Redundancy Reduction with Information Preserving Nonlinear Maps

Lucas Parra[1,2], Gustavo Deco[2], Stefan Miesbach[2]
[1] Siemens AG, Corporate Research and Development, ZFE ST SN 41
Otto-Hahn-Ring 6, 81739 Munich, Germany
[2] Ludwig-Maximilian-Universität, Institut für Medizinische Optik,
Munich, Germany

Abstract

The basic idea of linear Principal Component Analyses (PCA) consists in decorrelating coordinates by an orthogonal linear transformation. In this paper we generalize this idea to the nonlinear case. Simultaneously we will drop the usual restriction to gaussian distributions. The linearity and orthogonality condition of linear PCA is substituted with the condition of volume conservation in order to avoid spurious information generated by the nonlinear transformation. This leads us to a still very general class of nonlinear transformations, called symplectic maps. Further on, instead of minimizing the correlation, we minimize the redundancy measured at the output coordinates. This generalizes second order statistics being only valid for gaussian output distributions to higher order statistics. The proposed paradigm implements Barlow's redundancy reduction principle for unsupervised feature extraction. The resulting factorial representation of the joint probability distribution presumably facilitates density estimation and is especially applied to novelty detection.

1 Introduction

The aim of PCA is to linearly decorrelate features by rotating the coordinates of the feature space. Among the research done in the Neural Network community there is a vast body of results related to linear PCA. The greatest part of it demonstrates how various types of neural structures can perform PCA, and how PCA can be derived from information theoretic principles. The latter results are especially important to gain insight in the various learning paradigms, at least in the simple case of linear transformations. However - mostly in the context of information theory - the results were restricted to gaussian distributions of the input signals. The aim of this paper is to make use of the more general principle

of redundancy reduction instead of linear decorrelation, furthermore to get rid of the restriction to linear transformations for PCA.

Independently of each other Baldi and Hornik (1989) and Bourlard and Kamp (1988) proved, that Back Propagation in a linear auto-association net finds in the hidden layer the orthogonal projection onto the subspace spanned by the first principal eigenvectors of the auto-correlation matrix. Oja (1989) formulated a hebbian type learning rule that extracts the first principal component and leads to a orthonormal linear transformation. Rubner and Tavan (1989) proved that a network composed of linear neurons can be trained to perform PCA, if the synaptic adaptation is Hebbian in a vertical sense, i.e. from inputs to outputs, and anti-Hebbian for the inhibitory lateral connections between the output units. A similar concept was proposed by Földiák (1989). The Hebbian and anti-Hebbian form of the learning rules can be derived from first principles by applying information theoretic concepts (Linsker, 1988; Kuehnel & Tavan, 1990).

On the other hand, within the last years there has been an increasing interest in theories and models for unsupervised detection of statistical dependence in a sensorial environment. One of the most important theories of feature extraction is the one proposed by Barlow (1959, 1989). Barlow describes the process of cognition as a preprocessing of the sensorial information performed by the nervous system in order to extract the statistically relevant and independent features of the inputs without loosing information. As a learning strategy Barlow (1959, 1989) formulated the principle of redundancy reduction. This kind of learning is called factorial learning, since in the case of zero redundancy the joint probability distribution can be expressed by the product of the single coordinate probability distributions. Recently Atick and Redlich (1990, 1992) and Redlich (1993a, 1993b) concentrate on the original idea of Barlow yielding a interesting formulation of early visual processing and factorial learning.

In the signal processing literature a closely related problem, called "blind separation of sources", has recently been addressed quite frequently. Here it is assumed that statistical independent signals are linearly superposed, and the task is to find the original signals. An outstanding work is that of Comon (1994) where he formalizes his Independent Component Analyses (ICA) proposed in (Comon, Jutten, & Herault, 1991). He restricts himself to linear transformations, but solves the problem of non-gaussian distributions by measuring the redundancy in terms of higher order cumulants. Some papers have been published addressing the nonlinear case, but they lack either in generality or in their theoretical foundation (Karhunen & Joutsensalo, 1994; Burel, 1992). At last we want to mention the work of Hastie and Stuetzle (1989) that proposed a statistical technique for finding "Principal Curves". The drawback of the proposed algorithm is the lack of an analytic representation of the curves found.

Obradovic and Deco (1993) apply Barlow's principle to a class of general linear transformations which conserve the input information but minimize the mutual information between the outputs in order to make them statistically

independent. Among other special cases, PCA is included in this class satisfying the additional condition of norm or energy preservation.

