Dimensionality reduction and generalization

Conference Paper · January 2007

DOI: 10.1145/1273496.1273579 · Source: DBLP

3 authors, including:

Lorenzo Rosasco
Massachusetts Institute of Technology
203 PUBLICATIONS **5,041** CITATIONS

SEE PROFILE

Alessandro Verri
Università degli Studi di Genova
203 PUBLICATIONS **9,187** CITATIONS

SEE PROFILE

Dimensionality Reduction and Generalization

Sofia Mosci MOSCI@DISI.UNIGE.IT
Difi & Disi, Università di Genova, via Dodecaneso 33, 16146 Genova, Italy

Lorenzo Rosasco ROSASCO@DISI.UNIGE.IT
Alessandro Verri VERRI@DISI.UNIGE.IT
Disi, Università di Genova, via Dodecaneso 35, 16146 Genova, Italy

Abstract

In this paper we investigate the regularization property of Kernel Principal Component Analysis (KPCA), by studying its application as a preprocessing step to supervised learning problems. We show that performing KPCA and then ordinary least squares on the projected data, a procedure known as kernel principal component regression (KPCR), is equivalent to spectral cut-off regularization, the regularization parameter being exactly the number of principal components to keep. Using probabilistic estimates for integral operators we can prove error estimates for KPCR and propose a parameter choice procedure allowing to prove consistency of the algorithm.

1. Introduction

Principal component analysis (Hastie et al., 2001) is a very common statistical tool for dimensionality reduction and in its linear version it consists in the projection of the data on the directions of highest variance, namely the principal components. A non-linear version of the same procedure, namely kernel principal component analysis (KPCA), has been also proposed in Scholkopf et al. (1999) where the projection is performed in a (possibly) high dimensional feature space hence enabling to exploit nonlinearity of the data. The free parameter in the algorithm is the number of components to keep and a fundamental question is then if it exists an "optimal" number of components for a given task.

This last question naturally leads to another question that is *how to measure* how effective is KPCA. If the reconstruction error is used as a criterion, recent results (Shawe-Taylor et al., 2004; Zwald et al., 2007) suggest that no such an optimal choice exists and the more components we keep the better.

On the other hand it is worth noting that one of the main uses of KPCA is as a preprocessing for supervised learning algorithms. In this case one might expect the dimensionality reduction step to influence also the generalization performance since some information is discarded. Going further one might ask if, after KPCA, any kind of regularization is needed at all since again some shrinking of the available information already occurred (see Scholkopf et al. (1999) and the discussion in Blanchard et al. (2004)). These issues have been recently addressed in Blanchard et al. (2004) where an algorithm, called kernel projection machine, was proposed which essentially amounts to a KPCA step and then empirical risk minimization with hinge loss function on the projected data. Note that ERM is unpenalized and the only free-parameter is the number of components in the dimensionality reduction. In this case the authors empirically show that indeed an optimal number of components exists when we look at how the generalization performance depends on the dimensionality reduction procedure. As a byproduct they also argued that using some further regularization, for example support vector machines, after KPCA is somewhat redundant and not really necessary.

The main idea of our study is to give a proof of such empirical evidences. In fact considering a similar approach, with the hinge loss replaced by the square loss, we can follow Bauer et al. (2006) and prove that indeed the number m of principal components kept *is* a regularization parameter and that an optimal parameter choice exists. As a main mathematical tool we use estimates of integral operators based on vector valued law of large numbers, to derive the probabilis-

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

tic error estimates which are the keys to understand the role of m. Indeed such error estimates are made by two error terms, sample and approximation errors, and the best choice for m is the one balancing out the two terms. Our results are indebted to Bauer et al. (2006) and results of a similar flavor can also be found in Massart et al. (1999) for the case of regression with Gaussian white noise and fixed design (see also Blanchard et al. (2004)). Our analysis is developed in the usual context of statistical learning where the design is random and for the sake of simplicity we consider bounded outputs, though more general kind of noise-such as sub-Gaussian noise- can also be treated. A related analysis can be found in Zhang (2005) who considers empirical risk minimization in a reproducing kernel Hilbert space. Indeed the results in such paper show that the number of principal components controls the performance of the algorithm yet the subject of model selection via sample/approximation trade-off is not considered. In this view we specialize and further develop the above reasoning giving explicit parameter choices leading to consistency. We also note that the bound we obtain conveys the correct qualitative behavior of the error w.r.t. the number of components and can be shown to be essentially optimal under the given assumptions (cfr. Caponnetto & De Vito, 2006). Nonetheless since the bound is basically distribution independent in typical applications it will be too pessimistic. For this reason in the experiments section we will consider data-driven model selection via cross validation.

