Statistics and Probability Letters 171 (2021) 109019

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

# Revisiting the predictive power of kernel principal components

Ben Jones, Andreas Artemiou [*]

*School of Mathematics, Cardiff University, United Kingdom of Great Britain and Northern Ireland*

ARTICLE INFO

*Article history:*
Received 10 August 2020
Received in revised form 17 November 2020
Accepted 5 December 2020
Available online 19 December 2020

*Keywords:*
Model-free regression
Conditional independence
Hilbert spaces
Dimension reduction
Principal components

ABSTRACT

In this short note, recent results on the predictive power of kernel principal component in a regression setting are extended in two ways: (1) in the model-free setting, we relax a conditional independence model assumption to obtain a stronger result; and (2) the model-free setting is also extended in the infinite-dimensional setting.

## 1. Introduction

The efficient statistical analysis of high-dimensional datasets is considered one of the most challenging problems in recent years. The ability to collect and store massive amounts of data in a cheap way has allowed scientists to collect a lot of high-dimensional data. In an effort to reduce the dimensionality of the datasets, researchers are resorting to a preprocessing step which allows them to reduce the dimensionality of the dataset. In a regression setting, when the predictors are high dimensional and there is need to reduce the dimension of the dataset for an efficient analysis a number of preprocessing steps have been proposed in the literature. One of these approaches is the principal component analysis (PCA) which reduces effectively the dimensionality of the predictors (see for example Chiaromonte and Martinelli, 2002).

During the 20th century, a long debate in the Statistics community, evolved around the effectiveness of using PCA to reduce the dimensionality in a regression setting (and more generally — in a supervised setting). The debate on this topic, had some prominent statisticians taking opposing sides (see for example Mosteller and Tukey, 1977; Cox, 1968). Cook (2007) gives a very detailed overview of this debate.

Following the discussion of this topic in Cook's Fisher Lecture (Cook, 2007) by (Li, 2007) a number of researchers tried to give a probabilistic answer on the predictive potential of principal components, that is on the probability that the higher order principal components will be more correlated with the response rather than the lower order principal components. Artemiou and Li (2009) discussed this in a linear model under the assumption of an orientationally uniform covariance matrix $\boldsymbol{\Sigma} = \text{var}(\boldsymbol{X})$ and they proved that the probability of a higher order principal component having higher correlation with the response than a lower order one is greater than 1/2. Ni (2011) extended the result showing that the exact probability is $(2/\pi)E(\arctan\sqrt{\lambda_i/\lambda_j})$ where $\lambda_i$ is the $i$th eigenvalue of $\boldsymbol{\Sigma}$ and $i, j$ where $i < j$ are the subscripts of the two principal components. Artemiou and Li (2013) expanded the results for a more general regression settings. Jones and Artemiou (2020) discussed the results in the context of functional principal components and Jones et al. (2020)

[*] Corresponding author.
*E-mail address:* ArtemiouA@cardiff.ac.uk (A. Artemiou).

https://doi.org/10.1016/j.spl.2020.109019

discussed this for the kernel principal components. The more general result in Jones et al. (2020) essentially shows that in a model-free setting, if someone randomly chooses a measure for the conditional distribution of $Y|\boldsymbol{X}$ similar results as the one proved by Artemiou and Li (2009) and Ni (2011) hold.

In this paper we generalize further some of the results in Jones et al. (2020) for the predictive potential of kernel principal components on different directions. First of all, we generalize the results in the model-free setting, that is, we extend the results in the case where the conditional independence $Y \perp\!\!\!\perp \boldsymbol{\Sigma}|\boldsymbol{X}$ is relaxed to $g(\boldsymbol{Y}) \perp\!\!\!\perp \boldsymbol{\Sigma}|\boldsymbol{X}$ and then we propose a way to extend the results in the case that we have infinite-dimensional kernels. In Jones et al. (2020) the model free setting was discussed only in the case where we have finite dimensional kernels. In this work, we incorporate an extra assumption to demonstrate that the results can be extended to infinite dimensional kernels.

