y \sin(\theta))^2/(4\sigma_d^2)) \quad (7.1)$$

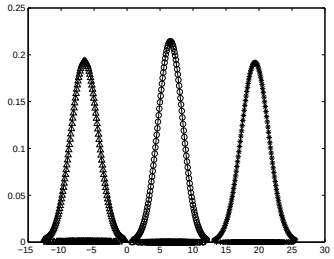
where $\mu_{ij}^x = \mu_i(1) - \mu_j(1)$ is the difference in the x-coordinates of the means of class i and j, $\mu_{ij}^y = \mu_i(2) - \mu_j(2)$ represents the difference in the y-coordinates of the means between class i and j, and θ is the angle defining the one-dimensional linear transformation $A = [\cos(\theta), \sin(\theta)]$. In the analytic formulation we can evaluate the projected covariance in one-dimension as $\sigma_d^2 = \cos(\theta) + 49\sin(\theta)$. We now compare the sample estimated cost and the analytic cost across all angles of the one-dimensional projection. A plot of the estimated cost obtained by evaluating based on samples in a non-parametric manner and the analytic expression (7.1) is provided in Fig. (7.2(e)). We note the correspondence between the analytic cost function and the cost evaluated using data samples. It can be noted that the analytic cost evaluated is higher than the estimated cost in regions where the cost function is high. This occurs due to error in the sample based approximation of analytic integral becoming significantly noticeable when classes are further apart. The QMI cost function achieves maximum value at 177° (see Fig. 7.2(e)). The PDF associated with the one-dimensional projection of the data for $\theta = 177^\circ$ is shown in Fig. (7.1(c)). We notice significant overlap in classes and that the resulting linear projection of

the data is significantly different than that in Fig. 7.1(b), which is achieved by a projection associated with $\theta = 193°$. We propose the following explanation. The QMI cost function for this toy example (7.1) is inversely related to the projected class variance σ_d^2 which is minimized at $\theta = 0°$ or $180°$ (result of our initial choice of the class means). This drives the maximizer of the QMI objective towards $\theta = 0°$ or $180°$. While this may seem an artifact of the example presented here, it is a deeper issue that is associated with the replacement of Shannon MI, which is used to bound probability of error, with the QMI. One property of statistical divergences (such as the MI) is that they are invariant to invertible linear transformations. This property does not hold for QMI. The cost function used by our proposed algorithm (5.17) is minimized at $193°$ as seen in Fig. (7.2(b)). Our method seem to balance well the within-class covariance and the between class covariance. We see the true training error for the toy data being discussed in Fig. (7.2(a)) and see the correspondence with the CUB cost function.

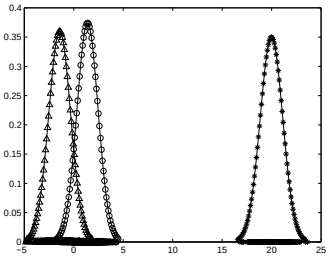
We analyze the cost function of NCA on the same dataset. NCA yields a projection plane that distinctly separates the three classes as shown in Fig. 7.1(d). Since no constraints are applied to the matrix A in NCA, NCA allows for a magnitude scale. In some case, this could lead to increased distance between points to the point that the second nearest neighbor is significantly further away as compared to the first neighbor. This lead to a non smooth cost function that is dominated by first-nearest neighbor distances. This of course is counter to the intention of the approach, which is to provide a smooth alternative to a non-smooth first-nearest neighbor based objective. From Fig. 7.2(c), we observe that for lower magnitudes of the transformation matrix the cost remains smooth. However, the cost becomes non-smooth for higher magnitudes. This could result in difficulties during the opti-



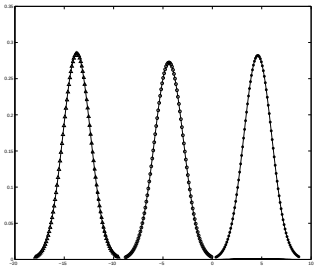
(a) 2D toy dataset



(b) Ideal projection at 193°



(c) Projection obtained by QMI at 177°



(d) Projection obtained by NCA

FIGURE 7.1: Comparison of the projection of a 2-dimensional data in Fig. (7.1(a)) on a 1-dimensional projection plane obtained by QMI and KDE for a toy dataset.

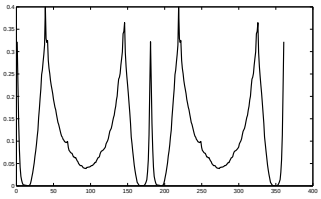
mization of the NCA objective.

7.2. Numerical Performance Evaluation

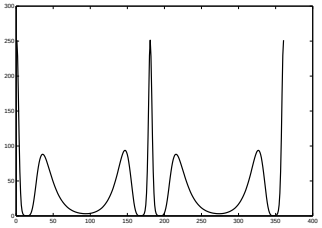
In the following, we present the setup used for evaluating the proposed methods and for the comparison with alternative methods along with analysis of the results.

