INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Algoritmos De Análise Discriminativa Linear

Pedro Miguel Correia Guerreiro

Dissertação para obtenção do Grau de Mestre em

Engenharia Electrotécnica e de Computadores

Júri

Presidente:	Prof. Carlos Jorge Ferreira Silvestre
Orientadores:	Prof. João Manuel de Freitas Xavier
	Prof. Pedro Manuel Quintas Aguiar
Vogais:	Prof. José Manuel Bioucas Dias

Setembro 2008

INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Linear Discriminant Analysis Algorithms

Pedro Miguel Correia Guerreiro

A Dissertation submitted in fulfillment of the requirements for the degree of Master of Science in:

Electrical and Computer Engineering

September 2008

Agradecimentos

Agradeço aos meus pais por todo o apoio que me deram ao longo destes anos e do seu constante incentivo a fazer sempre melhor. Agradeço ao Prof. João Xavier e Prof. Pedro Aguiar pela ajuda absolutamente essencial e sem a qual não teria sido possível a escrita desta tese. Agradeço ainda à instituição Instituto Superior Técnico e a todo o seu corpo docente pela formação que me facultaram.

Resumo

Propõem-se novos algoritmos para o cálculo de discriminantes lineares usados na redução de dimensão de dados de $\mathbb{R}^n$ para $\mathbb{R}^p$, com $p < n$. São apresentadas alternativas ao critério clássico da Distância de Fisher, nomeadamente, investigam-se novos critérios baseados em: Distância de Chernoff, J-Divergência e Divergência de Kullback-Leibler. Os problemas de optimização que emergem do uso destas alternativas são não convexos e consequentemente difíceis de resolver. No entanto, apesar da não convexidade, os algoritmos desenvolvidos garantem que o discriminante linear é globalmente optimo para $p = 1$. Tal foi possível devido a reformulações do problema e a recentes resultados na teoria da optimização [8],[9]. Uma abordagem suboptima é desenvolvida para $1 < p < n$.

Palavras-Chave: Discriminantes Lineares, Redução de Dimensão de Dados, Distância de Fisher, Distância de Chernoff, resultados não convexos de dualidade forte, Divergência de Kullback-Leibler.

Abstract

We propose new algorithms for computing linear discriminants to perform data dimensionality reduction from $\mathbb{R}^n$ to $\mathbb{R}^p$ with $p < n$. We propose alternatives to the classical Fisher's Distance criterion, namely, we investigate new criterions based on the: Chernoff-Distance, J-Divergence and Kullback-Leibler Divergence. The optimization problems that emerge of using these alternative criteria are non-convex and thus hard to solve. However, despite the non-convexity, our algorithms guarantee global optimality for the linear discriminant when $p = 1$. This is possible due to problem reformulations and recent developments in optimization theory [8],[9]. A greedy suboptimal approach is developed for $1 < p < n$.

Keywords: Linear Discriminants, Data Dimensionality Reduction, Fisher's Distance, Chernoff-Distance, Nonconvex strong duality results, Kullback-Leibler Divergence.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Linear Discriminant Analysis (LDA) is a very important tool in a wide variety of problems. It is commonly used in machine learning problems like, pattern recognition [1],[2], face recognition [4], feature extraction [3] and in data dimensionality reduction.

A problem that is treated in LDA is the binary class assigning problem: given one sample in a high-dimensional space $\mathbb{R}^n$, say $x \epsilon \mathbb{R}^n$, decide to which class C_0 or C_1 it belongs to. Usually the two classes C_0 and C_1 represent two random sources. The classification process can be made in high dimension, i.e. in $\mathbb{R}^n$, using therefore all information available. However this might be computationally heavy for for certain real time applications. So, instead of using all the n entries of the sample x directly, an appropriate linear combination of them is made. With this linear combination, we try to capture some data features (hopefully those where C_0 and C_1 differ most), and then perform the data classification. Making these linear combinations, lead generically to information loss, and consequently increases the probability of erroneous classifications. However, this problem can be attenuated, by making more than one linear combination, and collect them in a vector y, to perform the classification. The number of linear combinations is denoted by p. That is

$$y = Qx \tag{1.1}$$

where $Q \epsilon \mathbb{R}^{p \times n}$ is called the linear discriminant, $y \epsilon \mathbb{R}^p$ is the vector that collects the p linear combinations, and $x \epsilon \mathbb{R}^n$ is the sample to be classified. The classification process is made trough the low-dimensional vector $y \epsilon \mathbb{R}^p$, which works like a signature of the sample x.

The key issue here is the design of the linear discriminant Q. This design process is generically formulated as an optimization problem, where the objective function measures class separability in the projected space $\mathbb{R}^p$, i.e.

$$\max_{Q \epsilon \mathbb{R}^{p \times n}} \quad f(Q). \tag{1.2}$$

The choice of the cost function in (1.2) plays a critical role. An obvious proposal for such cost function, would be $f(Q) = -P_(e)(Q)$, where $P_(e)(Q)$ stands for the probability of error of the optimum detector in $\mathbb{R}^p$, for the given setup, the minus sign has to do with the fact, that the optimization problem in (1.2), has been written as a maximization problem. However, in general there is no closed form expression for $P_(e)(Q)$. This motivates the introduction of alternative suboptimum choices, which are nonetheless tractable.

1.2 Previous Work

We now give a precise formulation of the problem to be solved and review previous works in this area.

Problem Statement. In what follows, the two classes C_0 and C_1 introduced in section 1.1 are identified with two random sources, that are here denoted by S_0 for source 0, and by S_1 for source 1. We focus on the Gaussian case.

Given the two independent n-dimensional Gaussian sources

$$\begin{aligned} S_0 &: x \sim F_0 = N(\mu_0, \Sigma_0) \\ S_1 &: x \sim F_1 = N(\mu_1, \Sigma_1) \end{aligned} \tag{1.3}$$

we wish to find a linear discriminant Q, for data dimensionality reduction, minimizing erroneous classification of the samples generated by these sources in low dimension.

Being $x \,\epsilon\, \mathbb{R}^n$ a sample generated by one of the n-dimensional sources S_0 or S_1 that are considered to be equally probable, a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^p$ is made with the linear discriminant $Q \,\epsilon\, \mathbb{R}^{p\times n}$, i.e.

$$y = Qx.$$

Due to this linear mapping, we have

$$\begin{aligned} s_0 &: y \sim f_0 = N(Q\mu_0, Q\Sigma_0 Q^T) \\ s_1 &: y \sim f_1 = N(Q\mu_1, Q\Sigma_1 Q^T) \end{aligned} \tag{1.4}$$

where s_0 and s_1 denote the p-dimensional sources that result from the dimensionality reduction induced by the linear discriminant $Q \,\epsilon\, \mathbb{R}^{p\times n}$

Whenever a sample x is available, it has to be classified. The classification is made with the maximum likelihood criteria, that is more well known in this context as the Neyman-Pearson detector. The linear map Q is applied to the sample x, forming $y = Qx$, and then the maximum likelihood criteriion is applied to the random variable y. If $N(Q\mu_0, Q\Sigma_0 Q^T)(y) > N(Q\mu_1, Q\Sigma_1 Q^T)(y)$, y is considered to have been generated by the p-dimensional source s_0 and x is therefore considered to have been generated by the n-dimensional source S_0 and vice-versa.

Previous Work. In the following we discuss several proposals for the cost function $f(Q)$ in (1.2), and we analyze the strengths and weaknesses of previous works that utilize such cost functions.

Fisher's Distance Maximization Criterion. A popular choice is the Fisher's Distance Maximization Criterion, which is now reviewed.

We wish to optimally separate in Fisher's sense, the signatures y from s_0, from the signatures from s_1. Intuitively this is equivalent, to separate as much as possible, the respective probability density functions f_0 and f_1 defined in (1.4).

The general optimization problem in (1.2) under Fisher's Distance Maximization criterion (see [6]) is

$$\max_{Q \,\epsilon\, \mathbb{R}^{p\times n}} \quad \mathrm{tr}\{(Q(\Sigma_0 + \Sigma_1)Q^T)^{-1}(Q(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T Q^T)\} \tag{1.5}$$

where the objective function is the Fisher's Distance between the low dimensional distributions f_0 and f_1.

In order to better understand what Fisher's Distance measures, the case where $Q \epsilon \mathbb{R}^{1\times n}$ is presented. Putting $Q = [q^T]$, where $q \epsilon \mathbb{R}^n$, (1.5) boils down to:

$$\max_{q \epsilon \mathbb{R}^n} \quad \frac{q^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T q}{q^T(\Sigma_0 + \Sigma_1)q} \tag{1.6}$$

Now, it's easy to understand that, the outer class variance $q^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T q = [q^T\mu_0 - q^T\mu_1]^2$ is being maximized while the total inner class variance $q^T(\Sigma_0 + \Sigma_1)q$, is being minimized.

The solution Q for (1.5), can be obtained by doing the eigenvalue decomposition of $(\Sigma_0+\Sigma_1)^{-\frac{1}{2}}(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T(\Sigma_0 + \Sigma_1)^{-\frac{1}{2}}$, and taking for the p rows of Q, the p eigenvectors associated to the p largest eingenvalues, see [6]. However, since $(\Sigma_0 + \Sigma_1)^{-\frac{1}{2}}(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T(\Sigma_0 + \Sigma_1)^{-\frac{1}{2}}$ has rank 1, it is easy to see that the optimum discriminant for $p > 1$ achieves the same performance, as measured by (1.5), as the optimum discriminant for $p = 1$. That is, there is no gain in projecting to spaces whose dimension $p > 1$. For $p = 1$, the optimum descriminant is $Q = [q^T]$ where q is a solution of (1.6), that is

$$q = (\Sigma_0 + \Sigma_1)^{-1}(\mu_0 - \mu_1). \tag{1.7}$$

In sum, Fisher's Distance Maximization criterion enjoys a closed form solution and a very intuitive interpretation. However, it only allows dimensionality reduction to $p = 1$.

Other Criteria. It was said previously that in general there is no closed form expression for the classification error rate. This leads to the utilization of suboptimal measures for it. The theoretical basis for the cost functions or measures used in [6] and [7] is now presented.

Stein's Lemma. [10] Suppose we have k statistically independent samples from the same source, and the classification is made trough the maximum-likelihood detector, then we have

$$\lim_{k\to+\infty} \frac{\log P_F(k)}{k} = -D_{KL}(f_0||f_1) \text{ for fixed } P_M \tag{1.8}$$

$$\lim_{k\to+\infty} \frac{\log P_e(k)}{k} = -C(f_0, f_1) \tag{1.9}$$

$$\lim_{k\to+\infty} \frac{\log P_e(k)}{k} \geq -JD(f_0, f_1) \tag{1.10}$$

where $P_F(k)$ is the probability of false alarm, $P_e(k)$ is classification error probability and P_M is the missing probability, when k samples from the same source are used to make the classification. Note that f_0 and f_1 are the p-dimensional probability density functions, resultant from the dimensionality reduction induced by the linear discriminant Q.

These probabilities are well known from the hypothesis tests. The probability of false alarm $P_F(k)$, is the probability of detecting s_0 when y was generated by s_1. The missing probability P_M is the probability of detecting s_1 when y was generated by s_0. $P_e(k)$ is simply the probability of wrong classification of the sample. Note that were used the p-dimensional sources s_0 and s_0, whit which the classification process is performed.

