Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

On principal components regression with hilbertian predictors

Ben Jones · **Andreas Artemiou**

Received: date / Revised: date

Abstract We demonstrate that, in a regression setting with a Hilbertian predictor, a response variable is more likely to be more highly correlated with the leading principal components of the predictor than with trailing ones. This is despite the extraction procedure being unsupervised. Our results are established under the conditional independence model, which includes linear regression and single-index models as special cases, with some assumptions on the regression vector. These results are a generalisation of earlier work which showed that this phenomenon holds for predictors which are real random vectors. A simulation study is used to quantify the phenomenon.

Keywords conditional independence · hilbertian random variables · principal components regression · elliptical distributions · cauchy distribution

1 Introduction

Theoretical and computational issues are common in regression settings that have a large number of predictors. To address this, methods that reduce the number of predictors have been proposed. There are two main classes of such methods: *feature selection*, and *dimension reduction*. Feature selection works by choosing a subset of the original variables, whereas dimension reduction creates a set of functions of them.

The most commonly used dimension reduction method is *principal component analysis*. This procedure extracts linear combinations of the predictors which have maximal variance. In *principal component regression*, one uses a subset — conventionally the first few — of the principal components as the new predictors on which to regress the response. This technique has been questioned over the years, for example by Cox (1968), as there is no obvious reason for the first few principal components to be more correlated with the response variable than the last few. While this is the case, the practice is still common because often — as observed in simulations and real-world analyses — higher-ranking components actually are the most correlated with the response. A thought-provoking historical account of this long running debate surrounding principal component regression can be found in Cook (2007).

It is well-known that principal component regression is not guaranteed to select the components most correlated with the response but the phenomenon has only recently been quantified. For example, Hall and Yang (2010) considered whether selecting a different subset of the principal components, rather than the first few, is a better option given no other information about how the response and the principal components of the predictors are related. Under a linear regression model, they gave a minimax result by establishing that the largest mean squared difference between the fitted values and the signal possible

Ben Jones
School of Mathematics, Cardiff University
E-mail: JonesBL7@cardiff.ac.uk

Andreas Artemiou
School of Mathematics, Cardiff University
E-mail: ArtemiouA@cardiff.ac.uk

at each step is minimised by choosing the conventional subset of principal components. Moreover, they emphasise that this result holds for all sample sizes, not just asymptotically.

B.Li (2007), taking a different perspective, posed the following conjecture.

Conjecture 1 If nature selects a random covariance matrix for the predictor $\mathbf{X}$ and, independently, randomly selects a linear relation between $\mathbf{X}$ and Y, then the first principal component of $\mathbf{X}$ tends to have the largest correlation with Y among all the principal components.

Artemiou and Li (2009) proved a weaker version of this conjecture (see Theorem 1) by showing that, in a linear regression setting, the principal components of $\mathbf{X}$ with higher ranks tend to have stronger correlations with Y than those of lower ranks. This result is established under an exchangability assumption (see Assumption 1) on the eigenvalues and eigenvectors of the covariance matrix of $\mathbf{X}$. Ni (2011), following up on that result, used a rotational invariance assumption on the regression coefficients (see Assumption 2), and then on the covariance matrix of $\mathbf{X}$ (see Assumption 3) to derive exact forms for this probability (see Theorems 2 and 3). These results were further developed in Artemiou and Li (2013) to cover the more general conditional independence model (see Theorem 4). We note in passing that the conditional independence model is frequently used in the field of sufficient dimension reduction — a supervised framework for dimension reduction — and subsumes the linear model as a subcase. For an exposition of this field, consult Li (2018).

Both principal components analysis and principal components regression can be formulated more generally than most presentations of them detail. Ramsay and Silverman (2005) gave them suitable formulations for when the data are elements of a function space. These procedures work in essentially the same way as the classical formulations except that the inner product of $\mathbb{R}^p$ is replaced with the inner product of the function space in which the data lie. These procedures can be further generalised to separable Hilbert-space valued random variables in the same manner (the spaces are defined over real numbers for most statistical purposes). The separable Hilbert-space valued random variable setting subsumes the functional data scenario as a special subcase — so, while the functional data case is our main interest, we work in the more abstract setting in this paper. For an exposition of the theory behind functional data analysis, consult Hsing and Eubank (2015). Other dimension reduction methods for functional data have been proposed — see, for example, Ferré and Yao (2003) and Li and Song (2017) as examples in the sufficient dimension reduction framework.

In this paper, we establish a similiar predictive tendency for principal components regression when the predictor is a random variable in a separable infinite-dimensional Hilbert space. We assume that a conditional independence model relates the response and the predictor. The previous works relied on the notion of a spherical distribution of the regression coefficients. As we discuss in Appendix A, the infinite-dimensionality presents a challenge here as a spherical distribution does not exist in such spaces because of the noncompactness of the identity operator. We take the space as infinite-dimensional to motivate this discussion — it is not a requirement for the results we derive. We instead use similar assumptions to obtain our weaker, but alike in spirit, results (Theorems 5 and 6).

In Section 2, we review some of the most important previous results. In Section 3, we first present some lemmas needed to prove our results which are then given. We conclude with a discussion in Section 4. Essential definitions are provided in Appendix A, and proofs are supplied in Appendix B.

Remark 1 For real random vector data, throughout this paper, we use $\mathbf{X}$ as our p-dimensional predictor, β as the p-dimensional vector of regression coefficients (we use this terminology even when not considering a linear regression model), and $\mathbf{\Sigma}$ as the covariance matrix of $\mathbf{X}$. We use $\mathcal{H}$ to denote a separable infinite-dimensional Hilbert space. In the infinite-dimensional setting, we use X as the predictor and we use g and $\mathbf{\Gamma}$ as the analogues of β and $\mathbf{\Sigma}$ respectively. In both settings, we use Y as the response variable. We note that β, g, $\mathbf{\Sigma}$, or $\mathbf{\Gamma}$ may be random — we will specify when this is the case. When treated as random, the regression coefficients and the covariance matrix/operator will be assumed independent. The regression coefficients will also be assumed to be independent of the predictor.

2 Review of previous results

Artemiou and Li (2009) gave Definition 1 as a notion of a uniform distribution on the set of $p \times p$ positive definite matrices.

Definition 1 A $p \times p$ positive definite random matrix $\mathbf{M}$ has an *orientationally uniform distribution* if it can be decomposed as $\mathbf{M} = \sum_{i=1}^{p} \lambda_i (\mathbf{v}_i \otimes \mathbf{v}_i)$ where $(\lambda_1, \ldots, \lambda_p)$ are positive exchangeable random variables, $(\mathbf{v}_1, \ldots, \mathbf{v}_p)$ are exchangeable random vectors, and $(\lambda_1, \ldots, \lambda_p) \perp\!\!\!\perp (\mathbf{v}_1, \ldots, \mathbf{v}_p)$.

Intuitively, this means that the relative positions of the eigenvalues and eigenvectors of $\mathbf{M}$ can be freely permutated without changing the distribution of $\mathbf{M}$. This led them naturally to Assumption 1.

Assumption 1 *The covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}$ is assumed to be an orientationally uniform random matrix. In other words, $\mathrm{Var}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{\Sigma}$ almost surely and $\mathbf{\Sigma}$ has an orientationally uniform distribution. Here we are saying that $\mathbf{\Sigma}$ will be fixed, but that it has an orientationally uniform distribution. I think this is what confused one of the reviewers*

Assumption 1 implies that if $\mathbf{X}$ satisfies $\mathbb{E}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{\Sigma}$ almost surely then any random variable among $\mathbf{v}_1{}^{\mathrm{T}}\mathbf{X}, \ldots, \mathbf{v}_p{}^{\mathrm{T}}\mathbf{X}$, where $\mathbf{v}_k$ is the k^{th} most dominant eigenvector of $\mathbf{\Sigma}$, is equally likely to be the $1^{\mathrm{st}}, 2^{\mathrm{nd}}, \ldots,$ or p^{th} principal component of $\mathbf{X}$. Using Assumption 1, and assuming the regression coefficients are random, Artemiou and Li (2009) proved Theorem 1.