The use of principal component analysis for regression is standard in classical statistics where finite dimensional linear models are usually considered. The kernel extension to the non linear case makes apparent the relation with other algorithms such as SVM or regularized least squares, still theoretical results on regularization property of principal component regression in the learning setting were lacking so far. In the usual statistical setting the input points are fixed so that the problem of sample complexity is usually not taken into account.

Interestingly the algorithm we consider, which was recently proposed in Rosipal et al. (2000) and called kernel principal component regression, can be shown to be mathematically equivalent to truncated singular value decomposition or spectral cut-off, which is possibly the most famous regularization scheme for linear ill-posed problems.

Though we just considered the supervised case the conclusions we draw can be of interest in the context of semi-supervised learning since recently proposed techniques are based on the use of the principal components of data driven kernel for function approximation (Belkin & Niyogi, 2004; Coifman & Lafon, 2006). The extension of our analysis in the case where unlabeled data are available is an interesting direction for future work.

The plan of the paper follows. In the next section we introduce some notation and background on learning theory. In section 3 we prove the equivalence between principal component regression and truncated singular value decomposition (Engl et al., 1996), by showing with simple algebraic tools that the solution of the two algorithms are point-wise equal. In section 4 we derive error estimates and prove the existence of an optimal choice of the regularization parameter for KPCR as well as consistency. In section 5 we present some numerical experiments on real and simulated data that confirm theoretical results from the previous section. In the last section we discuss our results.

2. Setting

We consider the setting of supervised learning where we have to find an unknown input-output relation given a finite number of input-output instances. More precisely we assume that a training set $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = (x_1, y_1), \dots, (x_n, y_n)$ is sampled according to an unknown distribution $\rho(x, y) = \rho(y|x)\rho_X(x)$, where $x \in X \subset \mathbb{R}^d$ and $y \in Y = [-M, M] \subset \mathbb{R}$. The idea is to find a function f such that $f(x) \sim y$ and, considering least squares, this can be formalized saying that we look for a function with small expected error

$$\mathcal{E}(f) := \int_{X \times Y} (y - f(x))^2 d\rho(x, y).$$

Among all measurable functions one can easily show that the one that minimizes the expected error is the regression function $f_\rho := \int_Y y d\rho(x, y)$. Then, given a training set $\mathbf{z}$, the goal is to build an estimator $f_\mathbf{z}$ whose error is close to $\mathcal{E}(f_\rho)$. In particular a first important property is (weak) consistency

$$\lim_{n \to \infty} \Pr\left(\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho) \geq \epsilon \right) = 0 \qquad \forall \epsilon > 0$$

ensuring that, if we have enough data, we can eventually reach the best possible solutions for any probability distribution. A second crucial property concerns rate for the above convergence and is typically studied via probabilistic error estimates such that with probability at least $1 - \eta$

$$\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho) \leq \varepsilon(n, \eta) \tag{1}$$

where $\varepsilon(n, \eta)$ is a suitable bound depending on the number of samples and the confidence.

We also recall that for classification problems rather than the expected error we want to estimate the misclassification error

$$R(f) := Pr(yf(x) < 0)$$

whose minimizer R^*, namely the Bayes risk, is achieved by the Bayes rule

$$b(x) = \begin{cases} +1 & \text{if} \quad p(1|x) > \frac{1}{2} \\ -1 & \text{if} \quad p(1|x) \leq \frac{1}{2}. \end{cases} \tag{2}$$

In this case we wish to find a classification rule such that with probability at least $1 - \eta$ we have

$$R(f_{\mathbf{z}}) - R^* \leq \varepsilon(n, \eta) \tag{3}$$

and derive convergence of the misclassification error of our classification rule to the Bayes risk, namely Bayes consistency. Finally, we note that considering least squares estimates, a plug-in classification rule can be obtained taking $sign f_{\mathbf{z}}$ and moreover, since y is ±1, we get $f_\rho(x) = 2\rho(1|x) - 1$ so that the bayes rule is simply $sign f_\rho$. Interestingly the error measured via expected error and the misclassification error are related (Bartlett et al., 2003)

$$R(f) - R^* \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho)} \tag{4}$$

so that consistency w.r.t. to expected errors implies Bayes consistency.

