Electronic Journal of Statistics
Vol. 0 (0000) 1–22
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

On the predictive potential of kernel principal components

Ben Jones and Andreas Artemiou*

School of Mathematics, Cardiff University,
e-mail: JonesBL7@cardiff.ac.uk; ArtemiouA@cardiff.ac.uk

Bing Li

Department of Statistics, Pennsylvania State University,
e-mail: bing@stat.psu.edu

Abstract: We give a probabilistic analysis of a phenomenon in statistics which, until recently, has not received a convincing explanation. This phenomenon is that the leading principal components tend to possess more predictive power for a response variable than lower-ranking ones despite the procedure being unsupervised. Our result, in its most general form, shows that the phenomenon goes far beyond the context of linear regression and classical principal components — if an arbitrary distribution for the predictor X and an arbitrary conditional distribution for $Y|X$ are chosen then any measureable function $g(Y)$, subject to a mild condition, tends to be more correlated with the higher-ranking kernel principal components than with the lower-ranking ones. The "arbitrariness" is formulated in terms of unitary invariance then the tendency is explicitly quantified by exploring how unitary invariance relates to the Cauchy distribution. The most general results, for technical reasons, are shown for the case where the kernel space is finite dimensional. The occurency of this tendency in real world databases is also investigated to show that our results are consistent with observation.

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: Cauchy distribution, Dimension Reduction, Nonparametric Regression, Kernel Principal Components, Unitary Invariance.

Contents

*Corresponding author.

1. Introduction

Kernel principal component analysis (Schölkopf, Smola, and Müller, 1997, 1998) is one of the most widely used methods for unsupervised dimension reduction. It captures directions of maximal variation among arbitrary nonlinear paths by performing an eigendecomposition of the kernel covariance operator thereby, with its nonlinearity, greatly expanding the scope and power of classical principal component analysis (Jolliffe, 2002). A core feature of this method is its use of the "kernel trick". This trick implies that if an operation depends only on inner products then a lower dimensional nonlinear projection of the data can be extracted without dealing directly with the projection coefficients. This versatile idea also appears in other settings, such as the support vector machine and, more recently, sufficient (or supervised) dimension reduction. See Vapnik (1998), Fukumizu, Bach, and Jordan (2004, 2009), Yeh, Huang, and Lee (2009), Hsing and Ren (2009), Shi, Belkin, and Yu (2009), Li, Artemiou and Li (2011), Lee, Li and Chiaromonte (2013), Artemiou and Dong (2016), and Li and Song (2017). Some other variations of principal component analysis include principal curves (Hastie and Stuetzle, 1989) and functional principal component analysis (Rice and Silverman, 1991 and Silverman, 1996).

We consider the following question: if we perform kernel principal component analysis (KPCA) on a sample of vector-valued predictors, without regard to any response variable, would regressing a response on the leading components be more useful than regressing on the lower-ranking components? The question is rooted in, and subsumes, a well-known long enduring debate concerning the predictive value of principal components in linear regressions. Researchers have specifically debated the common practice of regressing a scalar response variable on the first few principal components of the predictors. Principal component analysis (PCA) is an unsupervised dimension reduction technique — i.e. the extraction process does not use any information from the response — so there is the possibility that the first few principal components are the least related to the response. There have been arguments both for and against this practice. See, for example, Hotelling (1957), Kendall (1957, page 75), Cox (1968), Hocking (1976), Mosteller and Tukey (1977, page 307), Jolliffe (1982), Hawkins and Fatti (1984), and Scott (1992). See Cook (2007) for a thought-provoking account of this debate. This problem has received renewed interest due to the increasing ubiquity of high-dimensional datasets, especially those with small sample sizes, as principal component analysis is often used as a prescreening method to reduce the dimension of a dataset before a regression analysis is applied. See, for example, Alter, Brown, and Botstein (2000), Chiaromonte and Martinelli (2002), Bura and Pfeiffer (2003), and L. Li and H. Li (2004).

Let X be a p-dimensional random vector and Y a real random variable. Li (2007) conjectured, in a discussion of Cook (2007), that if the linear model $Y = \beta^\intercal X + \varepsilon$ holds then the first principal component of X is the most likely to be the most correlated with Y. This is assuming that: (1) X and ε are independent, and (2) the coefficient vector β and the covariance matrix Σ of X are chosen randomly and independently.

Artemiou and Li (2009) rigorously formulated and proved a variation on this conjecture. Let $U_1, \ldots, U_p$ be the 1st, $\ldots$, pth principal components of the p-dimensional random vector X and let $\mathrm{Corr}(A, B) = \mathrm{Cov}(A, B)/\sqrt{\mathrm{Var}(A)\mathrm{Var}(B)}$ be the correlation of real random variables A and B.

They showed that, under some assumptions on β and Σ, the following holds when $i < j$.

$$P\left(\text{Corr}^2(Y, U_i|\beta, \Sigma) > \text{Corr}^2(Y, U_j|\beta, \Sigma)\right) > 1/2.$$

Ni (2011) refined and strengthened this result, using stronger assumptions, by showing the following.

$$P\left(\text{Corr}^2(Y, U_i|\beta, \Sigma) > \text{Corr}^2(Y, U_j|\beta, \Sigma)\right) = (2/\pi)E\left(\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\right), \tag{1}$$

where λ_i is the eigenvalue corresponding to the i^{th} principal component U_i. The right hand side of (1) is at least $1/2$, since $\lambda_i \geq \lambda_j$, and increases with the ratio λ_i/λ_j. We note that Ni (2011) made a tacit assumption on the distribution of the covariance matrix to obtain this result. This assumption is discussed in Jones and Artemiou (2019+) and is similar to our notion of unitary invariance. The above theoretical probability, under a reasonable assumption given in Li (2007) and with $p = 2$, was calculated to be $2/\pi$ by Ni (2011). This refined the empirical estimate of 0.65 computed by simulation in Li (2007). These results confirm theoretically that principal components, though they are derived without reference to any response, do in fact tend to contain *some* information about the response with the higher-ranking ones tending to have more predictive value than the lower-ranking ones.

These results were then broadened to more general settings by Artemiou and Li (2013). The most significant setting considered was that of the very general conditional independence model. This model frequently appears in the sufficient dimension reduction framework so it worth exploring the predictive power of principal components, in this setting, to provide a basis for method comparison.

Hall and Yang (2010), from an alternative perspective, considered the question — in the linear regression context — of whether it is better to select a subset of principal components other than the first few. They proved that the conventional choice achieves a minimax result in that the largest mean squared difference between the fitted values and the signal possible at each step is minimised by choosing the principal components in the usual order. It is notable that this result holds for all sample sizes as opposed to merely asymptotically.

We show, in this paper, that higher-ranking kernel principal components also have the tendency to be more informative of the response than lower-ranking ones. A crucial assumption that was needed to prove the results of (Artemiou and Li, 2009, 2013) is that the random regression coefficients have a spherically symmetric distribution. Spherical distributions, and the related concept of Haar measures, have been studied extensively in the literature — see e.g. Fang et al (1990), and Muirhead (1982). Some of the most well-known kernels (e.g. the Gaussian kernel) induce infinite-dimensional feature spaces and some other kernels (e.g. for example polynomial kernels) use finite-dimensional feature spaces. It is known that spherical distributions do not exist in infinite-dimensional spaces (see Jones and Artemiou (2019+)) because the identity operator is not compact on such spaces. The work in section 4 focuses on finite-dimensional kernels to avoid the complications of this limitation. The infinite-dimensional case is nevertheless considered in section 5 by presenting a result that relies on a different assumption.

The most far-reaching result of this paper, theorem 4 in section 4.3, states that if nature picks — without favouritism — an arbitrary distribution for X and an arbitrary conditional distribution for

$Y|X$ then any measureable function $g(Y)$, subject to a mild condition, tends to have more correlation with the higher-ranking kernel principal components. We assume that, similarly to Artemiou and Li (2009) and Ni (2011), the random operator Σ has an arbitrary orientation. This is formulated in terms of unitary invariance meaning that Σ is the same random element independently of the coordinate system from which it is observed. The result of theorem 4 requires no restrictive assumption meaning that no statistical models, parametric or nonparametric, are imposed on the relationship between X and Y.

We begin in section 2 by showing empirically, through investigating three databases, that the described phenomenon holds in real-world datasets. We then, in section 3, outline the construction of KPCA and layout the goal of this paper in technical terms. We present and prove the main results in section 4. Some discussion on the case with infinite-dimensional kernels is provided in section 5. We close with some concluding remarks in section 6 and give acknowledgements in section 7.