The exponents $D_{KL}(f_0||f_1)$, $JD(f_0, f_1)$, $C(f_0, f_1)$ in (1.8), are the Kullback-Leibler Divergence, the

J-Divergence and the Chernoff Distance, whose definitions for generic p-dimensional probability density functions f_0, f_1 are

$$D_{KL}(f_0||f_1) = \int_{\mathbb{R}^p} f_0(y) \log \frac{f_0(y)}{f_1(y)} dy \tag{1.11}$$

$$JD(f_0, f_1) = \frac{D_{KL}(f_0||f_1) + D_{KL}(f_1||f_0)}{2} \tag{1.12}$$

$$C(f_0||f_1) = \max_{0 \leq t \leq 1} -\log\left(\int_{\mathbb{R}^p} f_0(y)^t f_1(y)^{1-t} dy\right) \tag{1.13}$$

respectively.

Particularizing these expressions for the p-dimensional Gaussian probability density functions, resultant from the dimensionality reduction performed by the linear discriminant Q, $f_0(Q) = N(Q\mu_0, Q\Sigma_0 Q^T)$ and $f_1(Q) = N(Q\mu_1, Q\Sigma_1 Q^T)$, we have

$$D_{KL}(f_0||f_1)(Q) = \frac{1}{2}\left(\log \frac{|Q\Sigma_1 Q^T|}{|Q\Sigma_0 Q^T|} + tr((Q\Sigma_1 Q^T)^{-1}(Q\Sigma_0 Q^T)) + (\mu_0 - \mu_1)^T Q^T (Q\Sigma_1 Q^T)^{-1} Q(\mu_0 - \mu_1) - p\right) \tag{1.14}$$

$$\begin{aligned} JD(f_0, f_1)(Q) = \; & \frac{1}{4}\left(tr((Q\Sigma_1 Q^T)^{-1}(Q\Sigma_0 Q^T) + (Q\Sigma_0 Q^T)^{-1}(Q\Sigma_1 Q^T))\right) \\ & + \frac{1}{4}\left((\mu_0 - \mu_1)^T Q^T [(Q\Sigma_0 Q^T)^{-1} + (Q\Sigma_1 Q^T)^{-1}] Q(\mu_0 - \mu_1) - 2p\right) \end{aligned} \tag{1.15}$$

$$\begin{aligned} C(f_0, f_1)(Q) = \max_{0 \leq t \leq 1} \; & \frac{t(1-t)}{2}\left((\mu_0 - \mu_1)^T Q^T [tQ\Sigma_0 Q^T) + (1-t)Q\Sigma_1 Q^T]^{-1} Q(\mu_0 - \mu_1)\right) \\ & + \frac{1}{2}\left(\log \frac{|tQ\Sigma_0 Q^T + (1-t)Q\Sigma_1 Q^T|}{|Q\Sigma_0 Q^T|^t |Q\Sigma_1 Q^T|^{1-t}}\right) \end{aligned} \tag{1.16}$$

For $p = 1$, and attending to $Q = [q^T]$, the expressions simplify further an become

$$D_{KL}(f_0||f_1)(q) = \frac{1}{2}\left(\frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} - \log \frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} + \frac{[q^T(\mu_0 - \mu_1)]^2}{q^T\Sigma_1 q} - 1\right) \tag{1.17}$$

$$JD(f_0, f_1)(q) = \frac{1}{4}\left[\frac{(q^T\Sigma_0 q + q^T\Sigma_1 q)^2}{q^T\Sigma_0 q q^T\Sigma_1 q} + \frac{q^T\Sigma_0 q + q^T\Sigma_1 q}{q^T\Sigma_0 q q^T\Sigma_1 q}(q^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T q) - 4\right] \tag{1.18}$$

$$C(f_0, f_1)(q) = \max_{0 \leq t \leq 1} \frac{1}{2}\left(t(1-t)\frac{q^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T q}{tq^T\Sigma_0 q + (1-t)q^T\Sigma_1 q} + \log \frac{tq^T\Sigma_0 q + (1-t)q^T\Sigma_1 q}{(q^T\Sigma_0 q)^t (q^T\Sigma_1 q)^{1-t}}\right) \tag{1.19}$$

Stein's Lemma gives asymptotic expressions for $P_F(k)$ and $P_e(k)$. The heuristic behind Stein's Lemma utilization is that, even tough, be necessary a large number k of samples for the asymptotic expressions give a good approximation for $P_F(k)$ and $P_e(k)$, it is expected that they behave well when k is small or even equal to one. Stein's Lemma fills heuristically the lack of closed form expressions for the probabilities.

The Kullback-Leibler divergence has a very easy interpretation.

$$D_{KL}(f_0||f_1) = \int_{\mathbb{R}^p} f_0(y)\, log\, \frac{f_0(y)}{f_1(y)} dy$$

This can be interpreted as maximizing the expected value of the log-likelihood ratio $log\, \frac{f_0(y)}{f_1(y)}$, under f_0, which is equivalent to maximize the number of correct detections of samples from s_0, when such samples were generated by s_0. Note that this criteria is asymmetric.

The J-Divergence is simply the symmetrization of Kullback-Leibler Divergence. With this symmetrization, the two sources are treated equally.
From (1.10)

$$JD(f_0, f_1) = \frac{D_{KL}(f_0||f_1) + D_{KL}(f_1||f_0)}{2}$$

it can be seen that by maximizing J-Divergence, an asymptotic lower bound for the classification error rate is minimized, it is expected that by minimizing this lower bound, the classification error rate is also minimized.

The Chernoff distance is related with the geodesic distance between the probability density distributions f_0 and f_1 in the probability manifold (see [6]).

There is a characteristic that the J-Divergence, the Chernoff Distance and Kullback-Leibler Divergence share. That characteristic will reveal to be the main advantage of this methods when compared to Fisher's Distance. That characteristic is the capability of these criteria for discriminate the probability density distributions f_0 and f_1 by their variance. Looking at the expressions of these criteria in (1.17) to 1.19, it is easy to see, that when maximized in q, very different variances for f_0 and f_1 will emerge. Looking to the Kullback-Leibler Divergence expression it can be seen that the parcel

$$\frac{q^T \Sigma_0 q}{q^T \Sigma_1 q} = \frac{\sigma_0^2}{\sigma_1^2} \tag{1.20}$$

will contribute for this phenomenon. For the J-Divergence and Chernoff Distance this is done by

$$\begin{matrix} \frac{(q^T\Sigma_0 q + q^T \Sigma_1 q)^2}{q^T \Sigma_0 q q^T \Sigma_1 q} \\ \log \frac{t q^T \Sigma_0 q + (1-t) q^T \Sigma_1 q}{(q^T \Sigma_0 q)^t (q^T \Sigma_1 q)^{1-t}} \end{matrix} \tag{1.21}$$

respectively, that can be interpreted as the arithmetic mean over the geometric mean of ($q^T \Sigma_0 q$,$q^T \Sigma_1 q$), where in the second are weighted by t. Attending to the fact that this quotient (arithmetic mean over the geometric mean) has a minimum when the quantities involved are equal, it's explained why, when these parcels are maximized the variances will be different.

Figures 1.1 and 1.2 show what happened using Fisher's Distance Maximization criterion and the J-Divergence Maximization criterion Figures 1.1 and 1.2 prove unequivocally this capability.

Looking at Stein's Lemma statement in (1.8), it can be seen that in order to minimize $P_F(k)$ and $P_e(k)$, $D_{KL}(f_0||f_1)$, $JD(f_0, f_1)$, $C(f_0, f_1)$ must be maximized. This is precisely what is done in [7] where the J-Divergence is maximized and [6] where the same happens with Chernoff-Distance.

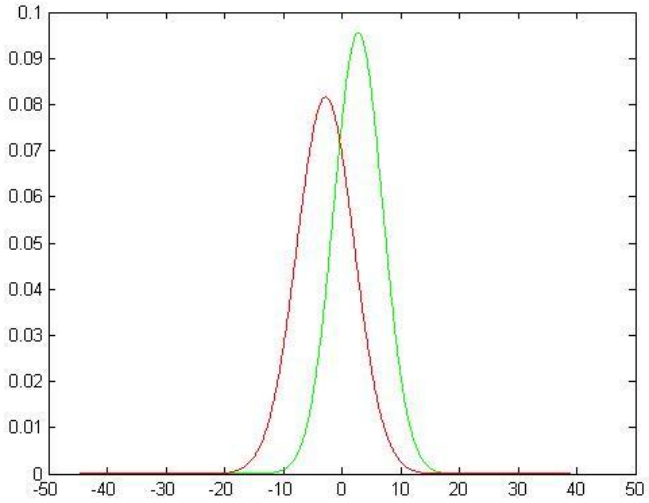

Figure 1.1: 1-dimensional pdf's f_0 and f_1 obtained through Fisher's Distance.

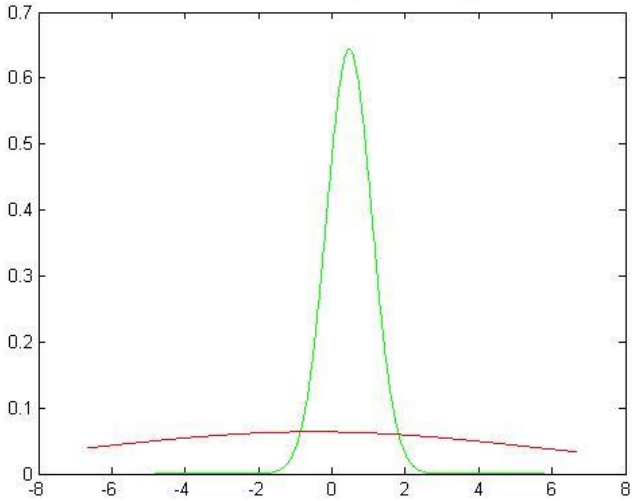

Figure 1.2: 1-dimensional pdf's f_0 and f_1 obtained through J-Divergence.

1.3 Contribution

In this thesis is treated the class or source assigning problem for the case where the sources S_0 and S_1 or classes C_0 and C_1 are Gaussian distributed. The criteria or cost functions utilized are the Chernoff Distance, the Kullback-Leibler Divergence and J-Divergence. The major problem with the chose of these criteria is that the respective optimization problems are very hard to solve.

The works already developed, namely the ones presented in [7] and [6] don't solve the problems with full generality or utilize methods that don't guarantee global optimality. The work developed in this thesis gives the next step by solving the optimization problems resultant from the utilization of Chernoff Distance, Kullback-Leibler Divergence and J-Divergence with full generality and guaranteing global

optimality. Global optimality is just guaranteed when reducing to one dimension i.e: $p = 1$.

1.4 Thesis Outline

Chapter 2. In this chapter are presented the algorithms that compute the linear discriminants, that maximize the Kullback-Leibler Divergence, J-Divergence and Chernoff-Distance when projecting the n-dimensional samples to $\mathbb{R}$.

Chapter 3. In this chapter is presented the suboptimal approach for computing the linear discriminants when projecting the n-dimensional samples to $\mathbb{R}^p$.

Chapter 4. This chapter presents the results of the performances, i.e. correct classifications of the n-dimensional samples for the several criteria used in the linear discriminants computation.

