Unsupervised Slow Subspace-Learning from Stationary Processes

Andreas Maurer

Adalbertstr. 55
D-80799 München
andreasmaurer@compuserve.com

Abstract. We propose a method of unsupervised learning from stationary, vector-valued processes. A low-dimensional subspace is selected on the basis of a criterion which rewards data-variance (like PSA) and penalizes the variance of the velocity vector, thus exploiting the short-time dependencies of the process. We prove error bounds in terms of the β-mixing coefficients and consistency for absolutely regular processes. Experiments with image recognition demonstrate the algorithms ability to learn geometrically invariant feature maps.

1 Introduction

Some work has been done to extend the results of learning theory from independent, identically distributed input variables to more general stationary processes ([19], [8], [16]). For suitably mixing processes this extension is possible, with an increase in sample complexity caused by dependencies which slow down the estimation process. But some of these dependencies also provide important information on the environment generating the process and can be turned from a curse to a blessing, in particular in the case of unsupervised learning, when side information is scarce and the sample complexity is not as painfully felt.

Consider a stationary stochastic process modeling the evolution of complex sensory signals by a sequence of zero-mean random variables X_t taking values in a Hilbert-space H. Let $\mathcal{P}_d$ be the class of d-dimensional orthogonal projections in H. From observation of $X_0, ..., X_m$ we seek to find some $P \in \mathcal{P}_d$ such that the projected stimulus PX on average captures the significance implied by the primary stimulus $X \in H$. To guide this search we will invoke two principles of common sense.

The first principle states that *significant signals should have a large variance.* In view of the zero-mean assumption this classical idea suggests to maximize $\mathbb{E}\left[\|PX_0\|^2\right]$, which coincides with the objective of PSA[1]([9], [10], [15]) seeking to give the perspective with the broadest view of the distribution.

[1] Principal Subspace Analysis, sometimes Principal Component Analysis (PCA) is used synonymously

The second principle, the principle of *slowness* (introduced by Földiak [2], promoted and developed by Wiskott [17]), states that *sensory signals vary more quickly than their significance*. Consider the visual impressions caused by a familiar complex object, like a tree on the side of the road or a person acting in a movie. Any motion or deformation of the object will cause rapid changes in the states of retinal photoreceptors (or pixel-values). Yet the identities of the tree and the person in the movie remain unchanged. When a person speaks, the communicated ideas vary much more slowly than individual phonemes, let alone the air pressure amplitudes of the transmitted sound signal.

The slowness principle suggests to minimize $\mathbb{E}\left[\left\|P\dot{X}_0\right\|^2\right]$ (here $\dot{X}$ is the velocity process $\dot{X}_t = X_t - X_{t-1}$), and combining both principles leads to the objective function

$$L_\alpha(P) = \mathbb{E}\left[\alpha\|PX_0\|^2 - (1-\alpha)\left\|P\dot{X}_0\right\|^2\right],$$

to be maximized, where the parameter $\alpha \in [0,1]$ controls the trade-off between two potentially conflicting goals. In section 4 we will further justify the use of this objective function and show that for $\alpha \in (0,1)$ maximizing L_α minimizes an error bound for a simple classification algorithm on a generic class of classification problems, and that $\sqrt{\alpha}$ can be interpreted as a scale-parameter. When there is no ambiguity we write $L = L_\alpha$.

As the details of the process X are generally unknown, the optimization has to rely on an empirical basis. Let $(X)_0^m = (X_0, ..., X_m)$ be $m+1$ consecutive observations of the process X and define an empirical analogue $\hat{L}(P)$ of the objective function L

$$\hat{L}(P) = \frac{1}{m}\sum_{i=1}^{m}\left(\alpha\|PX_i\|^2 - (1-\alpha)\left\|P\dot{X}_i\right\|^2\right).$$

We now propose to seek $P \in \mathcal{P}_d$ to maximize $\hat{L}(.)$. This optimization problem, its analysis, algorithmic implementation and preliminary experimental tests are the contributions of this paper.

Existence of Solutions. We will require the general boundedness assumption that $\|X_t\| \leq 1/2$ a.s. Define an operator T on H by

$$Tz = \mathbb{E}\left[\alpha\langle z, X\rangle X - (1-\alpha)\left\langle z, \dot{X}\right\rangle \dot{X}\right] \text{ for } z \in H. \tag{1}$$

Then $T = \alpha C_X - (1-\alpha)C_{\dot{X}}$, where C_X and $C_{\dot{X}}$ are the covariance operators corresponding to X and $\dot{X}$ respectively. The empirical counterpart to T is $\hat{T}$ defined by

$$\hat{T}z = \frac{1}{m}\sum_{i=1}^{m}\left(\alpha\langle z, X_i\rangle X_i - (1-\alpha)\left\langle z, \dot{X}_i\right\rangle \dot{X}_i\right). \tag{2}$$

The operators T and $\hat{T}$ are central objects of the proposed method. They are both symmetric and compact, T is trace-class and $\hat{T}$ has finite rank. If $\alpha \in (0,1)$ they will tend to have both positive and negative eigenvalues. The following Theorem (see section 2) shows that a solution of our optimization problem can be obtained by projecting onto a dominant eigenspace of $\hat{T}$.

Theorem 1. *Fix $\alpha \in [0,1]$ and let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq 0$ be the nonnegative eigenvalues of $\hat{T}$, and (e_i) the sequence of associated eigenvectors. Then*

$$\max_{P \in \mathcal{P}_d} \hat{L}(P) = \sum_{i=1}^{d} \hat{\lambda}_i,$$

the maximum being attained when P is the orthogonal projection onto the span of $e_1, ..., e_d$.

