A Novel Feature Extraction Algorithm for Asymmetric Classification

David Lindgren, Per Spångeus

Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden
WWW: <http://www.control.isy.liu.se>
E-mail: david@isy.liu.se

August 16, 2002

Report no.: LiTH-ISY-R-2457
Submitted to the Sensors Journal

Technical reports from the Control & Communication group in Linköping are
available at <http://www.control.isy.liu.se/publications>.

A Novel Feature Extraction Algorithm for Asymmetric Classification

David Lindgren

August 16, 2002

Abstract

A linear feature extraction technique for asymmetric distributions is introduced, the *asymmetric class projection* (ACP). By *asymmetric classification* is understood discrimination among distributions with different covariance matrices. Two distributions with unequal covariance matrices do not in general have a symmetry plane, a fact that makes the analysis more difficult compared to the symmetric case. The ACP is similar to *linear discriminant analysis* (LDA) in the respect that both aim at extracting discriminating features (linear combinations or projections) from many variables. However, the drawback of the well known LDA is the assumption of symmetric classes with separated centroids. The ACP, in contrast, works on (two) possibly concentric distributions with unequal covariance matrices. The ACP is tested on data from an electronic nose with the purpose of distinguish bad grain from good.

1 Introduction

The background to this work is the signal processing for an *electronic nose* (EN) that was used to detect bad grain samples. An EN is a gas or smell sensor that is sensitive to a wide range of gas mixtures. Usually an EN consists of a whole set of discrete sensors (for instance *semiconductor*, *conductive polymer* (CP), *surface acoustic wave* (SAW) and *quartz crystal micro balance* (QCM) sensors) to give high potential selectivity among some set of gas mixtures. The signal from a semiconductor sensor, for instance, could be recorded under an interval of time when a gas mixer switches from clean dry air to the gas that should be identified. This sampled (time discrete) signal is compared to other signals taken at other time instants and due to other gases. Typically, an EN outputs a large amount of data for each measurement (or sniff) and it is up to the signal processing to make sense of this data. Although this work was inspired by the EN, the results are general and not limited to sensor systems. The discussion will thus be held on a rather general level. For more on the EN specifically, see for instance [15]. For an overview of signal processing methods for sensors, see [8, 12].

We assign to the vector $x_i \in \mathbb{R}^n$ the EN output data from measurement i . Since we do not know exactly what we measure, and since the measurement probably is perturbed with random noise, we assume that it is actually a sample of the random variable x . The dimensionality n is usually very large – say

$n \approx 1000$. From x_i we like to extract the feature y_i with *high accuracy*. y_i is also drawn from a random variable y , since the feature that should be predicted is unknown at the time the measurement is conducted. The feature could for instance be a *concentration of some chemical* (quantification), the selection of a category or *class* from a set of classes (classification), it could be a gas *detection* or it could be the distinction between *good and bad quality*. The signal processing that maps the measurement x to the feature y we denote $f(\cdot)$, and by *high accuracy* is meant that the magnitude of the residual r in the regression

$$y = f(x) + r \quad (1)$$

should be small in general. A common measure for residual magnitude is the *expected quadratic* $E[r^2]$. Usually one assumes that x has a particular distribution, multi-normal for instance, and then one estimates the mean vector(s) and covariance matrix (matrices) of the distribution from a calibration data set obtained from measurements x_i on specimen with *a priori* known features y_i .

By *classification* we define the regression problem (1) where y is a bit vector $\{0,1\}^q$ with one entry for each of q categories, populations, distributions or *classes*. Every possible x is associated by $f(x)$ with one distinct class j , $1 \leq j \leq q$, and further with the bit vector

$$y = \begin{bmatrix} 0_1 \\ \vdots \\ 0_{j-1} \\ 1_j \\ 0_{j+1} \\ \vdots \\ 0_q \end{bmatrix} \quad (2)$$

in such a way that the general probability of misclassification (the error rate) $E[r^T r]/2$ is as small as possible. Here x is a random variable that stems from distribution j with probability $p(y_j)$. Thus, there are q distributions, and $f(\cdot)$ has to find the one that best matches a measurement x_i .

1.1 Regression in a Subspace

Since the dimensionality n of the measurement data x is very high and the entries correlated, the regression in (1) is often difficult to solve directly by *least squares* or by *maximum likelihood* methods, see [19]. A common approach is to use a set of k linear combinations of x comprised as the rows of the k -by- n matrix S . A new k -dimensional random variable x_S is thus calculated by the matrix-vector multiplication $x_S = Sx$. The modified regression is now

$$y = \tilde{f}(Sx) + r, \quad (3)$$

and the objective is to find the linear combinations S and function $\tilde{f}(\cdot)$ that make the magnitude of the residual r small. This is known as *subspace regression* since the regressor is projected onto a linear subspace. How to find an appropriate $\tilde{f}(\cdot)$ will not be discussed here, see instead [5].