Theorem 1 *Suppose the following:*

1. *Assumption 1 holds*
2. $\mathbb{E}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{0}$ *and* $\mathrm{Var}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{\Sigma}$ *almost surely*
3. $Y = \beta^T \mathbf{X} + \varepsilon$ *where* $\beta \perp\!\!\!\perp (\mathbf{X}, \mathbf{\Sigma})$, $\varepsilon \perp\!\!\!\perp (\beta, \mathbf{X}, \mathbf{\Sigma})$, $\mathbb{E}(\varepsilon) = 0$, *and* $\mathrm{Var}(\varepsilon)$ *is finite*
4. $\mathbb{P}(\beta \in G) > 0$ *for any nonempty open set* $G \subset \mathbb{R}^p$

Let $\mathbf{v}_k$ be the k^{th} most dominant eigenvector of $\mathbf{\Sigma}$. Then, letting $\rho_k(\beta, \mathbf{\Sigma}) = \mathrm{Corr}^2(Y, \mathbf{v}_k{}^T\mathbf{X} \mid \beta, \mathbf{\Sigma})$, for $i < j$ we have the following:

$$\mathbb{P}(\rho_i(\beta, \mathbf{\Sigma}) \geq \rho_j(\beta, \mathbf{\Sigma})) > \frac{1}{2}$$

Under Assumptions 2 and 3 (considered separately), Ni (2011) proved Theorems 2 and 3. The second assumption was used tacitly. Notice that Theorem 2 gives a uniformity assumption for the regression coefficients and conditions on the covariance matrix, whereas Theorem 3 puts the uniformity on the covariance matrix and conditions on the regression coefficients.

Assumption 2 *The regression vector β has a spherically symmetric distribution — that is $\beta \overset{\mathcal{D}}{=} A\beta$ for any $A \in \mathbb{O}(p)$. Equivalently the characteristic function of $\beta - \mathbb{E}(\beta)$ has the form*

$$\psi(s) = \varphi(s^T s)$$

for all $s \in \mathbb{R}^p$, where p is the dimension of the space in which β lies, and φ is a univariate function.

Assumption 3 *The random covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}$ is symmetric and its distribution is invariant under orthogonal transformations — that is, for any $U \in \mathbb{O}(p)$, $\mathbf{\Sigma} \overset{\mathcal{D}}{=} U\mathbf{\Sigma}U^T$. Moreover the eigenvalues of $\mathbf{\Sigma}$ are positive and distinct.*

Theorem 2 *Suppose the following:*

1. *Assumption 2 holds*
2. $\mathbb{E}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{0}$ *and* $\mathrm{Var}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{\Sigma}$ *almost surely*
3. $Y = \beta^T \mathbf{X} + \varepsilon$ *where* $\beta \perp\!\!\!\perp (\mathbf{X}, \mathbf{\Sigma})$, $\varepsilon \perp\!\!\!\perp (\beta, \mathbf{X}, \mathbf{\Sigma})$, $\mathbb{E}(\varepsilon) = 0$, *and* $\mathrm{Var}(\varepsilon)$ *is finite*

Let $\mathbf{v}_k$ and λ_k be the k^{th} most dominant eigenvector and eigenvalue of $\mathbf{\Sigma}$. Then, letting $\rho_k(\mathbf{\Sigma}) = \mathrm{Corr}^2(Y, \mathbf{v}_k{}^T\mathbf{X} \mid \mathbf{\Sigma})$, for $i < j$, provided $\lambda_j > 0$, we have the following:

$$\mathbb{P}(\rho_i(\mathbf{\Sigma}) \geq \rho_j(\mathbf{\Sigma})) = \frac{2}{\pi}\mathbb{E}\left(\arctan\left(\sqrt{\frac{\lambda_i}{\lambda_j}}\right)\right) \geq \frac{1}{2}$$

Theorem 3 *Suppose the following:*

1. *Assumption 3 holds*

2. $\mathbb{E}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{0}$ *and* $\mathrm{Var}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{\Sigma}$ *almost surely*
3. $Y = \beta^T\mathbf{X} + \varepsilon$ *where* $\beta \perp\!\!\!\perp (\mathbf{X}, \mathbf{\Sigma})$, $\varepsilon \perp\!\!\!\perp (\beta, \mathbf{X}, \mathbf{\Sigma})$, $\mathbb{E}(\varepsilon) = 0$, *and* $\mathrm{Var}(\varepsilon)$ *is finite.*

Let $\mathbf{v}_k$ *and* λ_k *be the* k^{th} *most dominant eigenvector and eigenvalue of* $\mathbf{\Sigma}$. *Then, letting* $\rho_k(\beta) = \mathrm{Corr}^2(Y, \mathbf{v}_k{}^T\mathbf{X} \mid \beta)$, *for* $i < j \leq p$ *(recall* p *is the dimension of the space) we have the following:*

$$\mathbb{P}(\rho_i(\beta) \geq \rho_j(\beta)) = \frac{2}{\pi}\mathbb{E}\left(\arctan\left(\sqrt{\frac{\lambda_i}{\lambda_j}}\right)\right) > \frac{1}{2}$$

In Theorems 1, 2 and 3, the authors assumed a linear regression setting. Artemiou and Li (2013) examined this probabilistic tendency under the more general conditional independence model $Y \perp\!\!\!\perp \mathbf{X} \mid \beta^{\mathrm{T}}\mathbf{X}$, which subsumes the linear regression model as a special case. The most general result they showed was Theorem 4.

Theorem 4 *Suppose:*

1. $Y \perp\!\!\!\perp \mathbf{X} \mid (\beta^T\mathbf{X}, \beta, \mathbf{\Sigma})$
2. *Almost surely,* $\mathrm{Var}(Y \mid \beta, \mathbf{\Sigma})$ *is finite and* $\mathrm{Cov}(Y, \beta^T\mathbf{X} \mid \beta, \mathbf{\Sigma})$ *is nonzero*
3. $\mathbb{E}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{0}$ *and* $\mathrm{Var}(\mathbf{X} \mid \mathbf{\Sigma}) = \mathbf{\Sigma}$ *almost surely*
4. $\beta \perp\!\!\!\perp (\mathbf{X}, \mathbf{\Sigma})$
5. $\mathbb{E}(\mathbf{X} \mid \beta^T\mathbf{X}, \beta, \mathbf{\Sigma})$ *is linear in* $\beta^T\mathbf{X}$
6. *Either assumption 2 or 3 holds.*

Let $\mathbf{v}_k$ *and* λ_k *be the* k^{th} *most dominant eigenvector and eigenvalue of* $\mathbf{\Sigma}$. *Let* $\rho_k(\beta, \mathbf{\Sigma}) = \mathrm{Corr}^2(Y, \mathbf{v}_k{}^T\mathbf{X} \mid \beta, \mathbf{\Sigma})$. *Then for* $i < j$, *provided* $\lambda_j > 0$, *we have the following:*

$$\mathbb{P}(\rho_i(\beta, \mathbf{\Sigma}) \geq \rho_j(\beta, \mathbf{\Sigma})) = \frac{2}{\pi}\arctan\left(\sqrt{\frac{\lambda_i}{\lambda_j}}\right) \geq \frac{1}{2}$$

Remark 2 We note that condition 5 is commonly assumed in the sufficient dimension reduction literature. It is known to hold for $\mathbf{X}$ with an elliptically symmetric distribution — see e.g. Y.Li (2007).