2. Some empirical evidence

We present here some empirical evidence, from three databases, that shows how the predictive tendency of the kernel principal components analysis (when using a kernel that has a finite-dimensional feature space) manifests itself in naturally collected data sets. We select data sets, from the databases, according to three pre-specified criteria: (i) they have univariate responses; we choose randomly if the response is multivariate, (ii) they have no categorical predictors, and (iii) they are not artificially constructed. The first database, from which we select 33 data sets according to the above criteria, is that provided in the *Arc* software (http://www.stat.umn.edu/arc/software.html). This database is also used in Artemiou and Li (2009) in the context of classical principal components analysis. The second database, from which we select 53 data sets, consists of data sets from a multivariate analysis textbook by Johnson and Wichern (2007), The third database, from which we select 54 data sets, is the CMU StatLib database (http://lib.stat.cmu.edu/index.php).

We use a second-order polynomial kernel with unit scale and offset to construct the boxplots (figure 1) for the first 5 principal components of the 3 databases. We use the *kernelMatrix* function in the *kernlab* R library (Karatzoglou et al, 2018) to compute the principal components. The absolute values of the sample correlation between each of the first 5 kernel principal components and the response are calculated. We obtain, e.g. from the *Arc* database, 33 correlations (one for each dataset) for each of the 5 kernel principal components. It is evident from the boxplots that higher-ranking kernel principal components tend to have stronger correlations with the response. The first kernel principal components, in particular, have considerably stronger correlations with the response than the other components. We again stress — as is evident in the figure — that the tendency is *probabilistic*.

3. Technical construction of KPCA

The following notation is adopted throughout the rest of this paper. Let U, V and W be generic random variables defined on some probability space. The notation $U \perp\!\!\!\perp V$ will mean that U and V are independent and, similarly, $U \perp\!\!\!\perp V|W$ will mean that U and V are independent conditioning on W. Capital letters such as X, Y, Z will denote random variables or vectors; capital letters such as

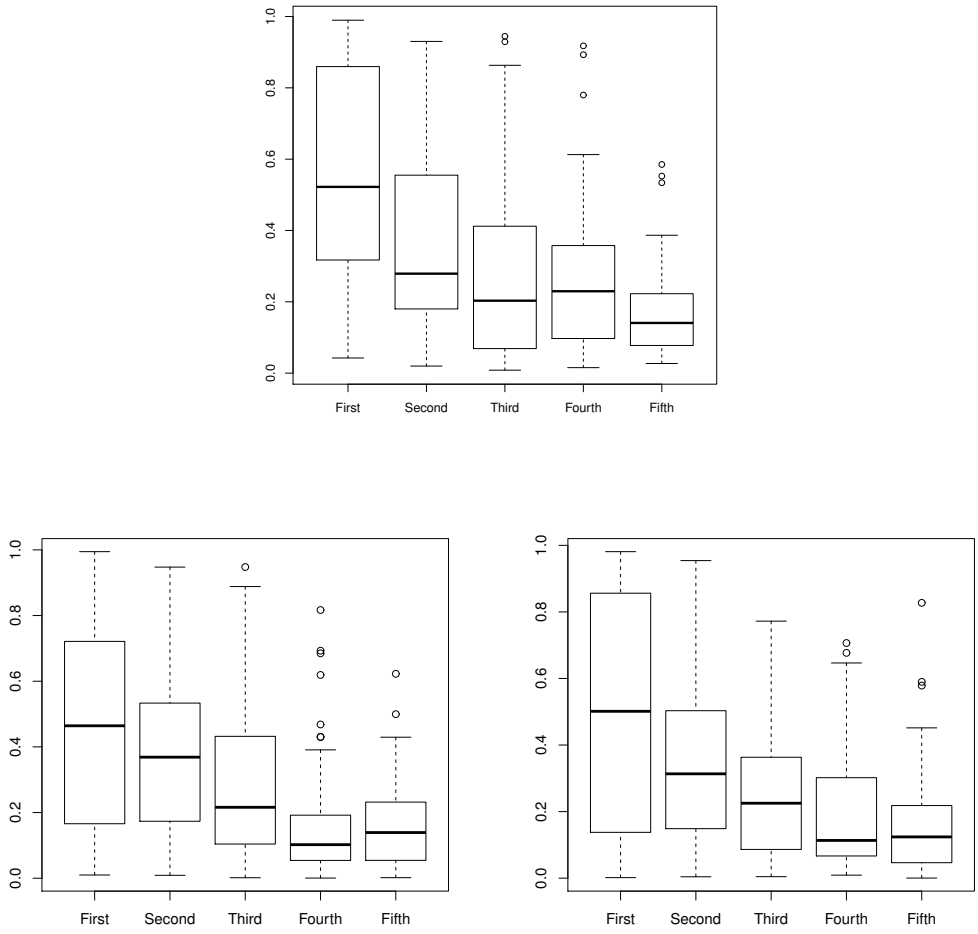

Fig 1: Boxplots for the absolute correlations between the response and the first 5 kernel principal components (second-order polynomial kernel with unit scale and offset) of the predictors in three databases. Upper panel: 33 data sets from the *Arc* database. Lower-left panel: 53 data sets from Johnson and Wichern (2007). Lower-right panel: 54 data sets from CMU StatLib database.

A, B, C will denote sets or matrices; script letters such as $\mathcal{H}, \mathcal{L}, \mathcal{M}$ denote collections of functions or measures; fraktur letters such as $\mathfrak{R}, \mathfrak{F}, \mathfrak{G}$ denote collections of sets. The symbol $\mathbb{R}$ denotes the set of real numbers; the symbol $\mathbb{N}$ denotes the set of natural numbers $\{1, 2, \ldots\}$. The notation $\overset{\mathcal{D}}{=}$ means "equal in distribution".

We suppose that X and Y are defined on a probability space $(\Omega, \mathfrak{F}, P)$ and let $\Omega_X = \{X(\omega) : \omega \in \Omega\}$ denote the range of X. Let $\mathcal{H}$ be a separable Hilbert space, defined over $\mathbb{R}$, whose members are real-valued functions defined on Ω_X. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denote the inner product in $\mathcal{H}$ and let $\| \cdot \|_{\mathcal{H}}$ denote the induced norm.

KPCA is traditionally formulated in the following manner. The first kernel principal component is the function u_1 in $\mathcal{H}$ that solves

$$\underset{f \in \mathcal{H}, \|f\|_{\mathcal{H}}=1}{\operatorname{argmax}} \operatorname{Var}[f(X)] \tag{2}$$

For $k = 2, 3, \ldots$, the kth kernel principal component u_k is the solution to (2) subject to the orthogonality constraints:

$$\operatorname{Cov}[u_k(X), u_i(X)] = 0, \quad i = 1, \ldots, k-1.$$

This is much more general than the classical (linear) principal component analysis because the maximization is carried out among all functions in $\mathcal{H}$ rather than just functions of the form $a^{\mathrm{T}} X$.

The term "kernel" comes from the fact that $\mathcal{H}$ may be taken to be a reproducing kernel Hilbert space derived from a positive definite mapping, or kernel function, $K : \Omega_X \times \Omega_X \to \mathbb{R}$. A kernel function induces the kernel space $\mathcal{H}$ to be the closed linear span of functions with the form

$$a_1 K(\cdot, x_1) + \cdots + a_m K(\cdot, x_m), \quad x_1, \ldots, x_m \in \Omega_X, \ a_1, \ldots, a_m \in \mathbb{R}.$$

The inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is specified by $\langle K(\cdot, x_1), K(\cdot, x_2) \rangle_{\mathcal{H}} = K(x_1, x_2)$. More details are provided in Aronszajn (1950). This particular form of $\mathcal{H}$ has no bearing on the question we are investigating so we only assume $\mathcal{H}$ to be a separable Hilbert space. We nevertheless adopt the term for ease of writing.

