Statistical Independence and Novelty Detection with Information Preserving Nonlinear Maps

Lucas Parra, Gustavo Deco, Stefan Miesbach

Siemens AG, Corporate Research and Development, ZFE ST SN 41

Otto-Hahn-Ring 6, 81739 Munich, Germany

Abstract. According to Barlow (1989), feature extraction can be understood as finding a statistically independent representation of the probability distribution underlying the measured signals. The search for a statistically independent representation can be formulated by the criterion of minimal mutual information, which reduces to decorrelation in the case of Gaussian distributions. If non-Gaussian distributions are to be considered, minimal mutual information is the appropriate generalization of decorrelation as used in linear Principal Component Analyses (PCA). We also generalize to nonlinear transformations by only demanding perfect transmission of information. This leads to a general class of nonlinear transformations, namely symplectic maps. Conservation of information allows us to consider only the statistics of single coordinate. The resulting factorial representation of the joint probability distribution gives a density estimation. We apply this concept to the real world problem of electrical motor fault detection treated as a novelty detection task.

1 Information Preserving Nonlinear Maps

Unless one has *a priori* knowledge about the environment, *i.e.* the distribution of the input signals, it can be difficult to find criteria for separating noise from useful information. To extract structure from the signals, one applies statistical decorrelating transformations to the input variables. In order to avoid a loss of information, these transformations have to preserve entropy. According to Shannon (1948) entropy is defined as $H(\boldsymbol{x}) = -\int P(\boldsymbol{x}) \ln P(\boldsymbol{x})\, d\boldsymbol{x}$ of a continuous distribution $P(\boldsymbol{x})$, with $\boldsymbol{x} \in R^n$. Continuous entropy is sensitive to scaling. Scaling coordinates changes the amount of information (or entropy) of a distribution. More general, for an arbitrary mapping on R^n: $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ condition $det(\partial \boldsymbol{f} / \partial \boldsymbol{x}) = 1$ yields $H(\boldsymbol{y}) = H(\boldsymbol{x})$ (Papoulis 1991), *i.e.* local conservation of volume guarantees constant entropy from the input $\boldsymbol{x}$ to the output $\boldsymbol{y}$. To avoid spurious information generated by a transformation, we consider therefore volume-conserving maps, *i.e.* those with unit Jacoby determinant.

The goal of this paper is to present a special neural-network like structure for building volume preserving transformations. Two approaches may be used to achieve this goal. First, one may prestructure the neural network in such a way that volume preservation is guaranteed independent of the network weights (Deco and Brauer, 1995; Deco and Schürmann, 1995). Alternatively, weight constraints may be used to restrict the learning algorithms to volume conserving network solutions.

In this paper we present a new prestructuring technique which is based on symplectic geometry in even-dimensional spaces ($n = 2m$). The core of symplectic geometry is the idea that certain “area elements” are the analogue of “length” in standard Euclidean geometry (Siegel, 1943). Transformations which preserve these area elements are referred to as symplectic. Symplectic transforms do also preserve volume. However, the converse is not true, i.e. volume preservation is not sufficient for symplecticity.

The advantage of symplectic transforms is the fact that they can be parametrized by arbitrary scalar functions $S(\boldsymbol{z})$; $\boldsymbol{z} \in R^{2m}$ due to the implicit representation[1]

$$\boldsymbol{y} = \boldsymbol{x} + \mathbf{J}\frac{\partial}{\partial \boldsymbol{z}} S\left(\frac{\boldsymbol{x}+\boldsymbol{y}}{2}\right); \qquad \mathbf{J} = \begin{bmatrix} 0 & -\mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix} = -\mathbf{J}^{-1} \tag{1.1}$$

where the $\mathbf{I}$ denotes the n-dimensional identity matrix, and $\boldsymbol{x}, \boldsymbol{y} \in R^{2m}$. Any non-reflecting symplectic transform ($det\left(\boldsymbol{I} - \frac{d\boldsymbol{f}}{d\boldsymbol{x}}\right) \neq 0$) can be generated by an appropriate function S, and also the converse is true: Any twice differentiable scalar function, e.g. an arbitrary standard neural network, leads to a symplectic transform in (1.1).

Consequently, to obtain a set of symplectic transforms that is as general as possible, we use a one-hidden-layer neural network NN as a general function approximator (Hornik et. al., 1989) for the generating function S:

$$S(\boldsymbol{z}) = NN(\boldsymbol{z}, \boldsymbol{w}, W) = \boldsymbol{w} \cdot g(W\boldsymbol{z}), \tag{1.2}$$

where W denotes the input-hidden weight matrix, $\boldsymbol{w}$ the hidden-output weights and g the activation function. Eq. (1.1) has to be solved numerically. We use either fixed-point iteration or a homotopy-continuation method (Stoer and Bulirsch 1993).

