Sankhyā : The Indian Journal of Statistics
2019, Volume 81-A, Part 1, pp. 214-243

Inference on Covariance Operators via Concentration Inequalities: k-sample Tests, Classification, and Clustering via Rademacher Complexities

Adam B. Kashlak, John A. D. Aston and Richard Nickl
University of Cambridge, Cambridge, UK
Adam B. Kashlak
University of Alberta, Edmonton, Canada

Abstract

We propose a novel approach to the analysis of covariance operators making use of concentration inequalities. First, non-asymptotic confidence sets are constructed for such operators. Then, subsequent applications including a k sample test for equality of covariance, a functional data classifier, and an expectation-maximization style clustering algorithm are derived and tested on both simulated and phoneme data.

AMS (2000) *subject classification.* Primary 62G05; Secondary 62G15.
Keywords and phrases. Functional data analysis, Manifold data, Non-asymptotic confidence sets, Concentration of measure.

1 Introduction

Functional data spans many realms of application from medical imaging, Jiang et al. (2016), to speech and linguistics, Pigoli et al. (2014), to the movement of DNA molecules, Panaretos et al. (2010). General inference techniques for functional data have received much attention in recent years from the construction of confidence sets, to other topics such as k-sample tests, classification, and clustering of functional data. Most testing methodology treats the data as continuous L^2 valued functions and subsequently reduces the problem to a finite dimensional one through expansion in some orthogonal basis such as the often utilized Karhunen-Loève expansion (Horváth and Kokoszka, 2012). However, inference making use of non-Hilbert norms has received much less attention. We propose a novel methodology for performing fully functional inference through the application of concentration

inequalities, which is furthermore a single methodology applicable to a wide variety of inference problems; for general concentration of measure results, see Ledoux (2001) and Boucheron et al. (2013). Special emphasis is given to inference on covariance operators, which offers a fruitful way to analyze functional data.

As an example, imagine multiple samples of speech data collected from multiple speakers. Each speaker will have his or her own sample covariance operator taking into account the unique variations of his or her speech and language. An exploratory researcher may want to find natural clusters amidst the speakers perhaps corresponding to gender, language, or regional dialect. Meanwhile, a linguist studying the similarities between languages may want to test for the equality of such covariances. A computer scientist may need to implement an algorithm that when given speech data quickly identifies what language is being spoken and furthermore parses the sound clip and identifies each individual phoneme in order to process the speech into text. Our proposed method has the versatility to yield statistical tests that address all of these questions as well as others.

Past methods for analyzing covariance operators (Panaretos et al., 2010; Fremdt et al., 2013) rely on the Hilbert-Schmidt setting for their inference. However, the recent work of Pigoli et al. (2014) argues that the use of the Hilbert-Schmidt metric ignores the geometry of the covariance operators which lie on a manifold and that more statistical power can be gained by using alternative metrics. The main drawback of their research is their reliance on permutation based tests, which are computationally intensive and, in some instances, incapable of achieving an acceptable level of accuracy in a reasonable amount of time. In the age of *big data*, if p-values less than 1/1000 are desired, this can become computationally intractable with permutation methods; see Fig. 1. Hence, we approach such inference for covariance operators by using a non-asymptotic concentration of measure approach, which can incorporate arbitrary norms. This has previously been used in nonparametric statistics and machine learning, sometimes under the name of 'Rademacher complexities' (Koltchinskii, 2001, 2006; Bartlett et al., 2002; Bartlett and Mendelson, 2003; Giné and Nickl, 2010; Arlot et al., 2010; Lounici and Nickl, 2011; Kerkyacharian et al., 2012; Fan, 2011). These concentration inequalities provide a natural way to construct non-asymptotic confidence regions and, subsequently, statistical tests. Our approach can classify as well as k-nearest neighbours, cluster as well as k-means, and can test for equality of covariance as well as a permutation test. These methods are available in the R package `fdcov` (Cabassi and Kashlak, 2016).

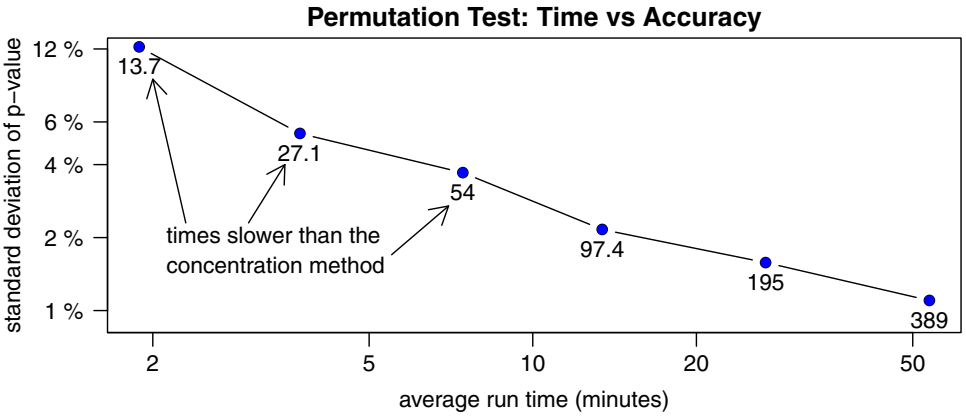

Figure 1: Plotted are the run times against the accuracy of the permutation test for testing for equality of covariance given five samples of 30 curves each. The procedure requires over 50 minutes of computation time to get a standard deviation of around 1% for the estimated p-value. Adjacent to each point is the number of times slower the permutation test is when compared to the concentration test. The average run times were clocked on an Intel(R) Core(TM) i3-3217U CPU @ 1.80GHz

2 Definitions and Notation

Generally, we will consider functional data to be in the Hilbert space $L^2(I)$ for $I \subset \mathbb{R}$. While the methods we outline could be used in a number of settings, we will concentrate on covariance operators which are operator valued random variables. Let

$$Op(L^2) = \big\{T : L^2(I) \to L^2(I) \mid \text{there exists } M \geq 0, \|T\phi\|_{L^2} \leq M\|\phi\|_{L^2} \text{ for all } \phi \in L^2(I)\big\}$$

denote the space of all bounded linear operators mapping L^2 into L^2, the space which contains the covariance operators of interest.

The metrics that will be investigated are those that correspond to the p-Schatten norms. When $p \neq 2$, these are not Hilbert norms.

Definition 1 (p-Schatten Norm)**.** *Given separable Hilbert spaces H_1 and H_2, a bounded linear operator $\Sigma : H_1 \to H_2$, and $p \in [1, \infty)$, then the p-Schatten norm is $\|\Sigma\|^p = \mathrm{tr}((\Sigma^*\Sigma)^{p/2})^{1/p}$. For $p = 1$, this is often referred to as the trace or nuclear norm. For $p = 2$, it is the Hilbert-Schmidt norm. For $p = \infty$, the Schatten norm is the operator norm: $\|\Sigma\|_\infty = \sup_{\|f\|_{H_1}=1} \|\Sigma f\|_{H_2}$.*

In the case that Σ is compact, self-adjoint, and trace-class, then given the associated eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$, the p-Schatten norm coincides with the ℓ^p norm of the eigenvalues:

$$\|\Sigma\|_p = \begin{cases} \|\lambda\|_{\ell^p} = (\sum_{i=1}^{\infty} |\lambda_i|^p)^{1/p}, & p \in [1, \infty) \\ \max_{i \in \mathbb{N}} |\lambda_i|, & p = \infty \end{cases}.$$

In order to construct a covariance operator from a sample of functional data, the notion of tensor product is required. Let $f, g \in L^2(I)$ and ϕ in the dual space $L^2(I)^*$ with inner product $\langle f, \phi \rangle = \phi(f)$. The tensor product, $f \otimes g$, is the rank one operator defined by $(f \otimes g)\phi = \langle g, \phi \rangle f = \phi(g) f$.

Secondly, we will implement a Rademacher symmetrization technique in the concentration inequalities. This requires the use of the namesake Rademacher random variables.

Definition 2 (Rademacher Distribution)**.** *A random variable $\varepsilon \in \mathbb{R}$ has a Rademacher distribution if* $\mathrm{P}(\varepsilon = 1) = \mathrm{P}(\varepsilon = -1) = 1/2$.

One particularly fruitful avenue of functional data analysis is the analysis of covariance operators. Such an approach to functional data has been discussed by Panaretos et al. (2010) for DNA microcircles, by Fremdt et al. (2013) for the egg laying psractices of fruit flies, and by Pigoli et al. (2014) with application to differentiating spoken languages.

Definition 3 (Covariance Operator)**.** *Let $I \subseteq \mathbb{R}$, and let f be a random function (variable) in $L^2(I)$ with $\mathrm{E}\|f\|_{L^2}^2 < \infty$ and mean zero. The associated covariance operator $\Sigma_f \in Op(L^2)$ is defined as $\Sigma_f = \mathrm{E} f^{\otimes 2} = \mathrm{E}(\langle f, \cdot \rangle f)$.*

As a particular special case, if $I = \{i_1, \ldots, i_m\}$ has finite cardinality, then $f = (f_1, \ldots, f_m)$ is a random vector in $\mathbb{R}^m$ and for some fixed vector $v \in \mathbb{R}^m$, $\mathrm{E}\langle f, v \rangle f = \mathrm{E}\left(f f^{\mathrm{T}}\right) v$ where $\Sigma_f = \mathrm{E}\left(f f^{\mathrm{T}}\right)$ is then the usual covariance matrix. More generally, covariance operators are integral operators with the kernel function $c_f(s,t) = \mathrm{cov}\{f(s), f(t)\} \in L^2(I \times I)$. Such operators are characterized by the result that for $f \in L^2(I)$, Σ_f is a covariance operator if and only if it is trace-class, self-adjoint, and compact on $L^2(I)$ where the symmetry follows immediately from the definition and the finite trace norm comes from Parseval's equality.