The rest of the paper is structured as follows. First, we revisit some key results and notation from Jones et al. (2020) in Section 2. In Section 3 we discuss the extensions of the previous results. We close with a short discussion in Section 4. Here we emphasize that the proofs are very similar to the ones presented in Jones et al. (2020) and therefore they can be omitted completely. In any case we provide the proof of the more general result in this work which is the most general result to this day on the predictive potential to kernel principal components. (Essentially this is the most general result on the predictive potential of any form of principal component analysis).

## 2. Predictive power of kernel principal components

In this section we revisit the most general results from Jones et al. (2020) which are the ones for the predictive potential of Kernel Principal Components in the model-free setting. The key assumption in their results is the $Y \perp\!\!\!\perp \Sigma|X$ which we relax in the next section.

**Theorem 1.** *Suppose that:*

1. $Y \perp\!\!\!\perp \Sigma|X$
2. *$\mathcal{H}$ has finite dimension*
3. *$\Sigma$ is a random covariance operator where the distribution is invariant under unitary transformation. In other words, $\Sigma$ has the same distribution as $U\Sigma U^{-1}$ for any unitary $U : \mathcal{H} \to \mathcal{H}$. It is assumed that, almost surely, the non-zero eigenvalues have unit multiplicity.*
4. *$g(Y)$ is a real-valued measurable function of $Y$ such that the function $x \mapsto E[g(Y)|X = x]$ belongs to $\mathcal{H}$.*

*Then, with probability* 1,

$$P\left\{\mathrm{Corr}^2[g(Y), u_i(X)|\Sigma] \geq \mathrm{Corr}^2[g(Y), u_j(X)|\Sigma]\right\} = (2/\pi)E\left\{\arctan[(\lambda_i/\lambda_j)^{1/2}]\right\}$$

*for any two eigen-pairs $(\lambda_i, u_i)$ and $(\lambda_j, u_j)$ of $\Sigma$ satisfying $i < j$ and*

$$\mathrm{Cov}[g(Y), u_i(X)|\Sigma] \neq 0,\ \ \mathrm{Cov}[g(Y), u_j(X)|\Sigma] \neq 0.$$

The next theorem gives the more general result Jones and Artemiou (2020) proved where it states that we can arbitrarily choose a conditional distribution for $Y|X$ and the result still holds. In Section 4 we will extend this, as well as the above result, to allow for the Hilbert space $\mathcal{H}$ to be infinite dimensional.

**Theorem 2.** *Suppose that:*

1. *$\Sigma$ is a random covariance operator where the distribution is invariant under unitary transformation. In other words, $\Sigma$ has the same distribution as $U\Sigma U^{-1}$ for any unitary $U : \mathcal{H} \to \mathcal{H}$. It is assumed that, almost surely, the non-zero eigenvalues have unit multiplicity.*
2. *$\mathcal{H}$ has finite dimension*
3. *$\nu$ is a random conditional distribution for $Y|X$ such that $P(\nu \in K_0) = 0$ where $K_0$ denotes the set of conditional distributions for which $X$ and $Y$ are independent*
4. $Y|(X, \nu) \sim \nu, \quad \nu \perp\!\!\!\perp (X, \Sigma), \quad Y \perp\!\!\!\perp \Sigma|(X, \nu)$
5. *$g$ is a real-valued measurable function of $Y$ such that the random function $m_\nu(\cdot) = \int g\, \nu(d\omega, \cdot)$ belongs to $\mathcal{H}$ almost surely and, with probability* 1,

$$\mathrm{Cov}[g(Y), u_i(X)|\nu, \Sigma] \neq 0,\ \ \mathrm{Cov}[g(Y), u_j(X)|\nu, \Sigma] \neq 0.$$

*Then for any $i < j$,*

$$P\{\mathrm{Corr}^2[g(Y), u_i(X)|\nu, \Sigma] \geq \mathrm{Corr}^2[g(Y), u_j(X)|\nu, \Sigma]\} = (2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{1/2}]\}.$$

## 3. Model free setting

In this section we present the most important results of this paper. We first demonstrate how one can relax the assumption in Theorems 1 and 2 to extend the results in a more general setting. Then we also demonstrate how the model free results can be extended in the infinite dimensional Hilbert space $\mathcal{H}$. We emphasize here that the results in this section are the most general to date in the predictive power of kernel principal components. More importantly the results on the infinite dimensional Hilbert space $\mathcal{H}$ were not addressed at all in Jones and Artemiou (2020).

