JMLR: Workshop and Conference Proceedings 44 (2015) 19-29 NIPS 2015

The 1st International Workshop "Feature Extraction: Modern Questions and Challenges"

A Dimension-Independent Generalization Bound for Kernel Supervised Principal Component Analysis

Hassan Ashtiani MHZOKAEI@UWATERLOO.CA
School of Computer Science
University of Waterloo
Waterloo, ON, Canada

Ali Ghodsi AGHODSIB@UWATERLOO.CA
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, ON, Canada

Editor: Dmitry Storcheus

Abstract

Kernel supervised principal component analysis (KSPCA) is a computationally efficient supervised feature extraction method that can learn non-linear transformations. We start the study of the statistical properties of KSPCA, providing the first bound on its sample complexity. This bound is dimension-independent, which justifies the good performance of KSPCA on high-dimensional data. Another observation is that in the kernelized version, the number of parameters of KSPCA grows linearly with the sample size. While this potentially increases the risk of over-fitting, KSPCA works well in practice. In this work, we justify this compelling characteristic of KSPCA by providing a guarantee indicating that KSPCA generalizes well even when the number of parameters is large, as long as they have small norms.

Keywords: Kernel Supervised PCA, Sample Complexity, Generalization Bound

1. Introduction

Kernel supervised principal component analysis (KSPCA) (Barshan et al., 2011) is a dimensionality reduction and feature extraction method that has found many applications in data visualization, regression and classification (Fewzee and Karray, 2012; Parsaei et al., 2012; Samadani et al., 2013; Wu et al., 2013; Adamiak et al., 2015; Wu et al., 2014). Building on the idea of Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005), KSPCA provides a supervised extension of principal component analysis (PCA) and kernel PCA (Schölkopf et al., 1997).

KSPCA has a number of desirable characteristics that makes it both theoretically and practically interesting. First of all, the algorithm is intuitive and easy to understand and implement. In fact, it finds a linear mapping of the instances to a new space so that the cross-correlation between the instances and their labels (in the classification setting) is maximized. The algorithm can also be "kernelized" to allow non-linear mappings as well. Furthermore, it can be shown that the optimization problem boils down to singular value decomposition for which efficient solvers exist. These qualities along with the fact

that—unlike many other counterparts—KSPCA is a supervised method make it especially attractive for feature extraction.

However, KSPCA has not been analyzed from a statistical point of view. Particularly, we are not aware of any generalization bound that assists us in knowing the number of required training samples for learning a good mapping. This becomes more crucial when we observe that in the kernelized version, the number of parameters can grow linearly with the number of instances. Hence, one may expect that KSPCA is prone to over-fitting. However, quite surprisingly, KSPCA still works well in practice. It is therefore important to understand this compelling feature of KSPCA.

In this paper, we study KSPCA from a statistical perspective and prove a generalization bound for it. The bound suggests that KSPCA generalizes well—even when the number of parameters is very large—as long the parameters are "small in size". In order to prove such a result, we will establish a dimension-independent generalization bound that can work in infinite dimensional Hilbert spaces. In this regard, we see our results in line with dimension-independent generalization bounds for different algorithms in Hilbert spaces, including SVMs (Vapnik and Kotz, 1982), k-means clustering (Biau et al., 2008) and PCA (Shawe-Taylor et al., 2005; Blanchard et al., 2007; Rosasco et al., 2010). Moreover, a related bound for neural networks suggests that "it is the size of the weights that matters, not the size of the network." (Bartlett, 1998). Nevertheless, proving such results for KSPCA is still a challenge, which we will address by using a different proof technique.

In the next section, we will review the Hilbert-Schmidt independence criterion. Building on that, we will define KSPCA in Section 3. We present the main result in Section 4 and provide its proof in Section 5, which will be followed by the conclusions in Section 6.