Furthermore, working under the assumption that $\mathrm{E}\|f\|_{L^2}^4 < \infty$, we will require tensor powers of covariance operators denoted as $\Sigma^{\otimes 2} : Op(L^2) \to Op(L^2)$. For a basis $\{e_i\}_{i=1}^{\infty} \in L^2(I)$ with corresponding basis $\{e_i \otimes e_j\}_{i,j=1}^{\infty}$ for $Op(L^2(I))$, the previous definition is extended to $\Sigma^{\otimes 2} = \langle \Sigma, \cdot \rangle \Sigma$ where for $\Sigma_1, \Sigma_2 \in Op(L^2)$ with $\Sigma_1 = \sum_{i,j=1}^{\infty} \lambda_{i,j} e_{i,j}$ and $\Sigma_2 = \sum_{i,j=1}^{\infty} \gamma_{i,j} e_{i,j}$, then

$\langle \Sigma_1, \Sigma_2 \rangle = \sum_{i,j}^{\infty} \lambda_{i,j}\gamma_{i,j}$. Specifically for covariance operators, the tensor power takes on a similar integral operator form with kernel $c_\Sigma(s,t,u,v) = \mathrm{cov}\left(f(s), f(t)\right)\mathrm{cov}\left(f(u), f(v)\right)$.

Given an Hilbert space H with inner product $\langle \cdot, \cdot \rangle$, the adjoint of a bounded linear operator $\Sigma : H \to H$, denoted as Σ^*, is the unique operator such that $\langle \Sigma f, g \rangle = \langle f, \Sigma^* g \rangle$ for $f, g \in H$, the existence of which is given by the Riesz representation theorem for self-adjoint operators, such as the covariance operators of interest, $\Sigma = \Sigma^*$.

We begin with a sample of functional data. Let $f_1, \ldots, f_n \in L^2(I)$ be independent and identically distributed observations with mean zero and covariance operator Σ. Let the sample mean be $\bar{f} = n^{-1}\sum_{i=1}^n f_i$ and the empirical estimate of Σ be $\hat{\Sigma} = n^{-1}\sum_{i=1}^n (f_i - \bar{f}) \otimes (f_i - \bar{f})$. The initial goal is to construct a confidence set for Σ with respect to some metric $\mathrm{d}(\cdot,\cdot)$ of the form $\{\Sigma : \mathrm{d}(\hat{\Sigma}, \Sigma) \leq r(n, \hat{\Sigma}, \alpha)\}$, which has coverage $1-\alpha$ for any desired $\alpha \in [0,1]$ and a radius r depending only on the data and α. Such a confidence set can be utilized for a wide variety of statistical analyses.

3 Confidence Sets for Covariance Operators

To construct a confidence set for covariance operators, let our functional data $f_i \in L^2(I)$ and $f_i^{\otimes 2} = f_i \otimes f_i \in Op(L^2)$, the Hilbert space of bounded linear operators mapping L^2 to L^2, such that $(f_i \otimes f_i)\phi = \langle f, \phi \rangle f$ for some $\phi \in L^2$. The construction of our confidence set is based on Talagrand's concentration inequality (Talagrand, 1996) with explicit constants, which can be thought of as a more general version of Bernstein's inequality (Boucheron et al., 2013, Chapter 2). This inequality is typically stated for empirical processes (Giné and Nickl, 2016, Theorem 3.3.9 and 3.3.10), but applies to random variables with values in a separable Banach space $(B, \|\cdot\|_B)$ as well by simple duality arguments (Giné and Nickl, 2016, Example 2.1.6). More details on this construction can be found in Appendix A. For some desired p-Schatten norm, $\|\cdot\|_p$, with $p \in [1,\infty)$ and with conjugate $q = p/(p-1)$, we require the following terms

$$Z = \Big\| \frac{1}{n}\sum_{i=1}^n f_i \otimes f_i - \mathrm{E} f_i \otimes f_i \Big\|_p = \|\hat{\Sigma} - \Sigma\|_p, \sigma^2 = \frac{1}{n}\sum_{i=1}^n \sup_{\|\Pi\|_q \leq 1} \mathrm{E}\left\{ \langle f_i^{\otimes 2} - \mathrm{E} f_i^{\otimes 2}, \Pi \rangle^2 \right\}$$

for the supremum being taken over a countably dense subset of the unit ball of $\Pi \in Op(L^2)$. For some $U \geq \|f_i^{\otimes 2}\|_{L^2}^2$ and $v_n = 2U\mathrm{E}Z + n\sigma^2$, the initial level $(1-\alpha)$ confidence set constructed is

$$C_{n,1-\alpha} = \left[\Sigma : Z \leq \mathrm{E}Z + \left\{ -2v_n \log(2\alpha)/n \right\}^{1/2} - U\log(2\alpha)/(3n) \right].$$

To make this confidence usable on real data, the norm of the Rademacher average, $R_n = n^{-1}\sum_{i=1}^n \varepsilon_i\{(f_i - \bar{f})^{\otimes 2} - \hat{\Sigma}\}$, will be used as a proxy for the unknown $\mathrm{E}Z$, which is justified by the symmetrization inequality also detailed in Appendix A. Note that R_n is also in $Op(L^2)$, because for any $\phi \in L^2(I)$ and for some $M \in \mathbb{R}$, $\|R_n\phi\|_{L^2} \leq |\varepsilon_i|\|\{(f_i - \bar{f})^{\otimes 2} - \hat{\Sigma}\}\phi\|_{L^2} \leq M\|\phi\|_{L^2}$ since $\{(f_i - \bar{f})^{\otimes 2} - \hat{\Sigma}\}$ is a bounded operator and $|\varepsilon_i| = 1$. For the bound U, in the case that there exists a fixed $c \in \mathbb{R}$ with $\|f_i\|_{L^2} \leq c$ for all i corresponding to a physical bound on the energy of f_i, $\|f_i^{\otimes 2}\|_p = \|f_i\|_{L^2}^2 \leq c^2 = U$. It will be determined in Appendix B that $U \geq \sigma$ in this case. In general, setting $U = \sigma$ gives good experimental results when f_i is Gaussian as will be discussed in later sections. This results in $v_n \approx \sigma^2/n$. For any $p \in [1, \infty)$ and $\alpha \in [0, 1/2]$, the proposed $(1-\alpha)$-confidence set for covariance operators is

$$C_{n,1-\alpha} = \left[\Sigma : \|\hat{\Sigma} - \Sigma\|_p \leq \|R_n\|_p + \sigma\left\{-2\log(2\alpha)/n\right\}^{1/2} - \sigma\log(2\alpha)/(3n)\right]. \tag{3.1}$$

where σ depends on the distribution on the functional data. As a rule of thumb for the choice of σ^2, as shown in Appendix B, is to note that $\sigma^2 \leq \|\mathrm{E}(f^{\otimes 4}) - \Sigma^{\otimes 2}\|_p$ and to estimate this bound empirically by $\hat{\sigma}^2$. For example, when the f_i are from a Gaussian process $\hat{\sigma} \leq 2^{1/2}\|\Sigma\|_p$ as explained in detail in Appendix B.3. In practice, $\|\Sigma\|_p$ is replaced with the consistent estimator $\|\hat{\Sigma}\|_p$. Consistency of the estimate follows from the central limit theorem in Banach spaces.

Constructing confidence sets in this way will lead to sets that are too large. That is, our $(1-\alpha)-$confidence set may have a coverage greater than the desired $1-\alpha$. While the level increases more quickly than desired, it does not increase too quickly to be useful as will be discussed in the applications of Section 4. Figure 2 displays the empirical coverage for five different operators. Specifically, for the five operators derived from the phoneme data sets of Section 5.1, 35 curves were generated as realizations of a mean zero Gaussian process with given covariance, the confidence set was constructed, and it was tested whether or not the true covariance operator lied within this set. This was repeated 10,000 to produce the estimates in Fig. 2. The choice of parameters makes the confidence set idea for such moderate sized datasets. If n is quite small, say $n < 10$, then the proposed confidence set is too small and would need to be adjusted. Similarly, if $n > 100$, then the confidence set is too conservative and can be tightened.

