# Classification-Optimal Feature Transforms

Sergios Petridis*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications,
NCSR "Demokritos"
petridis@iit.demokritos.gr

**Abstract.** Supervised learning of feature vector transforms is a common practice in statistical pattern recognition applications. This study considers optimality criteria for learning both dimensionality reduction and invertible transforms. Starting from the Bayes loss and mutual information, we show the derivation of widely used methods such as linear discriminant analysis and sensitivity analysis. The effective feature vector concept is introduced to extend the applicability of these criteria to invertible transforms. Our approach allows us to derive two novel algorithms for linear feature extraction and non-linear feature scaling respectively.

**Keywords**: linear feature extraction, mutual information maximization, Bayes loss sensitivity, isotropicity, feature scaling

## 1 Introduction

A fundamental issue in machine learning is solving classification problems, in which $d$–dimensional real-valued vectors $\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{T}}$ are mapped to one among $K$ classes $\{\mathcal{C}_k\}$. Ideally, this mapping is summarized by the conditional probability distribution of the class $p\left(\mathcal{C}_k \middle| \mathbf{x}\right)$, which is to be estimated from a limited size sample set. In the context of pattern recognition, $\mathbf{x}$ is commonly referred as *feature vector* since it is assembled from distinct features $x_i$ of the observation to be classified. A common practice is to first preprocess the feature vector by applying optimal transforms of the form

$$\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}, \quad d, d' \in \mathbb{N}^*. \tag{1}$$

The rationale behind this preprocessing is typically (a) to improve problem inspection, by allowing visualization of samples in a space of lower dimension, convenient axis and informative samples distances, (b) to facilitate any following processing steps, by reducing their complexity and memory requirements and (c) to alleviate the initial feature generation step (i.e. the construction of the original feature vector), by indicating what would constitute useless to generate features.

* Dissertation Advisors: Sergios Theodoridis, Professor - Dr. Stavros Perantonis

Although a plethora of algorithms exist to cope with particular computation constraints, finding optimal transforms seems to be driven by more or less heuristic criteria. Here, we bring together results that have been independently studied in the pattern recognition and machine learning literature, to provide a unifying framework for transform learning. This framework allow us to define additional criteria formulations resulting in novel algorithms. Namely, we first introduce a framework for supervised learning criteria comparison for feature extraction, defining an hierarchy of classification problem models where particular goals are feasible. We then study the mutual information maximization criterion comparatively to the criteria of linear discriminant and heteroscedastic discriminant analysis. We end up, on one hand, to the establishment of a criteria hierarchy and, on the other, to compatible mathematical expressions, thus unifying the originally different approaches and deriving linear discriminant analysis based on information theory [7]. In the same framework, we define the Bayes loss sensitivity maximization criterion and show its optimality for linear feature extraction. The criterion leads to the derivation of the SPCA algorithm [5], which comes down to extracting principal components of suitably preprocessed samples. Evaluating the algorithm with a wide range of real datasets demonstrates its advantages against linear discriminant analysis.

The above framework is further extended to account for invertible transforms. To that end, we model the partially known continuous density probability of the feature vector by defining the corresponding effective one. By introducing further assumptions regarding the desired homogeneity and isotropicity of the feature space, we derive the isotropic loss minimization and isotropic information maximization criteria, as the generalization of the corresponding non-isotropic ones. This offers a unified framework for the learning of both linear feature extraction and invertible transforms. Finally, the criterion of isotropic information maximization is explored in view of its application to the learning of non-linear feature scaling transforms. The criterion leads to the formulation of the isotropic non-linear feature scaling algorithm which applies to the non-linear grid transforms [6]. Evaluation of the algorithm to real data shows its utility to the increase of classification generalization performance.

## 2 Optimal Dimensionality Reduction

A variety of feature transforms families exist, from which one may choose to concentrate optimality search, in accordance with ones specific motives and computational constraints. We may organize these families based on three properties: *linearity*, where the transformed feature space is a linear map of the original one, *axis preservation*, where each output feature is a function of only one input feature and *invertibility*, where recovering the input feature vector from the output one is possible. A number of such families and their corresponding properties are shown in Table 1. In this section, we discuss linear transforms resulting in feature vectors of reduced dimension. Invertible transforms, both linear and non-linear, are discussed in Section 3.

| feature vector transforms | | | |
|---|---|---|---|
| transform families | transform properties | | |
| | linearity | axis–preservation | invertibility |
| feature scaling | ✓ | ✓ | ✓ |
| feature selection | ✓ | ✓ | |
| linear feature scaling | ✓ | | ✓ |
| linear feature extraction | ✓ | | |
| non-linear feature scaling | | ✓ | ✓ |
| non-linear feature extraction | | | |