2. Hilbert-Schmidt Independence Criterion

Hilbert-Schmidt Independence Criterion (HSIC), proposed by Gretton et al. (Gretton et al., 2005), measures the dependence between two random variables based on the norm of the covariance operator defined over the associated Reproducing Kernel Hilbert Spaces (RKHSs). HSIC, when equipped with appropriate kernels, is an effective tool for "measuring" (non-linear) dependence between two random variables.

Notations. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^q$ be two domain sets, with $\mathcal{F}$ and $\mathcal{G}$ being their corresponding (separable) RKHSs. Denote by $\phi : \mathcal{X} \mapsto \mathcal{F}$ and $\psi : \mathcal{Y} \mapsto \mathcal{G}$ two feature maps with their associated kernels being $k(.,.)$ and $l(.,.)$ respectively. Let $\mathcal{D}_{xy}$ be a Borel probability measure over $\mathcal{X} \times \mathcal{Y}$.

Definition 1 (Hilbert-Schmidt Independence Criterion (HSIC))
HSIC is defined as follows.

$$HSIC(\mathcal{D}_{xy}, \mathcal{F}, \mathcal{G}) := \left\| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{xy}} [(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)] \right\|^2_{HS} \tag{1}$$

where μ_x and μ_y are the mean values of $\phi(x)$ and $\psi(x)$ respectively, $\otimes$ stands for the tensor product, and HS-norm is the Hilbert-Schmidt norm[1].

1. I.e., the generalization of Frobenius norm to Hilbert Spaces.

Working directly with—possibly infinite dimensional—RKHSs is not practical. The following lemma (Gretton et al., 2005) gives an equivalent formulation of HSIC in terms of the kernel functions.

Lemma 2 (HSIC in terms of kernels)

$$HSIC(\mathcal{D}_{xy}, \mathcal{F}, \mathcal{G}) = \mathop{\mathbb{E}}_{x,y,x\prime,y\prime}[k(x,x\prime)l(y,y\prime)] + \mathop{\mathbb{E}}_{x,x\prime}[k(x,x\prime)] \mathop{\mathbb{E}}_{y,y\prime}[l(y,y\prime)] - 2 \mathop{\mathbb{E}}_{x,y}\left[\mathop{\mathbb{E}}_{x\prime}[k(x,x\prime)] \mathop{\mathbb{E}}_{y\prime}[l(y,y\prime)]\right] \quad (2)$$

where the (x, y) and $(x\prime, y\prime)$ are generated iid from $\mathcal{D}_{xy}$.

The connection between HSIC and dependence of two random variables is established in the following lemma (Gretton et al., 2005). We do not provide the definition of universal kernels (Steinwart, 2002), but note that many natural kernels (e.g., Gaussian) are universal.

Theorem 3 (HSIC and Independence)
Assume k and l are universal kernels on compact domains $\mathcal{X}$ and $\mathcal{Y}$. Assume $\|f\|_\infty \leq 1$ and $\|g\|_\infty \leq 1$ for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Then x and y are independent if and only if $HSIC(\mathcal{D}_{xy}, \mathcal{F}, \mathcal{G}) = 0$.

As it can be seen, computing HSIC requires having access to the joint distribution. However, in practice, we only have a set of samples generated from the distribution. The following method gives an empirical estimation of HSIC based on the data.

Definition 4 (Empirical HSIC)
Let $S = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\} \subseteq \mathcal{X} \times \mathcal{Y}$ be a sample set independently drawn from $\mathcal{D}_{xy}$. Then, empirical HSIC is defined by

$$HSIC(S, \mathcal{F}, \mathcal{G}) = \frac{\mathbf{tr}(KHLH)}{(m-1)^2} \quad (3)$$

where $K_{i,j} := k(x_i, x_j)$, $L_{i,j} := l(y_i, y_j)$ and H is the centering matrix $H = I - \frac{1}{m}\mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ is a vector of all 1s of size m.

Finally, one needs to show that the empirical HSIC is a good estimator of HSIC. The following large deviation bound is also due to Gretton et al. (2005).