This leads to a straightforward batch algorithm: Observe and store a realization of $(X_0, ..., X_m)$, construct $\hat{T}$, find eigenvectors and eigenvalues and project onto the span of d orthonormal eigenvectors corresponding to the largest eigenvalues.

Such a solution P need not be unique. In fact, if $\alpha = 0$ and $\dim(H) = \infty$, then $\hat{T}$ is a nonpositive operator with infinite dimensional nullspace, and there is an infinity of mutually orthogonal solutions, from which an arbitrary choice must be made. This can hardly be the way to extract meaningful signals, and the utility of the objective function with $\alpha = 0$ is questionable for high-dimensional input spaces. Except for very pathological cases, this extreme degeneracy is absent in the case $\alpha > 0$. In the generic, noisy case all nonzero eigenvalues will be distinct and if m is large then there are more than d positive eigenvalues of $\hat{T}$, so that the solution will be unique.

Estimation. Having found P to maximize $\hat{L}(.)$, can we be confident that $L(P)$ is also nearly maximal, and how does this confidence improve with the sample size?

These questions are complicated by the interdependence of observations, in particular by the possibility of being trapped for longer periods of time. Since we want to estimate an expectation on the basis of a temporal average, some sort of ergodicity property of the process X will be relevant. Our bounds are expressed in terms of the mixing coefficients $\beta(a)$, which roughly bound the interdependence of past and future variables separated by a time interval of duration a. Combining the techniques developed in [11] and [19] we arrive at the following result:

Theorem 2. *With the assumptions already introduced above, fix $\delta > 0$ and let $m, a \in \mathbb{N}$, $a < m/2$ and $l = \lfloor m/2a \rfloor$ and $\beta(a) < \delta/(2l)$. Then with probability greater $1 - \delta$ in the sample $(X)_0^m = (X_0, ..., X_m)$ we have*

$$\sup_{P \in \mathcal{P}_d} \left| \hat{L}(P) - L(P) \right| \leq \frac{4}{\sqrt{l}} \left(\sqrt{d} + \sqrt{\frac{1}{2} \ln \frac{1}{\delta/2 - l\beta(a-1)}} \right).$$

If the mixing coefficients β are known, then the right hand side can be minimized with an appropriate choice of a, which in general depends on the sample size (or *total learning time*) m. For easy interpretation assume $\beta(a) = 0$ for $a \geq a_0$. Then we can interpret a_0 as the mixing time beyond which all correlations vanish. If we set $a = a_0 + 1$ above, the resulting bound resembles the bound for the iid case with an effective sample size $l = \lfloor m/(2(a_0+1)) \rfloor$. This shows the ambiguous role of temporal dependencies: Over short time intervals they are beneficial, providing us with information which allows us to go beyond PSA by using the slowness principle. Over long periods of time they get in the way of mixing and become detrimental to learning.

Often the mixing coefficients are unknown, but one knows (or assumes or hopes) that X is absolutely regular, that is $\beta(a) \to 0$ as $a \to \infty$. We can then still establish learnability in the sense of convergence in probability:

Theorem 3. *If X is absolutely regular then for every $\epsilon > 0$ we have*

$$\lim_{m\to\infty} \Pr\left\{ \sup_{P\in\mathcal{P}_d} \left| \hat{L}(P) - L(P) \right| > \epsilon \right\} = 0.$$

We will prove both theorems in section 3.

A major problem caused by large observation times is the accumulating memory requirement to store the sample data, as long as we adhere to the batch algorithm sketched above. For this reason we use an online-algorithm for our experiments in image processing. The algorithm, a modification of an algorithm introduced by Oja [9], is briefly introduced in section 5. We apply it either directly to the image data or to train the second layer of a two-layered radial-basis-function network.

The experiments reported in section 6 involve processes with specific geometric invariants: Consider rapidly rotating views of a slowly changing scene. The projection returned by our algorithm then performs well as a preprocessor for rotation invariant recognition. An analogous behaviour was observed for scale-invariance, and it might be conjectured that similar mechanisms could account for the ubiquity of scale invariant perception in biological vision.

A similar technique has been proposed by Wiskott [17]. It is missing an analogue of a positive variance term in the objective function. The problem of potentially trivial solutions is circumnavigated by an orthonormalization prescription (whitening) of the covariance matrix prior to the subspace search, which then essentially seeks out a minimal subspace of the velocity covariance. In high (or infinite) dimensions minimal subspace analysis of (compact positive) operators should cause the above-mentioned degeneracy problem, because the eigenvalues will concentrate at zero. In [17] a corresponding problem is in fact mentioned. Also the orthonormalization increases the norms of the input vectors as the dimension grows, making it difficult to analyse the generalisation behaviour. In our approach all these problems are eliminated by a positive variance term, corresponding to $\alpha > 0$.

2 Preliminaries

For the next sections H will be a real separable infinite-dimensional Hilbert space with norm $\|.\|$ and inner product $\langle .,. \rangle$. In practice H will be finite dimensional, but as the dimension is large and should not enter into our results we may as well assume infinite-dimensionality, which will also eliminate some complications.