Some nonlinear extensions of PCA for decorrelation of input signals were recently introduced. These follow very closely Barlow's original ideas of unsupervised learning. Deco and Parra (1994) defined a stochastic neural network of Ising spins that performs statistical decorrelation of Boolean outputs by nonlinear transformations using information theoretic concepts. Deco and Brauer (1994) formulated a neural paradigm for the statistical decorrelation of the output components by using a special volume conserving architecture and higher order cumulant expansions. Deco and Schürman (1994) applied triangular volume conserving architectures and information based decorrelation learning algorithm for modeling chaotic time series. Atick and Redlich (1992) and especially the two papers of Redlich (1993a, 1993b) use similar information theoretic concepts and reversible cellular automata architectures in order to perform nonlinear decorrelation. Taylor and Coombes (1993) presented an extension of Oja's learning rule (1989) for higher order neural networks.

The aim of our work is to formulate a connectionist network architecture that performs Barlow's unsupervised redundancy reduction learning in the most general fashion. The basic idea is to define an architecture that assures perfect transmission without loss of information. Thereby we make use of a special class of nonlinear transformations, the so-called symplectic transformations, the key property of which is volume preservation. Unless one has a priori knowledge about the environment (i.e. distribution of the input signals) it is difficult to find criteria for separating noise from useful information. Without having a good model of the probability distributions of the environment, one should preserve the information of the measured signal. The second reason for conserving the information passed through a system, is more technical and arises from the definition of the Shannon information of continuous distributions. Shannon information is sensitive to scaling. By scaling continuous coordinates the amount of information of a signal distribution changes. To avoid spurious information generated by such a transformation, we require volume preserving maps.

In section two we will first introduce the class of symplectic transformations in context of the information carried by a continuous random variable. In section three we will explain our approach of reducing redundancy by minimizing a simple upper bound of the entropy. In section four we present a parametrization of symplectic maps by a connectionist network structure and perform gradient calculations. In the following section we describe how the nonlinear redundancy reduction technique presented in the paper can be used for density estimation and novelty detection.

2 Entropy Preserving Nonlinear Maps

The entropy $H\{\mathbf{x}\}$ of a distribution $P(\mathbf{x})$ of a continuous random variable $\mathbf{x}$ in R^n is defined as (Shannon, 1948):

$$H\{\mathbf{x}\} \equiv H[P(\mathbf{x})] = -\int_{R^n} P(\mathbf{x}) \ln P(\mathbf{x})\, d\mathbf{x} \tag{1}$$

Unlike the discrete case this entropy might be negative, e.g. the one point distribution, in the continuous case represented by Dirac's delta distribution, has corresponding to (1), infinite negative entropy. Consider an arbitrary transformation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ of the random variable $\mathbf{x}$ in R^n. The corresponding transform of the entropy of the distribution $P(\mathbf{x})$ can be express by (Papoulis, 1991)

$$H\{\mathbf{y}\} \leq H\{\mathbf{y}\} + \int_{R^n} P(\mathbf{x}) \ln \left(det \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right) \right) d\mathbf{x} \tag{2}$$

If $\mathbf{f}(\mathbf{x})$ is invertible, equality holds in (2). The Jacoby determinant in the integral yields the scaling caused by the transformation $\mathbf{f}(\mathbf{x})$. For an arbitrary distribution $P(\mathbf{x})$ the transformation would generate spurious information, unless the determinant is equal one everywhere

$$det \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right) = 1 \tag{3}$$

As an unsupervised method, that doesn't use any *a priori* knowledge about the environment, linear PCA accomplishes this requirement in a very natural manner. It performs only rotations, which are obviously volume conserving. In order to generalize to nonlinear transformations, the only condition to be maintained is volume conservation. A general class of transformations that fulfill condition (3) are the symplectic maps (Abraham & Marsden, 1978) originally introduced by Siegel (1943) in the context of multivariate function theory. In classical mechanics this class coincides with the well-known class of canonical transformations. A very interesting and for our purpose important fact is that any non-reflecting [1] symplectic transformation can be expressed implicitly in terms of a scalar function

$$\mathbf{y} = \mathbf{x} - J^{-1} \frac{\partial}{\partial \mathbf{z}} S \left(\frac{\mathbf{x}+\mathbf{y}}{2} \right) \quad ; \qquad J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} = -J^{-1} \tag{4}$$

where I denotes the unity matrix in $R^{n/2}$. The gradient is to be taken with respect to the argument $\mathbf{z} = (\mathbf{x}+\mathbf{y})/2$ of $S(\mathbf{z})$. [2]

[1] A transformation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ in R^n is called non-reflecting iff $det(I - \frac{\partial \mathbf{f}}{\partial \mathbf{x}}) \neq 0$ i.e. iff the Jacobian of $\mathbf{f}(\mathbf{x})$ has no eigenvalue -1.

