Chapter 4

CC-ICA Feature Selection

4.1 Introduction

The problem of feature selection for classification can be stated as, given a set of features used to represent our data, select a subset of these features such that working with the reduced set proves advantageous for our task. Selecting a feature subset can be seen as a linear transform. If the original number of features is D and we select a subset of M features, this is equivalent to multiplying the data with an $M \times D$ matrix with its rows formed by D-dimensional canonical vectors with the M ones pointing to each of the selected features. As with any linear transformation, the Bayes error of the resulting dataset will be, in the best case, equal to the original Bayes error [83]: if class-conditional densities can be accurately estimated in domain space, feature selection will negatively affect classification. Of course, this situation is unfrequent and if a *good* feature subset is chosen this fact, together with the dimensionality reduction, can benefit estimation and thus, classification.

First of all we must define what a *good* feature subset is. This requires a criterion stating the goodness of any subset. Once a decision has been made over the criterion, selecting the best subset might imply evaluating it on all available subsets, a problem that grows combinatorially on the dimension. Probably being this the most delicate issue concerning feature selection. One way of avoiding the exhaustive search for feature subsets is to use iterative algorithms that successively reduce search space. Examples of these algorithms are *backward* and *forward* selection which do not guarantee the selection of the best subset [78], or branch and bound algorithms which optimize combinatorial problems and do guarantee a global optimum [106, 48, 140]. Another possibility is to use a criterion that does not require an exhaustive search at all. While this converges to the global optimum, such criterions are difficult to find.

In classification, feature selection criterions are based on class separability and, sometimes, directly on classification accuracy. The first approach states that a good feature subset should preserve or enhance the separation among classes, the second one simply uses as feature selection criterion the objective of the feature selection. Using classification accuracy as criterion is a very delicate issue since much care has to be taken in order to give statistical significance to the results. Among all, it requires careful training and test set selection procedures to avoid overlearning and allow generalization. A major drawback with this criterion is that in most cases an exhaustive search is required. Moreover, each feature subset also requires costly and completely new calculations for each possible subset. On the other hand, class separability measures are generally simpler and, in some cases as ours, can be linked with the classifier of choice. These measures can be made from straightforward metric considerations such as the distance between class elements or using more complex statistical or even information theoretic approaches. Statistical class separability measures take into account the distance among distributions. Since distribution estimation is not always possible, parametric approximations (for instance, assuming the classes are Gaussian) are frequently introduced. Some of the most used class separability measures will be enumerated in the next section. We will then focus on the measure of *divergence*, which can be interpreted from both a statistical and information theoretic approach. Divergence has a strong theoretical basis and does not make any assumption on the class-conditional distributions which also turns out to be its main drawback, since accuracy of divergence is related with the accuracy of the conditional density estimation. But we have just presented an algorithm which greatly simplifies this estimation by allowing the independence assumption to hold, so densities can be estimated in projected space. Divergence can be calculated using this, hopefully accurate, approximation. More important, we will show through a simple property of order that if conditional independence is true, the class separability measure of divergence does not require an exhaustive search for selecting the best feature subset of any cardinality.

CC-ICA has a particularity that should be contemplated when focusing on feature selection. Since representations are class dependent, the feature subsets will also depend on the class label. Feature, for instance, number 1 on the ICA representation corresponding to a certain class, has no relationship with the same feature number in any of the other representations. This can give place to situations in which certain features are good for separating their corresponding class from the others, and completely different features are necessary for other classes. Fortunately, by interpreting divergence in terms of expected log-likelihood ratios we can formulate a divergence-based criterion good for evaluating local features and obtaining class-dependent optimal feature subsets. So divergence can be naturally be introduced within the CC-ICA framework.

The point now is why should we select features within CC-ICA if, as we hold until now, conditional density estimation is quite accurate thanks to the independence assumption. Reducing dimensionality preserving discriminability can give us more than improved accuracy. The computational load of our CC-ICA scheme is high since we need to perform as many projections as classes there are, before evaluating all the marginal densities. Strongly reducing the number of features will very likely improve the speed of our algorithm. So there are cases in which we will accept to sacrifice small variations in the accuracy for a greater speed and efficiency. The experiments will show that, being CC-ICA with naive Bayes a Bayesian classifier, it is very difficult to select features and improve the accuracy, though this happens in more than one case. But they will also show that the accuracy is preserved for even feature subsets with very low cardinality. In some cases, it is enough to keep a couple of filters from each of the class-conditional projections matrices to obtain the same accuracy as using the whole CC-ICA spaces. This turns CC-ICA into a very fast method.

4.2 Class Separability Measures

Class separability measures (noted by $\mathcal{S}$) provide simple and efficient criterions for measuring the goodness of feature subsets. In this section we will expose some of the most spread out measures, and consider their expression for the two class case ($K = 2$). The most frequently used feature selection criterions, for their simplicity, are those based on statistics of the distributions instead of the distributions themselves. In almost all cases, only up to second order statistics are used. A very simple criterion is to use the Mahalanobis distance (section 1.2) between the two class means, $\mathcal{S} = (\boldsymbol{\mu}^1 - \boldsymbol{\mu}^2)^T \Sigma^{-1} (\boldsymbol{\mu}^1 - \boldsymbol{\mu}^2)$ with Σ the common covariance matrix. We mentioned that class separability measures can be frequently linked to the classifier. In this case, under Gaussian class-conditional densities, the probability of error is inversely proportional to the Mahalanobis distance [38], so we observe that this criterion ensures optimal features for the quadratic classifier.

The Mahalanobis distance only takes into account the global distribution and a rough approximation of the intra-class distances through the difference between the means. Slightly more complex approaches consider intra- and extra-class properties: desirable features should have compactly represented and widely separated classes. Most of these techniques result from variations of the classical Fisher ratio [41, 38]. In its most common version it is defined in terms of within- and between-class scatter matrices (Σ_b and Σ_w). The within class scatter matrix can be defined as the weighted sum of the class covariance matrices, where the weights correspond to the class priors. The between class scatter matrix can be defined as the scatter of the class mean around the common mean, also weighted with the prior class probabilities. The Fisher type class separability measures are then statistics obtained from these matrices. For instance, $\mathcal{S} = trace(\Sigma_w^{-1}\Sigma_b)$. For details and variations of this approach refer to [48]. In the next chapter, we will also observe that this criterion has been more spread out as an objective function for extracting features than for selecting features: instead of selecting feature subsets that maximize this measure of class separability, use it to analytically find the linear transform such that projected features optimize the criterion.

Another criterion is the *Bhattacharyya distance* or *bound* [73], a popular measure of similarity between two distributions which arises from the problem of finding an upper bound for the Bayes error for normally distributed classes,

$$\begin{aligned}\mathcal{S}_{\mathcal{B}} = & \frac{1}{2} \log \frac{\left|\frac{\Sigma^1+\Sigma^2}{2}\right|}{\sqrt{|\Sigma^1|}\sqrt{|\Sigma^2|}} \\ & + \frac{1}{8} (\boldsymbol{\mu}^1 - \boldsymbol{\mu}^2)^T \left(\frac{\Sigma^1 + \Sigma^2}{2}\right)^{-1} (\boldsymbol{\mu}^1 - \boldsymbol{\mu}^2)\end{aligned} \tag{4.1}$$

where $\boldsymbol{\mu}^k$ and Σ^k correspond to the class mean and covariance with $k = 1, 2$. Notice that the second term of this equation considers the distance between the class means in a Mahalanobis fashion, while the first term only involves second order information.

The exposed criterions all have the same inconvenient we exposed for PCA, in section (2.2.1), they are *blind* beyond second order statistics. While it might not be true to say they assume conditional Gaussian distributions, it is certainly true that they are unable to tell the difference. Attempts have been made to generalize these criterion to more general class-conditional distributions. For instance, a generalized Fisher ratio can be obtained by considering Gaussian mixtures to model the conditional densities [80]. If the above approaches could be regarded as parametric criterions, this last one would clearly be a semiparametric criterion. A nonparametric and more general alternative is to consider objective functions which make direct use of the conditional densities and charge the complexity to the density estimation method. These approaches directly measure the distances between a pair of distributions.

The Jeffries-Matusita (JM) distance between a pair of probability distributions [158] is defined as,

$$\mathcal{S}_{\mathcal{JM}} = \int_{\boldsymbol{x}} \big(\sqrt{p(\boldsymbol{x}|C^1)} - \sqrt{p(\boldsymbol{x}|C^2)}\big)^2 d\boldsymbol{x} \tag{4.2}$$

and can be seen as a measure of the average distance between the two class densities. For normally distributed classes, the JM distance becomes

$$\mathcal{S}_{\mathcal{JM}} = 2(1 - e^{-\mathcal{S}_{\mathcal{B}}}) \tag{4.3}$$

so it is directly related with the Bhattacharyya bound. Moreover, we observe an exponentially decreasing weight as the class separation increases. This is because the JM distance is asymptotic to 2. This saturating behaviour is a highly desirable property because a value of 2 for the JM distance already ensures 100% classification accuracy, a consequence of the fact that (4.2) is equivalent to

$$\mathcal{S}_{\mathcal{JM}} = 2\big(1 - \int_{\boldsymbol{x}} \sqrt{p(\boldsymbol{x}|C^1)p(\boldsymbol{x}|C^2)}d\boldsymbol{x}\big) \tag{4.4}$$

Divergence is another measure of distribution separability that has its basis in their degree of overlap. Divergence is based on the Kullback-Leibler (2.46) divergence between the class-conditional distributions. To avoid confusion with the version of divergence we are now introducing we will refer to (2.46) as the *Kullback-Leibler distance.* Adapted to class-conditional densities, this distance can be expressed as,

$$\mathcal{KL}(C^1, C^2) = \int_{\Omega} p(\boldsymbol{x}|C^1) \log \frac{p(\boldsymbol{x}|C^1)}{p(\boldsymbol{x}|C^2)} d\boldsymbol{x} \tag{4.5}$$

The asymmetry of Kullback-Leibler motivates the symmetric measure of divergence, long ago used for feature selection [96], defined as

$$\mathcal{D}(C^1, C^2) = \mathcal{KL}(C^1, C^2) + \mathcal{KL}(C^2, C^1) \tag{4.6}$$

Besides being symmetric, divergence is zero between a distribution and itself, always positive, monotonic on the number of features and, for the monotonic two class case, divergence provides an upper bound for the classification error [73] since,

$$\varepsilon > \frac{1}{8} e^{-\mathcal{D}/2}$$

The two main drawbacks of divergence are that it requires density estimation and has a nonlinear relationship with classification accuracy. The first drawback is specially problematic in high dimensional spaces, precisely where we would require feature selection techniques. The second one is related with the fact divergence increases without bound as class separability increases. Swain and Davis [146] heuristically solved this inconvenient by introducing the following transformed divergence,

$$\hat{\mathcal{D}}(C^1, C^2) = 2[1 - \exp(-\frac{\mathcal{D}ij}{8})] \tag{4.7}$$

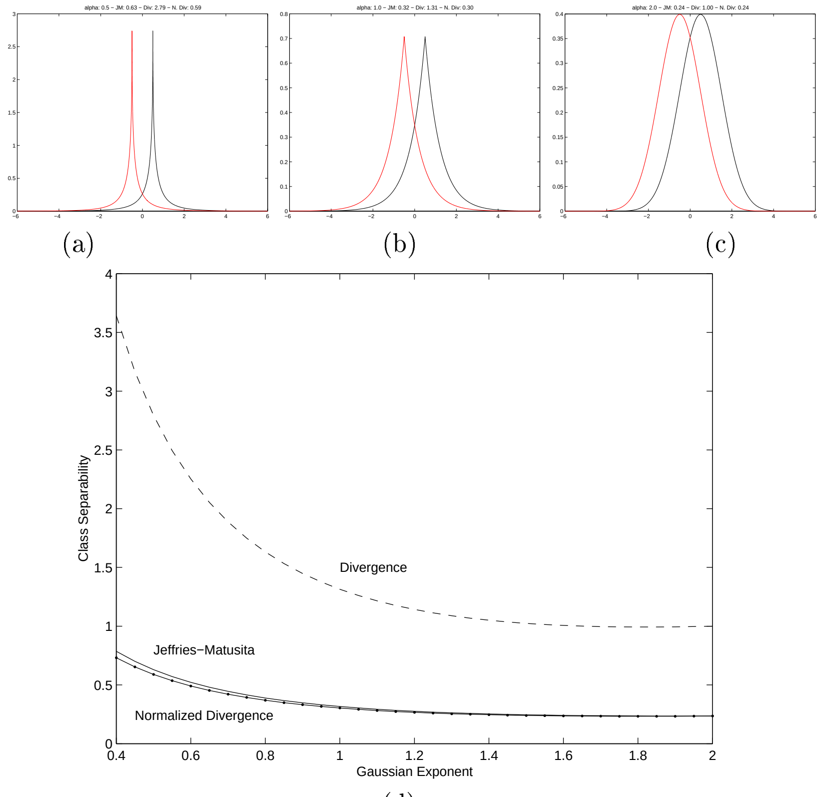
As the JM distance, transformed divergence has a saturating behaviour asymptotic to 2. It has also been shown that transformed divergence is computationally more economical, comparably as effective as the JM distance and considerably better than simple divergence or than the Bhattacharyya distance [146, 147, 99].

Divergence, as the JM distance, also has an analytic expression for Gaussian classes. This expression shares with the Bhattacharyya bound the property of having one term which only involves covariances and a second term which is the square of a covariance normalized distance between the means of the distributions,

$$\begin{aligned}\mathcal{D}(C^1, C^2) = {} & \frac{1}{2} trace\big((\Sigma^1 - \Sigma^2)({\Sigma^2}^{-1} - {\Sigma^1}^{-1})\big) \\ & + \frac{1}{2} trace\big(({\Sigma^1}^{-1} + {\Sigma^2}^{-1})(\boldsymbol{\mu}^1 - \boldsymbol{\mu}^2)(\boldsymbol{\mu}^1 - \boldsymbol{\mu}^2)^T\big)\end{aligned} \tag{4.8}$$

An additional property of divergence, straightforward to see from (4.8), is invariance for invertible linear transformations. So if we should find an invertible transformation that simplifies the density estimation, calculating the divergence in transformed space is equivalent to the calculation in domain space.

Since sparse data is of special interest for our work, we have chosen to illustrate the properties of divergence through an example with this kind of data. For this purpose, we considered two classes represented by a single feature with the same generalized Gaussian distribution (2.41)on each class, except for the means, located in -0.5 and 0.5 respectively. The standard deviation of the conditional densities was fixed to 1, such that the only free parameter is the Gaussian exponent α. We made this exponent take values ranging from 0.4 (highly supergaussian distribution) to 2 (Gaussian distribution). Then we calculated the JM distance, the divergence, and the normalized divergence between these two distributions as a function of the exponent. Results are exposed in fig. (4.1). In (4.1.a), (4.1.b) and (4.1.c) we can see the artificial situation for three different choices of α: 0.5, 1 (Laplacian densities) and 2 (Gaussian densities). From these figures we can already notice that as the value of α approximates 2, the degree of overlap between the distributions increases, so the distribution distances decrease. The overlap is not as large as the figures suggest at a first glance, since the y-axis is different for each figure. In fig. (4.1.d) the values of the class-separability measures are plotted as a function of α. Here we confirm that class separability is inversely proportional to the exponent value. We can also observe the unbounded nature of divergence, and the close relationship between normalized divergence and the JM distance. Notice that, for this particular problem, any class separability measure solely based on second order statistics of the classes would be useless since these statistics are constant for all chosen exponents.



(a) (b) (c)

(d)

Figure 4.1: Class separability measures for two unidimensional classes with super-gaussian distributions. In (a), (b) and (c) the distribution for three sample Gaussian exponents: 0.5, 1 and 2. In (d) the separability measures as a function of the Gaussian exponent.

4.2.1 Divergence and Independence

For simplicity, we will note divergence between classes C^i and C^j as $\mathcal{D}^{ij} \overset{def}{=} \mathcal{D}(C^i, C^j)$ and marginal divergences will be noted by $\mathcal{D}_d^{ij}$ with $d = 1, \ldots, D$, being D the data dimensionality. The most important consequence upon divergence derived from assuming class-conditional independence is that this expression results additive in the features,

$$\mathcal{D}^{12} = \sum_{d=1}^{D} \mathcal{D}_d^{12} \tag{4.9}$$

Since divergence is defined as a linear combination of terms of the form $A(\boldsymbol{x}) = \int f(\boldsymbol{x}) \log g(\boldsymbol{x}) d\boldsymbol{x}$ with f and g probabilities, proving (4.9) is equivalent to proving that it holds for A. Also, we restrict ourselves to $D = 2$ noting $\boldsymbol{x} = (x, y)$ since extension to more dimensions is straightforward. By using the definition of independence

(3.3) and the fact the marginal probabilities integrate to 1, we have

$$\begin{aligned}
A(x,y) &= \int_{xy} f(x,y)\log g(x,y)\\
&= \int f_x(x)f_y(y)\log g_x(x)g_y(y)\\
&= \int_{xy} f_x(x)f_y(y)\big(\log g_x(x) + \log g_y(y)\big)\\
&= \int_x f_x(x)\int_y f_y(y)\log g_y(y)dy + \int_y f_y(y)\int_x f_x(x)\log g_x(x)\\
&= \int_y f_y(y)\log g_y(y)dy + \int_x f_x(x)\log g_x(x)\\
&= A(x) + A(y).
\end{aligned}$$

This additivity property does not hold for transformed divergence so, if we choose to work with this alternative expression, we should take care of calculating it after the marginal divergences have been added or directly using the fact that in this case

$$\hat{\mathcal{D}}^{12} = 2(1 - \prod_{d=1}^{D} e^{-\frac{\mathcal{D}_d^{12}}{8}}). \tag{4.10}$$

From (4.9) and the fact divergence is nonnegative it is clear that divergence increases with dimensionality. It is also straightforward to calculate the divergence of any feature subset $S \subseteq \{1, ..., D\}$, which we note by $\mathcal{D}_S^{12}$ since we simply need to add the marginal divergences pointed out by S. We can also observe the following property of monotonicity in divergence

$$(d_1 \notin S, d_2 \notin S) \wedge (\mathcal{D}_{d_1}^{12} \leq \mathcal{D}_{d_2}^{12}) \Rightarrow (\mathcal{D}_{S\bigcup d_1}^{12} \leq \mathcal{D}_{S\bigcup d_1}^{12}). \tag{4.11}$$

This property of order suggests that, at least for the two class case, the global best feature subset is the one that contains the features with maximum marginal divergence, and thus provides a very simple rule for feature selection using the divergence criterion under the assumption of conditional independence without involving any search procedure: given a feature subset size, preserve only those features with maximal marginal divergence.

4.2.2 Divergence in the Multiclass Case

Although, divergence only provides a measure for the distance between two classes there are several ways of extending it to the multiclass case, providing an effective feature selection criterion. The most common is taking the average over all class pairs

$$\mathcal{D}_d^A = \frac{2}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\mathcal{D}_d^{ij}. \tag{4.12}$$

$\mathcal{D}_d^A$ represents the average divergence present in feature d, with $d = 1, \ldots, D$. This approach (noted by D^A) is simple and preserves the exposed property of order for feature subsets, but it is not reliable as the variance of the pairwise divergences increases.

A more robust approach is to sort features by their maximum minimum (two-class) divergence,

$$\mathcal{D}_d^M = \max_{i=1\ldots K} \min_{j\neq i} \mathcal{D}_d^{ij} \tag{4.13}$$

This works fine when the number of features we wish to select is small, but soon decays as the size of the subset increases: sorting features by maximum minimum divergence is a very conservative election.