Datasets

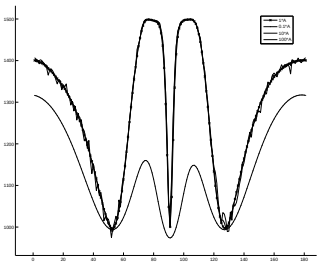
To conduct a numerical performance study, we consider several datasets. We selected the following datasets: Landsat, Phoneme and Optdigits. The phoneme



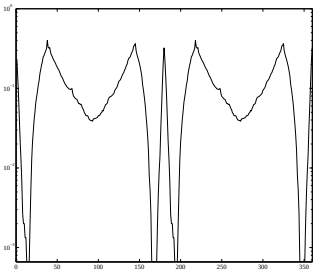
(a) True train error vs all angles



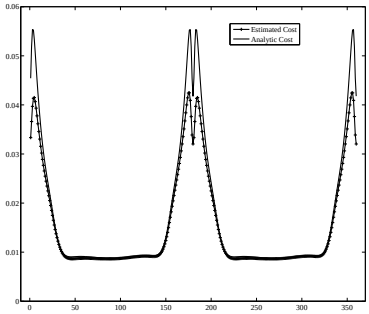
(b) CUB:KDE: Training error vs Angle of projection



(c) NCA: Cost function vs Angle of projection for different magnitudes of projection plane



(d) QMI: Training error vs Angle of projection



(e) Cost vs Angle evaluated from data

FIGURE 7.2: Plots of cost vs 1D angle of projection for QMI, KDE-CUB and NCA for the dataset in Fig. (7.1(a))

dataset was obtained from Stanford database[2], the rest of the datasets were obtained from UCI machine learning repository[3]. The datasets have been selected from a wide range of dimensions and sizes in order to bring out the advantages and disadvantages of DR methods with dimensions and class size. Table 7.2. lists the datasets along with the following parameters: data dimension, data size (total number of available points), and the number of classes. The following criteria were considered when selecting the datasets: i) no. of classes - to illustrate the applicability of the method to the multiclass case we consider datasets with more than two classes and ii) data structure - the datasets picked varied to provide good fit to any either of the models considered in this thesis: Gaussian, Gaussian mixtures, and kernel density estimates.

Name	Dimension	Size	no.Classes
Landsat	36	6435	6
Phoneme	256	4509	5
Optical Digits	64	5620	10

TABLE 7.1: Datasets used in simulations

7.2.1 Dimension reduction methods

For each data model, we considered a different collection of DR methods. For each DR method, we considered the corresponding model based classifier (e.g., a GMM likelihood ratio test for the GMM data model). The following DR methods we considered (data models assumed for each DR are included in parentheses):

---

[2]http://www-stat.stanford.edu/ tibs/ElemStatLearn/

[3]http://archive.ics.uci.edu/ml/

1. **PCA**: Principle component analysis [1] (all data models). We constructed the covariance matrix of the data and applied an eigendecomposition to it. We then selected the d principle vectors (i.e., the eigenvectors corresponding the d-largest eigenvalues) for the rows of A. We included the results of PCA for all data models.

2. **LDA**: Linear discriminant analysis [19] (all models)

   There are numerous variants of LDA in literature for the multiclass case. We considered the variant of LDA, which is based on the generalized eigendecomposition of the between-class covariance matrix to the within class covariance. As in PCA, the eigenvectors corresponding to the d-largest eigenvalues were used to construct the rank d linear projection matrix A. For implementation details, we refer the reader to the introduction section in [53]

3. **RH**: The algorithm proposed by Rueda and Hererra in [36] based on the maximization of the sum of Chernoff distance (Gaussian data model) for multiclass LDR. In our implementation we used $s = 0.5$ to evaluate the Chernoff distance.

4. **MMI**: Maximum mutual information [44] (GMM data model). A mutual information based cost criterion that is maximized using gradient methods. The DR method is suitable for data modeled as Gaussian mixtures but the actual cost function is based on sampled evaluations.

5. **QMI**: Quadratic mutual information [37] (non-parametric)

   The method is based on a cost function called the quadratic mutual information and fits the DR setup where a non-parametric assumption of the data holds.

6. **NCA**: Neighborhood component analysis [35] (non-parametric)

   A gradient based method that aims at learning a distance metric for the data resulting in maximizing the probability of data in a given class. This method is suitable for the non-parametric LDR.

7. **GCUB**: CUB algorithm for Gaussian data (Gaussian data model)

   The cost function is based on the bound developed in section 5.1..

8. **GMM-CUB**: CUB algorithm for GMM (GMM data model)

   The cost function is based on the bound developed in section 5.2..

9. **CUB-KDE**: CUB for KDE (non parametric)

   The cost function is based on the bound developed in section 5.3..