Before we outline the result we explain the assumption we use in the model free setting. In Jones and Artemiou (2020) the assumption $Y \perp\!\!\!\perp \Sigma|X$ was used. To extend Theorem 1, a random conditional distribution can be chosen for $g(Y)|X$ rather than for $Y|X$. See below an example why such relaxation is important.

**Remark 1.** An example of a model for which $g(Y) \perp\!\!\!\perp \Sigma|X$ holds but $Y \perp\!\!\!\perp \Sigma|X$ fails is given by:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

Take $g(Y) = Y_3$. Then $g(Y) \perp\!\!\!\perp \Sigma|X$, but not $Y \perp\!\!\!\perp \Sigma|X$.

### 3.1. Finite dimensional setting

Although both Theorems 1 and 2 can be extended in this setting we focus on the extension of Theorem 2 which presents the more general setting. the extension for Theorem 1 is straight forward.

**Theorem 3.** *Suppose that:*

1. *Let $\mathcal{H}$ be a finite dimensional Hilbert space*
2. *$g$ is a real-valued measurable function of $Y$ such that the random function $m_\nu(\cdot) = \int g\, \nu(d\omega, \cdot)$ belongs to $\mathcal{H}$ almost surely and, with probability* 1,
$$\mathrm{Cov}[g(Y), u_i(X)|\nu, \Sigma] \neq 0, \;\; \mathrm{Cov}[g(Y), u_j(X)|\nu, \Sigma] \neq 0.$$
3. *$\Sigma$ is a random covariance operator where the distribution is invariant under unitary transformation. In other words, $\Sigma$ has the same distribution as $U\Sigma U^{-1}$ for any unitary $U : \mathcal{H} \to \mathcal{H}$. It is assumed that, almost surely, the non-zero eigenvalues have unit multiplicity.*
4. *$\nu$ is a random conditional distribution for $g(Y)|X$ such that $P(\nu \in K_0) = 0$ where $K_0$ denotes the set of conditional distributions for which $X$ and $g(Y)$ are independent*
5. $g(Y)|(X, \nu) \sim \nu, \quad \nu \perp\!\!\!\perp (X, \Sigma), \quad g(Y) \perp\!\!\!\perp \Sigma|(X, \nu)$

*Then for any $i < j$,*

$$P\{\mathrm{Corr}^2[g(Y), u_i(X)|\nu, \Sigma] \geq \mathrm{Corr}^2[g(Y), u_j(X)|\nu, \Sigma]\} = (2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{1/2}]\}.$$

### 3.2. Infinite dimensional kernels

In this section we show how one can address the model-free setting in the infinite dimensional kernel case which was not considered in Jones and Artemiou (2020). By making a uniformity assumption on a "restriction" of $\Sigma$, the results can be extended.

**Assumption 1.** Suppose that $\Sigma$ is a random compact covariance operator. There exists a set of integers $V = \{v_1, \ldots, v_l\}$, for some $l \in \mathbb{N}$, such that

$$\Sigma_* = \sum_{v_i \in V} \lambda_{v_i}(u_{v_i} \otimes u_{v_i})$$

is invariant under unitary transformations. Without loss of generality, it will be assumed that $v_1 < v_2 \ldots < v_l$.

The following theorem says that you can choose any measure to define the relationship between $X$ and $Y$ as long as the two are not independent. This is the more general theorem of the predictive power of kernel principal components. For this reason, we provide its proof (although it is very similar to the one in Jones and Artemiou, 2020). (Similarly to the previous section we show the extension of Theorem 2 under the new Assumption. One can adjust Theorem 1 simlarly).