2.1 Hilbert Schmidt operators

With H_2 we denote the real vector space of symmetric operators on H satisfying $\sum_{i=1}^{\infty} \|Te_i\|^2 < \infty$ for every orthonormal basis $(e_i)_{i=1}^{\infty}$ of H. For $S, T \in H_2$ the number $\langle S, T \rangle_2 = Tr(TS)$ defines an inner product on H_2, making it into a Hilbert space with norm $\|T\|_2 = \langle T, T \rangle_2^{1/2}$. The members of H_2 are compact and called Hilbert-Schmidt operators (see Reed and Simon [12] for background on functional analysis). For every $v \in H$ we define an operator Q_v by

$$Q_v x = \langle x, v \rangle v \text{ for all } x \in H.$$

The set of d-dimensional, orthogonal projections in H is denoted with $\mathcal{P}_d$. The following facts are easily verified (see [5]):

Lemma 1. *Let $x, y \in H$ and $P \in \mathcal{P}_d$. Then (i) $Q_x \in H_2$ and $\|Q_x\|_2 = \|x\|^2$, (ii) $\langle Q_x, Q_y \rangle_2 = \langle x, y \rangle^2$, (iii) $\langle P, Q_x \rangle_2 = \|Px\|^2$ and (iv) $\|P\|_2 = \sqrt{d}$.*

In terms of the Q-operators we can rewrite the operators T and $\hat{T}$ in (1) and (2) as

$$T = \mathbb{E}\left[\alpha Q_X - (1-\alpha) Q_{\dot{X}}\right] \text{ and } \hat{T} = \frac{1}{m} \sum_{i=1}^{m} \left(\alpha Q_{X_i} - (1-\alpha) Q_{\dot{X}_i}\right).$$

Using (iii) above, the objective functionals $L(.)$ and $\hat{L}(.)$ become

$$L(P) = \langle T, P \rangle_2 \text{ and } \hat{L}(P) = \left\langle \hat{T}, P \right\rangle_2.$$

Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq 0$ be any nonincreasing enumeration of the nonnegative eigenvalues of $\hat{T}$, counting multiplicities, and (e_i) a corresponding orthonormal sequence of eigenvectors. Note that the sequence is necessarily infinite because $\hat{T}$ has finite rank and thus an infinite-dimensional null-space. Now let $P \in \mathcal{P}_d$. Since P has the eigenvalue 1 with multiplicity d and all its other eigenvalues are zero, it follows from Horn's theorem [14, Theorem 1.15] that

$$\left\langle \hat{T}, P \right\rangle_2 \leq \sum_{i=1}^{d} \hat{\lambda}_i.$$

If P is the projection onto the span of $e_1, ..., e_d$ then this becomes an equality. This shows that any such maximal projection P is also a maximizer for $\hat{L}(P)$ and that

$$\max_{P \in \mathcal{P}_d} \hat{L}(P) = \sum_{i=1}^{d} \hat{\lambda}_i,$$

thus proving Theorem 1.

These arguments are fairly standard, but in the infinite dimensional case there are some pitfalls resulting from non-positivity. For example the above is not generally true for the operator T corresponding to the true objective functional L, because it may happen that T has fewer than d nonnegative eigenvalues, or none at all. Since all negative eigenvalues converge to 0, the supremum might not be attained.

2.2 Mixing coefficients and inequalities

Let $\xi = \{\xi_t\}_{t \in \mathbb{Z}}$ be a stationary stochastic process with values in a measurable space (Ω, Σ) and with law μ. For $A \subseteq \mathbb{Z}$ let σ_A denote the σ-algebra generated by the variables ξ_t with $t \in A$, and use μ_A to denote the marginal distribution of μ on $\left(\Omega^A, \sigma_A\right)$.

Definition 1. *For $k \in \mathbb{N}$ define the mixing coefficient*

$$\beta_\xi(k) = \mathbb{E}\left[\sup\left\{\left|\mu\left(B|\sigma_{\{t:t\leq l\}}\right) - \mu(B)\right| : B \in \sigma_{\{t:t\geq l+k\}}\right\}\right].$$

The process ξ is called absolutely regular or β-mixing if $\beta_\xi(k) \to 0$ as $k \to \infty$.

The interpretation is as follows: The random variable

$$\sup\left\{\left|\mu\left(B|\sigma_{\{t:t\leq l\}}\right) - \mu(B)\right| : B \in \sigma_{\{t:t\geq l+k\}}\right\}$$

gives the largest change in the probability of any future event B occurring when a specific realization of the past is unveiled. It therefore measures the maximal dependence of the future $\{t \geq l+k\}$ on the past $\{t \leq l\}$, as a function of the past. Taking the expectation of this variable leads to a quantity which is itself independent of the past but takes the probabilities of different realizations of the past into account (see the book by Rio [13] for a general theory of weakly dependent processes). From this definition one can prove the following (Yu [19]):

Lemma 2. *Let $\xi = \{\xi_t\}_{t \in \mathbb{Z}}$ be stationary with values in a measurable space (Ω, Σ) and $B \in \sigma_{\{1,...,m\}}$. Then*

$$\left|\mu_{\{1,...,m\}}(B) - \left(\mu_{\{1\}}\right)^m(B)\right| \leq (m-1)\,\beta_\xi(1).$$

We will also need the following lemma of Vidyasagar [16, Lemma 3.1]:

Lemma 3. *Suppose $\beta(k) \downarrow 0$ as $k \to \infty$. It is possible to choose a sequence $\{a_m\}$ such that $a_m \leq m$, and with $l_m = \lfloor m/a_m \rfloor$ we have that $l_m \to \infty$ while $l_m \beta(a_m) \to 0$ as $m \to \infty$.*

3 Generalization

We first prove a general result for vector-valued processes. For two subsets $V, W \subseteq H$ of a Hilbert space H we introduce the following notation

$$\|V\| = \sup_{v \in V} \|v\| \text{ and } |\langle V, W \rangle| = \sup_{v \in V, w \in W} |\langle v, w \rangle| .$$

Theorem 4. *Let $V, W \subset H$ and $X = \{X_t\}_{t \in \mathbb{Z}}$ a stationary, mean zero process with values in V.*

1. Fix $\delta > 0$ and let $m, a \in \mathbb{N}$, $a < m/2$ and $l = \lfloor m/2a \rfloor$ and $\beta(a) < \delta/(2l)$. Then with probability greater than $1 - \delta$ we have

$$\sup_{w \in W} \left| \frac{1}{m} \sum_{i=1}^{m} \langle w, X_i \rangle \right| \leq \frac{2}{\sqrt{l}} \left(\|V\| \|W\| + |\langle V, W \rangle| \sqrt{\frac{1}{2} \ln \frac{1}{\delta/2 - l\beta_X(a)}} \right).$$