**Table 1.** Feature vector transforms families and properties

### 2.1 Feature vector optimality

A basic issue when seeking for classification-optimal feature vector transforms, is to specify measures against which feature vectors are compared. Decision theory defines, as optimal such measure, the expected loss we risk when attributing a class to a feature vector using the (optimal) Bayes classification rule, referred to as the *Bayes loss*:

$$\mathcal{B}_{\mathrm{L}}[\mathcal{C}\,|\,\mathbf{x}] = \mathbb{E}_{\mathbf{x}}\left[\min_{\mathcal{C}_k} l(\mathcal{C}_k|\mathbf{x})\right] \tag{2}$$

where $l(\mathcal{C}_k|\mathbf{x})$ is the expected loss of mapping point $\mathbf{x}$ to class $\mathcal{C}_k$

$$l(\mathcal{C}_k|\mathbf{x}) = \sum_{k'=1}^{K} L_{kk'}\,p\left(C_{k'}\,|\,\mathbf{x}\right) \tag{3}$$

and $L_{kk'}$ is the $(k,k')$ element of the loss matrix L, defining the loss when misstaking class $C_k$ for class $C_{k'}$. When $\mathrm{L} = 1_K - \mathrm{I}_K$, the Bayes loss is commonly referred to as Bayes error. Roughly, if one could choose between two different feature vectors to solve a classification problem, one should optimally prefer the one that yields lower Bayes loss. We will refer to this as the Minimization of Bayes Loss (MBL) criterion.

In a certain sense, Bayes loss measures our uncertainty about the unknown class value even when the feature vector value is known. Information theory gives an alternative definion of the uncertainty as the feature vector-conditional class entropy:

$$\mathrm{H}[\mathcal{C}\,|\,\mathbf{x}] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{C}}\left[-\log p\left(\mathcal{C}_k\,|\,\mathbf{x}\right)\right]\right]. \tag{4}$$

The above measure, when substracted by the entropy of the class, which is constant for a given classification problem, gives rise to the mutual information between the class and the feature vector:

$$\mathrm{I}[\mathcal{C};\mathbf{x}] = \mathrm{H}[\mathcal{C}] - \mathrm{H}[\mathcal{C}\,|\,\mathbf{x}]. \tag{5}$$

So minimizing $\mathrm{H}[\mathcal{C}\,|\,\mathbf{x}]$ in respect to the feature vector is equivalent to maximize $\mathrm{I}[\mathcal{C};\mathbf{x}]$. We will refer to this as the the Maximization of Mutual Information (MMI) criterion.

Comparison of feature vectors using the MMI criterion is not optimal in the sense of Bayes loss. However, a certain near-optimality is guaranteed since the Bayes loss is bounded by mutual information. Namely, for the Bayes error, it holds that:

$$\frac{\mathrm{H}[\mathcal{C}] - \mathrm{I}[\mathcal{C};\mathbf{x}] - 1}{\log(K-1)} \leq \mathcal{B}_{\mathrm{e}}[\mathcal{C}\,|\,\mathbf{x}] \leq \frac{\mathrm{H}[\mathcal{C}] - \mathrm{I}[\mathcal{C};\mathbf{x}]}{2} \tag{6}$$

### 2.2 Lossy vs Lossless transforms

Bayes loss and mutual information may assist us to clarify what one may expect from dimensionality reduction, depending on ones priorities. Here, we focus on linear transforms, expressed as $\mathrm{A}^{\mathrm{T}}\mathbf{x}$ where $\mathrm{A}_{[d\times d']}$, $d' < d$, is a $d \times d'$ matrix, but this discussion also extends to the non-linear case. When visualization or complexity handling issues are a priority, feature extraction may be constraint by an upper limit $\bar{d}$ of the features to extract. In this case, ideally, one should optimally extract those that yield the lower Bayes loss:

$$\hat{\mathrm{A}} = \operatorname*{argmin}_{\mathrm{A}_{[d\times d']},\, d'\in[1\ldots\bar{d}]} \mathcal{B}_{\mathrm{L}}[\mathcal{C}\,|\,\mathrm{A}^{\mathrm{T}}\mathbf{x}] \tag{7}$$