[2] This representation of symplectic maps is a special case of the generating function theory

Unfortunately, there is no general way of constructing $S(\mathbf{z})$ analytically with a given symplectic $f(\mathbf{x})$ and vice versa. We could either define a function $f(\mathbf{x})$ and prove the volume conservancy, or we could define an arbitrary scalar function $S(\mathbf{z})$ and realize the map by solving (4) numerically. The focus of this work is to have the most general possible map. Therefore we do not restrict to a certain structure given by an ad hoc definition of $f(\mathbf{x})$. We will instead use a network structure that has shown to be a general function approximator (Hornik, Stinchcombe, & White, 1998), in order to have a flexible expression of $S(\mathbf{z})$.

3 Mutual Information and Factorial Representation

Let us now focus the attention on statistical independence. Consider a random variable $\mathbf{y}$ in R^n distributed according to a joint probability distribution $P(\mathbf{y})$. In general the components of the input are statistically dependent. Statistical independence occurs when the joint probability distributions factorizes,

$$P(\mathbf{y}) = \prod_i^n P(y_i) \tag{5}$$

The definition of mutual information relies strongly on this definition of independence. When random variables are statistically independent one would expect vanishing mutual information. A distance measure for entropies is the relative entropy. The relative entropy between the left- and right-hand side of (5) defines the mutual information of a multidimensional random variable:

$$0 \leq MI\{\mathbf{y}\} = -H\{\mathbf{y}\} + \sum_i^n H\{y_i\} \tag{6}$$

Redundancy is then defined as the normalized mutual information $R\{\mathbf{y}\} = MI\{\mathbf{y}\}/H\{\mathbf{y}\}$. Note that (5) is equivalent to $MI\{\mathbf{y}\} = 0$ (Atick & Redlich, 1992). How does this relates to linear decorrelation used in linear PCA? In the case of Gaussian distributions diagonalizing the correlation matrix of the linearly transformed variables has proven to be equivalent to reducing redundancy (Papoulis, 1991). But for general distributions decorrelation does not imply statistical independence of the coordinates. Starting from the principle of minimum redundancy Deco and Brauer (1994) formulated criteria for higher order decorrelation. Another approach with higher order cumulants for the case of linear transformations was studied by (Comon, 1994).

developed in full generality by Feng and Qin (1985). A proof of the representation (4) and a discussion of its role for the numerical integration of Hamiltonian Systems can be found in (Miesbach & Pesch, 1992)

In this work we make use of the more general principle of redundancy reduction instead of pairwise decorrelation. Combining this with the preservation of the input information by using symplectic maps, we arrive at the principle of feature extraction formulated by Barlow (1959, 1989). This principle states, that the preprocessing of cognition tries to find a factorial representation of the environment while preserving the information contained in the sensory stimulus.

In order to find a factorial representation in the output coordinates , we use the mutual information (6) as a cost function. Since for the symplectic map $H\{\mathbf{x}\} = H\{\mathbf{y}\}$ holds, we are left with the task of minimizing the sum of the single coordinate entropies (second term in the left-hand side of (6)). Thus we get not involved in computational expensive multi-coordinate statistics. On the other hand, we need an analytic expression or at least a measurement of the single coordinate entropies. Since we are given only a set of data points, drawn according to the output distributions, this is still a difficult task. But fortunately we know a feasible upper bound for these entropies,

$$
\begin{aligned}
0 \leq MI\{\mathbf{y}\} &= -H\{\mathbf{x}\} - \sum_i^n \int_R P(y_i) ln P(y_i)\, dy_i && (7)\\
&\leq -H\{\mathbf{x}\} - \sum_i^n \int_R P(y_i) ln G(y_i)\, dy_i && (8)\\
&= -H\{\mathbf{x}\} + \frac{n}{2}\ln(2\pi) + \frac{1}{2}\sum_i^n \int_R P(y_i)(y_i - \langle y_i\rangle)^2\, dy_i && (9)
\end{aligned}
$$