2. If X is absolutely regular then for every $\epsilon > 0$

$$\Pr \left\{ \sup_{w \in W} \left| \frac{1}{m} \sum_{i=1}^{m} \langle w, X_i \rangle \right| > \epsilon \right\} \to 0 \text{ as } m \to \infty.$$

If we let W be the unit ball in H we immediately obtain the following

Corollary 1. *Under the first assumptions of Theorem 4 we have with probability greater $1 - \delta$ that*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} X_i \right\| \leq \frac{2\|V\|}{\sqrt{l}} \left(1 + \sqrt{\frac{1}{2} \ln \frac{1}{\delta/2 - l\beta_X(a)}} \right).$$

If in addition X_t is absolutely regular then $\|(1/m) \sum_{i=1}^{m} X_i\| \to 0$ in probability.

Here is a practical reformulation with trivial proof:

Corollary 2. *Theorem 4 and Corollary 1 remain valid if the mean-zero assumption is omitted, X_i is replaced by $X_i - \mathbb{E}[X_1]$ and $\|V\|$ and $|\langle V, W \rangle|$ are replaced by $2\|V\|$ and $2|\langle V, W \rangle|$ respectively.*

To prove Theorem 4 we first establish an analogous result for iid X_i (essentially following [11]) and then adapt it to dependent variables.

Lemma 4. *Let $V, W \subset H$ be and $X_1, ..., X_m$ iid zero-mean random variables with values in V. Then for ϵ and m such that $\|W\| \|V\| < \sqrt{m}\epsilon$ we have*

$$\Pr \left\{ \sup_{w \in W} \left| \frac{1}{m} \sum_{i=1}^{m} \langle w, X_i \rangle \right| > \epsilon \right\} \leq \exp \left(\frac{-\left(\sqrt{m}\epsilon - \|V\| \|W\|\right)^2}{2|\langle V, W \rangle|^2} \right).$$

Proof. Consider the average $\bar{\mathbf{X}} = (1/m)\sum_1^m X_i$. With Jensen's inequality and using independence we obtain

$$\left(\mathbb{E}\left[\left\|\bar{\mathbf{X}}\right\|\right]\right)^2 \leq \mathbb{E}\left[\left\|\bar{\mathbf{X}}\right\|^2\right] = \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}\left[\left\|X_i\right\|^2\right] \leq \left\|V\right\|^2/m.$$

Now let $f : V^m \to \mathbb{R}$ be defined by $f(\mathbf{x}) = \sup_{w\in W}\left|(1/m)\sum_1^m \langle w, x_i\rangle\right|$. We have to bound the probability that $f > \epsilon$. By Schwartz' inequality and the above bound we have

$$\mathbb{E}\left[f(\mathbf{X})\right] = \mathbb{E}\left[\sup_{w\in W}\left|\left\langle w, \bar{\mathbf{X}}\right\rangle\right|\right] \leq \left\|W\right\|\mathbb{E}\left[\left\|\bar{\mathbf{X}}\right\|\right] \leq \left(1/\sqrt{m}\right)\left\|W\right\|\left\|V\right\|. \tag{3}$$

Let $\mathbf{x} \in V^m$ be arbitrary and $\mathbf{x}' \in V^m$ be obtained by modifying a coordinate x_k of $\mathbf{x}$ to be an arbitrary $x'_k \in V$. Then

$$\left|f(\mathbf{x}) - f(\mathbf{x}')\right| \leq \frac{1}{m}\sup_{w\in W}\left|\langle w, x_k\rangle - \langle w, x'_k\rangle\right| \leq \frac{2}{m}\left|\langle V, W\rangle\right|.$$

By (3) and the bounded-difference inequality (see [7]) we obtain for $t > 0$

$$\Pr\left\{f(\mathbf{X}) > \frac{\left\|W\right\|\left\|V\right\|}{\sqrt{m}} + t\right\} \leq \Pr\left\{f(\mathbf{X}) - \mathbb{E}\left[f(\mathbf{X})\right] > t\right\} \leq \exp\left(\frac{-mt^2}{2\left|\langle V, W\rangle\right|^2}\right).$$

The conclusion follows from setting $t = \epsilon - \left(1/\sqrt{m}\right)\left\|W\right\|\left\|V\right\|$ ■

The proof of Theorem 4 now uses the techniques introduced by Yu [19] (see also Meir [8] and Lozano et al [3]).

Proof (of Theorem 4). Select a time-scale $a \in \mathbb{N}$, $2a < m$ and represent the discrete time axis as an alternating sequence of blocks

$$\mathbb{Z} = \left(..., H_{-1}, T_{-1}, H_0, T_0, H_1, T_1, ..., H_k, T_k, ...\right),$$

where each of the H_k and T_k has length a,

$$H_k = \left\{2ka, ..., 2ka + a - 1\right\} \text{ and } T_k = \left\{(2k+1)a, ..., (2k+1)a + a - 1\right\}.$$