2.1. Learning with Kernels

The search for possible solutions is often restricted to an hypotheses space $\mathcal{H}$. In the following we consider hypotheses spaces that are reproducing kernel Hilbert (RKH) spaces (Aronszajn, 1950). Recall that these are Hilbert spaces of functions which are completely determined by a symmetric positive definite function $K(x, s)$. In particular we make use of the following well-known properties:

- reproducing property: for $f \in \mathcal{H}$ it holds

$$f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}} ; \tag{5}$$

- feature map: we can consider a mapping $\Phi : X \to \mathcal{H}$ which can be seen as a data parameterization related to the kernel through the following equality

$$\langle \Phi(x), \Phi(s) \rangle_{\mathcal{H}} = K(x, s), \quad x, s \in X.$$

For technical reasons we will assume the kernel to be continuous and bounded, i.e.

$$\kappa^2 = \sup_{x \in X} K(x, x) < \infty.$$

It is interesting to recall the derivation of the solution to empirical risk minimization (ERM) algorithm

$$f_{\mathbf{z}} = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2, \tag{6}$$

when $\mathcal{H}$ is a RKH space. If we consider the feature map

$$\Phi(x) = K(x, \cdot) =: K_x$$

the function in $\mathcal{H}$ can be written as $f(x) = \langle w, \Phi(x) \rangle$ and we can simply differentiate the empirical risk with respect to w to get a normal equation

$$\frac{1}{n} \sum_{i=1}^{n} \langle w, \Phi(x_i) \rangle_{\mathcal{H}} \Phi(x_i) = \frac{1}{n} \sum_{i=1}^{n} y_i \Phi(x_i).$$

Interestingly if the data are centered then we have that

$$T_{\mathbf{x}} := \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \otimes \Phi(x_i) = \frac{1}{n} \sum_{i=1}^{n} \langle \cdot, \Phi(x_i) \rangle_{\mathcal{H}} \Phi(x_i) \tag{7}$$

is simply the (uncentered) covariance operator and the solution can be written as

$$w = T_{\mathbf{x}}^{\dagger} h_{\mathbf{z}} \tag{8}$$

with $h_{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^{n} y_i \Phi(x_i)$ and $T_{\mathbf{x}}^{\dagger}$ denotes the generalized inverse of the covariance operator. We use this equation extensively in the next section, but we note here that from a practical point of view when the Hilbert space is *not* finite dimensional one usually prefers to use the fact that the solution can also be written as

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$$

where $\alpha = \mathbf{K}^{\dagger} \mathbf{y}$ and $\mathbf{K}^{\dagger}$ is the generalized inverse of the kernel matrix, $[\mathbf{K}]_{ij} = K(x_i, x_j)$.

3. Principal Component Regression and Spectral Cut-Off

In this section we show the equivalence between principal component regression (Rosipal et al., 2000) and the regularization algorithm known as spectral cut-off or truncated singular value decomposition (TSVD) (Engl et al., 1996). First, we briefly recall the principal component regression algorithm, or rather its *kernel* version. Second we review TSVD regularization. Third we discuss a straightforward connection between the two.

We previously note that under our assumptions the covariance operator in the feature space is known to be positive and self-adjoint. In particular we let $(\sigma_i, v_i)_{i \in I}$ be the associated eigensystem[1]. We will assume throughout the data to be centered in the feature space so that the v_i's are the principal components.

[1] We always assume the eigenvalues to be arranged in decreasing order.

Remark 1. *When the data are not centered we cannot asses the equivalence between principal component regression and truncated singular value decomposition unless we consider a modified kernel which corresponds to the features covariance operator*

$$T_{\mathbf{x}} \to \hat{T}_{\mathbf{x}} = (I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n)T_{\mathbf{x}}(I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n).$$

Spectral cut-off on the non recentered kernel is still an efficient algorithm but it is not evident its connection with principal component analysis because the eigenvectors of $T_{\mathbf{x}}$ and $\hat{T}_{\mathbf{x}}$ may be different.

Again we note that from the computational point of view rather than working with $T_{\mathbf{x}}$ one usually considers the kernel matrix since it can be shown that they share the same spectrum and their eigenfunctions/eigenvectors are related. For theoretical purposes it is convenient to consider simply $T_{\mathbf{x}}$.

Kernel Principal component regression can be seen as a two steps algorithm: the first step amounts to an unsupervised dimensionality reduction via (kernel) principal component analysis and the second step is simply ERM on the projected data. As it is often done in practice we control the projection of the data choosing a threshold λ on the magnitude of the eigenvalues. In other words we only keep $m = m(\lambda)$ components corresponding to eigenvalues bigger than λ. We will show in the following, that such a threshold plays the role of regularization parameter controlling the complexity of the KPCR solution. More in details KPCR can be described in the following steps:

1. decomposition of $T_{\mathbf{x}}$ to obtain $(\sigma_i; v_i)$;
2. projection of the data on the first m components such that $\sigma_m > \lambda$ for fixed $\lambda > 0$,
$$\Phi(x) \to \vec{\varphi}^{\,m}(x) = \sum_{j=1}^{m} \langle \Phi(x), v_j\rangle \, \vec{e_j}$$
where $\vec{\varphi}^{\,m}(x) \in \mathbb{R}^m$ and $(\vec{e_j})_j$ is a canonical basis in $\mathbb{R}^m$;
3. ERM, $$\min_{\vec{w}\in\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n}(y_i - \vec{w}\cdot\vec{\varphi}^{\,m}(x_i))^2$$
whose solution is given by $\vec{w} \in \mathbb{R}^m$
$$\begin{aligned}\vec{w} &= \sum_{j=1}^{m}([(\hat{\varphi}^{m})^{\,T}\hat{\varphi}^{m}]^{\dagger}(\hat{\varphi}^{m})^{\,T}\mathbf{y})_j\,\vec{e}_j = \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{y_i}{\sigma_j}\langle\Phi(x_i), v_j\rangle_{\mathcal{H}}\,\vec{e}_j\end{aligned}$$
where $[\hat{\varphi}^{m}]_{ij} = \vec{\varphi}_j^{\,m}(x_i)$ and $[(\hat{\varphi}^{m})^{\,T}\hat{\varphi}^{m}]_{ij} = \sigma_i\delta_{ij}$ (by the definition of (σ_i, v_i)).

The KPCR solution can then be written as

$$f_{\mathbf{z}}^{(PCR)}(x) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{y_i}{\sigma_j}\langle\Phi(x_i), v_j\rangle_{\mathcal{H}}\langle\Phi(x), v_j\rangle_{\mathcal{H}}\,.$$

We emphasize that in this case the solution $\vec{w}$ is an m dimensional vector.

To describe the **spectral cut-off regularization** it is convenient to remember that from the formulation of ERM in the feature space we can rewrite the solution (8) on the spectrum of $T_{\mathbf{x}}$ to get

$$w = \sum_{j=1}^{\infty}\sum_{i=1}^{n}\frac{y_i}{\sigma_j}\langle\Phi(x_i), v_j\rangle_{\mathcal{H}}\,v_j.$$

The above problem is possibly ill-posed (Engl et al., 1996) and the TSVD regularization simply cuts-off unstable components, that is only $m = m(\lambda)$ components are kept corresponding to eigenvalues bigger than λ. This way we get $w^m \in \mathcal{H}$ such that

$$w^m = \sum_{j=1}^{m}\sum_{=i}^{n}\frac{y_i}{\sigma_j}\langle\Phi(x_i), v_j\rangle_{\mathcal{H}}\,v_j. \qquad (9)$$

We emphasize that in this case w^m is a function in a possibly infinite dimensional space. The solution can then be written as

$$f_{\mathbf{z}}^{(TSVD)}(x) = \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{y_i}{\sigma_j}\langle\Phi(x_i), v_j\rangle_{\mathcal{H}}\langle\Phi(x), v_j\rangle_{\mathcal{H}}$$

which shows that the solution of principal component regression and spectral cut-off are point-wise equal. The theory of RKH spaces ensures that the obtained solutions are identical, in fact for any $g, f \in \mathcal{H}$ the reproducing property (5) ensures

$$f(x) = g(x) \;\; \forall \;\; x \quad \Leftrightarrow \quad \langle f - g, K_x\rangle_{\mathcal{H}} = 0 \;\; \forall \;\; x$$

and this implies that f and g are the same function.

4. Dimensionality Reduction and Generalization

In this section we prove that if $f_\rho \in \mathcal{H}$ we can derive error estimates of the form (4) as well as consistency (and Bayes consistency) of KPCR. Alternatively one should replace f_ρ with the best in the model $f_{\mathcal{H}} = \min_{f\in\mathcal{H}}\mathcal{E}(f)$ (see also Bauer et al. (2006)). To this aim we note that the parameter we have to choose is the threshold λ on the eigenvalues so that it is convenient to use the notation $f_{\mathbf{z}}^{\lambda}$ in place of $f_{\mathbf{z}}^{(PCR)}$.

Theorem 1. *We let $n \in \mathbb{N}$ and $0 < \eta \leq 1$. Moreover we assume that $f_\rho \in \mathcal{H}$ and $\|f_\rho\|_{\mathcal{H}} \leq R$. Then with probability at least $1 - \eta$ we have*

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \mathcal{E}(f_\rho) \leq \frac{16\sqrt{2}}{\sqrt{n}}(\kappa^2R^2 + (M+R)^2)\log\frac{4}{\eta} \qquad (10)$$

where we choose

$$\lambda_n = \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa^2 \log\frac{4}{\eta}.$$

We give the proof in the next section and add some comments. As we previously mentioned, an important consequence of theorem 1 is the existence of an optimal value m_n for the number of principal components which depends on the size of the training set and corresponds to the optimal choice for the parameter λ, that is $m_n = m(\lambda_n)$. At first sight this may appear in contrast with the results in Zwald et al. (2007), where the authors discuss the behavior of the true reconstruction error which should decrease with the number of dimensions D, the parameter m in our conventions. The reason for this apparent contrast is due to the fact that the reconstruction error quantifies the effect of PCA in an unsupervised setting whereas our bound is on the expected error of a supervised problem. Indeed in a supervised setting if we keep too few components we are oversmoothing whereas if we add too many of them we risk to incur into overfitting thus spoiling the generalization performance.