2 Mutual Information and Statistical Independence

The components of a multidimensional random variable $\boldsymbol{y} \in R^n$ are said to be statistical independent if the joint probability distribution $P(\boldsymbol{y})$ factorizes, i.e. $P(\boldsymbol{y}) = \prod_i^n P(y_i)$. Here, $P(y_i)$ represents the distribution of the individual coordinates y_i, $i = 1, \ldots, n$ of the random variable $\boldsymbol{y}$. Statistical independence can be measured in terms of the mutual information $MI(\boldsymbol{y})$,

1. This representation of symplectic maps is a special case of the generating function theory developed in full generality by Feng Kang, Qing Meng-zhao (1987). A proof of the representation (1.1) and a discussion of its role for Hamiltonian Systems can be found in Abraham and Marsden (1978) and Miesbach (1992).

$$0 \le MI(\boldsymbol{y}) = -H(\boldsymbol{y}) + \sum_{i}^{n} H(y_i) \,, \tag{2.1}$$

Zero mutual information indicates statistical independence. Here, $H(y_i) = -\int P(y_i) \ln P(y_i)\, dy_i$ denotes the single coordinate entropies.

In the case of Gaussian distributions, linear decorrelation, i.e. diagonalizing the correlation matrix of the output $\boldsymbol{y}$, has been proven to be equivalent to minimizing the mutual information (Papoulis 1991) and corresponds to the standard Principal Component Analysis (PCA) method. However, for general distributions, decorrelation does not imply statistical independence of the coordinates.

Starting from the principle of minimum mutual information, Deco and Brauer (1994) formulated criteria for decorrelation by means of higher orders cumulants. A similar approach, that considers the distance to the Gaussian distribution (standardized mutual information) but restricts itself to linear transformations, was studied by Comon (1994). Redlich (1993) suggested the use of reversible cell automata in the context of nonlinear statistical independence. Instead of preserving information the invertibility of the map was considered. While invertibility indeed assures constant information when dealing with discrete variables, for continuous variables, conservation of volume is necessary.

In the case of binary outputs, maximum mutual information has been proposed instead (Schmidthuber, 1992, Deco and Parra, 1994). In the context of the blind separation problem, Bell and Sejnowski (1994) proposed a technique for the separation of continuous output coordinates with a single layer perceptron. But the authors admit, that the information maximization criterion they use, does not necessarily lead to a statistical independent representation. In parallel Nadal and Parga (1994) based this idea on a more rigorous discussion.

In this paper, we make use of the more general principle of minimal mutual information (statistical independence) instead of the decorrelation used in PCA.

For the symplectic map, the identity $H(\boldsymbol{x}) = H(\boldsymbol{y})$ holds, and therefore we are left with the task of minimizing the sum of the single coordinate entropies (second term in the left-hand side of (2.1)). Since we are given only a set of data points, drawn according to the output distributions, this is still a difficult task. But fortunately, there is a feasible upper bound for these entropies (Parra et.al., 1995),

$$MI(\boldsymbol{y}) \leq -H(\boldsymbol{x}) + \frac{n}{2}\ln(2\pi) + \frac{1}{2}\sum_{i}^{n}\int P(y_i)\,(y_i - \langle y_i \rangle)^2 dy_i \tag{2.2}$$

where $\langle y_i \rangle = \int P(y_i)\, y_i dy_i$. Using only the second order moments for estimating the mutual information might be seen as a strong simplification. At the expense of computational efficiency, higher order cumulants may be included to increase accuracy. An interesting property of eq. (2.2) is that, if the transformation $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ is flexible enough, this cost function will produce Gaussian distributions at the output. Using a variational approach it can be shown that under the constraint of constant entropy a circular Gaussian distribution minimizes the sum of variances in (2.2) (Parra et.al., 1995). This will be useful for the density estimation addressed next. We will observe there some limitations of the continuous volume concerving map in transforming arbitrary distributions into Gaussians.

The training of the network (1.2) can be performed with standard gradient descent techniques. The gradient of the output coordinates with respect to the parameters of the map can be calculated by implicitly differentiating equation (1.1). This leads to a system of linear equations for the gradient. The overall computational complexity of the optimization

algorithm is then $O(n^4)$ for each data point. This restricts this approach to a low dimensional space (in practice $n \leq 30$).

3 Density estimation and novelty detection

If one knows that a joint distribution factorizes, then the problem of finding an estimation of the joint probability $P(\boldsymbol{x})$ in an n-dimensional space is reduced to the task of finding the one-dimensional probability distributions $P(y_i)$. As stated before, the Gaussian upper bound cost function favors Gaussian distributions at the output, provided that the symplectic map is general enough to transform the given distribution. Figure 1 demonstrates this ability.

If the training succeeds, we might estimate the distributions by the straightforward assumption of independent Gaussian distributions at the output:

$$P(\boldsymbol{y}) \approx \prod_i^n \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(\frac{-(y_i - \langle y_i \rangle)^2}{2\sigma_i^2} \right) \tag{3.1}$$

Estimation reduces then to the measurement of the output variances σ_i^2.