3 Main Results

We start this section by giving two lemmas (Lemmas 1 and 2) which are needed to prove our main results (Theorems 5 and 6). We first recall a more general form of the linearity assumption, Condition 5 in Theorem 4, which is common in the literature. Y.Li (2007) has shown, in analogy to the case of vector data, that it holds when X has an elliptically symmetric distribution.

Assumption 4 *For all* $f \in \mathcal{H}$, $\mathbb{E}(\langle f, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma})$ *is linear in* $\langle g, X\rangle_{\mathcal{H}}$. *That is, for any fixed* $f \in \mathcal{H}$, *there is a constant* $\alpha \in \mathbb{R}$ *which gives the following:*

$$\mathbb{E}(\langle f, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma}) = \alpha\langle g, X\rangle_{\mathcal{H}}$$

We could instead make the more general assumption that $\mathbb{E}(\langle f, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma})$ is affine in $\langle g, X\rangle_{\mathcal{H}}$. In other words, there are constants $\alpha_0 \in \mathbb{R}$ and $\alpha_1 \in \mathbb{R}$ such that

$$\mathbb{E}(\langle f, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma}) = \alpha_0 + \alpha_1\langle g, X\rangle_{\mathcal{H}}.$$

If we do this, we can show first that:

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(\langle f, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma})) &= \mathbb{E}\left(\langle f, \mathbb{E}(X \mid \langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma})\rangle_{\mathcal{H}}\right) \\
&= \langle f, \mathbb{E}(\mathbb{E}(X \mid \langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma}))\rangle_{\mathcal{H}} \\
&= \langle f, \mathbb{E}(X \mid g, \mathbf{\Gamma})\rangle_{\mathcal{H}} \\
&= \langle f, \mathbb{E}(X \mid \mathbf{\Gamma})\rangle_{\mathcal{H}} \\
&= \langle f, 0\rangle_{\mathcal{H}} = 0
\end{aligned} \tag{1}$$

where the first and second equalities follow from Equation 5, the third equality from the law of total expectation, the fourth follows as $g \perp\!\!\!\perp (X, \mathbf{\Gamma})$, and the fifth follows as X is centered given $\mathbf{\Gamma}$. We can also show that:

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(\langle f, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma})) &= \mathbb{E}(\alpha_0 + \alpha_1 \langle g, X\rangle_{\mathcal{H}}) \\
&= \alpha_0 + \alpha_1 \mathbb{E}(\langle g, X\rangle_{\mathcal{H}}) \\
&= \alpha_0 + \alpha_1 \mathbb{E}(\mathbb{E}(\langle g, X\rangle_{\mathcal{H}} \mid \mathbf{\Gamma})) \\
&= \alpha_0 + \alpha_1 \mathbb{E}(\langle g, \mathbb{E}(X \mid \mathbf{\Gamma})\rangle_{\mathcal{H}}) \\
&= \alpha_0 \ (\text{as } \mathbb{E}(X \mid \mathbf{\Gamma}) = 0) \qquad (2)
\end{aligned}$$

PROPOSED ALTERNATIVE FROM LINE 2 OF ABOVE:

$$\begin{aligned}
\mathbb{E}(\langle g, X\rangle_{\mathcal{H}}) &= \mathbb{E}(\mathbb{E}(\langle g, X\rangle_{\mathcal{H}} \mid g)) \\
&= \mathbb{E}(\mathbb{E}(\langle g, X\rangle_{\mathcal{H}} \mid g, \mathbf{\Gamma})) \\
&= \mathbb{E}(\langle g, \mathbb{E}(X \mid g, \mathbf{\Gamma})\rangle_{\mathcal{H}}) \\
&= \mathbb{E}(\langle g, \mathbb{E}(X \mid \mathbf{\Gamma})\rangle_{\mathcal{H}}) \\
&= 0
\end{aligned}$$

where the first equality holds by the law of total expectation, the second as $g \perp\!\!\!\perp (X, \mathbf{\Gamma})$, the third by Equation 5, the fourth as $g \perp\!\!\!\perp (X, \mathbf{\Gamma})$, and the fifth as $\mathbb{E}(X \mid \mathbf{\Gamma}) = 0$. Let me know if I am putting nonsense here

Combining these two, we see that α_0 is necessarily zero. We thus assume the linear version without loss of generality.

Lemma 1 is an adaptation of Theorem 1 from Dauxois et al. (2001) allowing for random g and $\mathbf{\Gamma}$.

Lemma 1 *Suppose that Assumption 4 holds. Then we have, almost surely, the following:*

$$\mathbb{E}(X \mid \langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma}) \in \text{Span}(\mathbf{\Gamma} g)$$

For reasons discussed in Appendix A, a spherical distribution cannot be defined directly on an infinite-dimensional space. We propose two ways around this issue: (i) we can consider the coefficients of a random element, in the principal component basis, and suppose that the first n of them are spherically distributed whatever the value of n and (ii) we can use elliptical distributions instead (see Appendix A) and aim for a more general, but weaker, result.

The main results of this paper make use of Theorem 1 from Arnold and Brockett (1992), which states that the ratio of any two components of a spherically distributed random vector has a standard Cauchy distribution. To use that result, we make Assumption 5. A similar assumption is used by Kingman (1972) who showed that if S is a sequence of random variables with all finite truncated subsequences being spherically symmetric, then there exists a random variable V such that, when conditioned on V, all the terms of S are independent and normally distributed with mean 0 and variance V.

Assumption 5 *g is such that for all $n \in \mathbb{N}$, $T_n := (\langle \phi_1, g\rangle_{\mathcal{H}}, \langle \phi_2, g\rangle_{\mathcal{H}}, \ldots, \langle \phi_n, g\rangle_{\mathcal{H}})^T$ is spherically distributed where ϕ_k is the k^{th} most dominant eigenvector of $\mathbf{\Gamma}$*

The following lemma demonstrates that when we have an sequence that is elliptically distributed then any terminating subsequence is also elliptically distributed.

Lemma 2 *Assume that g has an elliptically symmetric distribution. Then the sequence $S := (\langle \phi_k, g\rangle_{\mathcal{H}})_{k \in \mathbb{N}}$ is elliptically distributed. Furthermore, for all $n \in \mathbb{N}$, $T_n := (\langle \phi_1, g\rangle_{\mathcal{H}}, \langle \phi_2, g\rangle_{\mathcal{H}}, \ldots, \langle \phi_n, g\rangle_{\mathcal{H}})^T$ is elliptically distributed. ϕ_k is the k^{th} most dominant eigenvector of $\mathbf{\Gamma}$.*

We now, having given these supporting lemmas, present the main results of this paper. First we give a result using Assumption 5 and then we replace that with the ellipticity of g to get a more general result.