3.1. Discussion The constructed confidence set relies on many different facets. A main tool is the Rademacher average, which uses symmetrization to approximate and bound the unknown expectation. Such a technique

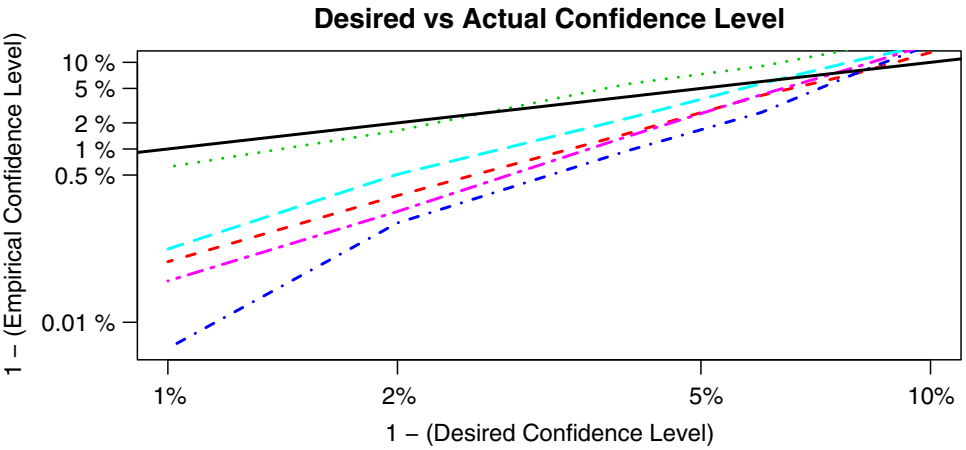

Figure 2: The empirical confidence level of the set from Eq. 3.1 for five different operators given a sample size of 35 curves generated from a Gaussian process. The black line is where the desired and empirical levels are equal. The desired level ranges from $\alpha = 1\%$ to $\alpha = 10\%$. 10,000 replications were used to produce these curves. The red, green, blue, cyan, and purple lines refer respectively to the phonemes /ɑ/ , /ɔ/ , /d/ , /i/ , and /ʃ/ , which are discussed in detail in Section 5.1. Of note is the green line which seems to differ from the others. This is the plosive /d/ , which is produced via a stop in air ow through the mouth as opposed to the other four phonemes considered

was used in Lounici and Nickl (2011) for wavelet deconvolution density estimators. This term can be simulated in practice by generating random Rademacher variates and computing the sum based on the observed data. While the simulation can be repeated for increased accuracy, often a single random draw is sufficient due to the concentration behaviour of the Rademacher sum.

As this confidence set is conservative, we chose to set the upper bound U at its most optimistic value being the weak variance. Weak variances, as well as other stronger variances achieved by permuting the expectation, summation, and supremum, arise often in the concentration literature (Boucheron et al., 2013). Computation of such weak variances under different norms and in the Gaussian case can be found in Appendix B. Experiments with heavier tailed and noisy data can be found in Appendix C. As an example, the difference between the weak variance for Gaussian and t distributed data is a multiplicative factor based on the degrees of freedom of the latter distribution. This method is adaptable to most settings. However, care must

be taken as the confidence set is often too conservative at first for practical use. Tuning the confidence set for specific applications is discussed in the subsequent section.

4 Applications

4.1. k-sample Comparison Testing for the equality of means among multiple sets of data is a common task in data analysis. In the functional setting, there has been recent work on performing such a test on covariance operators in order to test whether or not k sets of curves have similar variation. Panaretos et al. (2010) propose such a method for a two sample test on covariance operators given data from Gaussian processes. Similarly, Fremdt et al. (2013) propose a non-parametric two sample test on covariance operators. Both of these approaches make use of the Karhunen-Loève expansion and, hence, the underlying Hilbert space geometry. Pigoli et al. (2014) take a comparative look at a variety of metrics to rank their statistical power when used in a two sample permutation test.

Following from the results of Pigoli et al. (2014), our method uses the p-Schatten norms with the concentration inequality based confidence sets of the previous section to compare covariance operators. In the two sample setting, we are able to achieve similar statistical power to that of the permutation test after proper tuning of the coefficients in the inequalities. Furthermore, the analytic nature of the concentration approach leads to a significant reduction in computing time, which offers an even more significant savings for larger values of k as was already displayed in Fig. 1.

From the confidence set constructed in the previous section, we can devise a test for comparing the empirical covariance operators generated from k samples of functional data. Let the k samples be $f_1^{(1)}, \ldots, f_{n_1}^{(1)}, \ldots, f_1^{(k)}, \ldots, f_{n_k}^{(k)}$ where for each sample i and all elements $j = 1, \ldots, n_i$, $f_j^{(i)}$ has covariance $\Sigma^{(i)}$. Our goal is to design a test for the following two hypotheses:

$$\mathrm{H}_0: \quad \Sigma^{(1)} = \ldots = \Sigma^{(k)} \mathrm{H}_1: \quad \text{there exists } i, j \text{ such that } \Sigma^{(i)} \neq \Sigma^{(j)}.$$

To achieve this, a pooled estimate of the weak variance is computed as a weighted average of each sample's individual weak variance in similar style to that of a standard t-test. Let the total data size be $N = n_1 + \ldots + n_k$ and σ_i^2 be the weak variance for sample i, then the pooled variance is defined as $\sigma_{\text{pool}}^2 = N^{-1} \sum_{i=1}^k n_i \sigma_i^2$. Given Gaussian data and the p-Schatten norm, for example, this reduces to $\sigma_{\text{pool}}^2 = 2N^{-1} \sum_{i=1}^k n_i \|\Sigma^{(i)}\|_p^2$. In practice, σ_{pool}^2 is estimated from the data for the following confidence regions in order to have those regions only depend on the data.

Taking inspiration from the standard analysis of variance (Casella and Berger, 2002, Chapter 11), let $\hat{\Sigma}^{(i)}$ be the empirical estimate of the covariance operator for the ith sample, and let $\hat{\Sigma}$ be the estimate of the covariance operator for the total data set. Making use of the confidence sets for covariance operators from Section 3 gives the rejection region

$$\mathcal{C} = \left\{ f : \sum_{i=1}^{k} \|\hat{\Sigma}^{(i)} - \hat{\Sigma}\|_p > \sum_{i=1}^{k} \| \sum_{j=1}^{n_i} \varepsilon_{i,j} \left(f_j^{(i)\otimes 2} - \hat{\Sigma} \right) \|_p \right.$$
$$\left. + \left(\sum_{i=1}^{k} \frac{\sigma_{\text{pool}}^2}{n_i} \right)^{1/2} (-2\log 2\alpha)^{1/2} + \left(\sum_{i=1}^{k} \frac{\sigma_{\text{pool}}}{n_i} \right) \frac{\log 2\alpha}{3} \right\},$$

which under the null hypothesis will have size no greater than the desired α.

The size of the test induced by this rejection region is significantly less than the target size α due to the use of multiple concentration inequalities. Hence, tuning the inequalities is required to yield a useful test. Many experiments were run on simulated data sets generated as samples from a Gaussian process with randomly generated covariance operators whose eigenvalues were chosen to decay at a variety of rates. In this setting, the coefficients of $1 - k^{-1/2}$ for the Rademacher term and $(k+2)/(k+3)$ for the deviation term were determined experimentally to improve the size of the confidence region in the Gaussian process data setting:

$$\mathcal{C} = \left[f : \sum_{i=1}^{k} \|\hat{\Sigma}^{(i)} - \hat{\Sigma}\|_p > \left(1 - k^{-1/2}\right) \sum_{i=1}^{k} \| \sum_{j=1}^{n_i} \varepsilon_{i,j} \left(f_j^{(i)\otimes 2} - \hat{\Sigma} \right) \|_p \right.$$
$$\left. + \left(\frac{k+2}{k+3} \right) \left\{ \left(\sum_{i=1}^{k} \frac{\sigma_{\text{pool}}^2}{n_i} \right)^{1/2} (-2\log 2\alpha)^{1/2} + \left(\sum_{i=1}^{k} \frac{\sigma_{\text{pool}}}{n_i} \right) \frac{\log 2\alpha}{3} \right\} \right]. \tag{4.1}$$

The goal of these tweaked coefficients is to achieve to correct empirical size for the rejection region. The values were determined through running extensive simulations of Gaussian process data for a variety of operators, sample sizes n, and categories k, and choosing coefficients to tune the confidence sets to the desired sizes. The values of k were tested from 2 to 12; hence, a dataset containing many dozen categories may require more care. Ultimately, they should be used as a heuristic or a starting place for fine tuning this method to a specific problem of interest. For example, in Appendix C, we see that this approach can still apply to heavier t-distributed data with $(k+2)/(k+3)$ replaced with 1.

4.2. Classification of Operators Classification of functional data has been an area of heavy research over the last two decades. James and Hastie (2001) extend linear discriminant analysis to functional data. Hall et al. (2001) and Glendinning and Herbert (2003) classify with principal components. Ferraty and Vieu (2003) implement kernel estimators. General linear models for functional data are discussed by Müller and Stadtmüller (2005). Delaigle and Hall (2012) analyze the asymptotic properties of the centroid based classifier. Wavelet based classification is detailed by Berlinet et al. (2008) and Chang et al. (2014).

One application of our method beyond classification of functional data is the classification of covariance operators. In the setting of speech analysis, consider multiple speakers and multiple samples of speech from each speaker. The speech samples can be combined into a single sample covariance operator for each speaker. Then, our method can be employed, for example, to classify the covariance operators by speaker gender or speaker language. Evidence that this is a fruitful approach can be found in the analysis of Pigoli et al. (2014) and Pigoli et al. (2015) where a variety of metrics are compared for their efficacy when performing inference on covariance operators. These articles detail the discrepancy between sample covariance operators produced by speakers of different romance languages.