The kernel principal components analysis can, similarly to the classical procedure, be represented as an eigen-decomposition problem. Consider the bilinear form $b : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ defined by

$$b(f, g) = \operatorname{Cov}[f(X), g(X)].$$

If b is bounded then there is, see e.g. Hsing and Eubank (2015), a bounded and self-adjoint linear operator $\Sigma : \mathcal{H} \to \mathcal{H}$ which satisfies

$$b(f, g) = \langle f, \Sigma g \rangle_{\mathcal{H}} = \langle \Sigma f, g \rangle_{\mathcal{H}}.$$

This operator Σ is called the covariance operator of X. Under the assumption that Σ is a compact operator, which is always so when $\mathcal{H}$ is finite-dimensional, it has a discrete spectral decomposition $\sum_{i=1}^{\infty} \lambda_i P_i$, where $\lambda_1 > \lambda_2 > \cdots \geq 0$ are real numbers and P_i is the projection onto the linear subspace

$$\ker(\Sigma - \lambda_i I) = \{f \in \mathcal{H} : \Sigma f = \lambda_i f\},$$

where $I : \mathcal{H} \to \mathcal{H}$ is the identity operator and ker denotes the kernel of a linear operator. These projections are orthogonal to each other; i.e. $P_i P_j = 0$ whenever $i \neq j$. It can be shown that any function in $\ker(\Sigma - \lambda_i)$ is the ith kernel principal component (up to rescaling).

The central question pursued in this paper is considered under two levels. The first level is the fully nonparametric regression model

$$Y = f(X) + \varepsilon. \tag{3}$$

It is assumed that $f : \mathbb{R}^p \to \mathbb{R}$ is an arbitrary function in $\mathcal{H}$ and $\varepsilon \perp\!\!\!\perp X$. Our question here is: would the higher-ranking kernel principal components, given a randomly selected regression function f and a randomly selected covariance operator Σ for X, tend to be more correlated with the response than the lower-ranking ones?

The second level is the most general. Suppose Y and X are dependent but the dependence is not restricted by any model (parametric or nonparametric). Our question here is: would the higher-ranking kernel principal components, given a randomly selected conditional distribution for $Y|X$ and a randomly selected covariance operator Σ for X, tend to be more correlated with the response or, more generally, measureable functions of the response than the lower-ranking ones?

The result at the first level is not technically, despite first appearances, a special case of the result at the second level because the conditions we assume to investigate the question are different. We solve the nonparametric regression problem first because it involves most of the techniques needed for the other model.

4. Predictive power of KPCA with finite-dimensional kernels

We first give the definitions of unitarily invariant random functions and operators then we state some of the desirable properties these entities possess. We then establish the predictive potential of finite-dimensional KPCA in a nonparametric regression setting and then in an arbitrary X-Y relationship. We assume that all Hilbert spaces referred to, in this setting, are finite-dimensional so that we can use some sort of uniformity either on the regression coefficient vector or on the covariance operator.

4.1. Unitarily invariant random functions and operators

We give rigorous definitions, based on the notion of unitary invariance, of an arbitrary function $f \in \mathcal{H}$ and an arbitrary covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$. The results in this section depend on some basic facts from linear algebra so proofs are omitted.

Let $n \in \mathbb{N}$ and let $\mathbb{N}_n$ denote $\{1, \ldots, n\}$. Let $\mathcal{H}$ be an n-dimensional Hilbert space with orthonormal basis $\{u_i : i \in \mathbb{N}_n\}$. The inner product between two members of $\mathcal{H}$ is

$$\left\langle \sum_{i \in \mathbb{N}_n} \alpha_i u_i, \sum_{j \in \mathbb{N}_n} \beta_j u_j \right\rangle_{\mathcal{H}} = \sum_{i \in \mathbb{N}_n} \alpha_i \beta_i.$$

Any $f \in \mathcal{H}$ can be expressed as $\sum_{i\in\mathbb{N}_n} \alpha_i u_i$. The sequence of Fourier coefficients (the Fourier sequence) of an element $f \in \mathcal{H}$, with respect to the orthonormal basis of $\mathcal{H}$, is the sequence

$$\{\langle f, u_i\rangle_{\mathcal{H}} : i \in \mathbb{N}_n\} = \left\{ \left\langle \sum_{j\in\mathbb{N}_n} \alpha_j u_j, u_i \right\rangle_{\mathcal{H}} : i \in \mathbb{N}_n \right\} = \left\{ \sum_{i\in\mathbb{N}_n} \alpha_i \delta_{ij} : i \in \mathbb{N}_n \right\} = \{\alpha_i : i \in \mathbb{N}_n\}.$$

We call the α_i's the weights f assigns to the u_i's.

An arbitrary function in $\mathcal{H}$ should intuitively have equal probability of assigning any coefficients to the basis. The magnitude of f is, furthermore, irrelevant as we are concerned with quantities such as $\mathrm{Corr}(Y, u(X)|f)$ rather than f itself so what matters is the relative weights that f gives to each u_i. We have so far considered arbitrariness of f in terms of a basis but it is desirable that a formal definition of f be basis-free. These considerations lead us to the following definition.

Definition 1. *An $\mathcal{H}$-valued random variable f is said to be unitarily invariant if, for any unitary operator $U : \mathcal{H} \to \mathcal{H}$, we have $f \overset{\mathcal{D}}{=} U(f)$.*

The next proposition makes clear why unitary invariance characterises "arbitrariness". We call, in the following, the vector $A = (\alpha_1, \ldots, \alpha_n)^\intercal$ — for any $f = \alpha_1 u_1 + \ldots + \alpha_n u_n \in \mathcal{H}$ — the coordinate of f with respect to $\{u_1, \ldots, u_n\}$. We will, if the basis is obvious from the context, simply say the coordinate of f.

Proposition 1. *An $\mathcal{H}$-valued random variable is unitarily invariant if and only if its coordinate with respect to any orthonormal basis of $\mathcal{H}$ is a spherically distributed $\mathbb{R}^n$-valued random variable.*

We note that, while omitted, the proof of Proposition 1 relies on Lemmas 1 and 2 which are, in some sense, dual to each other.

Lemma 1. *Suppose that $u_1, \ldots, u_n$ form an orthonormal basis for $\mathcal{H}$ and let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a unitary operator. The operator $U : \mathcal{H} \to \mathcal{H}$ defined by*

$$U(h) = \sum_{j\in\mathbb{N}_n} T_j(C) u_j$$

is a unitary operator on $\mathcal{H}$. C is the coordinate of h and $T_j(C)$ denotes the j-th component of $T(C)$.

Lemma 2. *Suppose that $u_1, \ldots, u_n$ form an orthonormal basis for $\mathcal{H}$ and let $U : \mathcal{H} \to \mathcal{H}$ be a unitary operator. Let h be an arbitrary element of $\mathcal{H}$ with coordinate C, and let D be the coordinate of $U(h)$. Define the operator $T : \mathbb{R}^n \to \mathbb{R}^n$ by $T(C) = D$. We have then that T is a unitary operator on $\mathbb{R}^n$.*

Proposition 1 tells us that a unitarily invariant random function f has equal probability of assigning any weights to the basis provided that the norm of the weight vector remains constant. A unitarily invariant random function is therefore, taking into consideration that the norm of f is irrelevant for our discussion, fully arbitrary.

We now define an arbitrary covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$. Our motivation comes from the definition of an orientationally uniform random matrix given in Artemiou and Li (2009, 2013).

Let $\mathfrak{R}$ be the σ-field of Borel sets in $\mathbb{R}$ and let $\mathcal{L}(\mathcal{H})$ be the space of linear operators on $\mathcal{H}$. A random linear operator A is a mapping from Ω to $\mathcal{L}(\mathcal{H})$ such that, for any $f_1, f_2 \in \mathcal{H}$, the function $\omega \mapsto \langle A(\omega)f_1, f_2\rangle_{\mathcal{H}}$ is measurable with respect to $\mathfrak{F}/\mathfrak{R}$. See Skorohod (1976, 1984). We define a random covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ as a bounded and self-adjoint random linear operator such that, for any $f_1, f_2 \in \mathcal{H}$, $\langle \Sigma(\omega)f_1, f_2\rangle_{\mathcal{H}} \geq 0$ almost surely P.

Definition 2. *A random covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ is said to be unitarily invariant if, for any unitary operator $U : \mathcal{H} \to \mathcal{H}$, we have $\Sigma \overset{\mathcal{D}}{=} U\Sigma U^{-1}$.*

This definition intuitively means that the operator Σ is the same random object irrespective of the coordinate system for $\mathcal{H}$ from which it is observed. The eigenspaces, in other words, are equiprobable to have any orientation. This implies that every unit norm function in $\mathcal{H}$ has an equal probability of being the 1st, ..., nth kernel principal component of the operator Σ. This is a fitting description that the distribution of X is chosen without regard to any response variable Y. Notice that, unlike Artemiou and Li (2009), this definition does not require that the eigenvalues and eigenvectors of Σ are independent.