This result may look similar to the error bound presented in Massart et al. (1999) and recalled in Blanchard et al. (2004) where the authors investigate the effect of regularization performed by (kernel) PCA through dimensionality reduction. However it can be noted that such result deals with Gaussian white noise regression in a fixed design setting, whereas we consider random design.

Finally as a direct consequence theorem 1 leads to weak consistency for spectral cut-off regularization in regression

$$\lim_{n\to\infty} \Pr\left(\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \mathcal{E}(f_\rho) \geq \epsilon\right) = 0 \qquad \forall \epsilon > 0$$

and in classification

$$\lim_{n\to\infty} \Pr\left(R(f_{\mathbf{z}}^{\lambda_n}) - R^* \geq \epsilon\right) = 0 \qquad \forall \epsilon > 0$$

where we used (4).

4.1. Proof of the Error Estimates

In this section we give the proof of the main results. We follow the same approach as in Bauer et al. (2006) but the proofs adapted to our setting are considerably simplified. We previously need some notation and facts. First we note that, comparing the ERM solution (8) with (9), we can rewrite the solution of KPCR as

$$f_{\mathbf{z}}^\lambda = g_\lambda(T_{\mathbf{x}}) h_{\mathbf{z}}$$

where g_λ can be seen via spectral theory as a function on the spectrum of $T_{\mathbf{x}}$ such that $g_\lambda(\sigma) = \frac{1}{\sigma}$ if $\sigma \geq \lambda$ and 0 otherwise. Second, we denote with

$$T := \int_X \langle \cdot, \Phi(x)\rangle \Phi(x) d\rho_X(x) = \mathbf{E}[T_{\mathbf{x}}]$$

the expected covariance operator and we also denote with

$$h = T_{\mathbf{x}} f_\rho \tag{11}$$

Third we recall the following lemma from Caponnetto and De Vito (2006).

Lemma 1. *Let* $\kappa = \sup_{x\in X} \|K_x\|_{\mathcal{H}}$, $\|f_\rho\|_{\mathcal{H}} \leq R$ *and* $y \in [-M, M]$. *For* $0 < \eta \leq 1$ *and* $n \in \mathbb{N}$ *let*

$$G_\eta = \{\mathbf{z} \in (X\times Y)^n : \|h - h_{\mathbf{z}}\|_{\mathcal{H}} \leq \delta_1, \quad \|T - T_{\mathbf{x}}\| \leq \delta_2\},$$

with

$$\begin{aligned} \delta_1 := \delta_1(n,\eta) &= \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa(M+R)\log\frac{4}{\eta} \\ \delta_2 := \delta_2(n,\eta) &= \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa^2 \log\frac{4}{\eta}. \end{aligned}$$

then

$$\Pr(G_\eta) \geq 1 - \eta.$$

Recalling De Vito et al. (2005) that we have

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \left\|\sqrt{T}(f - f_\rho)\right\|_{\mathcal{H}}^2 \tag{12}$$

for all $f \in \mathcal{H}$, in order to prove theorem Thm. 1 we first derive a bound on $\left\|\sqrt{T}(f_{\mathbf{z}}^\lambda - f_\rho)\right\|_{\mathcal{H}}^2$ for fixed λ (Thm. 2) and then choose the value $\lambda_n = \lambda(n)$ optimizing the bound

Theorem 2. *We let* $n \in \mathbb{N}$ *and* $0 < \eta \leq 1$. *We assume that* $\lambda < 1$ *and*

$$\lambda \geq \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa^2 \log\frac{4}{\eta}. \tag{13}$$

Moreover we assume that $f_\rho \in \mathcal{H}$ *and* $\|f_\rho\|_{\mathcal{H}} \leq R$. *Then with probability at least* $1 - \eta$ *we have*

$$\mathcal{E}(f_{\mathbf{z}}^\lambda) - \mathcal{E}(f_\rho) \leq 8(\lambda R^2 + \frac{C}{\lambda n}) \tag{14}$$

where $C = C(\eta, \kappa, M, R) = 8\kappa^2(M+R)^2(\log\frac{4}{\eta})^2$ *does not depend on* λ *and* n.