Given k possible labels and n samples of labeled data (Y_i, f_i) with label $Y_i \in \{1, \ldots, k\}$ and observation $f_i \in L^2(I)$, our goal is to determine the probability that a newly observed $g \in L^2(I)$ belongs to label $Y = j$. Given such a g, the Bayes classifier chooses the label $y = \arg\max_j \mathrm{P}(Y = j \mid g)$ where $\mathrm{P}(Y = j \mid g) = \mathrm{P}(g \mid Y = j)\mathrm{P}(Y = j)/\mathrm{P}(g)$.

Beginning with a training set of n samples with n_j samples of label j, the sample mean of each category is computed: $\bar{f}_j = n_j^{-1} \sum_{i:Y_i=j} f_i$. The probability $\mathrm{P}(g \mid Y = j)$ above is replaced with $\mathrm{P}\left(\|\bar{f}_j - g\|_{L^2} > \mathrm{E}\|\bar{f}_j - \mathrm{E}\bar{f}_j\|_{L^2} + r\right)$ with the goal of making a decision based on how much more $\bar{f}_j$ differs from g than $\bar{f}_j$ differs from its expectation $\mathrm{E}\bar{f}_j$. Similar techniques to those is Section 3 as used. Define the Rademacher sum, R_j, and the empirical weak variance, $\hat{\sigma}_j^2$, for label j to be, respectively,

$$R_j = \frac{1}{n_j} \sum_{i:Y_i=j} \varepsilon_i (f_i - \bar{f}_j), \hat{\sigma}_j^2 = \| \frac{1}{n_j} \sum_{i:Y_i=j} f_i^{\otimes 2} - \bar{f}_j^{\otimes 2} \|_p$$

where ε_i are independent and identically distributed Rademacher random variables. The tail bound for the above probability is then

$$\mathrm{P}\left(\|\bar{f}_j - g\|_{L^2} - \|R_j\|_{L^2} > r\right) < \exp\left(\frac{-n_j r^2}{4\|R_j\|_{L^2} U + 2\hat{\sigma}_j^2 + 2rU/3}\right), \quad (4.2)$$

where U is an upper bound on $\|f_i\|_{L^2}$. However, this can be approximated by the Gaussian tail $\exp\left(-n_j r^2/2\sigma_j^2\right)$. In the simulations of Section 5.3, this approximation actually achieves a better correct classification rate on both Gaussian and t-distributed data. This specifically works on t-distributed data as the estimate in Eq. 4.3 below is merely concerned with comparing the tail bounds rather than their specific values. Consequently, the tail for every category is underestimated in the t case, but the ratio remains valid for comparison purposes.

Assuming uniform priors on the labels, the estimate for the probability expression in the Bayes classifier is achieved by replacing the r on the right hand side of Eq. 4.2 with the observed $\|\bar{f}_j - g\|_{L^2} - \|R_j\|_{L^2}$. The result is

$$\mathrm{P}(Y{=}j \mid g) \approx \frac{\phi_j(g)}{\sum_{l=1}^{k} \phi_l(g)}, \phi_j(g) = \exp\left\{ -\frac{n_j}{2} \left(\frac{\|\bar{f}_j - g\|_{L^2} - \|R_j\|_{L^2}}{\hat{\sigma}_j} \right)^2 \right\}. \quad (4.3)$$

This can be extended to the case where an unlabeled observation is a collection of curves $g_1, \ldots, g_m$ by replacing $\|\bar{f}_j - g\|_{L^2}$ in the above expression with $\|\bar{\Sigma}_j - \hat{\Sigma}_g\|_p$ where $\bar{\Sigma}_j$ is the sample covariance of the f_i with label j and $\hat{\Sigma}_g$ is the sample covariance of the g_i. The Rademacher and weak variance terms would also be updated accordingly. The result would be a classifier that incorporates the covariance structure of the data into the decision.

4.3. Clustering of Operator Mixtures Closely related to the problem of classification is the problem of clustering. Given a sample of functional data, we want to assign one of a finite collection of labels to each curve. For example, in speech processing, one may want to cluster sound clips based the language of the speaker, or, to be discussed in Section 5.4, one may want to separate unlabeled phoneme curves into clusters.

There have been many recently proposed methods for clustering functional data. Many approaches begin by constructing a low dimensional representation of the data in some basis such as modelling the data with a B-spline basis followed by clustering the spline representations with k-means (Abraham et al., 2003). A similar approach makes use of the eigenfunctions of the covariance operator instead of B-splines (Peng and Müller, 2008). In contrast, we will attempt to cluster functions or operators directly via a concentration of measure approach similar to the previously described classification procedure.

Consider the same setting to the previous section of multiple observations from multiple categories. However, now the category labels are missing. This is a functional mixture model where each observed functional datum

is a stochastic process with one of k possible covariance operators. In the below experiments, the data will be simulated from a Gaussian process. The goal is to correctly separate the data into k sets. To achieve this, an expectation-maximization style algorithm is implemented.

Let the observed operator data be $S_1, \ldots, S_n \in Op(L^2)$ where each $S_i = \mathrm{cov}(f_1^{(i)}, \ldots, f_{m_i}^{(i)})$ is a rank m_i operator produced from m_i functional observations. Let the latent label variables be $Y_1, \ldots, Y_n \in \{1, \ldots, k\}$. Assuming no prior knowledge on the proportions of data in each category, the algorithm is initialized with the Jeffreys prior for the Dirichlet distribution by randomly generating $\rho_{i,\cdot}^{(0)} \sim \mathrm{Dirichlet}\,(1/2, \ldots, 1/2)$, the initial probability vector that $\mathrm{P}(Y_i = * \mid f_i)$.

Assuming t iterations of the algorithm have completed, we have a label probability vector $\rho_{i,\cdot}^{(t)}$ for each of the n observations. Given this collection of vectors, the expected proportions of each category can be estimated as $\tau_j^{(t+1)} = n^{-1} \sum_{i=1}^n \mathrm{E}\,(\mathbf{1}_{Y_i=j}) = n^{-1} \sum_{i=1}^n \rho_{i,j}^{(t)}$. Similarly, a weighted sum of the data, $\hat{\Sigma}_j^{(t+1)}$, and a weighted Rademacher sum, $R_j^{(t+1)}$, can be used to update the estimated covariance operators for each label j:

$$\hat{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^n \rho_{i,j}^{(t)} S_i}{\sum_{i=1}^n \rho_{i,j}^{(t)}}, \qquad R_j^{(t+1)} = \frac{\sum_{i=1}^n \rho_{i,j}^{(t)} \varepsilon_i \left(S_i - \hat{\Sigma}_j^{(t+1)} \right)}{\sum_{i=1}^n \rho_{i,j}^{(t)}}.$$

Lastly, a pooled weak variance is required, which is used in place of each individual category weak variance. Otherwise, in practice, one single category captures all of the data points. By defining the pooled covariance operator as $\hat{\Sigma}_{\mathrm{pool}}^{(t+1)} = \sum_{j=1}^k \tau_j^{(t+1)} \hat{\Sigma}_j^{(t+1)}$, then the pooled weak variance in the Gaussian case, for example, is estimated by $2\|\hat{\Sigma}_{\mathrm{pool}}^{(t+1)}\|_p$.

As a result, the label probability vectors $\rho_{i,\cdot}^{(t)}$ can be updated given the $t+1$st collection of estimated covariance operators, Rademacher sums, and the pooled covariance operator. From the previous section, Eq. 4.3 can be used to determine $\rho_{i,j}^{(t+1)} = \mathrm{P}(Y_i = j \mid S_i, \hat{\Sigma}_1^{(t+1)}, \ldots, \hat{\Sigma}_k^{(t+1)})$, the probability that observation i belongs to the jth category. This process can be iterated until a local optimum is reached.

5 Numerical Experiments

5.1. Simulated and Phoneme Data To test each of the above three applications, experiments were first run on simulated data. These data sets were generated as mean zero observations from Gaussian or t-distributed processes with randomly selected covariance operators. These were selected

by choosing a specific decay rate for the eigenvalues in a diagonal operator D, by generating a random orthonormal basis U, and then combining them as $\Sigma = UDU^{\mathrm{T}}$.

Secondly, the phoneme data to be tested (Ferraty and Vieu, 2003; Hastie et al., 1995) is a collection of 400 log-periodograms for each of five different phonemes: /ɑ/ as in the vowel of "dark"; /ɔ/ as in the first vowel of "water"; /d/ as in the plosive of "dark"; /i/ as in the vowel of "she"; /ʃ/ as in the fricative of "she". Each curve contains the first 150 frequencies from a 32 ms sound clip sampled at a rate of 16-kHz.

5.2. k-sample Comparison The above confidence set in Eq. 4.1 comparing k samples can be used to refute the null hypothesis that all covariance operators are equal. A two sample permutation test was performed in Pigoli et al. (2014). Given two samples of functional data, $f_1^{(1)}, \ldots, f_n^{(1)}$ and $f_1^{(2)}, \ldots, f_m^{(2)}$ with associated covariance operators $\Sigma^{(1)}$ and $\Sigma^{(2)}$, respectively, the desired hypotheses to test are

$$\mathrm{H}_0 : \quad \Sigma^{(1)} = \Sigma^{(2)} \qquad \mathrm{H}_1 : \quad \Sigma^{(1)} \neq \Sigma^{(2)}.$$

When using a permutation test, the labels are randomly reassigned M times, and each time, the distance between the two new covariance operators is computed. For sufficiently large M, this procedure will return the exact significance level of the observations with respect to the data set.