Theorem 5 (Large Deviation Bound for Empirical HSIC)
Assume that k and l are non-negative and bounded almost everywhere by 1. Then for all $\mathcal{D}_{xy}$, if S is a sample of size m generated iid from $\mathcal{D}_{xy}$, then with probability at least $1 - \delta$ we have

$$|HSIC(S, \mathcal{F}, \mathcal{G}) - HSIC(\mathcal{D}_{xy}, \mathcal{F}, \mathcal{G})| = O\left(\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{m}}\right) \quad (4)$$

Therefore, HSIC is a useful tool to measure the dependence between two random variables. As we will see in the next section, it is the basis of the KSPCA algorithm.

3. Kernel Supervised Principal Component Analysis

Kernel supervised principal component analysis (KSPCA) (Barshan et al., 2011) is a feature extraction method that gives a representation of data under which the dependence between the input variable x and the response variables y is maximized.

Consider a binary classification problem[2] where we are given a training sample set $S = \{(x_i, y_i)\}_{i=1}^m$. Now for fixed kernel functions k and l over the input and the labels respectively, the kernel matrices over the sample are defined by $K_{i,j} = k(x_i, x_j)$ and $L_{i,j} = l(y_i, y_j)$. Using the empirical HSIC criterion (3), we can measure the dependence between the two random variables x and y. For the simplest case where the kernel functions are just the inner products in the original spaces, we can measure the dependence by

$$\frac{\mathbf{tr}(KHLH)}{(m-1)^2} = \frac{\mathbf{tr}(X^T XHYY^T H)}{(m-1)^2} \tag{5}$$

where $X_{:,i} = x_i$ is the input matrix and $Y_i = y_i$ is the label vector.

Assume that the input data lies in $\mathbb{R}^d$. In the non-kernelized version, supervised PCA tries to find a linear transformation from $\mathbb{R}^d$ to $\mathbb{R}^q$ (represented by the matrix $U_{d\times q}$), such that the dependence between the new representation $U^T X$ and the response matrix Y is maximized. More specifically, it solves the following optimization problem.

$$\begin{aligned} \underset{U}{\text{maximize}} \quad & \mathbf{tr}((U^T X)^T (U^T X) HYY^T H) \\ & \text{s.t. } U^T U = I \end{aligned} \tag{6}$$

This method can be readily generalized to the kernelized version, where the goal is to find a linear mapping over a RKHS. A representer theorem makes sure that the transformation U can be written in terms of data points (i.e., $U = \phi(X)B$ for some B of size $d \times m$). The optimization problem then boils down to

$$\begin{aligned} \underset{B}{\text{maximize}} \quad & \mathbf{tr}(B^T KHLHKB) \\ & \text{s.t. } B^T KB = I \end{aligned} \tag{7}$$

where the new representation of the data is found by computing $X_{new} = B^T K$.

A key advantage of KSPCA is that this optimization problem can be solved efficiently. In fact, solving KSPCA boils down to generalized singular value decomposition (Barshan et al., 2011).

It is important to observe that for supervised PCA, the number of parameters (i.e., elements of U) is dq, which depends on the dimensionality of the input. This can be potentially problematic, espesially for the kernelzed version where the dimension of the Hilbert space is infinite. Note that based on the representer theorem, the number of required parameters is not actually infinite. In other words, the number of parameters is the size of B (i.e., mq) which grows linearly with the size of the training set. Nevertheless, KSPCA seems to be prone to over-fitting. In the next section, we will establish a generalization bound for KSPCA that can work even for infinite dimensional Hilbert spaces.

2. In principle, KSPCA can be applied to regression problems as well, but we limit our discussion to binary classification.

4. A Dimension-Independent Generalization Bound of KSPCA

In this section, our goal is to show that, under certain conditions, KSPCA will not over-fit. In particular, we will prove a uniform convergence result for supervised PCA. Since our bound is dimension-independent, it can be readily applied to the kernelized version (i.e, KSPCA) as well.

In order to provide a generalization bound, we need to show that the solution which is learned based on the training data (denoted by $\hat{U}$) works well over the whole (unknown) distribution—almost as good as the best possible solution U^*.