We introduce an additional criterion, that attempts to rank the features maximizing the divergence in all classes, minimizing the influence of already well separated classes over the choice of features. To express this in a clearer way, if a certain subset of features already separates certain class or classes, why should these classes influence the choice of subsequent features? Let Ψ^{ij} be an ordering of $\{1, ..., D\}$ in descending order of $\mathcal{D}_d^{ij}$, and note by $\boldsymbol{\Psi}_1 = \Psi_1^{ij}$ the matrix where each element contains the features that best discriminate C^i from C^j. We first select those features within $\boldsymbol{\Psi}_1$ sorted by frequency in descending order. The first features to appear in this listing are those that best discriminate the larger amount of classes. If the number of features present in this ordering is larger or equal than the size of the subset we are searching for, we stop. This will not be the general case since it is common to find single features good for separating many classes. If this is the case, we place those features within $\boldsymbol{\Psi}_2$ (second best discriminative features) that have not been already included, also in descending frequency order. We repeat this process on $\boldsymbol{\Psi}_j$ until all features are ranked. By this simple algorithm we ensure that all two-class separations are considered, and the best discriminant features placed at the beginning of the rank. Also, features adequate for separating several classes are emphasized. We will call this criterion $\mathcal{D}^X$ and in many cases it outperforms $\mathcal{D}^A$ and $\mathcal{D}^M$.

4.3 Divergence and CC-ICA

If we are able to ensure class-conditional independence of our data then divergence, as exposed in the last section, should prove an effective criterion for feature selection. Usually this is not the case so, as we have seen in the last chapter, one option is to take advantage of the CC-ICA representation. In this case we have K linear representations, one per class, each one making the variables in random vector $\boldsymbol{x}|C_k$ independent. With this class-conditional approach, it is not possible, or at least not optimal, to select the same feature subset for all classes. We are forced to consider each class in turn in order to determine which feature subset best separates this class from the others. An insight on the definition of divergence will allow us to adapt this concept to this class-conditional situation

The *log-likelihood ratio* ($\mathcal{L}$) is defined as,

$$\mathcal{L}^{ij}(\boldsymbol{x}) = \log p(\boldsymbol{x}|C^i) - \log p(\boldsymbol{x}|C^j) \tag{4.14}$$

$L_{ij}(\boldsymbol{x})$ is not symmetric and punctually measures the overlap of the class-conditional densities in $\boldsymbol{x}$. Notice that the Kullback-Leibler distance is nothing but the class-conditional expectation of the log-likelihood ratio so (4.6) can be rewritten as,

$$\mathcal{D}^{ij} = E^i\{\mathcal{L}^{ij}(\boldsymbol{x})\} + E^j\{\mathcal{L}^{ji}(\boldsymbol{x})\}, \tag{4.15}$$

where E^i is the expectancy operator conditional on class C^i. We can replace by (4.14) and rearrange the terms in this last equation to obtain,

$$\begin{aligned}\mathcal{D}^{ij} =& E^i\{\log p(\boldsymbol{x}|C^i)\} - E^j\{\log p(\boldsymbol{x}|C^i)\} + \\ & E^j\{\log p(\boldsymbol{x}|C^j)\} - E^i\{\log p(\boldsymbol{x}|C^j)\}\end{aligned} \tag{4.16}$$

We can now replace the estimate for the logarithm of the class conditional probabilities provided by the CC-ICA model (3.12),

$$\begin{aligned}\mathcal{D}^{ij} =& E^i\{\sum_{m=1}^{M^i} \log p^i({\boldsymbol{w}_m^i}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^i)) + \nu^i\} - E^j\{\sum_{m=1}^{M^i} \log p^i({\boldsymbol{w}_m^i}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^i)) + \nu^i\} \\ &+ E^j\{\sum_{m=1}^{M^j} \log p^j({\boldsymbol{w}_m^j}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^j)) + \nu^j\} - E^i\{\sum_{m=1}^{M^j} \log p^j({\boldsymbol{w}_m^j}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^j)) + \nu^j\}\end{aligned} \tag{4.17}$$

Where $\boldsymbol{w}_m^k$ is the m-th row of the filter matrix learnt for class C^k, and $\overline{\boldsymbol{x}}^k$ the estimated class mean. Normalization constants are cancelled and the sum can be taken out of the expectation operators,

$$\begin{aligned}\mathcal{D}^{ij} =& \sum_{m=1}^{M^i}\Big(E^i\{\log p^i({\boldsymbol{w}_m^i}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^i))\} - E^j\{\log p^i({\boldsymbol{w}_m^i}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^i))\}\Big) + \\ & \sum_{m=1}^{M^j}\Big(E^j\{\log p^j({\boldsymbol{w}_m^j}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^j))\} - E^i\{\log p^j({\boldsymbol{w}_m^j}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^j))\}\Big)\end{aligned} \tag{4.18}$$

This equation can be simplified defining

$$\mathcal{B}_m^{ij} = \Big(E^i\{\log p^i({\boldsymbol{w}_m^i}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^i))\} - E^j\{\log p^i({\boldsymbol{w}_m^i}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^i))\}\Big) \tag{4.19}$$

such that

$$\mathcal{D}^{ij} = \sum_{m=1}^{M^i}\mathcal{B}_m^{ij} + \sum_{m=1}^{M^j}\mathcal{B}_m^{ji} \tag{4.20}$$

Observe that $\mathcal{B}_m^{ij}$ is the result of projecting all available samples from C^i and C^j into the m-th filter of the class-conditional representation obtained for C^i and then subtracting their respective expected log-likelihood values. So $\mathcal{B}_m^{ij}$ measures the separability of the m-th independent component of the representation obtained for C^i, between classes C^i and C^j. We have reexpressed the divergence between two classes in terms of two summands such that each summand only affects projections on one of the class-conditional representations. Divergence is maximized by maximizing both terms in (4.20) and each of the terms is maximized by selecting adequate features for its corresponding representation. Maximization of each of the terms, independently, as with divergence under class-conditional independence, is the result of preserving only those features with maximum marginal divergences. Also maximization of each of the terms will provide different feature subsets on each class representation, meaning that, while certain features might be appropriate for separating class C^i from class C^j in the i^{th} representation, possibly distinct features will separate class C^j from class C^i in the j^{th} representation. In practice, the empirical expectation operator can be used. So we can use the following approximation

$$E^i\{\log p^j({\boldsymbol{w}_m^j}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^j))\} \approx \frac{1}{\#C_i}\sum_{x\in C_i}\log p^j({\boldsymbol{w}_m^j}^T(\boldsymbol{x}-\overline{\boldsymbol{x}}^j)). \tag{4.21}$$

From this scheme we have obtained a divergence-based measure of discriminability between two classes for each class-conditional component. As with divergence, we have to decide how we combine these pairwise separability measures into a single measure of separability for this component. This can be done in exactly the same way as with divergence but for the value $\mathcal{B}_m^{ij}$. That is, using the average ($\mathcal{B}^A$), the maximum minimum ($\mathcal{B}^M$) or the proposed criterion ($\mathcal{B}^X$).

4.3.1 CC-ICA Feature Selection Algorithm

Table (4.1) includes the results obtained in the last section in a learning algorithm (CC-ICA-FS) that, given a fixed number of features we wish to select, outputs the most discriminative features for each class. This algorithm should be applied after training our CC-ICA representation (CC-ICA-Train). After obtaining the set of features for each of the representations, CC-ICA-Test should be applied using only the selected features.

4.4 Experiments

As in the last chapter, experiments were performed over artificial, benchmark and real-world data.

4.4.1 Artificial Data

A first experiment is performed on the artificial two-class example Trunk used to illustrate the curse of dimensionality [151]. In this experiment, Trunk considered two classes in a D-dimensional space with class-conditional densities,

$$p(\boldsymbol{x}|C^1) \sim N(\boldsymbol{\mu},\boldsymbol{I}) \ , \ p(\boldsymbol{x}|C^2) \sim N(-\boldsymbol{\mu},\boldsymbol{I}) \tag{4.22}$$

where

$$\mu = \left[\frac{1}{\sqrt{1}},\frac{1}{\sqrt{2}},\ldots,\frac{1}{\sqrt{D}}\right]^T \tag{4.23}$$

For this data, Trunk showed that, if the mean vector $\boldsymbol{\mu}$ is known in advance the probability of classification error decreases while dimensionality increases, and if $\boldsymbol{\mu}$

```
1. given $F$, the number of features we wish to select per class
2. for each class $i = 1 \dots K$
    (a) for each class $j = 1 \dots K$, $j \neq i$
        i. for each $m = 1 \dots M^i$
            A. Obtain the $m$-th component value for the $n = 1, \dots, N^i$ available samples in $C^i$
                $x_n = {\boldsymbol{w}_m^i}^T (\boldsymbol{x}_n - \overline{\boldsymbol{x}}^i)$
            B. Obtain the $m$-th component value for the $n = 1, \dots, N^j$ available samples in $C^j$
                $y_n = {\boldsymbol{w}_m^i}^T (\boldsymbol{x}_n - \overline{\boldsymbol{x}}^i)$
            C. Calculate $\mathcal{D}_m^{ij}$, using the empirical expectation and the marginal probabilities $p_m^k$ learnt in
               CC-ICA-Train,
                $\mathcal{D}_m^{ij} = \frac{1}{N^i} \sum_{n=1}^{N^i} \log p_m^k(x_n) - \frac{1}{N^j} \sum_{n=1}^{N^j} \log p_m^k(y_n)$
            D. end loop
        ii. end loop
    (b) Generalize the pairwise class separability measures to multiclass separability measures using $\mathcal{D}^A$, $\mathcal{D}^M$ or
        $\mathcal{D}^X$. For instance, if we use $\mathcal{D}^A$ we have
            $\mathcal{D}_m^i = \frac{1}{K-1} \sum_{j \neq i}^{K} \mathcal{D}_m^{ij}$
    (c) Select the $F$ features with the highest values of $\mathcal{D}_m^i$. These features are the $F$ most discriminant features
        for the $i$-th representation (class $C^i$) of the CC-ICA model. Forming set DiscFeat$^i = \{m_1, \dots, m_F\}$
    (d) end loop
3. end loop
```

Table 4.1: Class-conditional ICA feature selection training algorithm (CC-ICA-FS).

is unknown and estimated from a fixed number of samples the probability of error shows a peaking behaviour converging to the worst case error $1/2$. From a different perspective, in a recent survey on feature selection [66], Jain and Zongker propose this example to investigate the quality of certain feature subsets considering that for this dataset the optimal F-feature subset is known in advance: since it has equal dispersion on every direction and the distance from the means decreases for each consecutive dimension, the first F features are always the best subset. Their research was aimed towards evaluating the effect of training set size on feature selection, so dimensionality was fixed to $D = 20$. In order to perform this comparison a measure of quality for a given feature subset is required. Their proposed measure takes the number of commonalities between the proposed subset and the optimal subset: features that were included in both subsets, and features that were excluded from both subsets. This count is divided by the number of dimensions and that value averaged over values of F, from 1 to 19 inclusive to give a final quality value for the feature set. The maximum possible value for this average quality is one, meaning that the 19 possible feature subsets were the optimal subset for the five data sets. The authors remark that this is not a measure of the classification error, but rather a measure of the difference between the subset produced by a feature selection method and the ideal feature subset. Data sets of different training sizes were produced, ranging from 10 to 5000 samples. For each sample number, 5 datasets were artificially generated and the results averaged. The feature selection criterion employed by the author is the Mahalanobis distance between the two class means, and the features were selected using branch and bound and sequential forward search [78].

Notice that this data set is actually an ICA space: the class-conditional densities are uncorrelated Gaussians, thus independent. So there is no need to transform the data and the equation of divergence for class-conditional independent data (4.9) can be used directly. Also considering the data is Gaussian, each marginal divergence can be calculated using (4.8). In Fig. (4.2) we reproduce the results in [66] using the optimal branch and bound feature subset selection algorithm. We also plot the results of our method, estimating the marginal densities with a 2-Gaussian Mixture Model (no prior knowledge of the data assumed) and with a Gaussian with unknown mean and covariance (Gaussian data assumed).

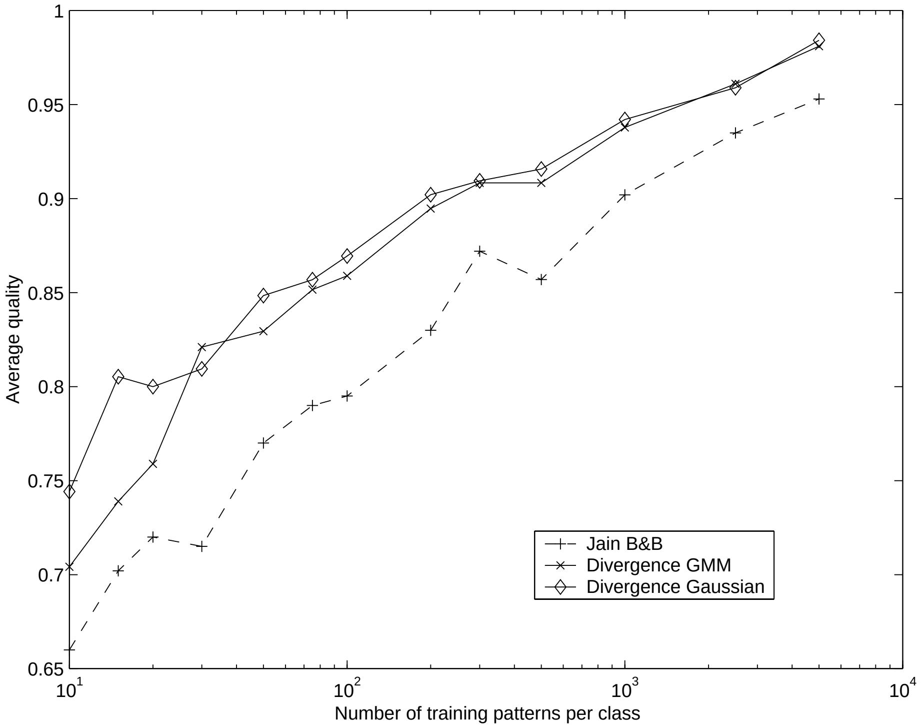


Figure 4.2: Quality of selected feature subsets as a function of the size of the training data.

From fig. (4.2) we observe divergence is a fairly robust criterion with performance above Jain's criterion, even though both make use of only first and second order statistics. Gaussian Mixture Models do not perform well when the number of samples is similar to the dimensionality but soon recovers, meaning that we can do without the prior knowledge on the data distribution without seriously affecting the results.

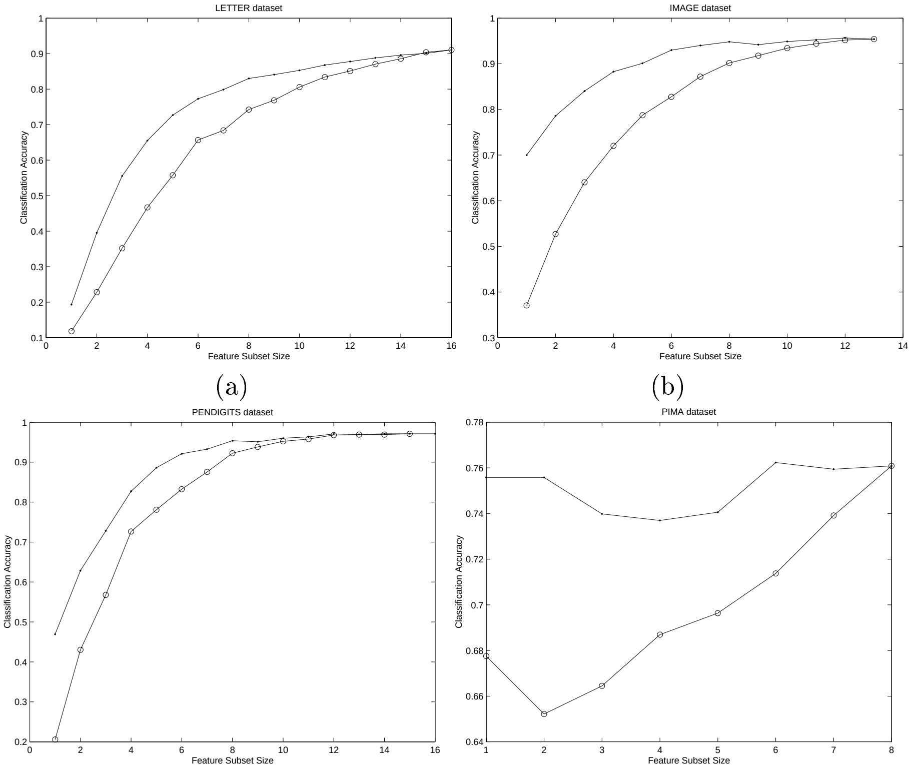
4.4.2 Benchmark Data

For illustrating the performance of our feature selection algorithm, we chose the same benchmark databases detailed in section (3.5.2): LETTER, IMAGE, PENDIGITS and PIMA. For each of these databases, the classification accuracy against the number of selected features is plotted in fig. (4.3). The straight line with dots is the result of using the CC-ICA-FS algorithm. Since, in the last chapter we already observed that CC-ICA was in all cases the Bayesian classifier with highest accuracy, and the Bayes error decreases with dimensionality we can fairly assume that it is very unlikely that any other of the proposed schemes would significantly improve their performance for a reduced number of features. Actually, we can already observe in the figure, that for our Bayesian scheme (CC-ICA) maximum accuracy is always achieved in high-dimensional subsets. For this reason, no comparison is made between our scheme and feature selection for other statistical classifiers. Instead, we chose to compare the effectiveness of our feature selection criterion with an alternative criterion requiring an exhaustive search. As with divergence the main inconvenient is adapting the chosen criterion to work with class-conditional representations.

If the class-conditional independence assumption is dropped we are forced to evaluate divergence on multivariate probabilities. This gives unstable results and, as we mentioned parametric approaches are advised. So we decided to use the Bhattacharyya bound (4.1) and the version of divergence for Gaussian variables (4.8) as criterions. Class-conditional feature subsets were then obtained by maximizing the criterion on each of the representations. In this case, we cannot avoid the exhaustive search on each class and this is how we selected the best feature subsets. The Bhattacharyya bound yielded nonsignificant but slightly better results than Gaussian divergence. In fig. (4.3), the accuracy obtained using Bhattacharyya feature selection with exhaustive search for feature subsets is illustrated with a straight line interpolated with circles.

Though we still have to evaluate our feature selection on high dimensional data, some conclusions can be already extracted from this experiment. We first observe that, in all cases and for all choices of subset size, the divergence criterion outperforms the Bhattacharyya criterion. Though an exhaustive search has been performed, the cause of these poor results is probably related with the false assumption of Gaussianity of the projected data. We also observe that the accuracy using divergence, except for PIMA, increases almost monotonically, generally reaching the highest accuracy before reaching full dimensionality. This has to do with the Bayesian nature of our classifier. In the case of PIMA, a single feature per representation already yields an accuracy of 75.6%, almost the best reported result for this database. In all databases except for LETTER, there existed a subset with higher accuracy than the full dimensional case, but these accuracies were clearly nonsignificant. Never more than a 0.5% difference.

Even though an exhaustive search was performed to select features using the



(a) (b)

(c) (d)

Figure 4.3: Comparison, on benchmark databases, of divergence-based independent feature selection without an exhaustive search (interpolated dots) with Bhattacharyya exhaustive feature selection (interpolated circles).