Theorem 5 *Suppose:*

1. $Y \perp\!\!\!\perp X \mid (\langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma})$
2. *Almost surely,* $\mathrm{Var}(Y \mid g, \mathbf{\Gamma})$ *is finite and* $\mathrm{Cov}(Y, \langle g, X\rangle_{\mathcal{H}} \mid g, \mathbf{\Gamma})$ *is nonzero*
3. *Almost surely,* $\mathbb{E}(X \mid \mathbf{\Gamma}) = 0$ *and* $\mathrm{Var}(X \mid \mathbf{\Gamma}) = \mathbf{\Gamma}$
4. $g \perp\!\!\!\perp (X, \mathbf{\Gamma})$
5. *Assumptions 4 and 5 hold*

Let ϕ_k *and* λ_k *be the* k^{th} *most dominant eigenvector and eigenvalue of* $\mathbf{\Gamma}$*. Let* $\rho_k(g, \mathbf{\Gamma}) = \mathrm{Corr}^2(Y, \langle \phi_k, X\rangle_{\mathcal{H}} \mid g, \mathbf{\Gamma})$*. Then, for* $i < j$ *with* $\lambda_j > 0$*:*

$$\mathbb{P}(\rho_i(g, \mathbf{\Gamma}) > \rho_j(g, \mathbf{\Gamma})) = \frac{2}{\pi} \arctan\left(\sqrt{\frac{\lambda_i}{\lambda_j}}\right)$$

In the proof of Theorem 5, Assumption 5 is used at the last step only in order to make use of Theorem 1 from Arnold and Brockett (1992). We now consider what happens when Assumption 5 is replaced by the ellipticity of g. This case was not studied previously in the literature as spherical distributions are available in finite-dimensional spaces. While we have assumed that $\mathcal{H}$ is infinite-dimensional throughout this paper, it is not necessary for the proofs of our results. As, for finite-dimensional spaces, the class of elliptical distributions contains the class of spherical distributions, we believe that it is of interest to consider under what conditions this larger class has the desired lower bound.

We first revisit Theorem 2 from Arnold and Brockett (1992). If we let $A = (A_1, \dots A_n)^{\mathrm{T}}$ be an elliptically distributed random vector, then for any i and j:

$$A_{(ij)} = \begin{pmatrix} A_i \\ A_j \end{pmatrix} = \mathbf{C}_{(ij)} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$$

where $\binom{B_1}{B_2}$ — denoted henceforth by B — has a spherical distribution and $\mathbf{C}_{(ij)}$ is some upper triangular matrix

$$\mathbf{C}_{(ij)} = \begin{pmatrix} a_{ij} & b_{ij} \\ 0 & c_{ij} \end{pmatrix}.$$

Theorem 2 of Arnold and Brockett (1992) states that A_i/A_j has a noncentral Cauchy distribution with scale and location parameters

$$\gamma_{ij} = a_{ij}/c_{ij} \text{ and } \kappa_{ij} = b_{ij}/c_{ij} \tag{3}$$

We note that since B is spherically distributed then

$$\mathbf{\Sigma}_B = \begin{pmatrix} \sigma_B^2 & 0 \\ 0 & \sigma_B^2 \end{pmatrix}$$

and

$$\mathrm{Var}\left(A_{(ij)}\right) = \mathbf{\Sigma}_{A_{(ij)}} = \mathbf{C}_{(ij)} \mathbf{\Sigma}_B \mathbf{C}_{(ij)}{}^{\mathrm{T}} = \begin{pmatrix} \left(a_{ij}^2 + b_{ij}^2\right)\sigma_B^2 & b_{ij} c_{ij} \sigma_B^2 \\ b_{ij} c_{ij} \sigma_B^2 & c_{ij}^2 \sigma_B^2 \end{pmatrix}$$

Therefore $\mathrm{Var}(A_i)/\mathrm{Var}(A_j) = a_{ij}^2 + b_{ij}^2/c_{ij}^2 = \gamma_{ij}^2 + \kappa_{ij}^2$.

The above analysis is used in the proof of Theorem 6.

Theorem 6 *Suppose:*

1. $Y \perp\!\!\!\perp X \mid (\langle g, X\rangle_{\mathcal{H}}, g, \mathbf{\Gamma})$
2. *Almost surely,* $\mathrm{Var}(Y \mid g, \mathbf{\Gamma})$ *is finite and* $\mathrm{Cov}(Y, \langle g, X\rangle_{\mathcal{H}} \mid g, \mathbf{\Gamma})$ *is nonzero*
3. *Almost surely,* $\mathbb{E}(X \mid \mathbf{\Gamma}) = 0$ *and* $\mathrm{Var}(X \mid \mathbf{\Gamma}) = \mathbf{\Gamma}$
4. $g \perp\!\!\!\perp (X, \mathbf{\Gamma})$
5. *Assumption 4 holds*
6. *g has an elliptical distribution*

Let ϕ_k and λ_k be the k^{th} most dominant eigenvector and eigenvalue of $\mathbf{\Gamma}$. Let $\rho_k(g,\mathbf{\Gamma}) = \text{Corr}^2(Y, \langle \phi_k, X\rangle_{\mathcal{H}} \mid g, \mathbf{\Gamma})$. Then, for $i<j$ with $\lambda_j > 0$:

$$\mathbb{P}(\rho_i(g,\mathbf{\Gamma}) \geq \rho_j(g,\mathbf{\Gamma})) = \frac{2}{\pi}\arctan\left(\frac{d_{ij,1}}{d_{ij,2}+\sqrt{d_{ij,1}^2+d_{ij,2}^2}}\right)$$

where $d_{ij,1} = 2\gamma_{ij}\sqrt{\lambda_i/\lambda_j}$ and $d_{ij,2} = \kappa_{ij}^2+\gamma_{ij}^2-(\lambda_i/\lambda_j)$. κ_{ij} and γ_{ij} are the results of applying Theorem 2 of Arnold and Brockett (1992) to the ratio $\langle \phi_i, g\rangle_{\mathcal{H}}/\langle \phi_j, g\rangle_{\mathcal{H}}$ (see Equation 3).

We note that when $\mathcal{H}$ has finite-dimension, we can have spherical distributions. In this case, Theorem 6 reduces to Theorem 5 as a result of $\kappa_{ij}=0$ and $\gamma_{ij}=1$.

Theorem 6 is not as strong as the result in Theorem 5 as we cannot ensure the lower bound that the probability is greater than 1/2. To achieve this, we need to add the extra assumption that $d_{ij,2}$ is negative.

This is equivalent to:

$$\frac{\text{Var}(\langle \phi_i, X\rangle_{\mathcal{H}} \mid \mathbf{\Gamma})}{\text{Var}\left(\langle \phi_j, X\rangle_{\mathcal{H}} \mid \mathbf{\Gamma}\right)} = \frac{\lambda_i}{\lambda_j} > \gamma_{ij}^2+\kappa_{ij}^2 = \frac{a_{ij}^2+b_{ij}^2}{c_{ij}^2} = \frac{\text{Var}(\langle \phi_i, g\rangle_{\mathcal{H}})}{\text{Var}\left(\langle \phi_j, g\rangle_{\mathcal{H}}\right)} \tag{4}$$

One might ask about the physical meaning of this assumption. Equation 4 shows that it is restricting the ratio of the axes lengths of the ellipsoid of $\langle \phi_i, g\rangle_{\mathcal{H}}$ against $\langle \phi_j, g\rangle_{\mathcal{H}}$ to be less than the ratio of the axes lengths of the ellipsoid between $\langle \phi_i, X\rangle_{\mathcal{H}}$ and $\langle \phi_j, X\rangle_{\mathcal{H}}$ after conditioning on $\mathbf{\Gamma}$. Comparing this against the finite-dimensional results, we are essentially saying that the distribution of g is not too far away from being what one intuitively understands as spherical. When $\mathcal{H}$ is finite-dimensional and g has a spherical distribution, the location and scale parameters are 0 and 1 respectively — Equation 4 is then just the familiar assumption $\lambda_i/\lambda_j > 1$.