We will use the following technical assumption on Σ, in addition to unitary invariance, when it is treated as a random operator.

Assumption 1. *With probability one, each nonzero eigenvalue of Σ has multiplicity 1.*

This ensures that a nonzero eigenvalue, and its corresponding unit-norm eigenvector, are uniquely determined by Σ (modulo sign for the eigenvector). We note that if u is a random vector with a spherical distribution then so is $-u$. Taking this into consideration along with the fact that squared correlations are unaffected by sign means that taking the eigenvector modulo sign has no bearing on our results. This condition is made to achieve technical clarity and simplicity and we believe it can be avoided by a more elaborate analysis than presented here.

4.2. Nonparametric regression case

We now tackle the first problem stated in section 3. We first give the distribution of the ratio of two Fourier coefficients of a unitarily invariant random function.

Lemma 3. *If $f = \sum_{i\in\mathbb{N}_n} \alpha_i u_i$ is a unitarily invariant random function in $\mathcal{H}$ then the ratio between two random weights of f has a standard Cauchy distribution.*

Remark 1. *We assume that $P(f = 0) = 0$ as if $f = 0$ then $Y = \varepsilon$ so $Y \perp\!\!\!\perp X$. We want to exclude the case of independence. This is assumed throughout this section.*

Proof. The weights of f are, by proposition 1, a spherically symmetric random vector. The ratio of any two components of such a vector, by theorem 1 of Arnold and Brockett (1992), has a standard Cauchy distribution. □

The next theorem assumes f to be a unitarily invariant random function in $\mathcal{H}$ and the covariance operator Σ to be fixed.

Theorem 1. *Let $\lambda_1 > \lambda_2 > \cdots > \lambda_k$ be the k distinct nonzero eigenvalues of Σ. Let u_i be, for each i in $\{1, \cdots, k\}$, the eigenvector corresponding to eigenvalue λ_i. Suppose the nonparametric*

regression model (3) holds where f is a unitarily invariant random element in $\mathcal{H}$ and $f \perp\!\!\!\perp X$. Suppose $\epsilon \perp\!\!\!\perp (X, f)$, $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \tau^2 < \infty$. For $i < j \leq k$, the following holds.

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

Proof. Let $f_1, \ldots, f_n$ be a basis for $\mathcal{H}$ so that $f = \sum_{i \in \mathbb{N}_n} \alpha_i f_i$ where $(\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$ is a spherically distributed random vector. We can hence write $f(X) = \alpha^{\mathrm{T}}\psi(X)$ where $\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$ and $\psi(X) = (f_1(X), \ldots, f_n(X))^{\mathrm{T}}$. Conditioning on f is equivalent to conditioning on α. Similarly $u_i(X) = \beta_i^{\mathrm{T}}\psi(X)$. It can be shown that β_i is the i-th eigenvector of the covariance matrix of $\psi(X)$ and the corresponding eigenvalue is λ_i. Replacing the β in Ni (2011) with our α, the X with our $\psi(X)$, and the U_i^{T} with our β_i, we apply theorem 2 of that paper to obtain the result. The conditional independences assumed there are implied by those we assume here. The result does not depend on the choice of basis as the basis was chosen arbitrarily. □

The interpretation of this theorem is that if nature chooses an arbitrary function f from $\mathcal{H}$ for the nonparametric regression model (3) then, for any $i < j$, the magnitude of the correlation between Y and u_i is larger than the magnitude of the correlation between Y and u_j in $(2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}] \times 100$ percent of the time. We now extend this result to the situation where Σ is also random.

Theorem 2. *Suppose that model (3) holds where X is a random vector whose covariance operator is Σ and Σ is a random covariance operator satisfying assumption 1. Suppose f is a unitarily invariant random function in $\mathcal{H}$ and $\varepsilon \perp\!\!\!\perp X|(f, \Sigma)$, $(f, \Sigma) \perp\!\!\!\perp \varepsilon$, $f \perp\!\!\!\perp (X, \Sigma)$. Suppose that $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \tau^2 < \infty$. For $i < j \leq k$ (k being the number of nonzero eigenvalues of Σ), the following holds.*

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

We briefly, before proving the theorem, explain the various conditional independences assumed. The conditional independence $\varepsilon \perp\!\!\!\perp X|(f, \Sigma)$ is the usual regression assumption except that we have taken into consideration that f and Σ are random. The assumption $(f, \Sigma) \perp\!\!\!\perp \epsilon$ means the parameter of the distribution of X and the regression function are independent of the error. This is reasonable considering that these quantities are usually assumed nonrandom. $f \perp\!\!\!\perp (X, \Sigma)$ means that nature chooses f completely independent of X and its orientation. This then characterises exactly the arbitrariness we desire.

Proof of theorem 2. We apply, by a similiar argument to used to show theorem 1, theorem 2 from Ni (2011) to obtain the result. □

4.3. Arbitrary conditional distribution case

We now deal with the general situation where X and Y are dependent but the dependence is not restricted by any model. This requires a different set of conditions from those assumed in theorem 1. We assume, instead of taking f to be unitarily invariant, Σ to be a unitarily invariant random operator.

We do need, while we assume no model for the relation between X and Y, the following conditional independence

$$Y \perp\!\!\!\perp \Sigma|X. \tag{4}$$

This implies that Y is related to X through its value only and does not depend on its covariance operator. This is a very mild assumption. Consider, e.g., the following scenario (where g is an unknown function and $\epsilon \perp\!\!\!\perp X$)

$$Y = g(X, \varepsilon).$$

The distribution of Y, in this case, conditional on X depends only on the distribution of ϵ and the value of X and not on Σ (except through X). The nonparametric regression model (3) clearly satisfies this condition. The following, where $\mu(\cdot)$ and $\sigma(\cdot)$ are unknown functions, is another example

$$Y = \mu(X) + \sigma(X)\varepsilon, \quad \varepsilon \perp\!\!\!\perp X.$$

We also note that $Y \perp\!\!\!\perp X|g(X)$, for some function g, is a special case of (4). This model is used extensively in the literature on nonlinear sufficient dimension reduction (SDR) and accommodates both of the examples listed above.

We will need the following lemma.

Lemma 4. *Suppose that u_1, u_2 are random functions in $\mathcal{H}$ such that: (1) $\langle u_1, u_2\rangle_{\mathcal{H}} = 0$ and (2) $(u_1, u_2) \overset{\mathcal{D}}{=} (U(u_1), U(u_2))$ for any unitary operator $U : \mathcal{H} \to \mathcal{H}$. For any (nonrandom) function $f \in \mathcal{H}$, $f \neq 0$, the ratio $\langle f, u_1\rangle_{\mathcal{H}}/\langle f, u_2\rangle_{\mathcal{H}}$ has a standard Cauchy distribution.*

This lemma, despite its appearance, is quite different from lemma 3 because, since u_1 and u_2 are random and f is fixed, the vector $(\langle f, u_1\rangle_{\mathcal{H}}, \langle f, u_2\rangle_{\mathcal{H}})^{\mathrm{T}}$ may not have a spherically contoured distribution so the result of Arnold and Brocket (1992) cannot be directly applied. The idea of the proof is to introduce an artificial random function $\tilde{f}$ and then condition on u_1, u_2, so that we "transfer" randomness from (u_1, u_2) to $\tilde{f}$. The method for proving lemma 3 can then be applied.