Proof of Thm. 2. In this proof we use the inequalities in the above lemma which holds with probability at least $1-\eta$ with $0 < \eta \leq 1$. Recalling (12), we consider the following error decomposition

$$\begin{aligned} &\left\|\sqrt{T}(f_{\mathbf{z}}^\lambda - f_\rho)\right\|_{\mathcal{H}}^2 \leq \qquad (15) \\ &\leq 2\left\|\sqrt{T}(f_{\mathbf{z}}^\lambda - f^\lambda)\right\|_{\mathcal{H}}^2 + 2\left\|\sqrt{T}(f^\lambda - f_\rho)\right\|_{\mathcal{H}}^2 \end{aligned}$$

where

$$f^\lambda = g_\lambda(T_{\mathbf{x}})h \quad \text{with } h \text{ given by } \;(11)\;.$$

We now separately bound the two terms in the right-hand side. The first term can be decomposed as

$$\begin{aligned}\sqrt{T}(f^\lambda_{\mathbf{z}} - f^\lambda) = \sqrt{T}g_\lambda(T_{\mathbf{x}})(h_{\mathbf{z}} - h) = \qquad (16)\\ = \sqrt{T_{\mathbf{x}}}g_\lambda(T_{\mathbf{x}})(h_{\mathbf{z}} - h) + (\sqrt{T} - \sqrt{T_{\mathbf{x}}})g_\lambda(T_{\mathbf{x}})(h_{\mathbf{z}} - h).\end{aligned}$$

We note that the inequality

$$\left\| \sqrt{T} - \sqrt{T_{\mathbf{x}}} \right\| \leq \sqrt{\|T - T_{\mathbf{x}}\|} \leq \sqrt{\delta_2} \leq \sqrt{\lambda} \tag{17}$$

follows from Theorem 8.1 in Mathe and Pereverzev (2002), lemma 1 and Ass. (13).
Moreover from the definition of operator norm and standard results of spectral theory

$$\|g(\mathrm{A})\| = \sup_{\sigma \in \ \Lambda(\mathrm{A})} g(\sigma) \tag{18}$$

where $\Lambda(\mathrm{A})$ is the set of the eigenvalues of the operator $\mathrm{A} : \mathcal{H} \to \mathcal{H}$, it is easy to see that

$$\|g_\lambda(T_{\mathbf{x}})\| \leq \frac{1}{\lambda} \qquad \left\| \sqrt{T_{\mathbf{x}}} g_\lambda(T_{\mathbf{x}}) \right\| \leq \frac{1}{\sqrt{\lambda}}.$$

If we now take the norm in (16) we get

$$\left\| \sqrt{T}(f^\lambda_{\mathbf{z}} - f^\lambda) \right\|_{\mathcal{H}} \leq \frac{2}{\sqrt{\lambda}} \|h_{\mathbf{z}} - h\|_{\mathcal{H}} \leq \frac{2}{\sqrt{\lambda}} \delta_1. \tag{19}$$

We now deal with the second term in the r.h.s. of (15). We can write

$$\begin{aligned}\sqrt{T}(f^\lambda - f_\rho) &= \sqrt{T}(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})f_\rho \\ &= \sqrt{T_{\mathbf{x}}}(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})f_\rho + \qquad (20)\\ &\quad + (\sqrt{T} - \sqrt{T_{\mathbf{x}}})(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})f_\rho.\end{aligned}$$

We can bound this term recalling that by assumption $\|f_\rho\|_{\mathcal{H}} \leq R$ and noting that definition (18) implies

$$\|I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}\| \leq 1 \quad \text{and} \quad \left\|(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})\sqrt{T_{\mathbf{x}}}\right\| \leq \sqrt{\lambda}.$$

We note that operator $g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}$ is exactly the projection operator on the subspace spanned by the eigenvectors of $T_{\mathbf{x}}$ with eigenvalue greater or equal to λ, whereas $I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}$ is the projection operator on the orthogonal subspace. We can now take the norm of (20) and use (17) to get

$$\left\| \sqrt{T}(f^\lambda - f_\rho) \right\|_{\mathcal{H}} \leq 2\sqrt{\lambda}R. \tag{21}$$

The estimate in (14) follows plugging (21) and (19) into (15) and using the definition of δ_1. □

We are now ready to give the proof of Thm. 1.

Proof of Thm. 1. The proof of the theorem is straightforward. In fact since the sample error increases with λ while the approximation error decreases, in order to get the best error we should take the value of λ which gives a good trade-off between the two terms. To this end we set the two terms to be of the same order

$$\lambda_n = \frac{1}{\lambda_n n} \quad \Rightarrow \quad \lambda_n = O(\frac{1}{\sqrt{n}}).$$

Then, in order to be consistent with condition (13), we can choose the following value for λ_n

$$\lambda_n = \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa^2 \log \frac{4}{\eta}.$$

Substituting λ_n in (14) we obtain the rate (10). □

5. Numerical Experiments

In this section we present some numerical results to illustrate the behavior of principal component regression on real and simulated data.