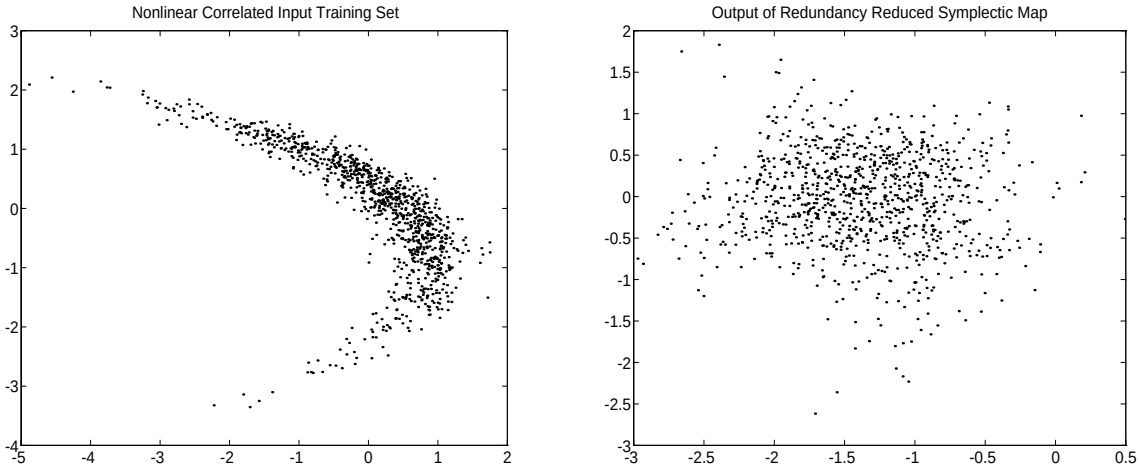


FIGURE 1. Nonlinear correlated and non Gaussian joint input distribution (left) is transformed into almost independent normal distributions (right). The input distribution was generated by mapping a one dimensional exponential distribution with additive Gaussian noise onto a circle.

The cost function was reduced by 68% in 300 training steps. The "network" contained 6 parameters ($w \in R^2$ **and** $W \in R^2 \times R^2$ **).**

We now address the closely related task of novelty detection. Given a set of samples corresponding to a prior distribution, one has to decide whether or not a new sample corresponds to this distribution. Putting it into other words the question is: "How probable is an observed new sample according to what we have seen so far?" Given a certain decision threshold, novelty detection is based on the corresponding contour of the density of the data points previously seen. If the contour is required for an arbitrary threshold, we need the complete estimation of the density. As a solution to this problem we propose the presented symplectic factorization with the a posterior Gaussian density estimation (3.1). The decision surface for the novelty detection is then just a hypersphere in the output of the symplectic map after reducing the mutual information according to the given sample set.

Figure 2 demonstrates this idea. The symplectic map was trained to reduce mutual information on the samples '+'. The samples 'o' are to be discriminated. The procedure transforms the output distribution to a Gaussian distribution as closely as possible, in order to use a circular contour of the density as a decision boundary. As a side effect, volume conservancy tends to separate regions not belonging to the training set from those corresponding to it. The former regions are mapped far away from the gap area. Obviously, taking a circular decision measure at the output distribution will give a fair solution. We show the performance of the proposed technique in figure 3 (left) by showing the standard graph of misclassification and false-alarm rates. For this illustrative example we could also obtain good results with a simple Gaussian mixture (Duda & Hart 1973) of two Gaussian spots.

This example also demonstrates one of the possible limitations of the technique as a general density estimation procedure. Perfect transformation into a single Gaussian spot requires a singularity in order to map the two spots arbitrary close together. Because of the property of local conservation of volume, vanishing distance in one direction implies unbounded stretching in the orthogonal direction, which will not be possible with a continuous map. More generally speaking, the combination of a continuous and volume conserving map together with a unimodal distribution is best suited for distributions spread over connected regions rather than for disjoint distributions. For the novelty detection this behavior is clearly an advantage since it separates known distributions from unknown regions.

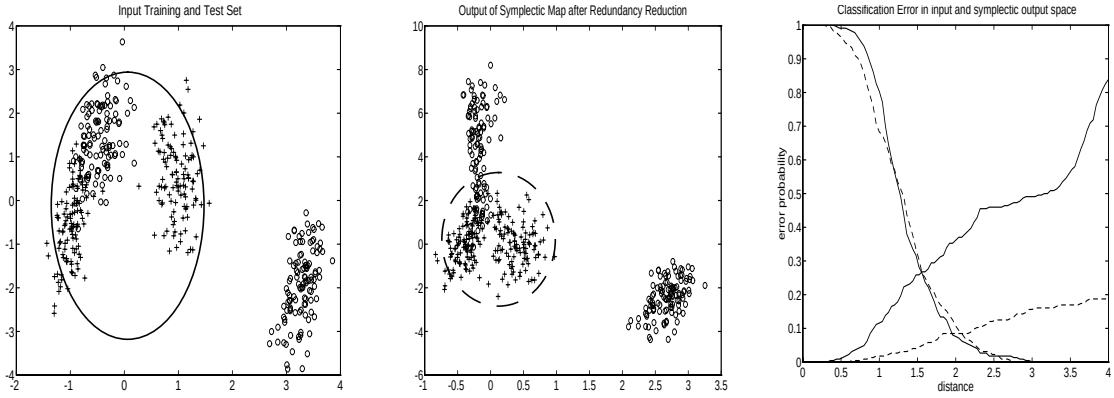


FIGURE 2. '+' training samples. 'o' test samples. Left: Input signals; Center: Output signals of the trained symplectic map. The symplectic map partially transforms a bimodal training distribution into a unimodal distribution. The map used again 6 parameters. Ellipses indicate possible classification boundaries for the '+' samples. Right: Rate of misclassification and false-alarm. We used in both cases (input and output) an 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 curves gives the false-alarm rate, while the increasing curves denote the rate of missing the "novel" data points.