A power analysis was performed between the permutation method and our proposed concentration approach using Eq. 4.1. Two different operators $\Sigma^{(1)}$ and $\Sigma^{(2)}$ were randomly generated by first generating a random basis of eigenvectors. Let $M^{(1)}$ and $M^{(2)}$ be $m \times m$ matrices with iid standard normal entries. Let $U^{(i)}$ be the matrix of eigenvectors from the symmetric matrix ${M^{(i)}}^{\mathrm{T}} M^{(i)}$. Then, $\Sigma^{(i)} = U^{(i)} \Lambda {U^{(i)}}^{\mathrm{T}}$ where Λ is a diagonal matrix of eigenvalues which is the same for both $\Sigma^{(1)}$ and $\Sigma^{(2)}$. In the below experiments, $m = 15$ and the eigenvalues are $500 \times j^{-4}$ and $500 \times j^{-2}$ for $j = 1, \ldots, 15$ for Figs. 3 and 4, respectively.

Given $\Sigma^{(1)}$, $\Sigma^{(2)}$ and a $\gamma \geq 0$, an interpolation between the two operators is constructed as $\Sigma^{(\gamma)} = [(\Sigma^{(1)})^{1/2} + \gamma\{S(\Sigma^{(2)})^{1/2} - (\Sigma^{(1)})^{1/2}\}][(\Sigma^{(1)})^{1/2} + \gamma\{S(\Sigma^{(2)})^{1/2} - (\Sigma^{(1)})^{1/2}\}]^*$, where S is an operator minimizing the Procrustes distance, between $\Sigma^{(1)}$ and $\Sigma^{(2)}$, which is $d_{\mathrm{Proc}}(\Sigma^{(1)}, \Sigma^{(2)})^2 = \inf_{S \in U\{L^2(I)\}} \|R^{(1)} - R^{(2)} S\|_2^2$ where $\Sigma^{(i)} = (R^{(i)})(R^{(i)})^*$ and $U\left\{L^2(I)\right\}$ is the space of unitary operators on $L^2(I)$ (Pigoli et al., 2014).

Monte Carlo simulations were run in order to estimate the power of each test. Two operators $\Sigma^{(1)}$ and $\Sigma^{(2)}$ with similar eigenvalue decay were compared. with a sample size $n = 50$ and $\gamma \in \{0, .1, .2, .3, .4, .5\}$. For each

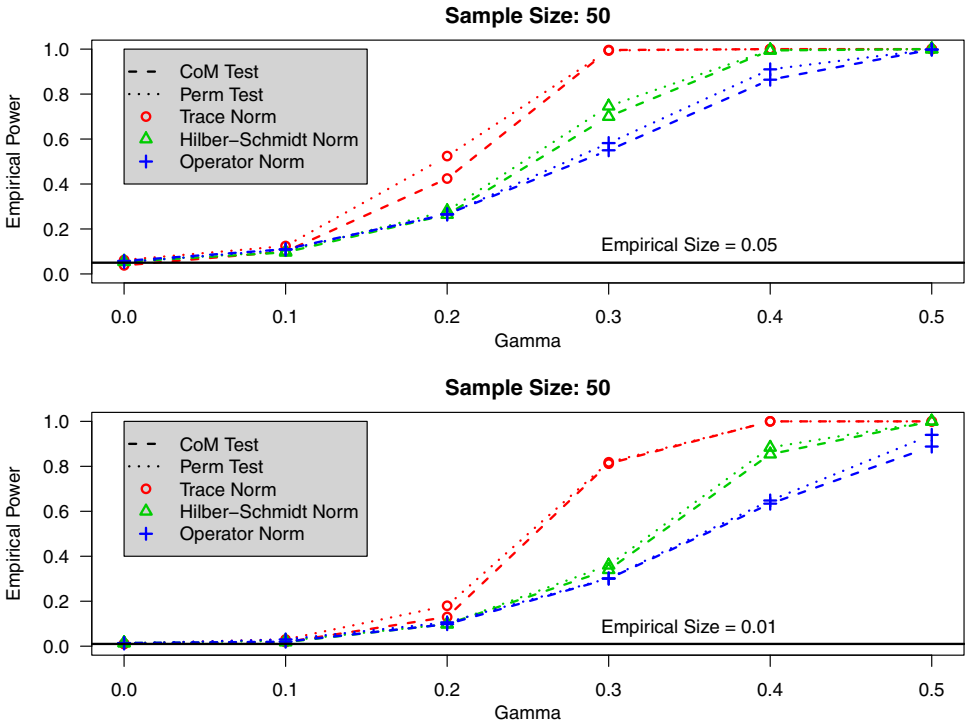

Figure 3: A power analysis for testing whether or not operator $\Sigma^{(1)} = \Sigma^{(\gamma)}$ comparing the permutation method (short dashed lines) with the concentration approach (long dashed lines). The size $\alpha = 0.05$ in the top plot, and $\alpha = 0.01$ in the bottom. The eigenvalues of the operators decay at a rate $O(k^{-4})$. The red circle, green triangle, and blue plus lines respectively correspond to the trace class, Hilbert-Schmidt, and operator norms

γ, 5000 samples of size n were generated for $\Sigma^{(1)}$ and $\Sigma^{(\gamma)}$. Equation 4.1 and the permutation method (Pigoli et al., 2014) were both implemented to estimate the empirical power.

Figures 3 and 4 display the results for operators whose eigenvalues decay at a quartic and quadratic rate, respectively. The short dashed lines indicate the power of the permutation test, and the long dashed lines indicate the power of our concentration approach. The colors red, green, and blue and the points circle, triangle, and plus correspond to the three norms trace, Hilbert-Schmidt, and operator, respectively.

In most cases, the concentration approach is able to achieve the similar power to reject the null as does the permutation test. The notable exception

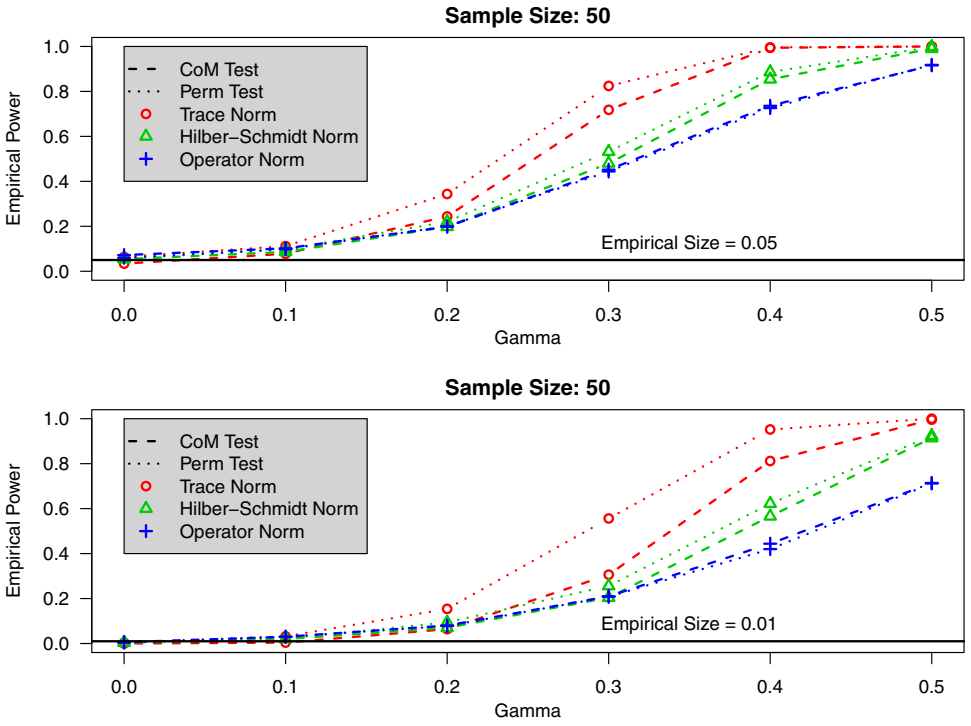

Figure 4: A power analysis for testing whether or not operator $\Sigma^{(1)} = \Sigma^{(\gamma)}$ comparing the permutation method (short dashed lines) with the concentration approach (long dashed lines). The size $\alpha = 0.05$ in the top plot, and $\alpha = 0.01$ in the bottom. The eigenvalues of the operators decay at a rate $O(k^{-2})$. The red circle, green triangle, and blue plus lines respectively correspond to the trace class, Hilbert-Schmidt, and operator norms

is for the trace norm when the eigenvalues decay slowly, which is the lower plot in Fig. 4. The added benefit to the concentration approach is the speed with which it executes. Across all of the Monte Carlo simulations, our concentration approach ran on average 140.7 times faster than the permutation method based on running the method with 500 permutations. This was computed by tracking the amount of computation time each method spent while producing the plots in Figs. 3 and 4, which corresponds to 6 values of γ, 2 values of α, 3 different norms, and 5000 replications each resulting in 180,000 function calls for both the permutation and concentration methods. Unlike the other norms, the Hilbert-Schmidt norm can be calculated without explicit computation of the eigenvalues. For each evaluation of the permutation

test, 500 permutations of the data were generated, which corresponds to 500 random draws and 500 eigenvalue computations. More accuracy would require even more permutations. In comparison, our concentration approach requires only $3k$ eigenvalue computations and no random draws and hence is only dependent on the number of samples regardless of data size or α.