Chapter 2

Algorithms for Dimensionality Reduction to $\mathbb{R}$

In this chapter are presented the algorithms to compute the linear discriminant $Q = [q^T]$ that performs dimensionality reduction to $\mathbb{R}$ by maximizing the several criteria presented in chapter 1, i.e: Kullback-Leibler Divergence ($D_{KL}(f_0||f_1)(Q)$), J-Divergence ($JD(f_0, f_1)(Q)$) and Chernoff Distance ($C(f_0, f_1)(Q)$).

As mentioned in chapter 1, the probability density functions f_0, f_1 present in these expressions, are those that characterize the output $y = Qx$ of the 1-dimensional sources s_0, s_1 resultant from the dimensionality reduction process.

2.1 Kullback-Leibler Divergence Maximization

In chapter 1, the expression for the Kullback-Leibler Divergence between the 1-dimensional probability density functions f_0, f_1 was found to be (see equation (1.17))

$$D_{KL}(f_0||f_1)(q) = \frac{1}{2}\left(\frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} - \log\frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} + \frac{[q^T(\mu_0-\mu_1)]^2}{q^T\Sigma_1 q} - 1\right) \tag{2.1}$$

The goal is to find the global maximizer q of (2.1), i.e.

$$q = \underset{q\neq 0}{\arg\max}\ \frac{1}{2}\left(\frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} - \log\frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} + \frac{[q^T(\mu_0-\mu_1)]^2}{q^T\Sigma_1 q} - 1\right). \tag{2.2}$$

It's easy to verify that (2.1) doesn't depend on the norm of q. So, a restriction that doesn't eliminate any direction for q is admissible.

In order to simplify the objective function of the optimization problem in (2.2) and without eliminating any direction for q, the restriction $q^T\Sigma_1 q = 1$ is chosen. Applying the restriction, the optimization problem

in (2.2) becomes

$$q = \begin{array}{l}\arg\max \\ q^T\Sigma_1 q = 1\end{array} \quad q^T\Sigma_0 q - \log q^T\Sigma_0 q + [q^T(\mu_0 - \mu_1)]^2 \tag{2.3}$$

$$= \begin{array}{l}\arg\max \\ q^T\Sigma_1 q = 1\end{array} \quad q^T\Sigma_0 q - \log q^T\Sigma_0 q + q^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T q \tag{2.4}$$

$$= \begin{array}{l}\arg\max \\ q^T\Sigma_1 q = 1\end{array} \quad q^T[\Sigma_0 + (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T]q - \log q^T\Sigma_0 q \tag{2.5}$$

In what follows, $\Sigma_0 + (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$ is substituted by R, resulting:

$$q = \begin{array}{l}\arg\max \\ q^T\Sigma_1 q = 1\end{array} \quad q^T R q - \log q^T\Sigma_0 q. \tag{2.6}$$

Problem Reformulation. The optimization problem in (2.6) is non-convex, so a reformulation is made by introducing the variables x and y

$$x = q^T R q \tag{2.7}$$

$$y = q^T\Sigma_0 q \tag{2.8}$$

resulting for (2.6) in

$$\begin{array}{l}max \\ (x,y)\,\epsilon\, C = \{(q^T R q, q^T\Sigma_0 q) :\ q^T\Sigma_1 q = 1\}\end{array} \quad x - \log y. \tag{2.9}$$

Reformulating the optimization problem this way, the optimization is just made in two variables $(x, y)\,\epsilon\, C$. However, the complexity of the original problem is hidden in the definition of the set C. The strategy to solve (2.6) consists in finding the solution (x^*, y^*) for (2.9), and then computing a corresponding q, i.e, a q that solves the following system of quadratic equations:

$$q: \begin{array}{l} q^T R q = x^* \\ q^T\Sigma_0 q = y^* \\ q^T\Sigma_1 q = 1 \end{array} . \tag{2.10}$$

The set C is compact and connected. It results from a continuous quadratic mapping of an ellipsoid, implying that the variables x and y considered separately, belong to closed intervals on $\mathbb{R}$. It is needed to compute the closed interval on $\mathbb{R}$ for the x variable, since it is a connected set, it is just needed to calculate the ends of the interval, i.e:

$$x_{\min} = \begin{array}{l}\min \\ q^T\Sigma_1 q = 1\end{array} \quad q^T R q \tag{2.11}$$

$$x_{\max} = \begin{array}{l}\max \\ q^T\Sigma_1 q = 1\end{array} \quad q^T R q \tag{2.12}$$

The solutions to (2.52) and (2.53) are $x_{\min} = \lambda_{\min}(\Sigma_1^{-\frac{1}{2}} R \Sigma_1^{-\frac{1}{2}})$ and $x_{\max} = \lambda_{\max}(\Sigma_1^{-\frac{1}{2}} R \Sigma_1^{-\frac{1}{2}})$ respectively, and thus $x\,\epsilon\,[x_{\min}, x_{\max}]$.

Knowing this, the strategy to solve (2.9) consists in discretizing the above interval fixing a value for x,

and optimizing over the y variable. Given the objective function in (2.9), this corresponds to minimize y. This procedure has to be done for all points x of the discretization of $[x_{\min}, x_{\max}]$. Once this procedure is finished, the best pair (x^*, y^*) is chosen and the corresponding q defined in (2.10), is the one that solves (2.6).

Fixing a value for $x \,\epsilon\, [x_{\min}, x_{\max}]$ and attending to (2.7), the problem related with the y variable optimization is

$$\begin{array}{cc} \min & q^T \Sigma_0 q \\ q^T R q = x & \\ q^T \Sigma_1 q = 1 & \end{array} \tag{2.13}$$

This problem is non-convex and will be solved trough duality theory presented in appendix (A.1).

In the process of finding the pair (x^*, y^*) that solves the optimization problem in (2.9), for a fixed value of x, it is just needed to know the value of the best attainable value of y (calculated as in (2.13)). For $x \epsilon] x_{\min}, x_{\max} [$ strong duality exists for (2.13), the values of y variable are calculated trough the dual problem that is:

$$\begin{array}{cc} \max & -\lambda_1 x - \lambda_2 1 \\ \Sigma_0 + \lambda_1 R + \lambda_2 \Sigma_1 \geq 0 & \\ \text{var} : (\lambda_1, \lambda_2) \,\epsilon\, R^2 & \end{array} \tag{2.14}$$

As explained in appendix (A.1), the dual problem is an optimization problem in just two variables. Once the pair (x^*, y^*) is computed, it is needed to compute the optimal q^* that solves 2.6. For this process the bi-dual problem of (2.13) is used. From this process we know that the set of optimal points that contain the solution q, is such that $q^T R q = x^*$, so this restriction is represented in the Bi-Dual Problem (2.15) through $tr(RQ) = x^*$

$$\begin{array}{cc} \min & tr(\Sigma_0 Q) \\ tr(RQ) = x^* & \\ tr(\Sigma_1 Q) = 1 & \\ Q \succeq 0 & \end{array} \tag{2.15}$$

Provided Slater conditions are verified i.e: $x \epsilon] x_{\min}, x_{\max} [$ and the uniqueness of the solution for the bi-dual problem, Q is a rank-1 semidefinite positive matrix and its only eigenvector is the solution for the problem i.e., it is the linear discriminant that optimizes the Kullback-Leibler Divergence criterion.

2.2 J-Divergence Maximization

In chapter 1, the expression for the J-Divergence between the 1-dimensional probability density functions f_0, f_1 was found to be (see equation (1.18))

$$JD(f_0||f_1)(q) = \frac{1}{4}\left[\frac{(q^T\Sigma_0 q + q^T\Sigma_1 q)^2}{q^T\Sigma_0 q q^T\Sigma_1 q} + \frac{q^T\Sigma_0 q + q^T\Sigma_1 q}{q^T\Sigma_0 q q^T\Sigma_1 q}(q^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T q) - 4\right]. \tag{2.16}$$

The goal is to find the global maximizer q of (2.16), i.e.

$$q = \underset{q \neq 0}{\arg\max} \quad \frac{1}{4}\left[\frac{(q^T\Sigma_0 q + q^T\Sigma_1 q)^2}{q^T\Sigma_0 q q^T\Sigma_1 q} + \frac{q^T\Sigma_0 q + q^T\Sigma_1 q}{q^T\Sigma_0 q q^T\Sigma_1 q}(q^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T q) - 4\right] \tag{2.17}$$

The expression for the J-Divergence in (2.16) doesn't depend on the norm of q, but just on its direction. Taking advantage of this property the following restriction is added

$$q^T\Sigma_0 q q^T\Sigma_1 q = 1. \tag{2.18}$$

As in the previous optimization problem, this restriction is admissible in the sense that it doesn't eliminate any direction. Given any $q \,\epsilon\, \mathbb{R}^n$ it is possible to scale it without changing its direction till verifies (2.18). Applying this restriction and dropping the multiplicative and constant parts of the objective function, the optimization problem in (2.17) becomes

$$q = \underset{q^T\Sigma_0 q q^T\Sigma_1 q = 1}{\arg\max} \quad (q^T\Sigma_0 q + q^T\Sigma_1 q)^2 + (q^T\Sigma_0 q + q^T\Sigma_1 q)(q^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T q) \tag{2.19}$$

The restriction in (2.18) can be written as $q^T\Sigma_1 q = \frac{1}{q^T\Sigma_0 q}$, so $q^T\Sigma_1 q$ is substituted by $\frac{1}{q^T\Sigma_0 q}$ in (2.19) resulting in

$$q = \underset{q^T\Sigma_0 q q^T\Sigma_1 q = 1}{\arg\max} \quad (q^T\Sigma_0 q + \frac{1}{q^T\Sigma_0 q})^2 + (q^T\Sigma_0 q + \frac{1}{q^T\Sigma_0 q})(q^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T q) \tag{2.20}$$

Problem Reformulation. The optimization problem in (2.20) is non-convex, so, a reformulation of the same is made by introducing the variables x and y

$$x = q^T\Sigma_0 q \tag{2.21}$$
$$y = q^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T q \tag{2.22}$$

resulting for (2.20) in

$$\underset{(x,y)\,\epsilon\, C = \{(q^T\Sigma_0 q,\, q^T M q):\; q^T\Sigma_0 q q^T\Sigma_1 q = 1\}}{max} \quad \left(x + \tfrac{1}{x}\right)^2 + \left(x + \tfrac{1}{x}\right) y \tag{2.23}$$

where $M = (\mu_0-\mu_1)(\mu_0-\mu_1)^T$

The strategy to solve (2.20) as previously seen, consists in finding the solution (x^*, y^*) for (2.23) and

then computing the corresponding q, i.e:

$$q: \quad \begin{array}{l} q^T\Sigma_0 q = x^* \\ q^T M q = y^* \\ q^T\Sigma_0 q q^T\Sigma_1 q = 1 \end{array} \quad . \tag{2.24}$$

In appendix C is shown that set the C is compact and connected, implying that the variables x and y considered separately, belong to closed intervals on $\mathbb{R}$. It is needed to compute the closed interval on $\mathbb{R}$ for the x variable, since it is a connected set, it is just needed to calculate the ends of the interval i.e:

$$x_{\min} = \begin{array}{c} \min \\ q^T\Sigma_0 q q^T\Sigma_1 q = 1 \end{array} \quad q^T\Sigma_0 q \tag{2.25}$$

$$x_{\max} = \begin{array}{c} \max \\ q^T\Sigma_0 q q^T\Sigma_1 q = 1 \end{array} \quad q^T\Sigma_0 q \tag{2.26}$$

The solutions to 2.25 and 2.26 are

$$x_{\min} = \frac{1}{\sqrt{\lambda_{max}(\Sigma_0^{-\frac{1}{2}}\Sigma_1\Sigma_0^{-\frac{1}{2}})}} \tag{2.27}$$

$$x_{\max} = \frac{1}{\sqrt{\lambda_{min}(\Sigma_0^{-\frac{1}{2}}\Sigma_1\Sigma_0^{-\frac{1}{2}})}} \tag{2.28}$$

So, $x\,\epsilon\,[x_{\min}, x_{\max}]$. In section (2.2.1) is shown how these expressions for the extremal points of the interval in $\mathbb{R}$ for the x variable, were obtained.