4 Motor fault detection

In this section, we show that the proposed concept of novelty detection provides encouraging results in a high dimensional real world problem. In motor fault detection, the task

consists of noting early irregularities in electrical motors by monitoring the electrical current. The spectrum of the current is used as a feature vector. The motor failure detector is trained with data supplied by a healthy motor and should notify, if the motor is going to fail.

Typically, one deals here with at least 100 and up to 1000 dimensions. Applying the outlined procedure to the complete feature vector is not manageable because of the high computational costs of our training procedure. On the other hand, it is hard to believe that 100 or more coordinates are altogether nonlinearly correlated. More likely, we expect most of the coordinates to be (if at all) only linearly correlated. Therefore, we first transform the spectrum with a linear PCA. We use 230 coordinates in the spectrum between 20Hz and 130Hz.

We observed that a few of the first principal components are nonlinearly correlated. No pairwise nonlinear structure could be observed between coordinates others than the first 10 or 15 first principal components. We assume that all other principal components are uncorrelated, unimodal, and symmetrically distributed. They can be fairly well approximated by a normal distribution (see figure 4). We know that for normal distributions, linear decorrelation is the best that can be done in order to minimize mutual information. Therefore, we can assume that these lower principal components are statistical independent. We apply the symplectic factorization only to the first few components. Figure 4 shows how two of the first 10 principal components have been transformed by a 10-20-10 symplectic map trained with 800 samples ($w \in R^{20}$, $W \in R^{20} \times R^{10}$). The net reduced variance by 65% in 650 training steps.

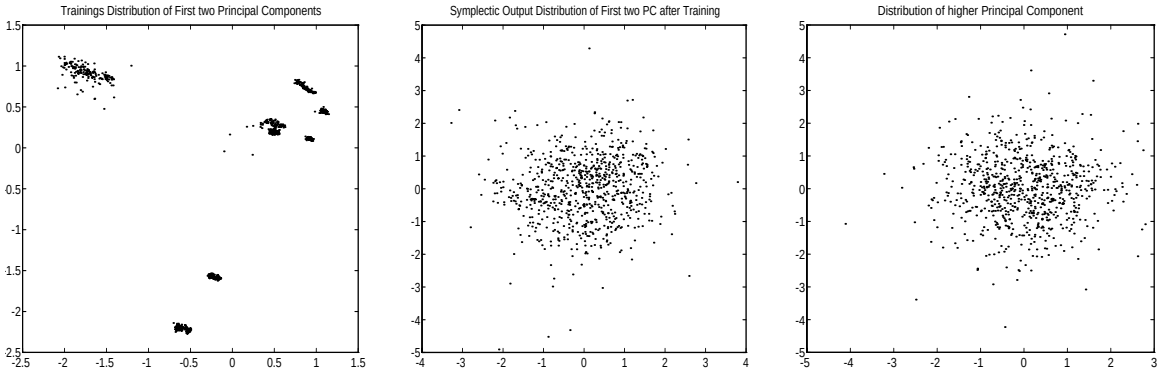


FIGURE 3. Left: Distribution of the 2 first principal components demonstrates a clear nonlinear dependency. Center: resulting distribution of the same first 2 components after reducing the redundancy in the first 10 components. Right: For any other component higher than the 15th no pairwise dependency could be observed. Here, we arbitrarily plot components 50 and 100.

Now we use this result to classify “good” vs. “bad” motors, according to (3.1). Since the performance may vary for different types of faults, we plot the performance curve for the three failure modes occurring in our test data (unbalanced, bearing race hole, and broken rotor bar). In figure 5 we compare the performance with a maximum measure ($max_i\,(y_i - \langle y_i \rangle)$) on the complete 230-dimensional principal component space (left), but use only the Gaussian estimates of the 10 nonlinear transformed coordinates (center). Furthermore, we analyze to what extent a given coordinate separates the ‘good motor’ and ‘bad motor’ distributions by measuring the ratio of the corresponding variances. This analysis reveals that by including the low variance linear normalized PCA coordinates the classification measure can be improved further. With ‘normalized’ we express the fact that we normalize the variance before performing PCA. Best results were obtained by including between 5 and 20 low variance PCA coordinates (see figure 4, right).