Notice that there is a set of matrices $\hat{\mathrm{A}}$ satisfying eq. (7), rather than just one, since optimality is related to the feature subspace engendered by the matrix rather than to the matrix itself. It is also important to stress that, in contrast to what one may suggest by considering common experimental resultsreducing the dimension can't lead to a decrease of the Bayes loss, i.e.:

$$\forall \mathrm{A}_{[d\times d']} : \quad \mathcal{B}_{\mathrm{L}}[\mathcal{C}\,|\,\mathrm{A}^{\mathrm{T}}\mathbf{x}] \geq \mathcal{B}_{\mathrm{L}}[\mathcal{C}\,|\,\mathbf{x}], \tag{8}$$

In other words, dimensionality reduction is, in general, lossy. Hence, a more precise interpretation of eq. (7) is *to seek for optimal subspaces where the Bayes loss is increased as less as possible, while satisfying the dimension upper limit requirement.*

On the other hand, when priority is given to the classification performance, one should avoid putting an upper limit for dimension and require instead that dimensionality reduction does not lead to increasing the Bayes loss at all, i.e. $\mathcal{B}_{\mathrm{L}}[\mathcal{C}\,|\,\mathrm{A}^{\mathrm{T}}\mathbf{x}] = \mathcal{B}_{\mathrm{L}}[\mathcal{C}\,|\,\mathbf{x}]$. To examine when and how a reduction in dimension is still possible, let us introduce a generative view, according to which the feature vector $\mathbf{x} \in \mathbb{R}^{d}$ is the result of mixing up a "usefull" *source* vector $\mathbf{s} \in \mathbb{R}^{d'}$ with a "useless" *noise* vector $\mathbf{n} \in \mathbb{R}^{(d-d')}$, by means of an unknown invertible matrix B, as follows:

$$\mathbf{x} = \mathrm{B}^{\mathrm{T}} \begin{bmatrix} \mathbf{s} \\ \mathbf{n} \end{bmatrix}. \tag{9}$$

The feature extraction problem then formulates as recovering the $d'$ *latent* featuresof the source vector, based on specific assumptions between $\mathbf{s}$, $\mathbf{n}$ and $\mathcal{C}$. Two such possible assumptions are the following:

**Zero Information Loss (ZIL)** The noise vector is conditionally independent to the class given the source vector:

$$\mathbf{n} \perp\!\!\!\perp \mathcal{C} \mid \mathbf{s} \tag{10}$$

**Zero Bayes-Loss Loss (ZBLL)** At any point of the feature space, the expected Bayes loss for the optimally chosen class given both the source and noise vectors equals the expected Bayes loss given only the source vector:

$$\forall \mathbf{s} \in \mathbb{R}^{d'}, \mathbf{n} \in \mathbb{R}^{(d-d')} : \quad l(\hat{\mathcal{C}} \,|\, \mathbf{s}, \mathbf{n}) = l(\hat{\mathcal{C}} \,|\, \mathbf{s}) \tag{11}$$

Despite their dissimilar expression, the consequences of these two assumptions differ only in whether one has knowledge of the loss matrix at the time the extraction is done. In particular, it holds that (a) the ZIL assumption is the weakest assumption for extracting $d'$ features without increasing the Bayes loss, with *unknown* loss matrix and (b) the ZBLL assumption is an assumption, weaker than the ZIL one, for extracting $d'$ features without increasing the Bayes loss, for a *known* loss matrix. Importantly, the ZIL and ZBLL assumptions are equivalent for two-class problems.

Let us point out that the MMI criterion is also optimal under the ZIL assumption. In other words, in contrast with lossy dimensionality reduction (see eq. 6), *the MMI and MBL criteria are equivalent for* lossless *dimensionality reduction with* unknown *loss matrix.*

### 2.3 Parametric models

We will show now how the Linear Discriminant Analysis (LDA) and Heteroscedastic Discriminant Analysis (HDA) [4] methods can be derived from the MMI criterion for loseless dimensionality reduction, by making further assumptions regarding the underlying parametric probability model. To begin with, following a path common in Independent Component Analysis (ICA), we analyze feature vector entropy as a difference between a gaussian entropy term, i.e. entropy due to the second order moments of the probability distribution, and a negentropy term, i.e. entropy due to all the other (higher) moments:

$$\begin{aligned} \mathrm{H}[\mathrm{A}^{\mathrm{T}}\mathbf{x}] &= \mathrm{H}_g[\mathrm{A}^{\mathrm{T}}\mathbf{x}] - \mathrm{J}[\mathrm{A}^{\mathrm{T}}\mathbf{x}] \\ \mathrm{H}_g[\mathrm{A}^{\mathrm{T}}\mathbf{x}] &= \frac{1}{2}\log{(2\pi \mathrm{e})^{d'}} \left|\mathrm{cov}[\mathrm{A}^{\mathrm{T}}\mathbf{x}]\right| \end{aligned} \tag{12}$$