Bhattacharyya criterion, this cannot be considered a *full* exhaustive search. This is because, class-conditional features were chosen independently of the other classes. A complete exhaustive search should consider all possible feature combinations for all possible classes. This requires a huge number of operations. For instance, consider selecting 5 features for each of the 26 16-dimensional classes in the LETTER database. For this case, we have $\binom{16}{5} = 4368$ possible feature subsets for each class, so we have a total $4368^{26} \approx 2^{312}$ possible combinations of feature subsets. There is no way we can evaluate all these combinations. Nevertheless a simple experiment can be made in order to heuristically check the almost global nature of the solution provided by our algorithm. For this same dataset (LETTER) and feature subset size (5) our method yields an accuracy of 72.7%. Randomly generating 100000 sets of possibly different features for each class the maximum accuracy obtained was 66.3% and the mean accuracy 58.7%. The mean accuracy value already stabilized to 5 decimal places after 1000 tries, so we can predict that this value is close to the mean after evaluating all possible feature subsets. The maximum accuracy can still improve but it is still far from our obtained accuracy. Similar results were obtained when this experiment was repeated for a number of feature subset sizes and for the different databases. In no case a randomly generated feature subset improved the accuracy of the feature subset obtained by CC-ICA-FS.

4.4.3 Real-world Data

As in section 3.5.3 we will attempt to classify from images using the high dimensional color signature as a descriptor. Once again, the samples consist in 512-dimensional vectors made from the three 8-bin histograms that correspond to each color spectrum. These histograms were extracted from different representative regions of 948 images belonging to the Corel Database [28]. The regions belong to ten different classes corresponding to clouds, grass, ice, leaves, rocky mountains, sand, sky, snow mountains, trees and water. Some of these representative regions, for all ten classes, are illustrated on fig. (4.4). In this figure we can already predict the high confusion between classes derived from only using color signature. Grass, trees and leaves can all present similar green patterns; sky, ice and water have strong blue distributions; rock and sand are also similar, at least visually. In addition, there are portions of the clouds which are only sky, or portions of the snow mountains which are mainly rock. Of course, additional information such as textural information would aid discriminability, but we will choose not to avoid the overlap among classes to evaluate CC-ICA and CC-ICA-FS performance.

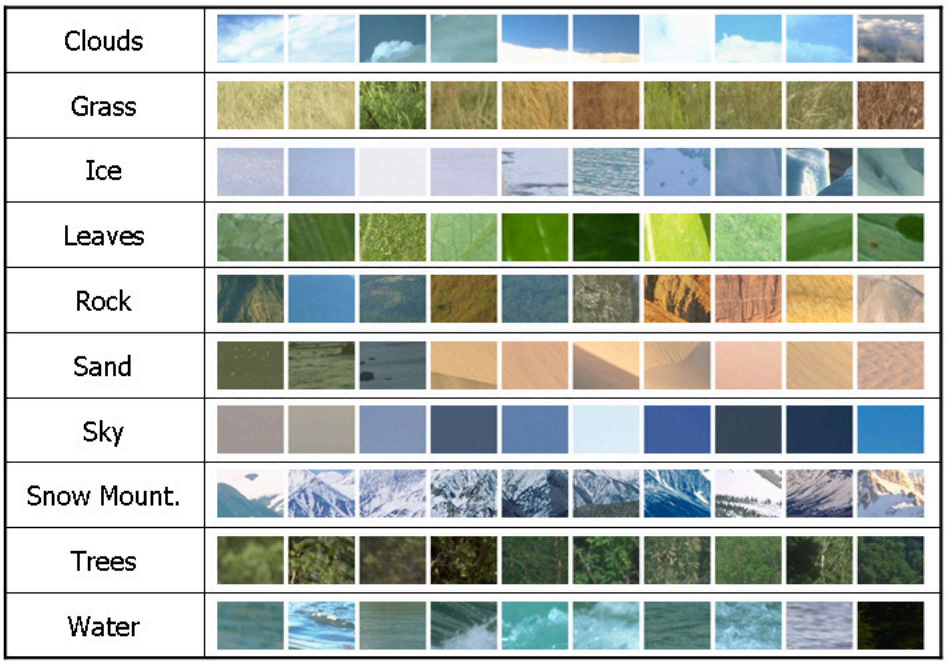


Figure 4.4: Sample image regions and their corresponding class labels.

Finally a total of 40000 possibly overlapped regions were extracted and the histograms calculated. The number of samples per class was equal for all classes: 4000. Of these class samples, 3000 were randomly selected for training our classifiers and feature selectors and 1000 were used for evaluation. For all classes, the true class dimensionality was considerably below 512, 150 in average. This means that classes have very restricted color variation since there are many colors that do not appear in any of the class samples. CC-ICA was performed after PCA dimensionality reduction and whitening. In this case we chose to preserve 98.5% of the class variation. Preserving different percentages of variance does not have a strong impact on performance, except for the 100% case, where accuracy drops: directions of extremely low variation do not contain discriminability information and can be regarded as noise. For instance, a single red flower in one of the 4000 images of leaves. An interesting point is that all CC-ICA representations have different dimensionality, a fact that does not prevents us from still applying the Bayesian scheme for classification. The lowest dimension corresponded to the ice class (42) and the highest dimension corresponded to the leaves class (85). Feature selection was performed using CC-ICA-FS and component densities were estimated using a mixture of 3 Gaussians. We also tried some other methods for comparison. Our main interest is comparing the performance of CC-ICA with the naive Bayes classifier applied to different representations. Results with a global ICA were bad as we could have already predicted from the results over artificial data on the last chapter. So we applied naive Bayes to a PCA representation using a forward search feature selection scheme and the Bhattacharyya bound as separability criterion (PCA-SFS). We also applied a Gaussian maximum likelihood classifier on PCA using the hierarchy PCA itself imposes on the features (PCA). Results with this classifier drop to zero once the covariance matrix for a certain class becomes rank-deficient. We also applied a 1-NN classifier on the original representation using the mean Bhattacharyya distance as feature selection criterion. All these results are illustrated in (4.5). We observe that CC-ICA outperforms all the other evaluated methods for all feature subset sizes. It is also interesting to see very few class-conditional features already yield results close to the optimum, achieved for 31 features per class.

Nevertheless, accuracy results in fig. (4.5) are not impressive: below 70% in the best case. In the color indexing literature, it is not unfrequent to use the average match percentile (AMP) to evaluate the results [145], where a rank in the classification proves sufficient. The match percentile for a certain sample is defined as [145],

$$MP = \frac{K - R}{K - 1} \tag{4.24}$$

Where R stands for rank in which the sample was classified (in our case, obtained from the value of the posterior probability). From (4.24) that the match percentile ranges from zero (the correct sample label is considered the least probable label) to one (the correct sample label is the most probable label). The AMP is the average of match percentiles for all evaluated samples. For instance, an average match percentile of 95, informs us that, on average, the correct match scored higher than 95% of the other models. The AMP for our best case (31 features) is 92.64%. Results using AMP instead of accuracy are exposed in fig. (4.6).

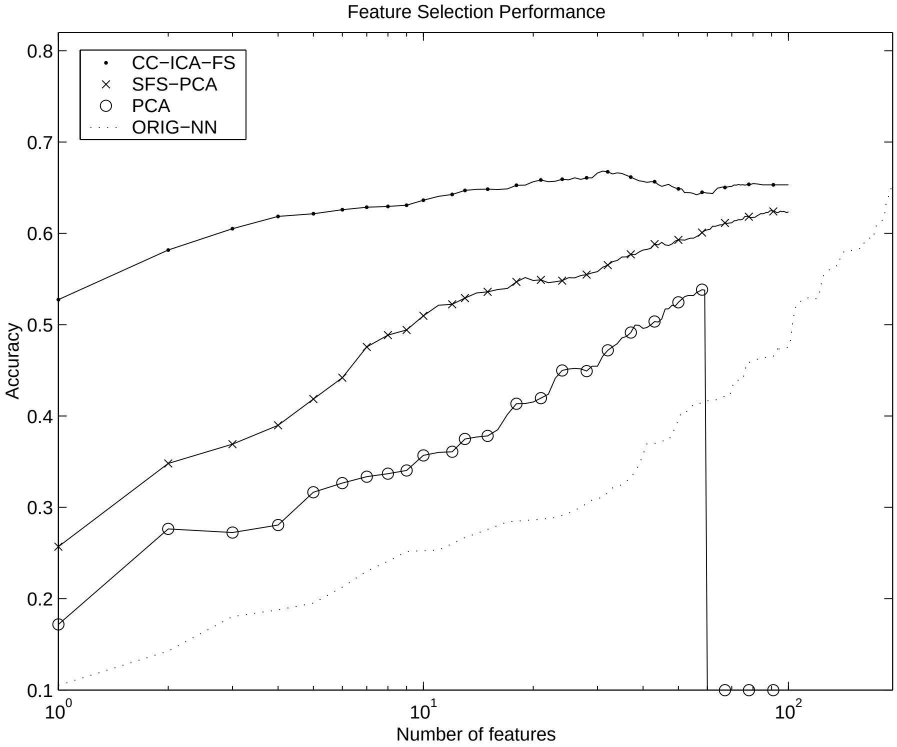


Figure 4.5: Classification accuracy against the feature subset size for a color indexing problem.

A distinctive characteristic of CC-ICA-FS applied to this problem is that selecting a single feature per class already provides an accuracy (or AMP) close to the maximum achieved accuracy. This means that we have quite a lot discriminative information in a single (class-conditional) feature or, equivalently, in a single ICA basis. Visualization of this basis as a color histogram is not straightforward since it might have negative values. But we can have an idea of the colors affected by the basis by considering its absolute value as a histogram. In fig. (4.6) we have artificially generated colored squares with the distributions of the most discriminative basis of each class. The proceeding was similar to that performed in section (2.3.2), except for the fact that the basis in that section could be considered histograms because they were obtained through an NMF representation and in consequence had only nonnegative values. It is not strange that in the figure we observe only colors belonging to particular classes, since each basis was obtained exclusively from its corresponding class samples. The interesting point is that the colors observed are those less repeated from class to class. For instance, consider the grass, trees and leaves classes. Their CC-ICA representations probably had highly correlated basis vectors, mainly greenish. But these bases precisely correspond to components with no discriminative information. Instead, our single feature selection procedure chose mainly yellows for grass, green

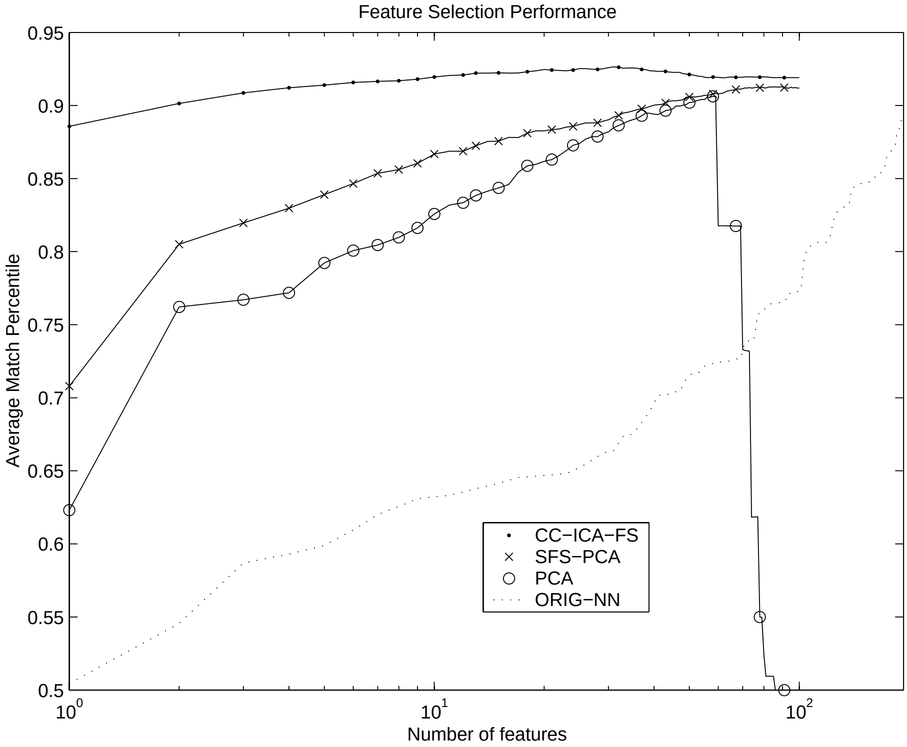


Figure 4.6: Average Match Percentile against the feature subset size for a color indexing problem.

for leaves and brown for trees, which is the distribution that probably will allow discriminating the highest amount of images belonging to these three classes. A similar phenomenon can be observed in sky, ice and water. Observe that water in this basis is more represented by the colors of rocks or sand present underneath or by the water then by the blue tones that might also be present. Comparison of fig. (4.7) with fig. (4.4) might result helpful for understanding these facts.

4.5 Conclusions

In this and the previous chapter we have introduced a statistical approach which covers the different stages of a general pattern classification scheme: feature extraction, feature selection and classifier. More importantly, the proposed theory attempts to integrate these phases within a unified framework. Once the initial assumptions are made, feature extraction, selection and finally classification are naturally associated among each other. This is opposed to the usual approach where no relationship among the stages is assumed and where, the common procedure is to fix any one of the stages and than evaluating results on another stage. For instance, test nearest neighbor performance on different representations such as PCA, ICA, LDA, etc.

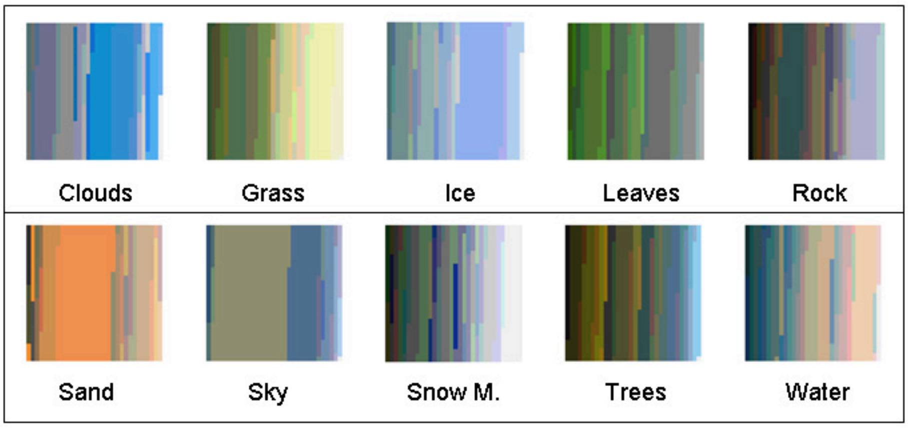


Figure 4.7: Artificial images with color distribution given by the normalized absolute values of the most discriminant class-conditional bases.

Or evaluate distance metrics on a given representation. Our unifying assumption has been that each class is a linear mixture of statistically independent sources. If this is true, CC-ICA will effectively separate the sources and working in this new (class-conditional) representation will greatly simplify density estimation. Bayesian classification can take advantage of this simplification and, always under the independence assumption, the classifier under this hypothesis is the naive Bayes classifier. Moreover, independence also has a positive effect which facilitates the selection of discriminative features without the need for an exhaustive search. This can be done using the statistical and information theoretic class separability measure of divergence. This measure can also be adapted to the situation in which class-conditional representations are found.

This unified framework requires the estimation of a number of parameters (mainly the ICA parameters and densities) so, besides fulfillment of the assumptions, the amount of available data should be adequate for the estimation of these parameters. We can then pose the two major drawbacks of our method. If the assumptions hold, we have to count on a proper amount of data (number of samples per class) to estimate our parameters. This dataset size is closely related to dimensionality. If the assumptions do not hold, the whole process is invalidated. In practice, if we have reasons to think that the available data is sufficient but still, results are not satisfactory, it is highly unprobable that we will be able to improve performance by taking a different approach on any single stage. For instance, changing the classifier while preserving the CC-ICA features will probably worsen the situation since the proposed classifier is actually the natural classifier for this kind of features. The same situation arises when using other criterions for feature selection or, as we have seen in the experiments, when applying the proposed classifier to different representations.

Beneath the particularities of our method underlies a general line of reasoning which holds that the different stages of a pattern recognition scheme should be considered as a whole. That a particular choice on any of these stages involves assumptions that should affect all other steps of the process. The next chapter extends this line of reasoning to a completely nonparametric framework.

Chapter 5

Nonparametric Discriminant Analysis and Classification

5.1 Introduction

In section (4.5) we observed that the statistical approach exposed in the last two chapters covers almost all the stages of a general pattern classification scheme except, possibly, for data preprocessing. These phases are integrated under the umbrella of a set of initial hypotheses based on independence which, once accepted, gives place to a unified and interrelated framework for feature selection, extraction and classification. Two drawbacks which negatively affect performance of the proposed approach were mentioned: sample size and unmet assumptions.

Parametric methods normally restrict data to a particular model based on the chosen parameters and then attempt to fit the model to a given set of observations using one among many possible parameter estimation techniques. In general, restricting data to a model involves making more or less specific assumptions on the data. General assumptions usually require more information for robust parameter estimation. Nonparametric methods, instead, make no assumption on the data distribution and solely make use of the available set of observations for making their predictions. Though their performance improves with a larger number of samples, they still are the only option when this is not the case. In this chapter, we extend the line of reasoning held in the previous chapters to a nonparametric context. That is, consider the pattern recognition scheme as a whole and try to answer the following question: in the same way CC-ICA provides naive Bayes with features that adapted to meet its assumptions, given a nonparametric classifier, can we also find a (nonparametric) representation which results optimal for the classifier?

The family of nonparametric classifiers is huge so in the first place we will restrict our analysis to the family of nearest neighbor classifiers [42, 29]. This rule, besides being very intuitive and of straightforward application, has very desirable convergence properties with an asymptotic error rate of at most twice the Bayes error [29]. The simplest and most well known version of this classifier is the 1-nearest neighbor (1-NN) provided with the Euclidean distance. Since its introduction many modifications and improvements have been suggested to enhance this rule and make it suitable for a wide range of problems [30]. More neighbors can be considered giving rise to the K-NN rule when K neighbors are considered; any other distance than the Euclidean can be used, or local distances adapted, for example, to the distribution of the data can be considered [135, 148]; other neighborhood definitions can be used [128]; data editing can frequently enhance performance [34]; etc.

On the other side, Fukunaga's Nonparametric Discriminant Analysis (NDA) [49] is a linear discriminant feature extraction technique based on nonparametric extensions of commonly used scatter matrices. In contrast to parametric discriminant analysis these matrices are generally full rank so the extracted number of features can exceed the number of classes and no assumption on the class distributions is made. The procedure can work with high dimensional data and does not necessarily require a high number of samples per class. We explore the link between NDA and the NN classifier, observing that a slight modification of NDA results in a representation very likely to improve NN performance. Feature selection is not necessary in this stage since NDA sorts features according to their discrimination, and usually the first F features will be those that yield best results.

The resulting method proves to be a possible alternative to CC-ICA as a coherent pattern recognition scheme since it should work in many occasions where CC-ICA fails. It makes absolutely no assumption on the distribution of the data and can be implemented even in those cases small sample sets do not allow robust parameter estimation. Figure (5.1) is a diagram which illustrates both approaches and their relationship with a standard pattern classification training process.

First, a general framework for discriminant analysis is introduced following the structure proposed by Fukunaga in [48]. Both parametric and nonparametric approaches are detailed and placed within this framework. We then analyse the nonparametric algorithm from within the nearest neighbor perspective showing that a slight modification of NDA yields a maximum of a criterion function good for predicting nearest neighbor performance. The resulting method, which we call NDA-NN, is evaluated on artificial, benchmark and real-world databases. The artificial database provides an insight on the nature of our representation and compares results with the classical NDA approach. Experiments on benchmark databases compare our approach with classical NDA, parametric discriminant analysis and with the CC-ICA framework of previous chapters. We have already observed, in chapter 3 that NN performs well in at least two of the tested benchmark databases. We will see that NDA enhances NN performance to an unprecedented extent in these cases. We will also see that, in those cases where initial NN performance is far from acceptable, improvement is not significant. Finally, we test the performance of NDA-NN to real-world data where CC-ICA is unable to work properly due to unmet assumptions or small sample sizes.