Each DR method requires explicitly or implicitly the parametrization of the data model. The parameters for each data model were obtained as follows. For the Gaussian data model, class mean and covariance parameters were computed using the training data for each class. For the GMM case, the expectation maximization algorithm for GMM was used to obtain the mean, covariance, and prior probability for each Gaussian mixture component for each class. In our setup of the problem, finding the optimum number of class mixtures is not a factor in comparison of the methods since the same GMM was used for all GMM based DR methods. Initially we identified a suitable range of mixtures for each class based on a maximum likelihood criterion. We then selected a particular configuration and stored that to be used by all methods. The number of mixtures were not optimized to offer any improvement in classification results. The GMM error rates are sole representative of the DR method and could be improved if appropriate GMM configuration is identified. For the non-parametric case, a Gaussian kernel was adopted for KDE with

a covariance of $\sigma^2 I$. For each dataset, the bandwidth parameter σ^2 was obtained using the maximum likelihood principle [54]. For each dataset, the same value of the bandwidth parameter was used for all DR methods and all classifiers.

7.2.2 Classifiers compared

In this thesis we developed DR algorithms specifically for cases where data classification is desired after dimension reduction. our performance evaluation criterion for a DR method is thus the error produced by the dataset after the DR step. We adopted Bayes classifier for the data model under consideration. The general form of a Bayes classifier is given by (3.3). Its specific form for each data model is listed as follows:

1. **Quadratic classifier (Q):**

   For the Gaussian data model, Bayes classifier simplifies to the quadratic classifier (Q) given by

   $$y = \arg\max_i \ \log \pi_i - \frac{1}{2}\log\det(2\pi\Sigma_i) - \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i), \qquad (7.2)$$

   where μ_i and Σ_i are the mean and covariance of class i, respectively. Note that while the classifier was obtain by taking the log of the Gaussian PDF, the resulting form is quadratic in x.

2. **GMM Bayes classifier (GMC)**:

   For data that is distributed according to Gaussian mixture model, we utilized the GMM classifier (GMC). The general form of the GMC is given by

   $$y = \arg\max_i \pi_i \sum_{k=1}^{k_i} \alpha_{ik}\mathcal{N}(x;\mu_{ik},\Sigma_{ik}), \qquad (7.3)$$

where k_i, α_{ik}, μ_{ik}, and Σ_{ik} denote the number of Gaussian mixtures, the k component prior probability, the k component mean, and the kth component Covariance, for class i, respectively.

3. **KDE Bayes classifier (KC)**:

   For data that is modeled use the KDE, we considered the KDE Bayes classifier (KC) given by

$$y = \arg\max_i \ \pi_i \frac{1}{N} \sum_{k=1}^{N} I(y_k = i) K(x - x_k). \tag{7.4}$$

4. **k-nearest neighborhood classifier (KNN)**:

   Motivated primarily by NCA [35]. We considered the k-nn estimator [55]. In our simulations, we selected $k = 1$ in the k-nn classifier.

7.2.3 Cross-validation setup

All our simulations involve a DR step followed by a classification step. Hence, the performance evaluation is based on the error produced by the classifier for the given DR method and dataset. To robustly estimate the probability of error for the given dataset we employ a cross-validation scheme. For datasets which were split into training and test subsets, we first combine the two subsets into a single dataset. We then partition the complete dataset into 75% training data and 25% test data at random. For each dataset, we construct 15 different data partitions. For each partition, probability of error was calculated and from the collection of 15 values of probability of error a mean value and standard deviation were obtained and reported. A particular cross-validation setup remained the same over the analysis of multiple DR methods to ensure fair comparison.

7.3. Analysis of results

We proceed with the evaluation and the analysis of the proposed LDR methods and provide a comparison to classical and state-of-the-art LDR techniques. As the probability of error depend on the classifier applied, we group the results of various LDR techniques based on the classifier used.

We present the DR results for each dataset used in individual tables: Landsat (7.3.), Optical digits (7.3.) and Phoneme (7.3.). Each table is split into four sections horizontally. Each section correspond to a choice of a data model assumption and the classifier applied. For section of the table, we compare various classifiers that are based (implicitly or explicitly) on the data model assumption. The tables are split into four sections: i) Gaussian assumption and quadratic classifier (Q), ii) Gaussian mixture model (GMM) assumption and Gaussian mixture classifier (GMC) iii) non-parametric data model (KDE) with a KDE Bayes classifier (KC) and iv) non-parametric data model (KDE) with a k-nn classifier.

Along a particular row of the table, we list the probability of error as a function of the dimension of projection. The reported error in each cell is in the format of mean $\pm$ standard deviation, which were obtained through 15 independent cross-validation trials.