The proposed k-sample test was also used to compare samples of log-periodogram curves from the spoken phonemes /ɑ/ and /ɔ/ . As one can imagine, these vowels can be hard to distinguish; see Section 5.4 for further evidence of this. For $k \in \{2,3,4,5,6\}$, $k-1$ disjoint sets of 40 /ɑ/ curves and one set of 40 /ɔ/ curves were randomly sampled from the data set. This was replicated 500 times, and each time (4.1) was used to decide whether or not the k covariance operators were equivalent at the $\alpha = 0.05$ level. The resulting estimated statistical power for each k is

k	2	3	4	5	6
Power	0.00	0.018	0.228	0.656	0.936

The low power for small values of k results from the conservative nature of this test, but also from the fact that the phonemes /ɑ/ and /ɔ/ are quite difficult to separate unlike other pairings.

In the null setting, the above experiment was rerun except that every disjoint set of curves came from the /ɑ/ set. The resulting experimentally computed test sizes are

k	2	3	4	5	6
Size	0.00	0.00	0.00	0.004	0.072

For small values of k, we see that the sizes are significantly below the desired size. Hence, the test is too conservative, which corresponds to the lack of power. As the number of categories increases, we roughly achieve the desired test size.

5.3. Binary and Trinary Classification Our concentration of measure (CoM) method is implemented on covariance operators making use of the trace norm $\|\cdot\|_{\text{tr}}$ where for a covariance operator Σ with eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$, $\|\Sigma\|_{\text{tr}} = \sum_{i=1}^{n} |\lambda_i|$. The trace norm was chosen based on the analysis of the preceding section as well as that of Pigoli et al. (2014) where it achieved the best performance when compared with the other p-Schatten norms. The CoM approach to classification of operators is tested in a variety of simulations against other standard approaches to functional classification. The methods used for comparison are k-nearest neighbours (Ferraty and Vieu, 2006), classification using kernel estimators (Ferraty and Vieu, 2003), general linear model (Müller and Stadtmüller, 2005), and regression trees.

The first simulation asks each method to classify observed mean zero Gaussian process data or mean zero t-process data with 4 degrees of freedom. The two covariance operators in question, Σ_1 and Σ_2, are the sample covariances of the male and of the females of the Berkeley growth curve data (Ramsay and Silverman, 2005). In particular, n collections of k curves were generated from each of Σ_1 and Σ_2 as a training set, and m collections of k curves were generated as a test set. The CoM method was trained on the set of n sample covariances and used to classify each of the m test covariances. The remaining classification methods were trained and tested in two separate ways: By treating each sample covariance as a function and classifying as usual, and by training on all $n \times k$ observations and testing each of the m collections by classifying each constituent curve individually and taking a majority vote with ties settled by a uniform random draw.

For group sizes $k = 1, 2, 4, 8, 16$, 100 simulations were run with $n = 100$ sets of k training curves. To compare the accuracy of each approach $m = 100$ sets of k testing curves were generated for each operator. The accuracy of each method is tabulated in Table 1.

The concentration method performed well against the alternatives. Its performance was on par with the kernel method applied to each covariance operator as a function. Our method was only consistently outperformed by the kernel method implementing the majority vote approach. However, the two operators in question have very similar weak variances. The next simulation demonstrates how the concentration method adapts naturally when the variances of each label significantly differ.

Continuing from the previous simulation, a third operator is constructed from Σ_1 and Σ_2 by averaging these two and then scaling up the non-principal eigenvalues by a factor of 5. This, in some sense, creates a third operator between the first two, but with higher variance. The simulation is carried out precisely as before, but incorporating all three operators. In this setting, our concentration approach demonstrates the best performance. The results are listed in Table 2.

These five methods tested on simulated data were also tested against phoneme data. Across 50 iterations, each set of 400 curves was partitioned at random into an 100 curve training set and a 300 curve testing set. The five classifiers were trained and run on each of the 300×5 curves individually. For our concentration of measure approach, the rank one operator associated to each individual curve was compared with the covariance operator formed from the 100×5 training curves. The results are detailed in Table 3. Our concentration of measure approach only uniformly outperforms the

Table 1: A comparison of the performances of five classification methods including our concentration of measure approach (CoM), k-nearest-neighbours (KNN), kernel method (Kernel), generalized linear model (GLM), and regression trees (Tree), on a binary classification problem

	Gaussian					t-distributed				
k	1	2	4	8	16	1	2	4	8	16
CoM	62 (5)	62 (5)	76 (8)	87 (6)	96 (3)	59 (5)	62 (5)	75 (7)	86 (6)	95 (4)
KNN	52 (4)	44 (4)	57 (4)	76 (4)	91 (2)	42 (4)	45 (4)	58 (4)	76 (3)	92 (2)
KNN′	·	47 (3)	59 (4)	74 (3)	89 (2)	·	45 (4)	58 (4)	72 (3)	89 (3)
Kernel	**65** (4)	64 (5)	75 (3)	87 (3)	96 (2)	**65** (4)	64 (5)	75 (4)	87 (2)	96 (2)
Kernel′	·	**70** (4)	**82** (2)	**92** (2)	**99** (1)	·	**68** (4)	**80** (3)	**92** (2)	**99** (1)
GLM	51 (4)	62 (4)	74 (4)	86 (2)	94 (2)	50 (3)	62 (3)	74 (4)	86 (3)	94 (2)
GLM′	·	50 (4)	50 (3)	50 (4)	50 (4)	·	50 (3)	50 (3)	51 (4)	50 (3)
Tree	57 (4)	54 (4)	59 (3)	66 (4)	75 (3)	54 (4)	54 (4)	59 (3)	67 (3)	75 (3)
Tree′	·	55 (4)	59 (4)	60 (5)	60 (9)	·	54 (4)	57 (4)	59 (5)	57 (6)

The first entry for each method corresponds to classifying the covariance operators as functions. The prime entry corresponds to classifying curves with a majority vote. The estimated percent of correct classification is listed in the table with the sample standard deviation in brackets. The left block comes from Gaussian process data, and the right comes from t-process data with 4 degrees of freedom. The highest percentage of each column is marked in bold

Table 2: A comparison of the performances of five classification methods as in Table 1, but with three potential classes from which to choose

	Gaussian					t-distributed				
k	1	2	4	8	16	1	2	4	8	16
CoM	51 (4)	55 (4)	**75** (5)	**89** (5)	**97** (3)	46 (5)	50 (6)	63 (5)	75 (8)	85 (7)
KNN	50 (3)	52 (3)	61 (3)	75 (3)	90 (2)	46 (3)	49 (3)	57 (3)	67 (3)	78 (2)
KNN′	·	55 (3)	68 (3)	80 (2)	90 (2)	·	50 (3)	64 (3)	**77** (2)	**87** (2)
Kernel	**54** (3)	52 (3)	64 (3)	77 (3)	92 (2)	**50** (3)	48 (3)	57 (3)	68 (3)	80 (2)
Kernel′	·	**58** (3)	69 (2)	81 (3)	92 (2)	·	**53** (3)	**66** (3)	**77** (2)	85 (2)
GLM	36 (4)	41 (4)	49 (4)	57 (3)	65 (3)	35 (3)	41 (4)	46 (4)	53 (3)	58 (4)
GLM′	·	35 (4)	36 (4)	36 (5)	35 (5)	·	35 (3)	36 (4)	36 (4)	36 (6)
Tree	44 (3)	44 (3)	45 (3)	50 (3)	55 (3)	42 (3)	44 (3)	45 (3)	46 (3)	49 (3)
Tree′	·	46 (3)	51 (4)	51 (7)	47 (7)	·	43 (3)	47 (4)	48 (6)	46 (8)

The estimated percent of correct classification is listed in the table with the sample standard deviation in brackets. The left block comes from Gaussian process data, and the right comes from t-process data with 4 degrees of freedom. The highest percentage of each column is marked in bold

Table 3: Percentage of correct classification of the five phonemes against the five methods

	/ɑ/	/ɔ/	/d/	/i/	/ʃ/
CoM	76.9	76.8	96.6	**98.5**	99.4
KNN	72.4	79.1	**98.5**	97.4	**100.**
Kernel	72.0	**80.5**	98.4	97.2	99.9
GLM	**79.0**	72.3	98.2	95.9	99.2
Tree	70.8	69.4	95.6	87.8	92.6

The highest percentage of each column is marked in bold

regression tree classifier, but has comparable performance to the other three methods, and none of the competing methods uniformly outperforms ours.

5.4. The Expectation-Maximization Algorithm in Practice The experiments described and depicted below make use of the trace norm only. It was determined through experimentation that the expectation-maximization algorithm we propose in Section 4.3 does not perform well under the topology of either the Hilbert-Schmidt or operator norms as they give more emphasis to the principal eigenvalue at the expense of the others. The usual behavior under these norms is for all estimates to converge to the average of all of the data points. This is in contrast to the better performance of the algorithm making use of the trace norm, which is somewhat more uniform in its treatment of the eigenstructure.