Analyzing accordingly the class conditional feature-vector entropy, mutual information decomposes as:

$$\begin{aligned} \mathrm{I}[\mathcal{C}; \mathrm{A}^{\mathrm{T}}\mathbf{x}] &= \mathrm{H}[\mathrm{A}^{\mathrm{T}}\mathbf{x}] - \mathrm{H}[\mathrm{A}^{\mathrm{T}}\mathbf{x}|\mathcal{C}] \\ &= \frac{1}{2}\log\frac{\left|\mathrm{cov}[\mathrm{A}^{\mathrm{T}}\mathbf{x}]\right|}{\prod_{k=1}^{K}\left|\mathrm{cov}[\mathrm{A}^{\mathrm{T}}\mathbf{x}|\mathcal{C}_k]\right|^{p(\mathcal{C}_k)}} - \left(\mathrm{J}[\mathrm{A}^{\mathrm{T}}\mathbf{x}] - \mathrm{J}[\mathrm{A}^{\mathrm{T}}\mathbf{x}|\mathcal{C}]\right) \end{aligned} \tag{13}$$

Now, let assume that the feature space is such that (a) the ZIL assumption holds and (b) the conditional probability distribution of the feature vector given the class is gaussian, i.e. $\mathbf{x}|\mathcal{C}_k \sim \mathcal{N}\left(\mathbb{E}[\mathbf{x}|\mathcal{C}_k], \mathrm{cov}[\mathbf{x}\,|\,\mathcal{C}_k]\right)$. It may then be shown that the difference of negentropies term may be safely neglected while maximizing eq.(13). Moreover, it turns out that the covariance ratio left is the one maximized by HDA, when the covariances are replaced by their samples estimates. Going one step further and assuming that the class-conditional feature vector probability distributions share the same covariance matrix (homoskedasticity), HDA further simplifies to LDA:

$$\mathrm{LDA}[\mathcal{C};\, \mathrm{A}^{\mathrm{T}}\mathbf{x}] = \log \frac{\left|\mathrm{cov}[\,\mathrm{A}^{\mathrm{T}}\mathbf{x}]\right|}{\mathbb{E}_{\mathcal{C}}\left|\mathrm{cov}[\,\mathrm{A}^{\mathrm{T}}\mathbf{x}\,|\,\mathcal{C}_k]\right|}. \tag{14}$$

Using this derivation, we may now comment on the optimality of both HDA and LDA based on the optimality of MMI. In particular, for lossless dimensionality reduction with unknown loss matrix, HDA is optimal under the ZIL + gaussianity assumption, whereas LDA is optimal under the ZIL + gaussianity + homoskedasticity assumption.

Last, let us point out that maximizing eq. (13) under the gaussianity assumption is also equivalent to maximize the negentropy on the extracted subspace. This creates a bridge between supervised and unsupervised learning, since, by assuming that the noise vector is gaussian, we can simply perform multidimensional ICA [1] to reject it.

### 2.4 Sensitivity analysis

Looking back at eq. (2), we notice that a feature vector is meaningless to the overall loss, if a *change* of its value does not result to a *change* of the loss function. A measure of sensitivity of loss in respect to the feature subspace engendered by matrix $\mathrm{A}_{[d\times d']}$ may be defined as:

$$\mathcal{B}_{\mathrm{L}}^{(\mathrm{A})}[\mathcal{C}\,|\,\mathbf{x}] = \mathbb{E}_{\mathbf{x}}\left\|\frac{\partial}{\partial \mathrm{A}} l(\hat{\mathcal{C}}\,|\,\mathbf{x})\right\|^2. \tag{15}$$

where $\hat{\mathcal{C}}$ is the optimal class decision for $\mathbf{x}$ and we have assumed that the loss function is differentiable. Therefore, instead of seeking feature vectors that minimize Bayes loss, we may as well seek those against which the Bayes loss sensitivity is maximized. We define this as the Maximization of Bayes Loss Sensitivity (MBLS) criterion. It turns out that, for performing *lossless* linear dimensionality reduction with *known* loss matrix, the MBL and MBLS are equivalent. The Bayes loss sensitivity in the source subspace will then equal to the overall Bayes loss sensitivity, i.e. $\mathcal{B}_{\mathrm{L}}^{(\mathrm{A})}[\mathcal{C}\,|\,\mathbf{x}] = \mathcal{B}_{\mathrm{L}}^{(\mathrm{I}_d)}[\mathcal{C}\,|\,\mathbf{x}]$.