where $\langle y_i\rangle = \int_R P(y_i)\, y_i\, dy_i$. The inequality in (7) expresses the positiveness of the relative entropy between two arbitrary distributions, in that case between P and G where we used the Gaussian distribution $G(y) = 1/\sqrt{\pi}\exp(-(y - \langle y\rangle)^2)$. The first two terms of the expression (9) are constants. Thus we can reduce the problem of statistical decorrelation to the problem of minimizing the upper bound of the entropies $\sum_i H\{y_i\}$, i.e. the sum of the output variances. We could also use gaussian distributions $G(y_i)$ with different variances for each coordinate, namely the variance of the distributions $P(y_i)$. This is the idea of the second Gibbs theorem. This theorem gives the entropy of the Gaussian distribution having the variance of $P(y_i)$ as an upper bound for the entropy $P(\mathbf{y})$. In this case the upper bound is the sum of the logarithms of the variances:

$$
MI\{\mathbf{y}\} \leq -H\{\mathbf{x}\} + \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\sum_i^n \ln\left(\int_R P(y_i)(y_i - \langle y_i\rangle)^2\, dy_i\right) \qquad (10)
$$

These two different cost functions yield slightly different learning goals. In case of constant volume the direct sum of variances (9) favors equal variances in the different coordinates, while the sum of the logarithm of the variances (10)

does not favor any special scaling between the variances. In all experiments we used cost function (9). It might be seen as a strong simplification, to use only the second order moment for minimizing the entropy. One should rather try to include higher order cumulants to get a more accurate estimate of the distributions and therefore of the corresponding entropies. But in a first stage we want to restrict to a computationally efficient solution. Beside of this, if the transformation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ is flexible enough this cost function will tend to produce gaussian distributions at the output. With a variational approach it can be shown that under the constraint of constant entropy a circular gaussian distribution extremizes the sum of variances. If the transformation is general enough to assume arbitrary output distribution, then the cost function (9) can be seen as a functional of $P(\mathbf{y})$ to be extremized. We introduce the side conditions $\int_{R^n} P(\mathbf{y})\, d\mathbf{y} = 1$ and $\int_{R^n} P(\mathbf{y}) \ln P(\mathbf{y}) d\mathbf{y} = const$ with Lagrangian multipliers λ_1 and λ_2. The functional - say $J\{P(\mathbf{y})\}$ - now reads,

$$J\{P(\mathbf{y})\} = \int_{R^n} \left(P(\mathbf{y}) \|\mathbf{y}\|^2 + \lambda_1 P(\mathbf{y}) + \lambda_2 P(\mathbf{y}) \ln P(\mathbf{y}) \right) d\mathbf{y} \tag{11}$$

The Euler-Lagrange equation is then,

$$\|\mathbf{y}\|^2 + \lambda_1 + \lambda_2 (\ln P(\mathbf{y}) + 1) = 0 \tag{12}$$

which gives together with the normalization condition,

$$P(\mathbf{y}) = \frac{1}{\sqrt{\pi \lambda_2}}\, e^{-\|\mathbf{y}\|^2/\lambda_2} \tag{13}$$

Therefore the circular gaussian distribution is optimal and the training will tend to transform the input distribution into a gaussian distribution. This will be useful for the density estimation addressed in chapter five.

4 Optimizing a parametrized symplectic map

As we pointed out at the end of chapter two, we use a connectionist network structure that has shown to perform a general function approximation in order to have a flexible expression of the scalar function $S(\mathbf{z})$ in (4). Consider now the specific class of symplectic maps that results from using a conventional network structure for the scalar function $S(\mathbf{z})$ in (4),

$$\mathbf{y} = \mathbf{x} + J \frac{\partial}{\partial \mathbf{z}} S\left(\frac{\mathbf{x} + \mathbf{y}}{2}, \mathbf{w}, W \right) \tag{14}$$

$$S(\mathbf{z}, \mathbf{w}, W) = \mathbf{w} \cdot g(W\mathbf{z}) \tag{15}$$

Note that if $S(\mathbf{z})$ is a quadratic function, (14) represents a linear transformation. We parametrize S($\mathbf{z}$, $\mathbf{w}$, W) by the network parameters $\mathbf{w} \in R^m$ and

$W \in R^m \times R^n$. For J to be well defined in (4), has n to be an even number. This is not a severe restriction, because one may add a dummy coordinate to the problem if necessary. The function g applies to each component of the vector $W\mathbf{z}$ and may be chosen as a sigmoidal shaped function. In this case, the function (15) realizes a 3-layer perceptron.