We now define the blocked processes X^H and X^T with values in $\mathrm{co}(V)$ by $X_t^H = (1/a)\sum_{j\in H_t} X_j$ and $X_t^T = (1/a)\sum_{j\in T_t} X_j$. By stationarity the X_i^H and X_i^T are identically distributed and themselves stationary. Because of the gaps of size a we have $\beta_{X^H}(1) = \beta_{X^T}(1) = \beta_X(a)$. We can now write

$$(1, ..., m) = \left(H_1, T_1, H_2, T_2, ..., H_l, T_l, R\right),$$

where the number l of block-pairs is chosen so as to minimize the size of the remainder R, so $l = \lfloor m/(2a) \rfloor$ and $|R| < 2a$. For arbitrary $\epsilon > 0$ we obtain

$$\begin{aligned}
&\Pr\left\{\sup_{w\in W}\left|\frac{1}{2al}\sum_{i=1}^{2al}\langle w, X_i\rangle\right| > \epsilon\right\} \\
&= \Pr\left\{\sup_{w\in W}\left|\frac{1}{2l}\sum_{i=1}^{l}\left\langle w, X_i^H\right\rangle + \frac{1}{2l}\sum_{i=1}^{l}\left\langle w, X_i^T\right\rangle\right| > \epsilon\right\} \\
&\leq \Pr\left\{\sup_{w\in W}\left|\frac{1}{2l}\sum_{i=1}^{l}\left\langle w, X_i^H\right\rangle\right| + \sup_{w\in W}\left|\frac{1}{2l}\sum_{i=1}^{l}\left\langle w, X_i^T\right\rangle\right| > \epsilon\right\} \\
&= 2\Pr\left\{\sup_{w\in W}\left|\frac{1}{l}\sum_{i=1}^{l}\left\langle w, X_i^H\right\rangle\right| > \epsilon\right\} \\
&\leq 2\exp\left(\frac{-\left(\sqrt{l}\epsilon - \|V\|\,\|W\|\right)^2}{2\left|\langle V, W\rangle\right|^2}\right) + 2l\beta_X(a).
\end{aligned}$$

The last inequality follows from the mixing Lemma 2, $\beta_{X^H}(1) = \beta_X(a)$, the iid case Lemma 4 and the fact that $\|\mathrm{co}(V)\| = \|V\|$ and $|\langle \mathrm{co}(V), W\rangle| = |\langle V, W\rangle|$. To deal with the remainder R, note that

$$\Pr\left\{\sup_{w\in W}\left|\frac{1}{m}\sum_{i=1}^{m}\langle w, X_i\rangle\right| > \epsilon\right\} \leq \Pr\left\{\sup_{w\in W}\left|\frac{1}{2al}\sum_{i=1}^{2al}\langle w, X_i\rangle\right| + \frac{\|V\|\,\|W\|}{l} > \epsilon\right\}.$$

We thus obtain

$$\begin{aligned}
&\Pr\left\{\sup_{w\in W}\left|\frac{1}{m}\sum_{i=1}^{m}\langle w, X_i\rangle\right| > \epsilon\right\} \\
&\leq 2\exp\left(\frac{-\left(\sqrt{l}\epsilon - \left(1+\frac{1}{\sqrt{l}}\right)\|V\|\,\|W\|\right)^2}{2\left|\langle V, W\rangle\right|^2}\right) + 2l\beta_X(a). \qquad (4)
\end{aligned}$$

Solving for ϵ and using $\left(1 + 1/\sqrt{l}\right) \leq 2$ gives the first conclusion.

If X is absolutely regular then $\beta(a) \downarrow 0$ as $a \to \infty$. Choosing a subsequence a_m as in Lemma 3 we have $l_m = \lfloor m/(2a) \rfloor \to \infty$ and $l_m \beta(a_m) \to 0$. Substituting l_m for l and a_m for a above, the bound (4) will go to zero as $m \to \infty$, which proves the second conclusion ■

Now it is easy to prove the bounds in the introduction by applying Theorem 4 to the stationary operator-valued stochastic process

$$A_t = \alpha Q_{X_t} - (1-\alpha) Q_{\dot{X}_t}, \qquad (5)$$

which we reinterpret as a vector-valued process with values in the Hilbert space H_2 of Hilbert-Schmidt operators. Note that $T = \mathbb{E}[A_1]$ and $\hat{T} = (1/m)\sum_1^m A_i$.

Proof (of Theorem 2 and Theorem 3). : First note that $\beta_A(a) = \beta_X(a-1)$, because A_t depends also on X_{t-1}, and that A is absolutely regular if X is. Set $W = \mathcal{P}_d$ and define $V \subset H_2$ by

$$V = \{\alpha Q_x - (1-\alpha) Q_x : \|x\| \leq 1 \text{ and } \|y\| \leq 1\}.$$

Then $A_t \in V$ a.s. By Lemma 1 (i), V is contained in the unit ball in H_2 and

$$\begin{aligned} |\langle V, W\rangle_2| &= \sup_{P\in\mathcal{P}_d} \sup\{|\langle P, \alpha Q_x - (1-\alpha) Q_x\rangle_2| : \|x\| \leq 1, \ \ \|y\| \leq 1\} \\ &\leq \sup_{P\in\mathcal{P}_d} \sup\left\{\alpha \|Px\|^2 + (1-\alpha)\|Py\|^2\right\} \leq 1. \end{aligned}$$

By Lemma 1 (iv) $\|W\|_2 = \sqrt{d}$. We also have

$$\sup_{P\in\mathcal{P}_d} \left|\hat{L}(P) - L(P)\right| = \sup_{P\in\mathcal{P}_d} \left|\frac{1}{m}\sum_{i=1}^{m} \langle P, A_i - \mathbb{E}[A_1]\rangle_2\right|.$$

Applying Corollary 2 to the process $A_t - \mathbb{E}[A_1]$ gives both Theorem 2 and 3 ■.

4 A Generic Error Bound

Now we show that maximizing L minimizes an error-bound for all classification tasks posessing a certain continuity property. We fix a stationary process $\xi = \{\xi_t\}_{t\in\mathbb{Z}}$ with values in a measurable space (Ω, Σ), law μ and marginal distributions μ_I for $I \subset \mathbb{Z}$.