When we are talking about the optimal solution, we need to be clear about the objective function. For a distribution $\mathcal{D}_{xy}$ and a transformation U, the objective is the limit of HSIC criterion when the number of training points goes to infinity. In other word, the objective function f is defined by

$$f(U, \mathcal{D}_{xy}) = \lim_{m\to\infty} \frac{\mathbf{tr}(X_m^T U U^T X_m H Y_m Y_m^T H)}{(m-1)^2} \tag{8}$$

where X_m and Y_m are data matrices of m examples, generated iid from $\mathcal{D}_{xy}$.[3]

Theorem 6 (Main Result)

Let $\mathcal{R} = \{x \in \mathbb{R}^d : \|x\|_2^2 \leq 1\}$, and $\mathcal{D}_{xy}$ be any probability measure over $\mathcal{R} \times [-1, 1]$. Let (X_m, Y_m) be an iid sample set of size m generated from $\mathcal{D}_{xy}$. Denote by $\hat{U} = A(X_m, Y_m)$ the output of SPCA algorithm on the training data. Then there is a constant C (independent of $\mathcal{D}_{xy}$) such that with probability at least $1 - \delta$ we have

$$f(U^*, \mathcal{D}_{xy}) - f(\hat{U}, \mathcal{D}_{xy}) \leq C\sqrt{\frac{\log \frac{1}{\delta}}{m}} \tag{9}$$

where U^* is optimal solution with respect to $\mathcal{D}_{xy}$. In other words,

$$\left(\max_{\substack{U \\ s.t.\ U^T U = I}} f(U, \mathcal{D}_{xy}) \right) - f(A(X_m, Y_m), \mathcal{D}_{xy}) = O\left(\sqrt{\frac{\log \frac{1}{\delta}}{m}}\right) \tag{10}$$

Note that the result is presented as a one-sided inequality, because the goal is to maximize the objective function. A nice property of this result is that it does not depend on the dimension d. Therefore, it can be used for infinite dimensional Hilbert spaces as well. In other words, it also holds for the kernelized SPCA, as long as the kernels are bounded[4]. The assumption that enables us to prove this dimension-independent bound is the boundedness of (i) the size of the input, i.e., $\|x\|_2^2 \leq 1$ and (ii) the size of the parameters, i.e., $U^T U = I$. In the next section, we provide the proof of this result.

3. Note that a good solution should also satisfy the constraint $U^T U = I$. However, we consider this as a necessary requirement for any solution, as it is the case for KSPCA algorithm.
4. Note that even for unbounded kernels like Gaussian kernel, if the inputs are bounded, then the output of the kernel is still bounded on the working domain

5. Proofs

In order to prove Theorem 6, we will establish a uniform convergence result. For standard loss functions (e.g., 0-1 loss for classification), this is usually done through bounding the covering number[5] (Vapnik and Chervonenkis, 1971). However, for other scenarios (e.g., other loss functions), it is sometimes easier to bound the Rademacher complexity (Bartlett and Mendelson, 2003) directly (e.g., Maurer and Pontil (2010); Biau et al. (2008))[6].

While bounding the Rademacher complexity seems to be a plausible approach, we employ a different proof technique which is more direct. In the next section, we will provide a useful statistical tool which paves the way for the final proof.

5.1 A Concentration Inequality for Hilbert Spaces

Hoeffding's inequality (Hoeffding, 1963) is a useful concentration inequity which can be used to bound the deviation of the empirical mean of a random variable from its true mean. This result is based on the assumption that the random variable is real-valued and bounded. However, in our analysis we need a more general result—one that would work for random *vectors*. In order to prove such a result, we start by providing some background about U-Statistics (Hoeffding, 1963).

Definition 7 (U-Statistic)
A one-sample degree r U-statistic is defined as

$$u := \frac{(m-n)!}{m!} \sum_{\mathbf{i}_r^m} g(x_{i_1}, x_{i_2}, ..., x_{i_r}) \tag{11}$$

where $\mathbf{i}_r^m$ is the set of all permutations of r elements from $\{1, 2, .., m\}$ and g is called the kernel of U-statistic. The following theorem gives a concentration inequality for U-statistics (Hoeffding, 1963).