5.2 Discriminant Analysis

Discriminant analysis is a feature extraction tool based on a criterion J and two square matrices $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$. These matrices generally represent the scatter of sample vectors

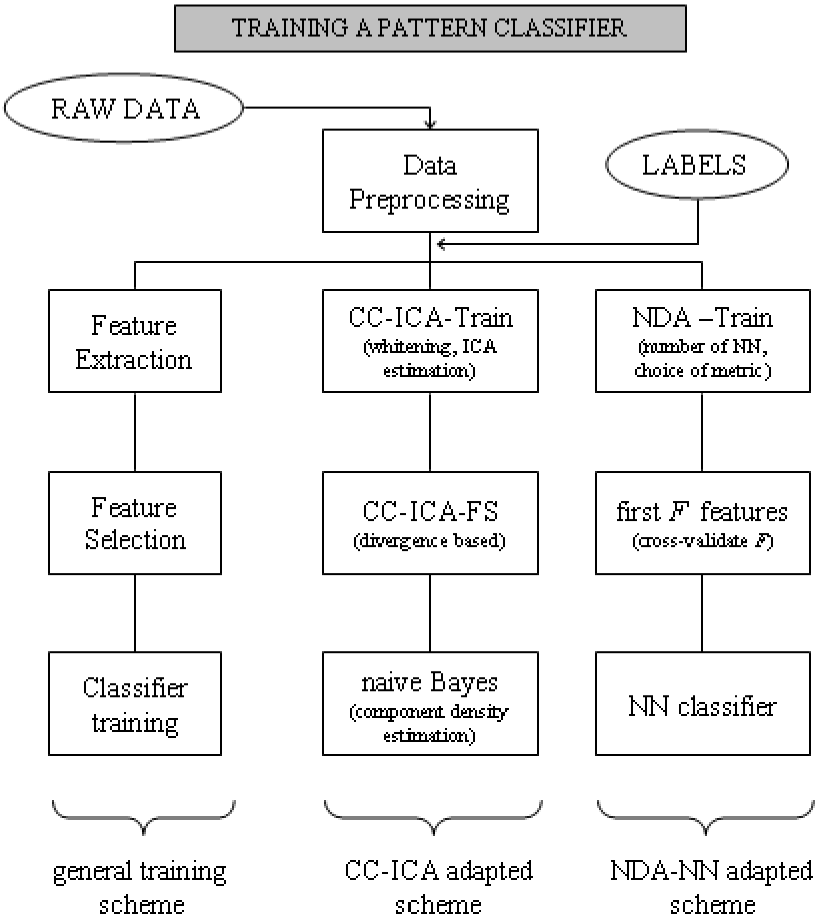

TRAINING A PATTERN CLASSIFIER
RAW DATA
Data Preprocessing
LABELS
Feature Extraction
CC-ICA-Train (whitening, ICA estimation)
NDA –Train (number of NN, choice of metric)
Feature Selection
CC-ICA-FS (divergence based)
first F features (cross-validate F)
Classifier training
naive Bayes (component density estimation)
NN classifier
general training scheme
CC-ICA adapted scheme
NDA-NN adapted scheme

Figure 5.1: A standard pattern classification training process and its adapted version to the CC-ICA and NDA-NN frameworks.

between different classes for $\boldsymbol{S}_b$, and within a class (or sometimes class independent scatter information) for $\boldsymbol{S}_w$. The most frequently used criterion, is to choose $J = trace(\boldsymbol{S}_w^{-1}\boldsymbol{S}_b)$.

It can be seen that, maximization of J is equivalent to finding the $D \times M$ linear transformation $\boldsymbol{W}$ such that

$$\hat{\boldsymbol{W}} = \arg \max_{\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}=\boldsymbol{I}} trace(\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}) \qquad (5.1)$$

where $\boldsymbol{I}$ is the identity matrix. It can be proven that, given N samples of D dimensional data $\boldsymbol{X}$ and discriminant space dimensionality M, the algorithm in table (5.1) solves the optimization problem given in equation (5.1) [48].

We can now turn to the definition of the within and between class scatter matrices.

1. Given $\boldsymbol{X}$ the matrix containing data samples placed as N D-dimensional columns, $\boldsymbol{S}_w$ the within class scatter matrix, and M maximum dimension of discriminant space,
2. Compute eigenvectors and eigenvalues for $\boldsymbol{S}_w$ and make Φ the matrix with the eigenvectors placed as columns and Λ the diagonal matrix with only the nonzero eigenvalues in the diagonal. Let $M_w = \{\text{number of non-zero eigenvalues}\}$.
3. Whiten the data with respect to S_w, to obtain M_w dimensional whitened data,
$$\boldsymbol{Z} = \Lambda^{-1/2}\Phi^T\boldsymbol{X}.$$
4. Compute S_b on the whitened data.
5. Compute eigenvectors and eigenvalues for $\boldsymbol{S}_b$ and make Ψ the matrix with the eigenvectors placed as columns and sorted by decreasing eigenvalue value.
6. Preserve only the first $M_b = \min\{M_w, M, \text{rank}(\boldsymbol{S}_b)\}$ columns, $\Psi_M = \{\psi_1, \ldots, \psi_{M_b}\}$ (those corresponding to the M_b largest eigenvalues).
7. The resulting optimal transformation is $\hat{\boldsymbol{W}} = \Psi_M^T\Lambda^{-1/2}\Phi^T$ and the projected data, $\boldsymbol{Y} = \hat{\boldsymbol{W}}\boldsymbol{X} = \Psi_M^T\boldsymbol{Z}$

Table 5.1: Given a method for calculating within and between class scatter matrices S_w and S_b, general algorithm for solving maximum discriminability optimization problem stated in equation (5.1).

5.2.1 Fisher Discriminant Analysis

As usual, the first and most widely spread approach is the one that makes use of only up to second order statistics of the data. This was done in a classic paper by Fisher [41]. The popularity of the technique introduced by Fisher has caused that this feature extraction method has become known as *discriminant analysis* or *linear discriminant analysis*. This is not exact because since Fisher several other techniques that can fairly be regarded as discriminant analysis have arisen and many of them, including the nonparametric version we are introducing, are linear. To avoid confusion, we will call this technique Fisher Discriminant Analysis (FDA). In FDA the within class scatter matrix is usually computed as a weighted sum of the class-conditional sample covariance matrices where the weights are given by the class prior probabilities,

$$\boldsymbol{S}_w = \sum_{k=1}^{K} P(C^k)\boldsymbol{\Sigma}^k \tag{5.2}$$

where $\boldsymbol{\Sigma}^k$ is the class-conditional covariance matrix, estimated from the sample set. On the other side, the most common way of defining the between class-scatter matrix is as,

$$\boldsymbol{S}_b = \sum_{k=1}^{K} P(C^k)(\boldsymbol{\mu}^k - \boldsymbol{\mu}^0)(\boldsymbol{\mu}^k - \boldsymbol{\mu}^0)^T \tag{5.3}$$

where μ^k is the class-conditional sample mean and μ^0 is the unconditional (global) sample mean. Many other less spread out forms, always based on sample means and class-conditional covariance matrices are also available for these two scatter matrices [48].

Without further experiments some observations can be readily made from these definitions. For classes tightly grouped about their means, the objective value should be sufficiently high. This is not the case when the means are close among each other, negatively affecting the between class scatter matrix. Another problem arises when the class-conditional covariances are very different from each other, each one contributing with scatter in different directions. Also observe that the rank of $\boldsymbol{S}_b$ is $K-1$, so the number of extracted features is, at most, one less than the number of classes. Some solutions have been proposed for solving this problem [44]. A first solution is to artificially increase the number of classes by, for instance, clustering within the classes. For those cases where multimodal behavior is present, and a clustering algorithm can be found that *properly* identifies the clusters, this might work well. A second possibility is, after determining the first $K-1$ features, remove them leaving a subspace orthogonal to the extracted features and repeat the algorithm.

A more fundamental problem with these scatter matrices is their parametric nature. Once again, given a certain problem, we run into a solution that results blind for nongaussian classes. If the class-conditional distributions are highly nongaussian (subgaussian, supergaussian or even multimodal), applying the algorithm (5.1) to the data will give equivalent results to applying the algorithm to Gaussian data with the same means and covariances as the original data. So, for this particular case, we cannot expect our method to accurately indicate which features should be extracted to preserve any complex classification structure.

In fig. (5.2) we illustrate the potential and limitations of FDA to extract a single feature from a couple of artificial 2-dimensional examples. In both plots the conditional and unconditional means are shown with circles and the straight line corresponds to the direction of the projection subspace. The extracted feature is the result of projecting the data in this direction. The extracted feature for the first example where Gaussian conditional densities are found is in effect optimal for discrimination. In this case, classes are well separated in projection space. This is not the case in the second example, where multimodality is shown to negatively affect the estimation. If a vertical direction was chosen instead, the classes would be perfectly separated.

5.2.2 Nonparametric Discriminant Analysis

In [49] Fukunaga and Mantock present a linear and nonparametric method for discriminant analysis in an attempt to overcome the limitations present in (FDA) [41], and name the technique Nonparametric Discriminant Analysis (NDA). Basically, the fact that FDA's resulting dimensionality is upper bounded by the number of classes and its parametric nature. Nonparametric Discriminant Analysis also makes use of a Fisher-like objective function but in this case the between-class scatter matrix is of nonparametric nature. This scatter matrix is generally full rank, thus loosening the bound on extracted feature dimensionality. Also, the nonparametric structure of this matrix inherently leads to extracted features that preserve relevant structures for classification. We briefly expose this technique, extensively detailed in [48].

In FDA the between-class scatter matrix is constructed from the differences among

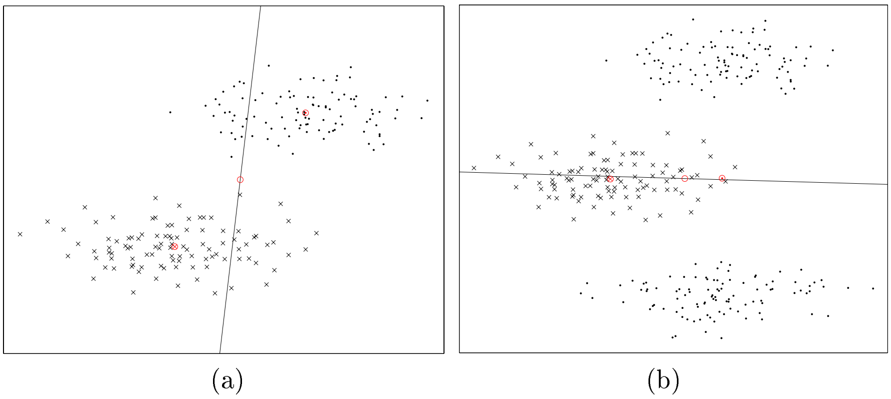


Figure 5.2: First and only direction of Fisher Discriminant projection space on two artificial datasets. Observe in (b) the problems caused by unfulfilled assumptions.

the class means. In NDA we define a between-class matrix as the scatter matrix obtained from vectors locally pointing to another class. This is done as follows. Given a norm $\|\|$ in the metric space where the samples live, the extraclass nearest neighbor for a sample $\boldsymbol{x} \in C^k$ is defined as

$$\boldsymbol{x}^E = \{\boldsymbol{x'} \in \overline{C^k} / \|\boldsymbol{x'} - \boldsymbol{x}\| \leq \|\boldsymbol{z} - \boldsymbol{x}\|, \forall \boldsymbol{z} \in \overline{C^k}\} \tag{5.4}$$

where $\overline{C^k}$ notes the complement set of C^k. In the same fashion we can define the intraclass nearest neighbor as

$$\boldsymbol{x}^I = \{\boldsymbol{x'} \in C^k / \|\boldsymbol{x'} - \boldsymbol{x}\| \leq \|\boldsymbol{z} - \boldsymbol{x}\|, \forall \boldsymbol{z} \in C^k\} \tag{5.5}$$

Both definitions (5.4) and (5.5) can be extended to the K nearest neighbors case by defining $\boldsymbol{x}^E$ and $\boldsymbol{x}^I$ as the mean of the K nearest extra or intra-class samples. From these neighbors or neighbor averages, the extraclass differences are defined as

$$\boldsymbol{\Delta}^E = \boldsymbol{x} - \boldsymbol{x}^E \tag{5.6}$$

and the intraclass differences as

$$\boldsymbol{\Delta}^I = \boldsymbol{x} - \boldsymbol{x}^I \tag{5.7}$$

Notice that $\boldsymbol{\Delta}^E$ points locally to the nearest class (or classes) that does not contain the sample. The nonparametric between-class scatter matrix is then defined as

$$S_b = \sum_{n=1}^{N} w_n (\boldsymbol{\Delta}_n^E)(\boldsymbol{\Delta}_n^E)^T \tag{5.8}$$

where $\boldsymbol{\Delta}_n^E$ is the extraclass distance for sample $\boldsymbol{x}_n$, w_n a sample weight defined as

$$w_n = \frac{\min\{\|\boldsymbol{\Delta}^E\|^\alpha, \|\boldsymbol{\Delta}^I\|^\alpha\}}{\|\boldsymbol{\Delta}^E\|^\alpha + \|\boldsymbol{\Delta}^I\|^\alpha} \tag{5.9}$$

and α is a control parameter between zero and infinity. This sample weight is introduced in order to de emphasize samples away from class boundaries. These samples generally have a larger extraclass difference magnitude exercising an undesirable influence on the scatter matrix: Precisely these samples are those that less information carry on nearest class direction. The sample weights in (5.9) take values close to 0.5 on class boundaries and drop to zero as we move away. The control parameter α adjusts how fast this happens. In the experiments we have not observed a uniform impact on performance of the sample weights. To facilitate comparison and, unless stated otherwise, uniform weights will be considered.

Unlike the between-class scatter matrix, the within-class scatter matrix is parametrically estimated in exactly the same fashion as in FDA (eq. 5.2). This choice is heuristically based on the observation that normalization (the first step of the detailed algorithm) should be as global as possible. Intuitively, whitened data can be naturally associated to the Euclidean distance so a global whitening matrix should benefit distance-based classifiers that make use of this metric.

In the calculation of both scatter matrices, for theoretical integrity, Fukunaga includes the class prior probabilities $P(C^k)$. In practice, when no such information is available uniform priors are assumed and the resulting formulas for between and within-class scatter matrices are (5.8) and (5.2). In order to show that NDA is actually a natural nonparametric extension of FDA, the behaviour of $\boldsymbol{S}_b$ when the number of neighbors considered reaches the total number of available class samples is also studied. It is observed that, for this particular case and restricting the problem to two classes, the features extracted are the same as in FDA.

Figure (5.3) shows the NDA solution to the same artificial datasets where we tested FDA. For this example a single nearest neighbor was used in the computation of the between-class scatter matrix and uniform sample weights were considered. Particularly interesting is the case illustrated in fig. (5.3.b). Though both within-class scatter matrices are equal, the bimodality of one of the classes displaces the estimate of the class mean used in the computation of the parametric between-class scatter matrix. This is the main source of error for FDA.

5.3 NDA and Nearest Neighbors

Making use of the introduced notation we will first examine the relationship between NN and NDA. Given a training sample $\boldsymbol{x}$, the accuracy of the 1-NN rule can be directly computed by examining the ratio $||\boldsymbol{\Delta}^E||/||\boldsymbol{\Delta}^I||$. If this ratio is more than one, $\boldsymbol{x}$ will be correctly classified. The fact this affirmation does not generalize for the K-NN rule, is due to our definition of intra and extra class nearest neighbor for this particular case, coherent with Fukunaga's approach. This situation, if problematic, can be easily overcomed by using the median according to the distance norm in the definitions. In the experiments, we have observed that this choice does not affect results significantly so the definition using the mean was preserved for proper of comparison.

Given the $M \times D$ linear transform $\boldsymbol{W}$, the projected distances are defined as

$$\boldsymbol{\Delta}_{\boldsymbol{W}}^{E,I} = \boldsymbol{W}\boldsymbol{x} - \boldsymbol{W}\boldsymbol{x}^{E,I} \tag{5.10}$$

Notice that this definition does not exactly agree with the extra and intraclass dis-

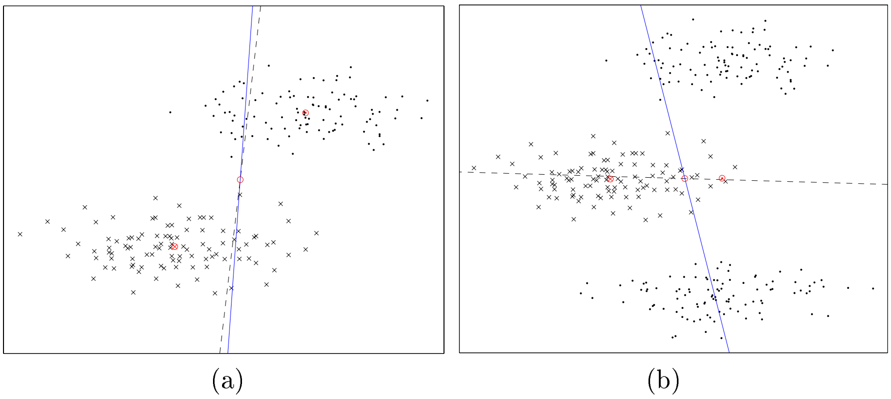


Figure 5.3: First direction of nonparametric discriminant projection space on two artificial datasets. In dashes the FDA direction. Compare in (b) where the FDA assumptions are not met.

tances in projection space since, except for the orthonormal transformation case, we have no warranty on distance preservation. Equivalence of both definitions is asymptotically true on the number of samples. By the above remarks it is expected, that optimization of the following objective function should improve or, at least not downgrade NN performance,

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \frac{E\{\|\boldsymbol{\Delta}_{\boldsymbol{W}}^{E}\|^2\}}{E\{\|\boldsymbol{\Delta}_{\boldsymbol{W}}^{I}\|^2\}} \tag{5.11}$$

Considering that [48],

$$E\{\|\boldsymbol{\Delta}_{\boldsymbol{W}}^{E,I}\|^2\} = trace(\boldsymbol{W}^T \boldsymbol{S}_{b,w} \boldsymbol{W}) \tag{5.12}$$

where, in this case, $\boldsymbol{S}_b$ (the between-class scatter matrix) agrees with (5.8), but the within-class scatter matrix is now defined in a nonparametric fashion,

$$S_w = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\Delta}_n^I {\boldsymbol{\Delta}_n^I}^T \tag{5.13}$$

From (5.12) we have that the algorithm in table (5.1) can also be applied to the optimization of our proposed objective function (5.11).

Considerations on sample weights and class priors can be plugged in the same fashion as with NDA. Theoretical considerations also hold: it can be seen that, if the all the class members are used and the mean nearest neighbor criterion used for defining the intra-class distances, (5.13) turns out to be (5.2). This is simply because, if all class members are considered, the nearest neighbor mean becomes the class mean $\boldsymbol{\mu}^k$.