There are many well-known and robust techniques to find a subspace S in which the actual linear or nonlinear regression can take place. One of the simplest one is to use a limited number of uncorrelated *principal components* chosen after an error-impact or a variance criterion (PCR), [11]. QR-decomposition is a similar way to orthogonalize the problem, [6]. One popular technique in chemometrics is the *Partial Least Squares* algorithm (PLS), which also can be formulated as a subspace regression method. In the original PLS algorithm, a set of linear combinations of the variables $x^{(j)}$ that have maximum covariance with the variable y is calculated iteratively and used in the regression, [18]. It has been shown, though, that the PLS-subspace can be found by a QR-decomposition of a Krylov matrix composed from the matrix $[x_i]_1^p$ and the vector $[y_i]_1^p$, [14]. A nonlinear extension to PLS is the *polynomial PLS*, which uses a polynomial inner relation between x and y , [17]. Instead of maximizing the covariance between x and y as in PLS, the criterion could be *correlation*, as in *Canonical Correlation Analysis* (CCA), [9] (does not work well applied directly to multicollinear data). In [16, 2] it is described how the ML subspace is calculated (not for multicollinear data). *Projection pursuit regression* (PPR), [3], iteratively finds the directions of the subspace, one at a time, that reduce the (residual) unexplained variance as much as possible.

1.2 Linear Discriminant Analysis

Particularly for classification problems, the *Fisher linear discriminant analysis* (LDA), [9] etc, is a well known technique to find linear combinations or *discriminants* that facilitates the ability to discriminate among q distributions. As a measure of this ability is used the *Fisher ratio*,

$$\Delta = \sqrt{\sum_{j=1}^q (\mu_j - \mu)^T \Sigma^{-1} (\mu_j - \mu)}, \quad (4)$$

where Σ is the covariance matrix, which is assumed to be invertible and equal for all class distributions, μ_j the mean of distribution j , and $\mu = E[x]$. The operator $E[\cdot]$ denotes expectation. In words, Δ is the amount of variation between distribution means with respect to the variation within the distributions, a measure maximized in the subspace given by LDA. For two normal distributions ($q = 2$) with equal covariance matrices and $p(y_1) = p(y_2) = 1/2$, the LDA subspace is the subspace with optimal error rate (minimal $E[r^T r]$). The LDA problem is numerically and computationally efficiently solved by a generalized singular value decomposition, see [13].

Among techniques to find good discriminative subspaces we can also mention [10], where an over-bound on the classification accuracy is locally maximized and [1], which introduces the *optimal linear transformation* (OLT).

1.3 The Asymmetric Classification Problem

The asymmetric classification problem concerns discrimination between (two) distributions with unequal covariance matrices. When the covariance matrices of the classes are *equal*, the optimal decision boundary to separate two normal distributions (minimal error rate) is given by the distributions symmetry plane,

Figure 1: A ACP of a data set with good observations (rings) and bad (crosses).

and it is not difficult to calculate the error rate $E[r^T r]/2$ and to find subspaces where this error rate is minimal.

When the covariance matrices are *unequal*, it exists no symmetry plane, and the classes are said to be *asymmetric*. Still, the optimal classifier is well-known for normal distributions (the *quadratic classifier*), but it is much more difficult to calculate the optimal error rate, especially in the multivariate case. Very little is found in literature about optimal subspaces.

The typical example of an asymmetric problem that we will focus on, is when the two classes are of the type, “good/bad”, “normal/abnormal”, “accept/reject”. Say, for instance, that an EN should be used at a dairy and be “trained” to detect bad milk. The objective of the signal processing is then to transform a point in the feature space into a binary output; *good* or *bad*. The mean of the classes are not necessarily unequal, but it is assumed that “good” can be defined as a restricted region in the feature space while “bad” is everything that does not lie in this region, see Figure 1. In this work we shall actually adopt the notion of a “good” class and a “bad” class, although the theory is general for asymmetric problems.

This work deals with the problem of how to find the best subspace S for asymmetric classification. The subspace can be used to compress data, to enhance the prediction or just to visualize asymmetric classes in a 2-dimensional plot.

1.4 Reader’s Guide

In the next section, the ACP will be introduced and the objective explained as well as how to compute the ACP. After that, a result on Bayes error optimality that concerns the ACP will be presented. In the 4th section the ACP, *Principal Component Analysis* (PCA), and LDA will be compared when applied on a known artificial data set. In section 5, a numerical example with real data from

an EN is subjected to ACP. Conclusions and acknowledgments are found last.

