CVC Tech.Rep. #048 September,2000

Independent Modes of Variation in Point Distribution Models

Marco Bressan

Centre de Visió per Computador

Advisor: Jordi Vitrià

Submitted to the Computer Vision Center
on September,2000

Computer Vision Center, 2000

Contents

Acknowledgements

I would like to thank Dr. Jordi Vitrià for always being available when I need him, for his patience and humour, for his directions and advice. But above all I want to thank Dr. Vitrià because he trusted me without knowing me, allowing me the opportunity of doing research in the CVC.

Very special thanks to my parents who, with their love, supported every decision I took. Even when these decisions pulled me away from them. Thanks to them for sharing with me their rigorous passion for scientific research.

Thanks to Anna, the greatest office mate one can have, talking about work when work needs to be talked and talking about life when life needs to be talked.

Thanks to David, a magnificent working colleague, always generous and incredibly efficient.

Actually, thanks to all the people in the CVC, willing to share their knowledge and ideas at all times. Thanks for giving me a place to work and making me feel at home from the first day.

Thanks to Yanina and Lucas, filling me with energy every day. Encouraging me to come to the CVC every morning and making me happy to leave on the evenings.

Finally, thanks to all people at the SIC, down in Bariloche. They taught me that human quality and professionality not only go hand in hand, but that it is the only way to get stuff done and feel it is worth while.

This work is supported by CICYT and EU grants TAP98-0631 and 2FD97-0220 and the Secretaría de Estado de Educación, Universidades, Investigación y Desarollo from the Ministerio de Educación y Cultura de España.

Introduction

The main motivation behind this work is to test the feasibility of statistically independent parameters to represent shape variation. This is achieved by using an Independent Component Analysis (ICA) representation as an alternative to Principal Component Analysis (PCA) for the representation of Point Distribution Models. Alongside with this objective, some natural problems arise, and they are also treated in this work.

ICA is a general-purpose statistical technique with a wide range of applications in neural computing, signal processing and statistics [41, 17, 8, 14]. The idea is to transform observed random data such that the transformed components are maximally independent from each other. In practice, ICA´s most widely spread version consists in searching a linear non-orthogonal coordinate system in multivariate data determined by second and higher order statistics. The first part of this work is devoted to the presentation of ICA, methods for its estimation, applications and advantages as a feature extraction technique. The problem of density estimation within the ICA representation is also addressed, and general solutions for this problem are introduced. These estimation techniques proved efficient in most of the experiments which were performed. The problem of classification under the ICA representation also confirmed the accuracy of the estimation. Classification was focused from a bayesian perspective and the basic underlying theory is also exposed in this work. A practical example comparing this classification scheme with more classical methods is briefly mentioned.

The Point Distribution Model (PDM) [19] is a shape description technique based on the vectorized representation of shapes to estimate a statistical model for shape variation. By modeling this distribution, we can generate new examples, similar to those in the original training set, and we can also examine the plausibility of new shapes. The construction of an appropriate PDM is heavily related with the selection of a good representation and the choice of an appropriate density estimation method for the distribution of the shapes within this representation. After this choice, the two most common problems which arise

are the analysis of shape feasibility, and the problem of given certain parameters, to find the nearest feasible shape. Chapter five of this work introduces the more classical, and also most popular, version of Point Distribution Models.

In chapter six, the main results of the dissertation are exposed. Here, ICA is proposed as a represention for PDMs. The parameters representing shapes in the ICA space will be known as *independent modes of variation*. Through them, the modeling of nonrigid shapes whose modes of variation are supposed statistically independent is possible. In all previous theory the modes of variation are given by uncorrelated projections of maximum variance. The assumption that these projections are optimal for modeling shape variation is not necessarily correct. In certain cases (the fingers of the hand can prove to be a good example) higher order relationships such as independence can be important for better modeling. Improvements in this sense can be considered from several perspectives such as manageability, simplicity, precision, etc. For certain particular problems all these objectives are achieved. Not only a higher control of the problem is gained, but also, the assumption of independence greatly simplifies statistical estimation, transforming an N-dimensional density estimation problem in N 1-dimensional estimations. To be honest, this is, probably the most important advantage of the ICA representation, providing both simplicity and a robust framework. The problem of shape feasibility and nearest feasible shape are addressed. Possible applications of the ICA representation for PDMs are also exposed.

Chapter seven exposes some of the experiments that were performed. The effectiveness of this new approach is presented experimenting with artificial and non-artificial shapes. All results are compared with existing approaches. Finally, the conclusions and proposals for further research are exposed.

Chapter 1

Principal Component Analysis

1.1 Feature Extraction

Feature extraction is generally considered the process of mapping the original measurements into more effective features [26]. If we replace the expression *original measurements* with *object of study*, the act of taking the original measurements can also be considered as a part of the feature extraction process. Effectiveness can be understood as a tradeoff between accuracy and simplicity. When extracting features what we are doing is choosing an appropriate representation for our object of study. The choice of an appropriate representation should not be underestimated since it is one of the most significant factors for the final performance of the system.

Imagine a quality control visual system with the task of separating bad from good objects. An initial decision on feature extraction is made when deciding upon the type of image we consider appropriate for representing the objects. If its going to be a range image, X-ray image, color or grayscale image, etc. If the choice were to be a 256×256 color image with 256 levels per color, each object would correspond to a 196608-dimensional vector taking values between 0 and 255. Maybe this is too much. Maybe this has more information than we need or can even handle. Imagine the goodness of a certain object had only to do with its color, or texture, or border shape, etc. In the first case, and having color consistency ensured, a histogram might be enough. If we, for instance, work with 8-bin color histograms, we are now representing our objects with 512-dimensional vectors. This new representation, might lose accuracy but is clearly simpler and more manageable. Additional prior information will surely provide even more effective features. If certain colors were known never to ocurr, or more subtly, if some colors were independent of the quality of an object, why include them

in the histogram? By knowing these colors in advance or by, mathematically speaking, having learnt a subspace where only those colors relevant to object quality lived, we can project the histograms in this subspace surely obtaining a considerable reduction of dimensionality. Also, since we are dealing with a classification problem, this last subspace can be mapped into another space where these relevant colors enhance their differences, making classification more robust and effective. In this coarse example we have extracted and chosen the feature *discriminated relevant colors*, among the features *pixel values*, *colors* and *relevant colors* as an appropriate representation for our objects. If this were to be once learnt and then applied, all the feature extraction process, mainly adquisition, histogram calculation, projection and mapping, can be done in a pre-processing stage, which leaves the data clean for the classification algorithm. We observe here how feature extraction can completely change the characteristics and complexity of a problem. We also observe how feature extraction obeys to generic postulates such as simplicity and dimensionality considerations, but is basically a problem-oriented process.

Feature extraction is the backbone of this whole work. The objective is quite simple: to model deformations of nonrigid shapes representing a certain object such as a hand, a pair of scissors, a face, etc; and to verify the performance of the obtained model. Performance will be tested in very concrete ways, such as density estimation. To find such a model is no more then to find an appropriate representation for the data.

The entire objective of this work is to treat PDMs using an ICA as a representation for the data. ICA is a statistically-based linear feature extraction technique for multivariate data, as PCA. PCA has a strong theorethical background and solid applications almost anywhere highly dimensional data is found. As a matter of fact the Point Distribution Model is based on PCA. Understanding PCA, its advantages and limitations, helps to understand ICA. This is the reason why this chapter is dedicated to PCA.

1.2 Linear Transforms

In general, features are extracted from measured data by a nonlinear mapping

$$\mathbf{s} = \mathbf{f}(\mathbf{x}) \tag{1.1}$$

where $\mathbf{x}$ denotes the m-dimensional random vector corresponding to the data, and $\mathbf{s}$ the n-dimensional random vector that corresponds to the feature values. We will say that $\mathbf{s}$ is our new representation of $\mathbf{x}$. If the mapping is linear, Eq.

(1.1) can be rewritten as

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{1.2}$$

where $\mathbf{W}$ is a n by m matrix. There are many reasons to restrict ourselves to linear mappings. If the mapping is linear, the mapping function is well defined and our task is to find the coefficients of the linear function so as to optimize a criterion. Optimality can be defined in terms of dimension reduction, statistical properties, simplicity of the components and other general or application oriented criteria. Well known mathematical theory and techniques can be used. Using linear transformations makes the problem computationally and conceptually simpler. We treat only linear transformations in this dissertation, even though most of the methods here exposed can be extended to the nonlinear case. These extensions are generally problem-specific, computationally expensive, and outside the scope of this work.

1.3 Principal Component Analysis

We have mentioned that effective features are those which represent our data both accurately and in a simple way. The problem now is what is accurate and what is simple. Let us assume that simple has only to do with dimension, so we are looking for a subspace. Let us also assume that accuracy is measured in terms of the mean square distance between the original data points and their projections on the subspace. We will also assume that the data is centered, that is $E\{\mathbf{x}\} = \bar{\mathbf{x}} = 0$. If this were not the case we traslate the origin to the mean obtaining centered data. If the dimension of the data is m, we are seeking for orthogonal vectors $\mathbf{v}_i$ with $i = 1, \ldots, n$ which minimize

$$\mathbf{v} = \arg \min_{\substack{v_i^t v_j = 0 \\ \|v_i\| = 1}} E\Big\{ \|\mathbf{x} - \sum_{i=1}^{n} (\mathbf{v}_i^t \mathbf{x}) \mathbf{v}_i \|^2 \Big\} \tag{1.3}$$

The solution to this can be found recursively. In the case $n = 1$, Eq. (1.3) can be rewritten as

$$\mathbf{v_1} = \arg \max_{\|v\|=1} E\Big\{ (\mathbf{v}^t \mathbf{x})^2 \Big\} \tag{1.4}$$

If $\Sigma = E(\mathbf{x}\mathbf{x}^t)$ is the covariance matrix of random vector $\mathbf{x}$, then it can be seen through standard linear algebra [45, 10] that the vector $\mathbf{v}_1$ from Eq. (1.4) is the the eigenvector corresponding to the maximum eigenvalue (λ_1) of matrix

Σ. Note that, since the covariance matrix is real and symmetric, this eigenvalue is real and positive. In fact,

$$\lambda_1 = E\{(\mathbf{v_1}^t \mathbf{x})^2\} \tag{1.5}$$

so the eigenvalue is the variance of the data in the direction of eigenvector $\mathbf{v}_1$. The remaining directions are found in the same fashion, recursively seeking the directions of maximum data, under the constraint of orthogonality to previously found directions. It turns out that they are the eigenvectors of Σ in decreasing order of the corresponding eigenvalues. If the n first directions are arranged in matrix $\mathbf{V}$, such that the $\mathbf{v}_i^t$ is the i-th row of $\mathbf{V}$ and $E\{\mathbf{x}\} = \bar{\mathbf{x}}$

$$\mathbf{s_P} = \mathbf{V}(\mathbf{x} - \bar{\mathbf{x}}) \tag{1.6}$$

The components of $\mathbf{s_P}$ are the components in the basis given by the eigenvectors. We will call them *principal components*. These are the extracted features. From this, what we have performed is known as Principal Component Analysis and Eq. (1.6) is known as the PCA transform. The transform matrix $\mathbf{V}$ will be called *PCA filter matrix*. Figure 1.1 contains a simple example of the principal components of an artificial two-dimensional dataset. The two eigenvectors of the covariance matrix of the data are shown. We can see how the largest variance of the data is along the direction of the first eigenvector. The value of this variance is the value of the corresponding eigenvalue (λ_i).

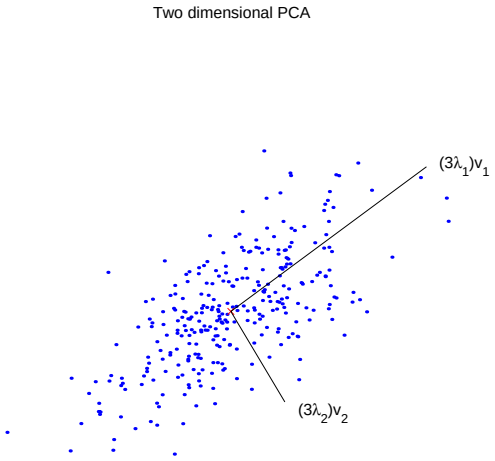

Figure 1.1: The PCA basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ for a two-dimensional artificial (normal) dataset.