What's more, by equivalently expressing Bayes loss sensitivity as

$$\mathcal{B}_{\mathrm{L}}^{(\mathrm{A})}[\mathcal{C}\,|\,\mathbf{x}] = \mathrm{trace}\left(\mathrm{A}^{\mathrm{T}}\mathrm{cor}[\nabla l(\hat{\mathcal{C}}\,|\,\mathbf{x})]\mathrm{A}\right), \tag{16}$$

| problem | | features | | classific. accuracy difference | | | |
|---|---|---|---|---|---|---|---|
| name | $K$ | original | optimal | | $d' = 1$ | $d' = 2$ | $d' = 3$ |
| cancer | 2 | 9 | 1 | 0.1 | 0.1 | −0.3 | **−0.7** |
| card | 2 | 51 | 2 | 0.1 | −0.1 | 0.7 | 0.3 |
| diabetes | 2 | 8 | 2 | -0.2 | 0.5 | −0.8 | 0.1 |
| glass | 6 | 9 | 6 | **1.2** | **1.4** | 1.6 | **1.9** |
| heart | 2 | 35 | 3 | **1.7** | **8.6** | **2.4** | **2.3** |
| horse | 2 | 47 | 2 | -0.1 | **2.2** | −0.1 | **2.2** |
| iris | 3 | 4 | 1 | 0.1 | 0.1 | **1.6** | 1.1 |
| sonar | 2 | 60 | 14 | **2.6** | 0.6 | 1.3 | 1.6 |
| soybean | 19 | 82 | 20 | **6.8** | **2.0** | **−13.8** | **−9.1** |
| xorrot | 2 | 8 | 2 | **14.0** | **20.3** | **33.4** | **28.1** |

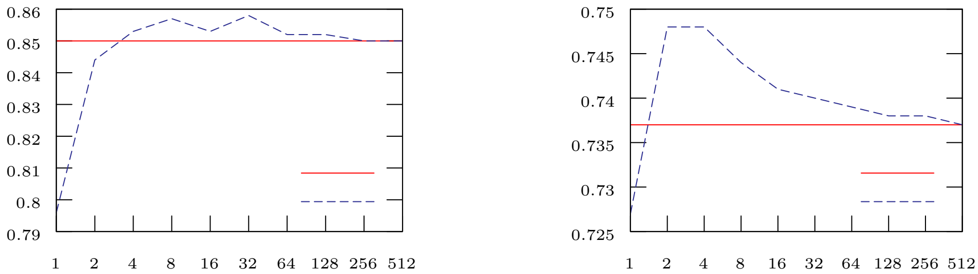
**Table 2.** Difference of generalization performance (average of 30 tries of random 80%-20% cross-validation test of K–NN) using linear feature extraction with the SPCA vs LDA algorithm. The first two columns contain the name of the problem and the number of classes. The following two columns, the original number of features and the optimal ones, obtained as those that maximized the generalization performance with either method. The last columns display the results for optimal and constrained (1, 2 and 3) features respectively. Statistically significant differences (t-test 0.90) are marked in bold.

we observe that applying MBLS comes down to solving an eigenvector problem, or equivalently, to applying Principal Component Analysis (PCA) by first replacing each original sample $\mathbf{x}$ with the corresponding samples sensitivities $\mathbf{x}' = [\ldots, \frac{\partial l(\hat{C}|\mathbf{x})}{\partial x_i}, \ldots]$. A number of algorithms in the pattern recognition literature are essentially using this method, while estimating the derivatives via local parametric models, neural networks or SVMs [8]. We explore here the potential of the method based on estimation of sensitivities via the standard Parzen method (the Parzen-SPCA algorithm ). Table 2 shows an increase in generalization accuracy, as compared to the one obtained using the popular LDA method, for both lossy and lossless dimensionality reduction.

## 3 Optimal invertible transforms

### 3.1 Invertible vs Non-invertible transforms

The study of optimal dimension reduction in Section 2 reveals two important issues regarding the Bayes loss and mutual information measures. First, these measures fail to explain why, often, dimensionality reduction increases generalization performance, as compared to just not worsen it (see eq. 8). The second issue relates to invertible transforms, i.e. those for which we may revover the original feature vector given the transformed one, such as scaling transforms [3]. Namely, both Bayes loss and the mutual information with the class variable stay



(a) card– LDA

(b) diabetes–SPCA

**Fig. 1.** Scaling vs rejection. The continuous line corresponds to the generalization performance of the K–NN algorithm using the LDA and SPCA optimal transforms. The dashed-line corresponds to the performance of the same algorithm using the initial feature space with the features suggested for reduction being scaled down. The scale-down factors are represented in the x-axis.

unchanged after application of an invertible transformand hence seem useless for optimal invertible transform search.