We have already seen in figures (5.2) and (5.3) the difference that taking a nonparametric approach on the between-class scatter matrix can make on the estimated projection space. In that example, the within-class scatter matrix was considered

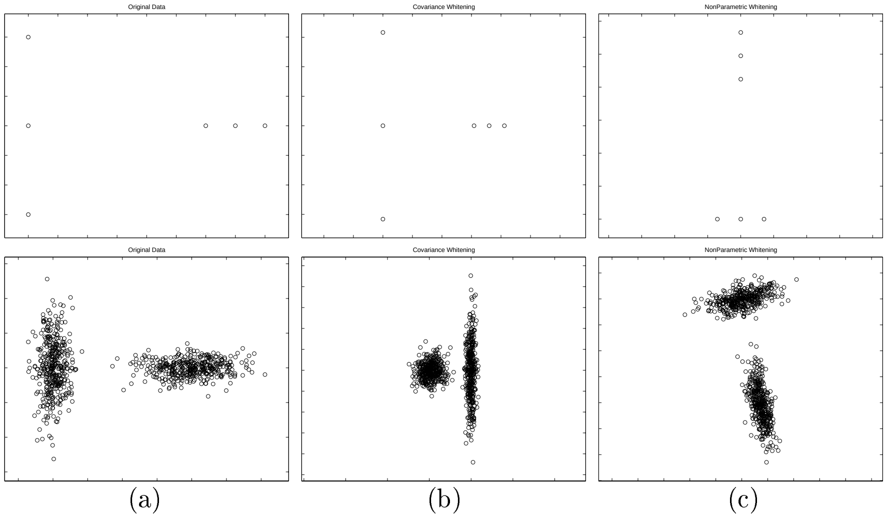


Figure 5.4: On each row a toy dataset. (a) original data. (b) Whitened data using the usual covariance-based within-class scatter matrix. (c) Whitened data using nonparametric within-class scatter matrix.

equal. Through another artificial with two simple toy datasets example we will compare the effect of whitening using a nonparametric within-class scatter matrix (5.13) or a parametric approach (5.2). The first choice whitens our data respect to combinations of class-dependent covariance matrices. This second-order statistic, measuring the mean distance to the mean, fails to represent classes with more complex distributions. The interpretation of (5.13) is quite straightforward. In the whitened data, the distribution of the intraclass nearest-neighbor distances are normalized, with the favourable consequences this has on the NN-rule. This can be observed in the two examples in fig. (5.4).

5.3.1 Related Works

A close approach to ours was introduced in [19]. In this article, the concept of intra- and extra-class differences is also used, but to obtain a Bayesian measure of face similarity through subspace analysis (PCA) of both difference spaces. The NN rule is then applied to this measure. The main differences with respect to NDA are mainly the parametric nature of the approach (Gaussian or Gaussian mixture assumption on the reduced spaces) and a dual analysis of the data within a Bayesian framework instead of a joint discriminant analysis. The main similarity, besides the use of class differences, lays on the dual eigenspaces, obtained from covariance matrices from which our scatter matrices are a particular case: the case in which only the nearest neighbors are used in the calculation.

Another close approach which uses local discriminant information is Discriminant Adaptive approach introduced by Hastie and Tibshirani [148]. In this case, the authors focus on an iterative scheme to obtain a local metric modifying the local neighborhoods.

5.4 Experiments

In all the experiments NDA and modified NDA (NDA2) were learnt using a single nearest neighbor and no sample weights. Nevertheless, we have observed that a proper adjustment of these parameters to each dataset can generally enhance results. Considering the results are illustrative enough and for the sake of comparison, we choose not to touch these settings.

5.4.1 Artificial Data

In previous sections we have already set up two very simple artificial experiments. The first one, illustrated in fig. (5.3) compares classical NDA with FDA. This experiment shows how NDA's nonparametric approach to the estimation of the between-class scatter matrix can prove advantageous in certain situations where the assumptions made by FDA are not met. The second toy dataset, illustrated in (5.4) compares classical NDA with our modified NDA (NDA2). In this case, the difference results from different approaches to estimating the within-class scatter matrix. In this case, the completely nonparametric nature of NDA2 allows preservation of the intra-class distances with the benefit this brings to NN classification. We will now further compare the two approaches reproducing an artificial experiment proposed by Fukunaga and Mantock in their original work on NDA [49]. The object of this comparison is to evaluate the robustness of both algorithms to the number of samples.

For this experiment, two groups of three-dimensional data were generated. The first two measurements were generated using random number with the uniform distribution as shown in fig. (5.5). The third dimension has zero mean and unit variance Gaussian distribution.

Notice that, for this dataset, the third dimension is irrelevant for classification. In their experiment, for which 100 samples were generated, Fukunaga and Mantock observed that the resulting eigenvalues (obtained in step 6 of the algorithm given in table 5.1) clearly indicate that only two features are needed and that first two rows of the filter matrix (obtained in step 7 of the same algorithm) practically exclude the third variable. In this experiment, the eigenvalues were normalized to add 1 so the third eigenvalue was 0.04 or equivalently, accounted for 4% of the between class scatter. From this, the authors reasonably conclude that NDA in effect learns that all discriminative information is in the first two variables, disregarding any structure present in the third variable. Fukunaga and Mantock used 3 nearest-neighbors and sample weights with $\alpha = 2$ in (5.9). We reproduced the experiment with a single nearest-neighbor and equal sample weights and results were very similar. We also applied NDA2 to this dataset with a single nearest-neighbor and equal weights. Results were slightly better when considering only the third eigenvalue but hard to interpretate due to the difficult in comparing the filters by their values. So we tried to generate a unique statistic to measure the *goodness* of the two first filters or projection vectors. While the third eigenvalue measures the relevance of the third projection vector and

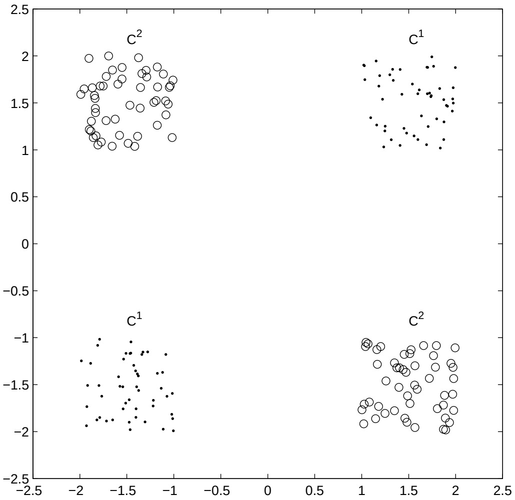


Figure 5.5: Distribution of the two classes along the first two dimensions in the artificial experiment [49].

consequently third feature, this statistic should measure the precision of the first two features.

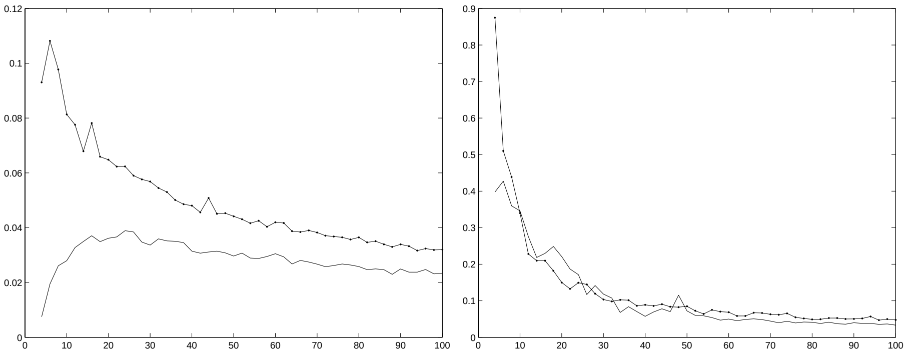
The statistic we use is based on the distance of the optimal 2-dimensional subspace to the obtained subspace. In this case, the optimal subspace is well known: the hyperplane spanned by the two first canonical vectors. So, if $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are the first two filter vectors estimated by NDA, this distance can be estimated, for instance, using the angle between the hyperplanes. A simple expression of this angle can be obtained if we use $\boldsymbol{\tau}$, the unit norm vector orthogonal to the obtained hyperplane,

$$\begin{aligned} \boldsymbol{w}_1^T \boldsymbol{\tau} &= 0 \\ \boldsymbol{w}_2^T \boldsymbol{\tau} &= 0 \\ \boldsymbol{\tau}^T \boldsymbol{\tau} &= 1 \end{aligned}$$

Considering that the orthogonal vector to the optimal subspace is $[0, 0, 1]$, the angle between hyperplanes is obtained by,

$$\theta = \arg\cos|\tau_3| \tag{5.14}$$

Since the object of this experiment is to evaluate the robustness of both algorithms with respect to the sample size we have generated datasets for different number of samples per class. This number ranged from 4 through 100 in steps of 2. For each sample size, we generated 20 artificial datasets, estimated NDA (classical approach) and NDA2 (our modified approach) and averaged the resulting eigenvalues and angles obtained from each representation. For the resulting representations, we computed both statistics: the value of the third eigenvalue and the angle given by (5.14). Notice that the maximum value for this angle is $\pi/2 \approxeq 1.571$. Results are shown in fig. (5.6). In fig. (5.6.a) the obtained eigenvalues for the third direction of projection are plotted against the number of samples. Notice that, for a low number of samples, the nonparametric estimation of the within-class scatter matrix (NDA2) results in considerably better estimation of the degree of relevance of the third feature. The angles are plotted in fig. (5.6.b) in this case, and except for a very low number of samples, differences are not significative. An interesting point to observe is the unstable response of NDA2. This is because the distance to the nearest neighbor might have strong variations for different randomly generated datasets. We can see that the situation stabilizes after 50 samples are considered, NDA2 yielding slightly better angles than NDA. From the plots we also notice that interpreting the results exclusively from the third eigenvalue might be misleading. For NDA2 very low eigenvalues are obtained for small number of samples. But for the same number of samples a high angle was computed. This means that while NDA2 considers the third feature irrelevant, still the first two feature are not properly estimated.



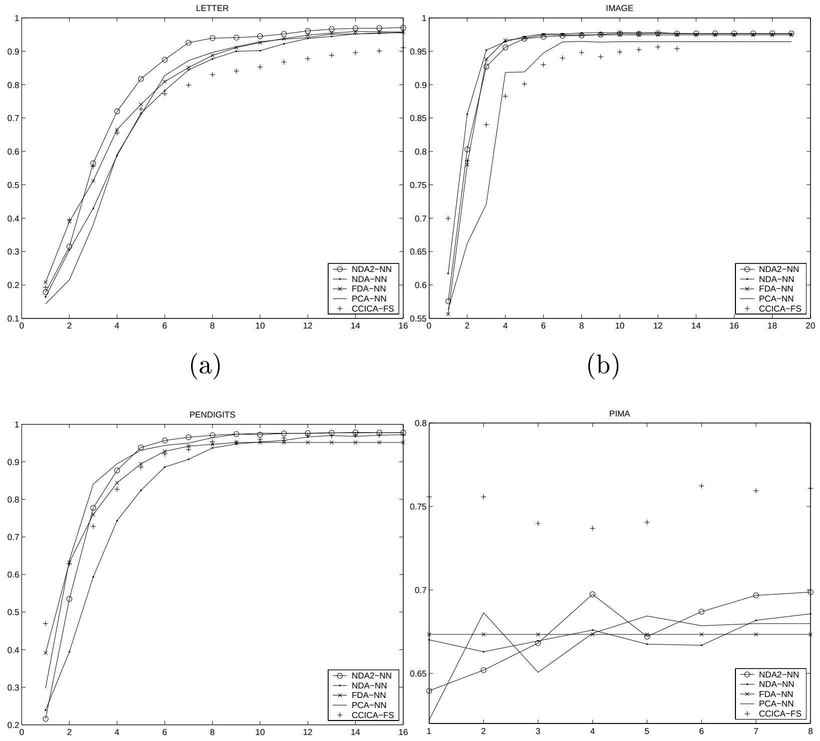
(a) (b)

Figure 5.6: Results on the Fukunaga dataset. As a function of the sample size: in (a) the eigenvalues; in (b) the angle between the obtained hyperplane and the optimal hyperplane. For both plots, the straight line indicates NDA2 and the dotted line NDA.

5.4.2 Benchmark Data

Once again, for illustrating the performance of our nonparametric approach to global classifier design, we chose the benchmark databases detailed in section (3.5.2): LETTER, IMAGE, PENDIGITS and PIMA. In this case, the (1)-NN classifier with NDA2 (NDA2-NN) was compared with the following classifiers: NN with classical NDA (NDA-NN), NN with FDA (FDA-NN), NN with PCA (PCA-NN) and, as a reference, the CC-ICA-FS results are also included.

Classification accuracy results are shown in (5.7). Training and test sets were the same than the sets used in all other experiments made with these benchmark databases, and are detailed in section (3.5.2). Plots are slightly confusing due to similar accuracies when higher dimensionalities are considered. In all cases, NDA2 achieved the highest classification accuracy, sometimes together with other methods. In the letter database, this occurs for all possible dimensionalities except the lower two. In general, NDA2 outperforms NDA and when this is not the case, both techniques yield very similar results. Nonparametric methods outperform CC-ICA-FS. The exception is the PIMA database, where the statistical approach clearly outperforms the metric approach. This could have been anticipated from table (3.5.2) in chapter 3. In table (5.4.2) we resume the maximum accuracies achieved for each benchmark database, the dimensionality where this accuracy was obtained, and the method that achieved this value. Most accuracies are very close to the maximum reported accuracy for the corresponding database. Notably, in the letter database, where the accuracy we obtained outperforms the highest reported accuracy for this dataset (95.7%).



(a)

(b)

(c)

(d)

Figure 5.7: Comparison of NDA2-NN classification with other representations and with CC-ICA-FS on benchmark databases.

In all cases, performing nonparametric discriminant analysis prior to classification with nearest neighbors enhances the results. In general nearest neighbor performance over the original measurements is achieved at low dimensionalities of the NDA space. For instance, in the LETTER database, nearest neighbor accuracy (95.7%) is already achieved when making use of only 11 out of 16 NDA components. For the IMAGE database this occurs when using 4 out of the 19 components. We can conclude that NDA ensures an improvement in nearest neighbor performance, even after strong reductions of dimensionality. We have also observed, that modifications in NDA estimation such as using a different number of neighbors, choosing a metric different than the Euclidean or including sample weights can further improve these results. Nevertheless, this improvement was generally problem specific in the sense that the same choice of parameters does not have equal effect on all databases.

	LETTER	IMAGE	PENDIGITS	PIMA
MAX ACC	97.1	97.8	97.9	76.2
DIM	16	10	14	6
CLASSIFIER	NDA2-NN	NDA-NN	NDA2-NN	CC-ICA-FS

Table 5.2: Maximum predictive accuracy for each benchmark database, together with the dimensionality where this accuracy was obtained and the classifier that achieved the result.

5.4.3 Real-world Data

For evaluating NDA on real-world data we chose a problem where high dimensional data is present but the number of available samples per class is low. Statistical methods, and in particular our CC-ICA method, are not applicable in this case since the number of samples is insufficient for robustly learning the necessary parameters. The problems we chose both come from the classical field of face recognition. Experiments were performed on the AR Face Database [97]. A first experiment on recognition subject to strong light variations and a second experiment on gender recognition.

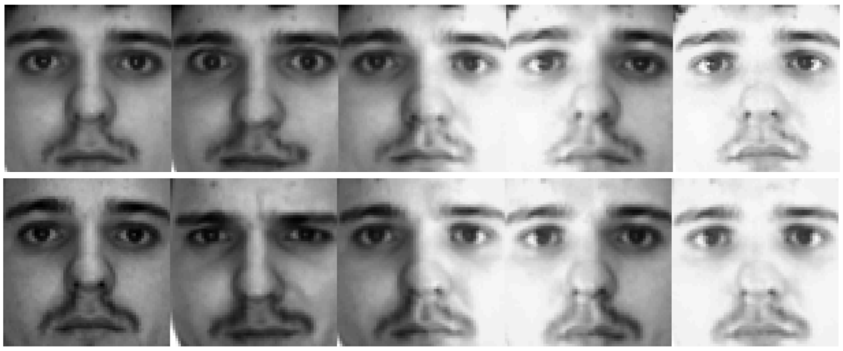


Figure 5.8: Training (top row) and test images chosen from the AR face database.

For the subject recognition experiment 5 training and 5 test images were chosen for each of the 115 subjects. Test and training were taken over different period of time and, as can be observed in fig.(5.8), the images are subject to strong light changes on each one of these sets. Images were subsampled to 24×24 pixels. We expect NDA to learn these light changes in the within-class whitening stage and reflect this normalization in the classification. Results are shown in Fig.(5.9). Here, NDA2 achieves the best results for practically all dimensions, being 87.7% the highest achieved accuracy. PCA instead, performs worse than any other discriminant technique. Classical NDA and FDA have very similar performance, below our modified NDA. This difference can only be blamed on the computation of the within-class scatter matrix and the normalization with respect to within-class variations. Actually, NDA and FDA share the same within-class scatter matrix, so it is the nonparametric approach NDA2 uses for computing this matrix what improves performance.

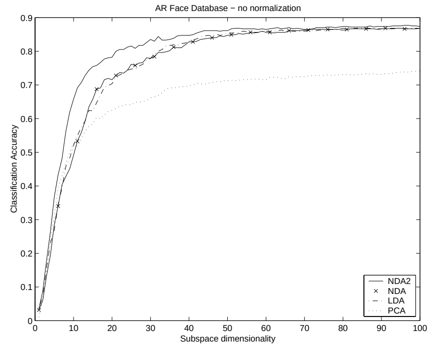
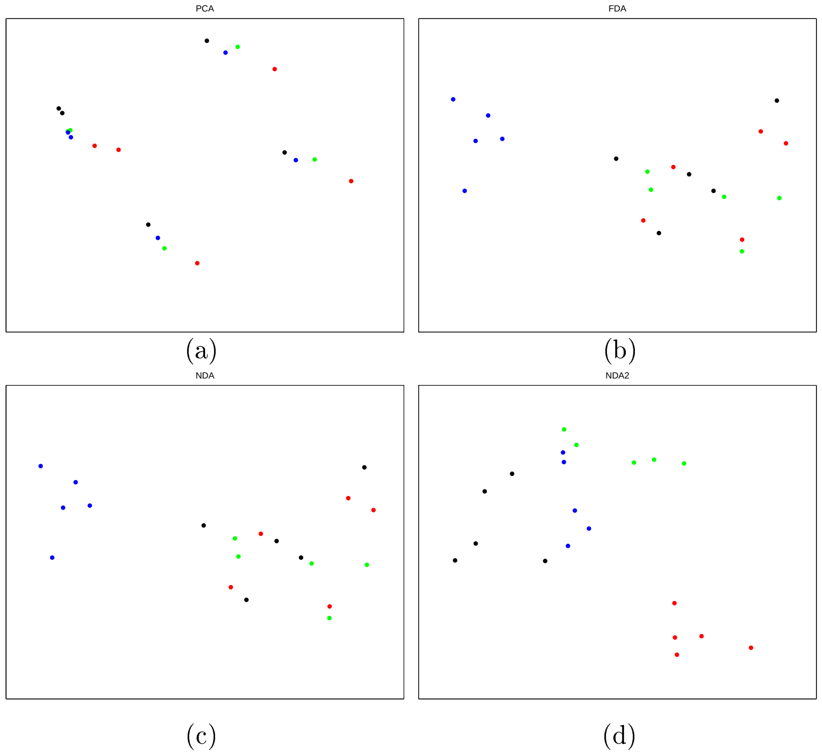


Figure 5.9: (a) Recognition accuracy on the AR face database with different subspace dimensionalities and no light normalization.