2 The ACP

As indicated above, the purpose of the ACP is to find a set of linear combinations or a *linear subspace* with as much relevant information as possible. This can alternatively be viewed as *data compression*, where we aim at retaining the ability to distinguish “good” from “bad” in the compressed data. The reduction from n to k dimensions ($k < n$) is defined by a set of k linear combinations, each of n variables, comprised in a k -by- n matrix. If S is that matrix, the compression of the measurement vector $x_i \in \mathbb{R}^n$ is calculated by the matrix multiplication $x_S = Sx$. The measurement vector is thus *projected* onto the rows of S , and we shall therefore denote this reduction by *projection* (although S is not a projection matrix in mathematical sense).

If we are interested in a projection of a space with n dimensions where some expectancy or mean vector μ and covariance matrix Σ are known, it is of course interesting to know the corresponding entities in the compressed k -dimensional space. The mean is calculated as $\mu_S = S\mu$, and the covariance matrix as $\Sigma_S = S\Sigma S^T$. In particular, if S is a row vector, both the mean vector and covariance matrix will be compressed to scalars with the mean and variance, respectively.

2.1 Objective

The fundamental assumption of the ACP is the existence of an *ideal* point in feature space. The degree of *good* will decay as we move away from that point. Thus, if we have two sets of data, one good and one bad, the measurements or *observations* in the good set will be well clustered around the ideal point while the bad observations will be more scattered and distant to the ideal point, see Figure 1. This is the basic property a classifier would exploit, and the property the ACP tries to retain in a projection.

Now, consider two n -dimensional random variables: g (the *good*) and b (the *bad*) with mean vectors

$$\mu_g = E[g], \quad \mu_b = E[b], \quad (5)$$

and covariance matrices

$$\Sigma_g = E[(g - \mu_g)(g - \mu_g)^T], \quad \Sigma_b = E[(b - \mu_b)(b - \mu_b)^T].$$

Assume that Σ_g is invertible and that x_i is a sample of either g or b with *a priori* equal probability. For a particular *1-dimensional* projection $x_s = s^T x$, $s \neq 0$, the quotient between bad class variance and good class variance is given by

$$\xi(s) = \frac{\text{Var}[s^T b]}{\text{Var}[s^T g]} = \frac{s^T \Sigma_b s}{s^T \Sigma_g s}. \quad (6)$$

This quotient is a quality measure of the projection onto s . Projections with larger ξ is preferred, since they facilitate the distinction between good and bad, as will be shown later.

In fact, the maximization of (the Rayleigh quotient) $\xi(s)$ is a well-known problem equivalent to the generalized eigenvalue problem, [7]. The generalized eigenvalue problem is to calculate the matrix E with eigenvectors e_i and the diagonal matrix D with eigenvalues λ_i , $1 \leq i \leq n$, such that

$$\Sigma_b E = \Sigma_g E D, \quad (7)$$

$e_i^T e_i = 1$ and $e_i^T \Sigma_g e_j = e_i^T \Sigma_b e_j = 0$ whenever $i \neq j$. We assume that the eigenvalues are ordered so that $\lambda_i \geq \lambda_{i+1}$. The eigenvector e_1 then solves

$$\xi_1 = \arg \max_s \frac{\text{Var}[s^T b]}{\text{Var}[s^T g]} \quad (8)$$

and the eigenvalue λ_1 identifies the optimum value. The vector e_1 gives the linear combination of b with largest variance compared to the variance of the same linear combination of g .

Thus, by solving the generalized eigenvalue problem (7) a subspace is obtained where the good class is well clustered, the bad class well scattered. That this probably is a very good subspace in a Bayes error sense will be shown in section 3. The solution to the generalized eigenvalue problem is very well known and numerically stable and fast algorithms exist, see [13].

2.2 Modified Covariance

The covariance matrix Σ_b is defined as

$$\Sigma_b = \text{E}[(b - \mu_b)(b - \mu_b)^T].$$

This is the standard definition of a covariance matrix, which means that the magnitude of the covariance is a measure of the variation or spread with respect to the mean. However, if the good and bad class are not concentric ($\mu_g \neq \mu_b$) it is more interesting for our purposes to measure the spread of the bad class with respect to the *good* class mean rather than to the bad class mean itself. This can be achieved by in (7) replacing Σ_b with

$$\tilde{\Sigma}_b = \text{E}[(b - \mu_g)(b - \mu_g)^T].$$

With this definition of bad class covariance, $s^T \tilde{\Sigma}_b s$ is a measure of how well the projection onto s spread the *bad class* with respect to the *good class mean*. Note that $\tilde{\Sigma}_b$ is actually not a covariance matrix.