The real data experiments have been carried out on two datasets available at `http://www.ics.uci.edu/~mlearn/MLSummary.html`. In the first one we analyzed the Wisconsin diagnostic breast cancer database on benign vs malignant classification. The dataset is made of $n = 569$ examples divided in two classes and described by $d = 30$ features. In the second experiment we examined the SPECTF heart database. This dataset is made of $n = 267$ instances (patients) and $d = 44$ attributes per instance. Each of the patients is classified into two categories: normal and abnormal.

In both experiments we first partitioned the dataset in two balanced subsets, training and test set. As for the parameter choice, despite optimality of the bound in practice it is going to be too pessimistic to be used with few examples. Indeed the theoretical results of sec. 4 highlight the regularization role of the number of dimensions but yet in practice we often need some data driven procedure, such as cross-validation, to choose it. Therefore we determined an optimal value for the regularization parameter via 5-fold cross validation on the training set. For each value of the parameter we estimated the average misclassification error and the median number of principal components that survived the thresholding. Hence we obtained a curve describing an estimate of the expected error as a function of either the regularization parameter λ or the corresponding number of selected components m (see figure 1), choosing the optimal value of the parameter λ_n and m_n as the minimum of such curve. Finally we run the

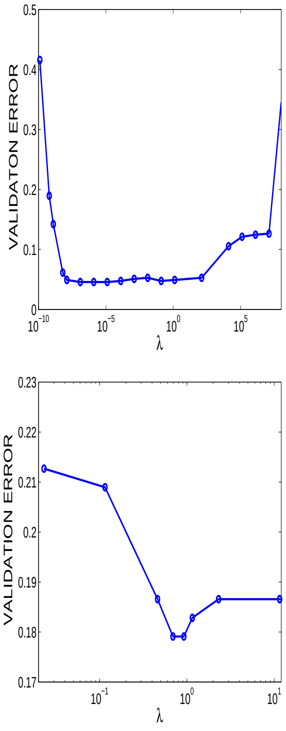

Figure 1. 5-fold cross validation error vs λ and m for the brain cancer dataset(above) and for the SPECTF heart dataset(below)

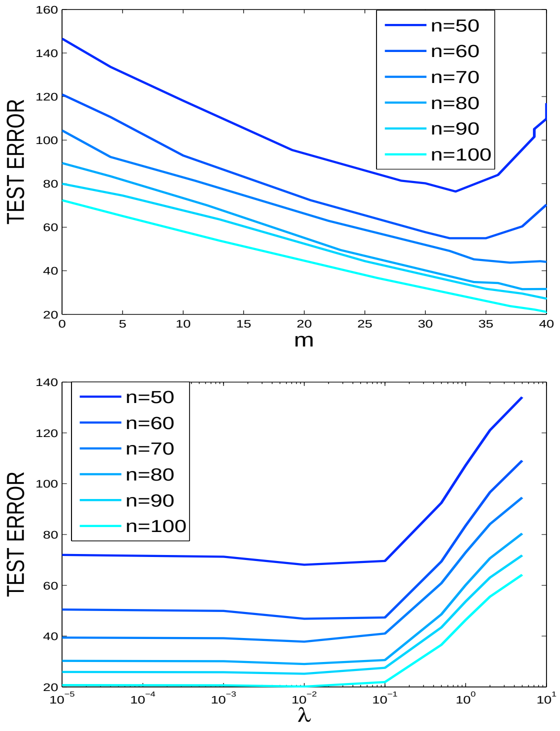

Figure 2. test error vs λ of TSVD(above) and of TSVD with regularized Least Squares(below) for the toy example

algorithm on the entire training set with the value for λ_n provided by the 5-fold cross validation, and computed the misclassification error on the test data. In order to obtain a more precise estimate of the test error we repeated the entire protocol for 50 different splits of the total dataset in training and test set and averaged the results on these repetitions.

Comparisons with the original results from these two data sets show a lower prediction accuracy (96% against 97.5%) for the breast cancer data set and a higher prediction accuracy (80% against 77%) for the SPECTF data set. However the main purpose of these experiments has been to empirically demonstrate the possibility of choosing an optimal value for the number of components rather than searching for an accurate predictor. In fact, figure 1 clearly indicates the existence of an optimal value for the threshold which corresponds to an optimal number of principal components to be used in the determination of the classifier. Taking into account more than m_n components can only increase the prediction error. We also investigated the effect of spectral cut-off on a toy example based on a Gaussian linear regression model $y = \beta x + \epsilon$, where $x \in \mathbb{R}^d$ and $d = 40$. We run the algorithm on training sets of increasing number of samples with different values of the parameter λ and evaluated the error on a test set of 5000 instances. As expected figure 2 clearly shows that m_n, corresponding to the minimum of the test error for different n, reaches the maximum number of components only for large data sets, whereas a limited number of training instances is better generalized with a limited number of principal components. In order to better understand the effect of further regularization after KPCR, we evaluated the test error committed by regularized least squares(RLS) on the first m_{opt} principal components. From figure 2 we can see that RLS do not improve prediction performance since the test error is always approximately equal or greater than the error committed with just spectral cut-off($\lambda = 0$).