Under this assumption, note that $-2d_{ij,1}d_{ij,2} > 0$ as $d_{ij,1}$ is positive because the eigenvalues and the scale parameter are both positive. This implies that $(d_{ij,1}-d_{ij,2})^2 > d_{ij,1}^2+d_{ij,2}^2$ and $d_{ij,1}-d_{ij,2}>0$ which means that:

$$d_{ij,1} > d_{ij,2}+\sqrt{d_{ij,1}^2+d_{ij,2}^2}.$$

Note also, that $d_{ij,1}^2+d_{ij,2}^2 > (-d_{ij,2})^2$ which implies that $d_{ij,2}+\sqrt{d_{ij,1}^2+d_{ij,2}^2} > 0$. Combining this with the above inequality we have that

$$\frac{d_{ij,1}}{d_{ij,2}+\sqrt{d_{ij,1}^2+d_{ij,2}^2}} > 1$$

which means that

$$\arctan\left(\frac{d_{ij,1}}{d_{ij,2}+\sqrt{d_{ij,1}^2+d_{ij,2}^2}}\right) > \frac{\pi}{4}$$

We conclude that, with the extra assumption that $d_{ij,2}$ is negative, we get the desired result that $\mathbb{P}(\rho_1(g,\mathbf{\Gamma}) \geq \rho_2(g,\mathbf{\Gamma})) > 1/2$.

To check how often this assumption holds, we ran a simulation study where we generated both X and g as standard Brownian motions. We simulated 500 observations of X and g which were assumed to have been observed at 100 equispaced points in the interval $[0,1]$. We calculated all eigevectors ϕ_i, $i=1,\ldots,100$. Then using all possible (i,j) pairs, we checked whether the assumption holds. We repeated the experiment 1000 times and we found the percentage of times this holds for each pair. We see that for about 97% of the pairs, where $i<j$, the assumption holds more in than half of the simulations. In Figure 1, we created a matrix of size 100×100 and — in the upper triangular region — we use a darker colour to indicate the pairs where the assumption was satisfied in more than half the simulations. The spaces on the upper triangular matrix not coloured indicate pairs where the assumption was satisfied less than 50% of the time. It is interesting to note here that, even in those occasions, the percentage

of times the assumption was met never fell below 45%. Another interesting feature of Figure 1, is that most of the pairs where the assumption does not hold are those where j is close to i and i is relatively small. It is reasonable to expect that there will be a smaller proportion of simulations where the ratio is satisfied if j is closer to i rather than further away as the ratio between the eigenvalues fluctuates more around 1 (for more evidence see Figure 2). Interestingly though, this result seems to be minimised as i and j are increased. Moreover, as is evident in Figure 2, there is an interesting behaviour as i increases when $j = 100$. It is not immediately clear to us why this happens.

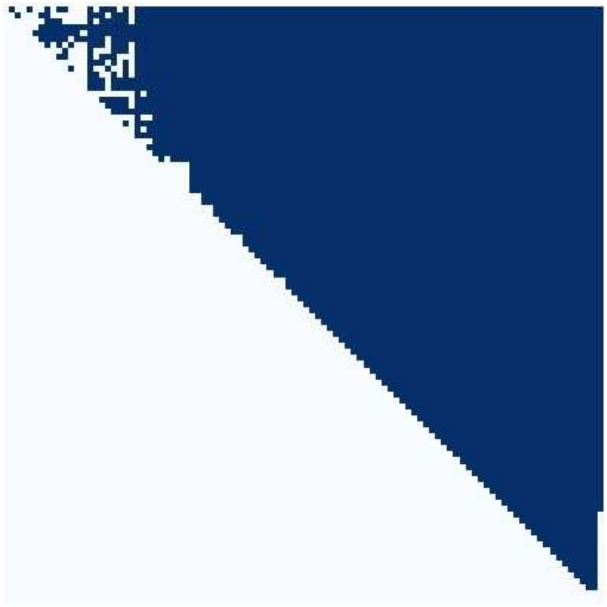

Fig. 1 Dark blue indicates the pairs (i, j) where more than half of the simulations satisfied equation 4

4 Discussion

In this paper, we extend the results from Artemiou and Li (2009), Artemiou and Li (2013), and Ni (2011) to a regression setting with Hilbertian predictors. We demonstrate that the predictive power of principal components is still valid in this setting — that is, the probability that a higher ranked principal component will have larger correlation with the response than a lower ranked one is greater than 1/2 under some assumptions. The work presented some challenges due to the infinite-dimensional setting used and the non-existence of a spherical distribution in this setting. We demonstrate that the result is valid under two conditions: first when Assumption 5 is used and second when we assume the weaker condition of ellipticity on g alongside a specific relationship between the ratio of the eigenvalues and the ratio of the variances of the inner products of the eigenvectors with g.

The question of the predictive potential of principal component regression was always part of the discussion among researchers. In this paper, we discuss this potential when the predictor is a Hilbertian random variable and the response is a scalar. It will be interesting to see if this relationship holds when the response Y is also a Hilbertian random variable. It would also be interesting to explore whether similar results hold in nonlinear principal component algorithms or other infinite-dimensional settings like kernel principal components regression (should we remove this sentence now that the kpca paper is out?).

A Essential Definitions

For the benefit of the reader, we present here some fundamental definitions in functional data analysis. These definitions can be found in Hsing and Eubank (2015) along with a deeper exposition of the field. We first define random variables and

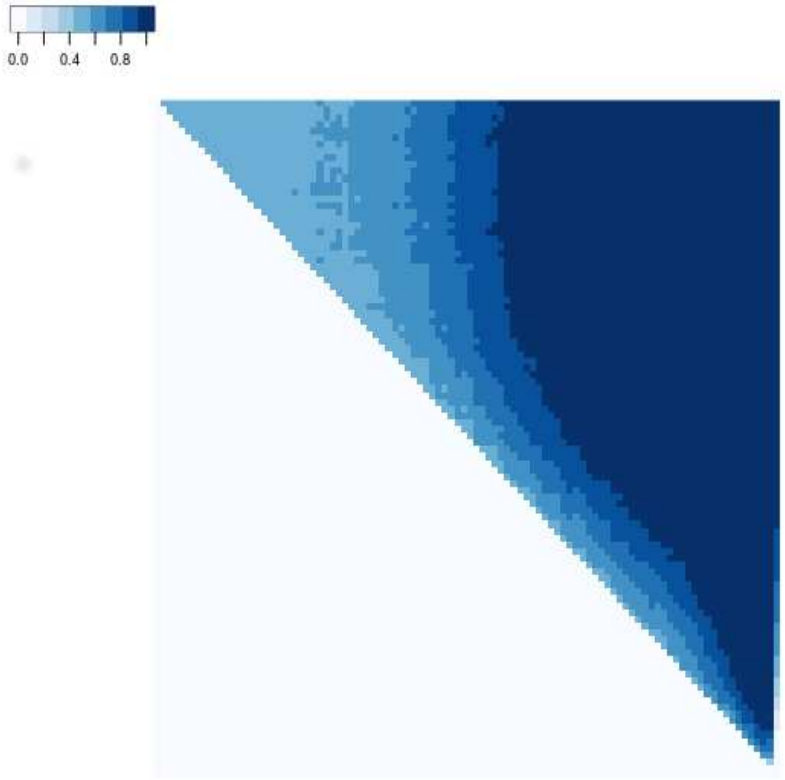

Fig. 2 This is a heatmap of the probabilities that each pair (i, j) satisfies equation (4). As expected the further away i is from j the bigger the probability.

nuclear operators in and on a Hilbert space respectively. Although our interest lies in the case where the random variables are random functions, the definitions are given for the more general setting of Hilbertian random variables. We note that this more abstract framework includes function spaces where the functions need not be univariate so this paper applies to, say, predictors which are random fields. This work therefore is relevant to a number of fields including: FMRI data analysis, spatial statistics, image processing, and speech recognition.