Definition 2. *Let ξ be as above. An (at most) countable partition $\Omega = \bigcup_k E_k$ of Ω into disjoint measurable E_k is continuous w.r.t. X if for all k and all $A, B \subseteq E_k$ we have*

$$\mu_{\{0\}}(A)\,\mu_{\{0\}}(B) \leq \mu_{\{0,1\}}(A, B).$$

So knowledge that E_k occurs at time 0 increases the probability at time 1 for any event A implying E_k. For an example let Ω be the unit interval, $\{E_k\}$ any partition of Ω into intervals of diameter less than $1/2$ and X_t a Gaussian random walk with periodic boundary conditions. Unlike the mixing properties relevant for generalization, the notion of continuity is concerned only with process dependencies on a microscopic time-scale.

We now assume that the process X has the form $X_t = \phi \circ \xi_t$, where $\phi : \Omega \to H$ is a a feature map with $\|\phi\| \leq 1/2$ and $\mathbb{E}[\phi \circ \xi_t] = 0$. One easily verifies $\beta_X(k) \leq \beta_\xi(k)$, for all k. The feature map ϕ may hide important information such as labels, for example if $\Omega = \mathcal{X} \times \mathcal{Y}$ and $\phi(x, y) = \psi(x)$.

Suppose now that $\{E_k\}$ is a partition of Ω, with each E_k defining some pattern class. Given a pair (ω_1, ω_2) drawn from $\mu^2_{\{0\}}$ we have to decide if ω_1 and ω_2 belong to the same class, that is to decide if there is some k such that $x \in E_k$

and $y \in E_k$. In the absence of other known structure we use a simple metric decision rule based on the projected input and the distance threshold $\sqrt{\alpha}$.

$$\omega_1 \text{ and } \omega_2 \text{ are in the same class iff } \left\|P\phi\left(\omega_1\right) - P\phi\left(\omega_2\right)\right\|^2 < \alpha.$$

Error bounds for this rule can be converted into error bounds for simple metric classifiers, whenever we are provided with examples for the various E_k.

Theorem 5. *With ξ, ϕ and X as above and $\alpha \in (0,1)$, if $\{E_k\}$ is continuous w.r.t. ξ, then the error probability for the above rule, as ω_1 and ω_2 are drawn independently from $\mu_{\{0\}}$, is bounded by*

$$Err \leq \frac{1}{1-\alpha}\left(1 - \frac{2}{\alpha}L_\alpha\left(P\right)\right) - R$$

where $R = \sum_k \left(\mu_{\{0\}}\left(E_k\right)\right)^2$.

The theorem implies a rule to select the trade-off parameter α: It should be chosen to minimize the first term in the bound above, so α should be close to 0, but a positive value for $L_\alpha\left(P\right)$ should still be obtained, corresponding to positive eigenvalues of the operator T.

Proof. We use the notation $\Delta = \Delta\left(\omega_1, \omega_2\right) := \left\|P\phi\left(\omega_1\right) - P\phi\left(\omega_2\right)\right\|^2$. Then

$$\begin{aligned}
Err &= \sum_{k,l:k\neq l} \mathbb{E}_{\mu^2_{\{0\}}}\left[1_{\Delta<\alpha} 1_{E_k\times E_l}\right] + \sum_k \mathbb{E}_{\mu^2_{\{0\}}}\left[1_{\Delta\geq\alpha} 1_{E_k\times E_k}\right] \\
&= \mathbb{E}_{\mu^2_{\{0\}}}\left[1_{\Delta<\alpha}\right] + 2\sum_k \mathbb{E}_{\mu^2_{\{0\}}}\left[1_{\Delta\geq\alpha} 1_{E_k\times E_k}\right] - R \\
&\leq \mathbb{E}_{\mu^2_{\{0\}}}\left[\frac{1-\Delta}{1-\alpha}\right] + 2\sum_k \mathbb{E}_{\mu^2_{\{0\}}}\left[\frac{\Delta}{\alpha}\ 1_{E_k\times E_k}\right] - R \\
&\leq \frac{1}{1-\alpha} - \frac{1}{1-\alpha}\mathbb{E}_{\mu^2_{\{0\}}}\left[\Delta\right] + \frac{2}{\alpha}\sum_k \mathbb{E}_{\mu_{\{0,1\}}}\left[\Delta\ 1_{E_k\times E_k}\right] - R.
\end{aligned}$$

The first inequality uses the bounds $1_{\Delta<\alpha} \leq \left(1-\Delta\right)/\left(1-\alpha\right)$ and $1_{\Delta\geq\alpha} \leq \Delta/\alpha$, which hold since $\Delta \in [0,1]$. The other inequality uses the continuity property of the E_k-system, because for any nonnegative function $g = g\left(\omega_1, \omega_2\right)$ and any k we have

$$\mathbb{E}_{\mu^2_{\{0\}}}\left[g\ 1_{E_k\times E_k}\right] \leq \mathbb{E}_{\mu_{\{0,1\}}}\left[g\ 1_{E_k\times E_k}\right],$$

as can be shown directly from Definition 2 by an approximation with simple functions. Now we use

$$\sum_k \mathbb{E}_{\mu_{\{0,1\}}}\left[\Delta\ 1_{E_k\times E_k}\right] \leq \mathbb{E}_{\mu_{\{0,1\}}}\left[\Delta\right] = \mathbb{E}\left[\left\|P\dot{X}_1\right\|^2\right] = \mathbb{E}\left[\left\|P\dot{X}_0\right\|^2\right]$$

and the identity $\mathbb{E}_{\mu^2_{\{0\}}}\left[\Delta\right] = 2\mathbb{E}\left[\left\|PX_0\right\|^2\right]$, which follows from the mean-zero assumption, to obtain

$$Err \leq \frac{1}{1-\alpha} - \frac{2}{1-\alpha}\mathbb{E}\left[\left\|PX_0\right\|^2\right] + \frac{2}{\alpha}\mathbb{E}\left[\left\|P\dot{X}_0\right\|^2\right] - R \ \blacksquare$$

5 An Online Algorithm

In practice H will be finite-dimensional. If the process X is slowly mixing, the learning time m can be quite large, leading to excessive storage requirements for any kind of batch algorithm. For this reason we used an online algorithm for principal subspace analysis, to which every successive realization of the operator valued variable $A_t = (1-\alpha) Q_{X_t} - \alpha Q_{\dot{X}_t}$ was fed, for $t = 1, ..., m$. This takes us somewhat astray from the results proved in this paper, and would require a different analysis in terms of stochastic approximation theory (see Benveniste et al [1]), an analysis which we cannot provide at this point. The principal goal of our first experiments was to test the value of our objective function L.