PCA performs considerably worse than all discriminant analysis techniques. Remember that PCA searches for directions of maximum variance and, as we noticed in section (2.2.2), illumination changes were frequently associated with maximum variance. In this case, where no normalization is performed this situation will be present with a negative effect in classification: for these principal components a subject will be closer to other subjects with similar illumination than to himself. Figure (5.10) illustrates this situation. For this figure, four evaluation subjects were chosen, with 5 samples per subjects. These samples were projected into each of the representations we used in the experiment: PCA, LDA, NDA and NDA2. We then plot the values for the first two components of each representation as dots in two dimensional space. Each dot corresponds to a subject and each of the subjects is identified with a unique color. Figure (5.10.a) shows the distribution for the first two principal components. As expected, subjects are clustered according to their illumination. Instead, the three discriminant techniques shown in figs. (5.10.a), (5.10.b) and (5.10.c) attempt to normalize this situation making use of the within-class scatter matrix. It can be seen that, at least visually, NDA2 is highly successful with this normalization. The projected samples are nearly clustered according to their class labels and these clusters have a positive effect on nearest neighbor classification.

For the second experiment, on gender recognition, images with strong light variation were discarded and both training and test sets merged in a single dataset that results in 256 male samples and 204 female samples taken from 115 subjects. In this particular case, a leave-one-out procedure was employed each time leaving out all the samples for a given subject. We chose this approach because we noticed that, in many cases, correct classification was achieved only thanks to the fact the same subject was found on the training set. Results shown in Fig.(5.9.b) obey the same procedure employed in the previous experiment, with subspace dimensionality con-



(a) (b) (c) (d)

Figure 5.10: First two components of the tested representations, corresponding to four different subjects on the evaluation set. There are five samples per subject, and each subject is identified with a unique color. Notice the positive effect of NDA on nearest neighbors, and the sensibility of PCA to illumination changes.

sidered from 1 to 100. In this case, images were normalized in variance. Once again, NDA2 outperforms the three other evaluated techniques, being 95.61% the highest achieved accuracy. It is quite surprising that using a single basis on this projection, accuracy already stands above 92%. Once again NDA and FDA perform similarly due to the limitations imposed by a parametric within-class scatter matrix. Learning gender was recently tested in [103], concluding support vectors were superior to any other traditional classifier in this task. So we applied this technique to our particular problem in order to compare results. After extensive testing, the best results were achieved with a RBF kernel with $\gamma = 3$: 93.86%.

5.5 Conclusions

The contribution in this chapter is related with the previous chapters in the sense we also search for a representation optimal for a given classifier. In this case, we focus on the nearest neighbor classifier. Searching a linear feature extraction technique that

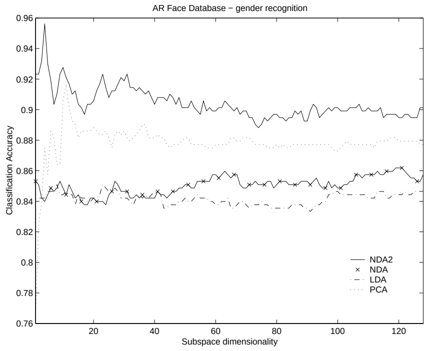


Figure 5.11: Gender recognition accuracy on the same database and using a ”leave subject samples out” procedure.

preserves nearest neighbor discriminability, results in a slight modification of nonparametric discriminant analysis. This modification affects only the within-class scatter matrix. The resulting technique has several advantages. The fact it works with intra and extra-class distances allows small class sample sizes (a minimum of two samples per class) and high dimensionality on the sample space. As all linear discriminant analysis techniques it provides a hierarchy on the features. Unlike Fisher discriminant analysis, the number of classes does not affect performance. More generally, its completely nonparametric nature implies no assumption on the class-conditional distributions, making it applicable to a wide range of situations.

Experiments with the obtained technique were performed on artificial, benchmark and real-world data. Artificial experiments allow comparison with the classical approach to nonparametric discriminant analysis and illustrate the difference with parametric discriminant analysis techniques such as Fisher discriminant analysis. Tests with benchmark databases show that our technique can in effect enhance nearest-neighbor estimation or, in the worst case, leave it unaltered. This improvement can be observed even when the projection is done on lower dimensional spaces. For one of the databases, unprecedented results were obtained.

The real-world experiments were made on the classical problems of face recognition robust to illumination variation and gender recognition. In the first experiment, we expect the nonparametric within-class scatter matrix to absorb the intra-class variations, unsupervisedly correcting the illumination. This seems to be the case and our approach performs considerably better than any other technique, particularly at low dimensions. Through this experiment, we can also observe the actual effect the projection has on the evaluation samples. Our approach to nonparametric discriminant analysis generates class clusters which are very convenient to the nearest neighbor classifier. Results are also satisfactory in the second experiment, gender recognition, where we have also compared our approach with a common choice of classifier in this problem: support vector machines.

Chapter 6

Concluding Remarks

6.1 Conclusions

This thesis is focused on the problem of linear feature extraction for the task of statistical classification of visual data, usually a particular case of classification of high-dimensional data. The main inconvenient faced by statistical classification when dealing with this kind of data is proper density estimation due to dimensionality concerns. Additionaly the data can be contaminated by noise, and not necessarily all measurements contribute to classification.

Linear feature extraction techniques are helpful for modeling low dimensional pattern structures present in high dimensional data. In our case these pattern structures, will only be useful if they benefit classification. It is a well known fact that, the Bayes error of the extracted features will be equal or higher than the Bayes error of the original data. From this fact, a highly relevant premise that drives this whole thesis is derived: a statistical classifier will only benefit from a linear transformation if the projection improves the estimation of the conditional densities. Linear transformations can provide this benefit, for instance, by reducing dimensionality preserving relevant information. This approach can be seen as a particular case of noise reduction, where anything not contributing to classification is considered noise. Discriminant Analysis or Principal Component techniques usually take this approach. A more direct method can result from considering linear transformations exclusively from the simplification they might provide on density estimation.

Independent component analysis can provide this simplicity in terms of density marginalization: density estimation in feature space is reduced to a number of unidimensional estimations. Additionally, higher-order dependencies between the features are removed allowing single feature interpretations. Our first contribution, in the field of shape analysis, is focused on these properties using independent component analysis for representing point distribution models. Towards this end,

- We have noticed that independent component analysis can provide unique shape descriptors with desirable predictive and intuitive properties based on the simplicity this technique provides to the density estimation and the statistical independence of the extracted features. These features were named independent modes of variation.

- By making use of the statistical model associated to independent component analysis we provide a robust solution for the problems of analysing shape feasibility and searching for the nearest feasible shape. Additional properties of our model for non-rigid shape variation can also unsupervisedly provide a set of shape prototypes useful for shape database querying.

We then propose to take advantage of independent component analysis in the context of statistical pattern classification. Since Bayesian classification makes use of the conditional densities, the choice of any representation oriented to simplifying density estimation necessarily implies the use of class-conditional representations. In this case, and under certain assumptions, independent component analysis can provide a framework where conditional independence can be assumed. Naive Bayes appears as the naturally associated classifier for this situation. Inversely, we can claim an optimal representation for naive Bayes classification. In this sense, the following contributions have been made,

- We have formalized the theory for linear class-conditional representations, adapting Bayesian decision to this scheme.

- When independent component analysis is introduced as a representation, we obtain the context we named as class-conditional independent component analysis (CC-ICA). Adapting the classifier to CC-ICA results in a modified naive Bayes classifier. These results are summarized in two algorithms, called CC-ICA-Train and CC-ICA-Test.

- Highly redundant data usually yields sparse independent components. If sparsity can be safely assumed, our classifier can make use of this prior knowledge through specific density functions. Classification can be also understood in terms of the sparse coding principle, as can be observed in the experiments.

The main disadvantage of class-conditional representations is that they fail to learn the relationship among classes. Some of this information, such as discriminability, is of particular importance for the task of classification. The ability of an extracted feature to discriminate among classes can be used as a criterion for feature selection. So we considered the problem of selecting discriminative features under the assumption of class-conditional independence, which is the situation within a CC-ICA representation. Feature selection can, in some situations, enhance classification. In most of the cases it can greatly reduce the computational load of our algorithm at small or no costs in classification accuracy. In this sense,

- The class separability measure of divergence has been introduced as a feature subset selection criterion adequate for our model. Our model assumes nongaussian data, and divergence makes no prior assumption on the data distribution. Our model assumes independence and divergence greatly benefits from this assumption becoming a sum of unidimensional divergences.

- It has been shown that, when the conditional independence assumption holds, feature subset selection using divergence can be performed without the need for an exhaustive search.

- Moreover, divergence can be understood in terms of log-likelihood ratios and adapted to class-conditional representations yielding a class separability measure adapted to this case. This measure contemplates the possibility of selecting different features for different classes.

Behind these contributions is the attempt to provide a unified framework for the design of a statistical pattern classifier, opposed to the alternative where each stage in a pattern classification process is considered independently from the other stages. In our case, once the initial assumptions are made, feature extraction, selection and finally classification are naturally associated among each other. For our proposed scheme we have the following initial assumptions

- The ICA assumptions should hold for each class. Mainly that each class is a linear mixture of independent components and of these components, at most one is Gaussian.

- We count with enough samples per class in order to perform a robust ICA estimation. The number of required samples is closely related to domain space dimensionality.

If these assumptions hold we have shown that each of the stages of the process is theoretically justified and intimately linked with the other stages. Actually if there is no dimensionality reduction, the error of our classifier is the Bayes error. If the assumptions do not hold, the whole process is invalidated. Nevertheless, there are ways of relaxing these assumptions, some of them we will mention in the next section.

The line of reasoning we used to relate independence with classification can also be stated as *finding a representation optimal for a classifier* and can be extended to the nonparametric case,

- We have shown that searching a linear feature extraction technique that preserves nearest neighbor discriminability results in a slight modification of nonparametric discriminant analysis. This modified algorithm joins the advantages of being a completely nonparametric approach (no assumption on the data distribution, the number of classes does not limit the maximum number of extracted features, it can work on reduced sample sets with any dimensionality) with the property of having a simple naturally associated classifier.

6.2 Future Work

The work covered in this thesis provides a number of areas of interest that may be worth further investigation. Among others, it may be interesting to consider the following lines for further research:

Independent Modes of Variation

Even though we presented a model for the analysis of shape feasibility through independent modes of variation, this model has yet to be validated. This can be done through experiments that evaluate performance in indexing shape databases, or by including our model into a more general framework for unsupervised classification of non-rigid shape deformations. Any new results on unsupervised generation of point distribution models from object images would surely impulse this line of research.

Extensions of Independent Component Analysis

When working with visual data, binary representations arise as a simple and economic way of extracting information from the image, and several computer vision and image processing algorithms have a binary response. This results in binary high-dimensional data. We believe that adapting the estimation of independent component analysis to this kind of data can be of utility. From our perspective, statistical classification on statistically independent binary data is straightforward. Moreover, data with binary or strong binary nature can arise from many other applications. A particularly interesting application is the problem of classifier combination, where it has been shown that the combination of statistically independent classifiers yields better results than the combination of classifiers with strong dependencies among each other. Given D classifiers and different training datasets, classifier response can be modeled as a binary vector indicating correct or incorrect decisions. In this case, binary independent component analysis could unsupervisedly learn a way of linearly combining these classifiers into a subset of statistically independent classifiers.

Another interesting extension for independent component analysis can result from its supervised implementation. In this case a direct modification of the objective function could be used to contemplate higher-order discriminability among classes. Consider, for instance, the problem of deciding between two classes with identical mean and covariance. An objective function based on measures of nongaussianity similar to those used in independent component analysis could provide a discriminative representation based on high-order information.

Relaxing the CC-ICA Assumptions

In general, relaxing the assumptions of class-conditional independent component analysis involves using a representation other than ICA for each class. A natural extension, would be to make use of mixtures of independent component analysers. This is comparable to extending the Gaussian maximum likelihood classifier to mixtures of Gaussians, but in the nongaussian case. Other similar extensions could contemplate nonlinear independent component analysis.

The close link between sparsity and independence could allow to hold the independence assumption in the presence of sparse data not necessarily obtained through independent component analysis. Prior assumptions on the model for nonnegative matrix factorization can force sparsity in the encodings. In this case, if an adequate density model is considered for the sparse and nonnegative data, statistical classification could be simply extended to the nonnegativity context. This would combine the advantages of NMF as a representation with the advantages characteristic of statistical classifiers.

In general class-conditional representations can be used under any linear feature extraction technique other than ICA, given that we have reasons to think that the technique can simplify or improve the estimation of the conditional densities. An example of this was given in the experiments when working with class-conditional PCA, which showed improved performance with respect to global PCA.

Modifying the Naive Bayes Classifier

Instead of relaxing the assumptions we could focus on modifying the classifier. If CC-ICA is used, this is a delicate issue since choosing a classifier other than naive Bayes is very likely to degrade all the simplicity we might have gained in terms of density estimation. Nevertheless, the independence assumption can still be of utility without using the maximum a posteriori decision scheme. An example of this can be found on pairwise classifiers. The main idea of this approach is to replace the global maximum on the posterior probabilities by pairwise class comparisons. In this case, after comparing the likelihood on all classes among each other, classification is performed through some kind of voting scheme. It can be seen that CC-ICA, and even CC-ICA-FS can be easily adapted to this framework.

Class Separability Criterions

Other class separability criterions such as the Jeffries Matusita distance or modified Fisher Ratio can also benefit from the independence assumption. In particular, the Jeffries Matusita distance, has desirable properties which are not present in divergence: its bounded nature. A thorough comparison of these criteria has still to be done.

There are also limitations derived from extending divergence to the multiclass case. Features selected according to their average discriminability among all classes are not necessarily the best features for distinguishing between two classes. But the CC-ICA context already contemplates the possibility of using different features per class. In this case, using a pairwise classifier would allow direct use of the pairwise divergence definition, obtaining optimal discrimination among any two classes.

Extensions to Nonparametric Discriminant Analysis

Natural extensions to NDA involve the evaluation of different distances, the actual effectivity of the proposed sample weights and robustness to dimensionality and sample set size. Also, there have been proposed several boosting algorithms which make use of discriminant directions in the construction of weak classifiers. To our knowledge none of these algorithms make use of nonparametric discriminant analysis. We believe that boosted nonparametric discriminant analysis is an interesting line of research that could eventually combine the advantages of a nonparametric approach with the advantages of boosting.

Appendix A

Notation

In general, boldface fonts indicate vectors and matrices, superscripts are used for specifying class attributes, and subscripts unambiguously indicate sample number or component number (for a vector). For instance, if our working data for class C^k is formed by N^k D-dimensional samples and a matricial notation is used, we might represent this data as a matrix $\boldsymbol{X}^k = \{x^k\}_{ij}$, where x^k_{ij} stands for the i-th component of the j-th sample for class C^k. Unless specifically indicated otherwise, table (A.1) indicates the chosen variable names or general notation principles used throughout this thesis.

K,k	number of classes and class iterator (also used for number of nearest neighbors)
C^k	class k
N,n	number of samples and sample iterator
N^k	number of class k samples
D,d	dimensionality and dimensionality iterator
M,m	feature dimensionality (after a linear projection)
M^k	class-conditional feature dimensionality
$\boldsymbol{x}$	random vector
$\boldsymbol{z}$	white random vector
$\boldsymbol{y}$	general extracted features
$\boldsymbol{s}$	independent components
$\boldsymbol{W}$	$M \times D$ projection or filter matrix
$\boldsymbol{w}_m$	m^{th} row of projection matrix
$\boldsymbol{A}$	$D \times M$ basis or mixture matrix
$\boldsymbol{V}$	$M \times D$ matrix with covariance eigenvectors as rows
$\boldsymbol{D}$	$M \times M$ matrix with M largest covariance eigenvalues on the diagonal
λ_m	m^{th} covariance eigenvalue, in descending order
Σ, Σ^k	global and class k covariance
μ, μ^k	global and class k mean
$\bar{\boldsymbol{x}}$, $\bar{\boldsymbol{x}}^k$	global and class k sample mean
$p(\boldsymbol{x}\|C^k)$	class-conditional density
$p^k(\boldsymbol{s})$	class-conditional density for k^{th} CC-ICA representation
ϵ,$\epsilon_{\boldsymbol{W}}$	Bayes error and Bayes error for projected data
$E^k\{\}$	class-conditional expectation operator
$\|\boldsymbol{W}\|$	matrix determinant (absolute value)
$\boldsymbol{W}^T$	transpose operator
$\mathcal{L}(\boldsymbol{\theta})$	likelihood, dependent on parameter vector $\boldsymbol{\theta}$
$g(\boldsymbol{x})$	score function
$\mathcal{H}(\boldsymbol{x})$	differential entropy
$\mathcal{J}(\boldsymbol{x})$	negentropy
$\mathcal{I}(\boldsymbol{x})$	mutual information
$\mathcal{KL}(p_1,p_2)$	Kullback-Leibler distance or divergence between probability distributions
$\mathcal{D}(p_1,p_2)$	divergence between probability distributions
$\mathcal{D}^{ij}$	divergence between class-conditional densities for classes C^i and C^j
$\hat{\mathcal{D}}(p_1,p_2)$	transformed divergence
$kurt(x)$	kurtosis for random variable x
$\mathcal{S}$	separability measure
$\mathcal{S}_{\mathcal{B}}$	bhattacharyya bound
$\mathcal{S}_{\mathcal{JM}}$	jeffries-Matusita distance
$\boldsymbol{S}_b$	between-class scatter matrix
$\boldsymbol{S}_w$	within-class scatter matrix
$\boldsymbol{x}^{I,E}$	intra- and extra- class nearest neighbor or nearest neighbor average
$\boldsymbol{\Delta}^{I,E}$	intra- and extra- class differences
$\boldsymbol{\Delta}_{\boldsymbol{W}}^{I,E}$	projected intra- and extra- class differences

Table A.1: Variable name convention and notation used throughout the thesis.

Appendix B

Multispectral Images

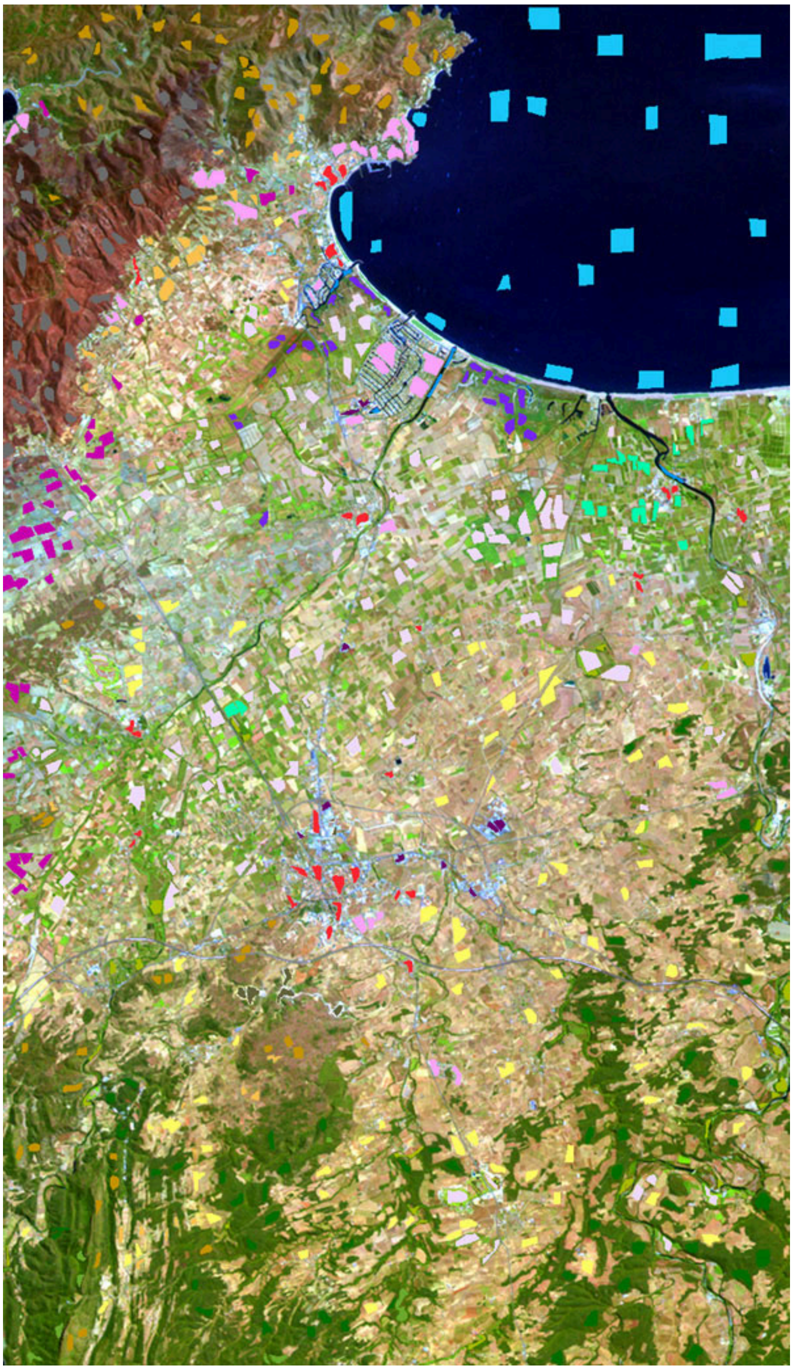


Figure B.1: False color composite of the region of study in the land use classification experiment from chapter 3. Training and test regions are overlapped and have themsleves a particular color coding.