It can be shown, however, that invertible transforms *do* influence the generalization accuracy. To illustrate this fact, we present a test-case involving linear invertible transforms , compactly expressed as a product of a full rank diagonal matrix and a full rank rotation matrix, as $\mathrm{A} = \mathrm{D}^{\mathrm{T}}\mathrm{Q}$. Notice that, by letting a subset $L = \{d_l\}$ of the diagonal elements of D being scaled down to zero,

$$\mathrm{A}' = \lim_{\forall d_l \in L : d_l \to 0} \mathrm{D}^{\mathrm{T}}\mathrm{Q}, \tag{17}$$

the resulting matrix will have reduced rank and will map the original feature space to a subspace. Thus, linear dimensionality reduction may actually be considered as a limit of these transforms. We can then verify that this convergence also applies to the corresponding generalization accuracy in respect to a given classification problem. Namely, by applying the LDA and SPCA algorithms to our set of benchmark tests and, instead of rejecting the dimensions with low suitability, scaling them down, we obtain the curves in Figure 1. One may see that the generalization does depend on feature scaling and converges, at the limit of very small scaling factors, to the one obtained by feature extraction.

A further finding from this experiment is that generalization performance may actually be superior when scaling down the features with an appropriate factor rather than rejecting them. This indicates that LDA and SPCA are somehow pertinent for invertible transforms, although they fail to determine the optimal scaling factors. In the following sections, we study extended forms of the Bayes and MMI criteria, converging to the plain ones at the non-invertibility limit and propose a new algorithm for learning invertible transforms.

### 3.2 Effective feature vector optimality

Exact knowledge of the feature vector probability distribution, as assumed by the Bayes loss and the mutual information criteria, does not hold in practice, and this is the only reason they may have for failing to explain the increase in generalization accuracy after a dimensionality reduction or invertible transform. Here, to formally account for limited sample size, we use a metaphor and postulate that any classification algorithm is "looking" at the feature space through a "blurring" gaussian *focus*:

$$g(\mathbf{s}\,|\,r,\,Q\,) \triangleq e^{-r\cdot\|Q^{\mathrm{T}}\mathbf{s}\|^2} \tag{18}$$

where $r \in \mathbb{N}^*$ is the focus magnitude and $Q, |Q| = 1$, is the focus angle matrix. Roughly, focus magnitude is related to samples density, whereas, importantly, focus angle describes a specific view of the feature space (at the limit, a projection) the classification algorithm concentrates on. The limited focus "blurs" the probability at $\mathbf{x}$, allowing only to see the probability of *a region around* $\mathbf{x}$:

$$P(\mathbf{x} = \mathbf{x}|r, Q) = \int_{\mathbb{R}^d} g(\,\mathbf{s}\,|\,r(\mathbf{x})\,,\,Q(\mathbf{x})\,)\cdot p(\,\mathbf{x} = \mathbf{x} + \mathbf{s}\,)\cdot d\mathbf{s}.$$

Conveniently, we may define the *effective feature vector* $\mathring{\mathbf{x}}_{r,Q}$ as a feature vector with probability distribution function the normalized probability of a region

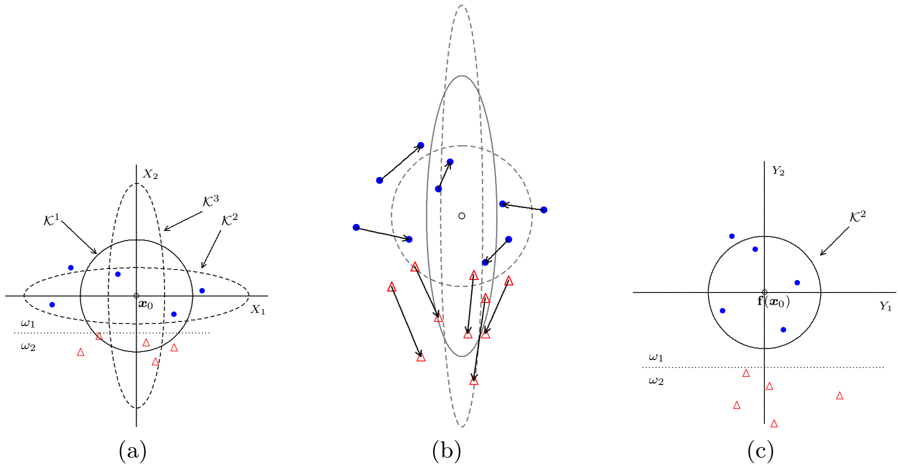
$$p(\,\mathring{\mathbf{x}}_{r,Q} = \mathbf{x}\,) \triangleq \frac{P(\mathbf{x} = \mathbf{x}|r, Q)}{\int_{\mathbb{R}^d} P(\mathbf{x} = \mathbf{x}|r, Q)\cdot d\mathbf{x}}. \tag{19}$$