If $\mathbf{v} = (v_1, ..., v_d)$ is an orthonormal basis for the range of some $P \in \mathcal{P}_d$, the Oja-Karhunen flow [9], is given by the ordinary differential equation

$$\dot{v}_k = (I - P_{\mathbf{v}}) T v_k,$$

where $P_{\mathbf{v}}$ is the projection onto the span of the v_k. If T is symmetric it has been shown by Yan et al [18] that a solution $\mathbf{v}(t)$ to this differential equation will remain forever on the Stiefel-manifold of orthonormal sets if the initial condition is orthonormal, and that it will converge to a dominant eigenspace of T for almost all initial conditions. Discretizing gives the update rule

$$v_k(t+1) = v_k(t) + \eta(t) \left(I - P_{\mathbf{v}(t)}\right) T v_k(t),$$

where $\eta(t)$ is a learning rate. Unfortunately a careful analysis shows that the Stiefel manifold becomes unstable if T is not positive. The simplest solution to this problem lies in orthonormalization. This is what we do, but there are more elegant techniques and different flows have been proposed (see e.g. [4]) to extract dominant eigenspaces for general symmetric operators. We now replace $T = E[A_t]$ by the process variable A_t to obtain the final rule

$$v_k(t+1) = v_k(t) + \eta(t) \left(I - P_{\mathbf{v}(t)}\right) \left((1-\alpha) Q_{X_t} - \alpha Q_{\dot{X}_t}\right) v_k(t), \qquad (6)$$

which, together with the orthonormalization prescription, gives the algorithm used in our experiments. The update rule (6) can be considered a combination of Hebbian learning of input data with anti-Hebbian learning of input velocity.

6 Experiments

We applied our technique to train a preprocessor for image recognition. In all these experiments we used the output dimension $d = 10$, and the trade-off parameter $\alpha = 0.8$.

To train the algorithm we generated different input processes ξ to produce sequences of 28x28-pixel, gray-scale images, normalized to unity in the euclidean norm of $\mathbb{R}^{28\times 28}$. These processes are described below.

We considered two possible architectures for the preprocessor: In the linear case we used the pixel vectors directly as inputs to our algorithm, that is $X = \xi$ and $H = \mathbb{R}^{28\times 28}$.

In the nonlinear case (RBF) we used our algorithm to train the second layer of a two-layered radial-basis-function network. In an initial training phase a large number (2000) of prototypes π_i for the first layer were chosen from the process ξ at time intervals larger than the mixing time and kept fixed afterwards. Define a kernel κ on $\mathbb{R}^{28\times 28} \times \mathbb{R}^{28\times 28}$ by

$$\kappa(\zeta_1, \zeta_2) = \exp\left(-\beta \|\pi_j - \xi\|^2_{28\times 28}\right),$$

where in practice we always use $\beta = 4$. The first network layer then implements the (randomly chosen) nonlinear map $\tau : \mathbb{R}^{28\times 28} \to \mathbb{R}^{2000}$ given by

$$\tau(\xi)_k = \sum_{j=1}^{2000} G^{-1/2}_{kj} \kappa(\pi_j, \xi), \text{ for } \xi \in \mathbb{R}^{28\times 28},$$

where G is the Gramian $G_{ij} = \kappa(\pi_i, \pi_j)$, which is generically non-singular. The transformation through $G^{-1/2}_{kj}$ is chosen to ensure that $\langle \tau(\pi_i), \tau(\pi_j) \rangle_{2000} = \kappa(\pi_i, \pi_j)$. We then applied the algorithm to the output of the first layer, so $X = \tau(\xi)$ and $H = \mathbb{R}^{2000}$.

The processes are designed to train specific geometric invariants. Fix a large image I with periodic boundary conditions. At any time t the 28x28-process image ξ_t is a mapped subimage of I and completely described by four parameters: The position $\mathbf{x}_t = (x_t, y_t)$ of ξ_t within the source image, a rotation angle r_t and a scale s_t in the interval $[1/2, 3/2]$. We can thus write $\xi_t = \xi(\mathbf{x}_t, r_t, s_t)$ and we initialize to $\xi_0 = \xi(\mathbf{0}, 0, 1)$. Given ξ_t we find ξ_{t+1} by

$$\xi_{t+1} = \xi(\mathbf{x}_t + D\mathbf{x}, r_t + Dr, s_t + Ds),$$

where it is understood that the additions on $\mathbf{x}_t$ and r_t respect the periodic boundary conditions, and the addition on s_t restricts to the interval $[1/2, 3/2]$. The $D\mathbf{x}, Dr, Ds$ are random variables defining the essential geometric properties of the process. Here we report two cases, corresponding to the training of rotation and scale invariance. There were no experiments with translation invariance yet.

To train rotation invariance: The distribution of Dr is uniform on $[-\pi, \pi]$ and the distribution of Ds is uniform on $[-0.01, 0.01]$. Rapidly changing orientation, small changes in scale.