6. Conclusions

In this paper we have shown the equivalence between principal component regression and spectral cut-off, observing that the solutions of the two algorithms are point-wise equal. Moreover we have emphasize the fact that principal component analysis, as a preprocessing step in supervised learning, is itself a regularization step and does not need any further regularization. Indeed (unpenalized) empirical risk minimization on the projected data does not incur in overfitting if the projection step is suitably tuned.

We also observe that even though in principal component regression the empirical risk minimization algorithm deals with shorter vectors, that is the m-dimensional projection of the data, most of the computation is performed in the preprocessing step which projects the data on the m principal component; therefore the benefit of dealing with smaller matrices is paid with the drawback of the computationally demanding projection. On the other hand, truncated singular value decomposition deals with possibly infinite dimensional vectors, but all the computation is confined to the construction and diagonalization of the covariance matrix or its dual kernel matrix.

Another important observation can be done on the choice of the the number of dimensions to keep in the dimensionality reduction step. In fact, we have shown that, when KPCA is used as a preprocessing to a supervised learning task, there exists an optimal value λ_n for the threshold on the eigenvalues and hence a corresponding optimal number of dimensions $m_n = m(\lambda_n)$ to be used in the projection step. This result apparently goes against the intuition that adding more dimensions, and therefore more information from the distribution, the result should improve. Indeed such an intuition is misleading when the data are finitely sampled from a probability distribution; in fact, from (14), we can observe that, when m increase (that is λ decreases), the approximation error decreases, but the sample error increases. Therefore an optimal number of dimensions, m_n, exists which depends on the size of the training set and such that using more than m_n dimensions will cause a decrease in the predicting power of the solution.

Acknowledgments

This work has been partially supported by the FIRB project LEAP RBIN04PARL and by the EU Integrated Project Health-e-Child IST-2004-027749.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, *68*, 337–404.

Bartlett, P., Jordan, M., & McAuliffe, J. (2003). *Convexity, classification, and risk bounds* (Technical Report 638). Department of Statistics, U.C. Berkeley.

Bauer, F., Pereverzev, S., & Rosasco, L. (2006). On regularization algorithms in learning theory. *Journal of complexity*, *23*, 52–57.

Belkin, M., & Niyogi, P. (2004). Semi-supervised learning on riemannian manifolds. *Machine Learning*, *56*, 209–239.

Blanchard, G., Massart, P., Vert, R., & Zwald, L. (2004). Kernel projection machine: a new tool for pattern recognition. *NIPS 2004* (pp. 1649–1656).

Caponnetto, A., & De Vito, E. (2006). Optimal rates for regularized least-squares algorithm. *Found. Comput. Math. In Press, DOI 10.1007/s10208-006-0196-8. Online August 2006.*

Coifman, R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets*, *21*, 5–30.

De Vito, E., Rosasco, L., Caponnetto, A., Giovannini, U. D., & Odone, F. (2005). Learning from examples as an inverse problem. *Journal of Machine Learning Research*, *6*, 883–904.

Engl, H., Hanke, M., & Neubauer, A. (1996). Regularization of inverse problems. *Mathematics and its Applications*, *375*.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer-Verlag.

Massart, P., Barron, A., & Birge, L. (1999). Risk bounds for model selection via penalization. *Proba.Theory Relat.Fields*, *113*, 301–413.

Mathe, P., & Pereverzev, S. (2002). Moduli of continuity for operator monotone functions. *Numerical Functional Analysis and Optimization*, *23*, 623–631.

Rosipal, R., Trejo, L., & Cichocki, A. (2000). *Kernel principal component regression with em approach to nonlinear principal components extraction* (Technical Report). CIS, University of Paisley.

Scholkopf, B., Smola, A., & Muller, K. (1999). Kernel principal component analysis. In *Advances in kernel methods - support vector learning*, 327–352. MIT Press.

Shawe-Taylor, J., Williams, C., Cristianini, N., & Kandola, J. (2004). On the eigenspectrum of the gram matrix and the generalisation error of kernel pca. *IEEE Transactions on Information Theory 51* (pp. 2510–2512).

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, *17*, 2077–2098.

Zwald, L., Bousquet, O., & Blanchard, G. (2007). Statistical properties of kernel principal component analyis. *Machine Learning*, *66*, 259–294.