Theorem 8 (Deviation Bound for U-statistic)
Let u be a degree r U-statistic with $0 \leq g \leq 1$ being the associated kernel. Then with probability at least $1 - \delta$

$$|u - \mathbb{E}[u]| = O\left(\sqrt{\frac{r \log \frac{1}{\delta}}{m}}\right) \tag{12}$$

Now we are ready to prove the concentration inequality for random variables taking values in Hilbert spaces.

5. Growth function in the case of binary classification
6. Also, Rademacher complexity has the advantage of being distribution-dependent, which can be tighter in practice (Mohri and Rostamizadeh, 2009; Cortes et al., 2010).

Theorem 9 (Generalized Hoeffding's Inequality)

Let $x_1, x_2, ..., x_n$ be a set of independent and identically distributed random vectors, taking values in the unit ball of the Hilbert space[7] *with their mean being $\bar{x}$. Then with probability at least $1-\delta$ we have*

$$\left\| \bar{x} - \frac{1}{n}\sum_{i=1}^{n} x_i \right\|_2^2 = O\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \tag{13}$$

Proof

$$\left\| \bar{x} - \frac{1}{n}\sum_{i=1}^{n} x_i \right\|^2 = \|\bar{x}\|_2^2 + \frac{<\sum_{i=1}^n x_i, \sum_{i=1}^n x_i>}{n^2} - \frac{2<\bar{x}, \sum_{i=1}^n x_i>}{n}$$

$$= \|\bar{x}\|_2^2 - 2\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} <\bar{x}, x_i>\right)}_{\text{degree one U-statistic}} + \frac{1}{n^2}\sum_{i=1}^{n}\|x_i\|_2^2 + \left(\frac{n-1}{n}\right)\underbrace{\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} <x_i, x_j>\right)}_{\text{degree two U-statistic}}$$

$$= \|\bar{x}\|_2^2 - 2\left(\|\bar{x}\|_2^2 \pm O\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)\right) + O\left(\frac{1}{n}\right) + \left(\frac{n-1}{n}\right)\left(\|\bar{x}\|_2^2 \pm O\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)\right)$$

$$= \|\bar{x}\|_2^2 - 2\|\bar{x}\|_2^2 + \|\bar{x}\|_2^2 \pm O\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \pm O\left(\frac{1}{n}\right) = O\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \tag{14}$$

Note that the kernels of the U-statistics, $g_1(x_i) =< x_i, \bar{x} >$ and $g_2(x_i, x_j) =< x_i, x_j >$, are bounded in $[-1, 1]$, rather than $[0, 1]$. But by a normalization argument, it can only introduce a factor of 2 to our analysis. Also, we used two deviation bounds simultaneously for two dependent random variables. This is also fine, because having union bound in mind, this can only introduce a constant factor to our computations. ■

5.2 Proof of the Main Result

As it was discussed before, the special form of the objective function f (see Eq. 8) makes it difficult to reuse the conventional uniform convergence results in our proof. One important characteristic of f is that it cannot be reformulated as the average of an objective over the training points. In other words, f is not a degree-1 U-statistic[8]. Therefore, we use a different proof technique.

The following lemma gives an equivalent form of the objective function which is easier to analyse.

7. That is $\{x : x \in \mathcal{H}, \|x\|_2 \leq 1\}$
8. It is actually a degree-2 U-statistic. Note that 0-1 loss for classification or mean squared loss for regression are degree-1 U-statistic.