Knowing this, the strategy to solve (2.23), consists in discretizing the above interval, fixing a value for x, and optimize over the y variable. The goal is to solve (2.23) and so, optimizing y consists in maximizing it. This procedure has to be done for all points x of the discretization of $[x_{\min}, x_{\max}]$. Once this procedure is finished, the best pair (x^*, y^*) is chosen, and the corresponding q according to (2.24), is the one that solves (2.20).

Fixing a value $x\,\epsilon\,[x_{\min}, x_{\max}]$ and attending to (2.21), the problem related with the y variable optimization is

$$\begin{array}{c} \max \quad q^T M q \\ q^T\Sigma_0 q = x \\ q^T\Sigma_1 q = \frac{1}{x} \end{array} \tag{2.29}$$

that written as a minimization problem becomes

$$\begin{array}{c} \min \quad q^T(-M) q \\ q^T\Sigma_0 q = x \\ q^T\Sigma_1 q = \frac{1}{x} \end{array} \tag{2.30}$$

This problem is solved trough duality theory presented in section A.1. From this point, everything follows the same procedure as for the Kullback-Leibler Divergence algorithm (see 2.1). It's important to note that strong duality for (2.30), only exist for $x\,\epsilon\,]x_{\min}, x_{\max}[$ce criteria.

2.2.1 Interval Computation

Here is shown how to compute the extremal points of the interval $[x_{\min}, x_{\max}] \subset \mathbb{R}$, where the x variable defined in (2.21) belongs.

The definition of this extremal points is given in (2.25) and (2.26) and is here again presented

$$x_{\min} = \min_{q^T\Sigma_0 q q^T\Sigma_1 q = 1} q^T\Sigma_0 q \tag{2.31}$$

$$x_{\max} = \max_{q^T\Sigma_0 q q^T\Sigma_1 q = 1} q^T\Sigma_0 q. \tag{2.32}$$

Introducing the variables a and b, defined as

$$a = q^T\Sigma_0 q \tag{2.33}$$

$$b = q^T\Sigma_1 q \tag{2.34}$$

the problems in (2.31) and (2.32) are equivalent to

$$x_{\min} = \min_{\substack{ab=1 \\ \text{var}: (a,b)\,\epsilon\, K = \{(q^T\Sigma_0 q, q^T\Sigma_1 q):\ q\,\epsilon\,\mathbb{R}^n\}}} a \tag{2.35}$$

$$x_{\max} = \max_{\substack{ab=1 \\ \text{var}: (a,b)\,\epsilon\, K = \{(q^T\Sigma_0 q, q^T\Sigma_1 q):\ q\,\epsilon\,\mathbb{R}^n\}}} a \tag{2.36}$$

The restrictions in the reformulated problems are

$$\begin{aligned} &ab = 1 \\ &(a,b)\,\epsilon\, K = \{(q^T\Sigma_0 q, q^T\Sigma_1 q):\ q\,\epsilon\, R^n\}. \end{aligned}$$

Due to the positive definiteness of the matrices Σ_0 and Σ_1 involved in the definition of the variables a and b, it's easy to see that the variables a and b belong to the first orthant.

Set K consists in a quadratic mapping from $\mathbb{R}^n$ to $\mathbb{R}^2$ and due to a theorem by Dines (see [9]), set K is a closed convex cone in the first orthant. With this graphical interpretation, it's easy to see that $x_{\min}$ and $x_{\max}$ are the points of intersection of the straight lines that delimitate set K with the graph of the hyperbola function.

In order to calculate these intersections, the mathematical expressions of the straight lines that delimitate set K are needed. Since these straight lines pass trough the origin, they are of the form $b = ma$.

From figure 2.1 can be seen, that in order to calculate $x_{\min}$, the slope of the upper straight line is needed, in order to do that a point (a, b) of this straight line must be computed. Attending to set K definition and fixing $a = 1$, b is equal to:

$$b = max_{q^T\Sigma_0 q = 1} q^T\Sigma_1 q = \lambda_{max}(\Sigma_0^{-\frac{1}{2}}\Sigma_1\Sigma_0^{-\frac{1}{2}}) \tag{2.37}$$

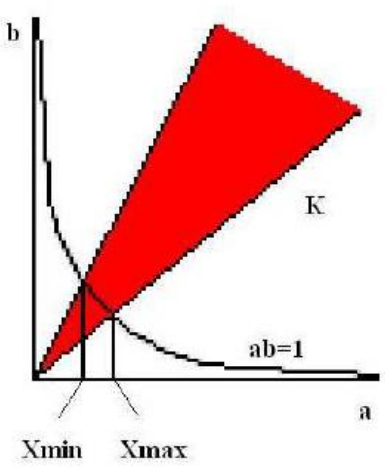

Figure 2.1: Geometrical interpretation of the set of restrictions.

The restriction $q^T \Sigma_0 q = 1$ in the above optimization problem needed to compute b, has naturally to do with $a = 1$. The choice of $a = 1$ was made in order to obtain directly the slope of the upper straight line.

With the slope of the upper straight line that delimitates set K, calculated it is needed to intersect this straight line with the graph of the hyperbola function, i.e, the following system of equations must be solved

$$ab = 1 \tag{2.38}$$

$$b = ma \tag{2.39}$$

where $m = \lambda_{max}(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}})$.

This results for a and b, in

$$a = \frac{1}{\sqrt{\lambda_{\max}(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}})}}$$
$$b = \sqrt{\lambda_{\max}(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}})}$$

where a is $x_{\min}$.

Following the same process to calculate $x_{\min}$, can be shown that $x_{\max} = \frac{1}{\sqrt{\lambda_{\min}(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}})}}$.

2.3 Chernoff Distance Maximization

In chapter 1 the expression for the Chernoff Distance between the 1-dimensional probability density functions f_0, f_1 was found to be (see equation (1.19))

$$C(f_0,f_1)(q) = \max_{0\leq t\leq 1} \frac{1}{2}\left(t(1-t)\frac{q^T(\mu_0-\mu_1)(\mu_0-\mu_1)^Tq}{tq^T\Sigma_0q+(1-t)q^T\Sigma_1q}+\log\frac{tq^T\Sigma_0q+(1-t)q^T\Sigma_1q}{(q^T\Sigma_0q)^t(q^T\Sigma_1q)^{1-t}}\right). \tag{2.40}$$

The goal is to find the global maximizer q of (2.40), i.e.

$$q = \arg\max_{q\neq 0} \max_{0\leq t\leq 1} \frac{1}{2}\left(t(1-t)\frac{q^T(\mu_0-\mu_1)(\mu_0-\mu_1)^Tq}{tq^T\Sigma_0q+(1-t)q^T\Sigma_1q}+\log\frac{tq^T\Sigma_0q+(1-t)q^T\Sigma_1q}{(q^T\Sigma_0q)^t(q^T\Sigma_1q)^{1-t}}\right). \tag{2.41}$$

The expression of the Chernoff-distance in (2.40), involves an intrinsic optimization in the variable t, so the optimization problem in (2.41) can be written expliciting that characteristic giving

$$(q,t) = \arg\max_{\substack{q\neq 0\\ 0\leq t\leq 1}} \frac{1}{2}\left(t(1-t)\frac{q^TMq}{tq^T\Sigma_0q+(1-t)q^T\Sigma_1q}+\log\frac{tq^T\Sigma_0q+(1-t)q^T\Sigma_1q}{(q^T\Sigma_0q)^t(q^T\Sigma_1q)^{1-t}}\right) \tag{2.42}$$

where $M=(\mu_0-\mu_1)(\mu_0-\mu_1)^T$.

The above optimization problem (2.42) is non-convex and it has to be optimized in two variables, $q\,\epsilon\,\mathbb{R}^n$ and $t\,\epsilon\,[0,1]$. Due to its non-convexity, it is useful to rewrite it in the following equivalent form

$$\max_{0\leq t\leq 1} \max_{q\neq 0} \frac{1}{2}\left(t(1-t)\frac{q^TMq}{tq^T\Sigma_0q+(1-t)q^T\Sigma_1q}+\log\frac{tq^T\Sigma_0q+(1-t)q^T\Sigma_1q}{(q^T\Sigma_0q)^t(q^T\Sigma_1q)^{1-t}}\right). \tag{2.43}$$

This equivalent optimization problem is solved by fixing t belonging to a discretization of $[0,1]$, and optimizing in the q variable. The main advantage of this written, is that the cost function in (2.43) is independent on the norm of q for a fixed t. Since this optimization problem is non-convex, it has necessarily to be solved by searching for every point t of the discretization of $[0,1]$ the best corresponding q. Once this procedure is finished, the best pair (q^*,t^*) is chosen.

Due to the independence on the norm of q for a fixed t, the optimization problem can be further simplified by introducing the following restriction

$$tq^T\Sigma_0q+(1-t)q^T\Sigma_1q=1 \tag{2.44}$$

resulting for (2.43) in

$$\max_{0\leq t\leq 1} \max_{tq^T\Sigma_0q+(1-t)q^T\Sigma_1q=1} \frac{1}{2}\left(t(1-t)q^TMq-log\,(q^T\Sigma_0q)^t(q^T\Sigma_1q)^{1-t}\right). \tag{2.45}$$

With this approach, for a fixed t we have to find the global maximizer q of the following subproblem.

$$\begin{array}{cc} \max & \frac{1}{2}\left(t(1-t)q^TMq - log\left(q^T\Sigma_0 q\right)^t\left(q^T\Sigma_1 q\right)^{1-t}\right) \\ tq^T\Sigma_0 q + (1-t)q^T\Sigma_1 q = 1 & \end{array} \tag{2.46}$$

It is expected that t^* belongs to the open interval $]0,1[$. Otherwise, the information about one of the covariance matrices Σ_0 or Σ_1 would be neglected. With this assumption, the endpoints of $[0,1]$ are not evaluated. This enables a rewritten of the restriction introduced at (2.44) that is the following

$$q^T\Sigma_1 q = \frac{1 - tq^T\Sigma_0 q}{1-t}.$$

Having this in consideration, and using the properties of the logarithm function, (2.46) becomes

$$\begin{array}{cc} \max & \frac{1}{2}\left(t(1-t)q^TMq - t\log q^T\Sigma_0 q - (1-t)\log\frac{1-tq^T\Sigma_0 q}{1-t}\right). \\ tq^T\Sigma_0 q + (1-t)q^T\Sigma_1 q = 1 & \end{array} \tag{2.47}$$

Subproblem reformulation. The optimization problem in (2.47) is non-convex, so as in the previous situations a reformulation of the same is made by introducing the variables x and y

$$x = q^T\Sigma_0 q \tag{2.48}$$
$$y = q^TMq \tag{2.49}$$

resulting for (2.47) in

$$\begin{array}{cc} \max & t(1-t)y - t\log x - (1-t)\log\left(\frac{1-tx}{1-t}\right). \\ (x,y)\,\epsilon\, C = \{(q^T\Sigma_0 q, q^TMq):\ tq^T\Sigma_0 q + (1-t)q^T\Sigma_1 q = 1\} & \end{array} \tag{2.50}$$

Again, the strategy to solve (2.47) consists in finding the solution (x^*, y^*) for (2.50) and then computing the corresponding q, i.e. a q that solves the following system of quadratic equations:

$$q: \begin{array}{l} q^T\Sigma_0 q = x^* \\ q^TMq = y^* \\ tq^T\Sigma_0 q + (1-t)q^T\Sigma_1 q = 1 \end{array}. \tag{2.51}$$

The set C is compact and connected. It results from a continuous quadratic mapping of an ellipsoid, implying that the variables x and y considered separately, belong to closed intervals on $\mathbb{R}$.