If we understand the trace of Σ as the total variance of the data, by projecting

in the n first eigenvectors, we are preserving

$$100 \times \frac{\sum_{i=1}^{n} \lambda_i}{Tr(\Sigma)}\% \tag{1.7}$$

of the data variation in the new representation. This criteria is frequently chosen to decide the dimension of the principal components.

1.4 Some properties

The linear procedure we have just described is called Karhunen-Loéve or Hotelling transform or Principal Component Analysis (PCA) and is discussed at length by Jollife [39]. Since it relies entirely in the input data without reference to the target data, it can be regarded as a form of unsupervised learning. It is probably the most widely used data-adaptive transformation and it is of great practical significance. Its success is largely due to the fact that the solution can be found in an analytical way and, in very highly dimensional problems, through an efficient iterative algorithm.

To *whiten* or *sphere* is to transform the data linearly so the components of the transformed vector are uncorrelated and have unit variance. Since PCA uncorrelates the data, one of its applications is to perform whitening. We define $\mathbf{D}$ as the diagonal matrix with the first (the largest) n eigenvalues so $\mathbf{D} = diag\{\lambda_1, \ldots, \lambda_n\}$. In this case if $\mathbf{V}$ is the PCA filter matrix the following transform whitens the data:

$$\tilde{\mathbf{x}} = \mathbf{D}^{-\mathbf{1/2}}\mathbf{V}(\mathbf{x} - \bar{\mathbf{x}}) \tag{1.8}$$

If $n = m$ the dimension of the white data is equal to dimension of the original data. Choosing any value of $n < m$ not only whitens but also reduces the dimension. In any case we have that $cov\{\tilde{\mathbf{x}}\} = E\{\tilde{\mathbf{x}}^t\tilde{\mathbf{x}}\} = \mathbf{I}$. Small variances in the data are often associated to noise. So, under certain simple assumptions, we can state that PCA reduces noice as well as dimension. A minor but frequently used advantage is that, by projecting data into a two or three dimensional space, visualization is possible. Viewing the data is not negligible and more than often saves a lot of time.

Even though, when using PCA we should never forget the natural limitations derived from its definition. First, unsupervised learning might give results substancially less than optimal. Also, PCA fails to distinguish high order relationships between the data. This affects many problems where this information is precisely what we need. Data relevance frequently goes beyond a standard

deviation value in the same way data independence trascends correlation. In this context, dropping information we *consider* irrelevant, might cause us to lose information. On the other side, the information we keep might not be important for our problem. An example of this can be seen in figure 1.2 where we observe data grouped in two clusters. Consider a classification problem of this data after reducing dimension to one. If dimensionality reduction was performed with PCA, since the direction of maximum variance would be vertical, the projection on the principal component would remove all ability to discriminate the two classes. A projection in the other principal component would give optimal class separation with no loss of discriminatory information.

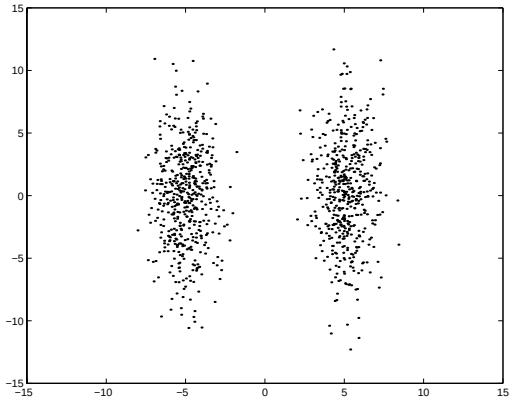

Figure 1.2: In this particular case, PCA discards the discriminatory information by projecting in the vertical axis.

Chapter 2

Independent Component Analysis

2.1 Statistical independence

Given a random vector $\mathbf{x}$, we can intuitively say that its components are independent if information on any one of the components gives no information on any one of the others. In this case, the value of one variable cannot be predicted if any other variable value is known. Technically this can be expressed in terms of the probability density function (p.d.f.). For simplicity let us assume that the dimension of $\mathbf{x}$ is two and denote $p_1(x_1)$ and $p_2(x_2)$ as the probability densities of the components of $\mathbf{x}$ and $p(\mathbf{x})$ as the joint probability of $\mathbf{x}$. Then, x_1 and x_2 are said to be independent if and only if the joint p.d.f. is factorizable, that is

$$p(\mathbf{x}) = p_1(x_1)p_2(x_2) \tag{2.1}$$

A very important property, for independent random variables derived directly from this definition is that given two functions $h_1(x)$ and $h_2(x)$

$$E\big\{h_1(x_1)h_2(x_2)\big\} = E\big\{h_1(x_1)\big\}E\big\{h_2(x_2)\big\} \tag{2.2}$$

In the particular case where $h_1(x)$ and $h_2(x)$ are linear, due to the linearity of the expectancy, Eq. (2.2) can be written as

$$E\big\{x_1x_2\big\} - E\big\{x_1\big\}E\big\{x_2\big\} = 0 \tag{2.3}$$

and variables x_1 and x_2 are said to be *uncorrelated*. The above result is equivalent to saying that $Cov(x, y) = 0$ and it is important to keep in mind that even though independence implies uncorrelatedness, the reciprocal is *not true*. Covariance

only captures the second-order dependencies between x_1 and x_2, independence is about all the orders. To take a simple example, let us consider a variable z distributed uniformely in the interval $[0, 2\pi)$ and the variables $x_1 = \cos(z)$ and $x_2 = \sin(z)$. The covariance and therefore correlation of x_1 and x_2 is zero

$$E\{x_1 x_2\} - E\{x_1\}E\{x_2\} = E\{\cos(z)\sin(z)\} - 0 = 0$$

If we take $h_1(x) = h_2(x) = x^2$ we can see that Eq. (2.3) is not verified and thus, the variables are not independent. In fact,

$$E\{x_1^2 x_2^2\} - E\{x_1^2\}E\{x_2^2\} = -0.127$$

In this particular case, the dependence of the variables is already clear from their definition. Also, if we decided to plot the variables this would result on points all living in the unit circumference so the value of any one of the variables can be easily predictable from the value of the other variable, contradicting the intuitive definition of independence.

In the particular case where random vector $\mathbf{x}$ has a Gaussian distribution, it can be seen that, if the components of $\mathbf{x}$ are uncorrelated (the covariance matrix is diagonal) then they are independent. This is deduced directly from the Gaussian p.d.f. and Eq. (2.1) and it is mentioned here due to its consequences on the identifiability of the ICA model.

2.2 The linear ICA model

In the literature, at least three different basic definitions for the linear ICA model can be found [17, 41]. The noise-free ICA model can be stated as follows: The ICA of an m dimensional random vector $\mathbf{x}$ is the estimation of the following generative model for the data:

$$\mathbf{x} - \bar{\mathbf{x}} = \mathbf{A}\mathbf{s} \tag{2.4}$$

where the latent variables or *independent components* s_i in the vector $\mathbf{s} = (s_1, \ldots, s_n)^t$ are assumed independent. These variables are also called *sources*. The matrix $\mathbf{A}$ is a constant $m \times n$ *mixing* matrix, and $\bar{\mathbf{x}}$ represents the mean of $\mathbf{x}$.

This representation in terms of independence proves useful in an important number of applications such as data analysis and compression, blind source separation, blind deconvolution, denoising, etc.

The pseudoinverse of $\mathbf{A}$ which we will represent as $\mathbf{W}$, is called the filter or projection matrix and provides an alternative generative model for ICA.

$$\mathbf{W}(\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{s} \tag{2.5}$$

Depending of the dimension of $\mathbf{x}$ (m) compared to that of $\mathbf{s}$ (n), we have three cases:

$\mathbf{m = n}$ The ideal or *complete* case. No information is lost in the estimation and if $\mathbf{W}$ is taken as the inverse of $\mathbf{A}$ the model given by Eq. (2.5) is equivalent that given by Eq. (2.4). Throughout this work we will assume the complete case for the noise-free linear ICA Model.

$\mathbf{m > n}$ This is known as the *overdetermined* case. In this case $\mathbf{W}$ can be completely determined but $\mathbf{A}$ contains uncertainty. A possible solution is to initially perform a dimensionality reduction using e.g. PCA to obtain the previous case.

$\mathbf{m < n}$ In this *underdetermined* case there are more independent components or sources than data. We will not treat this case. As a matter of fact, the solutions given so far to this problem are far from being satisfactory.

In addition to the fundamental assumption of independence and the stated dimensionality assumption, Comon shows that if at least $n - 1$ components are non-Gaussian, the identifiability of the ICA Model is ensured [17]. The problem here is that, for Gaussian random variables, mere uncorrelatedness implies independence. For these Gaussian variables, the ICA model can be estimated only up to an orthogonal transformation (orthogonal transformations preserve uncorrelatedness). The consequence is that $\mathbf{A}$ is not identifiable.

The ICA model also contains some ambiguities. From Eq. (2.4) we know that the independent components are zero-centered but we cannot determine the variances of the independent components. This is due to the fact that both $\mathbf{s}$ and $\mathbf{A}$ are unknown so any scalar multiplier on one of the sources can be cancelled by dividing the corresponding column of $\mathbf{A}$ by the same scalar. This motivates the assumption that each independent component s_i has unit variance: $E\{s_i^2\} = 1$. This restriction leaves an ambiguity in the sign, because the multiplication of an independent component by -1 does not affect the model.

The other important ambiguity in the model is that the order in the components cannot be determined. Given any permutation matrix $\mathbf{P}$ the model given in Eq. (2.4) is equivalent to $\mathbf{x} - \bar{\mathbf{x}} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s}$.

In many applications an order for the sources is necessary, so different ordering criterions can be used. The norm of the columns $\mathbf{A}$ can be understood as the contributions of the different sources to the variance of $\mathbf{x}$, so an order reminiscent to that of PCA would be to number the independent components in the decreasing order of the norm of the columns of mixing matrix A. As we shall see, measures of nongaussianity play a significative role in ICA estimation.

So another possibility is to order the sources according to their nongaussianity. The order here obtained would be related with that order given by projection pursuit. Nevertheless, none of these approaches is definitive and ordering of the independent components is absolutely problem-dependent.

2.3 Estimation of the ICA model

The estimation of the ICA model is done choosing an objective function somehow related to the statistical independence we wish to achieve, and then maximizing or minimizing this function. This objective function can involve, progressively, each one of the rows of the filter matrix or the whole filter matrix at the same time. Depending on the case we will say that we are performing one-unit or progressive estimation and multi-unit or simultaneous estimation. In this section we sketch the most frequently found estimation techniques employed in the literature.

2.3.1 Simultaneous estimation

A common approach is to estimate the model by a *maximum likelihood* method. Given T samples of random vector $\mathbf{x}$ and denoting by $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^t$ the log-likelihood in Eq. (2.5) takes the form

$$L = \sum_{t=1}^{T} \sum_{i=1}^{n} \log f_i(\mathbf{w}_i^t \mathbf{x}^t)) + T \log |\mathbf{W}| \tag{2.6}$$

where the f_i are the density functions of the s_i [52]. Since these functions are unknown, the implementation of a maximum likelihood approach requires also a succesive estimation of the probability densities.

A closely related approach consists in maximizing the output *entropy* or information flow of a neural network with nonlinear outputs. Differential entropy of random vector $\mathbf{y} \sim f(\mathbf{y})$ is defined as

$$H(\mathbf{y}) = -\int_{-\infty}^{\infty} f(\mathbf{y}) \log(f(\mathbf{y})) d\mathbf{y} \tag{2.7}$$

The entropy of a random variable can be interpreted as the degree of information that the observation of the variable gives. The more unpredictable and unstructured the variable is, the larger its entropy. So if $\mathbf{x}$ is the input to the neural network whose outputs are of the form $g_i(\mathbf{w}_i^t \mathbf{x})$ where the g_i are some kind of

non-linear scalar functions, and $\mathbf{w}_i$ the weight vectors of the neurons, we wish to maximize is

$$L_H = H(g_1(\mathbf{w}_1^t\mathbf{x}), \ldots, g_n(\mathbf{w}_n^t\mathbf{x})) \tag{2.8}$$

This principle is known as *infomax*, and it can be seen that if the g_i are chosen such that $g_i' = f_i$ then entropy maximization reduces to maximum likelihood estimation [14].