In the Gaussian setup, we constantly see that CUB outperforms the classical methods of PCA and LDA in Landsat and Optdigit datasets. It also outperforms the distance based criterion developed by RH. This supports the argument in the Section 4.3. suggesting that a bound based approach outperforms the sum of between class distance maximization approach. In the phoneme dataset PCA seems to perform as good as CUB for higher dimensions. This is due to the fact that the intrinsic

dimensions of the data is completely captured by the number of dimensions of projection.

Gaussian mixture models have the power to describe the data more accurately. In cases where the data follows the Gaussian model, a GMM approach may not offer significant improvement. In cases where the data follows the GMM, a DR method utilizing the GMM model offers better result. We observe that GMM-CUB outperforms CUB for Landsat dataset and in higher dimensions of phoneme. GMM-CUB performs competitively better than MMI for all dimensions in Landsat, Optical digits and the Phoneme datasets.

In the non-parametric setup, we observe that KDE-CUB outperforms state-of-the-art method such as NCA. In general, KDE-CUB offers lower error rate than QMI or NCA for most datasets and dimensions. KDE-CUB offers the best classification error in dimensions 6 and 8 of the Optdigits dataset. Although NCA was initially proposed under the premise that it will improve k-nn performance, it seems that KDE classifier yield significantly better results. The NCA cost function adds up the contribution of each point in the negative exponent fashion and KDE classifier offers a similar formulation. Thus KDE classifier offers better result for NCA compared to results of using the first nearest neighbor for classifications. Significant efforts have been made in the implementation of all DR methods including QMI and NCA. Significant consideration was given to the optimization involved in these methods. In-spite of using similar gradient approaches to all the methods, our experience form simulations suggest that NCA converges very slowly compared to QMI and KDE-CUB. We observed CUB-KDE converged the fastest and resulted in competitive results compared to QMI and NCA in landsat, phoneme and higher dimension of optical digits.

	2	4	6	8
PCA/Q	17.55 ± 0.7	14.55 ± 0.4	14.30 ± 0.6	14.97 ± 0.7
LDA/Q	19.57 ± 0.6	14.26 ± 0.6	13.59 ± 0.6	13.49 ± 0.5
RH/Q	16.28 ± 0.61	14.87 ± 0.6	14.63 ± 0.8	14.80 ± 0.4
CUB/Q	**15.31 ± 0.5**	**14.08 ± 0.9**	14.42 ± 0.7	14.55 ± 0.8
PCA/GMC	16.41 ± 0.01	13.02 ± 0.01	11.39 ± 0.1	11.59 ± 0.1
LDA/GMC	20.47 ± 0.01	13.09 ± 0.01	11.97 ± 0.3	11.67 ± 0.2
MMI/GMC	16.12 ± 0.10	12.04 ± 0.02	11.22 ± 0.1	11.03 ± 0.4
GCUB/GMC	**15.95 ± 0.02**	**11.70 ± 0.02**	**10.77 ± 0.3**	**10.66 ± 0.3**
PCA/KC	16.13 ± 0.6	**12.25 ± 0.5**	**9.66 ± 0.3**	**9.10 ± 0.5**
LDA/KC	18.96 ± 0.6	12.25 ± 0.6	10.8 ± 0.7	10.65 ± 0.8
NCA/KC	18.3 ± 0.7	13.87 ± 0.8	11.90 ± 0.7	10.66 ± 0.3
QMI/KC	17.05 ± 1.6	13.89 ± 0.63	12.93 ± 0.3	12.60 ± 0.01
KCUB/KC	**15.90 ± 0.7**	12.5 ± 0.5	10.50 ± 0.4	9.91 ± 0.5
PCA/KNN	**21.36 ± 0.7**	**13.58 ± 0.5**	**10.58 ± 0.5**	**9.72 ± 0.5**
LDA/KNN	24.29 ± 1.1	12.83 ± 0.8	12.82 ± 0.8	12.80 ± 0.9
NCA/KNN	25.25 ± 3.3	17.7 ± 0.1	15.71 ± 0.9	15.49 ± 0.6
QMI/KNN	22.70 ± 1.9	24.19 ± 0.6	23.44 ± 0.7	16.69 ± 0.01
KCUB/KNN	21.98 ± 1.2	15.21 ± 0.6	11.5 ± 0.5	10.8 ± 0.7