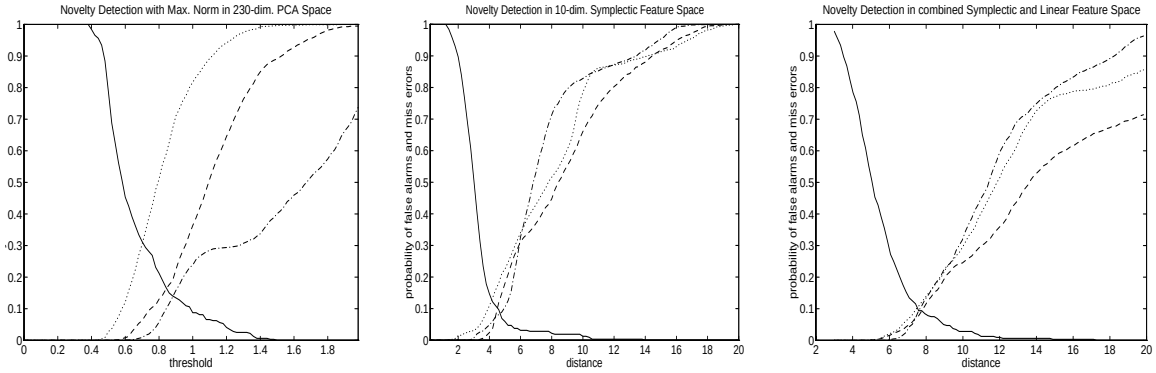


FIGURE 4. left: maximum measure on the 230-dim. principal component space. Center: circular distance measure on the 10 symplectic mapped first linear principal component. Right: combined symplectic and linear features space: 10 symplectic transformed first PC and last 7 PC of the normalized spectrum. The decreasing curves give the false alarm rate. Each of the three increasing curves provides the rate of missing the fault for three different fault situation. (- no fault, .. bearing race hole, -- unbalance, .- broken rotor bar)

One possible measurement of the quality of the classification technique is the decision error at the optimal decision threshold. The proposed technique achieves a decision error of $10 \pm 0.5\,\%$. This result is comparable with different approaches that have been applied to this problem at SCR[1] including, among others, MLP (11%) and RBF (10%) autoassociators, nearest neighbor (18%-32%), and hypersphere (37%) clustering, PCA (12%), or maximum measure (in roughly 2000 dimensions) (11%).

5 Conclusions

The factorization of a joint probability distribution has been formulated as a minimal mutual information criterion under the constrain of volume conservation. Volume conservation has been implemented by a general class of nonlinear transformations - the symplectic maps. A gaussian upper bound leads to a computational efficient optimization technique and favors normal distributions at the output as optimal solutions. This in turn, facilitates density estimation, and can be used particularly for novelty detection. The pro-

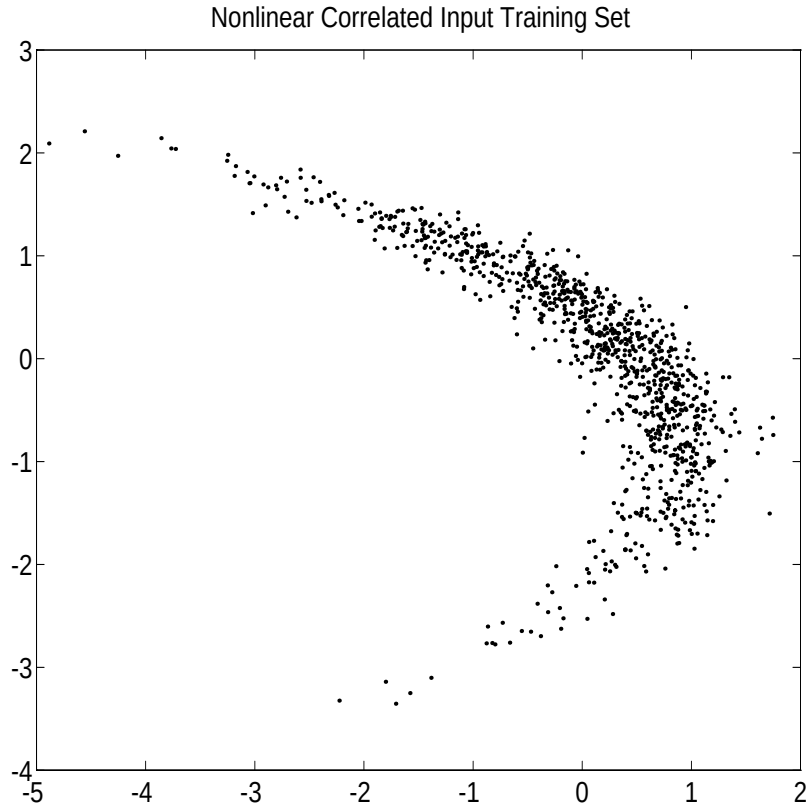
1. Siemens Corporate Research, Inc., 755 College Road East, Princeton, NJ 08540