A very natural approach for estimating ICA is through the minimization of the *mutual information* which, for random vector $\mathbf{y} = (y_1, \ldots, y_n)$, is defined as

$$I(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} H(y_i) - H(\mathbf{y}) \tag{2.9}$$

The Kullback-Leibler divergence for probability densities f_1 and f_2 is defined as

$$D(f_1 \| f_2) = \int_{-\infty}^{\infty} f_1(\mathbf{y}) \log \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} d\mathbf{y} \tag{2.10}$$

It is an (asymmetric) *distance* between probability densities. Using Eq. (2.7) it can be seen that the mutual information is the Kullback-Leibler divergence between the joint and the marginal densities. Considering Eq. (2.1), this makes mutual information a very natural measure for independence. As a matter of fact the mutual information is always positive and is zero if and only if the components of random vector $\mathbf{y}$ are independent. An important property for mutual information is the way it behaves under a linear transformation such as $\mathbf{y} = \mathbf{W}\mathbf{x}$

$$I(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} H(y_i) - H(\mathbf{x}) + log|\mathbf{W}| \tag{2.11}$$

It can be seen that if the density is accurately estimated as a part of the maximum likelihood approach, then this approach is equivalent to mutual information minimization [35, 47]. Both methods have the strong drawbacks that the density estimation is needed and that the methods are very sensitive to outliers. In the case of mutual information, polynomial density expansions for Eq. (2.11) such as the Edgeworth or Hermite expansions [46] can be used. Also, cumulant-based approximations have been proposed [1, 17].

Other methods used for simultaneous estimation worth mentioning are non-linear PCA [44, 51] and higher order cumulant matrices [15].

2.3.2 Progressive estimation

As we mentioned, progressive or one-unit estimation consists in estimating the independent components one by one. This approach has several advantages. Prior knowledge of the number of independent components is not needed and the algorithms are frequently simpler and cheaper than for simultaneous estimation. Our objective is to estimate one vector, say $\mathbf{w}$, such that $\mathbf{w}^t\mathbf{x} = \mathbf{s}$ is one of the independent components. We will refer to $\mathbf{w}$ as the direction corresponding to the independent component $\mathbf{s}$ and this is equivalent to the succesive estimation of the rows of $\mathbf{W}$.

Eq. (2.11) already suggests that the independent components correspond to directions in which the differential entropy is minimized. The problem is that, as reflected in the same equation, differential entropy is not invariant for affine transformations. The concept of negentropy, defined as

$$J(\mathbf{y}) = H(\mathbf{y_{Gauss}}) - H(\mathbf{y}) \tag{2.12}$$

provides an invariant version of entropy [17]. Here, $\mathbf{y_{Gauss}}$ refers to the Gaussian random vector with the same mean and covariance as $\mathbf{y}$. It can be seen that the negentropy of $\mathbf{y}$ is the Kullback-Leibler divergence between $\mathbf{y}$ and $\mathbf{y_{Gauss}}$. So, besides invariance, negentropy is always positive and only zero if $\mathbf{y}$ has a Gaussian distribution. Negentropy is a natural measure of nongaussianity. We will now relate nongaussianity with independence. If $\mathbf{y}$ is restricted to be uncorrelated, we have the following expression for mutual information in terms of negentropy

$$I(y_1, y_2, \ldots, y_n) = J(\mathbf{y}) - \sum_{i=1}^{n} J(y_i) \tag{2.13}$$

We can now conclude, due to the invariance property, that finding maximum negentropy directions is equivalent to finding a representation in which mutual information is minimized. This can also be read as *independence is found in the directions of maximum nongaussianity.*

In order to calculate negentropy, the densities of the independent components have to be estimated so the problems we had with mutual information hold for negentropy. As for mutual information in the multi-unit case, approximations of negentropy in terms of higher-order cumulants are proposed [40]. In [33] it is argued that these approximations result quite inaccurate and are very sensitive to outliers so the following approximations for the unidimensional case were introduced and their efficiency tested,

$$J(\mathbf{y}) \approx \|E(G(\mathbf{y})) - E(G(\nu))\|^p \tag{2.14}$$

where G is a non-quadratic function, ν is a standarized Gaussian variable and p is usually chosen such that $1 \leq p \leq 2$. The following prove to be efficient problem-dependent choices for G

$$\begin{aligned} G_k(y) &= y^4 \\ G_t(y) &= \frac{1}{a}\ln(\cosh(ay)) \\ G_e(y) &= \exp(-\frac{y^2}{2}) \end{aligned}$$

If G_k and $p = 1$ is used in Eq. (2.14) what we obtain is the modulus of kurtosis for an approximation of negentropy. Kurtosis is defined as

$$kurt(y) = E\{y^4\} - 3E\{y^2\}^2 \tag{2.15}$$

A direct approximation of negentropy by the modulus of kurtosis is inaccurate but it makes sense, since kurtosis is also a widely used measure for nongaussianity. As a matter of fact, random variables that have a negative kurtosis are called subgaussian or platykurtotic and they are called supergaussian or leptokurtotic if they have positive kurtosis. An additional result which we will informally expose makes kurtosis interesting for the estimation of the independent components. It seems reasonable to restrict ourself to directions which result in unit variance independent components ($E\{(\mathbf{w}^t\mathbf{x})^2\} = 1$) and to assume that $\mathbf{x}$ is zero centered. Defining $\mathbf{z} = \mathbf{A^t w}$, using Eq. (2.4) and remembering that $E\{\mathbf{ss^t} = \mathbf{I}\}$, we have that

$$1 = E\{(\mathbf{w}^t\mathbf{x})^2\} = \mathbf{w^t A A^t w} = \mathbf{z^t z} = \|\mathbf{z}\|^2$$

On the other side, using the properties of kurtosis

$$kurt(\mathbf{w}^t\mathbf{x}) = kurt(\mathbf{w}^t\mathbf{A}\mathbf{s}) = kurt(\mathbf{z}^t\mathbf{s}) = \sum_{i=1}^{n} z_i^4 kurt(s_i) \tag{2.16}$$

Using this, it can be seen that the optimization landscape for the problem

$$\max_{\|z\|=1} |kurt(\mathbf{w}^t\mathbf{x})| \tag{2.17}$$

has $2n$ local maxima, corresponding to the values $\mathbf{z} = \pm\mathbf{e}_j$. So one effectively obtains $\mathbf{w}^t\mathbf{x} = \pm s$ which is one of the independent components. We have already mentioned that the sign of the independent component cannot be determined.

Kurtosis has been widely used in ICA, but it provides a poor estimator due to its sensitivity and asymptotic variance. So in one-unit algorithms the approximations provided by functions other than G_k are frequently used.

2.4 FastICA: The algorithm we used

There is a wide variety of ICA algorithms. In the experiments performed, several ICA algorithms were tested. Bell and Sejnowski introduced an algorithm using the infomax principle [8]. A proper implementation of this algorithm requires some prior statistical information on the sources. Mainly if they are negatively or positively kurtotic. The computational performance of this algorithm in the problems here focused is quite poor, because of the dimensions involved.

Another extendedly used algorithm is JADE which stands for Joint Approximate Diagonalization of Eigenmatrices and was introduced by Cardoso and Soulomicac [15]. This algorithm solves the ICA problem by computing the eigenvectors of the cumulant matrices. The fact it performs simultaneous estimation affects its capacity to deal with highly dimensional data. Its results were similar to those of fixed point algorithms.

In [36] a fixed point algorithm with fast convergence properties is presented. This algorithm proved efficient regardless of the dimension and density of the sources. Moreover, its progressive essence makes it adaptable and robust. Unless otherwise stated this was our choice when implementing ICA. The name of this algorithm is FastICA and its basic features are presented here.

As with many algorithms it first requires some preprocessing of the data. Centering and whitening of the data was performed as presented in Section 1.4. If $\tilde{\mathbf{x}}$ is the whitened data ($cov(\tilde{\mathbf{x}}) = \mathbf{I}$) we have that restricting ourselves to unit variance directions is equivalent to restricting ourselves to $\|\mathbf{w}\| = 1$. Another very important property for whitened data is that the mixing and filter matrix for the whitened data are orthogonal. The reason for this is that they relate two spatially white vectors,

$$\mathbf{I} = E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^t\} = \tilde{\mathbf{A}}E\{\mathbf{s}\mathbf{s}^t\}\tilde{\mathbf{A}}^t = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^t \tag{2.18}$$

Working with white data, what FastICA does is to find directions of maximum nongaussianity in the unit hypersphere, restricting the directions to be orthogonal to those previously found. It is a progressive algorithm, and the nongaussianity measures are based on the negentropy approximation given by Eq. (2.14). The FastICA algorithm finds the k-th direction by

$$\mathbf{w}_k = \arg \max_{\substack{w^t w_i = 0 \\ \|w\| = 1}} E\left\{G((\mathbf{w}^t\tilde{\mathbf{x}}))\right\} \tag{2.19}$$

with $i < k$. This optimization problem is solved introducing Kuhn-Tucker conditions and than applying a Newton-Raphson algorithm. An approximation of the Jacobian matrix is also necessary. It can be shown that by starting on each

direction on a random initial point the algorithm converges at least quadratically to the desired optima if the ICA Model holds [36]. A simultaneus estimation version of FastICA can be implemented using symmetric decorrelation. In this case the method is essentially equivalent to a Newton method for maximum likelihood estimation and we can see that the FastICA algorithm can be used to optimize both one-unit and multi-unit contrast functions.

Given T samples of the m-dimensional random vector $\mathbf{x}$, the steps of an Independent Component Analysis as performed in this dissertation are as follow,

1. Whiten the centered data using Eq. (1.8) and obtain $\tilde{\mathbf{x}}$.

2. Apply FastICA to the whitened data, estimating $\tilde{\mathbf{W}}$ such that

$$\tilde{\mathbf{W}}\tilde{\mathbf{x}} = \mathbf{s}$$

   Because of its robustness and properties [36] and the characteristics of our experimental data, unless stated, the nonlinearity chosen was $G_t(\mathbf{y}) = \frac{1}{a}\ln(\cosh(a\mathbf{y}))$.

3. The ICA filter matrix for model (2.5) is $\mathbf{W} = \tilde{\mathbf{W}}\mathbf{D}^{-\mathbf{1/2}}\mathbf{V}$ and the ICA mixing matrix for model (2.4) is $\mathbf{A} = \mathbf{V}^t\mathbf{D}^{\mathbf{1/2}}\tilde{\mathbf{W}}^t$.

Chapter 3

ICA applications and connections

3.1 Blind Source Separation

Blind Source Separation (BSS, also known as Blind Signal Separation) is the classical application of the ICA Model, and the main motor of all initial research on ICA [41]. BSS consists in recovering unobserved signals or *sources* from several observed mixtures. The *cocktail-party problem*, is the typical BSS problem and it provides a clarifying picture of the ICA context. There is a room with n people talking simultaneously and m microphones placed in different directions. We choose to represent by $s_i, i = 1, \dots, n$ the original speech signals and by $x_k, k = 1, \dots, m$ the sound signals recorded by each of the microphones. The problem is to estimate the original speech signals from the recorded signals. This is a particular Blind Source Separation problem. We omit time delays, noise and other extra factors to simplify our model.

It is not unrealistic to assume that the speech signals are statistically independent. This is equivalent to assuming that what one person says has no relationship with the speech of another person. It turns out to be quite true in not few cocktail parties. Under this assumption, ICA provides a solution for this problem. In this case, the *sources* or *independent components* are the original speech signals. The *mixture matrix* is the matrix that mixes or gives a weight to the speeches, outputting the recorded signals. Each independent component is given by taking the inner product of the observed vector with a row vector of matrix $\mathbf{W}$. In other words, what we are doing is filtering the sample signal with one of the rows to obtain the original speech. For this reason, matrix $\mathbf{W}$ is called the *filtering matrix* and its rows the *ICA filters*.