It is needed to compute the closed interval on $\mathbb{R}$ for the x variable, since it is a connected set, it is just needed to calculate the ends of the interval, i.e:

$$x_{\min} = \begin{array}{cc} \min & q^T\Sigma_0 q \\ tq^T\Sigma_0 q + (1-t)q^T\Sigma_1 q = 1 & \end{array} \tag{2.52}$$

$$x_{\max} = \begin{array}{cc} \max & q^T\Sigma_0 q \\ tq^T\Sigma_0 q + (1-t)q^T\Sigma_1 q = 1 & \end{array} \tag{2.53}$$

Knowing this, the strategy to solve (2.50) consists in discretizing the above interval fixing a value for x, and optimizing over the y variable. Given the objective function in (2.50), this corresponds to maximize y.

This procedure has to be done for all points x of the discretization of $[x_{\min}, x_{\max}]$. Once this procedure is finished, the best pair (x^*, y^*) is chosen and the corresponding q defined in (2.51), is the one that solves (2.47).

Fixing a value for $x \,\epsilon\, [x_{\min}, x_{\max}]$ and attending to (2.7), the problem related with the y variable optimization is

$$\begin{array}{cc} \min & q^T M q \\ q^T \Sigma_0 q = x & \\ t q^T \Sigma_0 q + (1-t) q^T \Sigma_1 q = 1 & \end{array} \tag{2.54}$$

This problem is solved trough duality theory presented in section A.1. From this point, everything follows the same procedure as for the Kullback-Leibler Divergence algorithm (see 2.1). It's important to note that strong duality for (2.54), only exist for $x \,\epsilon\,]x_{\min}, x_{\max}[$.

Observation. In all the three previous algorithms, it was said that strong duality holds for $x \,\epsilon\,]x_{\min}, x_{\max}[$ and consequently the extreme points were not evaluated. This doesn't represent a problem, since the objective functions in (2.9), (2.23) and (2.50) are continuous functions for the respective closed intervals as well as the optimized y variable as a function of the fixed x.

Chapter 3

Greedy Algorithms

The algorithms presented in chapter 2 perform a dimensionality reduction from n dimensions to one dimension. This is done by linear mapping the n-dimensional samples $x \epsilon \mathbb{R}^n$ through the linear discriminant q. This drastic dimensionality reduction may induce an acceptable loss of information making the projected distributions almost indistinguishable. The consequence is significant classification error rate. In order to make the process of dimensionality reduction less drastic and hopefully lowering the classification error rate, this chapter considers dimensionality reduction to $\mathbb{R}^p$ with $p > 1$. The algorithms to be developed here perform a dimensionality reduction from n dimensions to p dimensions, where $p > 1$ through the linear discriminant matrix $Q \epsilon R^{p \times n}$, i.e.

$$y = Qx \tag{3.1}$$

where x is the n-dimensional sample, Q is the linear discriminant matrix, and $y \epsilon \mathbb{R}^p$ is the signature of the sample x, used in the classification procedure.

In order to better understand the greedy version of the algorithms presented in chapter 2 it is presented the greedy algorithm that maximizes the Kullback-Leibler Divergence. For the other criteria the algorithms follow a similar pattern which will not be repeated here.

The optimal linear discriminant $Q \epsilon \mathbb{R}^{p \times n}$, that maximizes the Kullback-Leibler Divergence between the p-dimensional probability density functions $f_0(Q) = N(Q\mu_0, Q\Sigma_0 Q^T)$ and $f_1(Q) = N(Q\mu_1, Q\Sigma_1 Q^T)$, is found by solving the optimization problem

$$\max_{Q \epsilon \mathbb{R}^{p \times n}} \quad D_{KL}(f_0||f_1)(Q) \tag{3.2}$$

where

$$D_{KL}(f_0||f_1)(Q) = \frac{1}{2}\left(\log\frac{|Q\Sigma_1 Q^T|}{|Q\Sigma_0 Q^T|} + tr((Q\Sigma_1 Q^T)^{-1}(Q\Sigma_0 Q^T)) + (\mu_0 - \mu_1)^T Q^T (Q\Sigma_1 Q^T)^{-1} Q(\mu_0 - \mu_1) - p\right). \tag{3.3}$$

The main problem with this approach is the non-convexity of the objective function. Although, the case $p = 1$ could be treated trough a series of reformulations and simplifications which made possible finding the solution efficiently, we were not able to extend this procedure for $p > 1$. So a sub-optimal approach to solve (3.2) is taken. This approach consists in compute the p rows of $Q \epsilon \mathbb{R}^{p \times n}$ one by one, by solving p 1-dimensional optimization problems, like the one in (2.2) for the case of the Kullback-Leibler Divergence.

We start by noting that, without loss of optimality, the matrix Q in (3.2) can be taken to be Stiefel, i.e.,

with orthonormal rows. This is proved in appendix B. The fact that Q can be a Stiefel matrix motivates, the following procedure to compute its p rows.

Computation of the rows of Q.

$$Q = \begin{bmatrix} - q_1^T - \\ \vdots \\ - q_p^T - \end{bmatrix}$$

The first row q_1^T, coincides with the linear discriminant q transposed, for the 1-dimension problem (see 2.2), i.e.

$$q_1 = \quad q = \quad \underset{q \neq 0}{argmax} \quad \frac{1}{2}\left(\frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} - \log\frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} + \frac{[q^T(\eta_0 - \eta_1)]^2}{q^T\Sigma_1 q} - 1\right) \tag{3.4}$$

The second row is computed by running again the algorithm, but now imposing that such row is orthogonal to the first, i.e:

$$q_2 = O_1 g \tag{3.5}$$

where $O_1 \,\epsilon\, \mathbb{R}^{n\times(n-1)}$ is a matrix, whose columns generate the orthogonal complement of the subspace generated by q_1, and $g \,\epsilon\, \mathbf{R}^{n-1}$ is the vector that collects the coefficients of the linear combination of the columns of O_1.

In order to compute q_2, a modified version of (3.4) is solved, i.e.

$$g = \quad \underset{g \neq 0}{argmax} \quad \frac{1}{2}\left(\frac{g^T O_1^T\Sigma_0 O_1 g}{g^T O_1^T\Sigma_1 O_1 g} - \log\frac{g^T O_1^T\Sigma_0 O_1 g}{g^T O_1^T\Sigma_1 O_1 g} + \frac{[g^T O_1^T(\eta_0 - \eta_1)]^2}{g^T O_1^T\Sigma_1 O_1 g} - 1\right) . \tag{3.6}$$

The modification introduced, was the substitution of the q in (3.4), by $O_1 g$. This imposes orthogonality condition. Note that this optimization problem has exactly the same form of the 1-dimensional problem in (3.4), being therefore solved in exactly the same way.
Note that $q_2 = O_1 g$.

To solve for row i, it's just a matter of substituting O_1 by O_{i-1}. Being O_{i-1}, the matrix that generates the orthogonal complement to the subspace generated by the $i-1$ rows, previously calculated.

It's important to note, that the complexity of the sub-problems solved to compute the p rows is decreasing. This is due to the fact that the optimization is being made in subspaces whose dimensions are decreasing.

Chapter 4

Computer Simulations

4.1 Hit Rates

In this section we compare the performance of the four criteria used to construct linear discriminants: Kullback-Leibler Divergence (KLD), J-Divergence (JD), Chernoff Distance (CHF), Fisher's Distance (FLDA). The index of performance is the hitrate, i.e., the percentage of correct decisions in the lower-dimensional space $\mathbb{R}^p$. We consider the following simulation scenarios: dimensionality reduction to $\mathbb{R}$, ($p = 1$) (4.1.1), dimensionality reduction to $\mathbb{R}^p$ (4.1.2), and the using of $k > 1$ samples from the same source in the detection process (4.1.3). The hit-rates were computed by Monte Carlo simulations. We generated 100000 samples from each Gaussian source in $\mathbb{R}^n$. Each sample is projected to $\mathbb{R}^p$ by each of the four linear discriminants and classified by the optimum detector (Maximum Likelihood Detector). The hit rate for a given linear discriminant corresponds to the average of correct decisions over the 200000 samples from both sources.

4.1.1 Dimensionality Reduction to $\mathbb{R}$

The results of the simulations are presented for three distinct cases concerning the parameters of the sources: distinct means and distinct covariance matrices (table 4.1), equal means and distinct covariance matrices (table 4.2), distinct means and equal covariance matrices (table 4.3), with increasing data dimensionality $n = 10, 20, 30, 40, 50$.

Table 4.1: Distinct Means and Distinct Covariance Matrices

	JD	KLD	CHF	FLDA
n=10	0.8508	0.8309	0.8607	0.5969
n=20	0.9870	0.9870	0.9415	0.6412
n=30	0.9010	0.9013	0.9376	0.7056
n=40	0.9935	0.9936	0.9426	0.6931
n=50	0.9891	0.9893	0.9430	0.7088

Table 4.2: Equal Means and Distinct Covariance Matrices

	JD	KLD	CHF	FLDA
n=10	0.9852	0.9853	0.9404	0.6006
n=20	0.9827	0.9830	0.9411	0.5159
n=30	0.9820	0.9821	0.9400	0.5255
n=40	0.9867	0.9868	0.9403	0.5243
n=50	0.9583	0.9586	0.9376	0.5053

Table 4.3: Distinct Means and Equal Covariance Matrices

	JD	KLD	CHF	FLDA
n=10	0.6740	0.6743	0.6743	0.6737
n=20	0.9255	0.9254	0.9253	0.9253
n=30	0.9203	0.9202	0.9204	0.9206
n=40	0.9665	0.9667	0.9666	0.9668
n=50	1.0000	1.0000	1.0000	1.0000

As we can see, the FLDA always corresponds to the worst performance tables 4.1 and 4.2. In table 4.3 all criteria performed about the same as predicted by the theory (see section D).

4.1.2 Dimensionality Reduction to $\mathbb{R}^p$

The next tables collect the results obtained when reducing the n-dimensional data to $\mathbb{R}^p$ with $p > 1$, through the linear discriminants that maximize the Chernoff Distance, Kullback-Leibler Divergence, J-Divergence and Fisher's Distance. For each criterion, there is a table with the classification hit-rates, for varying data dimensionality $n = 10, 20, 30, 40, 50$ and increasing p. Once again, we present the results for the three different cases concerning the parameters of the sources.