2.3 Generalization to More than One Dimension

The quotient of the modified variance of the bad and the variance of the good distribution is

$$\xi = \frac{\text{E}[(b - \mu_g)^2]}{\text{E}[(g - \mu_g)^2]} = \frac{\tilde{\sigma}_b^2}{\sigma_g^2} = \text{E}[(b - \mu_g)\sigma_g^{-2}(b - \mu_g)]. \quad (9)$$

As described earlier, this is used as a measure of discrimination in 1 dimension. The generalization to *more* than 1 dimensions we define as

$$\begin{aligned}
\xi &= \mathbb{E} \left[(b - \mu_g)^T \Sigma_g^{-1} (b - \mu_g) \right] \\
&= \mathbb{E} \left[\text{trace} \left[\left(\Sigma_g^{-\frac{1}{2}} (b - \mu_g) \right) \left((b - \mu_g)^T \Sigma_g^{-\frac{1}{2}} \right) \right] \right] \\
&= \text{trace} \left[\Sigma_g^{-\frac{1}{2}} \mathbb{E} \left[(b - \mu_g)(b - \mu_g)^T \right] \Sigma_g^{-\frac{1}{2}} \right] = \text{trace} \left[\Sigma_g^{-\frac{1}{2}} \tilde{\Sigma}_b \Sigma_g^{-\frac{1}{2}} \right] \\
&= \text{trace} \left[\Sigma_g^{-1} \tilde{\Sigma}_b \right].
\end{aligned} \tag{10}$$

Now, let E and D be the solution to the generalized eigenvalue problem,

$$\tilde{\Sigma}_b E = \Sigma_g E D \quad E^T \Sigma_g E, E^T \Sigma_b E \text{ diagonal.} \tag{11}$$

be the solution to the generalized eigenvalue problem, then

$$\xi = \text{trace} \left[\Sigma_g^{-1} \tilde{\Sigma}_b \right] = \text{trace} [E D E^T] = \text{trace} [D] = \sum_{i=1}^n \lambda_i, \tag{12}$$

since $e_i^T e_i = 1$. Furthermore, $e_i^T \Sigma_g e_j = e_i^T 0 \Sigma_b e_j = 0$ whenever $i \neq j$, which means that the components of $E^T g$ and $E^T b$ are uncorrelated. Thus, if we should chose among components that are uncorrelated with respect to g and b , it is evident that we should take the k ones with greatest eigenvalues. Thus, we define the k -dimensional ACP-projection by

$$S_{\text{ACP}} = [e_1 \quad e_2 \quad \cdots \quad e_k]^T, \tag{13}$$

and in this subspace the discrimination measure is

$$\xi(S_{\text{ACP}}) = \sum_{i=1}^k \lambda_i. \tag{14}$$

2.4 Relation to LDA

As mentioned in Section 1.2, LDA finds the subspace with maximum

$$\Delta^2 = \sum_{j=1}^q (\mu_j - \mu)^T \Sigma^{-1} (\mu_j - \mu). \tag{15}$$

For *two* random variables (classes) g and b with equal *a priori* likelihood, and with equal covariance $\Sigma_g = \Sigma_b = \Sigma$, straight-forward calculation gives

$$\begin{aligned}
\Delta^2 &= 0.5 \cdot (\mu_b - \mu_g)^T \Sigma^{-1} (\mu_b - \mu_g) \\
&= 0.5 \cdot \mathbb{E} [b - \mu_g]^T \Sigma^{-1} \mathbb{E} [b - \mu_g]
\end{aligned} \tag{16}$$

which is known as the (half, square) *Mahalanobis distance* between the classes. This should be compared to what is maximized by the ACP:

$$\xi = \mathbb{E} [(b - \mu_g)^T \Sigma_g^{-1} (b - \mu_g)]. \tag{17}$$

Figure 2: PDF of g , $p(x, 1)$, and b , $p(x, \xi)$. The grayed area identifies the classification error rate when the classification decision boundaries are $-a$ and a , respectively. Thus, if $|x| > a$, x most likely belongs to b .

It is seen that the major difference is that the expectation for the ACP is quadratic, $\xi = \mathbb{E}[m^T m]$ for $m = (b - \mu_g)\Sigma_g^{-\frac{1}{2}}$, while for LDA it is $\Delta^2 = \mathbb{E}[m]^T \mathbb{E}[m]$ for $m = (b - \mu_g)\Sigma^{-\frac{1}{2}}$, which of course is fundamentally different. LDA does not work for equal means, while ACP does not need any mean difference. Furthermore, ACP exploits the difference in covariance, while LDA assumes equal covariance. Although both the LDA and ACP can be expressed as generalized eigenvalue problems, the solutions are different.