Figure 3.1 illustrates the situation. In this case the original signals shown in figure 3.1(a) were artificially generated, so they are not original speech signals. A 4×4 random mixing matrix was generated, and the signals were mixed. The resulting signals are displayed in figure 3.1(b). Figure 3.1(c) shows the estimated sources through ICA. As can be seen, the estimated sources are very close to the original source signals, maybe except for the ambiguities already mentioned.

Applications of ICA and BSS can be found in the processing of communication signals [16, 2]; in biomedical applications [22, 48, 54, 61]; as an alternative to Principal Component Analysis [9, 43, 50, 4]; among many others.

3.2 Sparse Coding

Sparse coding is a coding of the data such that, for any given input vector, only a few components of the code will be significantly active. Which means that most of the components will be close to zero and only seldom significantly non-zero. A sparse code can be obtained, for instance, as a result of distributing the redundancy evenly between all the components. A sparse distribution, i.e. the distribution corresponding to a single component of a sparse code is typically supergaussian. This kind of distribution presents a peak in zero and heavy tails.

We have seen that a suitable estimation of the independent components can be done searching for nongaussian projections of the data. This was done optimizing nongaussianity measures such as kurtosis, negentropy or approximations of these. Very similar to what we need in order to obtain a sparse code but not the same, though. Negatively kurtotic directions are useless in sparse coding but may represent a source in ICA. Nevertheless it is true that if, by ICA, we obtain supergaussian projections than we are obtaining a sparse code for our data. In particular, a sparse code with components which are, at least uncorrelated and at most, statistically independent. There exist a large amount of problems where a high sparsity is observed in the independent components [24, 34, 37, 61, 9, 13].

Sparse coding, as a way for representing data, also proves useful for density estimation and as a discrimination technique for classification [13]. Still, its main application is redundancy reduction [6, 24].

3.2.1 Sparse coding and some biological considerations

It is said that an important amount of the sensory processing taking place in the brain is redundancy reduction [6]. In particular, in the human visual system, a large amount of research has been made concerning the sensory coding [3, 42, 24]. It has been suggested, and still argued, that a major function of

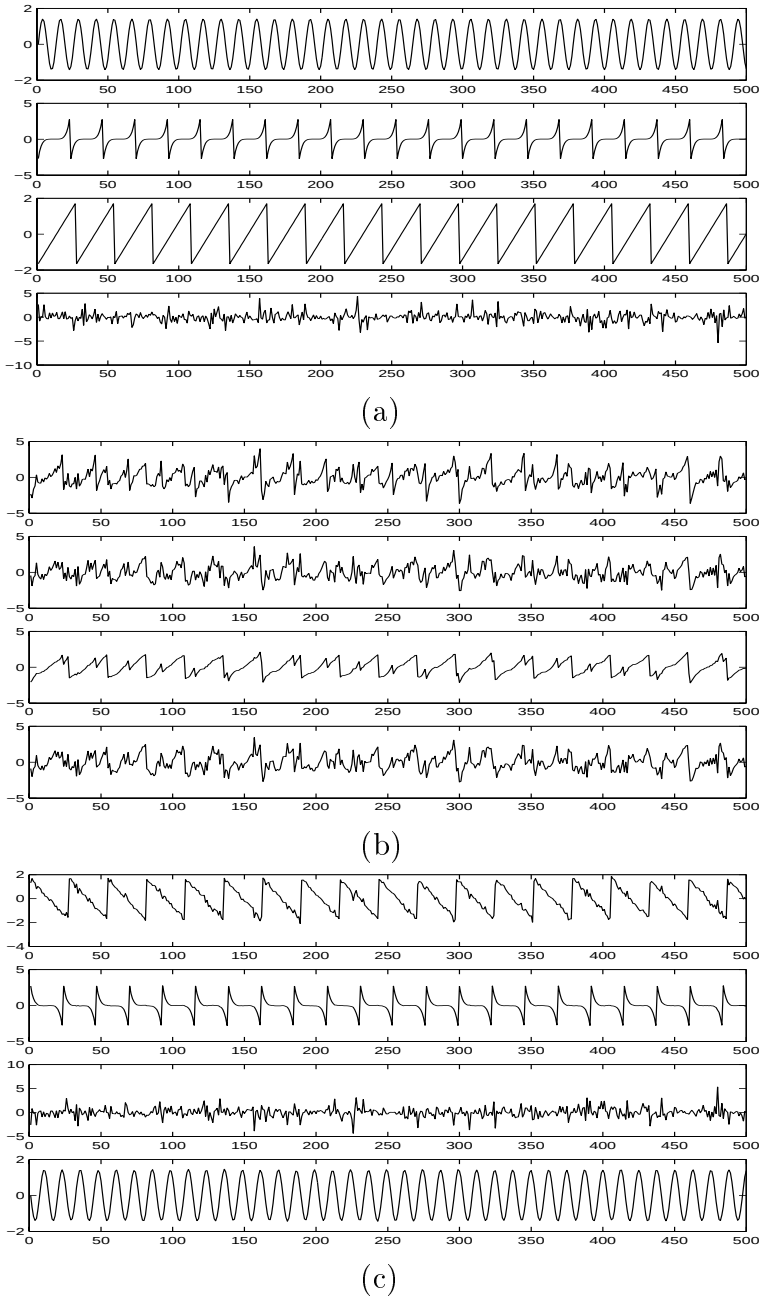

Figure 3.1: (a) The original signals (sources). (b) The sources are mixed using a random mixing matrix. (c) ICA estimation of the sources, using only the mixed signals.

the first two layers of our idealized visual system, the retina and the primary visual cortex, is to reduce the statistical redundancy in input. Some studies have claimed that important functions of the retina are the decorrelation of the input information and suppression of noise. We can (lightly) relate this to the PCA dimensionality reduction and whitening stage of the ICA model. On the other side, the activity patterns of the primary visual cortex can be thought to represent features of natural images, and the neurons may be seen as *feature detectors*. The neurons of this layer respond to various basic image structures like edges, lines and bars. If the relationship between ICA and these layers were true, than the training of ICA as a feature detector on natural images should filter these simple structures. Several works on natural images confirm these results [60, 5], showing that features extracted through ICA correspond closely with those observed in the primary visual cortex. In particular, a systematical comparison between the ICA features and the properties of the simple cells in the macaque primary visual cortex was conducted [60], where the authors find a good match for most of the parameters. The obtained features are also closely connected to those offered by wavelet theory and Gabor analysis [21, 49].

3.3 Projection Pursuit

Projection pursuit [25, 32, 40] is a technique developed for finding *interesting* projections of multidimensional data. Its applications range from simple visualization to discrimination, estimation and regression. Figure 1.2 shows how a method such as PCA would fail in a classical projection pursuit problem. As previously stated, the solution of this problem would be provided by a projection in the horizontal direction which obtained an optimal separation of the clusters. In the projection pursuit context it has been argued [32, 40] that the Gaussian distribution is the least interesting one. This is used in the definition of an index that defines the interestingness of a given direction. Usually this index is a measure of nongaussianity as those we have seen. Since a possible estimation of the ICA model is done by projecting in directions which maximize nongaussianity, the connection between projection pursuit and ICA is evident.

Chapter 4

Statistics within the ICA context

4.1 Density families for the sources

One of the most important characteristics of the ICA representation is that the assumption of independence of the sources greatly simplifies statistical operations such as density estimation, classification or regression. Density estimation is simplified by transforming an N-dimensional estimation problem in N 1-dimensional estimations due to the following relationship derived from Eq. (2.1)

$$p(\mathbf{x}) = |\mathbf{W}|p(\mathbf{s}) = |\mathbf{W}| \prod_{i=1}^{n} p(s_i) \tag{4.1}$$

Since we are working on a single dimension, the complexity of the method employed for density estimation is not relevant, but advantage can be taken of the nongaussianity of the sources. For instance, in Section 3.2 it is mentioned how in certain problems the supergaussianity of the independent components can be assumed. We will now mention a few unidimensional probability density families, useful when dealing with the ICA sources, represented by variable s.

4.1.1 Parametric methods

For supergaussian variables, we can use different density models depending on the level of sparsity. For moderately sparse variables, a simple and fairly good approach is the *Laplace* or *double-exponential* density,

$$p(s|\alpha) = \frac{1}{\sqrt{2}\alpha} e^{-\frac{\sqrt{2}|s|}{\alpha}} \tag{4.2}$$

The main disadvantage of the Laplace density is that it only has the parameter (α) so it is not suited for adapting the density to the data. As a solution to this, the generalized Gaussian model is used to model distributions that deviate from normality [12],

$$p(s|\mu,\sigma,\beta) = \frac{\omega(\beta)}{\sigma}\exp\left[-c(\beta)\left|\frac{x-\mu}{\sigma}\right|^{2/(1+\beta)}\right] \tag{4.3}$$

where

$$c(\beta) = \left[\frac{\Gamma(\frac{3}{2}(1+\beta))}{\Gamma(\frac{1}{2}(1+\beta))}\right]^{1/(1+\beta)} \tag{4.4}$$

and

$$\omega(\beta) = \frac{\Gamma(\frac{3}{2}(1+\beta))^{1/2}}{(1+\beta)\Gamma(\frac{1}{2}(1+\beta))^{3/2}} \tag{4.5}$$

In this form, the data's mean and standard deviation are given by μ and σ respectively. The parameter β is a measure of kurtosis and is related with Eq. (2.15)

$$kurt(\beta) = \frac{\Gamma(\frac{5}{2}(1+\beta))\Gamma(\frac{1}{2}(1+\beta))}{\Gamma(\frac{5}{2}(1+\beta))^2} - 3 \tag{4.6}$$

It can be seen that when $\beta = 0$, the distribution is the Gaussian and when $\beta = 1$, Laplacian. Also, as $\beta \to -1$, the distribution becomes uniform and as $\beta \to \inf$, it approximates the delta function at zero. A maximum a posteriori approach can be used for the estimation of β [12].

A quite robust parametrization which provides an accurate approximation of very sparse data was introduced by Hyvärinen [36],

$$p(s) = \frac{1}{2}\frac{(\alpha+2)[\alpha(\alpha+1)/2]^{(\alpha/2+1)}}{\sqrt{\alpha(\alpha+1)/2} + |s|^{(\alpha+3)}} \tag{4.7}$$

As $\alpha \to \infty$ this approaches the Laplace density. The parameters are estimated as follows,

$$\alpha = \frac{2 - k + \sqrt{k(k+4)}}{2k-1} \tag{4.8}$$

where $k = p(0)^2$ and $p(0)$ is estimated with a suitable kernel.

4.1.2 Nonparametric methods

Nonparametric models can be easily implemented and no computational problems arise, since we are dealing with one dimensional variables. The most straightforward nonparametric method is using the frequency histogram. An intermediate solution is to use kernel methods. A typical kernel method is to positions Gaussians at all the samples in the distribution. In our case, and if we have K samples $s_1, \ldots, s_K$, the corresponding density can be expressed as

$$P_{KER}(s) = \sum_{i=1}^{K} \frac{1}{K} G(s; s_i; \sigma) \tag{4.9}$$

where a good choice for σ if K is sufficiently high is $(\frac{12}{K})^{\frac{1}{5}}$ [55].

Nonparametric models have a strong disadvantage, which is worth mentioning. In the case of sparse data, some kernel methods cause that the probability drops to zero for most of the variable values. This is even more drastic when working with frequency histograms. In this case, there is few data far from zero, so the marginal probability of any value which differs in any way from the learnt values is generally zero. Thus, the total probability is also zero. This problem does not arise when the marginal densities are modelled by densities which have heavy tails.