Definition 2 Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space and $(\mathcal{H},\ \mathcal{B}(\mathcal{H}))$ be a measureable space where $\mathcal{H}$ is a Hilbert space and $\mathcal{B}(\mathcal{H})$ is its associated Borel σ-field. A measureable function $X : (\Omega, \mathfrak{F}, \mathbb{P}) \to (\mathcal{H},\ \mathcal{B}(\mathcal{H}))$ is called an *H-valued random variable*. We also say that X is a *Hilbertian random variable*.

Definition 3 Let $\mathcal{H}$ be a Hilbert space. A compact operator, that is one which is the operator norm limit of a sequence of finite rank operators, $L : \mathcal{H} \to \mathcal{H}$ is said to be a *nuclear* operator if the sum of its eigenvalues is finite.

Remark 3 The class of nuclear operators on a Hilbert space contains the class of all operators which have finitely many nonzero eigenvalues.

The expectation of a Hilbertian random variable is defined in terms of the Bochner integral — the construction is given in Hsing and Eubank (2015) and is similar to that for the Lebesgue integral so we will not present it here. For our purposes, it is enough to note that for a Hilbertian random variable a, the expectation $\mathbb{E}(a)$ is unique, an element of the space $\mathcal{H}$, and satisfies

$$\forall b \in \mathcal{H},\ \mathbb{E}\left(\langle b, a\rangle_{\mathcal{H}}\right) = \langle b, \mathbb{E}(a)\rangle_{\mathcal{H}} \tag{5}$$

Remark 4 Observe that the expectation on the left hand side is the expectation of a real random variable, whereas the expectation on the right side is the expectation of an $\mathcal{H}$-valued random variable.

We will also require a generalisation of the notion of variance for a Hilbertian random variable, but first we define a tensor product operation.

Definition 4 Let x_1, x_2 be elements of Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively. The *tensor product operator* $(x_1 \otimes_1 x_2) : \mathcal{H}_1 \to \mathcal{H}_2$ is defined by

$$(x_1 \otimes_1 x_2)\, y = \langle x_1, y\rangle_{\mathcal{H}_1} x_2$$

for $y \in \mathcal{H}_1$. If $\mathcal{H}_1 = \mathcal{H}_2$, we use $\otimes$ instead of $\otimes_1$.

In the case where $\mathcal{H}_1 = \mathcal{H}_2 = \mathbb{R}^p$, we have that $x_1 \otimes x_2 = x_2 x_1{}^{\mathrm{T}}$ so the usual covariance matrix can be written as

$$\mathbb{E}\left((\mathbf{X} - \mathbb{E}(\mathbf{X})) \otimes (\mathbf{X} - \mathbb{E}(\mathbf{X}))\right)$$

This notation will also be used for a covariance operator, but with X being an $\mathcal{H}$-valued random variable. We note that all covariance operators on a Hilbert space $\mathcal{H}$ are compact, non-negative definite, and self-adjoint. Proofs of these facts can be found in Pinelis and Molzon (2016).

Assuming that the covariance operator of some predictor X is nuclear gives meaning to the phrase "PCA captures most of the variability in the data" for the infinite-dimensional setting. This is because it supplies a notion of how much variance there is in total.

The notion of a spherical distribution was central to the work of Artemiou and Li (2013). In the case of data in an infinite-dimensional space, this notion cannot be generalised as is explained below but the idea of an elliptical distribution can. We will thus make use of this concept instead. The following definition is given by Y.Li (2007).

Definition 5 A Hilbertian random variable A, in a Hilbert space $\mathcal{H}$, has an *elliptically symmetric distribution* if the characteristic function of $A - \mathbb{E}(A)$ has the following form:

$$\psi(f) = \varphi\left(\langle f, \Psi f\rangle_{\mathcal{H}}\right)$$

for all $f \in \mathcal{H}$, where Ψ is a self-adjoint, non-negative definite, *nuclear* operator on $\mathcal{H}$, and φ is a univariate function.

We note that — in the infinite-dimensional Hilbert space setting — Ψ in Definition 5 cannot be the identity operator as it is noncompact and thus not nuclear. It can be shown that Ψ is, up to multiplication by a constant, the covariance operator (when it exists) of the Hilbertian random variable — the requirement then that the sum of the eigenvalues of A is finite is equivalent to the sum of the variances of the principal components of A being finite. We conclude that we cannot extend the notion of a spherically symmetric distribution to the entirety of an infinite-dimensional space. Note that we can have sphericity in a finite-dimensional subspace.

B Proofs

Proof (Lemma 1) Define Φ as an operator on $\mathcal{H}$ by $\Phi(x) = \langle g, x\rangle_{\mathcal{H}}$. This operator takes a fixed $x \in \mathcal{H}$ and returns a real random variable so it is a random operator. By the Riesz Representation Theorem, this random operator can be identified with a random element of the dual space $\mathcal{H}^*$ (that is the space of all continuous linear functions from the space $\mathcal{H}$ into the base field) so there is a unique random adjoint operator Φ^* such that for all fixed $x \in \mathcal{H}$ and $y \in \mathbb{R}$, $\Phi(x)\, y = \langle x, \Phi^*(y)\rangle_{\mathcal{H}}$. It is easy to see that for any fixed $y \in \mathbb{R}$, $\Phi^*(y) = yg$. We show that, almost surely, $\mathbb{E}(X \mid \Phi(X), g, \mathbf{\Gamma})$ is orthogonal to $\mathrm{Span}(\mathbf{\Gamma} g)^{\perp}$. For convenience, let T be the tuple $(\Phi(X), g, \mathbf{\Gamma})$. Let $x \in \mathrm{Span}(\mathbf{\Gamma} g)^{\perp}$, which is a random variable in $\mathcal{H}$, then we have the following:

$$\forall z \in \mathrm{Span}(\mathbf{\Gamma} g), \langle x, z\rangle_{\mathcal{H}} = 0 \implies \forall y \in \mathbb{R}, \langle x, y\mathbf{\Gamma} g\rangle_{\mathcal{H}} = 0$$

which implies that for any fixed $y \in \mathbb{R}$

$$\langle x, y\mathbf{\Gamma} g\rangle_{\mathcal{H}} = \langle x, \mathbf{\Gamma}(yg)\rangle_{\mathcal{H}} = \langle \mathbf{\Gamma} x, yg\rangle_{\mathcal{H}} = \langle \mathbf{\Gamma} x, \Phi^* y\rangle_{\mathcal{H}} = \Phi(\mathbf{\Gamma} x)\, y = 0$$

where the first and second equalities follow from the linearity and self-adjointedness of $\mathbf{\Gamma}$. The above now implies that $\Phi(\mathbf{\Gamma} x) = 0$ and therefore $\mathbf{\Gamma} x \in \mathrm{Ker}(\Phi)$.