There are different ways to solve (14) numerically. The fastest algorithm would be a Newton method for finding the zeros of $F(\mathbf{y}) = -\mathbf{y} + \mathbf{x} + J\partial S((\mathbf{x} + \mathbf{y})/2)/\partial \mathbf{z}$. The other, rather straight forward method is to regard equations (14) as a fixed-point iteration:

$$\mathbf{y}(t+1) = \mathbf{x} + J\frac{\partial}{\partial \mathbf{z}} S\left(\frac{\mathbf{x} + \mathbf{y}(t)}{2}, \mathbf{w}, W\right) \tag{16}$$

And finally we could use gradient methods such as conjugate gradient for minimizing $\|F(\mathbf{y})\|^2$. All these methods should start at $\mathbf{y}(0) = \mathbf{x}$ since the solutions will then yield a mapping "nearest" to the identity map. The metric defining the term "near" is determined by (15) but is hard to analyze. The iteration will succeed, if the starting point lies in the domain of attraction of the corresponding algorithm. In general this domain of attraction will be smaller with increasing dimension and with growing $\|\mathbf{w}\|$ and if g is a monotone increasing function, then also with growing $\|W\|$.

In the simulations fixed-point iteration and conjugate gradient has shown to be quite competitive with each other, while Newton methods exhibit poor stability. Fixed-point iteration is fast, since it converged mostly within 30 iterations, but failed to converge as the learning process increases the norms of the connection weights. Note that a necessary, local condition for fixed-point iteration to converge is $\|\partial \mathbf{y}(t+1)/\partial \mathbf{y}(t)\| \leq 1$. Conjugate gradient converges mostly within 10 steps but not necessarily to the desired global minimum solution. A more stable, globally convergent technique is the homotopy-continuation method (Stoer & Bulirsch, 1993) where starting from a known solution one gradually modifies the nonlinear equation while successively finding the new solutions. Usually one starts at the identity map and introduces the non-linearity gradually by a parameter α. In our case we have to find the solution of

$$\left\| -\mathbf{y} + \mathbf{x} + \alpha J\frac{\partial}{\partial \mathbf{z}} S\left(\frac{\mathbf{x} + \mathbf{y}(t)}{2}, \mathbf{w}, W\right)\right\|^2 = 0 \tag{17}$$

for a given α. During the homotopy-continuation we gradually increase $\alpha = 0 \rightarrow 1$ with step size $\Delta\alpha$, while using the solutions of the recent relaxation $\mathbf{y}(\infty, \alpha)$ as a the starting point of the next minimization $\mathbf{y}(0, \alpha + \Delta\alpha)$:

$$\mathbf{y}(0, \alpha + \Delta\alpha) = \mathbf{y}(\infty, \alpha) \tag{18}$$

This method may even accelerate the search, since for each step of increasing the parameter the new solution of (17) can be found in a few steps. On the other hand the gradual increase of the non-linearity helps to start always in

the domain of attraction of the global solution. But still, there is no guarantee for convergence. One may pass a bifurcation point and follow the path of a degenerating solution i.e. a solution that evolves to a relative minimum. For the training patterns we have a very natural way of staying in the basin of attraction. We keep track of the solutions of (17) while we modify the parameters that we incrementally change according our learning rule, i.e. we start the search for the solutions of (17) with new weight parameters $(\mathbf{w} + \Delta\mathbf{w}, W + \Delta W)$at the previous solution $\mathbf{y}(\infty, \mathbf{w}, W)$:

$$\mathbf{y}(0, \mathbf{w} + \Delta\mathbf{w}, W + \Delta W) = \mathbf{y}(\infty, \mathbf{w}, W) \tag{19}$$

This homotopy was used for the training, since we have natural access to the weight path, while the first homotopy is used for new data points not belonging to the original training set.

Corresponding to the cost function (9) we need the derivatives of the output, with respect to a parameter p in order to perform gradient descent:

$$\begin{aligned} \Delta p &= -\epsilon \tfrac{\partial}{\partial p} MI = -\tfrac{\epsilon}{2} \tfrac{\partial}{\partial p} \left\langle \| \mathbf{y} - \langle \mathbf{y} \rangle \|^2 \right\rangle \\ &= -\epsilon \left(\left\langle \mathbf{y} \cdot \tfrac{\partial \mathbf{y}}{\partial p} \right\rangle - \langle \mathbf{y} \rangle \cdot \left\langle \tfrac{\partial \mathbf{y}}{\partial p} \right\rangle \right) \end{aligned} \tag{20}$$