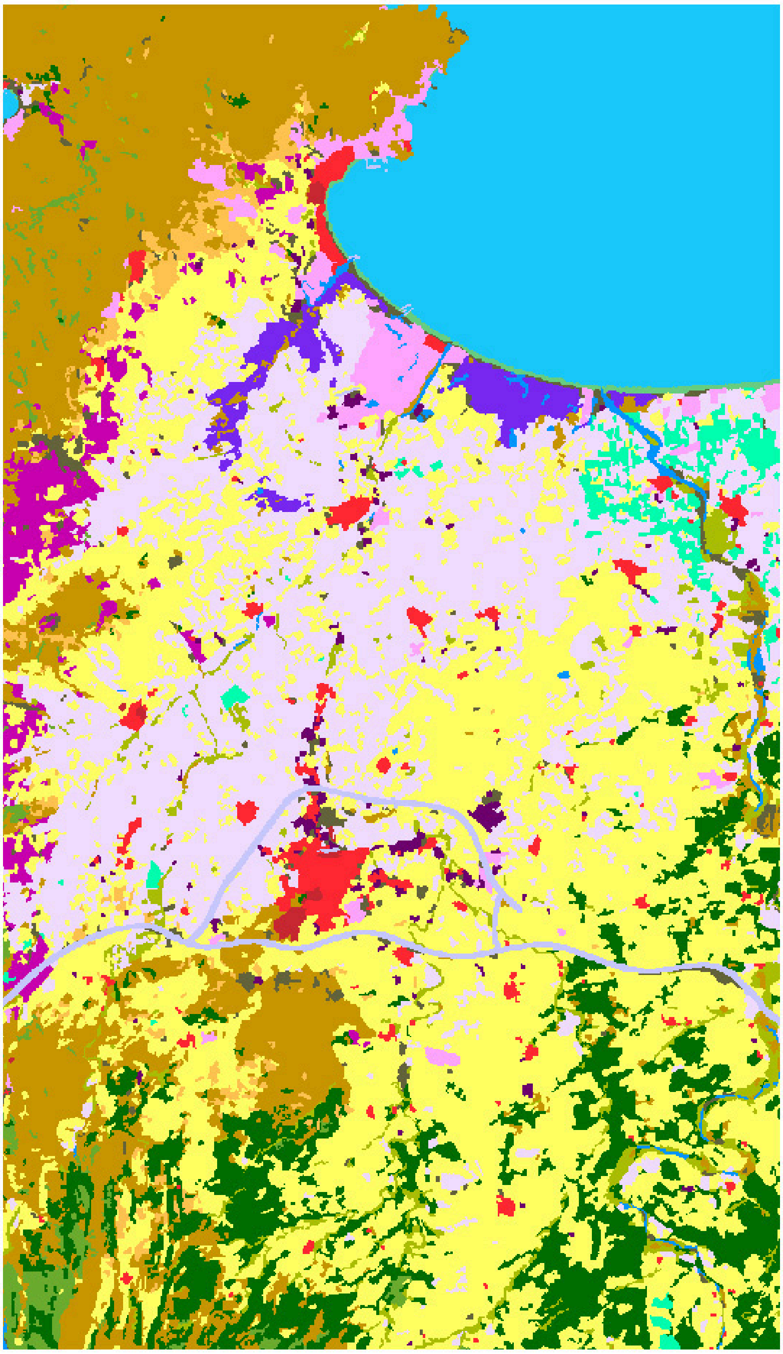


Figure B.2: Reference hand-made land use map from year 1998.

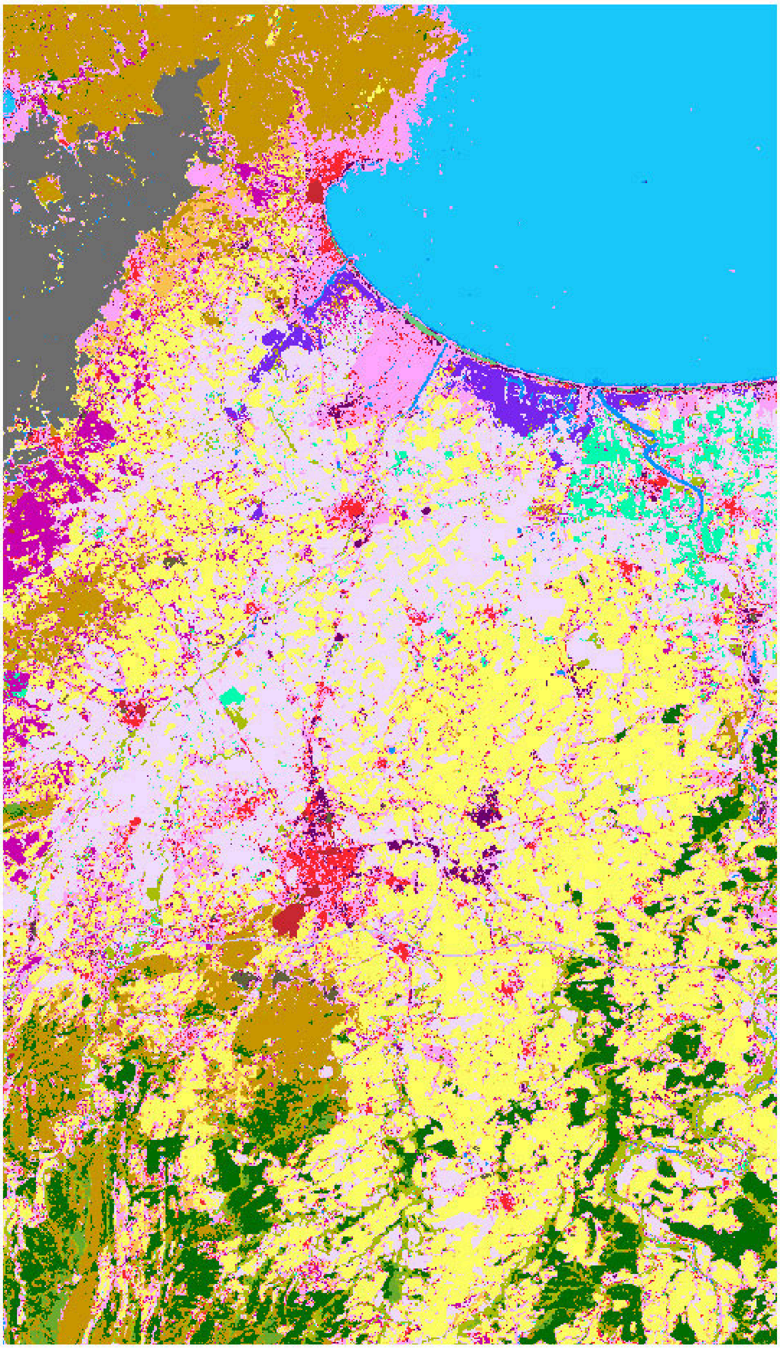
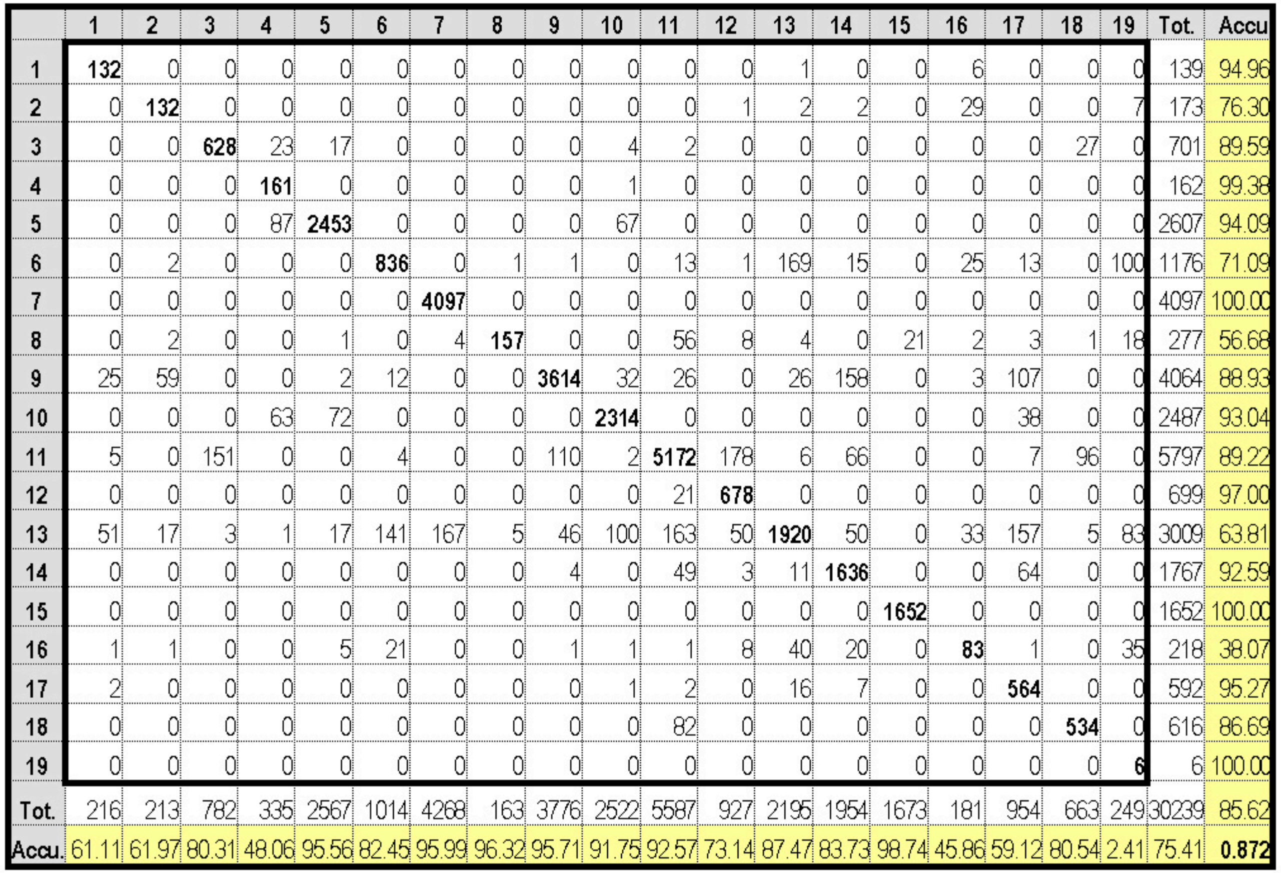


Figure B.3: Land use map obtained using the CC-ICA algorithm on multispectral images gathered in 1999 and 2000.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Tot.	Accu
1	**132**	0	0	0	0	0	0	0	0	0	0	0	1	0	0	6	0	0	0	139	94.96
2	0	**132**	0	0	0	0	0	0	0	0	0	1	2	2	0	29	0	0	7	173	76.30
3	0	0	**628**	23	17	0	0	0	0	4	2	0	0	0	0	0	0	27	0	701	89.59
4	0	0	0	**161**	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	162	99.38
5	0	0	0	87	**2453**	0	0	0	0	67	0	0	0	0	0	0	0	0	0	2607	94.09
6	0	2	0	0	0	**836**	0	1	1	0	13	1	169	15	0	25	13	0	100	1176	71.09
7	0	0	0	0	0	0	**4097**	0	0	0	0	0	0	0	0	0	0	0	0	4097	100.00
8	0	2	0	0	1	0	4	**157**	0	0	56	8	4	0	21	2	3	1	18	277	56.68
9	25	59	0	0	2	12	0	0	**3614**	32	26	0	26	158	0	3	107	0	0	4064	88.93
10	0	0	0	63	72	0	0	0	0	**2314**	0	0	0	0	0	0	38	0	0	2487	93.04
11	5	0	151	0	0	4	0	0	110	2	**5172**	178	6	66	0	0	7	96	0	5797	89.22
12	0	0	0	0	0	0	0	0	0	0	21	**678**	0	0	0	0	0	0	0	699	97.00
13	51	17	3	1	17	141	167	5	46	100	163	50	**1920**	50	0	33	157	5	83	3009	63.81
14	0	0	0	0	0	0	0	0	4	0	49	3	11	**1636**	0	0	64	0	0	1767	92.59
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1652**	0	0	0	0	1652	100.00
16	1	1	0	0	5	21	0	0	1	1	1	8	40	20	0	**83**	1	0	35	218	38.07
17	2	0	0	0	0	0	0	0	0	1	2	0	16	7	0	0	**564**	0	0	592	95.27
18	0	0	0	0	0	0	0	0	0	0	82	0	0	0	0	0	0	**534**	0	616	86.69
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**6**	6	100.00
Tot.	216	213	782	335	2567	1014	4268	163	3776	2522	5587	927	2195	1954	1673	181	954	663	249	30239	85.62
Accu.	61.11	61.97	80.31	48.06	95.56	82.45	95.99	96.32	95.71	91.75	92.57	73.14	87.47	83.73	98.74	45.86	59.12	80.54	2.41	75.41	**0.872**

Figure B.4: Confusion matrix for the evaluation set using the CC-ICA-NBGAUMIX classifier.

Appendix C

Publications

Independent component analysis is first related with point distribution models in the master's thesis,

- M. Bressan Linear point distribution models by independent component analysis. MsD Thesis, CVC Tech. Rep. 48, Centre de Visió per Computador, September 2000.

This research results in the introduction of independent modes of variation as shape descriptors. This contribution, explained in section (2.4.6), is presented as a book chapter in

- M. Bressan and J. Vitrià. Independent modes of variation in point distribution models. In L.P. Cordella and G. Sanniti di Baja, editors, *Visual Form 2001, 4th International Workshop on Visual Form*, LNCS 2059, pages 123–134. Springer Verlag, Italy, May 2001.

Further research on nonsupervised representations leads to nonnegative matrix factorization. The technique of weighted nonnegative matrix factorization (WNMF), detailed in section (2.3.1) appears in

- D. Guillamet, M. Bressan, and J. Vitrià. A weighted nonnegative matrix factorization for local representations. In *IEEE CSC in Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages 942–947, USA, December 2001.

The first results of pattern classification using an ICA representation applied to the problem of object recognition through local color histograms

- M. Bressan, D. Guillamet, and J. Vitrià. Using an ICA representation of local color histograms for object recognition. In *3rd Catalonian Conference on Artificial Intelligence*, pages 300–307, Spain, October 2000.

A formal exposition of these results, detailed in section (3.5.3), is published in

- M. Bressan, D. Guillamet, and J. Vitrià. Using an ICA representation of local color histograms for object recognition. *Pattern Recognition*, 36(3):691–701, March 2003.

The extension of ICA classification to the general case of high dimensional data and beyond the specific problem of object recognition gives rise to additional experiments on benchmark databases. Results are exposed in

- M. Bressan, D. Guillamet, and J. Vitrià. Using an ICA representation of high dimensional data for object recognition and classification. In *IEEE CSC in Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages 1004–1009, USA, December 2001.

A more applied perspective on ICA and classification and the first general remarks on class-conditional representations, and more specifically class-conditional independent component analysis (CC-ICA) are published in

- M. Bressan, D. Guillamet, and J. Vitrià. Multiclass object recognition using class-conditional independent component analysis. *Cybernetics and Systems*, 2003. IN PRESS.

Comparative analyses that state CC-ICA as a way of improving the naive Bayes classifier appear in

- M. Bressan and J. Vitrià. Independent component analysis and naive bayes classification. In J.J. Villanueva, editor, *2nd IASTED International Conference on Visualization, Imaging and Image Processing (VIIP 2002)*, volume 1, pages 496–501, Spain, September 2002.

and in the book chapter

- M. Bressan and J. Vitrià. Improving naive Bayes using class-conditional ICA. In F. Garijo, J. Riquelme, and M.Toro, editors, *Advances in Artificial Intelligence, VIII Iberoamerican conference on Artificial Intelligence (Iberamia 2002)*, LNAI 2527, pages 1–10. Springer Verlag, Spain, November 2002.

These last six publications form the theoretical core of CC-ICA, detailed in chapter 3 of this thesis. The experiment on the performance of this technique on multispectral data, detailed in section (3.5.3), is exposed in

- M. Bressan, P. Radeva, and J. Vitrià. Feasibility analysis for the nonsupervised generation of a land use map of Catalonia Tech. Report 58, Centre de Visió per Computador, November 2001.

The experiments and comparisons on CC-ICA applied to the visual inspection of cork stoppers, detailed in section (3.5.3) appear in

- P. Radeva, M. Bressan, A. Tobar, and J. Vitrià. Real-time inspection of cork stoppers using parametric methods in high dimensional spaces. In *IASTED International Conference on Signal and Image Processing (SIP 2002)*, volume 1, pages 480–484, USA, August 2002.

and in the book chapter,

- P. Radeva, M. Bressan, A. Tobar, and J. Vitrià. Bayesian classification for inspection of industrial products. In M.T. Escrig, F. Toledo, and E. Golobardes, editors, *5th Catalonian Conference on Artificial Intelligence, Topics on Artificial Intelligence*, LNAI 2504, pages 399–407. Springer Verlag, Spain, October 2002.

The first approach to feature selection in a space where class-conditional independence is met appears in the book chapter

- M. Bressan and J. Vitrià. Feature subset selection in an ICA space. In M.T. Escrig, F. Toledo, and E. Golobardes, editors, *5th Catalonian Conference on Artificial Intelligence, Topics on Artificial Intelligence*, LNAI 2504, pages 196–206. Springer Verlag, Spain, October 2002.

The theory and results in chapter 4 are an extension of the publication

- M. Bressan and J. Vitrià. On the selection and classification of independent features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003. IN PRESS.

This contribution explores the proposed feature selection technique, its relationship with CC-ICA and provides the feature selection algorithm CC-ICA-FS, found in section (4.3.1).

Finally, results from chapter 5, on finding the optimal representation for the nearest neighbor classifier are published in

- M. Bressan and J. Vitrià. Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recognition Letters*, 2003. ACCEPTED - UNDER REVIEW.

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *B. N. Petrox and F. Caski, Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, 1973.

[2] S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. *Advances for Neural Information Processing*, 8:757–763, 1996.

[3] S-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[4] S-I. Amari, T-P Chen, and A. Cichocki. Stability analysis of learning algorithms for blind source separation. *Neural Networks*, 10(8):1345–1351, 1997.

[5] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[6] A.D. Back and A.S. Weigend. A first application of independent component analysis to extracting structure from stock returns. *International Journal on Neural Systems*, 8(4):473–484, 1997.