For the same data dimension n (rows of the tables) and for the same sources setup, the different criteria may be compared. Ex: row 1 from table 4.4 can be compared with row 1 from tables 4.5, 4.6, 4.7.

Table 4.4: Chernoff Distance Criteria with Distinct Means and Distinct Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.9027	0.9827	x	x	x	x
n=20	0.9420	0.9995	0.9999	x	x	x
n=30	0.9419	0.9994	1.0000	1.0000	x	x
n=40	0.9409	1.0000	1.0000	1.0000	1.0000	x
n=50	0.9431	0.9994	1.0000	1.0000	1.0000	1.0000

Table 4.5: Fisher's Distance Criteria with Distinct Means and Distinct Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.5915	0.8523	x	x	x	x
n=20	0.6036	0.8200	0.9943	x	x	x
n=30	0.6396	0.7933	0.9735	1.0000	x	x
n=40	0.6439	0.7368	0.9302	0.9924	1.0000	x
n=50	0.6758	0.7677	0.9080	0.9863	0.9994	1.0000

Table 4.6: Kullback-Leibler Divergence Criteria with Distinct Means and Distinct Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.8426	0.9341	x	x	x	x
n=20	0.9905	0.9997	0.9998	x	x	x
n=30	0.9899	1.0000	1.0000	1.0000	x	x
n=40	0.9748	0.9998	1.0000	1.0000	1.0000	x
n=50	0.9893	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4.7: J-Divergence Criteria with Distinct Means and Distinct Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.8881	0.9808	x	x	x	x
n=20	0.9901	0.9997	0.9999	x	x	x
n=30	0.9898	0.9999	1.0000	1.0000	x	x
n=40	0.9747	0.9998	1.0000	1.0000	1.0000	x
n=50	0.9891	0.9999	1.0000	1.0000	1.0000	1.0000

For the case of distinct means and distinct covariance matrices for the sources, it can be seen, that the hit-rates are improved by increasing p. This is to be expected, as less information is being discarded. In chapter 1 in section 1.2, we saw that the pure Fisher's Distance Maximization criterion cannot handle dimensionality reduction to $\mathbb{R}^p$ with $p > 1$. However what can be seen from the tables

is that, Fisher´s Distance Maximization criterion (FLDA) benefits of applying the greedy technique we proposed in chapter 3.

Table 4.8: Chernoff Distance Criteria with Equal Means and Distinct Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.9013	0.9817	x	x	x	x
n=20	0.9413	0.9994	0.9999	x	x	x
n=30	0.9417	0.9995	1.0000	1.0000	x	x
n=40	0.9402	1.0000	1.0000	1.0000	1.0000	x
n=50	0.9406	0.9994	1.0000	1.0000	1.0000	1.0000

Table 4.9: Fisher's Distance Criteria with Equal Means and Distinct Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.5138	0.8480	x	x	x	x
n=20	0.5472	0.8111	0.9942	x	x	x
n=30	0.5473	0.7730	0.9719	1.0000	x	x
n=40	0.5173	0.6958	0.9237	0.9914	1.0000	x
n=50	0.5104	0.7183	0.8932	0.9846	0.9994	1.0000

Table 4.10: Kullback-Leibler Divergence Criteria with Equal Means and Distinct Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.8412	0.9327	x	x	x	x
n=20	0.9906	0.9997	0.9998	x	x	x
n=30	0.9904	1.0000	1.0000	1.0000	x	x
n=40	0.9749	0.9998	0.9999	1.0000	1.0000	x
n=50	0.9890	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4.11: J-Divergence Criteria with Equal Means and Distinct Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.8867	0.9801	x	x	x	x
n=20	0.9903	0.9997	0.9999	x	x	x
n=30	0.9905	0.9999	1.0000	1.0000	x	x
n=40	0.9748	0.9998	1.0000	1.0000	1.0000	x
n=50	0.9890	0.9999	1.0000	1.0000	1.0000	1.0000

In this case, the means of the sources are equal and the covariance matrices are distinct. All the criteria exhibit a better performance with increasing p. The most notorious case,correspond to FLDA: for $p = 1$, it can be seen (table 4.9) that the detector is as good as a random classificator, however for $p = 7$ the performance increases drastically. As before the other criteria outperform the FLDA criterion. This proves their ability for discriminating the sources through their covariance.

Table 4.12: Chernoff Distance Criteria with Distinct Means and Equal Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.8402	0.8402	x	x	x	x
n=20	0.7752	0.7753	0.7753	x	x	x
n=30	0.9685	0.9686	0.9686	0.9686	x	x
n=40	1.0000	1.0000	1.0000	1.0000	1.0000	x
n=50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4.13: Fisher's Distance Criteria with Distinct Means and Equal Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.8402	0.8402	x	x	x	x
n=20	0.7753	0.7753	0.7753	x	x	x
n=30	0.9686	0.9686	0.9686	0.9686	x	x
n=40	1.0000	1.0000	1.0000	1.0000	1.0000	x
n=50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4.14: Kullback-Leibler Divergence Criteria with Distinct Means and Equal Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.8402	0.8402	x	x	x	x
n=20	0.7753	0.7753	0.7753	x	x	x
n=30	0.9686	0.9686	0.9686	0.9686	x	x
n=40	1.0000	1.0000	1.0000	1.0000	1.0000	x
n=50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4.15: J-Divergence Criteria with Distinct Means and Equal Covariance Matrices

	p=1	p=7	p=17	p=27	p=37	p=47
n=10	0.8402	0.8402	x	x	x	x
n=20	0.7753	0.7753	0.7753	x	x	x
n=30	0.9685	0.9686	0.9686	0.9686	x	x
n=40	1.0000	1.0000	1.0000	1.0000	1.0000	x
n=50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

In this situation the parameters of the distributions are equal covariance matrices an distinct means. As shown in (D), all the criteria perform the same which is confirmed by analyzing tables 4.12 to 4.15. An intuitive interpretation of this is that apart from an invertible linear transformation of the samples, there is no difference of the situation where the covariance matrices are the identity and the means are distinct: for this situation it's obvious that the best possible linear discriminant is a vector that is aligned with $\mu_0 - \mu_1$; in the orthogonal directions the distributions are exactly the same and there is nothing to discriminate.

4.1.3 Asymptotic Behavior

In this last set of tables,

Table 4.16: Distinct Covariance Matrices and Distinct Means

	KLS	KLA	CHF	FLDA
k=1	0.8881	0.8426	0.9027	0.5915
k=3	0.9931	0.9934	0.9971	0.6709
k=5	0.9999	0.9999	0.9998	0.7584
k=7	1.0000	1.0000	1.0000	0.8595
k=9	1.0000	1.0000	1.0000	0.8634

Table 4.17: Distinct Covariance Matrices and Equal Means

	KLS	KLA	CHF	FLDA
k=1	0.9314	0.9314	0.8987	0.5044
k=3	1.0000	1.0000	0.9983	0.5465
k=5	1.0000	1.0000	0.9998	0.5512
k=7	1.0000	1.0000	1.0000	0.7413
k=9	1.0000	1.0000	1.0000	0.7114

Table 4.18: Equal Covariance Matrices and Distinct Means

	KLS	KLA	CHF	FLDA
k=1	0.6235	0.6235	0.6235	0.6235
k=3	1.0000	1.0000	1.0000	1.0000
k=5	1.0000	1.0000	1.0000	1.0000

as would be predictable, the performance of the detector increases when k is increased. It's important to note that the FLDA criterion never achieved a hit rate of 1, however in tables 4.16 and 4.17 the other criteria achieved a hit-rate of 1 for $k = 5$. In table 4.18 is again visible the criteria equivalence under equal covariance matrices.

4.2 ROC-Curves

The ROC curves give the probability of detection as a function of the probability of false alarm. In figures 4.1, 4.2, and 4.3 we present the results for the several criteria and for the three cases of the n-dimensional probability density functions parameters. The n-dimensional samples where classified by their 1-dimensional signatures obtained by the application of the linear discriminants for the several criteria. The dimension of the high dimensional samples n is 10.

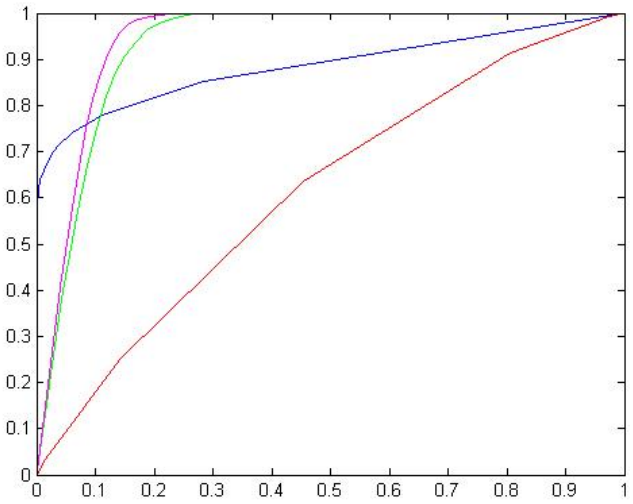

Figure 4.1: Distinct Means and Distinct Covariance Matrices. Legend: Magenta - Chernoff Distance , Red - Fisher's Distance, Blue - Kullback-Leibler Divergence, Green - J-Divergence.

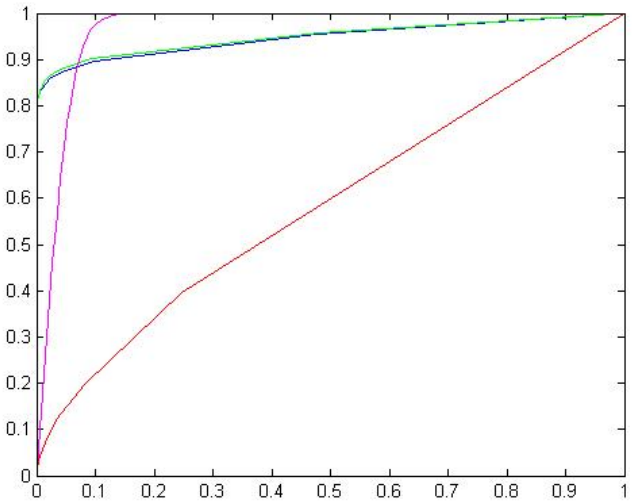

Figure 4.2: Equal Means and Distinct Covariance Matrices. Legend: Magenta - Chernoff Distance , Red - Fisher's Distance, Blue - Kullback-Leibler Divergence, Green - J-Divergence.

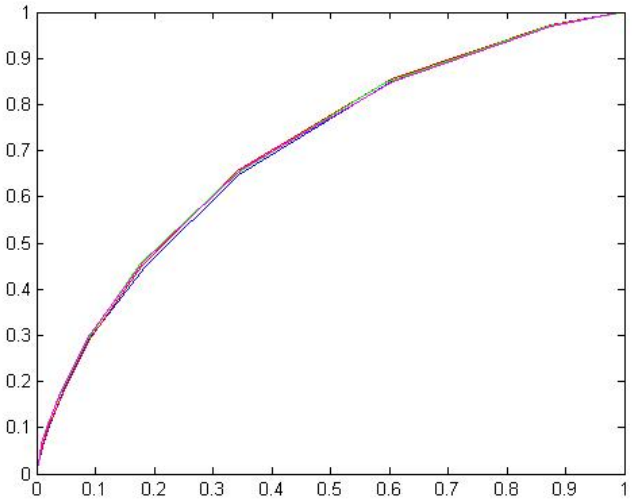

Figure 4.3: Distinct Means and Equal Covariance Matrices. Legend: Magenta - Chernoff Distance , Red - Fisher's Distance, Blue - Kullback-Leibler Divergence, Green - J-Divergence.