4.1.3 Mixture models

Mixture models [59] combine the advantages of both parametric and nonparametric methods, by not restricting estimation to rigid specific functional forms (as in the case of parametric estimation) and by assigning a model size related with the complexity of the problem and not only with the size of the data set (as with nonparametric estimation). As a counterpart, this *semiparametric* approach is, in general, computationally expensive. A mixture distribution model is formed from a linear combination of M basis functions, with $M \ll K$ and K is the number of samples,

$$p(s) = \sum_{j=1}^{M} p(s|j)P(j) \tag{4.10}$$

with $\sum P(j) = 1$, $0 < P(j) < 1$ and $\int p(s|j)ds = 1$. In a mixture model we are combining M parametrized densities and giving each one of them a weight. The training of a mixture model is to estimate from samples the parameters of $p(s|j)$ and the $P(j)$. The most common approach is based on maximizing the

likelihood of the parameters of the given data set. A frequently used algorithm for this optimization problem is to use the Expectation Maximization (EM) algorithm [23]. All parameter estimation for mixture models in this work has been done adapting the EM algorithm. Mixture models are good for approximating multimodal densities. In the experiments, except for a few exceptions in which bimodality was observed, the ICA sources were always unimodal. This prior information can be used in the estimation process.

The most commonly used mixture model is that where the individual component densities are Gaussian densities. In this case,

$$p(s|j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\Big(-\frac{(s-\mu_j)^2}{2\sigma^2}\Big) \tag{4.11}$$

For representing sparse and zero-centered unimodal densities [31], we can choose a mixture of two zero-centered Gaussians ($M = 2, \mu_1 = \mu_2 = 0$) of different variance.

We also implemented a Laplacian mixture model that proved simple, efficient and highly adaptive to strong variations in the level of sparsity,

$$p(s|j) = \frac{1}{\sqrt{2}\alpha_j} \exp\Big(-\frac{\sqrt{2}|s-\mu_j|}{\alpha_j}\Big) \tag{4.12}$$

As for the Gaussians, having some additional prior information we chose $M = 2$ and both individual components to be zero-centered. Since this density was a frequent choice a fast algorithm for the calculation of its log-likelihood, based on a Taylor approximation was also implemented.

Figure 4.1 illustrates the result of estimating the densities of two components obtained in the experiments with some of the proposed methods. The true histogram of the component and a standard gaussian are shown as reference. The component in figure 4.1(a),(b) is supergaussian and the component in figure 4.1(c),(d) is bimodal, supergaussian in each one of the modes.

4.2 Bayesian Decision

Given K classes $C_1, \ldots, C_K$ and an outcome x_{Test}, we wish to assign the outcome to a particular class. This is known as a classification problem and it can be focused from different perspectives. In a probabilistic framework one might think that a suitable solution for this problem is the one which minimizes the probability of misclassification [10]. It can be seen that the solution is achieved by assigning x_{Test} to that class which maximizes the *posterior probability*, that

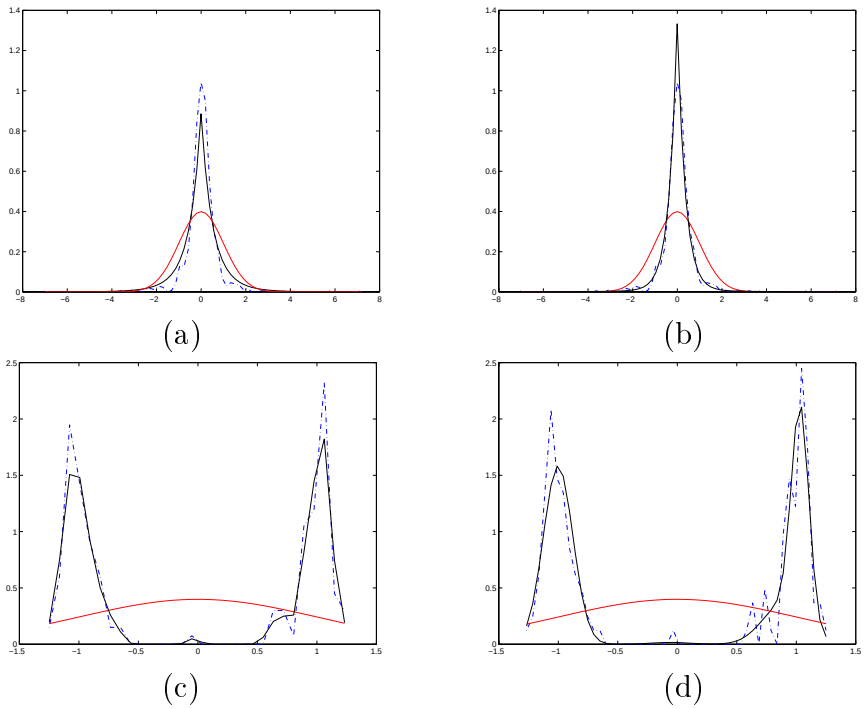

Figure 4.1: Estimation of some univariate densities from data. In all cases the histogram distribution of the component (dot-dash) and the standard normal (red) are plotted as a reference. (a) Density proposed by Hyvärinen, in Eq. (4.7). (b) A mixture of two zero-centered Laplacians. (c) Using a kernel method for a bimodal random variable. (d) Same, but using a Gaussian Mixture Model.

is the probability of the class given the outcome. This is called the Maximum a Posteriori or MAP solution and is formally expressed as

$$C_{MAP} = \arg \max_{k=1...K} \{P(C_k|x_{Test})\} \tag{4.13}$$

Bayes' theorem provides an alternative expression for the posterior probability

$$P(C_k|x_{Test}) = \frac{P(x_{Test}|C_k)P(C_k)}{P(x_{Test})} \tag{4.14}$$

On the right-hand side, $P(C_k)$ is usually called the *prior* probability, $P(x_{Test}|C_k)$ is referred to as the *class-conditional probability* or, when seen as a function of the parameters the *likelihood*, and $P(x_{Test})$ is referred to as the *unconditional probability* since it does not depend on the classes. This last quantity plays the role of a normalization factor, ensuring the posterior probabilities sum to unity.

It helps to read Bayes' theorem as: "The probability of x_{Test} belonging to class C_k (or posterior probability) is proportional to the probability of generating x_{Test} within C_k (or likelihood) weighted with the probability of class C_k (or prior)." The importance of Bayes' theorem lies in the fact that it re-expresses the posterior probabilities in terms of quantities which are often easier to calculate.

In practice, the class-conditional probabilities are modelled parametrically or non-parametrically from our training set. The priors, as their name indicate, are estimated from our prior knowledge of the problem. In some cases the training set is enough for this estimation but frequently more information is useful. Since the unconditional density does not depend on the classes, it can be dropped off from the maximization term. Having performed the estimation of priors and likelihood and using Eq. (4.13), Eq.(4.14) can be reformulated as

$$C_{MAP} = \arg \max_{k=1\ldots K} \{P(x_{Test}|C_k)P(C_k)\}. \tag{4.15}$$

4.2.1 Bayesian classification and ICA

If we assume that the hypotheses for ICA hold for each one of the classes, we can learn the ICA transform for each class from the training set. Suppose and $\mathbf{W}_k$ and $\mathbf{s}_k$ are, respectively, the projection matrix and the independent components for class C_k with dimensions $N_k \times M$ and N_k. Then, from Eq. (2.5),

$$\mathbf{s}_k = \mathbf{W}_k(\mathbf{x}_k - E(\mathbf{x}_k)) \tag{4.16}$$

We then learn the distribution for each of the N_k components of $\mathbf{s}_k$. Different methods for the unidimensional density estimation of the independent components are presented in Section 4.1. Having estimated the marginal densities, we can assume that the i-th component of vector $\mathbf{s}_k$ has density $P_k^i(\mathbf{s}, \theta)$ where θ represents the estimated parameters for the density, and $i = 1 \ldots N_k$ depends on the dimension of the ICA Model for class C_k. From the assumption of independence we have

$$(\mathbf{s}_k) \sim P_k(\mathbf{s}, \theta) = \prod_i P_k^i(s, \theta) \tag{4.17}$$

If we don´t have any a priori information we assume equiprobable priors for our classes so the Maximum a Posteriori rule becomes the Maximum Likelihood Rule and Eq. (4.15) becomes,

$$C_k = \arg \max_{j=1\ldots K} P_j(\mathbf{s}_j, \theta) = \arg \max_{j=1\ldots K} \prod_i P_j^i(s_j^i, \theta) \tag{4.18}$$

4.3 A practical example

The ICA Classification Model was applied and tested for pharmaceutical product classification through their color distributions. Figure 4.2 shows some of the products used in experiments. These distributions were learnt from local histograms extracted from the neighborhood of selected keypoints. The keypoints were selected using a structure tensor technique. Both the independence assumption and high sparsity of the independent components proved to be an important advantage for the modelling of the density, simplifying considerably the problem. On the other side, a straightforward implementation of a Bayesian criterium requires the estimation of the densities using nonparametric techniques such as Gaussian Kernels and an incorrect assumption of independence.

Figure 4.2: Subset of 25 pharmaceutical products used in our experiments.

The results of both approaches were compared and the improvement of the ICA approach is evident. This is shown in the experiments made for the problem of classifying 80 pharmaceutical products in a controlled environment. Through Bayesian ICA Classification, a total of 87.1% products were correctly classified, while the straightforward implementation only achieved 79.8% of correct classifications.

Chapter 5

Basic Concepts for PDMs

5.1 A Brief Introduction

As mentioned, the Point Distribution Model (PDM) [19] is a shape description technique based on the vectorized representation of shapes to estimate a statistical model for shape variation. By modeling this distribution, we can generate new examples, similar to those in the original training set, and we can also examine the plausibility of new shapes. It has been seen that this model succeeds in the treatment of non-rigid shapes, their analysis and synthesis. The statistical modeling for shape variation and its combination with several image processing techniques has generated an important number of applications in the last years. These applications include tracking, recognition, biomedical imaging, special effects for film and television and registration among others [29, 58, 7].

The construction of an appropriate PDM for a certain type of shape, requires both the selection of a good representation and of an appropriate density estimation method for the distribution of the shapes within this representation. For the representation we can use linear or nonlinear models. As usual, if we use a nonlinear model we can relax the hypotheses and obtain higher reliability. This precision has a high cost in the training stage due to the undeterministic characteristic of nonlinear models and also in the application stage, due to the fact that nonlinear algorithms are generally computationally expensive. This is justifiable for a large number of problems and several nonlinear models have been proposed [56, 57, 11, 30]. Even though, a linear representation is still a common choice for their speed and straightforward interpretability. As a matter of fact, most of the nonlinear representations are applied over a linear representantion which previously performs the dimensionality reduction. On the other side, even when the training set generates complex distributions, a linear repre-

sentation can be used and complexity charged to the statistical model [18]. The object of this dissertation is to present an alternative linear representation for the shapes, to show that, by choosing this representation, the complete framework is greatly simplified, and to provide an interpretation to the characteristics of the representation.

The most successful linear representation so far, for its simplicity and straightforward interpretation, is the one obtained through a Principal Component Analysis (PCA). PCA projects our data in an orthogonal subspace spanned by the uncorrelated directions of the training data´s maximum variance. By projecting a shape in a previously learnt PCA space, we have a set of coefficients or parameters (the principal components) which control the variation along these maximum variance directions. So we can naturally associate each principal component to a mode of variation of the shape.

5.2 The Point Distribution Model

If we use n points to describe a certain shape in d dimensions, we can represent this shape by a $N = nd$ dimensional vector by simply concatenating the point position values. Given K samples of a certain shape, we choose certain locations as key points or *landmarks points*, and obtain K vectors representing each shape of the training set. In order to be able to compare these points, a certain alignment in an approximate sense is necessary. Procrustes method [27] or modifications are frequently used in this stage. The selection of a correct criteria for alignment should not be underestimated since these operations will greatly affect the final distribution by introducing or avoiding nonlinearities. For PDMs to be used in Active Shape Models, Cootes [19] suggests aligning by minimizing a sum of squared distances. Mathematically, the expression $T_{\mathbf{t},\theta,s}(\mathbf{y})$ represents the application to shape $\mathbf{y}$ of a translation by $\mathbf{t}$, a rotation by θ and a scaling by s. Shape $\mathbf{y}$ is said to be aligned with reference shape $\mathbf{y}_{ref}$, if T minimizes the expression

$$|T_{\mathbf{t},\theta,s}(\mathbf{y}) - \mathbf{y}_{ref}|_W \tag{5.1}$$

Where $|\mathbf{y}|_W = \mathbf{y}^t W \mathbf{y}$ and $\mathbf{W}$ is a diagonal matrix of weights for each point. An iterative method for aligning a set of shapes is also provided in the cited article. This method, effective in the general case, can be improved for particular problems. For instance if perimeter invariance is known in advance, the scale of the aligned shapes can be forced to depend only in the perimeter. From here we will assume that our aligned training set is a sample of the random vector $\mathbf{x}$.