The parameter p represents one of the components of $\mathbf{w}$ or W. As usual, a step size ϵ was introduced. The derivatives are given by a linear equation:

$$\frac{\partial \mathbf{y}}{\partial p} = -J^{-1} \left(\frac{\partial^2 S}{\partial \mathbf{z} \partial \mathbf{z}} \bigg|_{\frac{\mathbf{x}+\mathbf{y}}{2}} \frac{\partial \mathbf{y}}{\partial p} + \frac{\partial^2 S}{\partial \mathbf{z} \partial p} \bigg|_{\frac{\mathbf{x}+\mathbf{y}}{2}} \right) \tag{21}$$

where $\partial^2 S / \partial \mathbf{z}^2$ is the Hessian matrix of $S(\mathbf{z})$ at $\mathbf{z} = (\mathbf{x} + \mathbf{y})/2$. This linear equation yields the derivatives at fixed $\mathbf{x}$ and $\mathbf{y}$ corresponding to (14).

The computational cost of finding a solution of (14) is $O(nmt)$ were t represents the number of steps the equation has to be iterated until convergence (in practice $10 < t < 200$). The computational complexity of solving (21) for all $n^2 + m$ parameters and performing the products in (20) is $O((n^2 + m)(1 + n + n^2/2) + n^3)$. One may think of approximations of (20) in order to update the parameters after each of N data points, but we performed the exact batch gradient descent, which implies an additional factor N in the computational cost of each gradient update step.

5 Density estimation and novelty detection

The proposed paradigm implements Barlow's redundancy reduction principle for unsupervised feature extraction. The resulting factorial representation of the joint probability distribution presumably facilitates density estimation and

will be applied now especially to novelty detection. If one knows that a joint distribution factorizes, then the problem of finding a estimation of the joint probability in an -dimensional space reduces to the task of finding the one-dimensional probability distributions . Once we found the estimate we can always calculate the corresponding distribution in the input space, since the symplectic map is invertible and differentiable. Corresponding to Papoulis (1991)

$$P(\mathbf{x}) = \left. \frac{P(\mathbf{y})}{det\left(\frac{\partial \mathbf{f}^{-1}}{\partial \mathbf{y}}\right)} \right|_{\mathbf{y}=\mathbf{f}(\mathbf{x})} = \left. \frac{\prod_i^n P(y_i)}{det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)} \right|_{\mathbf{y}=\mathbf{f}(\mathbf{x})} \tag{22}$$

The Jacobian of the inverse map is given by linear equations analogous to (21). But more likely than calculating the original distribution, in most applications one would map the entire problem to the new factorial space.

As we stated before the gaussian upper bound cost function favors at the output gaussian distributions with equal variance, provided that the symplectic map is general enough to transform the given distribution. Although not perfect, figure 1 demonstrates that ability.

If the training succeeds, we might estimate the distributions by the straight forward assumption of independent gaussian distributions at the output:

$$P(\mathbf{x}) = \prod_i^n \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(y_i - \langle y_i \rangle)^2}{2\sigma_i^2}\right) \tag{23}$$

A possible limitation of symplectic maps is continuity. In figure 2 the two '+' clusters of the training distribution can't be mapped into a unimodal gaussian distribution without having a discontinuity in the map. But even in that case, one is still free to choose a more suitable one-dimensional estimation technique, than the simple approximation by a single gaussian distribution.

Finally, we want to address the closely related task of novelty detection. Given a set of samples corresponding to *a priory* distribution one has to decide if a new sample correspond to that distribution or not. Therefore, the question is: "how probable is an observed new sample according to what we have seen so far?". Given a certain decision threshold, we would like to know the corresponding contour of the density of the data point previously seen. If the contour for an arbitrary threshold is required, we need the complete estimation of the density. As a solution to that problem we propose the presented symplectic redundancy reduction with the straightforward gaussian density estimation (23). The decision surface for the novelty detection is then just a hypersphere in the output of the symplectic map after reducing the redundancy according to the given sample set.

Figure 2 demonstrates that idea. The symplectic map was trained to reduce redundancy on the samples '+'. The samples 'o' are to be discriminated. It is obvious that taking a circular decision measure at the output distribution

Figure 1: Top: Nonlinear correlated and non gaussian joint input distribution. Bottom: After training the symplectic map the distribution is transformed into almost independent normal distributions. Cost function was reduced by 68% within 300 training steps. The "network" contained 6 parameters.