Once again, the J-Divergence , Chernoff Distance and Kullback-Leibler Divergence criteria, outperform Fisher's Distance criterion. In figure 4.3, is again clear the criteria equivalence for the case of equal covariance matrices. For the first two situations figure 4.1 and fig. 4.2 where the covariance matrices are different, it can be seen that bellow probabilites of false alarm of $0,1$, the kullback-Leibler Divergence Maximization criterion is the one that achieves a greater probability of detection. This fact is an implication of the asymmetrical character that this criterion exhibits. Looking at the Kullback-Leibler Divergence general definition

$$D_{KL}(f_0||f_1) = \int_{\mathbb{R}^p} f_0(y)\, log\, \frac{f_0(y)}{f_1(y)} dy,$$

this fact is very easily interpreted as maximizing the expected value of the log-likelihood ratio $\log \frac{f_0(y)}{f_1(y)}$, under f_0, which is equivalent to maximize the number of correct detections of samples from s_0, when such samples were generated by s_0.

Chapter 5

Conclusions and Future Work

In this thesis we proposed new criteria for designing linear discriminants for data dimensionality reduction prior to the application of a binary detector. We also developed algorithms to solve the non-convex optimization problems corresponding to the design of these new linear discriminants. These algorithms compute the linear discriminants that maximize the Chernoff Distance, the J-Divergence and Kullback-Leibler Divergence between the probability density functions that characterize the low-dimensional signatures of the original data.

The optimization problems that result from maximizing these measures of dissimilarity of the two sources are non-convex. However it was possible to solve them efficiently (global optimality), trough reformulations and the use of duality theory, for the case where the n-dimensional samples are mapped to $\mathbb{R}$. A suboptimal strategy was proposed for the case of mapping the samples to $\mathbf{R}^p$, with p greater than one.

The results present in chapter 4, proved uniquivocally that the new techniques outperform the Fisher's Distance Criteria. This is due to the fact that the new criteria can discriminates the probability density functions through their variance see figures 1.1 and 1.2. It's important to note that a Gaussian probability density function is characterized by its two first moments, the covariance matrix and the mean. Thus a good discriminator should use both to distinguish them. This is secured by Chernoff Distance, J-Divergence and Kullback-Leibler Divergence criteria.

Future Work. In this thesis, we focused on the Gaussian case. This framework may model many practical situations, however it is far from exhausting all practical applications. This observation leads immediately to two generalizations. The first generalization would be considering the case where the two sources, instead of following Gaussian distributions follow a Gaussian mixture. The importance of this generalization is clear: any regular probability density distribution can be well approximated by a mixture of Gaussians. The second generalization would be to study the multiclass problem. The main obstacles to these generalizations are: for the Gaussian mixture case, the computation of the closed form expressions for the Chernoff Distance, J-Divergence and Kullback-Leibler Divergence, and the non-convexity of the design; for the multiclass problem there is no asymptotic expresions as there are for the two class situation (Stein's Lemma), however may be useful trying to optimize pairwise Chernoff Distances and J-Divergences between the probability density functions that characterize the several classes.

Bibliography

[1] R.O. Duda, P.E. Hart, D.H. Stork. *Pattern Classification (2nd ed.)*. Wiley Interscience, (2000).

[2] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley-Interscience; New Ed edition (August 4, 2004).

[3] Jian Yang, Hui Ye, Zhang David. *A new LDA-KL combined method for feature extraction and its generalisation* Pattern Analysis and Applications, archive Volume 7 , Issue 2 (July 2004) table of contents Pages: 225 - 225 Year of Publication: 2004 ISSN:1433-7541

[4] Yongping Li , Josef Kittler , Jiri Matas. "Effective Implementation of Linear Discriminant Analysis for Face Recognition and Verification." Proceedings of the 8th International Conference on Computer Analysis of Images and Patterns, p.234-242, September 01-03, 1999.

[5] J.A.Thomas, T.M.Cover. *Elements of Information Theory*. John Wiley and Sons, Inc.,1991.

[6] Luis Rueda, Myriam Herrera. " A New Linear Dimensionality Reduction Technique Based on Chernoff Distance" IBERAMIA-SBIA 2006: 299-308

[7] Louis L. Scharf. *Statistical Signal Processing: Detection, Estimation and Time series Analysis*. Addision and Wesley, 1991.

[8] B. T. Polyak. "Convexity of Quadratic Transformations and Its Use in Control and Optimization". Journal of Optimization Theory and Applications. Vol. 99, No. 3, pp.553-583, December 1998.

[9] Lloyd L. Dines. "On the Mapping of Quadratic Forms." Bull. Amer. Math. Soc. Volume 47, Number 6 (1941), 494-498.

[10] Don H. Johnson, Sinan Sinanovic. *Symetrizing the Kullback-Leibler Distance*. Computer and Information Technology Institute. Department of Electrical Engineering. Rice University. Houston

Appendix A

Quadratic Program with Quadratic Constraints, Strong-Duality result

A.1 Introduction to Strong Duality

In all the optimization problems presented in chapter 2, emerged the necessity of solving a particular kind of optimization problem, a quadratic program with two quadratic constraints

$$\begin{array}{cc} min & q^T A q \\ q^T B q = b & \\ q^T C q = c & \\ \text{var} : q \,\epsilon\, \mathbb{R}^n & \end{array} \tag{A.1}$$

where A is a symmetric matrix, B and C are positive definite symmetric matrices, b and c are positive scalars.

This optimization problem in all its generality is very hard to solve due to its non-convex quadratic constraints. However, when the matrices involved in the quadratic expressions have some properties, strong duality exists.

Strong duality is a very important tool in optimization theory. Makes possible solving a very hard optimization problem, trough a easy one. The very hard problem is usually called the Primal Problem, and the easy one is called the Dual Problem. Strong duality says that under certain conditions, the Primal and Dual problems give the same result, that is: their objective functions attend the same value at their respective globally optimal points.

In optimization theory, the Primal Problem and the respective Dual Problem are:

Primal Problem.

$$\begin{array}{rcc} p^* = & \min & f(x) \\ & x \,\epsilon\, X & \\ & h(x) = 0 & \\ & g(x) \leq 0 & \\ & \text{var} : x \,\epsilon\, \mathbb{R}^n & \end{array} \tag{A.2}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$
$X \subset \mathbb{R}^n$

$h : \mathbb{R}^n \to \mathbb{R}^p$, $h = (h_1, h_2, ..., h_p)$
$g : \mathbb{R}^n \to \mathbb{R}^m$, $g = (g_1, g_2, ..., g_m)$

Dual Problem.

$$\begin{array}{rcl} d^* = & \max & L(\lambda, \mu) \\ & \mu \geq 0 & \\ & \text{var} : (\lambda, \mu) \, \epsilon \, \mathbb{R}^p \times \mathbb{R}^m & \end{array} \tag{A.3}$$

where

$$\begin{array}{rcl} L(\lambda, \mu) = & \inf & L(x, \lambda, \mu) \\ & x \, \epsilon \, X & \end{array} \tag{A.4}$$

being $L(x, \lambda, \mu) = f(x) + \lambda^T h(x) + \mu^T g(x)$ the Lagrangian function.

Applying the definition of the Dual Problem to the quadratic program with the quadratic constraints in (A.1) comes:

$$\begin{array}{c} \max \quad -\lambda_1 b - \lambda_2 c \\ A + \lambda_1 B + \lambda_2 C \succeq 0 \\ \text{var} : (\lambda_1, \lambda_2) \, \epsilon \, \mathbb{R}^2 \end{array} \tag{A.5}$$

which is an SDP (Semidefinite Program) and therefore convex, in just two variables independently of n that can be very large.

In the algorithms presented in section 2, is also needed the Bi-Dual Program. This is simply the dual problem of the first dual, since the first dual problem is convex, strong duality exists between them.

Bi-Dual Problem.

$$\begin{array}{c} \min \quad tr(AQ) \\ tr(BQ) = b \\ tr(CQ) = c \\ Q \succeq 0 \end{array} \tag{A.6}$$

where Q is a symmetric positive semi-definite matrix of dimension n.

The Bi-Dual Problem in (A.6) is very much harder to solve than the Dual Problem in (A.5), due to the dimensionality of the optimization variable $Q \, \epsilon \, \mathbf{S}_+^n$.

Provided strong duality exists for the problem in (A.1) and the Bi-Dual Problem has a single solution, its solution Q is a rank-1 matrix, from which the only eingenvector is the solution for the Primal Problem in (A.1). This is the way how the solution $q \, \epsilon \, \mathbb{R}^n$ is retrieved.

A.2 Strong Duality Result Demonstration

It is easy to show and a consequence of the definition of the Dual Problem, that $p^* \geq d^*$, so in order to show that there exists strong duality i.e. $p^* = d^*$, it is just needed to show $p^* \leq d^*$.

For this demonstration is of fundamental importance the following result, due to B. T. Polyak [8], related with quadratic mappings:

Theorem

$$\underset{(\mu_a,\mu_b,\mu_c)\,\epsilon\,\mathbb{R}^3}{\exists} : \mu_a A + \mu_b B + \mu_c C > 0 \tag{A.7}$$

implies that:

$$\{(x^T Ax, x^T Bx, x^T Cx) : x\,\epsilon\,\mathbb{R}^n\} \tag{A.8}$$

is a closed convex pointed cone, for $n \geq 3$.

Let $U = \{(f,u,v)\,\epsilon\ \mathbb{R}^3 : (f,u,v) = (x^T Ax, x^T Bx, x^T Cx), x\,\epsilon\,\mathbb{R}^n\}$ and $V = \{(f,u,v)\,\epsilon\ \mathbb{R}^3 : f < p^*, u = b, v = c\}$

The matrices B and C are positive definite, thus the conditions of the theorem in (A.7) are verified with $(\mu_a,\mu_b,\mu_c) = (0,1,1)$ and therefore, set U is a closed convex pointed cone. Set V is simply a half straight line, being also convex.

The sets U and V are disjoint. Supposing they are not, there is $x_0\,\epsilon\,\mathbb{R}^n$, such that $(x_0^T Ax_0, x_0^T Bx_0, x_0^T Cx_0)$ belonging to set U, belongs also to set V, verifying $x_0^T Ax_0 < p^*$, $x_0^T Bx_0 = b$ and $x_0^T Cx_0 = c$, in contradiction with the fact that:

$$p^* = \begin{array}{cc} \min & x^T Ax \\ x^T Bx = b & \\ x^T Cx = c & \end{array}$$

Because the sets U and V are convex and disjoint, there is an hyperplane that separates them i.e:

$$\underset{s\,\epsilon\,\mathbb{R}^3\backslash 0}{\exists} \quad s^T z \geq r\ \forall\ z\epsilon U\ \wedge\ s^T z \leq r\ \forall\ z\epsilon V \tag{A.9}$$

Attending to sets U and V definitions, and being $s = (s_f, s_u, s_v)$, the proposition in (A.9) is equivalent to the following two:

$$s_f x^T Ax + s_u x^T Bx + s_v x^T Cx \geq r\ \forall\, x\,\epsilon\,\mathbb{R}^n \tag{A.10}$$

$$s_f f + s_u b + s_v c \leq r\ \forall\, f \leq p_* \tag{A.11}$$

The proposition in (A.11) implies that $s_f \geq 0$. Otherwise and attending to the fact that $f\epsilon]-\infty, p^*[$, the product $s_f f$, could be made arbitrarily large by choosing f arbitrarily negative, and the proposition wouldn't be satisfied.