As a first test case, this algorithm was run given three target covariance operators, which were constructed by taking three randomly generated orthonormal bases U_i and a diagonal operator D of eigenvalues decaying at a rate $\lambda_k = O(k^{-4})$ and multiplying $\Sigma_i = U_i D {U_i}^{\mathrm{T}}$. Let the three target covariance operators be denoted as Σ_a, Σ_b, and Σ_c. For each of these operators, 500 rank four data points were generated from a zero mean Gaussian process. From the data, the algorithm initializes three estimates $\hat{\Sigma}_1^{(t)}$, $\hat{\Sigma}_2^{(t)}$, and $\hat{\Sigma}_3^{(t)}$, which attempt to locate the three target operators as the method iterates. After 15 iterations, the original 1500 data points were perfectly separated into three groups. To make the problem harder, a second test case was run identical to the first except that the observed operators are all of rank one. Here the algorithm had a harder time separating the data. The inaccuracy in the rank one setting is equivalent to the poor performance of classification of rank one operators detailed in Tables 1 and 2.

The resulting clusters from both tests as well as a comparison with the k-means method are in Table 4. The k-means algorithm was run with 50 iterations and 10 random starts. It still performed much worse than the concentration based method in the rank 4 setting. This is because the

Table 4: Clustering of simulated operators

	Rank 4 Operators			Rank 1 Operators		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
	Concentration					
Label a	500	0	0	318	0	182
Label b	0	500	0	0	335	165
Label c	0	0	500	295	0	205
	k-means					
Label a	261	0	239	219	0	281
Label b	0	290	210	0	179	321
Label c	0	0	500	211	0	289

The concentration approach performs better than k-means as it takes better account of the covariance structure present in each cluster

concentration approach focuses its clustering heavily on the covariance structure of the data whereas k-means does not. The concentration method arguably did better in the rank 1 case as well specifically in the cluster 2 column which more thoroughly captured the label b data.

For the phoneme data, all 400 sample curves from each of the five phoneme sets were clustered individually as curves. The algorithm was run for 20 iterations and told to partition the data into five clusters. The results are in Table 5. Clusters A and B partitioned almost all of the vowels /ɑ/ and /ɔ/ , which, recalling their definition in Section 5.1, are quite similar in sound. Clusters C, D, and E contain the majority of /d/ , /i/ , and /ʃ/ curves, respectively. Very similar results were achieved by the tried and true k-means clustering algorithm running with 50 iterations and 10 random starts. The proposed concentration based expectation-maximization algorithm is hence an effective method for the unsupervised clustering of phonemes.

Acknowledgements. JA is grateful that this research was supported by EPSRC grant EP/K021672/2.

Table 5: Clustering 2000 phoneme curves into 5 clusters

	Concentration					k-means				
Cluster	A	B	C	D	E	A	B	C	D	E
/ɑ/	281	119	0	0	0	281	119	0	0	0
/ɔ/	125	273	1	1	0	126	272	1	1	0
/d/	0	0	384	15	1	0	2	386	10	2
/i/	1	0	1	393	5	1	3	2	381	13
/ʃ/	0	0	0	3	397	0	0	0	2	398

Similar results achieved by both the concentration and k-means methods

References

ABRAHAM, C., CORNILLON, P.-A., MATZNER-LØBER, E. and MOLINARI, N. (2003). Unsupervised curve clustering using b-splines. *Scand. J. Stat.* **30**, 3, 581–595.

ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010). Some nonasymptotic results on resampling in high dimension, i: Confidence regions. *Ann. Stat.* **38**, 1, 51–82.

BARTLETT, P.L. and MENDELSON, S. (2003). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3**, 463–482.

BARTLETT, P.L., BOUCHERON, S. and LUGOSI, G. (2002). Model selection and error estimation. *Mach. Learn.* **48**, 1–3, 85–113.

BERLINET, A., BIAU, G. and ROUVIERE, L. (2008). Functional supervised classification with wavelets. In *Annales de l'ISUP*, volume 52.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

CABASSI, A. and KASHLAK, A.B. (2016). fdcov: Analysis of Covariance Operators. R package version 1.0.0.

CASELLA, G. and BERGER, R.L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove.

CHANG, C., CHEN, Y. and OGDEN, T. (2014). Functional data classification: a wavelet approach. *Comput. Stat.* **29**, 6, 1497–1513.

DE LA PENA, V. and GINÉ, E. (2012). Decoupling: From dependence to independence. Springer Science & Business Media.

DELAIGLE, A. and HALL, P. (2012). Achieving near perfect classification for functional data. *J. R. Statist. Soc. Series B (Statist. Methodol.)* **74**, 2, 267–286.

FAN, Z. (2011). Confidence regions for infinite-dimensional statistical parameters. Part III essay in Mathematics, University of Cambridge. http://web.stanford.edu/zhoufan/PartIIIEssay.pdf.

FERRATY, F. and VIEU, P. (2003). Curves discrimination: A nonparametric functional approach. *Comput. Statist. Data Anal.* **44**, 1, 161–173.

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.

FREMDT, S., STEINEBACH, J.G., HORVÁTH, L. and KOKOSZKA, P. (2013). Testing the equality of covariance operators in functional samples. *Scand. J. Stat.* **40**, 1, 138–152.

GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Stat.* **38**, 2, 1122–1170.

GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

GLENDINNING, R.H. and HERBERT, R.A. (2003). Shape classification using smooth principal components. *Pattern Recogn. Lett.* **24**, 12, 2021–2030.

HALL, P., POSKITT, D.S. and PRESNELL, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43**, 1, 1–9.

HASTIE, T., BUJA, A. and TIBSHIRANI, R. (1995). Penalized discriminant analysis. *Ann. Stat.*, 73–102.

HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*, volume 200. Springer Science & Business Media.

ISSERLIS, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika* **12**, 1/2, 134–139.

JAMES, G.M. and HASTIE, T.J. (2001). Functional linear discriminant analysis for irregularly sampled curves. J. R. Statist. Soc. Series B, Statist. Methodol., 533–550.

JIANG, C.-R., ASTON, J.A. and WANG, J.-L. (2016). A functional approach to deconvolve dynamic neuroimaging data. *J. Am. Stat. Assoc.* **111**, 513, 1–13.

KERKYACHARIAN, G., NICKL, R. and PICARD, D. (2012). Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probab. Theory Relat. Fields* **153**, 1–2, 363–404.

KOLTCHINSKII, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory* **47**, 5, 1902–1914.

KOLTCHINSKII, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *Ann. Stat.* **34**, 6, 2593–2656.

LEDOUX, M. (2001). *The Concentration of Measure Phenomenon*, volume 89. American Mathematical Soc.

LOUNICI, K. and NICKL, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators. *Ann. Stat.* **39**, 1, 201–231.

MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.*, 774–805.

PANARETOS, V.M., KRAUS, D. and MADDOCKS, J.H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *J. Am. Stat. Assoc.* **105**, 490, 670–682.

PENG, J. and MÜLLER, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Statist.*, 1056–1077.

PIGOLI, D., ASTON, J.A.D., DRYDEN, I.L. and SECCHI, P. (2014). Distances and inference for covariance operators. *Biometrika*, page asu008.

PIGOLI, D., HADJIPANTELIS, P.Z., COLEMAN, J.S. and ASTON, J.A.D. (2015). The analysis of acoustic phonetic data: exploring differences in the spoken romance languages. arXiv:1507.07587.

RAMSAY, J.O. and SILVERMAN, B.W. (2005). *Functional Data Analysis*. Springer, New York.

TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae* **126**, 3, 505–563.

Appendix A: Confidence Sets for the Mean in Banach Spaces

The goal of this section is to construct a non-asymptotic confidence region in the Banach space setting. This is specialized in Section 3 to our case of interest, covariance operators, when the X_i below are replaced with $f_i^{\otimes 2}$.

Let $X_1, \ldots, X_n \in (B, \|\cdot\|_B)$ be mean zero independent and identically distributed Banach space valued random variables with $\|X_i\|_B \leq U$ for all $i = 1, \ldots, n$ where U is some positive constant. Furthermore, let $\langle \cdot, \cdot \rangle : B \times B^* \to \mathbb{R}$ such that for $X \in B$ and $\phi \in B^*$ then $\langle X, \phi \rangle = \phi(X)$. Define

$$Z = \sup_{\|\phi\|_{B^*} \leq 1} \sum_{i=1}^{n} \langle X_i, \phi \rangle = \left\| \sum_{i=1}^{n} X_i \right\|_B, \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\phi\|_{B^*} \leq 1} \mathrm{E} \langle X_i, \phi \rangle^2,$$

where the supremum is taken over a countably dense subset of the unit ball of B^*. Furthermore, define $v_n = 2U\mathrm{E}Z + n\sigma^2$. Then, $\mathrm{P}\left(Z > \mathrm{E}Z + r\right) \leq \exp\{-r^2/(2v_n + 2rU/3)\}$. Rewriting Z as $n\left\|\bar{X} - \mathrm{E}\bar{X}\right\|_B$ results in

$$\mathrm{P}\left(\left\|\bar{X} - \mathrm{E}\bar{X}\right\|_B > \mathrm{E}\left\|\bar{X} - \mathrm{E}\bar{X}\right\|_B + r\right) < \exp\left(\frac{-n^2r^2}{2v_n + 2nrU/3}\right)$$

where $\|X_i\|_B < U$ and $v_n = 2nU\mathrm{E}\{\left\|\bar{X} - \mathrm{E}\bar{X}\right\|_B\} + n\sigma^2$.