3 Bayes Error

It will be shown that the projection onto s is Bayes error optimal if s is a solution to (8). The Bayes error identifies the error rate of an optimal classifier, see [4]. The assumptions are that both classes are normally distributed with the same mean (concentric) and that $0 < \text{Var}[s^T g] < \text{Var}[s^T b] \forall s \neq 0$, that is, for every linear combination, the variance of the bad class is larger than the variance of the good class (which is greater than zero). To start with, two normal *scalar* distributions g and b are considered that have been standardized so that $\text{Var}[g] = 1$ and $\mathbb{E}[g] = \mathbb{E}[b] = 0$. Thus, let ξ be the variance of the bad class, $\xi > 1$.

Figure 2 depicts the PDF (Probability Density Function) of g and b (standardized with respect to g). The normal PDF with zero mean and variance σ is

$$p(x, \sigma) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \quad (18)$$

and the upper tail integral (Error function) is

$$Q(x/\sigma) = \int_x^\infty p(t, \sigma) dt. \quad (19)$$

Q is used in the calculation of the classification error probability, which given a decision boundary $\pm a$ can be summed as the grayed areas in the figure, that is

$$\varepsilon(a, \xi) = 2Q(a) + 1 - 2Q(a/\xi). \quad (20)$$

Of course, this measure of classification accuracy is valid only if an unseen observation is equally likely to be good as bad. By the same figure one realizes that the optimal decision boundaries ($\pm x$) for classification are given by the intersections between $p(x, \xi)$ and $p(x, 1)$ (they minimize $\varepsilon(x, \xi)$; by moving the boundaries at $\pm a$ in the figure the sum of the grayed areas can only become larger). Solving this equation gives the optimal boundary as

$$a(\xi) = \sqrt{\frac{\ln \xi^2}{1 - 1/\xi^2}}. \quad (21)$$

A closed expression for the Bayes error thus comes out from (20) and (21) as

$$\varepsilon(\xi) = \varepsilon(a(\xi), \xi) = 2Q\left(\sqrt{\frac{\ln \xi^2}{1 - 1/\xi^2}}\right) + 1 - 2Q\left(\sqrt{\frac{\ln \xi^2}{\xi^2 - 1}}\right). \quad (22)$$

Lemma 1

If observations are a priori equally likely to be drawn from $g \in N_1(\mu, \sigma_g)$ as they are from $b \in N_1(\mu, \sigma_b)$ and $\xi = \sigma_b/\sigma_g > 1$, $\sigma_g \neq 0$, then the Bayes error decreases monotonically with the magnitude of ξ .

proof With no change in the Bayes error, g and b scaled with the same scaling factor so that $\sigma_g = 1$ and translated so that $\mu = 0$. It will now be shown that $\varepsilon(\xi)$ as defined in (22) decreases monotonically when $\xi > 1$ increases. More specifically it will be shown that if ε is differentiated with respect to ξ , the result is negative for every $\xi > 1$. Of course $dQ(a)/da = -p(a, 1)$ and by the chain rule

$$\frac{d\varepsilon}{d\xi} = -2\frac{da}{d\xi}p(a, 1) + 2\frac{d(a/\xi)}{d\xi}p(a/\xi, 1). \quad (23)$$

Differentiating (21) gives

$$\frac{da}{d\xi} = \frac{d}{d\xi} \sqrt{\frac{\ln \xi^2}{1 - 1/\xi^2}} = \frac{\xi}{D}(\xi^2 - 1 - \ln \xi^2) \quad (24)$$

and

$$\frac{d(a/\xi)}{d\xi} = \frac{d}{d\xi} \sqrt{\frac{\ln \xi^2}{\xi^2 - 1}} = \frac{1}{D}(\xi^2 - 1 - \xi^2 \ln \xi^2). \quad (25)$$

$D = \xi(\xi^2 - 1)^{\frac{3}{4}} \sqrt{\ln \xi^2}$ is a common denominator that apparently is positive for $\xi > 1$. Finally

$$\begin{aligned} \frac{d\varepsilon}{d\xi} &= -2 \frac{da}{d\xi} \frac{1}{\sqrt{2\pi}} e^{-\frac{\ln \xi}{1-1/\xi^2}} + 2 \frac{d(a/\xi)}{d\xi} \frac{1}{\sqrt{2\pi}} e^{-\frac{\ln \xi}{\xi^2-1}} \\ &= \sqrt{\frac{2}{\pi}} \left(-\frac{da}{d\xi} \frac{\xi^{-\frac{1}{\xi^2-1}}}{\xi} + \frac{d(a/\xi)}{d\xi} \xi^{-\frac{1}{\xi^2-1}} \right) \\ &= \sqrt{\frac{2}{\pi}} \xi^{-\frac{1}{\xi^2-1}} \frac{1}{D} (\ln \xi^2 - \xi^2 \ln \xi^2) < 0 \quad \forall \xi > 1. \end{aligned}$$