Remark 2. *We, before proving lemma 4, note that if $X(Z)$ is some random variable X, dependent on another random variable Z, such that $X(Z)|Z = z_1$ has the same distribution for any z_1 then $X(Z)$ has that same distribution unconditionally despite $X(z_1)$ and $X(z_2)$ possibly being different random variables for $z_1 \neq z_2$.*

Proof of lemma 4. Because U^{-1} is also a unitary operator, we have

$$(u_1, u_2) \overset{\mathcal{D}}{=} (U^{-1}(u_1), U^{-1}(u_2)).$$

Consequently,

$$\begin{aligned}(\langle f, u_1\rangle_{\mathcal{H}}, \langle f, u_2\rangle_{\mathcal{H}}) &\overset{\mathcal{D}}{=} \left(\langle f, U^{-1}(u_1)\rangle_{\mathcal{H}}, \langle f, U^{-1}(u_2)\rangle_{\mathcal{H}}\right) \\ &= \left(\langle U(f), u_1\rangle_{\mathcal{H}}, \langle U(f), u_2\rangle_{\mathcal{H}}\right).\end{aligned}$$

The distribution of $(\langle f, u_1\rangle_{\mathcal{H}}, \langle f, u_2\rangle_{\mathcal{H}})$ thus depends on f only through $\|f\|_{\mathcal{H}} \equiv a > 0$. Let $\tilde{f}$ be a random element in $\mathcal{H}$ that is independent of (u_1, u_2) and uniformly distributed on the sphere $\mathcal{S}(a) = \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} = a\}$. We have then that, for any Borel subset A of $\mathbb{R}$ and any nonrandom function f_0 with $\|f_0\|_{\mathcal{H}} = a$,

$$P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A|\tilde{f} = f_0) = P(\langle f_0, u_1\rangle_{\mathcal{H}}/\langle f_0, u_2\rangle_{\mathcal{H}} \in A). \tag{5}$$

This implies

$$P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A|\tilde{f}) = P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A). \tag{6}$$

The right hand side can be rewritten as

$$E[P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A|u_1, u_2)].$$

$\tilde{f}$ is unitarily invariant, because $\tilde{f} \perp\!\!\!\perp (u_1, u_2)$, when conditioning on (u_1, u_2). Moreover, $\langle u_1, u_2\rangle_{\mathcal{H}} = 0$. The ratio $\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}}$, conditioning on (u_1, u_2), has a standard Cauchy distribution — by lemma 3 — regardless of the value of (u_1, u_2). This means that the ratio $\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}}$ is independent of (u_1, u_2) and therefore has a standard Cauchy distribution unconditionally. It follows that

$$P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A) = P_C(A)$$

where $P_C(A)$ is the probability of A under the standard Cauchy distribution. By equalities (5) and (6), and the discussion preceding them, we have

$$\begin{aligned} P(\langle f, u_1\rangle_{\mathcal{H}}/\langle f, u_2\rangle_{\mathcal{H}} \in A) =& P(\langle f_0, u_1\rangle_{\mathcal{H}}/\langle f_0, u_2\rangle_{\mathcal{H}} \in A) \\ =& P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A) = P_C(A). \end{aligned}$$

Hence $\langle f, u_1\rangle_{\mathcal{H}}/\langle f, u_2\rangle_{\mathcal{H}}$ has a standard Cauchy distribution as claimed. □

We now establish the main result of this section.

Theorem 3. *Suppose that Σ is a unitarily invariant covariance operator satisfying assumption 1 and $Y \perp\!\!\!\perp \Sigma|X$. Let $g(Y)$ be any measurable function of Y such that the function $x \mapsto E[g(Y)|X = x]$ belongs to $\mathcal{H}$. We have, with probability 1,*

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

for any two eigen-pairs (λ_i, u_i) and (λ_j, u_j) of Σ satisfying $i < j$ and

$$\mathrm{Cov}[g(Y), u_i(X)|\Sigma] \neq 0,\ \mathrm{Cov}[g(Y), u_j(X)|\Sigma] \neq 0. \tag{7}$$

Remark 3. *We note, before establishing the theorem, that the sign of the eigenvectors is irrelevant as, for any random variables A and B, we have $\mathrm{Corr}^2(A, -B) = \mathrm{Corr}^2(A, B)$. We therefore take eigenvectors modulo sign.*

Proof. We begin by noting that condition (7) implies, with probability 1,

$$\lambda_i > 0,\ \lambda_j > 0,\ \langle f, u_i\rangle_{\mathcal{H}} \neq 0,\ \langle f, u_j\rangle_{\mathcal{H}} \neq 0.$$

We have

$$\begin{aligned} \mathrm{Corr}^2(g(Y), u_i(X)|\Sigma) &= \mathrm{Corr}^2(E(g(Y)|X, \Sigma), u_i(X)|\Sigma) \\ &= \mathrm{Corr}^2(E(g(Y)|X), u_i(X)|\Sigma) \\ &= \mathrm{Corr}^2(f(X), u_i(X)|\Sigma), \end{aligned}$$

where the second equality follows from $Y \perp\!\!\!\perp \Sigma | X$ and f is defined to be the function given by $f(x) = E(g(Y)|X = x)$. This is equal to, by the definition of correlation and the construction of the principal components,

$$\frac{\text{Cov}^2(f(X), u_i(X)|\Sigma)}{\text{Var}(f(X)|\Sigma)\text{Var}(u_i(X)|\Sigma)} = \frac{\langle f, \Sigma u_i\rangle^2}{\lambda_i \text{Var}(f(X)|\Sigma)} = \frac{\lambda_i \langle f, u_i\rangle^2}{\text{Var}(f(X)|\Sigma)}.$$

This implies that

$$\begin{aligned} P\left\{\text{Corr}^2[g(Y), u_i(X)|\Sigma] \geq \text{Corr}^2[g(Y), u_j(X)|\Sigma]\right\} &= P\left\{\frac{\text{Corr}^2[g(Y), u_i(X)|\Sigma]}{\text{Corr}^2[g(Y), u_j(X)|\Sigma]} \geq 1\right\} \\ &= P\left\{\frac{\lambda_i \langle f, u_i\rangle^2}{\lambda_j \langle f, u_j\rangle^2} \geq 1\right\} \\ &= P\left\{-\sqrt{\frac{\lambda_i}{\lambda_j}} \leq \frac{\langle f, u_j\rangle}{\langle f, u_i\rangle} \leq \sqrt{\frac{\lambda_i}{\lambda_j}}\right\}. \end{aligned}$$

By assumption 1, ignoring a probability null set, $(\lambda_i, \lambda_j, u_i, u_j)$ is uniquely determined (modulo sign) by Σ. $(\lambda_i, \lambda_j, u_i, u_j)$ then is a function of Σ. Write this function as $(\lambda_i(\Sigma), \lambda_j(\Sigma), u_i(\Sigma), u_j(\Sigma))$. We have, by the unitary invariance of Σ, $U\Sigma U^{-1} \stackrel{\mathcal{D}}{=} \Sigma$. This implies that

$$(\lambda_i(U\Sigma U^{-1}), \lambda_j(U\Sigma U^{-1}), u_i(U\Sigma U^{-1}), u_j(U\Sigma U^{-1})) \stackrel{\mathcal{D}}{=} (\lambda_i(\Sigma), \lambda_j(\Sigma), u_i(\Sigma), u_j(\Sigma)). \tag{8}$$

We note that

$$\begin{aligned} &\lambda_i(U\Sigma U^{-1}) = \lambda_i(\Sigma), \;\; \lambda_j(U\Sigma U^{-1}) = \lambda_j(\Sigma), \\ &u_i(U\Sigma U^{-1}) = U(u_i(\Sigma)), \quad u_j(U\Sigma U^{-1}) = U(u_j(\Sigma)). \end{aligned}$$

This implies that equality (8) reduces to

$$\{\lambda_i(\Sigma), \lambda_j(\Sigma), U(u_i(\Sigma)), U(u_j(\Sigma))\} \stackrel{\mathcal{D}}{=} \{\lambda_i(\Sigma), \lambda_j(\Sigma), u_i(\Sigma), u_j(\Sigma)\}.$$

The argument for all random elements is now Σ so we drop it in the notation. We can now rewrite the above as $\{\lambda_i, \lambda_j, U(u_i), U(u_j)\} \stackrel{\mathcal{D}}{=} (\lambda_i, \lambda_j, u_i, u_j)$. This implies

$$(u_i, u_j)|(\lambda_i, \lambda_j) \stackrel{\mathcal{D}}{=} [U(u_i), U(u_j)]|(\lambda_i, \lambda_j).$$

By lemma 4, as applied to the conditional probability given (λ_i, λ_j), the conditional distribution of the ratio $\langle f, u_j\rangle_{\mathcal{H}} / \langle f, u_i\rangle_{\mathcal{H}} | (\lambda_i, \lambda_j)$ has a standard Cauchy distribution. This implies

$$P\left(\frac{\langle f, u_j\rangle^2_{\mathcal{H}}}{\langle f, u_i\rangle^2_{\mathcal{H}}} \leq \frac{\lambda_i}{\lambda_j} \middle| \lambda_i, \lambda_j\right) = (2/\pi) \arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]. \tag{9}$$

Taking the unconditional expectation on both sides completes the proof. □

This theorem says that if nature selects an arbitrary covariance operator for X then, regardless of the form of dependence between X and Y, any measurable function $g(Y)$ tends to have a larger squared correlation with u_i than with u_j. The relative frequency of this tendency is

$(2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\} \times 100$ percent.