Lemma 10 *Let $\mathcal{D}_{xy}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$.*[9] *Then*

$$f(U, \mathcal{D}_{xy}) = \left\| U^T \left(\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] \right) \right\|_2^2 \tag{15}$$

Proof

$$\begin{aligned} f(U, \mathcal{D}_{xy}) &= \lim_{m\to\infty} \frac{\mathbf{tr}(X_m^T U U^T X_m H Y_m Y_m^T H)}{(m-1)^2} = \lim_{m\to\infty} \frac{\mathbf{tr}(U^T X_m H Y_m Y_m^T H X_m^T U)}{(m-1)^2} \\ &= \lim_{m\to\infty} \frac{\left\| U^T X_m H Y_m \right\|_{HS}^2}{(m-1)^2} = \lim_{m\to\infty} \frac{\left\| \sum_{i=1}^m U^T \left(x_i - \mathbb{E}[x] \right) y_i \right\|_{HS}^2}{(m-1)^2} \\ = \lim_{m\to\infty} \frac{\left\| U^T \left(\sum_{i=1}^m \left(x_i - \mathbb{E}[x] \right) y_i \right) \right\|_{HS}^2}{(m-1)^2} &= \lim_{m\to\infty} \left(\frac{m}{m-1} \right)^2 \left\| U^T \left(\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] \right) \right\|_2^2 \end{aligned} \tag{16}$$

■

Finally, we are ready to prove Theorem 6.

Proof

Let $\mathbb{E}$ denote the true mean w.r.t the distribution $\mathcal{D}_{xy}$, and $\hat{\mathbb{E}}$ denote the empirical mean w.r.t. the sample. Define $\epsilon_1 := \mathbb{E}[xy] - \hat{\mathbb{E}}[xy]$, $\epsilon_2 := \mathbb{E}[x] - \hat{\mathbb{E}}[x]$ and $\epsilon_3 := \mathbb{E}[y] - \hat{\mathbb{E}}[y]$. In the following, we start by using Lemma 10, and continue by exploiting the fact that $\hat{U}^T$ is the maximizer of the empirical objective.

$$\begin{aligned} f(U^*, \mathcal{D}_{xy}) - f(\hat{U}, \mathcal{D}_{xy}) &= \left\| U^{*T} \left(\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] \right) \right\|_2^2 - \left\| \hat{U}^T \left(\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] \right) \right\|_2^2 \\ = \left\| U^{*T} \left((\hat{\mathbb{E}}[xy] + \epsilon_1) - (\hat{\mathbb{E}}[x] + \epsilon_2)(\hat{\mathbb{E}}[y] + \epsilon_3) \right) \right\|_2^2 &- \left\| \hat{U}^T \left((\hat{\mathbb{E}}[xy] + \epsilon_1) - (\hat{\mathbb{E}}[x] + \epsilon_2)(\hat{\mathbb{E}}[y] + \epsilon_3) \right) \right\|_2^2 \\ \leq \underbrace{\left(\left\| U^{*T} \left(\hat{\mathbb{E}}[xy] - \hat{\mathbb{E}}[x]\hat{\mathbb{E}}[y] \right) \right\|_2^2 - \left\| \hat{U}^T \left(\hat{\mathbb{E}}[xy] - \hat{\mathbb{E}}[x]\hat{\mathbb{E}}[y] \right) \right\|_2^2 \right)}_{\leq 0} &+ \|\hat{U}^T \epsilon_1\|_2^2 + \|U^{*T} \epsilon_1\|_2^2 + \\ + \|\hat{U}^T \epsilon_2 \epsilon_3\|_2^2 + \|U^{*T} \epsilon_2 \epsilon_3\|_2^2 + \|\hat{U}^T \epsilon_2\|_2^2 + \|U^{*T} \epsilon_2\|_2^2 + \|\hat{U}^T \epsilon_3\|_2^2 + \|U^{*T} \epsilon_3\|_2^2 & \\ \leq 2\|\epsilon_1\|_1^2 + 2\|\epsilon_2\|_2^2 + 4\|\epsilon_3\|_2^2 \leq C\sqrt{\frac{\log \frac{1}{\delta}}{m}} & \end{aligned} \tag{17}$$

The last step follows from Theorem 9, which states that the empirical means of bounded vectors in Hilbert spaces are close to their true means[10].