□

Theorem 1

If observations are a priori equally likely to be drawn from $g \in N_n(\mu, \Sigma_g)$ as they are from $b \in N_n(\mu, \Sigma_b)$ and $0 < \text{Var}[s^T g] < \text{Var}[s^T b]$ for all non-zero $s \in \mathbb{R}^n$, then the vector

$$\hat{s} = \arg \max_s \xi(s) = \arg \max_s \frac{\text{Var}[s^T b]}{\text{Var}[s^T g]} \quad (26)$$

gives the Bayes error optimal projection to one dimension.

proof It will be shown that for two non-zero vectors s_1 and s_2 with $\xi(s_1) > \xi(s_2)$, the Bayes error is smaller in the projection onto s_1 compared to the projection onto s_2 .

It is well known that linear combinations of normally distributed variables are also normally distributed, thus $s_i^T g$ and $s_i^T b$ are normally distributed. Trivially $s_i^T \mu = \mathbb{E}[s_i^T g] = \mathbb{E}[s_i^T b]$ (in every projection, g and b have the same mean). Furthermore, $\text{Var}[s^T g] < \text{Var}[s^T b]$ why $\xi(s_i) > 1 \quad \forall s \neq 0$. Since $\xi(s_1) > \xi(s_2)$ the result follows directly from Lemma 1. □

3.1 k -dimensional Projection

The explicit calculations of Bayes error optimality in multi-dimensions will not be developed in this work. It shall be pointed out, though, that the solution to the generalized eigenvalue problem gives components that are uncorrelated with respect to both g and b , or equivalently, $E^T \Sigma_g E$ and $E^T \Sigma_b E$ are diagonal. The optimal dimensional extension to the principal eigenvector e_1 is thus $[e_2 \cdots e_k]$ if the components (linear combinations) should be uncorrelated in the sense $e_i^T \Sigma_g e_j = e_i^T \Sigma_b e_j = 0$ whenever $i \neq j$.

3.2 Non-Concentric Classes

Using the framework above, and the modified covariance explained in Section 2.2, it is necessary to show that for increasing values of $\xi = \mu^2 + \sigma^2$ the Bayes error can only decrease. Here the μ and σ denote the mean and variance of the bad class assuming the variate has been standardized so that the good class has zero mean and unit variance. Numerical experiments indicate that the Bayes error decreases monotonically whenever $\sigma^2 > \mu^2$, but this remains to be shown analytically.

4 Artificial Example

The ACP of a computer generated data set shall be studied and compared to well-known and common techniques for feature extraction, namely PCA and LDA. The data set is *not* designed to mimic real life measurements, but rather to illustrate an instance where the ACP is superior.

PCA (*Principal Component Analysis*) is possibly the most common processing method in chemometrics. It is an unsupervised technique that concentrate as much variance as possible into few uncorrelated components. With *unsupervised technique* is understood a technique that does not utilize class labels (Y). In other words, the knowledge of which observations that are good and which that are bad is not an input to PCA. LDA is a supervised technique that has been described earlier.

4.1 Artificial Data Set

The data set is originally 3-dimensional and 2-dimensional projections produced by PCA, LDA and ACP will be compared. The artificial data set describes two coaxial cylinders; the good class contained within the bad, 80 observations in each class. Of course, the best discriminative projection in this case is a radial section of the cylinders. However, to fool the PCA much variance is given the data set in the axial direction. To fool the LDA a slight mean displacement is present, this also in the axial direction. The classes are thus not concentric.

4.2 Comparison

In Figure 3 the outcome of the different methods are compared. As intended, the PCA favors the direction with much variance and the projection is thus aligned with the axes of the cylinders. In this 2-dimensional PCA-subspace, a quadratic classifier has an accuracy of 149 correctly classified observations among a total of 160 observations. Also the LDA favors the axes direction due to the mean difference. The classification accuracy for the LDA is 145/160. The ACP is more or less a radial section that very well concentrates the good class in the center, and spreads the bad class around it. The classification accuracy is 160/160.

5 Experimental Data

We study a data set obtained from measurements on 204 grain samples. A human test panel classifies each grain sample as *good* or *bad*. An *electronic nose* with 23 response variables measures on the same samples.