**Theorem 4.** *Suppose that:*

1. *$g$ is a real-valued measurable function of $Y$ such that the random function $m_\nu(\cdot) = \int g\, \nu(d\omega, \cdot)$ belongs to $\mathcal{H}$ almost surely and, with probability* 1,
$$\mathrm{Cov}[g(Y), u_i(X)|\nu, \Sigma] \neq 0, \;\; \mathrm{Cov}[g(Y), u_j(X)|\nu, \Sigma] \neq 0.$$

2. *$\Sigma$ is a random compact covariance operator satisfying Assumption* 1. *It is assumed that, almost surely, the non-zero eigenvalues have unit multiplicity.*
3. *$\nu$ is a random conditional distribution for $g(Y)|X$ such that $P(\nu \in K_0) = 0$ where $K_0$ denotes the set of conditional distributions for which $X$ and $g(Y)$ are independent*
4. *$g(Y)|(X,\nu) \sim \nu, \quad \nu \perp\!\!\!\perp (X,\Sigma), \quad g(Y) \perp\!\!\!\perp \Sigma|(X,\nu)$*

*Let $V = \{v_1, \dots, v_l\}$ $(v_1 < v_2 < \cdots < v_l)$ be such that $\sum_{v_i \in V} \lambda_{v_i} u_{v_i} \otimes u_{v_i}$ is unitarily invariant. Then for any $i < j \leq l$,*

$$P\{\mathrm{Corr}^2[g(Y), u_{v_i}(X)|\nu, \Sigma] \geq \mathrm{Corr}^2[g(Y), u_{v_j}(X)|\nu, \Sigma]\} = (2/\pi)E\{\arctan[(\lambda_{v_i}/\lambda_{v_j})^{1/2}]\}.$$

**Proof.** We begin similarly to the proof of theorem 11 by noting that for any $i$

$$\mathrm{Cov}[g(Y), u_{v_i}(X)|\nu, \Sigma] = \mathrm{Cov}\{E[g(Y)|\nu, \Sigma, X], u_i(X)|\nu, \Sigma\}.$$

Also note that

$$E[g(Y)|\nu, \Sigma, X] = E[g(Y)|\nu, X] = m_\nu(X).$$

because of the assumption $Y \perp\!\!\!\perp \Sigma|(X,\nu)$. We see that $\nu \perp\!\!\!\perp (X,\Sigma)$ implies $m_\nu \perp\!\!\!\perp (X,\Sigma)$. Thus, for any $\kappa \in \mathcal{K}$, we have that

$$\mathrm{Cov}[m_\nu(X), u_{v_i}(X)|\nu = \kappa, \Sigma] = \mathrm{Cov}[m_\kappa(X), u_{v_i}(X)|\Sigma] = \langle m_\kappa, \Sigma u_{v_i}\rangle_{\mathcal{H}} = \lambda_{v_i}\langle m_\kappa, u_{v_i}\rangle_{\mathcal{H}}.$$

which implies

$$\mathrm{Cov}[m_\nu(X), u_{v_i}(X)|\nu, \Sigma] = \lambda_{v_i}\langle m_\nu, u_{v_i}\rangle_{\mathcal{H}}.$$

Also by $\nu \perp\!\!\!\perp (X,\Sigma)$ we have

$$\mathrm{Var}[u_{v_i}(X)|\nu, \Sigma] = \mathrm{Var}[u_{v_i}(X)|\Sigma] = \lambda_{v_i}.$$

and therefore

$$\frac{\mathrm{Corr}^2[g(Y), u_{v_i}(X)|\nu, \Sigma]}{\mathrm{Corr}^2[g(Y), u_{v_j}(X)|\nu, \Sigma]} = \frac{\lambda_{v_i}\langle m_\nu, u_{v_i}\rangle_{\mathcal{H}}}{\lambda_{v_j}\langle m_\nu, u_{v_j}\rangle_{\mathcal{H}}}.$$