Consider now $\mathbb{E}\left(\langle x, \mathbb{E}(X \mid T)\rangle_{\mathcal{H}}^2\right)$. Showing this to be 0 gives the result, as it is the expectation of a squared random variable (I am not certain if there is an issue here or not - this is what I sent an image about).

$$\begin{aligned}
\mathbb{E}\left(\langle x, \mathbb{E}(X \mid T)\rangle_{\mathcal{H}}^2\right) &= \mathbb{E}\left(\langle x, \mathbb{E}(X \mid T)\rangle_{\mathcal{H}} \langle x, \mathbb{E}(X \mid T)\rangle_{\mathcal{H}}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\langle x, X\rangle_{\mathcal{H}} \mid T\right) \langle x, \mathbb{E}(X \mid T)\rangle_{\mathcal{H}}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\langle x, \mathbb{E}(X \mid T)\rangle_{\mathcal{H}} \langle x, X\rangle_{\mathcal{H}} \mid T\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\left\langle x, \langle x, \mathbb{E}(X \mid T)\rangle_{\mathcal{H}} X\right\rangle_{\mathcal{H}} \mid T\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\left\langle x, \mathbb{E}\left(\langle x, X\rangle_{\mathcal{H}} \mid T\right) X\right\rangle_{\mathcal{H}} \mid T\right)\right)
\end{aligned}$$

where the second equality follows from Equation 5 ; the third and fourth equalities follow by moving the second inner product into the expectation; the fifth equality uses Equation 5 again. Now by Assumption 4, there is a real constant A such that $\mathbb{E}\left(\langle x, X\rangle_{\mathcal{H}} \mid T\right) = A\Phi(X)$ (again, I'm not sure if x being random is an issue).

Therefore (Are we agreed that the proposed resolutions (in blue) are correct?)

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}\left(\left\langle x, \mathbb{E}\left(\langle x, X\rangle_{\mathcal{H}} \mid T\right) X\right\rangle_{\mathcal{H}} \mid T\right)\right) &= \mathbb{E}\left(\mathbb{E}\left(\langle x, A\Phi(X)\, X\rangle_{\mathcal{H}} \mid T\right)\right) \\
&= A\mathbb{E}\left(\mathbb{E}\left(\langle x, \Phi(X)\, X\rangle_{\mathcal{H}} \mid T\right)\right) \\
&= A\mathbb{E}\left(\mathbb{E}\left(\left\langle x, \langle g, X\rangle_{\mathcal{H}} X\right\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right)\right) \\
&= A\mathbb{E}\left(\mathbb{E}\left(\left\langle x, \langle g, X\rangle_{\mathcal{H}} X\right\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, \boldsymbol{\Gamma}\right)\right) \text{ as } g \text{ is independent of } X \\
&= A\mathbb{E}\left(\left\langle x, \langle g, X\rangle_{\mathcal{H}} X\right\rangle_{\mathcal{H}} \mid \boldsymbol{\Gamma}\right) \text{ by the law of total expectation} \\
&= A\left\langle x, \mathbb{E}\left(\langle g, X\rangle_{\mathcal{H}} X \mid \boldsymbol{\Gamma}\right)\right\rangle_{\mathcal{H}} \\
&= A\langle x, \boldsymbol{\Gamma} g\rangle_{\mathcal{H}} = A\langle \boldsymbol{\Gamma} x, g\rangle_{\mathcal{H}} = 0
\end{aligned}$$

Proof (Lemma 2) S is an element of l^2 because $\mathcal{H}$ and l^2 are isomorphic, up to isometry, and by the same reasoning S is elliptically distributed. Now let $P : l^2 \to \mathbb{R}^n$ be the operator which truncates a sequence at the n^{th} term. This operator is compact, and therefore bounded, so by Theorem 4 of Y.Li (2007), the vector T_n is elliptically distributed.

Proof (Theorem 5) From the definition of correlation:

$$\operatorname{Corr}^2\left(Y, \langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right) = \frac{\operatorname{Cov}^2\left(Y, \langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)}{\operatorname{Var}\left(Y \mid g, \boldsymbol{\Gamma}\right) \operatorname{Var}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)} \tag{6}$$

Now, recall that conditional expectation is a self-adjoint operator in the covariance inner product. That is for any random variables U_1, U_2, U_3, we have

$$\operatorname{Cov}\left(\mathbb{E}\left(U_1 \mid U_2\right), U_3\right) = \operatorname{Cov}\left(U_1, \mathbb{E}\left(U_2 \mid U_3\right)\right)$$

Consider:

$$\begin{aligned}
\operatorname{Cov}\left(Y, \langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right) &= \operatorname{Cov}\left(Y, \mathbb{E}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid g, X, \boldsymbol{\Gamma}\right) \mid g, \boldsymbol{\Gamma}\right) = \operatorname{Cov}\left(\mathbb{E}\left(Y \mid g, X, \boldsymbol{\Gamma}\right), \langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right) \\
&= \operatorname{Cov}\left(\mathbb{E}\left(Y \mid \langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right), \langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right) \\
&= \operatorname{Cov}\left(Y, \mathbb{E}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right) \mid g, \boldsymbol{\Gamma}\right)
\end{aligned} \tag{7}$$

where the third equality follows as $Y \perp\!\!\!\perp X \mid \left(\langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right)$. As Assumption 4 holds, there is a real constant α_i such that $\mathbb{E}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right) = \alpha_i \langle g, X\rangle_{\mathcal{H}}$, and similarly for j. Thus Equation 7 becomes:

$$\operatorname{Cov}\left(Y, \alpha_i \langle g, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right) = \alpha_i \operatorname{Cov}\left(Y, \langle g, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)$$

Substituting this into Equation 6, we find that

$$\operatorname{Corr}^2\left(Y, \langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right) = \frac{\alpha_i^2 \operatorname{Cov}^2\left(Y, \langle g, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)}{\operatorname{Var}\left(Y \mid g, \boldsymbol{\Gamma}\right) \operatorname{Var}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)}$$

Thus

$$\frac{\operatorname{Corr}^2\left(Y, \langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)}{\operatorname{Corr}^2\left(Y, \langle \phi_j, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)} = \frac{\alpha_i^2 \operatorname{Var}\left(\langle \phi_j, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)}{\alpha_j^2 \operatorname{Var}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)}$$

As $g \perp\!\!\!\perp (X, \boldsymbol{\Gamma})$, $\operatorname{Var}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right) = \operatorname{Var}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid \boldsymbol{\Gamma}\right) = \lambda_i$ and similarly for j. Thus

$$\frac{\operatorname{Corr}^2\left(Y, \langle \phi_i, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)}{\operatorname{Corr}^2\left(Y, \langle \phi_j, X\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)} = \frac{\alpha_i^2 \lambda_j}{\alpha_j^2 \lambda_i} \tag{8}$$

Now look back at Equation 7. By Equation 5, we see that

$$\operatorname{Cov}\left(Y, \mathbb{E}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right) \mid g, \boldsymbol{\Gamma}\right) = \operatorname{Cov}\left(Y, \left\langle \phi_i, \mathbb{E}\left(X \mid \langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right)\right\rangle_{\mathcal{H}} \mid g, \boldsymbol{\Gamma}\right)$$

By Lemma 1, $\mathbb{E}\left(X \mid \langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right) = c\boldsymbol{\Gamma} g$ for some constant c. Hence

$$\mathbb{E}\left(\langle \phi_i, X\rangle_{\mathcal{H}} \mid \langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right) = \alpha_i \langle g, X\rangle_{\mathcal{H}} = \left\langle \phi_i, \mathbb{E}\left(X \mid \langle g, X\rangle_{\mathcal{H}}, g, \boldsymbol{\Gamma}\right)\right\rangle_{\mathcal{H}} = c\langle \phi_i, \boldsymbol{\Gamma} g\rangle_{\mathcal{H}}$$

Now we have

$$c\langle \phi_i, \boldsymbol{\Gamma} g\rangle_{\mathcal{H}} = c\langle \boldsymbol{\Gamma}\phi_i, g\rangle_{\mathcal{H}} = c\lambda_i \langle \phi_i, g\rangle_{\mathcal{H}}$$