TABLE 7.2: Error rates for Landsat dataset

	2	4	6	8
PCA/Q	38.08 ± 0.8	18.17 ± 0.1	8.22 ± 0.6	5.11 ± 0.4
LDA/Q	33.9 ± 1.0	9.0 ± 0.6	5.3 ± 0.5	3.6 ± 0.3
RH/Q	40.06 ± 4.5	19.70 ± 2.7	9.73 ± 1.5	5.26 ± 0.8
CUB/Q	**21.7 ± 1.0**	**5.72 ± 0.5**	**2.46 ± 0.6**	**2.15 ± 0.3**
PCA/GMC	39.28 ± 0.01	17.98 ± 0.01	8.26 ± 0.01	5.83 ± 0.1
LDA/GMC	35.66 ± 0.30	12.40 ± 0.01	7.21 ± 0.31	5.75 ± 0.1
MMI/GMC	26.15 ± 0.10	8.16 ± 0.4	5.71 ± 0.5	**3.85 ± 0.6**
GCUB/GMC	**24.50 ± 0.30**	**8.01 ± 0.2**	**5.47 ± 0.3**	3.87 ± 0.1
PCA/KC	36.74 ± 0.1	16.44 ±	6.25 ± 0.6	3.45 ± 0.5
LDA/KC	42.23 ± 0.8	19.89 ± 1.2	19.54 ± 2.5	18.33 ± 0.5
NCA/KC	55.71 ± 0.7	15.46 ± 0.6	8.80 ± 1.7	6.3 ± 0.4
QMI/KC	**24.36 ± 1.7**	**8.01 ± 0.7**	8.43 ± 1.3	7.97 ± 0.7
KCUB/KC	28.14 ± 2.4	8.20 ± 0.7	**3.4 ± 0.4**	**3.2 ± 0.3**
PCA/KNN	46.22 ± 1.2	18.17 ± 0.1	6.80 ± 0.7	3.63 ± 0.5
LDA/KNN	41.9 ± 1.3	9.0 ± 0.7	6.6 ± 0.7	4.9 ± 0.6
NCA/KNN	66.05 ± 1.7	9.67 ± 0.1	6.63 ± 0.07	9.97 ± 0.1
QMI/KNN	**32.53 ± 1.7**	**9.6 ± 0.7**	9.22 ± 0.8	9.53 ± 0.08
KCUB/KNN	38.5 ± 0.3	9.9 ± 0.8	**3.5 ± 0.4**	**2.30 ± 0.8**

TABLE 7.3: Error rates for Optical Digit dataset

	3	7	10	14
PCA/Q	12.0 ± 1.0	7.8 ± 0.7	7.9 ± 0.6	8.30 ± 0.08
LDA/Q	12.40 ± 0.5	**7.21 ± 0.4**	**7.31 ± 0.3**	**7.49 ± 0.4**
RH/Q	11.28 ± 0.5	10.11 ± 0.63	9.59 ± 0.07	9.35 ± 0.09
CUB/Q	**7.9 ± 0.4**	8.8 ± 0.5	8.9 ± 0.5	8.3 ± 0.8
PCA/GMC	12.75 ± 0.01	7.14 ± 0.01	7.58 ± 0.1	8.28 ± 0.01
LDA/GMC	13.81 ± 0.01	**8.10 ± 0.2**	8.52 ± 0.3	8.27 ± 0.05
MMI/GMC	9.95 ± 0.1	11.01 ± 0.5	10.81 ± 0.04	12.42 ± 0.2
GCUB/GMC	**8.14 ± 0.5**	8.27 ± 0.7	**7.63 ± 0.2**	**7.82 ± 0.04**
PCA/KC	13.7 ± 0.5	10.83 ± 0.4	11.02 ± 0.9	11.11 ± 0.8
LDA/KC	13.35 ± 0.04	**7.2 ± 0.04**	**7.25 ± 0.04**	**7.14 ± 0.04**
NCA/KC	13.3 ± 0.9	12.49 ± 0.9	12.5 ± 0.1	12.6 ± 11.2
QMI/KC	13.12 ± 0.03	11.78 ± 0.1	12.01 ± 0.04	11.95 ± 0.08
KCUB/KC	**12.37 ± 0.7**	11.53 ± 0.6	10.8 ± 0.7	10.3 ± 0.8
PCA/KNN	15.42 ± 0.5	11.02 ± 0.6	11.15 ± 0.7	11.16 ± 0.2
LDA/KNN	16.06 ± 0.06	**9.4 ± 0.06**	**9.31 ± 0.06**	**9.53 ± 0.06**
NCA/KNN	17.5 ± 1.2	16.4 ± 0.1	16.42 ± 1.6	16.4 ± 0.1
QMI/KNN	15.05 ± 0.08	14.51 ± 0.1	14.60 ± 0.6	12.50 ± 0.08
KCUB/KNN	**14.8 ± 0.7**	11.59 ± 0.7	10.7 ± 0.5	10.4 ± 0.8

TABLE 7.4: Error rates for Phoneme dataset

8. CONCLUSION

8.1. Summary

In this thesis, we proposed a novel linear dimension reduction method based on a bound approach on the probability of error. The advantages of the bound we use are: i) it is in closed-form, ii) it is applicable for a range of data models (Gaussian, Gaussian mixture model, and kernel density estimates), and iii) it offers lower or comparable classification error rates in comparison with other methods. From our numerical study we observed that the CUB based LDR methods offer either better or competitive classification errors compared with state-or-the-art methods. Moreover, since the CUB approach is applicable to a range of data models, it offers a unified framework for dimension reduction based on the bound on error probability.