We next consider the situation where a random relation between X and Y is chosen in addition to choosing a covariance operator Σ for X. The randomness has to be imposed directly on the conditional distribution of $Y|X$ rather than on some aspect of it such as the regression function f in model (3). We therefore introduce the notion of a random conditional distribution of Y given X.

Let $\mathfrak{R}^p$ denote the σ-field of Borel sets in $\mathbb{R}^p$. We recall that a conditional distribution of $Y|X$ is a mapping

$$\kappa : \mathfrak{R} \times \Omega_X \to [0,1]$$

such that (i) for each $\omega \in \Omega$, the function $A \mapsto \kappa(A, X(\omega))$, $\mathfrak{R} \to [0,1]$ is a probability measure on $\mathfrak{R}$; (ii) for each $A \in \mathfrak{R}$, the function $\omega \mapsto \kappa(A, X(\omega))$, $\Omega \to [0,1]$ is a version of the conditional probability $P(Y \in A|X)$. Let $\mathcal{K}$ be the collection of all such mappings κ. We assume, for simplicity, that $\mathcal{H}$ is rich enough to contain all bounded measurable functions of X so that, for each $\kappa \in \mathcal{K}$ and each $A \in \mathfrak{R}$, $\kappa(A, \cdot) \in \mathcal{H}$. We define a random element in $\mathcal{K}$, or a random conditional distribution of $Y|X$, to be a mapping

$$\nu : \Omega \to \mathcal{K}, \quad \omega \mapsto \nu_\omega(\cdot, \cdot),$$

such that, for each $A \in \mathfrak{R}$, the function $\Omega \to \mathcal{H}$, $\omega \mapsto \nu_\omega(A, \cdot)$ is measurable $\mathfrak{B}/\mathfrak{R}^p$. We note that if $\mathcal{H}$ is a set of numbers, rather than a set of functions, then our definition reduces to the classical definition of a random probability measure. See, for example, Kingman (1967). We use the notation $Y|(X, \nu) \sim \nu$ to indicate that a ν is chosen from $\mathcal{K}$ to be the conditional distribution of $Y|X$.

If, for each $A \in \mathfrak{R}$, $\kappa(A, X)$ is almost surely constant, then κ represents the conditional distribution under which X and Y are independent. Let $\mathcal{K}_0$ be the collection of all such κ. The tendency described in this paper occurs only when X and Y are related in *some* way (as both correlations inside the probability are 0 otherwise) so we exclude the case of independence from consideration. This is formulated as $P(\nu \in \mathcal{K}_0) = 0$. This assumption is reasonable. Consider, for example, the simple case where X and Y are standard normal variables. The dependence of X and Y is completely determined by their correlation ρ. The probability of independence of X and Y is 0 if we assume ρ to have a continuous distribution.

Theorem 4. *Suppose that the covariance operator Σ of X is unitarily invariant and satisfies assumption 1. Suppose that ν is a random element of $\mathcal{K}$ such that $P(\nu \in \mathcal{K}_0) = 0$ and*

$$Y|(X, \nu) \sim \nu, \quad \nu \perp\!\!\!\perp (X, \Sigma), \quad Y \perp\!\!\!\perp \Sigma|(X, \nu). \tag{10}$$

Let g be any 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,

$$\mathrm{Cov}[g(Y), u_i(X)|\nu, \Sigma] \neq 0, \ \mathrm{Cov}[g(Y), u_j(X)|\nu, \Sigma] \neq 0. \tag{11}$$

We have then, for any $i < j$,

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

The independence and conditional independence in (10) have a similar interpretation to those in theorem 2: $Y \perp\!\!\!\perp \Sigma|(X,\nu)$ means that the distribution of $Y|(X,\nu)$ does not depend on Σ; $\nu \perp\!\!\!\perp (X,\Sigma)$ means that the relation between X and Y does not depend on X or its covariance operator Σ. The assumption $P(\nu \in \mathcal{K}_0) > 0$ is to ensure functions satisfying (11) exist.

Proof of theorem 4. We begin similarly to the proof of theorem 3 by noting that

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

We also have, since $Y \perp\!\!\!\perp \Sigma|(X,\nu)$, that

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

We see that $\nu \perp\!\!\!\perp (X,\Sigma)$ implies $m_\nu \perp\!\!\!\perp (X,\Sigma)$. We hence, for any $\kappa \in \mathcal{K}$, have that

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

This implies

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

We have, by $\nu \perp\!\!\!\perp (X,\Sigma)$, that

$$\mathrm{Var}[u_i(X)|\nu,\Sigma] = \mathrm{Var}[u_i(X)|\Sigma] = \lambda_i.$$

It follows that

$$\frac{\mathrm{Corr}^2[g(Y), u_i(X)|\nu,\Sigma]}{\mathrm{Corr}^2[g(Y), u_j(X)|\nu,\Sigma]} = \frac{\lambda_i\langle m_\nu, u_i\rangle_{\mathcal{H}}}{\lambda_j\langle m_\nu, u_j\rangle_{\mathcal{H}}}.$$

We see that $m_\nu \perp\!\!\!\perp (u_i, u_j, \lambda_i, \lambda_j)$ implies $m_\nu \perp\!\!\!\perp (u_i,u_j)|(\lambda_i,\lambda_j)$. We hence have, for any $\kappa \in \mathcal{K}$, that

$$P\left(\frac{\langle m_\nu, u_j\rangle^2_{\mathcal{H}}}{\langle m_\nu, u_i\rangle^2_{\mathcal{H}}} < \frac{\lambda_i}{\lambda_j}\middle|\, \nu=\kappa,\lambda_i,\lambda_j\right) = P\left(\frac{\langle m_\kappa, u_j\rangle^2_{\mathcal{H}}}{\langle m_\kappa, u_i\rangle^2_{\mathcal{H}}} < \frac{\lambda_i}{\lambda_j}\middle|\, \lambda_i,\lambda_j\right).$$

We see, by similiar reasoning to that used to show (9), that the right hand side is $(2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]$. We have thus proved that

$$P\left(\frac{\langle m_\nu, u_j\rangle^2_{\mathcal{H}}}{\langle m_\nu, u_i\rangle^2_{\mathcal{H}}} < \frac{\lambda_i}{\lambda_j}\middle|\, \nu,\lambda_i,\lambda_j\right) = (2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}].$$

Taking the conditional expectation on both sides of the above equality completes the proof. □

Remark 4. *The results of theorems 3 and 4 do not require Y to be univariate. It is apparent that Y can be multivariate by appropriately altering the definition of a random conditional distribution for $Y|X$. There may be some technicalities, not immediately apparent to us, worth investigating in future research for whether Y can be an infinite-dimensional object or, even more generally, a generic random variable where the codomain can be any measureable space.*

We now, to test how our theory holds up in real data sets, compare the estimated values of

$$\Pi_{ij} = (2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\}, \quad P_{ij} = P\{\mathrm{Corr}^2(Y, u_i|\nu, \Sigma) \geq \mathrm{Corr}^2(Y, u_j|\nu, \Sigma)\}$$

for each of the three databases described in section 2. These two values should be the same according to our theory. The values of Π_{ij} and P_{ij} are estimated as follows. Let $D_1, \ldots, D_m$ represent the data sets in each database where $m = 33, 53, 54$ for the respective databases. We compute, for each D_k, the ith empirical eigenvalues of the covariance operator induced by the second order polynomial kernel with unit scale and offset that was used as described in section 2. Denote these eigenvalues as $\hat{\lambda}_{ik}$. The value Π_{ij} is then estimated by

$$\hat{\Pi}_{ij} = \frac{2}{\pi m} \sum_{k=1}^{m} \arctan[(\hat{\lambda}_{ik}/\hat{\lambda}_{jk})^{\frac{1}{2}}].$$

The probability P_{ij} is estimated similarly. We compute, for each dataset D_k, the sample correlation between the ith kernel principal component and the response. Denote this correlation by $\hat{\rho}_{ik}$. P_{ij} is then estimated by

$$\hat{P}_{ij} = \frac{1}{m} \sum_{k=1}^{m} I(\hat{\rho}_{ik}^2 \geq \hat{\rho}_{jk}^2).$$

The results are presented in Table 1.