[7] H.B. Barlow. Possible principles underlying the transformations of sensory messages. *Sensory Communication*, pages 217–234, 1961.

[8] H.B. Barlow. *What is the computational goal of the neocortex?* MIT Press, Cambridge, MA, 1994.

[9] H.B. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12:241–253, 2001.

[10] A. Barros, A. Mansour, and N. Ohnishi. Removing artifacts from electrocardiographic signals using independent component analysis. *Neurocomputing*, 22:173–186, 1998.

[11] M. Bartlett, H. Lades, and T. Sejnowski. Independent component representations for face recognition. In T. Rogowitz and B. Pappas, editors, *Proceedings of the SPIE Symposium on Electonic Imaging: Science and Technology; Human Vision and Electronic Imaging III*, volume 3299, pages 528–539, San Jose, CA, January 1998. SPIE Press.

[12] A. Bell and T. Sejnowski. An information-maximization approach for blind signal separation. *Neural Computation*, 7:1129–1159, 1995.

[13] A. Bell and T. Sejnowski. Learing higher-order structure of a natural sound. *Network*, 7:261–266, 1996.

[14] A. Bell and T. Sejnowski. The 'independent components' of natural scenes are edge filters. *Neural Computation*, 11:1739–1768, 1999.

[15] R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, New Jersey, 1961.

[16] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1997.

[17] C.M. Bishop and N.D. Lawrence. Variational bayesian independent component analysis. Technical report, Computer Laboratory, University of Cambridge, 2000.

[18] C.L. Blake and C.J. Merz. Uci repository of machine learning databases, 1998.

[19] B.Moghaddam, T.Jebara, and A.Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, November 2000.

[20] J.F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4:112–114, 1997.

[21] J.F. Cardoso and A. Souloumiac. Blind beamforming for non gaussian signals. In *IEEE Proceedings-F*, volume 140(6), pages 362–370, 1993.

[22] O. Chapelle, P. Haffner, and V. Vapnik. SVM's for histogram-based image classification. *IEEE Transactions Neural Networks*, 10(5):1055–1065, September 1999.

[23] V. Cherkassky and F. Mulier. *Learning From Data.* Wiley - Interscience, New York, 1998.

[24] Peter Clark and Tim Niblett. The cn2 induction algorithm. *Machine Learning*, 3:261–283, 1989.

[25] P. Comon. Independent component analysis - a new concept? *Signal Processing*, 36:287–314, 1994.

[26] T.F. Cootes and C.J. Taylor. A mixture model for representing shape variation. In *Clark A.F., ed. British Machine Vision Conference 1997, BMVC'97*, volume 1, pages 110–119. University of Essex, UK:BMVA, 1997.

[27] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[28] Corel Corporation. Corel stock photo library. *Ontario, Canada*, 1990.

[29] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13:21–27, January 1967.

[30] B.V. Dasarathy. *NN Pattern Classification Techniques.* IEEE Computer Society Press, Los Alamitos, CA, 1990.

[31] A. P. Dawid. Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 41:1–31, 1979.

[32] V. Colin de Verdière and J. L. Crowley. Visual recognition using local appearance. In *Proceedings of the European Conference on Computer Vision (ECCV'98)*, pages 640–654, 1998.

[33] A.P. Dempster, N. Laird, and D. Rubin. Maximum likelihood for incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[34] P.A. Devijer and J. Kittler. On the edited nearest neighbor rule. In *Proceedings of the 5th International Conference on Pattern Recognition*, pages 72–80, 1980.

[35] K.I. Diamantaras and S.Y. Kung. *Principal Component Neural Networks: Theory and Applications.* Wiley, 1996.

[36] P. Domingos and M. J. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

[37] D.L. Donoho. Nature vs. math: Interpreting independent components in light of recent work in harmonic signals. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 459–470, Helsinki, Finland, 2000.

[38] R. Duda, P. Hart, and D. Stork. *Pattern Classication.* John Wiley and Sons, Inc., New York, 2nd edition, 2001.

[39] D.J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[40] G.D. Finlayson, S.S. Chatterjee, and B.V Funt. Color angular indexing. In *Proc. of ECCV'96*, volume 2, pages 16–27, 1996.

[41] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 37(II):179–188, 1936.

[42] E. Fix and J.L. Hodges. Discriminatory analysis: Nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, February 1951.

[43] T. Fogarty. First nearest neighbor classification on frey and slate's letter recognition problem. *Machine Learning*, 9:387–388, 1992.

[44] D.H. Foley and J.W. Sammon. An optimal set of discriminant vectors. *IEEE Transactions on Computing*, C-24:281–289, March 1975.

[45] P. W. Frey and D. J. Slate. Letter recognition using holland-style adaptive classifiers. *Machine Learning*, 6:161–182, March 1991.

[46] J.H. Friedman. An overview of predictive learning and function approximation. In V. Cherkassky, J.H. Friedman, and H. Wechsler, editors, *From Statistics to Neural Networks, NATO ASI Series F*, volume 136, pages 1–61, New York, 1994. Springer Verlag.

[47] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

[48] K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition.* Academic Press, Boston, MA, 1990.

[49] K. Fukunaga and J.M. Mantock. Nonparametric discriminant analysis. *IEEE Transactions on PAMI*, 5:671–678, November 1983.

[50] B. V. Funt and G. D. Finlayson. Color constant color indexing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995.

[51] M. Girolami. An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*, 10(8):2103–2114, 1998.

[52] M. Girolami. *Self-Organising Neural Networks - Independent Component Analysis and Blind Source Separation.* Springer-Verlag, 1999.

[53] J.C. Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

[54] D. Guillamet and J. Vitria. A comparison of local versus global color histograms for object recognition. In *Proceedings. ICPR 2000*, volume 2, pages 422–425, 2000.

[55] H.H. Harman. *Modern Factor Analysis.* University of Chicago Press, second edition, Chicago, IL, 1967.

[56] A.J. Heap and D. Hogg. 3d deformable hand models. In *Gesture Workshop, York, UK*, pages 131–139, March 1996.

[57] A.J. Heap and D. Hogg. Improving specificity in pdms using a hierarchical approach. In *Clark A.F., ed. British Machine Vision Conference 1997, BMVC'97*, volume 1, pages 80–89. University of Essex, UK:BMVA, 1997.

[58] J. Hilden. Statistical diagnosis based on conditional independence does not need it. *Comput. Biol. Med.*, 14(4):429–435, 1984.

[59] P.O. Hoyer. *Independent Component Analysis in Image Denoising.* PhD thesis, Helsinki University of Technology, 1999.

[60] P.O. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–200, 2000.

[61] G.F. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, IT-14(1):55–63, 1968.

[62] A. Hyvärinen. New approximatins of differential entropy for independent component analysis and projection pursuit. *Advances in Neural Processing Systems*, 10:273–279, 1998.

[63] A. Hyvärinen. The fastica algorithm, 2001.

[64] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley and Sons, 2001.

[65] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1999.

[66] Anil K. Jain and Douglas E. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[67] I.T. Jollife. *Principal component analysis.* Springer Verlag, New York, 1986.

[68] M. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society, series A*, pages 136–150, 1987.

[69] J. Joutsensalo and T. Ristaniemi. Learning algorithms for blind multiuser detection in cdma downlink. In MIT Press, editor, *Proc. IEEE 9th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC98)*, pages 267–270, Boston, MA, 1998.

[70] T.P. Jung, C. Humphries, T-W. Lee, S. Makeig, M.J. McKeown, V. Iragui, and T. Sejnowski. Extended ICA removes artifacts from electroencephalographic recordings. In MIT Press, editor, *Advances in Neural Information Processing Systems*, volume 10, 1998.

[71] C. Jutten and J. Herrault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[72] A. Kaban and M. Girolami. Clustering of text documents by skewness maximization. In MIT Press, editor, *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 435–440, Helsinki, Finland, 2000.

[73] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Trans. on Communication Technology COM-15*, 1:52–60, February 1967.

[74] J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear pca type learning. *Networks*, 7(1):113–127, 1994.

[75] M. Kearns and U.V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, Cambridge, MA, 1994.

[76] M. Kendall. *Multivariate Analysis.* Charles Griffin and Company, 1975.

[77] M. Kendall and A. Stuart. *The Advanced Theory of Statistics.* Charles Griffin and Company, 1958.

[78] J. Kittler. Feature set search algorithms. In C.H. Chen, editor, *Pattern Recognition and Signal Processing*, pages 41–60, 1978.

[79] T. Kolenda, L.K. Hansen, and S. Sigurdsson. Independent components in text. In M. Girolami, editor, *Advances in Independent Component Analysis*, pages 235–256. Springer-Verlag, 2000.

[80] S. Krishnan, K. Samudravijaya, and P.V.S. Rao. Feature selection for pattern classification with gaussian mixture models: A new objective criterion. *Pattern Recognition Letters*, 17:803–809, 1996.

[81] M. Kubat, D. Flotzinger, and G. Pfurtscheller. Discovering patterns in eeg-signals: Comparative study of a few methods. In *Proceedings of European Conference on Machine Learning*, pages 366–371, Berlin, 2001. Springer Verlag.

[82] J. L., T. S. Jayram, and I. Rish. Recognizing end-user transactions in performance management. In *Proceedings of AAAI-2000*, pages 596–602, Austin, Texas, 2000.

[83] G. Lugosi L. Devroye, L. Gyorfi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.

[84] J. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:154–174, 1977.

[85] P. Langley, W. Iba, and K. Thompson. An analysis of bayesian classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-9)*, pages 223–228, San Jose, CA, 1992. Morgan Kaufman.

[86] D.D. Lee and H.S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, July 1999.

[87] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural and Information Processing Systems*, 13:–, July 2001.

[88] T. Lee, M. Lewicki, and T. Seynowski. A mixture models for unsupervised classification of non-gaussian sources and automatic context switching in blind signal separation. *IEEE Transactions on PAMI*, 22(10):1–12, 2000.

[89] T.W. Lee, M. Girolami, and T.J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11:609–633, 1998.

[90] M. Lewicky and T.J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

[91] D. Lewis. Naive bayes at forty: The independence assumption in information retrieval. In Claire N'edellec and C'eline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, volume 1398.25, pages 4–15. Springer Verlag, Heidelberg, DE, 1998.

[92] C. Liu and H. Wechsler. Comparative assessment of independent component analysis (ICA)for face recognition. In *Second International Conference on Audio- and Video-based Biometric Person Authentication, AVBPA'99*, Washington D.C., 1999.

[93] N.K. Logothetis and D.L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19:577–621, 1996.

[94] A. Lopez, F. Lumbreras, J. Serrat, and J.J. Villanueva. Evaluation of methods for ridge and valley detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:327–335, 1999.

[95] D.G. Luenberger. *Linear and Nonlinear Programming, Second Edition.* Addison-Wesley, Reading, MA, 1984.

[96] T. Marill and D. Green. On the effectiveness of receptors in recognition systems. *IEEE Trans. on Information Theory*, 9:1–17, 1963.

[97] A. Martı́nez and R. Benavente. The AR face database. Technical Report 24, Computer Vision Center, June 1998.

[98] J. Matas, R. Marik, and J. Kittler. On representation and matching of multi-coloured objects. In *Proc. of ICCV'95*, pages 726–732, 1995.

[99] P.W. Mausel, W.J. Kramber, and J.K. Lee. Optimum band selection for supervised classification of multispectral data. *Photogrammetric Engineering and Remote Sensing*, 56:55–60, 1990.

[100] M.J. McKeown, S. Makeig, G.G. Brown, T.-P. Jung, A.J. Bell, and T.J. Sejnowski. Analysis of FMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–188, 1998.

[101] T. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[102] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, July 1995.

[103] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE Transactions on PAMI*, 24(5):707–711, 2002.

[104] F. Mokhtarian, S. Abbasi, and J. Kittler. Robust and efficient shape indexing through curvature scale space. In *In Proc. British Machine Vision Conference, BMVC'96, Edinburgh, UK*, pages 53–62, 1996.

[105] F. Mosteller and D.L. Wallace. *Applied Bayesian and Classical Inference.* Springer Verlag, New York, 2nd edition, 1984.

[106] A.N. Mucciardi and E.E. Gose. A comparison of seven techniques for choosing subsets of pattern recognition properties. *Trans. IEEE Computers*, 20:1023–1031, 1971.

[107] B. Noble and J. Daniel. *Applied Linear Algebra.* Prentice Hall, 3rd edition, 1988.

[108] B.A. Olshausen and D.J. Field. Natural image statistics and efficient coding. *Network*, 7(2):333–340, 1996.

[109] P. Paatero and U. Tapper. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 18:183–194, 1997.

[110] E.S. Palmer. Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9:441–447, 1977.

[111] L. Parra, C.D. Spence, P. Sajda, A. Ziehe, and K.-R. Muller. Unmixing hyperspectral data. In *Advances in Neural Information Processing Systems*, volume 12, pages 942–948, 2000.

[112] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[113] A. Passerini, M. Pontil, and P. Frasconi. From margins to probabilities in multiclass learning problems. In editor F. van Harmelen, editor, *Proceedings of the 15th European Conf. on Artificial Intelligence*, 2002.

[114] M. Pazzani. Searching for dependencies in bayesian classifiers. In D. Fisher and H. J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 239–248, New York, NY, 1996. Springer Verlag.

[115] B. Pearlmutter and L. Parra. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In M.C. Mozer, M.I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 613–619, 1997.

[116] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[117] P. Penev and J. Atick. Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

[118] P. Perona, A. Shiota, and J. Malik. Anisotropic diffusion. In B.M. ter Haar Romany, editor, *Geometry Driven Diffusion in Computer Vision*, pages 73–92. Kluwer Academic, 1994.

[119] D.T. Pham. Blind separation of instantaneous mixtures of sources via an independent component analysis. *IEEE Transactions on Signal Processing*, 44(11):2768–2779, 1996.

[120] M. Plumbey. Conditions for non-negative independent component analysis. *IEEE Signal Processing Letters*, page IN PRESS, 2002.

[121] J. Porrill and J.V. Stone. Undercomplete independent component analysis for signal separation and dimension reduction. Technical report, The University of Sheffield, Department of Psychology, 1998.

[122] A. Pujol, A.M. Lopez, J. Alba, and J. Villanueva. Ridges, valleys and hausdorff based similarity measures for face detection and matching. In Ana Fred and Anil K. Jain, editors, *Proceedings of the 1st International Workshop on Pattern Recognition in Information Systems (PRIS2001)*, pages 80–90, Setubal, Portugal, 2001. ICEIS Press.

[123] P. Radeva, J. Vitria, and X. Binefa. Eigenhistograms: using low dimensional models of color distribution for real time object recognition. In *8th International Conference on Computer Analysis of Images and Patterns*, pages 17–24. Springer Verlag LNCS 1689, September 1999.

[124] Rajesh P.N. Rao. *Dinamic Appearance-Based Vision*. PhD thesis, University of Rochester, 1997.

[125] J.A. Richards and X. Jia. *Remote Sensing Digital Image Analysis, third edition.* Springer, 1999.

[126] I. Rish, J. Hellerstein, and J. Thathachar. An analysis of data characteristics that affect naive bayes performance. In Claire N'edellec and C'eline Rouveirol, editors, *Proceedings of the Eighteenth Conference on Machine Learning -ICML2001*, pages –. Morgan Kaufmann, 2001.

[127] S.E. Robertson and K.Sparck Jones. Relevance weighting of search terms. *Journal of the American Society of Information Science*, pages 129–146, May-June 1976.

[128] J.S. Sanchez, F. Pla, and F.J. Ferri. On the use of neighborhood-based nonparametric classifiers. *Pattern Recognition Letters*, 18(11-13):1179–1186, November 1997.

[129] B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive fields. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[130] B. Schiele and A. Pentland. Probabilistic object recognition and localization. In *Proceedings of the International Conference in Computer Vision (ICCV'99)*, 1999.

[131] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. PAMI*, 19(5):530–534, 1997.

[132] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[133] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, pages –, 2002.

[134] S.Choi, A.Cichocki, and S.Amari. Flexible independent component analysis. *Journal of VLSI Signal Processing*, 26(1/2):25–38, August 2000.

[135] R.D. Short and K. Fukunaga. The optimal distance measure for nearest neighbor classification. *IEEE Transactions on Information Theory*, IT-27(5):622–627, September 1981.

[136] B. Silverman. *Density Estimation.* Chapman and Hall, 1986.

[137] E.P. Simoncelli and B.A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1215, 2001.

[138] E.H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Ser. B*, 13:238–241, 1951.

[139] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In IEEE Computer Society Press, editor, *Proceedings of the Symposium on Computer Applications and Medical Care*, pages 261–265, 1988.

[140] P. Somol, P. Pudil, F.J. Ferri, and J. Kittler. Fast branch and bound algorithm in feature selection. In B. Sanchez, J.M. Pineda, J. Wolfmann, Z. Bellahsene, and F.J. Ferri, editors, *Proceedings of World Multiconference on Systemics, Cybernetics and Informatics (SCI/ISAS 2000)*, volume 7, pages 646–651, 2000.

[141] P.D. Sozou, T.F. Cootes, C.J. Taylor, and E.C. Di-Mauro. A non-linear generalization of pdms using polynomial regression. In *Hancock E., ed. British Machine Vision Conference 1994, BMVC'94*, volume 1, pages 397–406. University of York, UK:BMVA, 1994.

[142] P.D. Sozou, T.F. Cootes, C.J. Taylor, and E.C. Di-Mauro. Non-linear point distribution modeling using a multi-layer perceptron. In *Pycock D., ed. British Machine Vision Conference 1995, BMVC'95*, volume 1, pages 107–116. University of Birmingham, UK:BMVA, 1995.

[143] C. Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

[144] M.B. Stegmann. On properties of active shape models. Technical report, Technical University of Denmark, Department of Mathematical Modeling, March 2000.

[145] M.J. Swain and D.H. Ballard. Color indexing. *Int. Journal of Computer Vision*, 7(1):11–32, 1991.

[146] P. Swain and R. King. Two effective feature selection criteria for multispectral remote sensing. In *Proceedings of the 1st International Joint Conference on Pattern Recognition, IEEE 73 CHO821-9*, pages 536–540, 1973.

[147] P.H. Swain and S.M. Davis. *Remote sensing: the quantitative approach.* McGraw-Hill, 1978.

[148] T.Hastie and R.Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:607–616, 1996.

[149] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions.* Wiley Publishers, New York, 1985.

[150] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, 1991.

[151] G. Trunk. A problem of dimensionality: A simple example. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1(3):306–307, July 1979.

[152] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[153] H.R. Turtle and W.B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.

[154] J.H van Hateren and D.L. Ruderman. Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. Royal Society, Series B*, 265:2315–2320, 1998.

[155] J.H van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Royal Society, Series B*, 265:359–366, 1998.

[156] R. Vigario, J. Särelä, V. Jousmäki, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Trans. Biomedical Engineering*, 47(5):589–593, 2000.

[157] E. Wachsmuth, M.W. Oram, and D.I. Perrett. Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. *Cerebral Cortex*, 4:509–522, 1994.

[158] A.G. Wacker. *The Minimum Distance Approach to Classification.* PhD thesis, Purdue University, West Lafayette, 1971.

[159] J. Weickert. *Anisotropic Diffusion in Image Processing.* PhD thesis, Universitat Kaiserlautern, 1996.

[160] M-H. Yang and D.J. Kriegman. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, January 2002.

[161] Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems. Kluwer Academic Press*, 2002.

[162] A. Ziehe, K-R. Muller, G. Nolte, B-M. Mackert, and G. Curio. Artifact reduction in magnetoneurography based on time-delayed second-order correlations. *IEEE Trans. Biomedical Engineering*, 47(1):75–87, 2000.