8.2. Contributions

In an effort to develop dimension reduction methods for better data classification, our contribution includes three novel DR methods. The following are the itemized description of our contributions.

1 A novel multiclass LDR method based on bound on probability of error for Gaussian data model developed in Section. 5.1. We illustrate how CUB addresses the drawbacks in other state-of-the-art class separation based LDR method [31] and demonstrated the superiority of our algorithm using numerical results on real datasets. We call this method CUB algorithm.

2 The novel bound described in CUB is not applicable to data modelled as Gaussian mixtures due to closed-form constraints. We the developed a generalized novel bound for error probability for Gaussian mixtures in Section. 5.2. We call this method as GMM-CUB. We demonstrated the superior performance of GMM-CUB by comparing with LDR methods based on mutual information [24].

3 To extended GMM-CUB to non-parametric models of data we re-formulated the GMM-CUB bound in Section. 5.3. Our non-parametric model is based on kernel density estimates and hence we call this method CUB-KDE. We demonstrated competitive performance of CUB-KDE with state-of-the-art methods like NCA [35] and QMI [28].

8.3. Publications

Following are the list of publications that were prepared for conferences and journals as a result of the work described in this thesis.

[1] M. Thangavelu and R. Raich, Multiclass linear dimension reduction via a generalized Chernoff bound, in proc. IEEE Workshop on Machine Learning for Signal Processing, Cancun, Mexico, 2008, pp. 350355.

[2] M. Thangavelu and R. Raich, On linear dimension reduction for multiclass classification of Gaussian mixtures, in proc. IEEE Workshop on Machine Learning for Signal Processing, Grenoble, France, 2009, pp. 16.

[3] On Error Bounds for Linear Feature Extraction
IEEE Transactions on Pattern Analysis & Machine Intelligence (In preparation).

8.4. Future work

The current drawback of the proposed method is much similar to other gradient based methods, which is, occurrence of local minimas. Though it has been tackled with multiple initializations, the fundamental issue is the non-convex nature of the problem. Future work in this direction would involve developing a convex counterpart to the CUB cost function that can truly offer a quick and competitive method of linear dimension reduction.

Our work developed bound on error probability based on the technique of upper bound. At each step of bounding, the difference between thue probabiltiy of error and the bound loosens. Even though the idea of a bound over the error probability is demonstrated to offer good LDR methods, adopting tighter bound for the same framework can result in better results. An exploration in this direction could well open up a new direction of research, a good place to start with is the Bayesian bound which is tighter than the Chernoff bound used in this work.

The superiority and competitiveness of the methods developed in this thesis has been demonstrated numerical analysis and comparison with other state-of-the-art methods. Future work could involve developing computationally fast libraries in languages such as C, C++ to aid development and further research based on these algorithms.

BIBLIOGRAPHY

1. K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, no. 11, pp. 559–572, 1901.

2. J.J. Dai, L. Lieu, and D. Rocke, "Dimension reduction for classification with gene expression microarray data," *Statistical Applications in Genetics and Molecular Biology*, vol. 5, no. 1, pp. 1147, 2006.

3. I. Mierswa and K. Morik, "Automatic feature extraction for classifying audio data," *Machine learning*, vol. 58, no. 2, pp. 127–149, 2005.

4. K.M. Carter, R. Raich, and AO Hero, "Fine: Information embedding for document classification," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*. Citeseer, 2008, pp. 1861–1864.

5. B.M. Sarwar, G. Karypis, J.A. Konstan, and J. Riedl, "Application of dimensionality reduction in recommender system–a case study," in *ACM WebKDD 2000 Web Mining for E-Commerce Workshop*. Citeseer, 2000.

6. D. Reduction, "Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 4, pp. 179, 1994.

7. K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking," *Signal Processing-Image Communication*, vol. 12, no. 3, pp. 263, 1998.

8. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

9. I. Guyon and A. Elisseeff, "An introduction to feature extraction," *Studies in fuzziness and soft computing*, vol. 207, pp. 1, 2006.

10. J.B. Tenenbaum, V. Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," 2000.

11. S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," 2000.

12. LJP van der Maaten, "Dimensionality Reduction: A Comparative," .

13. R. Raich, J.A. Costa, and AO Hero, "On dimensionality reduction for classification and its application," in *Proc. IEEE Intl. Conference on Acoustic Speech and Signal Processing*. Citeseer, 2006, vol. 5.

14. X. Geng, D.C. Zhan, and Z.H. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 6, pp. 1098–1107, 2005.

15. S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40, 2007.

16. B. Scholkopf, A. Smola, and K.R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

17. A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

18. J. Venna and S. Kaski, "Nonlinear dimensionality reduction as information retrieval," in *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007), San Juan, Puerto Rico*. Citeseer, 2007, pp. 568–575.