Table 1: Comparison of $\hat{\Pi}_{ij}$ and $\hat{P}_{ij}$ for three databases. "J & W" in the table stands for the database in Johnson & Wichern (2007).

	Arc		J & W (2007)		CMU StatLib	
(i, j)	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$
1 vs 2	0.936	0.727	0.917	0.585	0.969	0.667
2 vs 3	0.875	0.667	0.767	0.585	0.824	0.648
3 vs 4	0.817	0.455	0.771	0.660	0.774	0.593
4 vs 5	0.831	0.636	0.772	0.453	0.772	0.519

Table 1 shows reasonable agreements between $\hat{\Pi}_{ij}$ and $\hat{P}_{ij}$, at least in overall trends. It is interesting to see that $\hat{P}_{ij}$ seems to fluctuate more than $\hat{\Pi}_{ij}$ does. This is perhaps to be expected because, intuitively, Π_{ij} acts as a theoretical expectation of the relative predictive potentials of u_i and u_j based purely on the properties of the predictors themselves. It should also be noted that equality $P_{ij} = \Pi_{ij}$ is *marginal* in nature. A pair of eigenfunctions u_i, u_j are considered without reference to the other eigenfunctions. Perhaps this explains why, in Table 1, a relatively good agreement is sometimes followed by a relatively poor agreement and nonadjacent pairs seem to agree better. We believe, within our current theoretical framework, that it is possible to compute probabilities such as

$$P\{\mathrm{Corr}^2(Y, u_i|\nu, \Sigma) \geq \mathrm{Corr}^2(Y, u_j|\nu, \Sigma) \geq \mathrm{Corr}^2(Y, u_k|\nu, \Sigma)\}$$

for $i < j < k$, and such joint probabilities might improve the agreement.

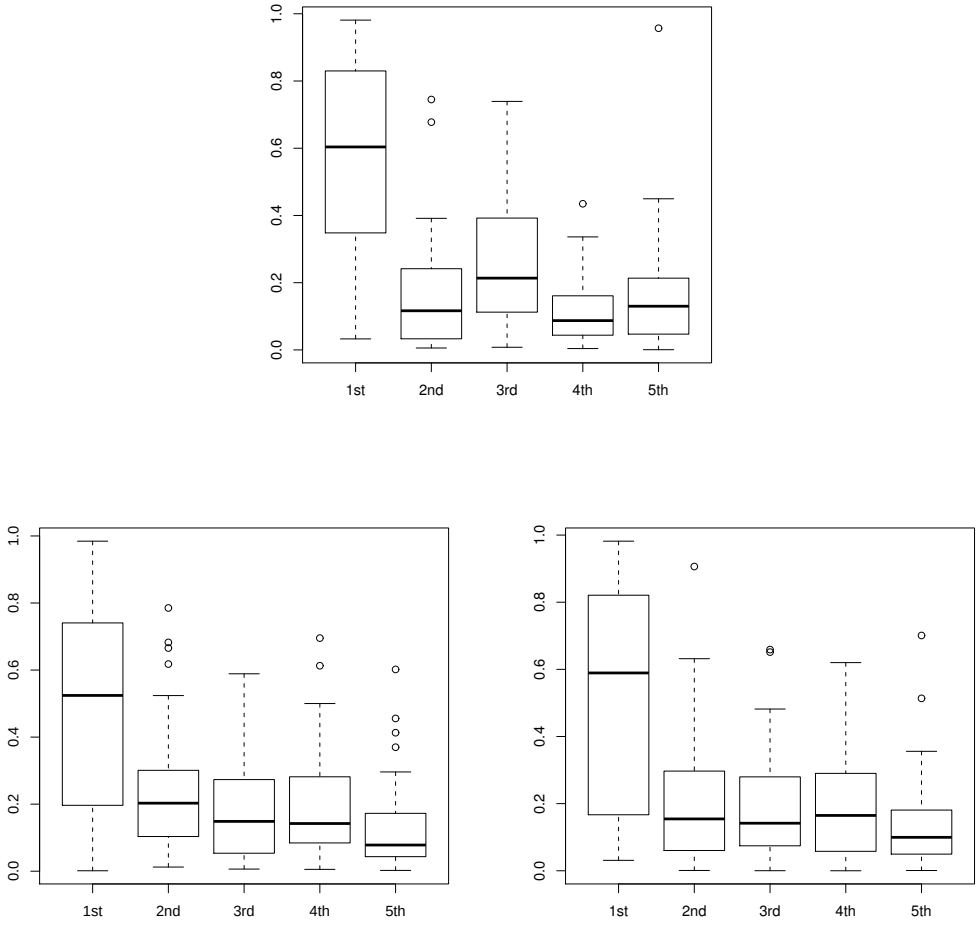

Fig 2: Boxplots for the absolute correlations between the response and the first 5 kernel principal components of the predictors in three databases. Upper panel: 33 data sets from the *Arc* database. Lower-left panel: 53 data sets from Johnson and Wichern (2007). Lower-right panel: 54 data sets from CMU StatLib database. The kernel is the Gaussian kernel with a data-adaptive value for σ.

5. Predictive power of KPCA with infinite-dimensional kernels

The difficulty in extending the previous results to infinite-dimensional KPCA arises from the fact that spherical distribution cannot be extended to infinite-dimensional spaces. One therefore cannot apply the notion of unitary invariance to functions or operators in those spaces as we did in the previous section. We discuss, in this section, a way to extend the results for nonparametric regression to infinite-dimensional KPCA by using a stronger assumption. All Hilbert spaces in this section will be assumed separable and infinite-dimensional although the results apply analogously in finite-dimensional spaces.

Assumption 2. *Let Σ be either a fixed or random compact operator on $\mathcal{H}$. This Σ will be the covariance operator of X. Let u_i be the i-th eigenvector of Σ. We assume that f is a random element of $\mathcal{H}$ such that there exist finite subvectors of the sequence $(\langle f, u_k\rangle)_{k\in\mathbb{N}}$ which are spherically distributed.*

This assumption removes some of the arbitrariness of f given by unitary invariance as it is no longer the case that f is equiprobable in assigning any weight sequence to the basis vectors. We rather have equiprobability for some finite subsets. The ratio of any two components of such a subset has a standard Cauchy distribution. This observation can be used to prove a similar result to the first we derived for the case of nonparametric regression.

Theorem 5. *Suppose that Σ is a compact operator. Let $\lambda_1 > \lambda_2 > \cdots$ be the distinct eigenvalues of Σ. Suppose the nonparametric regression model (3) holds and that f satisfies assumption 2 with $f \perp\!\!\!\perp X$. Suppose also that $\epsilon \perp\!\!\!\perp (X, f)$, $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \tau^2 < \infty$. Let $k \in \mathbb{N}$ and $V = (v_1, \ldots, v_k)$ be a subvector of $\mathbb{N}$ such that $v_1 < \ldots < v_k$ and $(\langle f, u_{v_1}\rangle, \ldots, \langle f, u_{v_k}\rangle)$ has a spherical distribution. We have the following whenever $i < j \leq k$ and $\lambda_{v_j} > 0$.*

$$P\left\{\mathrm{Corr}^2[Y, u_{v_i}(X)|f] \geq \mathrm{Corr}^2[Y, u_{v_j}(X)|f]\right\} = (2/\pi)\arctan[(\lambda_{v_i}/\lambda_{v_j})^{\frac{1}{2}}].$$

We now have the following result as a consequence of theorem 5 when Σ is also random. The proof is similar to theorem 2 with the only difference being that we replace the unitarily invariant assumption with assumption 2. The proof relies on Ni (2011).

Theorem 6. *Suppose that model (3) holds where X is a random vector whose covariance operator is Σ. Σ is assumed to be a random covariance operator satisfying assumption 1. Let $\lambda_1 > \lambda_2 > \cdots$ be the distinct eigenvalues of Σ. Suppose that f satisfies assumption 2 with $f \perp\!\!\!\perp X$. Suppose also that $\epsilon \perp\!\!\!\perp (X, f)$, $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \tau^2 < \infty$. Let $k \in \mathbb{N}$ and $V = (v_1, \ldots, v_k)$ be a subvector of $\mathbb{N}$ such that $v_1 < \ldots < v_k$ and $(\langle f, u_{v_1}\rangle, \ldots, \langle f, u_{v_k}\rangle)$ has a spherical distribution. We have the following whenever $i < j \leq k$ and $\lambda_{v_j} > 0$.*

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

To prove the results for an arbitrary X-Y relationship would require us to impose further restrictions on the covariance operator which may be undesirable. We will therefore leave this for further investigation in the future.