The next step is to find a proper representation for $\mathbf{x}$. In the choice of the representation simplicity, dimensionality reduction, statistical properties and interpretability should be considered.

5.3 The PCA Representation

From the training set, we can estimate both the mean of the data $\bar{\mathbf{x}}$ and its covariance Σ. From a simplified point of view, the covariance matrix tells us the way each landmark tends to move, as the others move. Using the covariance matrix, we obtain the PCA transform given by Eq. (1.6). So if $\mathbf{b} = \mathbf{s}_P$ is the vector of principal components, we can consider them as parameters ruling the displacement of the landmarks. Due to PCA theory, each parameter controls an uncorrelated variation (also called deformation), and the order of the parameters agrees with the degree of variation. In the shape context, the parameters or principal components, given by the components of vector $\mathbf{b}$ are often referred to as *modes of variation*. The choice of an appropriate value for the dimension of $\mathbf{b}$ (M) can be done in several ways, the most frequent is based on the proportion of variance we wish to capture in the subspace.

5.3.1 Density Models for the PCA Representation

We now choose a proper statistical density model for our shape representation and address the problem of examining the plausibility of new shapes or equivalently, generating new examples within the model. For each model we also present a solution for the problem of, given a certain shape, finding the nearest feasible shape within our model.

A reasonable definition for the plausibility of a shape can be the following. If we have estimated, from the training set, the distribution of the parameters $\mathbf{b} \sim p(\mathbf{b})$, a shape with parameters $\mathbf{b}'$ is said to be feasible if

$$p(\mathbf{b}') > p_t \tag{5.2}$$

where p_t is a certain threshold we consider appropriate. Since a threshold value based on the likelihood is not recommendable, it is usually chosen so that some proportion of the training set passes the threshold. If the parameters $\mathbf{b}$ are assumed Gaussian and independent, we have that

$$\log p(\mathbf{b}) = -\frac{1}{2}\sum_{i=1}^{M}\frac{b_i^2}{\lambda_i} + K \tag{5.3}$$

In this case, the threshold represents a likelihood which constrains feasible shapes to a hyperellipsoid. The size of the hyperellipsoid can be obtained considering that the sum of the square of gaussian variables has a chi-squared distribution,

$$\sum_{i=1}^{M} \frac{b_i^2}{\lambda_i} \sim \chi^2(M) \tag{5.4}$$

From Eq. (5.4) we can, given a certain probability value, obtain the desired threshold p_t. In finding the nearest feasible shape we first check the likelihood. If it is lower than our threshold, then our current shape is not feasible. The nearest feasible shape is that shape belonging to the intersection of the hyperellipsoid and the line passing through our current shape and the origin.

Another approach, is to choose hard limits on each direction [19]. This is related with the idea of statistical independence of the components of the parameter vector. It is equivalent to constraining feasible shapes to a hypercube. A good heuristical value for the threshold on each direction is 3 times the standard deviation on that direction. If we assume a gaussian distribution on each direction, this choice of limit values means that a shape is plausible if it belongs to the symmetrical mean-centered interval which has a marginal probability of 0.997. In this case, for each $i = 1, \ldots, M$, the feasibility of a shape is checked by $b_i' < 3\sqrt{\lambda_i}$ and the nearest feasible shape is obtained by

$$b_i^F = \text{sign}(b_i') * \min\left(3\sqrt{\lambda_i}, |b_i'|\right) \tag{5.5}$$

If a simple gaussian estimation is not enough we can use more complex models. A useful approach is to model $p(\mathbf{b})$ using Gaussian Mixture Models (GMM),

$$p_{GMM}(\mathbf{b}) = \sum_{l=1}^{L} w_l G(\mathbf{b}, \mu_l, C_l) \tag{5.6}$$

where L, μ_l, C_l are the parameters for the GMM and can be estimated with parameter estimation algorithm such as Expectation Maximization [23]. In this case, the plausibility of a shape is a more complex problem. Even though more precise solutions can be developed, a simple one consists on deciding that a shape is plausible if its likelihood is above the likelihood of a certain percentage of shapes in the training set. The percentage value is generally above 80%. When using a GMM, a general solution for the problem of finding the nearest feasible shape is not available so Monte Carlo and gradient descent methods are employed. Moreover, we have mentioned that estimating the GMM parameters on high dimensions is a highly unstable problem.

Chapter 6

ICA as a representation for PDMs

6.1 The ICA Representation

Assuming we have learnt the mixing and filter matrix for the ICA models (2.4) and (2.5), the ICA parameters $\mathbf{s}$ are obtained as for PCA

$$\mathbf{s} = \mathbf{W}(\mathbf{x} - \bar{\mathbf{x}}) \tag{6.1}$$

We will call $\mathbf{s}$ the independent components or *independent modes of variation*, and assume that the components s_i, $i = 1, \ldots, M$ are statistically independent. From Eq. (6.1) it can be seen that the independent components have zero mean and we can also assume, without loss of generality that they have unit variance [17]. The choice of dimension is not as straightforward as in PCA, where a natural hierarchy arises from the corresponding eigenvalues. While this is still an unsolved problem (how many independent components are *really* independent?) there exist several approaches [53]. Because of the shape problem in which it is logical to assign small variances to errors in the labelling process so, when required, we will reduce dimensionality by first performing PCA and then ICA. Nevertheless, we shall see that dimension is not as relevant as when we are using a PCA representation.

6.2 Statistical Density Models for the ICA Representation

Because of the assumption of independence, we need only to model the one-dimensional densities corresponding to the M independent components of pa-

rameter vector $\mathbf{s}$. This can be done through any of the methods presented in Section 4.1. The complexity of the method employed for density estimation is not relevant since we are working with a single dimension and the calculations need only be performed while training our PDM. Additionaly, because of the nature of PDMs, we will assume that the density of s (any component of $\mathbf{s}$) is likely to be one of a few particular densities. We informally base this assumption on the fact that distributions ruling a certain mode of variation cannot be entirely arbitrary.

The frequency of a certain position for a particular shape will affect the sub or supergaussianity of the modes of variation. A mode of variation of a shape which has a preferred position and seldom deforms will have a sparse or supergaussian distribution. On the other side, a mode of variation with almost equiprobable states along its variation range is clearly subgaussian. When preparing a training set for the generation of a PDM there is a tendency towards generating uniform distributions of the modes of variation. By doing this, we might be losing some prior information that can be useful, but this information does not affect our final objective of examining the plausibility of shapes. This tendency also favours subgaussian distributions. A shape has a point of equilibrium which is generally close to the mean shape of all the samples we can take. Symmetrical and gradual deformations of a part of the shape correspond to continuous symmetrical distribution. This is a frequent situation. In the particular shape problem skewness is more related with the sampling than with the real deformations of a shape. When a shape can be found in different states but doesn´t deform continuously from one state to the other we find clusters in the distribution of the mode of variation. Clusters can be also introduced by the incorrect identification of landmark points. We conclude that our density model should be open to both sub and supergaussian distributions, but particularly the first. Symmetry is very frequent, and it should also include multimodal densities, unless we have some other prior knowledge.

The broadest approach can be simply to use a histogram approximation. This is good but its discrete nature introduces complexity in the equations. Immediately related is to use a kernel method with Radial Basis Functions. If we consider the kernel method as excessive we can always use mixture models such as GMMs (see Eq. (5.6)) or other more specific models. This semiparametric methods are good for modeling both unimodal and multimodal densities but don´t succeed when we wish to model highly subgaussian densities such as a uniform density. All the different situations were tested and finally experiments were done using GMMs when possible, and RBF kernel-based methods otherwise.

6.3 Shape plausibility with ICA

As in PCA, the plausibility of a shape can be decided by evaluating its likelihood against a threshold value. In this section we will suggest a method for selecting the threshold which also provides a way for efficiently analysing shape plausibility and solves the problem of finding the nearest feasible shape. Suppose we have estimated the density for each of the independent modes of variation with one of the methods suggested above so that $s_m \sim p^m(s)$.

Given a certain probability value P_t between 0 and 1, let $p_t = P_t^{\frac{1}{M}}$. For each component there exists a union of disjoint intervals

$$I^m = [a_1^m, a_2^m] \bigcup [a_3^m, a_4^m] \bigcup \ldots \bigcup [a_{2t_m-1}^m, a_{2t_m}^m]$$

such that for all $m = 1 \ldots M$ I^m satisfies

$$\int_{I^m} p^m(s) ds = p_t$$

and

$$p(a_i^m) = l^m, \qquad \forall i = 1, \ldots, 2t_m$$

Once we have the set of intervals for each component, using the assumption of independence, it can be seen that

$$\int_{I^1 \otimes \ldots \otimes I^M} p(\mathbf{s}) d\mathbf{s} = \prod_{m=1}^{M} \int_{I^m} p_m(s_m) ds_m = \prod_{m=1}^{M} p_t = P_t \tag{6.2}$$

A constructive method which shows the existence of these intervals in a bimodal density obtained from experiments is exposed in figure 6.1. We first assume the probability density is continuous in $\mathbb{R}$. Given the likelihood value l, it can be seen that if the line $y = l$ intersects the function $y = p^m(s)$ it has to be in an even number of points. These points determine the interval borders. If the intersection is empty, we define I^m also as the empty set. The method consists in starting at a likelihood above the maximum and decreasing the likelihood value, thus increasing the probability, until the threshold is reached.

In practice, this can be implemented in several different ways, depending on the density model. For certain parametric and semiparametric models, both the likelihood and the interval borders can be obtained analytically. This is performed in the training stage. Any algorithm working on new shapes will need only the interval information for plausibility tests. Dividing each direction

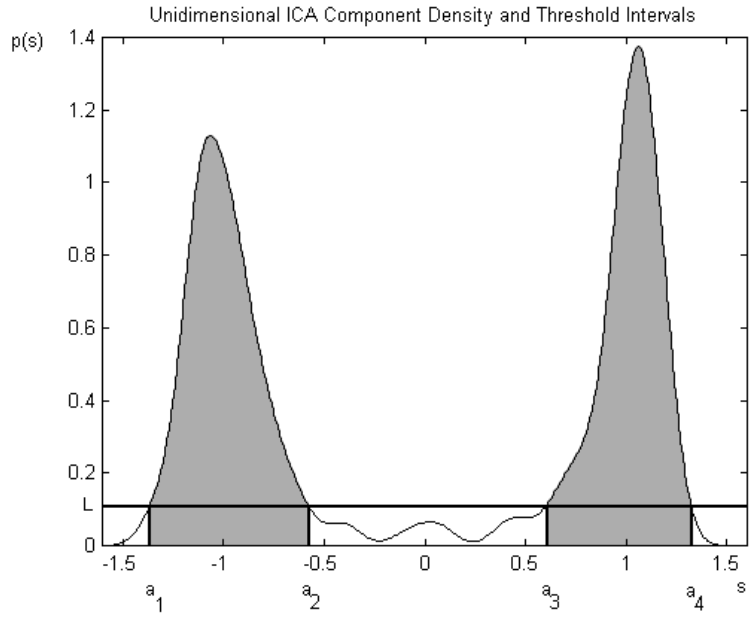

Figure 6.1: For a certain bimodal independent mode of variation, two intervals (to be used in shape feasibility) and a certain probability value corresponds to the likelihood value L.

in intervals divides the whole space into "hyperboxes" which have a geometric distribution reminiscent of that which arises from separable functions. In figure 6.2 the joint distribution of two independent directions obtained in experiments are plotted. Each marginal density (both clearly bimodal) was estimated with a kernel-based method. The product of the marginal densities is plotted with gray levels and contour lines. The rectangular boxes represent the cartesian product of the intervals estimated for each direction for $P_t = 0.95$.