Figure 2: '+' training samples. 'o' test samples. Left: Input signals; Right: Output signals of the trained symplectic map. The symplectic map partially transforms dual modal training distribution into unimodal distribution. The map needs 6 parameters. Perfect transformation would require a singularity of the map between the two spots and is therefore not possible with a symplectic map. Samples not corresponding to the training set, are mapped far away from the gap area, or remain not substantially changed within regions where no training samples are found. Ellipse and circle indicates possible classification boundaries for the plus samples.

will give a fair solution. Although for this illustrative example we could obtain also good results with a simple gaussian mixture (Duda & Hart, 1973) of two gaussian spots, we want to show the performance of the proposed technique. The usual performance measure for a binary classification problem is the rate of "miss classifications" and the rate of "false alarms" as shown in figure 3. Miss classification occurs when a sample that belongs to the "novel" set is missed and classified as normal. The opposite case where a "normal" sample is classified as "novel" is called a false alarm (see figure 3, left).

Novelty detections is a explicitly unsupervised task. It can also be understood as a two class decision making, where only one class is known. During training (i.e. finding the right decision surface) one is not allowed to use the distribution of "novel" samples, because in practise this distribution is unknown to us. If somebody would give us that knowledge, we could use a plenty of known supervised classification techniques. But still, during the design phase of any novelty detection algorithm we need to have the distribution of the "novel" samples in order to measure its performance and to compare it with others.

If the system performs well, after removing the nonlinear statistical dependency of the training distribution, and being left with only gaussian distributions, there is nothing we can do further unsupervised. At that point we might have a perfect representation of what is known. To go further we need to have some knowledge about the "novel" distribution. That knowledge might be implicit, e.g. a performance curve like the one in figure 3 which tell us, that for the given "normal" / "novel" distributions a given algorithm work better than an other. In that example the knowledge consists in the information, which of the extracted independent gaussian features separates the both distributions better (as demonstrated in figure 3, right). There one coordinate (x-axis) of the feature space basically represents noise to the decision we want to take, while the important information in order to separate the distributions is in the other coordinate (y-axis). This illustrative example demonstrate how noise and useful information can only be recognized as such, if we have some kind of supervision (performance of a decision algorithm).

This justifies again our approach of preserving the information by a map, if it is trained completely unsupervised. Unsupervised training without any other knowledge should not delete information.

6 Conclusions

We proposed a technique for finding a nonlinear transformation that produces a statistically independent representation of a probability distribution. This feature extraction in terms of minimal mutual information generalizes the decorrelation criterion of linear PCA to higher order but single coordinate statistics. In order to obtain a computationally feasible criteria we restricted ourselves in this paper to second order statistic. Future work will extend this paradigm to

Figure 3: Rate of miss classification and false alarm for the example of figure 2. The '+' training set is regarded as "normal" and the 'o' test samples are to be classified as "novel". We used in both cases (input and output) a elliptical distance measure as decision criteria for novelty, i.e. we classify as "normal" all points laying within an elliptical area around the center of the "normal" training set. All others are classified as "novel". The decreasing curve gives the false alarm rate, while the increasing curve denote the rate of missing the "novel" data points. The crossing point of this to error rates defines the minimum error rate, which gives a performance measure for the quality of the algorithm. Left: Solid and dashed lines correspond to the decision ellipse and circle in figure 2 respectively. Therefore the solid lines show the performance in the input space and the dashed lines the performance at the output of the symplectic transformation. Obviously minimum error rate is much smaller for the second. Right: Consider again figure 2. Now the right spot of the 'o' test set was not considered for the classification. The training set remained the same. Now solid lines correspond to the elliptical measure in both directions. Dashed lines correspond to measure in the y-axis direction only. In this way knowledge about which coordinates contains the desired information for classification can be obtained.

higher order single coordinate statistics. In comparison with other approaches we believe that the volume conserving nonlinear transformations are the information theoretic most meaningful class of nonlinear transformations that should be considered in statistically independent feature extraction. The approach of implicit symplectic maps represents a large class of such volume conserving nonlinear transforms that may be able to capture the nonlinear submanifold that underlies the environmental data. For these two reasons explicit volume conserving maps should be favored against other nonlinear approaches, were the transmission of information is guaranteed by heuristic penalty terms. We believe that almost like PCA this technique will have numerous applications in various machine learning fields. One possible application of density estimation and novelty detection has been outlined in this paper, and demonstrated again the importance of information preservation.

Reference

Abraham, R., & Marsden, J. (1978). *Foundations of Mechanics*. The Benjamin-Cummings Publishing Company, Inc., London.

Atick, J., & Redlich, A. (1990). Towards a theory of early visual processing. *Neural Computation*, *2*, 308–320.