19. PA Lachenbruch and M. Goldstein, "Discriminant analysis," *Biometrics*, pp. 69–85, 1979.

20. B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, 1992.

21. X. Wang and K.K. Paliwal, "A modified minimum classification error (MCE) training algorithm for dimensionality reduction," *The Journal of VLSI Signal Processing*, vol. 32, no. 1, pp. 19–28, 2002.

22. E. McDermott and S. Katagiri, "A new formalization of minimum classification error using a parzen estimate of classification chance," in *Proc. ICASSP*. Citeseer, vol. 3.

23. P.A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*, Prentice Hall, 1982.

24. J. M. Leiva-Murillo and A. Artes-Rodriguez, "Maximization of mutual information for supervised linear feature extraction," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1433–1441, 2007.

25. J. Peltonen and S. Kaski, "Discriminative components of data," *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 68–83, 2005.

26. U. Ozertem, D. Erdogmus, and R. Jenssen, "Spectral feature projections that maximize Shannon mutual information with class labels," *Pattern Recognition*, vol. 39, no. 7, pp. 1241–1252, 2006.

27. M. E. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. on Inform. Theory*, vol. 16, no. 4, pp. 368–372, July 1970.

28. K. Torkkola and W. M. Campbell, "Mutual information in learning feature transformations," in *Proc. International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 1015–1022.

29. Zoran Nenadic, "Information discriminant analysis: Feature extraction with an information-theoretic objective.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1394–1407, 2007.

30. E. Choi and C. Lee, "Feature extraction based on the Bhattacharyya distance," *Pattern Recognition*, vol. 36, no. 8, pp. 1703–1709, 2003.

31. L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space," *Pattern Recognition*, 2008.

32. Marco Loog and Robert P. W. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 732–739, 2004.

33. L. G. Rueda and M. Herrera, "A new linear dimensionality reduction technique based on Chernoff distance," in *Proc. Iberoamerican Congress on Pattern Recognition*, 2006, pp. 299–308.

34. AK Qin, PN Suganthan, and M. Loog, "Uncorrelated heteroscedastic LDA based on the weighted pairwise Chernoff criterion," *Pattern recognition*, vol. 38, no. 4, pp. 613–616, 2005.

35. Jacob Goldberger, Sam T. Roweis, Geoffrey E. Hinton, and Ruslan Salakhutdinov, "Neighbourhood components analysis," in *NIPS*, 2004.

36. L. G. Rueda and M. Herrera, "A new approach to multi-class linear dimensionality reduction," in *Proc. Iberoamerican Congress on Pattern Recognition*, 2006, pp. 634–643.

37. K. Torkkola, I. Guyon, and A. Elisseeff, "Feature Extraction by Non-Parametric Mutual Information Maximization," *Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1415–1438, 2003.

38. M. Thangavelu and R. Raich, "Multiclass linear dimension reduction via a generalized Chernoff bound," in *proc. IEEE Workshop on Machine Learning for Signal Processing*, Cancun, Mexico, 2008, pp. 350–355.

39. M. Thangavelu and R. Raich, " On linear dimension reduction for multiclass classification of Gaussian mixtures," in *proc. IEEE Workshop on Machine Learning for Signal Processing*, Grenoble, France, 2009, pp. 1–6.

40. K. Luebke and C. Weihs, "Improving feature extraction by replacing the Fisher criterion by an upper error bound," *Pattern recognition*, vol. 38, no. 11, pp. 2220, 2005.

41. G. Saon and M. Padmanabhan, "Minimum Bayes error feature selection for continuous speech recognition," *Advances in neural information processing systems*, pp. 800–806, 2001.

42. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 2 edition, 1990.

43. X. Wang and K.K. Paliwal, "Using minimum classification error training in dimensionality reduction," in *Neural networks in signal process proc IEEE*. Citeseer, 2000, vol. 1, pp. 338–345.

44. J. M. Leiva-Murillo and A. Artes-Rodriguez, "Linear dimensionality reduction with Gaussian mixture models," in *proc. IAPR Workshop on Cognitive Information Processing*, Santorini, Greece, 2008, pp. 48–53.

45. A. Tuerk, "Implicit softmax transforms for dimensionality reduction," in *Proc. ICASSP*, 2008, pp. 1973–1976.

46. M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal processing*, vol. 18, no. 4, pp. 349–369, 1989.

47. M. Loog, RPW Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 762–766, 2001.

48. B.C. Kuo and D.A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE transactions on geoscience and remote sensing*, vol. 42, no. 5, pp. 1096–1105, 2004.

49. G. Sfikas, C. Constantinopoulos, A. Likas, and N. P. Galatsanos, "An analytic distance metric for Gaussian mixture models with application in image retrieval," *Lecture notes in computer science*, vol. 3697, pp. 835, 2005.

50. F. D. Lorenzo-García, A. G. Ravelo-García, J. L. Navarro-Mesa, S. I. Martín-González, P. J. Quintana-Morales, and E. Hernández-Pérez, "A Chernoff-based approach to the Estimation of Transformation Matrices for Binary Hypothesis Testing," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Toulouse, France, 2006, pp. 753–756.