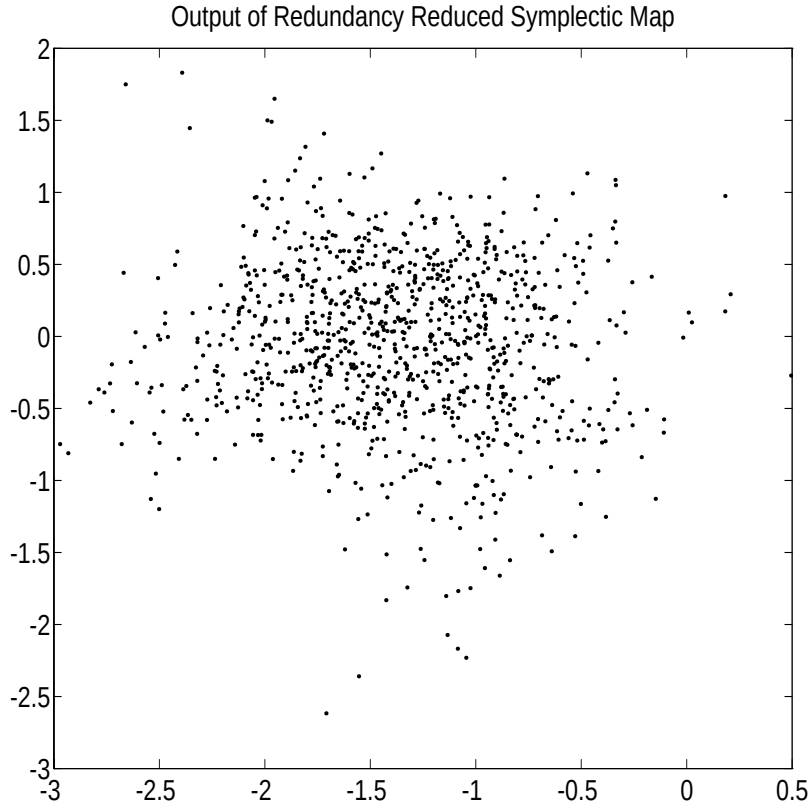
posed technique has been applied successfully to the real world problem of motor fault detection.

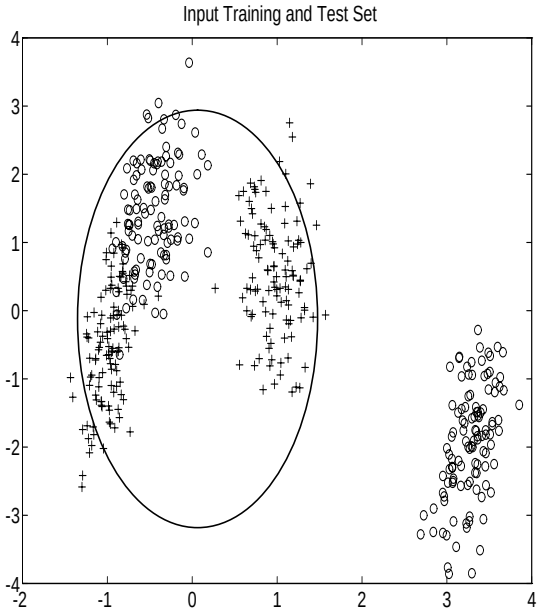
6 References

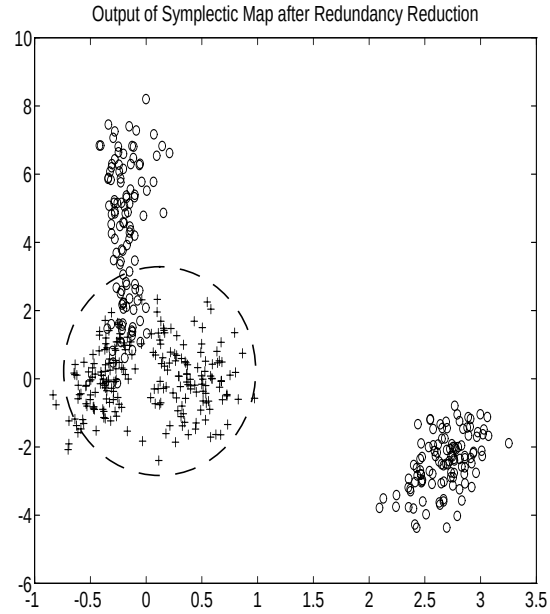
- Abraham R. & Marsden J., 1978, “Foundations of Mechanics,” The Benjamin-Cummings Publishing Company, Inc., London.
- Barlow H., 1989, “Unsupervised Learning,” Neural Computation, **1** (1), 295-311.
- Bell A. J, Sejnowski, T., 1994, “An information-maximization approach to blind separation and blind deconvolution,” submitted to Neural Computations
- Comon P., 1994, “Independent component analysis, a new concept.,” Signal Processing, **36**, 287-314.
- Hornik K., Stinchcombe M., White H., 1989, “Multilayer Feedforward Neural Networks are Universal Approximators,” Neural Networks, **2**, 359-366.
- Deco G., Brauer W., 1994, “Nonlinear Higher Order Statistical Decorrelation by Volume Conserving Neural Architectures,” Neural Networks, in press.
- Deco G. and B. Schürman, “Learning Time series Evolution by Unsupervised Extraction of Correlations”, Physical Review E, **51**, 2 , in press (1995).
- Deco G., Parra L, 1994, “Nonlinear Features Extraction by Redundancy Reduction with Stochastic Neural Networks,” submitted to Biological Cybernetics.
- Duda & Hart, 1973, “Pattern Classification and Scene Analysis”