Atick, J., & Redlich, A. (1992). What Does the Retina Know about Natural Scenes. *Neural Computation*, *4*, 196–210.

Baldi, P., & Hornik, K. (1989). Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima. *Neural Networks*, *2*, 53–58.

Barlow, H. (1959). Sensory Mechanism, the Reduction of Redundancy, and Intelligence. In *National Physical Laboratory Symposium*, Vol. 10. Her Majesty's Stationery Office, London. The Mechanization of Thought Processes.

Barlow, H. (1989). Unsupervised Learning. *Neural Computation*, *1*, 295–311.

Bourlard, H., & Kamp, Y. (1988). Auto-Association by Multilayer Perceptron and Singular Value Decomposition.. *Biological Cybernetics*, *58*, 291–294.

Burel, G. (1992). Blind Separation of Sources: A Nonlinear Neural Algorithm. *Neural Networks*, *5*, 937–947.

Comon, P. (1994). Independent component analysis, A new concept. *Signal Processing*, *36*, 287–314.

Comon, P., Jutten, C., & Herault, J. (1991). Blind separation of sources, Part II: Problem statement. *Signal Processing*, *24*, 11–20.

Deco, G., & Brauer, W. (1994). Higher Order Statistical Decorrelation by Volume Concerving Nonlinear Maps. *Neural Networks*, *?* submitted.

Deco, G., & Parra, L. (1994). Nonlinear Features Extraction by Redundancy Reduction with Stochastic Neural Networks. *Biological Cybernetics*, *?* submitted.

Deco, G., & Schürman, B. (1994). Learning Time Series Evolution by Unsupervised Extraction of Correlations. *Physical Review E*, *?* submitted.

Duda, R., & Hart, P. (1973). *Pattern Classification and Scene Analysis*. Wiley - Interscience.

Feng, K., & Qin, M.-z. (1985). The Symplectic Methods for the Computation of Hamiltonian Equations. In Zhu You-lan, G. B.-y. (Ed.), *Numerical Methods for Partial Differential Equations*. Proceedings of a Conference held in Shanghai, 1987. Lecture Notes in Mathematics. Vol. 1297, pp. 1-35. Springer, Berlin Heidelberg New York.

Földiák, P. (1989). Adaptive network for optimal linear feature extraction. In *IEEE/INNS International Joint Conference on Neural Networks*, pp. 401–405.

Hastie, T., & Stuetzle, W. (1989). Principal Curves. *Journal of the American Statistical Association*, *84*(406), 502–516.

Hornik, K., Stinchcombe, M., & White, H. (1998). Multilayer Feedforward Neural Networks are Universal Approximators. *Neural Networks*, *2*, 359–366.

Karhunen, J., & Joutsensalo, J. (1994). Representation and Separation of Signal Using Nonlinear PCA Type Learning. *Neural Network*, *7*(1), 113–127.

Kuehnel, H., & Tavan, P. (1990). The anti-Hebb Rule derived from Information Theory. In R. Eckmiller, G. H., & Hauske, G. (Eds.), *Parallel processing in neural systems and computers*, pp. 187–190. North-Holland: Elselvier Science, Amsterdam.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*, 105–.

Miesbach, S., & Pesch, H. (1992). Symplectic phase flow approximation for the numerical integration of canonical systems. *Numerical Mathematics*, *61*, 501–521.

Obradovic, D., & Deco, G. (1993). Generalized Linear Features Extraction: An Information Theory Approach. *Neural Computation*, *?*, ? submitted.

Oja, E. (1989). Neural Networks, Principal Components, and Subspaces. *International Journal of Neural Systems*, *1*(1), 61–68.

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes.* Third Edition, McGraw-Hill, New York.

Redlich, A. (1993a). Redundancy Reduction as a Strategy for Unsupervised Learning. *Neural Computation*, *5*, 289–304.

Redlich, A. (1993b). Supervised Factorial Learning. *Neural Computation*, *5*, 750–766.

Rubner, J., & Tavan, P. (1989). A Self-Organization Network for Principal-Component Analysis. *Europhysics Letters*, *10*, 693–698.

Shannon (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Siegel, C. (1943). Symplectic Geometry. *American Journal Mathematics*, *65*, 1–86.

Stoer, J., & Bulirsch, R. (1993). *Introduction to Numerical Analysis.* Springer, New York.

Taylor, J., & Coombes, S. (1993). Learning Higher Order Correlations. *Neural Networks*, *6*, 423–427.