51. D.W. Scott, *Multivariate density estimation: theory, practice, and visualization*, Wiley-Interscience, 1992.

52. S.C. Douglas, S. Amari, and SY Kung, "On gradient adaptation with unit-norm constraints," *IEEE Transactions on Signal Processing*, vol. 48, no. 6, pp. 1843–1847, 2000.

53. G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

54. MC Jones, JS Marron, and SJ Sheather, "A brief survey of bandwidth selection for density estimation.," *Journal of the American Statistical Association*, vol. 91, no. 433, 1996.

55. T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

APPENDIX

1 Appendix 1

Let $\|g\|_p$ be the L_p norm of a finite-dimensional vector g given by $\|g\|_p = \sum(|g_i|^p)^{\frac{1}{p}}$ for $p \geq 1$. Since $\|g\|_p$ is monotonically decreasing with p,

$$\|g\|_1 \;\; \geq \;\; \|g\|_p. \tag{.1}$$

By the L_p norm definition,

$$\sum |g_i| \;\; \geq \;\; \left(\sum |g_i|^p\right)^{\frac{1}{p}}. \tag{.2}$$

Replacing $|x_i| = |g_i|^p$ in (.2), yields

$$\left(\sum |x_i|^{\frac{1}{p}}\right) \geq \left(\sum |x_i|\right)^{\frac{1}{p}}. \tag{.3}$$

Finally, replacing $t = \frac{1}{p}$ (such that $0 < \mathrm{t} < 1$) into (.3), we obtain

$$\sum |x_i|^t \geq \left(\sum |x_i|\right)^t. \tag{.4}$$

2 Appendix 2

A multiclass data consisting of L classes is assumed to be drawn from distributions $p_1, p_2, \cdots p_L$ with prior probabilities $\pi_1, \pi_2, \cdots \pi_L$. The probability of error considering this multiclass case is represented as $P_e(p_1, p_2, \cdots, p_L, \pi_1, \pi_2, \cdots, \pi_L)$.

The probability of error constructed based on the bayes classifier rule for the multiclass setup as described in. 3.5.

$$P_e^* = \sum_{i=1}^{n} \int I\Big(\bigcup_{i \neq j} \big(\frac{\pi_j p_j(x)}{\pi_i p_i(x)} > 1\big)\Big) \pi_i p_i(x) dx. \tag{.5}$$

Applying the union bound $I(\cup_i A_i) \leq \sum_i I(A_i)$ to the above expression we obtain

$$\begin{aligned} P_e^* &\leq \sum_{i=1}^{n} \sum_{j \neq i} \int I\left(\frac{\pi_j p_j(x)}{\pi_i p_i(x)} > 1\right) \pi_i p_i(x) dx. & (.6) \\ &= \sum_{i=1}^{n} \sum_{j \neq i} (\pi_i + \pi_j) \int I\left(\frac{\pi_j p_j(x)}{\pi_i p_i(x)} > 1\right) \frac{\pi_i}{\pi_i + \pi_j} p_i(x) dx. & (.7) \end{aligned}$$

Let us introduce the terms $\tilde{\pi_i}, \tilde{\pi_j}$ as, $\tilde{\pi_i} = \frac{\pi_i}{\pi_i + \pi_j}$s and $\tilde{\pi_j} = \frac{\pi_j}{\pi_i + \pi_j}$. This the expression can be written as,

$$\begin{aligned} &= \sum_{i=1}^{n} \sum_{j \neq i} (\pi_i + \pi_j) \int I\left(\frac{\tilde{\pi_j} p_j(x)}{\tilde{\pi_i} p_i(x)} > 1\right) \tilde{\pi_i} p_i(x) dx. & (.8) \\ &= \sum_{i=1}^{n} \sum_{j > i} (\pi_i + \pi_j) \int I\left(\frac{\tilde{\pi_j} p_j(x)}{\tilde{\pi_i} p_i(x)} > 1\right) \tilde{\pi_i} p_i(x) dx. & (.9) \\ & & (.10) \end{aligned}$$

We know that

$$\int I\left(\frac{\pi_j p_j(x)}{\pi_i p_i(x)} > 1\right) \pi_i p_i(x) dx. \tag{.11}$$

$$\tag{.12}$$

describes the probability of error of a two class classification problem between the distributions i and j belonging to the distribution p_i and p_j having the prior probabilities π_i and π_j. Hence,

$$P_e(p_1, p_2, \cdots, p_L, \pi_1, \pi_2, \cdots, \pi_L) \leq \sum_{i=1}^{n} \sum_{j>i} (\pi_i + \pi_j) P_e(p_i, p_j, \tilde{\pi_i}, \tilde{\pi_j}). \tag{.13}$$

By the application of the union bound, a multiclass problem has been bounded by the sum of two class error probabilities (.13).