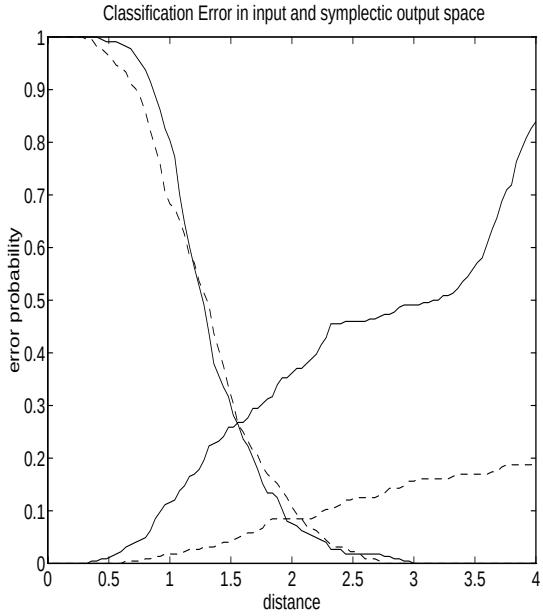
- Feng Kang, Qin Meng-zhao, 1985, “The Symplectic Methods for the Computation of Hamiltonian Equations” In: Zhu You-lan, Guo Ben-yu, eds., “Numerical Methods for Partial Differential Equations,” Proceedings of a Conference held in Shanghai, 1987. Lecture Notes in Mathematics. Vol. 1297, pp. 1-35. Springer, Berlin Heidelberg New York.
- Miesbach S., Pesch H.J., 1992, “Symplectic phase flow approximation for the numerical integration of canonical systems,” Numer. Math. 61, 501-521.
- Nadal J-P., Parga N., 1994 “Non-linear neurons in the low noise limit: a factorial code maximizes information transfer,” to appear in Network
- Papoulis A. 1991, “Probability, Random Variables, and Stochastic Processes,” Third Edition, McGraw-Hill, New York.
- Parra L., Deco G., Miesbach S. 1995, “Redundancy Reduction with Information Preserving Nonlinear Maps,” to appear in Network, **6**, 61-72.
- Redlich A.N., 1993, “Supervised Factorial Learning,” Neural Computation**, 5**, 750-766.
- Schmidhuber, 1992, “Learning Factorial Codes by Predictability Minimization,” Neural Computation, **4**, 6, 863-879.
- Shannon C., 1948, “A mathematical theory of communication,” Bell System Technical Journal, **7**, 379-423.
- Siegel, 1943, “Symplectic Geometry:, Am. J. Math., **65**, 1-86.
- Stoer J., Bulirsch R., 1993, “Introduction to Numerical Analysis,” Springer, New York.

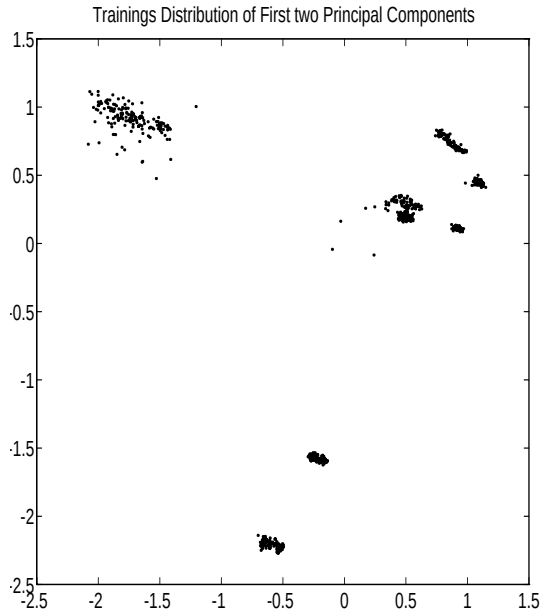

Nonlinear Correlated Input Training Set
3
2
1
0
-1
-2
-3
-4
-5
-4
-3
-2
-1
0
1
2

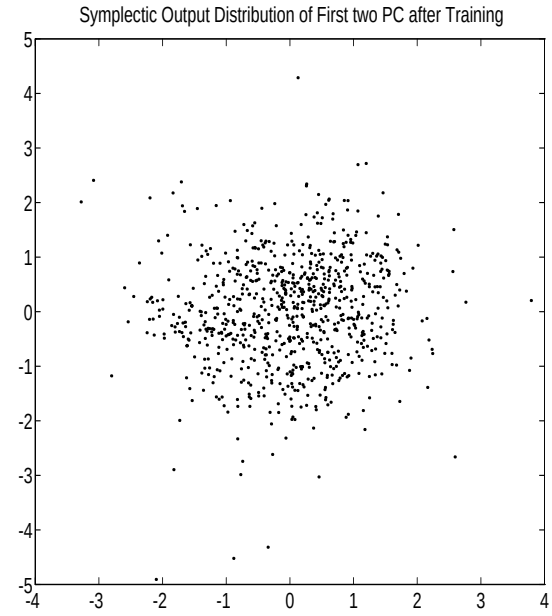

Output of Redundancy Reduced Symplectic Map
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5