■

9. For the binary classification task, $\mathcal{Y} = \{-1, 1\}$

10. Having union bound in mind, using this Theorem for multiple dependent random vectors is fine, as it can introduce only a constant factor to the deviation bound.

6. Conclusions

In this work we analysed the sample complexity of supervised principal component analysis (Barshan et al., 2011), providing a novel dimension-independent upper bound for it. This result explains why SPCA works well even on high-dimensional data. More importantly, it justifies the use of the kernelized version of SPCA (i.e., KSPCA), where the number of parameters grows linearly with the size of the training set. The key observation that enabled us to prove such a result was the fact that although the number of parameters is large, their "norms" are small.

On a different note, it was pointed out by Barshan et al. (2011) that Kernel PCA (Schölkopf et al., 1997) can be regarded as a special case of KSPCA, which can be obtained by substituting the kernel label by the identity matrix. Moreover, it was shown by Ham et al. (2004) that a number of important dimensionality reduction methods (including Isomap (Tenenbaum et al., 2000), graph Laplacian eigenmap (Belkin and Niyogi, 2003), and locally linear embedding (Roweis and Saul, 2000)) are special cases of Kernel PCA for appropriate choices of the kernel. Therefore, all of these methods can be regarded as special cases of KSPCA. However, our bound does not apply to them directly. The reason is that although we considered an arbitrary "input kernel" K over the data, we fixed the "label kernel" based on the labels, i.e., $L = Y^T Y$. Therefore, an interesting follow up question is to prove an upper bound for the sample complexity of KSPCA for arbitrary "label kernels".

References

Krzysztof Adamiak, Piotr Duch, Dominik Zurek, and Krzysztof Slot. Modifications of most expressive feature reordering criteria for supervised kernel principal component analysis. In *Cybernetics (CYBCONF), 2015 IEEE 2nd International Conference on*, pages 507–511. IEEE, 2015.

Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011.

Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *Information Theory, IEEE Transactions on*, 44(2):525–536, 1998.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Gérard Biau, Luc Devroye, and Gábor Lugosi. On the performance of clustering in hilbert spaces. *Information Theory, IEEE Transactions on*, 54(2):781–790, 2008.

Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294, 2007.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 247–254, 2010.

Pouria Fewzee and Fakhri Karray. Dimensionality reduction for emotional speech recognition. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 532–537. IEEE, 2012.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

Jihun Ham, Daniel D Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty-first international conference on Machine learning*, page 47. ACM, 2004.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Andreas Maurer and Massimiliano Pontil. K-dimensional coding schemes in hilbert spaces. *Information Theory, IEEE Transactions on*, 56(11):5839–5846, 2010.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, pages 1097–1104, 2009.

Hossein Parsaei, Mehrdad J Gangeh, Daniel W Stashuk, and Mohamed S Kamel. Augmenting the decomposition of emg signals using supervised feature extraction techniques. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2615–2618. IEEE, 2012.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *The Journal of Machine Learning Research*, 11:905–934, 2010.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Ali-Akbar Samadani, Ali Ghodsi, and Dana Kulić. Discriminative functional analysis of human movements. *Pattern Recognition Letters*, 34(15):1829–1839, 2013.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Artificial Neural NetworksICANN'97*, pages 583–588. Springer, 1997.

John Shawe-Taylor, Christopher KI Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *Information Theory, IEEE Transactions on*, 51(7):2510–2522, 2005.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research*, 2:67–93, 2002.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2): 264–280, 1971.

Vladimir Naumovich Vapnik and Samuel Kotz. *Estimation of dependences based on empirical data*, volume 40. Springer-verlag New York, 1982.

Hui Wu, Dustin M Bowers, Toan T Huynh, and Richard Souvenir. Biomedical video denoising using supervised manifold learning. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 1244–1247. IEEE, 2013.

Hui Wu, Toan T Huynh, and Richard Souvenir. Motion factorization for echocardiogram classification. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 445–448. IEEE, 2014.