The fact that $s_f \geq 0$ will lead to the strong duality result. This inequality is separated in two different situations. The first situation is $s_f > 0$, the second situation is $s_f = 0$. The first situation will lead directly to the strong duality result, and the second will generate a contradiction, by implying the inexistence of the separating hyperplane, that was proven to exist.

First situation: $s_f > 0$**.** For this situation the propositions in (A.10) and (A.11) can be rewritten as:

$$x^T Ax + \left(\frac{s_u}{s_f}\right) x^T Bx + \left(\frac{s_v}{s_f}\right) x^T Cx \geq \left(\frac{r}{s_f}\right) \forall\, x \,\epsilon\, \mathbb{R}^n \tag{A.12}$$

$$f + \left(\frac{s_u}{s_f}\right) b + \left(\frac{s_v}{s_f}\right) c \leq \left(\frac{r}{s_f}\right) \forall\, f \leq p_* \tag{A.13}$$

Defining $s_u^{'} = \frac{s_u}{s_f}$ and $s_v^{'} = \frac{s_v}{s_f}$ and adding term by term the two equations in the proposition in (A.12) and (A.13), follows:

$$x^T Ax + s_u^{'} x^T Bx + s_v^{'} x^T Cx \geq f + s_v^{'} c \forall\, x \,\epsilon\, \mathbb{R}^n,\ f \leq p_* \Leftrightarrow \tag{A.14}$$

$$x^T Ax + s_u^{'}(x^T Bx - b) + s_v^{'}(x^T Cx - c) \geq f \forall\, x \,\epsilon\, \mathbb{R}^n,\ f \leq p_* \Leftrightarrow \tag{A.15}$$

$$\inf_{x \epsilon \mathbb{R}^n} \quad x^T Ax + s_u^{'}(x^T Bx - b) + s_v^{'}(x^T Cx - c) \geq f \,\forall\, f < p_* \tag{A.16}$$

The left term in inequation (A.16) is the Dual function evaluated at point $(s_u^{'}, s_v^{'})$. It is known that the Dual function is always less or equal to p_*, but Attending to set V definition, f can be made arbitrarily close to p_* and so it follows the strong duality result.

Second situation: $s_f = 0$**.** This situation will lead to $s = 0$ implying the inexistence of the separating hyperplane for the sets U and V.

Since $s_f = 0$, the equations in the propositions in A.9 are equivalent to the following two:

$$s_u x^T Bx + s_v x^T Cx \geq r \,\forall\, x \,\epsilon\, \mathbb{R}^n \tag{A.17}$$

$$s_u b + s_v c \leq r \,\forall\, f \leq p_* \tag{A.18}$$

Adding the previous equations term by term, follows that:

$$s_u(x^T Bx - b) + s_v(x^T Cx - c) \geq 0 \,\forall\, x \,\epsilon\, \mathbb{R}^n \tag{A.19}$$

Supposing the existence of x_{b_+} and x_{b_-} such that:

$$\begin{gathered} x_{b_+}^T Bx_{b_+} - b > 0 \\ x_{b_-}^T Bx_{b_-} - b < 0 \\ x_{b_{+/-}}^T Cx_{b_{+/-}} - c = 0 \end{gathered} \tag{A.20}$$

the following propositions are true:

$$\begin{array}{ccccc} s_u(x_{b_+}^T Bx_{b_+} - b) + s_v(x_{b_+}^T Cx_{b_+} - c) \geq 0 & \Rightarrow & s_u \geq 0 \\ > 0 & & = 0 \\ s_u(x_{b_-}^T Bx_{b_-} - b) + s_v(x_{b_-}^T Cx_{b_-} - c) \geq 0 & \Rightarrow & s_u \leq 0 \\ < 0 & & = 0 \end{array} \tag{A.21}$$

The previous propositions combined imply that $s_u = 0$, the same reasoning implies $s_v = 0$. The conclusion is that considering $s_f = 0$ comes $s_u = 0$ and $s_v = 0$ and thus $s = 0$ implying the inexistence of the separating hyperplane for the sets U and V, that was proven to exist.

Appendix B

Stiefel Matrix Constraint Invariance

Here is shown that the linear discriminant $Q \epsilon \mathbb{R}^{p \times n}$ can be taken as a Stiefel Matrix. For convenience of notation, in this section the linear discriminant Q is denoted by A. This is proven by showing this result for the Kullback-Leibler Divergence whose expression

$$D_{KL}(f_0||f_1)(A) = \frac{1}{2}\left(\log\frac{|A\Sigma_1 A^T|}{|A\Sigma_0 A^T|} + tr((A\Sigma_1 A^T)^{-1}(A\Sigma_0 A^T)) + (\mu_0-\mu_1)^T A^T (A\Sigma_1 A^T)^{-1} A(\mu_0-\mu_1) - p\right) \tag{B.1}$$

was previously presented in (1.14).

The statement that is going to be proved is

$$D_{KL}(f_0||f_1)(A) = D_{KL}(f_0||f_1)(Q) \tag{B.2}$$

where $Q : Q^T Q = I_n$, i.e. Q is Stiefel.

Assuming linear independence between the rows of A, it's possible to write $A = RQ$. This is called the RQ factorization, where Matrix $R \epsilon \mathbb{R}^{p \times p}$ is an invertible upper triangular matrix and $Q \epsilon \mathbb{R}^{p \times n}$ is orthonormal.

Inserting $A = RQ$ in (B.1), comes

$$\begin{aligned} D_{KL}(f_0||f_1)(RQ) = \quad & \tfrac{1}{2}\left(\log\tfrac{|RQ\Sigma_1 Q^T R^T|}{|RQ\Sigma_0 Q^T R^T|} + tr((RQ\Sigma_1 Q^T R^T)^{-1}(RQ\Sigma_0 Q^T R^T))\right) \\ & + \tfrac{1}{2}\left((\mu_0-\mu_1)^T Q^T R^T (RQ\Sigma_1 Q^T R^T)^{-1} RQ(\mu_0-\mu_1) - p\right) \end{aligned} \tag{B.3}$$

Attending to the fact that R is invertible, it´s a matter of using the algebraic properties of the determinant and the trace of a matrix to see that R cancels out in (B.3), leading to (B.1).

Appendix C

Set Properties

In this section, we show that the set C defined in (2.23) as $C = \{(q^T\Sigma_0 q, q^T M q) : q^T\Sigma_0 q q^T\Sigma_1 q = 1\}$ is compact and connected.

Defining $A = \{q \epsilon \mathbb{R}^n : q^T\Sigma_0 q q^T\Sigma_1 q = 1\}$ and

$$\begin{array}{rcl} F : \mathbb{R}^n & \rightarrow & \mathbb{R}^2 \\ q & \mapsto & (q^T\Sigma_0 q, q^T M q) \end{array} \tag{C.1}$$

it follows that $C = F(A)$, where F is a continuous map. Thus the compactness and connectedness of set C will follow from the compactness and connectedness of A, which is now shown:

Defining $B = \{q \epsilon \mathbb{R}^n : ||q|| = 1\}$ and

$$\begin{array}{rcl} \Phi : \mathbb{R}^n & \rightarrow & \mathbb{R}^n \\ q & \mapsto & \frac{q}{\sqrt[4]{q^T\Sigma_0 q q^T\Sigma_1 q}} \end{array} \tag{C.2}$$

we see that $A = \Phi(B)$, where Φ is continuous function over B, and B is the compact and connected unit sphere, implying therefore that A is compact and connected.

In order to show $A = \Phi(B)$, it is proven $\Phi(B) \subset A$ and $A \subset \Phi(B)$.

Case $\Phi(B) \subset A$**.** Given $q \epsilon B$, it's a matter of algebra, to verify that $\Phi(q)$ verifies $\Phi(q)^T\Sigma_0\Phi(q)\Phi(q)^T\Sigma_1\Phi(q) = 1$, and thus belongs to set A.

Case $A \subset \Phi(B)$**.** Given $q^{'} \epsilon A$, and therefore verifying $q^{'T}\Sigma_0 q^{'} q^{'T}\Sigma_1 q^{'} = 1$, $q = \frac{q'}{||q'||} \epsilon B$, is such that $\Phi(q) \epsilon A$. So for every $q^{'} \epsilon A$ there is a point $q \epsilon B$ such that $\Phi(q) \epsilon A$, implying therefore $A \subset \Phi(B)$.

Appendix D

Criteria Equivalence

In this section it is shown that when $F_0 = N(\mu_0, \Sigma_0)$ and $F_1 = N(\mu_1, \Sigma_1)$ have equal covariance matrices $(\Sigma_0 = \Sigma_1)$ the Kullback-Leibler Divergence, J-Divergence and Chernoff Distance criteria are all equivalent to Fisher's Distance. This result is illustrated just for the Kullback-Leibler Divergence. The same line of reasoning shows this result is valid for the other criteria.

Demonstration. The Kullback-Leibler Divergence $D_{KL}(f_0||f_1)(Q)$ between the probability density functions $f_0(Q) = N(Q\mu_0, Q\Sigma_0 Q^T)$ and $f_1(Q) = N(Q\mu_1, Q\Sigma_1 Q^T)$ is

$$D_{KL}(f_0||f_1)(Q) = \frac{1}{2}\left(\log\frac{|Q\Sigma_1 Q^T|}{|Q\Sigma_0 Q^T|} + tr((Q\Sigma_1 Q^T)^{-1}(Q\Sigma_0 Q^T)) + (\mu_0-\mu_1)^T Q^T (Q\Sigma_1 Q^T)^{-1} Q(\mu_0-\mu_1) - p\right). \tag{D.1}$$

Inserting $\Sigma_0 = \Sigma_1 = \Sigma$ in (D.1), yields

$$D_{KL}(f_0||f_1)(Q) = \frac{1}{2}(\mu_0-\mu_1)^T Q^T (Q\Sigma Q^T)^{-1} Q(\mu_0-\mu_1). \tag{D.2}$$

Equation (D.2) coincides with Fisher's Distance presented in (1.5), whose expression is reproduced here for $\Sigma_0 = \Sigma_1 = \Sigma$:

$$\begin{array}{rl} FD(f_0||f_1)(Q) = & \frac{1}{2}\text{tr}\{(Q\Sigma Q^T)^{-1}(Q(\mu_0-\mu_1)(\mu_0-\mu_1)^T Q^T)\} \\ & \frac{1}{2}\text{tr}\{(\mu_0-\mu_1)^T)Q^T(Q\Sigma Q^T)^{-1}Q(\mu_0-\mu_1)\} \\ & \frac{1}{2}(\mu_0-\mu_1)^T)Q^T(Q\Sigma Q^T)^{-1}Q(\mu_0-\mu_1) \end{array} \tag{D.3}$$