and then define functionals of the effective feature vector, such as the Bayes loss $\mathcal{B}_{\mathrm{L}}[\,\mathcal{C}\,|\,\mathring{\mathbf{x}}_{r,Q}\,]$ and the mutual information $\mathrm{I}[\,\mathring{\mathbf{x}}_{r,Q}\,;\,\mathcal{C}\,]$. Notice that, as opposed to the original measures, by keeping the focus constant, any invertible transform of the feature space will also have an impact on these "effective" measures. What's more, as focus magnitude goes to infinity (i.e. at the limit of an infinitely dense sampled feature space) and for any focus angle, the effective feature vector converges to the original one:

$$\lim_{\forall \mathbf{x}:r(\mathbf{x})=\infty} \mathring{\mathbf{x}}_{r,Q} = \mathbf{x}, \tag{20}$$

thus allowing us to represent in a uniform way both "myopic" and complete knowledge of the probability distribution function and of its functionals.

We may say that non-parametric classification algorithms, such as the K–NN or RBF-Support Vector Machine (SVM) [2], when locally evaluating classification decisions, they have constraints regarding the focus angle. First, they don't adapt the angle locally, i.e. $Q(\mathbf{x}) = Q$ and, second, they equally weight all features, which translates as $Q = \mathrm{I}_d$. We may jointly call these properties *isotropicity.* It is thus to be expected that these algorithms will work better on feature space in which these two assumptions hold. It makes sense, therefore, to seek for transforms which tend to maximize the Bayes loss and mutual information in



(a) (b) (c)

**Fig. 2.** Local stretching

respect to isotropic focus angle. We will refer to those criteria as Minimization of Isotropic Loss (MIL)

$$\hat{\mathbf{f}} \triangleq \underset{\{\mathbf{f}\in\mathcal{F}\}}{\operatorname{argmin}} \mathring{\mathcal{B}}_{\mathrm{R}}(\mathcal{C}\,|\,\mathbf{f}(\mathbf{x}))$$

and Maximization of Isotropic Information (MII)

$$\hat{\mathbf{f}} \triangleq \underset{\{\mathbf{f}\in\mathcal{F}\}}{\operatorname{argmax}} \mathring{\mathcal{I}}(\mathbf{f}(\mathbf{x});\mathcal{C}).$$

which take, respectively do not take, into account a specific loss matrix. Figure 2 shows how a suitable invertible transform $\mathbf{f}$ may locally adapt the feature space so that subsequent classification with isotropically-constraint classifiers has more chances to be correct. Typically, since optimal $\mathbf{f}$ is hard to find, a general approach we suggest here, is to first locally estimate $\mathbf{f}$ on available samples, and then aggregate the local estimations to approximate a global overall transform, taking into account the constraints of the transform family (see Section 3.3). However, instead of finding an optimal $\mathbf{f}$, one may also optimally adapt kernels locally to improve classification performance.

### 3.3 Non-Linear Feature Scaling

A review of the benefits of feature vector transforms reveals that transforms other than ones that end up in reduced dimension may be actually worthy to consider. A first argument comes from questioning a popular claim regarding dimensionality reduction, namely that it reduces the complexity of further processing steps, since these will be performed on a smaller feature space. Even this sounds as an important reason, one should notice that the complexity needed to reduce the feature space in an optimal way should be added to the complexity of the steps to follow and, hence, it is not straightforward how the overall complexity gets reduced. A second argument, related to visualization of data, is that reducing the number of dimensions is not the only way to improve problem inspection. Other possibilities exist, as long as theses involve putting in evidence the distances of samples important to distinguish the classes. Last, let us point out that the commonly desired intepretability of a transform comes with the cost of low transform capacity. Here, we suggest that interpretability of the transform is linked to the easiness we have to associate each and every one of the final features with the the original ones: if a simple association of one original feature with the final is possible, then the transform may indeed be called interpretable.