To train scale invariance: The distribution of Dr is uniform on $[-0.01, 0.01]$ and the distribution of Ds is uniform on $[-1, 1]$. Rapidly changing scale, small changes in orientation.

The choice of the distribution of $D\mathbf{x}$ is critical, with qualitative aspects of the exploration-exploitation dilemma. If we chose $\mathcal{N}(0, \sigma^2)$ (normal, centered with width σ) the centers of ξ will take a random walk with average stepsize σ. If σ is large (rapid exploration) the translation obliterates the effect of rotation

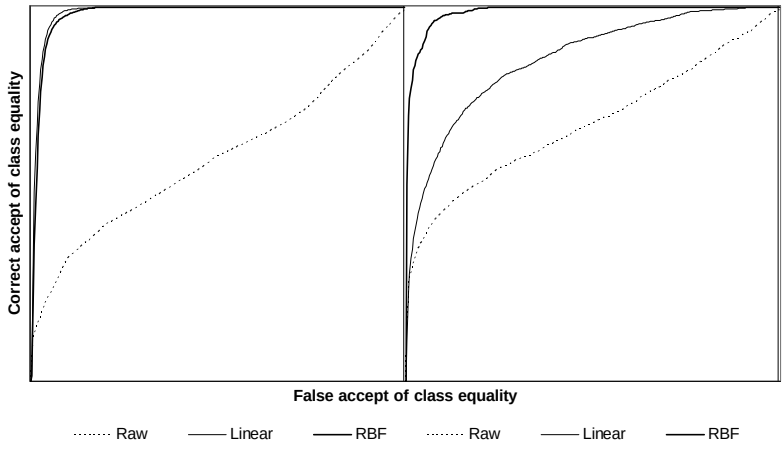

Fig. 1. ROC curves for the metric as a detector of class-equality for (left) rotation- and (right) scale-invariant character recognition.

or scaling, we loose continuity and the performance degrades. If σ is small (intense exploitation) the mixing time becomes large, causing exessive total learning times. We used $\sigma = 1/2$ in pixel units. With these parameter settings and a dynamic learning rate of $\eta(t) = \frac{10^2}{10^4+t}$ the system was trained on $m = 10^6$ observations.

The performance of the resulting preprocessors is tested on a real life problem, the rotation- (scale-)-invariant recognition of characters. To this end two test-sets were prepared containing images of the digits 0-8 (0-9) in 100 randomly chosen states of orientation (scaling between 1/2 and 3/2).

An important criterion for the quality of a preprocessor is the ability of the distance between preprocessed examples to serve as a detector for class-equality. Figure 1 shows corresponding receiver-operating-characteristics. The area under these curves then estimates the probability that for four independently drawn examples $\|a_1 - b_1\|_{10} \leq \|a_2 - b_2\|_{10}$, given that a_1 and b_1 belong to the same, and a_2 and b_2 to different classes. We also give a practical measure by recording the error rate of a *single-example-per-class* nearest-neighbour classifier, trained on a randomly selected example for each pattern class, *Error* in the following table.

Invariance Type	Method used	ROC-Area	Error
Rotation	Raw Data	0.597	0.716
	Linear	0.987	0.126
	RBF	0.983	0.138
Scaling	Raw Data	0.690	0.508
	Linear	0.866	0.421
	RBF	0.989	0.100

In the case of rotation invariance, the linear preprocessor architecture even slightly outperformed the RBF network. The latter showed stable good performance in both cases.

References

1. A. Benveniste, M. Métevier, Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer, 1987.
2. P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3: 194-200, 1991.
3. A. C. Lozano, S. R. Kulkarni, R. E. Shapire. Convergence and consistency of regularized boosting algorithms with stationary, β-mixing observations. *Advances in Neural Information Processing Systems* 18, 2006.
4. J.H. Manton, U. Helmke, I.M.Y. Mareels. A dual purpose principal and minor component flow. *Systems & Control Letters* 54: 759-769, 2005.
5. A. Maurer, Bounds for linear multi-task learning. *JMLR*, 7:117–139, 2006.
6. A. Maurer, Generalization Bounds for Subspace Selection and Hyperbolic PCA. *Subspace, Latent Structure and Feature Selection. LNCS* 3940: 185-197, Springer, 2006.
7. Colin McDiarmid, Concentration, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, p. 195-248. Springer, Berlin, 1998.
8. R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39, 5-34, 2000.
9. E. Oja. Principal component analysis. *The Handbook of Brain Theory and Neural Networks.* M. A. Arbib ed. MIT Press, 910-913, 2002.
10. S.Mika, B.Schölkopf, A.Smola, K.-R.Müller, M.Scholz and G.Rätsch. Kernel PCA and De-noising in Feature Spaces, in *Advances in Neural Information Processing Systems* 11, 1998.
11. J. Shawe-Taylor, N. Christianini, Estimating the moments of a random vector, *Proceedings of GRETSI 2003 Conference*, I: 47–52, 2003.
12. M. Reed, B. Simon. *Functional Analysis*, part I of *Methods of Mathematical Physics, Academic Press*, 1980.
13. E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants.* Springer 2000.
14. B. Simon. *Trace Ideals and Their Applications.* Cambridge University Press, London, 1979
15. J. Shawe-Taylor, C.K.I. Williams, N. Cristianini, J.S. Kandola: On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory* 51(7): 2510-2522, 2005.
16. M. Vidyasagar, *Learning and generalization with applications to neural networks.* Springer, London, 2003.
17. L. Wiskott, T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14: 715-770, 2003.
18. W. Yan, U. Helmke, J.B. Moore. Global analysis of Oja's flow for neural networks. *IEEE Trans. on Neural Networks* 5,5: 674-683, 1994.
19. B. Yu. Rate of convergence for empirical processes of stationary mixing sequences. Annals of Probability 22, 94-116, 1994.