Thus, for every observation we have 23 variables and the knowledge whether it is attributable to a *good* or a *bad* grain sample. The entities μ_g , μ_b , Σ_g and Σ_b are unknown and have to be estimated from the data set itself. We now want to find out if the sensor configuration in the electronic nose can be trained to make a distinction between good and bad similar to the one produced by the human test panel. We also want to compare the feature extraction of PCA, LDA and ACP.

Figure 3: The artificial data processed with PCA, LDA and ACP. Rings are *good* and crosses *bad*. The PCA selects the components with maximum variance, and thus loses the most relevant information needed to distinguish bad samples from good. The LDA does the same due to the mean difference direction, which in this data set is not optimal for discrimination. The ACP selects the cylinder radial section, which obviously is the best choice.

5.1 Validation

Since the means and covariances have to be estimated from the data set itself, random estimation/validation partitionings will be used to gain reliable results. The means, covariances and classification models are thus estimated on 75% of the observations, and the displayed projection and classification accuracy are due to the remaining 25%. We denote the estimation sets T_{good} and T_{bad} , respectively, and the validation sets V_{good} and V_{bad} , respectively. The sets are described by matrices, where the columns are observations. For instance,

$$T_{\text{good}} = [t_1^{\text{good}} \ t_2^{\text{good}} \ \dots \ t_q^{\text{good}}]$$

for the good class training observations, where $t_i^{\text{good}} \in \mathbb{R}^{23}$. $q = 76$ is the number of observations in the good class estimation set. $p = 77$ is the number of observations in the bad class estimation set.

5.2 Covariance Estimation

The data set means m_{good} and m_{bad} are simply the arithmetic means of the respective training set. The data set covariance and modified covariance matrices Σ_{good} and $\tilde{\Sigma}_{\text{bad}}$ are estimated as

$$C_{\text{good}} = \frac{1}{q-1} (T_{\text{good}} - m_{\text{good}} \mathbf{1}^T) (T_{\text{good}} - m_{\text{good}} \mathbf{1}^T)^T$$

$$\tilde{C}_{\text{bad}} = \frac{1}{p-1} (T_{\text{bad}} - m_{\text{good}} \mathbf{1}^T) (T_{\text{bad}} - m_{\text{good}} \mathbf{1}^T)^T,$$

where $\mathbf{1}^T = [1 \ 1 \ 1 \ \dots \ 1]$.

5.3 Projection Calculation

The optimal k -dimensional projection $S \in \mathbb{R}^{k \times 23}$ with respect to the quality measure $\text{trace} [C_{\text{good}}^{-1} \tilde{C}_{\text{bad}}]$ is calculated as (13), where e_i are the principal eigenvectors of the generalized eigenvalue problem $\tilde{C}_{\text{bad}} E = C_{\text{good}} E D$. The validation data sets V_{good} and V_{bad} are projected as $V_{\text{good}}^S = S V_{\text{good}}$ and $V_{\text{bad}}^S = S V_{\text{bad}}$. They are plotted in Figure 6. See [13] for numerically stable algorithms for solving the generalized eigenvalue problem.

5.4 Plots

For a particular estimation/validation partitioning of the data set, scatter plots of the 2-dimensional projection of LDA and ACP are studied. As a reference, the plot of a PCA is depicted in Figure 4. It is seen, that in this subspace the accurate detection of bad samples is almost impossible. Comparing the plots of the LDA in Figure 5 to the ACP in Figure 6, one can see that the data have fundamentally different structure. LDA tries to find two separated clusters of the good and bad class, while the ACP centers around the good class, and tries to spread the bad observations as much as possible. It is also seen that in both the ACP and the LDA subspace, distinction between good and bad can be done, although not very accurate.

Figure 4: PCA of the grain validation data set.

Figure 5: LDA of the grain validation data set.

Figure 6: ACP of the grain validation data set.

5.5 Results

As a quality measure of a projection is used the classification accuracy of a quadratic classifier. The quadratic classifier is known to give the highest (expected) classification accuracy for normal distributions (with known means and covariance matrices), see [9].

The classification accuracy in the original 23-dimensional feature space as well as the accuracy in the 2 and 4-dimensional subspaces from PCA, LDA and ACP is given in Table 1. The figures are based on 100 random estimation/validation partitionings of the available data set, and for each partitioning, the subspace and classification model are calculated from estimation data, and the number of correctly classified observations of the 51 observations in the validation data set is counted.

As seen in the table, classification in the PCA subspace is not very much better than the flipping of a fair coin. Among the 2-dimensional subspaces, the LDA gives highest accuracy, and the ACP almost equally good. The opposite holds for the 4-dimensional subspace. The classification accuracy in the unreduced 23 dimensions is 41 ± 2.9 .