A family of transforms effectively dealing with the above issues is Non-Linear Feature Scaling (NLFS),

$$\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d, \mathbf{f}(\mathbf{x}) = [f_1(x_1), f_2(x_2), \ldots, f_d(x_d)]^{\mathrm{T}},$$

where the original features are one-to-one mapped to the final features. In addition, the derivative of the transform in respect to any of the features is imposed to be strictly positive, which implies that $\mathbf{f}$ is invertible and thus no information is lost. The transform locally shrinks or expands each feature, depending on the value of the derivative (*stressing function*):

$$\frac{\partial}{\partial x_i} \mathbf{f}(x_i) \lesseqgtr 1.$$

Interpretability of the transform is supported by noticing that, while the distance between samples changes, their order along each feature stays unchanged. Transform complexity is controlled by forcing the stressing function to remain constant within a given interval of the feature values: constant stressing for the whole feature values range is equivalent to plain (linear) feature scaling, whereas, by repeatedly dividing the feature into small bins, we obtain a finer “grid” form of the transform. The goal of learning, then, is to find an optimal set of stressing factors for each one of the features.

To this end, we applied the MII criterion, under the condition that locally optimal matrices are diagonal. An analytical description of the resulting algorithm (Isotropic Non-Linear Feature Scaling (ISONLFS)) can be found in the author’s thesis (see also . The algorithm has been evaluated on the set of our benchmark tests against the SPCA method (see Section 2.4). The generalization performance obtained (see Table 3), justify both the use of invertible transforms for improving performance as well as the application of the proposed method to find optimal non-linear feature transforms.

## 4 Conclusions

We studied optimal feature transforms in a common framework though general optimality criteria (minimization of Bayes loss, maximization of mutual information, maximization of Bayes loss sensitivity). By introducing the concept of

| problem | methods | | comp. | |
|---|---|---|---|---|
| | SPCA | ISONLFS | diff. | t-test |
| cancer | 96.7 | 96.8 | +0.1 | |
| card | 85.7 | 86.1 | +0.1 | |
| diabetes | 74.9 | 76.7 | +1.8 | + |
| glass | 66.5 | 76.4 | **+9.9** | + |
| heart | 79.5 | 79.7 | +0.2 | |
| horse | 65.6 | 65.1 | -0.5 | |
| iris | 97.8 | 96.8 | - 1.0 | - |
| sonar | 80.9 | 89.5 | **+8.6** | + |
| soybean | 93.4 | 92.3 | -1.1 | - |

**Table 3.** ISONLFS: Generalization performance comparison with SPCA.

effective feature vector, we extended the application of this criteria to both dimensionality reduction transforms and invertible transforms. Exporing the relation between these criteria under specific assumptions reveals the connection of a variety of methods (MMI, LDA, HDA). Our framework allowed the derivation of new algorithms (SPCA, ISONLFS) with improved performance in respect to traditional feature extraction methods.

## References

1. Jean-Francois Cardoso. Multidimensional independent component analysis. In *Proceedings of ICASSP*, pages 1941–1944, 1998.
2. Olivier Chapelle, Jason Weston, Leon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *Advances in Neural Information Processing Systems*, volume 13, 2000.
3. Wiodzislaw Duch. Similarity based methodsq a general framework for classification approximation and association. *Control and Cybernetics*, XX(Y), 2000.
4. Nagendra Kumar and Sndreas G. Andreou. On generalizations of linear discriminant analysis. Technical Report JHU/ECE-96-07, Electrical and Computer Engineering, Johns Hopkins University, April 1996.
5. S. J. Perantonis, S. Petridis, and V. Virvilis. Supervised principal component analysis using a smooth classifier paradigm. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 2, pages 109–112, 2000.
6. S. Petridis and S. J. Perantonis. Feature deforming for improved similarity based learning. In *Proceedings of the 3rd Hellenic Conference on Artificial Intelligence*, volume 3025 of *Lecture Notes in Artificial Intelligence*, pages 201–209, 2004.
7. S. Petridis and S. J. Perantonis. On the relation between discriminant analysis and mutual information for supervised linear feature extraction. *Pattern Recognition*, 37:857–874, 2004.
8. Jiayong Zhang and Yanxi Liu. Svm decision boundary based discriminative subspace induction. *Pattern Recognition*, 38(10):1746–1758, October 2005.