Consequently

$$\alpha_i = \frac{c\lambda_i \langle \phi_i, g \rangle_{\mathcal{H}}}{\langle g, X \rangle_{\mathcal{H}}}$$

and similarly for α_j. So Equation 8 can be rewritten as

$$\frac{\mathrm{Corr}^2\left(Y, \langle \phi_i, X \rangle_{\mathcal{H}} \mid g, \mathbf{\Gamma}\right)}{\mathrm{Corr}^2\left(Y, \langle \phi_j, X \rangle_{\mathcal{H}} \mid g, \mathbf{\Gamma}\right)} = \frac{\lambda_i \langle \phi_i, g \rangle_{\mathcal{H}}^2}{\lambda_j \langle \phi_j, g \rangle_{\mathcal{H}}^2}$$

Now by Assumption 5, $\left\{\langle \phi_k, g \rangle_{\mathcal{H}}\right\}_{k \in \mathbb{N} \cap [1,n]}$ is spherically symmetric for any n. Therefore, by Theorem 1 of Arnold and Brockett (1992), $\frac{\langle \phi_i, g \rangle_{\mathcal{H}}}{\langle \phi_j, g \rangle_{\mathcal{H}}}$ has a standard Cauchy distribution. Thus

$$\mathbb{P}\left(\rho_i\left(g, \mathbf{\Gamma}\right) > \rho_j\left(g, \mathbf{\Gamma}\right)\right) = \mathbb{P}\left(-\sqrt{\frac{\lambda_i}{\lambda_j}} < \frac{\langle \phi_i, g \rangle_{\mathcal{H}}}{\langle \phi_j, g \rangle_{\mathcal{H}}} < \sqrt{\frac{\lambda_i}{\lambda_j}}\right) = \frac{2}{\pi} \arctan\left(\sqrt{\frac{\lambda_i}{\lambda_j}}\right)$$

□

Proof (Theorem 6) The proof is similar to that of Theorem 5 up to the point where we have shown that:

$$\mathbb{P}\left(\rho_i\left(g, \mathbf{\Gamma}\right) > \rho_j\left(g, \mathbf{\Gamma}\right)\right) = \mathbb{P}\left(-\sqrt{\frac{\lambda_i}{\lambda_j}} < \frac{\langle \phi_i, g \rangle_{\mathcal{H}}}{\langle \phi_j, g \rangle_{\mathcal{H}}} < \sqrt{\frac{\lambda_i}{\lambda_j}}\right)$$

Now as g has an elliptical distibution, we apply Lemma 2 and Theorem 2 of Arnold and Brockett (1992) to find that $\langle \phi_i, g \rangle_{\mathcal{H}} / \langle \phi_j, g \rangle_{\mathcal{H}}$ has a general Cauchy distribution with scale parameter γ_{ij} and location κ_{ij}. Thus:

$$\begin{aligned}
\mathbb{P}\left(\rho_i\left(g, \mathbf{\Gamma}\right) > \rho_j\left(g, \mathbf{\Gamma}\right)\right) =& \frac{1}{\pi} \arctan\left(\frac{\sqrt{\frac{\lambda_i}{\lambda_j}} - \kappa_{ij}}{\gamma_{ij}}\right) + \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{-\sqrt{\frac{\lambda_i}{\lambda_j}} - \kappa_{ij}}{\gamma_{ij}}\right) - \frac{1}{2} \\
=& \frac{1}{\pi}\left(\arctan\left(\frac{\sqrt{\frac{\lambda_i}{\lambda_j}} - \kappa_{ij}}{\gamma_{ij}}\right) - \arctan\left(\frac{-\sqrt{\frac{\lambda_i}{\lambda_j}} - \kappa_{ij}}{\gamma_{ij}}\right)\right)
\end{aligned}$$

Using $\arctan\left(-x\right) = -\arctan\left(x\right)$, we have that the above is equal to:

$$\frac{1}{\pi}\left(\arctan\left(\frac{\sqrt{\frac{\lambda_i}{\lambda_j}} - \kappa_{ij}}{\gamma_{ij}}\right) + \arctan\left(\frac{\sqrt{\frac{\lambda_i}{\lambda_j}} + \kappa_{ij}}{\gamma_{ij}}\right)\right)$$

Using $\arctan\left(u\right) + \arctan\left(v\right) = \arctan\left(\frac{u+v}{1-uv}\right)$ provided $uv \neq 1$ and the result is taken modulo π we have that the above probability is now:

$$\begin{aligned}
\frac{1}{\pi} \arctan\left(\frac{\left(\frac{\sqrt{\frac{\lambda_i}{\lambda_j}} - \kappa_{ij}}{\gamma_{ij}} + \frac{\sqrt{\frac{\lambda_i}{\lambda_j}} + \kappa_{ij}}{\gamma_{ij}}\right)}{1 - \frac{\sqrt{\frac{\lambda_i}{\lambda_j}} - \kappa_{ij}}{\gamma_{ij}} \frac{\sqrt{\frac{\lambda_i}{\lambda_j}} + \kappa_{ij}}{\gamma_{ij}}}\right) =& \frac{1}{\pi} \arctan\left(\frac{\frac{2\sqrt{\frac{\lambda_i}{\lambda_j}}}{\gamma_{ij}}}{1 - \left(\frac{\frac{\lambda_i}{\lambda_j} - \kappa_{ij}^2}{\gamma_{ij}^2}\right)}\right) \\
=& \frac{1}{\pi} \arctan\left(\frac{2\gamma_{ij}\sqrt{\frac{\lambda_i}{\lambda_j}}}{\gamma_{ij}^2 - \frac{\lambda_i}{\lambda_j} + \kappa_{ij}^2}\right)
\end{aligned}$$

We see that the numerator is equal to $d_{ij,1}$ and the denominator equal to $d_{ij,2}$. Then, using $\arctan\left(x\right) = 2\arctan\left(\frac{x}{1+\sqrt{1+x^2}}\right)$, we can rewrite the above and simplify to obtain:

$$\frac{2}{\pi} \arctan\left(\frac{d_{ij,1}}{d_{ij,2} + \sqrt{d_{ij,1}^2 + d_{ij,2}^2}}\right)$$

□

References

Arnold, B. C. and Brockett, P. L. (1992). On distributions whose component ratios are cauchy. *American Statistician*, 46(1):25–26.

Artemiou, A. and Li, B. (2009). On principal components regression: a statistical explanation of a natural phenomenon. *Statistica Sinica*, 19:1557–1565.

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.

Cook, R. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, 22(1):1–26.

Cox, D. R. (1968). Notes on Some Aspects of Regression Analysis. *Journal of the Royal Statistical Society Series A (General)*, 131(3):265–279.

Dauxois, J., Ferré, L., and Yao, A.-F. (2001). Un modèle semi-paramétrique pour variables aléatoires hilbertiennes. *CR Acad Sci Paris*, 333(1):947–952.

Ferré, L. and Yao, A. F. (2003). Functional sliced inverse regression analysis. *Statistics*, 37(6):475–488.

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: Statistical Methodology*, 72(1):93–110.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. West Sussex: Wiley, 1st edition.

Kingman, J. F. C. (1972). On Random Sequences with Spherical Symmetry. *Biometrika*, 59(2):492.

Li, B. (2007). Comment: Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, 22(1):32–35.

Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R*. Boca Raton: CRC Press, 1st edition.

Li, B. and Song, J. (2017). Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, 45(3):1059–1095.

Li, Y. (2007). A Note on Hilbertian Elliptically Contoured Distributions. Technical report.

Ni, L. (2011). Principal Component Regression Revisited. *Statistica Sinica*, 21:741–747.

Pinelis, I. and Molzon, R. (2016). Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, 10(1):1001–1063.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, 2nd edition.