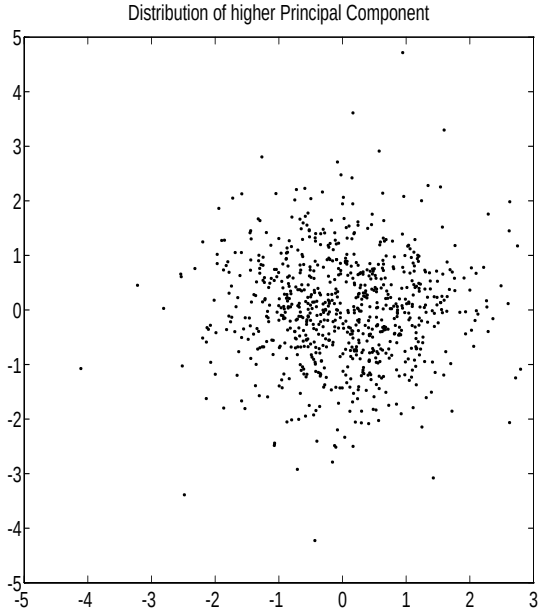

Input Training and Test Set

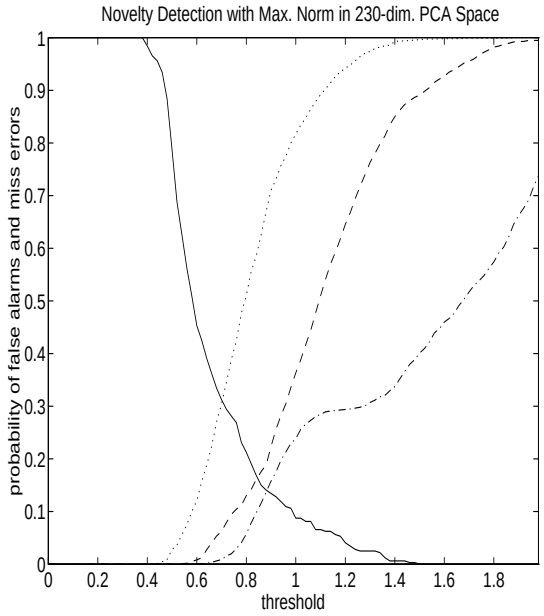

Output of Symplectic Map after Redundancy Reduction

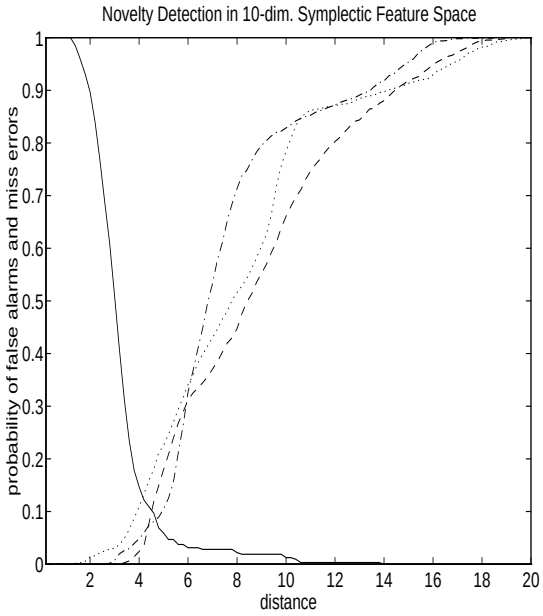

Classification Error in input and symplectic output space
error probability
distance

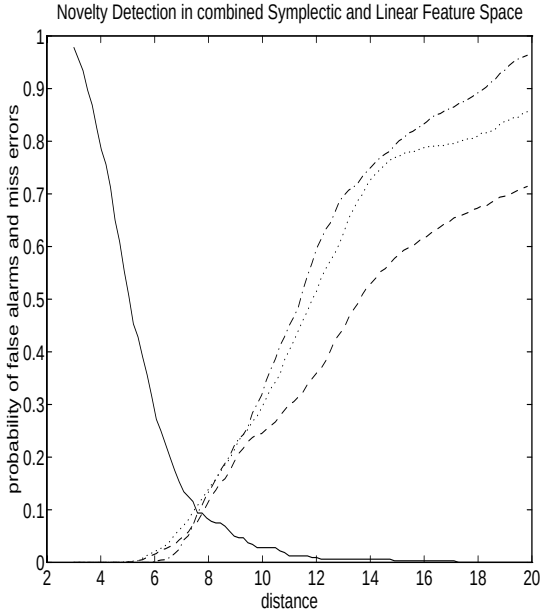

Trainings Distribution of First two Principal Components
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5


Symplectic Output Distribution of First two PC after Training
5
4
3
2
1
0
-1
-2
-3
-4
-5
-4
-3
-2
-1
0
1
2
3
4


Distribution of higher Principal Component
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3


Novelty Detection with Max. Norm in 230-dim. PCA Space
probability of false alarms and miss errors
threshold
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8


Novelty Detection in 10-dim. Symplectic Feature Space
probability of false alarms and miss errors
distance
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
4
6
8
10
12
14
16
18
20


Novelty Detection in combined Symplectic and Linear Feature Space
probability of false alarms and miss errors
distance
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
4
6
8
10
12
14
16
18
20