We demonstrate in figure 2 that similar results to the ones we had in figure 1 can be achieved with an infinite-dimensional kernel. These infinite-dimensional kernel principal components are computed using the centered Gram matrix described in Fukumizu, Bach, and Jordan (2009) with

the Gaussian kernel. The parameter σ for the Gaussian kernel is determined adaptively for each data set, according to the following procedure. Let $X_1, \ldots, X_n$ represent the observed predictors of a data set. We use the average of the Euclidean distances of $\{\|X_i - X_j\| : i, j = 1, \ldots, n, i < j\}$ as the value of σ.

6. Conclusion

This paper is an attempt at explaining, through a systematic probabilistic analysis and an empirical investigation of three databases, the phenomenon that the first few kernel principal components tend to have more predictive power for a response variable than lower-ranking ones. This occurs even though the response is not designed in any way to be associated with these components. This phenomenon has long been noticed, and was a focal point of a historical debate, in the context of linear regression and classical principal component analysis. This problem is even more important today because, in its most general form, it lies at the intersection of supervised, unsupervised, and semi-supervised dimension reductions.

This work is a continuation of Li (2007), Artemiou and Li (2009, 2013) and Ni (2011), but goes far beyond the linear regression and classical principal component analysis setting which these authors considered. The most general form of our result states, essentially, that if nature selects an arbitrary distribution for the predictor X and an arbitrary conditional distribution for $Y|X$ then the first few kernel principal components tend to have the most predictive power for a measureable function of Y. This tendency can be explicitly quantified with the Cauchy distribution.

Theorem 4 is the most far-reaching result of this paper but the other results are not special cases of it. This is because they are established under different sets of conditions and in different contexts. The result for nonparametric regression in section 4.2 is derived under the assumption that the regression function f is a unitarily invariant random function with no restrictions on the covariance operator Σ. The result for arbitrary X-Y relation in section 4.3 is based on the assumption that Σ is a unitarily invariant random covariance operator with virtually no restriction on the conditional distribution ν.

We must emphasise again that the tendency studied in this paper, and the previous works cited above, is *probabilistic*. The tendency clearly manifests itself in a collection of data sets but it cannot be used to draw a definite conclusion about any particular data set. It is indeed possible that the response may well be most strongly related to the least important principal components. We should incorporate response data, whenever available, in order to achieve stronger predictive power. This approach is taken, for example, in the field of sufficient dimension reduction.

We conclude with a brief discussion of some avenues for further investigation. We did not consider the case of arbitary X-Y relation in the infinite-dimensional setting as doing so would require the imposition of stricter assumptions on Σ. Future work could consider the naturalness of making such assumptions and what implications they have. We noted in remark (4) that the results for the finite-dimensional arbitrary X-Y relation case did not depend on Y being univariate and noted that future work could consider how abstract an object Y can be. The main results in this paper depend on the ratio of eigenvalues of some covariance operator. We know that, when $\mathcal{H}$ is induced by a reproducing kernel, the covariance operator depends on the choice of kernel. Future work could

consider whether the results can be strengthed for particular choices of the kernel or, more generally, consider how the choice of kernel affects the results. Dicker et al (2017) showed, along these lines, an upper bound on the risk of KPCR estimators with the conventional choice of the first k kernel principal components to be their "regularisation family". The work they did could be investigated under a different choice of what kernel components are retained.

7. Acknowledgements

We would like to thank the reviewers for their constructive comments and suggestions which strengthened an earlier version of the manuscript. We would like, in particular, to thank a reviewer for pointing out the connections between our results on the predictive potential of KPCR under nonparametric regression with the results of Ni (2011). Bing Li's work is supported in part by the U.S. National Science Foundation grant DMS1713078.

References

1. Alter, O., Brown, P. and Botstein, D. (2000). Singular value decomposition for gene-wide expression data processing and modelling. *Proceedings of the National Academy of Science*, **97**, 10101–10106.
2. Arnold, B. C. and Brockett, P. L. (1992). On distributions whose component ratios are Cauchy. *American Statistician*, **46**, 25 – 26.
3. Artemiou, A. and Dong, Y. (2016). Sufficient dimension reduction via principal Lq support vector machine. *Electronic Journal of Statistics*, **10**, 783–805
4. Artemiou, A. and Li, B. (2009). On principal components and regression: a statistical explanation of a natural phenomenon. *Statistica Sinica*, **19**, 1557 – 1566.
5. Artemiou A, and Li, B. (2013). Predictive power of principal components for single-index model and sufficient dimension reduction. *Journal of Multivariate Analysis*, **119**, 176–184
6. Bura, E. and Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, **19**, 1252–1258.
7. Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Math. Biosci.*, **176**, 123–144.
8. Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**, 1–40.
9. Cox, D. R. (1968). Notes on some aspects of regression analysis. *Journal of the Royal Statistical Society, Ser. A,*, **131**, 265–279.
10. Dicker, L., Foster, D. and Hsu, D. (2017). Kernel ridge vs principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, **11**, 1022–1047
11. Fang, K., Kotz, S., and Ng, K. (1990). *Symmetric multivariate and related distributions.* Chapman and Hall. Ltd., London.
12. Fukumizu, Bach, and Jordan (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *The Journal of Machine Learning Research*, **5**, 73–99.
13. Fukumizu, Bach, and Jordan (2009). Kernel dimension reduction in regression. *Annals of Statistics.* **4** 1871 – 1905.
14. Hall, P. and Yang, Y.-J. (2010). Ordering and selecting components in multivariate or functional data linear prediction. *Journal of the Royal Statistical Society, Series B*, **72**, 93 – 110.

15. Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*. **84**, 502 – 516.
16. Hawkins, D. M. and Fatti, L. P. (1984). Exploring multivariate data using the minor principal components. *The Statistician*, **33**, 325–338.
17. Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley and Sons, UK.
18. Hotelling, H. (1957). The relationship of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, **10**, 69–79.
19. Johnson, R. A. & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Education, Inc.
20. Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, **31**, 300–303.
21. Jolliffe, I. T. (2002). *Principal component analysis*. Springer.
22. Jones, B. and Artemiou, A. (2019+). On principal components regression with hilbertian predictors. To appear in the *Annals of the Institute of Statistical Mathematics*.
23. Kendall, M. G. (1957). *A course in Multivariate Analysis*. London: Griffin.
24. Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics*, **21**, 59 – 78.
25. Lee, K.-Y., Li, B. and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, **41**, 221-–249.
26. Li, B. (2007). Comment: Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**, 32–35.
27. Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machine for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, **39**, 3182–3210
28. Li, B. and Song, J. (2017). Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, **45**, 1059–1095
29. Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**, 3406–3412.
30. Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Reading, Massachusetts: Addison-Wesley.
31. Muirhead, R.J. (1982). *Aspects of multivariate statistical theory*. John Wiley and Sons, New York.
32. Ni, L. (2009). Principal component regression revisited. **21**, 741—747.
33. Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of Royal Statistical Society, Series B*, **53**, 233 – 243.
34. Schölkopf, B., Smola, A., Müller, K.-R. (1997). Kernel principal component analysis. *Artificial Neural Networks*. 583 – 588.
35. Schölkopf, B., Smola, A., Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*. **10**. 1299 – 1319.
36. Scott, D. (1992). *Multivariate Density Estimation*. New York: Wiley.
37. Shi, T., Belkin, M., and Yu, B. (2009). Data Spectroscopy: Eigenspaces of Convolution Operators and Clustering. *Annals of Statistics*, **37**, 3960 – 3984.
38. Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, **24**, 1 – 24.
39. Skorohod, A. V. (1976). Random operators in a Hilbert space. *Lecture Notes in Mathematics*.

550. 567 – 591.
40. Skorohod, A. V. (1984). *Random Linear Operators*. D. Reidel Publishing Company, Dordrecht, Holland.
41. Vapnik, V. (1998). *Statistical Learning Theory*. Wiley Intersience.
42. Yeh, Y. R., Huang S. Y., and Lee Y. J. (2009). Nonlinear Dimension Reduction with Kernel Sliced Inverse Regression. *IEEE transactions on Knowledge and Data Engineering*. **21**. 1590–1603