For this data set, the performance of LDA and ACP is thus about equally good. This is because there is sufficient mean difference between the good and the bad class for the LDA to operate well. The Mahalanobis distance is estimated to about 2 standard deviations, and this distance gives a theoretical classification accuracy of a *separating plane* of 85% or 43/51 (*a priori* equally likely normal distributions with equal covariance matrices assumed). The grain data thus have a structure that is not the best for the ACP.

The EN can (in conjunction with the algorithms mentioned above) only up

Table 1: Number of correctly classified observations out of 51 possible in different k -dimensional subspaces. The figures are the means $\pm$ standard deviations for 100 random estimation/validation partitionings of the dataset.

k	PCA	LDA	ACP
2	28 ± 3.3	38 ± 3.2	37 ± 3.1
4	29 ± 3.3	37 ± 2.9	38 ± 2.6

to a degree make the same distinction between good and bad as the human test panel. About 15% of the samples will be classified differently (if bad samples are as likely as good samples).

6 Conclusions

A method to find subspaces for asymmetric classification problems, *the Asymmetric Class Projection* (ACP), was introduced and compared to the well-known *Linear Discriminant Analysis* (LDA). The ACP has its main benefits when two distributions are nearly concentric and unequal in covariance. It was shown that for the concentric case (equal distribution means), the ACP is optimal for at least 1-dimensional projections. The LDA cannot be used to analyze concentric distributions at all. An artificial data set showed an instance where one can expect the ACP to be beneficial to use. However, tested on a real data set, the LDA performed equally good as the ACP. It was found that this was probably due to the assumption of near concentricity did not hold. The data set came from an electronic nose that should detect bad grain samples.

Acknowledgments

The ACP is a result of research at the *Division of Automatic Control* and the *Swedish Sensor Center* (S-SENCE). This work was partly sponsored by the *Vetenskapsrådet* (*Swedish Research Council*) and the *Swedish Foundation for Strategic Research* (the latter via the graduate school *Forum Scientum*). The contributions are gratefully acknowledged. We are also grateful to *AppliedSensor AB* for contributing with the grain dataset. The ACP algorithm (originally the GBP algorithm) as a calibration method for sensor systems is filed for a patent.

References

- [1] H. Brunzell. *Signal Processing Techniques for Detection of Buried Landmines using Ground Penetrating Radar*. PhD thesis, Chalmers University of Technology, 1998.
- [2] A. J. Burnham, J. F. MacGregor, and R. Viveros. A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *Journal of Chemometrics*, 13:49–65, 1999.
- [3] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

- [4] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition, 1990.
- [5] A. R. Gallant. *Nonlinear Statistical Models*. John Wiley & Sons, 1987.
- [6] G. H. Golub. *Matrix Computations, 3 ed*. Baltimore : Johns Hopkins Univ. Press, 1996.
- [7] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press Baltimore, MD., 2 edition, 1989.
- [8] R. Gutierrez-Osuna. Pattern analysis for machine olfaction: A review. *IEEE Sensors Journal*, 2(3):189–201, 2002.
- [9] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, Upper Saddle River, New Jersey, 4th edition, 1998.
- [10] R. Lotlikar and R. Kothari. Adaptive linear dimensionality reduction for classification. *Pattern Recognition*, 33:185–194, 2000.
- [11] H. Martens and T. Näs. *Multivariate Calibration*. John Wiley & Sons, Chichester, 1989.
- [12] M. Pardo and G. Sberveglieri. Learning from data: A tutorial with emphasis on modern pattern recognition methods. *IEEE Sensors Journal*, 2(3):203–217, 2002.
- [13] H. Park, M. Jeon, and P. Howland. Cluster structure preserving reduction based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, to appear, 2002.
- [14] D. Di Ruscio. A weighted view on the partial least-squares algorithm. *Automatica*, 36:831–850, 2000.
- [15] B. A. Snopok and I. V. Kruglenko. Multisensor systems for chemical analysis: state-of-the-art in electronic nose technology and new trends in machine olfaction. *Thin Solid Films*, (418):21–41, 2002.
- [16] P. Stoica and M. Viberg. Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions. *IEEE Transactions on Signal Processing*, 44(12):3069–3078, 1996.
- [17] S. Wold, N. Kettaneh-Wold, and B. Skagerberg. Nonlinear PLS-modeling. *Chemometrics and Intelligent Laboratory Systems*, 7:53–65, 1989.
- [18] S. Wold, H. Martens, and H. Wold. *The Multivariate Calibration Problem in Chemistry Solved by the PLS Method*. Springer Verlag, Heidelberg, 1983.
- [19] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn. The collinearity problem in linear regression. The partial least squares approach to generalized inversed. *SIAM J. Sci. Stat. Comput.*, 5:735–743, 1984.