We see that $m_\nu \perp\!\!\!\perp (u_{v_i}, u_{v_j}, \lambda_{v_i}, \lambda_{v_j})$ implies $m_\nu \perp\!\!\!\perp (u_{v_i}, u_{v_j})|(\lambda_{v_i}, \lambda_{v_j})$. Therefore, for any $\kappa \in \mathcal{K}$, that

$$P\left(\frac{\langle m_\nu, u_{v_j}\rangle^2_{\mathcal{H}}}{\langle m_\nu, u_{v_i}\rangle^2_{\mathcal{H}}} < \frac{\lambda_{v_i}}{\lambda_{v_j}} \middle| \nu = \kappa, \lambda_{v_i}, \lambda_{v_j}\right) = P\left(\frac{\langle m_\kappa, u_{v_j}\rangle^2_{\mathcal{H}}}{\langle m_\kappa, u_{v_i}\rangle^2_{\mathcal{H}}} < \frac{\lambda_{v_i}}{\lambda_{v_j}} \middle| \lambda_{v_i}, \lambda_{v_j}\right).$$

using the results to prove Theorem 3 in Jones and Artemiou (2020), the right hand side is $(2/\pi)\arctan[(\lambda_{v_i}/\lambda_{v_j})^{\frac{1}{2}}]$. So we have shown that

$$P\left(\frac{\langle m_\nu, u_{v_j}\rangle^2_{\mathcal{H}}}{\langle m_\nu, u_{v_i}\rangle^2_{\mathcal{H}}} < \frac{\lambda_{v_i}}{\lambda_{v_j}} \middle| \nu, \lambda_{v_i}, \lambda_{v_j}\right) = (2/\pi)\arctan[(\lambda_{v_i}/\lambda_{v_j})^{\frac{1}{2}}].$$

and by taking the conditional expectation on both sides of the above equality we get the desired result of the proof. □

## 4. Discussion

In this paper we extend recently proposed results in the literature on the predictive potential of kernel principal components. There are two important contributions in this work. The most important contribution of this paper is the relaxation of the assumption on the model free case. The new conditional independence proposed is more general than the previous assumption and therefore we demonstrate that the result holds in a much broader range of cases. The second important contribution is the extension of the model free approach in the infinite dimensional Hilbert space settings.

The results in this work enhance the discussion around a topic that has troubled Statisticians in the 20th century and has received renewed interest lately due to the volume of high-dimensional data collected nowadays, which forces researchers to perform variable screening in supervised settings using PCA approaches (which are unsupervised). In the last decade, a series of papers provided evidence of the predictive potential of principal components in various settings and in this work we expand the results on the predictive potential of kernel principal components which was first addressed in Jones and Artemiou (2020).

This discussion is still open and there are a lot of interesting questions one can try to address. One of the most obvious one is that we are measuring the relationship between $Y$ and $X$ on nonlinear regression models using correlation which measures linear relationship. It will be interesting to extend this into a different measure of association which is more appropriate for nonlinear relationships. One such approach is given in Artemiou (2021).

## Acknowledgments

The authors would like to thank the Editor and a reviewer for their useful comments. The work of B. Jones has been partially funded by EPSRC.

## References

Artemiou, A., 2021. Using mutual information to measure the predictive potential of principal components. In: Li, Bing, Bura, Estathia (Eds.), Festschrift for D. R. Cook (in press).

Artemiou, A., Li, B., 2009. On principal components and regression: a statistical explanation of a natural phenomenon. Statist. Sinica 19, 1557–1566.

Artemiou, A., Li, B., 2013. Predictive power of principal components for single-index model and sufficient dimension reduction. J. Multivariate Anal. 119, 176–184.

Chiaromonte, F., Martinelli, J., 2002. Dimension reduction strategies for analyzing global gene expression data with a response. Math. Biosci. 176, 123–144.

Cook, R.D., 2007. Fisher lecture: Dimension reduction in regression. Statist. Sci. 22, 1–40.

Cox, D.R., 1968. Notes on some aspects of regression analysis. J. R. Stat. Soc. Ser. A 131, 265–279.

Jones, B., Artemiou, A., 2020. On principal components regression with Hilbertian predictors. Ann. Inst. Statist. Math. 72, 627–644.

Jones, B., Artemiou, A., Li, B., 2020. On the predictive potential of kernel principal components. Electron. J. Stat. 14, 1–23.

Li, B., 2007. Comment: Fisher lecture: Dimension reduction in regression. Statist. Sci. 22, 32–35.

Mosteller, F., Tukey, J.W., 1977. Data Analysis and Regression. Addison-Wesley, Reading, Massachusetts.

Ni, L., 2011. Principal component regression revisited. Statist. Sinica 21, 741–747.