Working with these intervals has several advantages. Since the calculation of the intervals is performed in the training stage, all complex algebraic operations are removed from the working algorithms. This is because there is no need for the calculation of likelihoods once we have the interval limits. This interval structure also provides precision. It can be seen in figure 6.2 that, if we decided to use a GMM, the only way to improve the estimation would have been using more than four components in the mixture. This doesn´t sound too bad in a two dimensional context but is quite nasty as dimensions increase and we have no prior knowledge of the structure. Additionaly a GMM would need complex calculations throughout the algorithm.

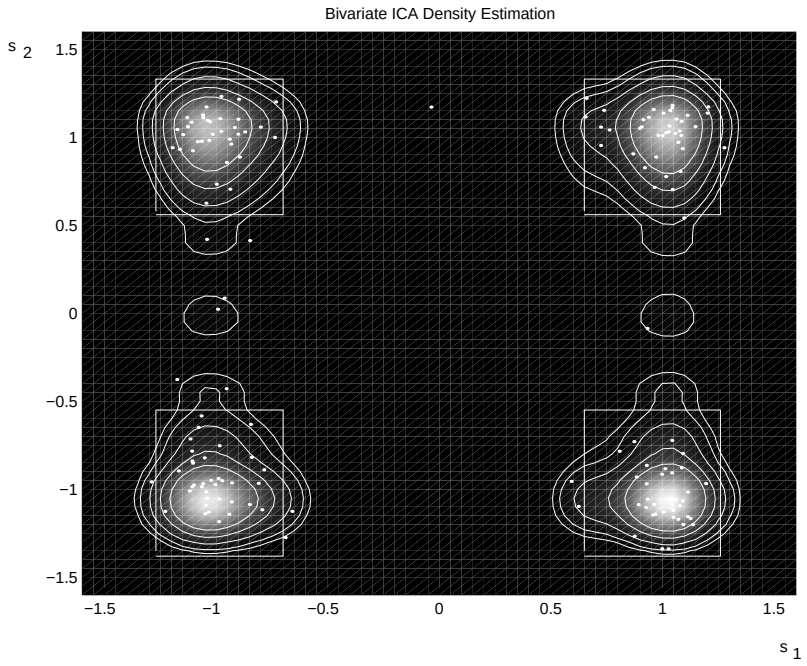

Figure 6.2: In the bivariate case, the ICA intervals for testing shape feasibility. The curves represent contour lines of the density estimation (obtained with a kernel method), and the rectangles represent the cartesian product of the intervals, capturing 95% of the probability.

In the interval context, plausibility is easily checked by first projecting the shape in the parameter space and then by verifying if $s^m \in I^m$ for all $m = 1, \ldots, M$.

6.4 Nearest feasible shape with ICA

Given the intervals I^m, and a shape with independent modes of variation s_T^m, the nearest feasible shape $\mathbf{s}_F$, with components s_F^m is

$$s_F^m = \begin{cases} s_T^m & \text{if } s_T^m \in I^m; \\ \arg\min_{1 \le i \le 2t_m} |s_T^m - a_i^m| & \text{otherwise.} \end{cases} \tag{6.3}$$

It is important to notice that the simplicity of this formulation would not be possible without the independence assumption, keeping in mind that the alternative method when using a PCA representation is an expensive Monte Carlo algorithm. Until now the truth is that a less rigid attitude is taken, and

independence is assumed within the PCA representation. As a consequence, to obtain correct results, we need a too rigid model.

Chapter 7

Experiments

7.1 Artificial set of shapes

First, an artificial set of shapes was created. In each shape we use 19 points to describe a fixed base and three deformable extensions of fixed length (see figure 7.1). Each extension can be found rotated in an angle between $-\frac{\pi}{4}$ and $\frac{\pi}{4}$. We created a training set of 400 shapes, choosing randomly the angle corresponding to each extension. The shapes have then, three independent degrees of freedom. The objective of this experiment is to observe if ICA effectively separates the independent variations, associating a component to each independent degree of freedom. The alignment step was skipped since the shapes were already aligned when created. Only centering and appropriate rescaling was necessary.

Figure 7.1 shows the three principal modes of variation assigning three values to each side of the mean. PCA decorrelates each of the movements but does not take in account statistics of higher order. The decorrelated movements have no relationship with the degrees of freedom chosen in the creation of the shapes.

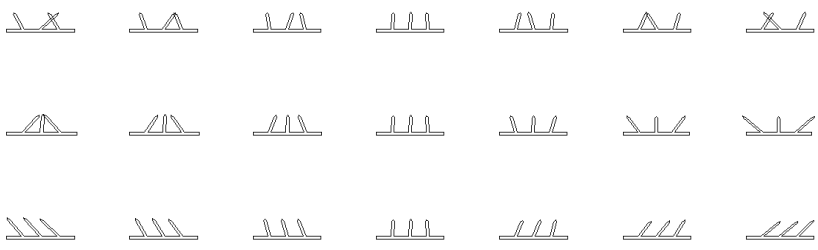

Figure 7.1: Three first modes of variation using PCA.

Instead, figure 7.2 shows the three independent modes of variation. We observe how ICA successfully separated the deformations corresponding to each one of the extensions. If the original problem would have been to classify a certain shape according to its deformations, it is observed how ICA would have successfully solved this problem. Also, if the problem would have been the artificial generation of movement through continuous plausible shape positions, it is observed how ICA would have allowed much more control over the variations.

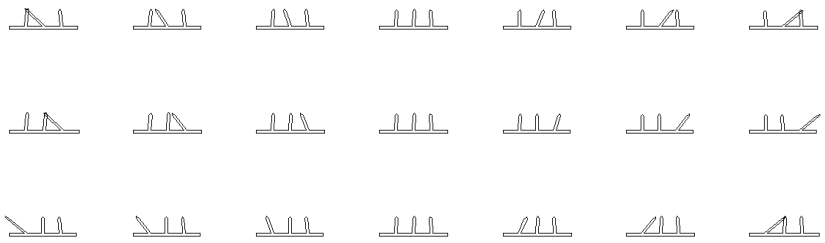

Figure 7.2: Three first modes of variation using ICA.

7.2 Open Hands

Experiments were also performed on a set of shapes representing hands. These hands were described by 55 points and were obtained from an image dataset not specifically generated for the shape problem. Only hands with extended fingers were considered and these were found in many positions, distances and planes from the camera, totalling 160 hands. Due to this lack of control, the PCA space of parameters needed a dimension of at least 37 to capture 99% of the data variation. In figure 7.3 we observe the two first principal modes of variations. The variation of the first mode moves all the fingers in the hand, the second is practically the same except for the fact the hand has undergone an affine transformation. We observe in the latter a rotation around the vertical axis of the palm.

The first two independent modes of variation are exposed in figure 7.4. These two modes were obtained by performing ICA on the principal modes of variation of dimension two. It has been observed that this can bring up corrupted independent components [53]. Even though this seems to be the case, we observe an interesting difference between both independent modes of variation. While the first one does practically the same as the first principal mode of variation,

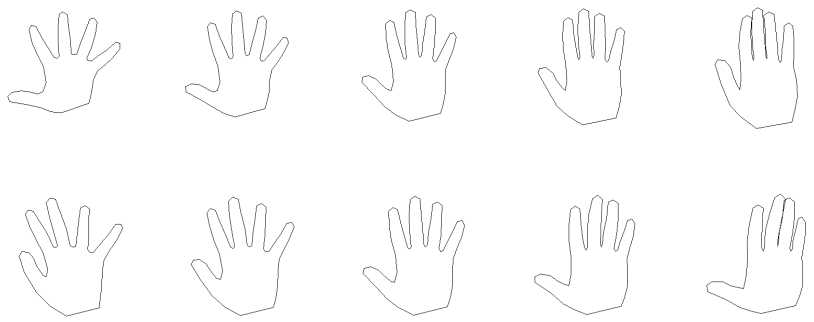

Figure 7.3: Two first modes of variation using PCA.

the variation of the second mode has the movement of the thumb practically isolated. The variation of the rest of the fingers is very small except, maybe, for that of the small finger which still is subtle.

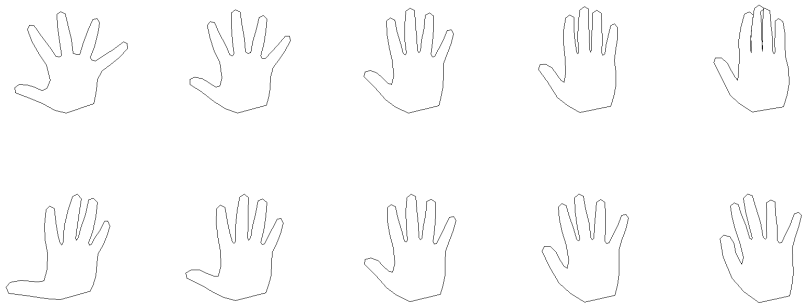

Figure 7.4: Two first modes of variation using ICA.

When working with more parameters (more modes of variation) results are also more difficult to interpret. The main reason for this is the bad quality and small size of the training set. This last point is evident from the fact that 55 two dimensional points are represented through vectors of dimension 110, and the size of our dataset is 160. This fact seriously affects the ICA performance, as we have observed in similar experiments. Nevertheless, some interesting remarks can be made. In all cases, there is an independent component which captures only the variation of the thumb. Modes of variation involving other fingers usually involve also the movement of the thumb, which is quite natural. For instance, it is very difficult to displace the index leaving the thumb static. By observing the possible variations in our open hands we might notice that the

displacement of single fingers has complex dependencies except for the case of the thumb.

In figure (7.5) we have plotted the first two modes of variation for ICA and PCA for the same dataset of hands. The limits for plausible shapes are also shown. In the PCA case, the rectangle responds to the choice of hard limits along three times the standard deviation. The ellipsoid results from assuming the distribution is normal bivariate. It can be seen that none of these assumptions hold. In the ICA case, the limits for plausibility were obtained with the interval method exposed. A kernel method was employed for the density estimation and the limits were chosen so that the intervals encompass for 98% of the probability. In the artificial case, the same situation is observed and the precision is much more drastical.

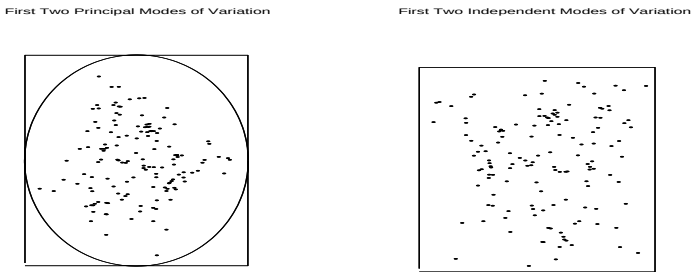

Figure 7.5: Two first modes of variation for the PCA and the ICA representation respectively. The limits shown are the limits for plausible shapes. In the PCA case assuming independence (square) and normality (ellipse). In the ICA case assuming independence. The higher precision of the latter is observed.

Chapter 8

Summary and Conclusions

8.1 Summary

The problem of modeling point distributions using the parameter representation given by Independent Component Analysis has been focused. In the first chapter, the importance of effective feature extraction is mentioned. ICA is presented as a feature extraction technique useful in a wide variety of problems. The most useful estimation algorithms for the independent components are presented along with the theory that motivates them. The estimation algorithm which proved most effective in the experiments, and consequently the one we used most, is FastICA and it is detailed briefly. After introducing the basic concepts, applications of ICA are exposed as well as the relation between ICA and other important theories in the signal processing area. Finally the statistics within the ICA context are considered. Mainly the problems of density estimation and classification. Several nongaussian and unidimensional density families which appear frequently within the ICA context are mentioned, and some new useful families are introduced. The problem of classification is considered from a bayesian point of view, and this scheme is adapted to the ICA environment. A practical example of classification is briefly mentioned.

In chapter five, the Point Distribution Model is presented and some considerations concerning the classical perspective are made. These considerations motivate, for certain applications, the introduction of the independent modes of variation which is done in chapter six. Under this framework, the two main problems when working with PDMs are shape plausibility and nearest feasible shape. Both problems are addressed.

Experiments were made with both artificial shape sets and real shape sets. The results are exposed in chapter seven.

8.2 Conclusions

The assumption that uncorrelated parameters corresponding to high variance directions are good for modeling shapes does not necessarily hold for all problems. Even when it holds it does not necessarily provide a simple and robust framework. In all these cases, ICA can prove to be an interesting alternative.

In an ICA framework, modes of variation no longer represent uniquely the deviations within the shape and can now be thought of as independent variations. Independence of variations can provide a higher and clearer control over parameters. This can be seen in the experiment performed with an artificial set of shapes, where the problem of classifying according to displacement is immediately solved. In an ideally open hand, the position of the middle finger is statistically independent of the position of the thumb, so the value of two independent modes of variation could be identified with each of these two fingers. Under this perspective, ICA is a good choice where there are reasons to believe that the shape deformations are not related between them.

Moreover, ICA does provide a simple and robust framework. This is completely based on the assumption of independence. Reducing a classicaly multidimensional problem to a single dimension allows straightforward application of complex and accurate methods. Plausibility is a good example of the importance of precise density estimation. On the other side, independence provides a particular distribution landscape which avoids complex optimization methods. The feasibility intervals take advantage this. The use of feasibility intervals with a PCA representation is widely spread but sensible and unprecise. A wrong assumption of independence causes unsatisfactory results and this can be seen in figure 7.5. Plain observation of plausible shapes generated using this model support this. Feasibility intervals also contemplate multimodality. If there are reasons to believe that both the landmarking and the alignment processes are correct, multimodality is very unusual within this context, and additional precision is possible.

In treating the original objective (modeling point distribution with *highly* independent parameters), some other problems arise. Unidimensional density estimation is necessary, and this is basically a decision over a certain density family. Nevertheless, special care is important in high dimensions because the product of the marginal densities can output very low values. This makes the histogram distribution unfeasible and kernel methods very problematic. If parametric estimation proves too rigid, then semiparametric estimation such as that done using mixture models may provide a good solution. We shouldn´t forget that all calculations are performed in the training stage. The problem of classification is also focused. In this case, there is some additional information which

can improve results. For instance, in the example with pharmaceutical products, high sparsity of all components was observed. Interpretation of the results from the perspective of sparsity can add useful information which can be used for fine tuning such as the addition of classification layers. ICA classification has some disadvantages. Even though the learning of the ICA representation and the estimation of the densities is done only once in the training stage, the storage of this information may require large amounts of space, depending on the number of objects we wish to classify. This can be improved by implementing some kind of dimensionality reduction. But dimensionality reduction preserving independence is yet an unsolved problem. Results are also quite sensible to small variations in the ICA representation. This is observed when comparing different algorithms, or when adjusting the parameter settings within a single algorithm.

From all these results we can finally conclude that the ICA representation can be successfully used when there are reasons to think that different shape deformations correspond to independent factors, and the shapes we observe are linear mixtures of these deformations. In this case, ICA can not only separate the deformations allowing control and classification, but can also provide a robust and simple density estimation framework, also given within this work. An interesting observation is that when a poor performance of ICA is observed, PCA is generally not better. Being poor performance related with factors such as nonlinearities which affect both representations. On the other side, ICA outperforms PCA in problems such as those mentioned here.

8.3 Further Work

More experiments dealing with real shapes satisfying the exposed assumption of independent variations should be made in order to validate the method. The problem is that working with point distribution models has the very important drawback of manual landmarking. A very time consuming stage of the whole process. Automatic approaches have been proposed but only work under very specifical circumstances.

The problem of shape deformation frequently involves nonlinear variations, even in the very simple case. This can be solved in several ways and we are actually performing experiments in order to test the possible solutions. One of the experiments is the modeling of scissors as the one shown in figure 8.1. Scissors have a central axis given by the junction of the two blades and have radial displacement around this axis. This is one deformation. Different scissors present a wide variety of possibilities concerning the size of blades and handles. These are other deformations, statistically independent from the first. Using a cartesian

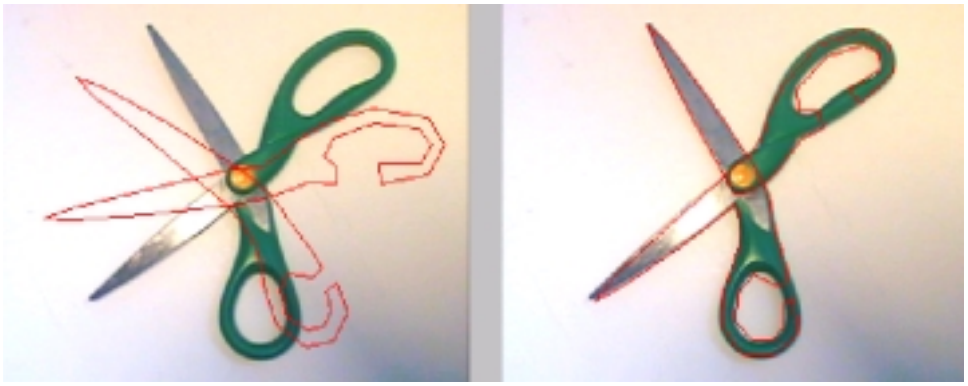

Figure 8.1: The shape of scissors using a point distribution model, in two stages of the same run of the algorithm of Multi-Resolution Active Shape Models [20].

coordinate system for the placement of the landmark points and performing the posterior alignment causes important nonlinearities in the parameter space. In these cases, ICA and PCA perform both badly. Surely, a cartesian-polar approach for the labelling of landmark points [28] would avoid the nonlinearity. Experiments are being done in this sense and results are encouraging. A more complex but also more general approach for avoiding nonlinearities can be developed using nonlinear ICA [38], or a combination of kernel PCA methods which provide a preprocessing for the ICA stage. The disadvantage of the first is that nonlinear ICA is not a developed area. Nevertheless much research is being done in this direction. The main disadvantage of kernel PCA is that projection in the original space is not possible. This makes results difficult to validate, being classification one of the few problems which could eventually be solved using this method.

Another difficulty arises from the problem of testing plausibility. There are few criteria other than the subjective direct observation. The testing of plausible shapes has to be made on real life problems such as quality control where a global measure of error can be obtained. This should also be done.

More experiments should be done using point distribution models for practical applications. In this sense, experiments were made within the context of Active Shape Models [19], specially in the case of hands. The results are not presented here because the differences with existing approaches is not significant. Surely one of the reasons for these poor results is that, in ASM, unaccuracy in the statistical model is allowed provided an efficient edge detection algorithm is used. Practical applications where testing can be interesting should be those that rely heavily on the statistical model, such as tagging.

From a theoretical perspective, the method of feasibility intervals is quite naive. It can surely be improved and research should be made in this sense. A general density estimation technique could easily be attached to ICA for a standard statistical model. This supposition is based on the close link between nongaussianity and ICA.

Bibliography

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

[2] K. Anand, G. Mathew, and V. Reddy. Blind separation of multiple cochannel bpsk signals arriving at an antenna array. *IEEE Signal Proc. Letters*, 2(9):176–178, Sept. 1995.

[3] J. Atick and A. Redlich. What does the retina know about natural scenes? *Neural Computation*, 4:196–210, 1992.

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

[5] B.A.Olshausen and D.J.Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7:333–339, 1996.

[6] H. Barlow, T. Kaushal, and G. Mitchison. Finding minimum entropy codes. *Neural Computation*, 1:412–423, 1989.

[7] A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models. Technical report, University of Leeds, School of Computer Studies, November 1994.

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

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

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

[11] R. Bowden, T. Mitchell, and M. Sahardi. Cluster based non-linear principle component analysis. *IEE Electronic Letters*, 33(22):1858–1858, 1997.

[12] G. Box and G. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1973.

[13] M. Bressan, D. Guillamet, and J. Vitriá. Using an ica representation of local color histograms for object recognition. Technical report, Centre de Visio per Computador, Universitat Autonoma de Barcelona, May 2000.

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

[15] J. Cardoso and A. Soloumiac. Blind beamforming for non-gaussian signals. In *IEE ProceedingsF*, volume 140(46), pages 362–370, 1993.

[16] E. Chaumette, P. Comon, and D. Muller. Ica-based technique for radiating sources estimation: application to airport surveillance. In *IEE Proc. -F*, volume 140 no. 6, pages 395–401, Dec. 1993.

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

[18] T. Cootes and C. 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.

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

[20] T. Cootes, C. Taylor, and A. Lanitis. Multi-resolution search with active shape models. In *International Conference in Pattern Recognition, ICPR94*, pages A:610–612, 1994.

[21] J. Daugman. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Transactions on Biomedical Engineering*, 36:107–114, 1989.

[22] L. de Lathauwer, B. de Moor, and J. Vandewalle. Fetal electrocardiogram extraction by source subspace separation. In *Proc. of IEEE SP workshop on Higher-Order Statistics*, volume 4, pages 134–138, 1995.

[23] A. 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.

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

[25] J. Friedman and J. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, C-23:881–889, 1974.

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

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

[28] A. Heap and D. Hogg. Automated pivot location for the cartesian-polar hybrid point distribution model. In *Pycock D., ed. British Machine Vision Conference 1995, BMVC'95*, volume 1, pages 97–106. University of Birmingham, UK:BMVA, 1995.

[29] A. Heap and D. Hogg. 3d deformable hand models. In *Gesture Workshop, York, UK*, March 1996.

[30] A. 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.

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

[32] P. J. Huber. Projection pursuit. Technical report, Dept. of Statistics. Research Report PJH-6, 1981.

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

[34] A. Hyvärinen. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11:1739–1768, 1999.

[35] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

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

[37] A. Hyvärinen, E. Oja, P. Hoyer, and J. Hurri. Image feature extraction by sparse coding and independent component analysis. In *Proceedings of Int. Conf. on Pattern Recognition 98, Brisbane, Australia*, pages 1268–1273, 1998.

[38] A. Hyvärinen and R. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

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

[40] M. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society*, ser. A 150:136, 1987.

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

[42] E. Kandel, J. Schwartz, and T. Jessel. *Essentials of Neural Science and Behavior*. Appleton and Lange, Norwalk, CU, 1995.

[43] J. Karhunen, A. Hyvärinen, R. Vigario, J. Hurri, and E. Oja. Applications of neural blind separation to signal and image processing. In *In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97), Munich, Germany*, pages 131–134, 1997.

[44] J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear pca type learning. *Neural Networks*, 7(3):113–127, 1994.

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

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

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

[48] S. Makeig, A. Bell, T. Jung, and T. Sejnowski. Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 8:145–151, 1996.

[49] S. Mallat. A theory of multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[50] J. Moody and L. Wu. What is the 'true price' ?— state space models for high frequency financial data. In *In Progress in Neural Information Processing. Proceedings of the International Conference on Neural Information Processing*, volume 2, pages 697–704. Springer-Verlag, 1996.

[51] E. Oja. The nonlinear pca learning rule in independent component analysis. *Neurocomputing*, 17(1):25–46, 1997.

[52] D. Pham, P. Garrat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *EUSIPCO*, volume 1, pages 771–774, 1992.

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

[54] T. Schreiber and D. Kaplan. Signal separation by nonlinear projections: The fetal electrocardiogram. *Phys. Rev.*, E 53(4326), 1996.

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

[56] P. Sozou, T. Cootes, C. Taylor, and E. 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.

[57] P. Sozou, T. Cootes, C. Taylor, and E. 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.

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

[59] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distribution.* John Wiley and Sons, San Diego, 1985.

[60] J. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. In *Proc.R.Soc.Lond.*, volume B 265, pages 359–366, 1998.

[61] R. Vigario, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Advances in Neural Information Processing Systems*, 10:229–235, 1998.