The above tail bound incorporates the unknown $\mathrm{E}(\|\bar{X} - \mathrm{E}\bar{X}\|_B)$. Consequently, a symmetrization technique is used. This term is replaced by the norm of the Rademacher average $R_n = n^{-1}\sum_{i=1}^n \varepsilon_i(X_i - \bar{X})$ where the ε_i are independent and identically distributed Rademacher random variables also independent of the X_i. This substitution is justified by invoking the symmetrization inequality (Giné and Nickl, 2016, Theorem 3.1.21),

$$\mathrm{E}Z = \mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n (X_i - \mathrm{E}\bar{X})\right\|_B \leq 2\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(X_i - \bar{X})\right\|_B = 2\mathrm{E}\|R_n\|_B.$$

If the data are symmetric about their mean, which is when $X_i - \mathrm{E}X_i$ and $\mathrm{E}X_i - X_i$ are equidistributed, the coefficient of 2 is unnecessary and can be dropped. This is because $X_i - \mathrm{E}X_i$ and $\varepsilon\{X_i - \mathrm{E}X_i\}$ are also equidistributed. In practice, the data may not be symmetric. However, averaging even a moderately sized data set has a symmetrizing effect on the sample mean. Assuming the data is not highly skewed, the coefficient of 2 can be safely dropped in practice to tighten the confidence set. In fact, considering the phoneme data from Section 5.1 in this setting results in the values displayed in Table 6, which shows that in the trace norm setting, the Rademacher average is much greater than half the size of $\mathrm{E}Z$, and that in the Hilbert-Schmidt and operator norm settings, the Rademacher average is actually marginally less than $\mathrm{E}Z$.

This symmetrization result allows us to replace the original expectation with the expectation of the Rademacher average. Furthermore, Talagrand's inequality also applies to R_n. Hence, the Rademacher average concentrates strongly about its expectation, which justifies dropping the expectation. In practice, one can use the intermediary $\mathrm{E}_\varepsilon\|R_n\|_B$, which can be approximated for reasonable sized data sets via Monte Carlo simulations of the ε_i. However, this is not strictly necessary, and for large data sets, a single random draw of ε_i will suffice (Giné and Nickl, 2016, Section 3.4.2).

Table 6: A comparison of the left and right hand sides of the symmetrization inequality and, hence, a justification for safely dropping the coefficient of 2 in the construction of confidence sets

	Trace		Hilbert-Schmidt		Operator	
	$\mathrm{E}Z$	$\mathrm{E}\|R_n\|$	$\mathrm{E}Z$	$\mathrm{E}\|R_n\|$	$\mathrm{E}Z$	$\mathrm{E}\|R_n\|$
/ɑ/	618.3	554.8	112.4	119.5	76.5	81.1
/ɔ/	591.3	525.2	108.7	112.2	70.9	74.2
/d/	506.8	450.7	105.6	115.0	83.3	92.1
/i/	610.4	545.9	107.6	111.2	63.5	72.3
/ʃ/	419.3	363.1	67.4	71.1	40.1	43.0

These numbers were computed for a sample size of $n = 60$ from the phoneme data set. The computation was repeated 100 times and averaged to approximate the following expectations

The resulting $(1-\alpha)$-confidence set is

$$\left\{ X : \left\| X - \bar{X} \right\|_B \leq \|R_n\|_B + \left\{ \frac{2}{n} \log(2\alpha) \left(\sigma^2 + 2U \|R_n\|_B \right) \right\}^{1/2} + \frac{U \log(2\alpha)}{3n} \right\}. \tag{A.1}$$

To make use of these results in practice, both the weak variance σ^2 must be estimated for the data and a reasonable choice of U must be made, and a main contribution of this present paper is to propose some theoretically motivated but practically useful non-asymptotic choices for these constants that work for the functional data applications we are investigating.

Appendix B: Calculation of the Weak Variance

B.1 The Weak Variance for $p \in [1, \infty)$

To calculate the weak variance σ^2, define $f^{\otimes n} = f \otimes \ldots \otimes f$ to be the n-fold tensor product of f with itself and extend the definition of $\langle \cdot, \cdot \rangle : (L^2)^{\otimes 4} \times \{(L^2)^{\otimes 4}\}^* \to \mathbb{R}$ such that $\left\langle f^{\otimes 4}, \phi^{\otimes 4} \right\rangle = \left\langle f^{\otimes 2}, \phi^{\otimes 2} \right\rangle^2 = \langle f, \phi \rangle^4 = \phi(f)^4$. For operators $\Pi \in \{(L^2)^{\otimes 2}\}^*$ and $\Xi \in \{(L^2)^{\otimes 4}\}^*$, the weak variance is

$$\begin{aligned} \sigma^2 &= \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\Pi\|_q \leq 1} \mathrm{E} \left\langle f_i^{\otimes 2} - \mathrm{E} f_i^{\otimes 2}, \Pi \right\rangle^2 \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\Xi\|_q \leq 1} \left\langle \mathrm{E} f_i^{\otimes 4} - \left\{ \mathrm{E} f^{\otimes 2} \right\}^{\otimes 2}, \Xi \right\rangle \leq \left\| \mathrm{E} f^{\otimes 4} - \Sigma^{\otimes 2} \right\|_p \end{aligned}$$

where the inequality stems from the fact that the supremum is being taken over a larger set. However, in the Hilbert space setting, the dual of the

tensor product does coincide with the tensor product of the dual space, and thus the above inequality can be replaced with an equality if the Hilbert-Schmidt norm, 2-Schatten norm, is used. Given a bound $\|f_i\|_{L^2}^2 \leq c^2 = U$, then $\sigma^2 \leq \|\mathrm{E} f^{\otimes 4}\|_p \leq \mathrm{E}\|f\|_{L^2}^4 \leq c^4 = U^2$.

B.2 The Weak Variance for $p = \infty$

Let E be a countable dense subset of the unit ball of $L^2(I)$. In the case $p = \infty$, we cannot use duality, but can still write Z and σ^2 as suprema over the countable set and achieve the same results as above.

$$
\begin{aligned}
Z &= \frac{1}{n} \sup_{e \in E} \sum_{i=1}^{n} \left\langle \left\{ f_i^{\otimes 2} - \mathrm{E} f_i^{\otimes 2} \right\} e, e \right\rangle = \sup_{e \in E} \left\langle (\hat{\Sigma} - \Sigma) e, e \right\rangle = \left\| \hat{\Sigma} - \Sigma \right\|_{\infty}, \\
\sigma^2 &= \frac{1}{n} \sum_{i=1}^{n} \sup_{e_1 \in E} \mathrm{E} \left\langle (f_i^{\otimes 2} - \Sigma) e_1, e_1 \right\rangle^2 \leq \frac{1}{n} \sum_{i=1}^{n} \sup_{e_1, e_2 \in E} \mathrm{E} \left\langle f_i^{\otimes 2} - \Sigma, e_1 \otimes e_2 \right\rangle^2 \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \sup_{e_1, e_2 \in E} \left\langle \left(\mathrm{E} f_i^{\otimes 4} - \Sigma^{\otimes 2} \right) (e_1 \otimes e_2), e_1 \otimes e_2 \right\rangle = \left\| \mathrm{E} f_i^{\otimes 4} - \Sigma^{\otimes 2} \right\|_{\infty}.
\end{aligned}
$$

As before, if $\left\| f_i^{\otimes 2} \right\|_{\infty} = \|f_i\|_{L^2}^2 \leq c^2 = U$, then $\sigma^2 \leq U^2$.

B.3 The Weak Variance for Gaussian Data

Similarly to the bounded case, we estimate $\left\| \mathrm{E} f^{\otimes 4} - \Sigma^{\otimes 2} \right\|_p$ for Gaussian data. Consider f from a Gaussian process with mean zero and covariance Σ. Strictly speaking these variables are not norm bounded, but similar concentration results for Gaussian processes can be derived. Indeed, let $f_1, \ldots, f_n$ be independent Gaussian processes with mean zero and covariance Σ. The empirical covariance kernel is $\hat{c}(s,t) = n^{-1} \sum_{i=1}^{n} f_i(s) f_i(t)$, which is a Gaussian polynomial. By the decoupling inequality (De la Pena and Giné, 2012, Theorem 4.2.27), there exists a $\kappa > 0$ such that

$$
\mathrm{P}\left(\|\hat{c}(s,t)\| \geq \mathrm{E}\|\hat{c}\| + r \right) \leq \kappa \, \mathrm{P}\left(\|\tilde{c}(s,t)\| \geq \mathrm{E}\|\hat{c}\| + r/\kappa \right)
$$

where $\tilde{c}(s,t) = n^{-1} \sum_{i=1}^{n} f_i(s) f_i'(t)$ with $f_1', \ldots, f_n'$ independent copies of the original f_i. Thus, our Gaussian polynomial can be thought of as a conditional Gaussian random variable. Now using concentration bounds for norms of Gaussian vectors (Giné and Nickl, 2016, Theorem 2.6.8) twice, an inequality similar to the one in the bounded case is obtained easily.

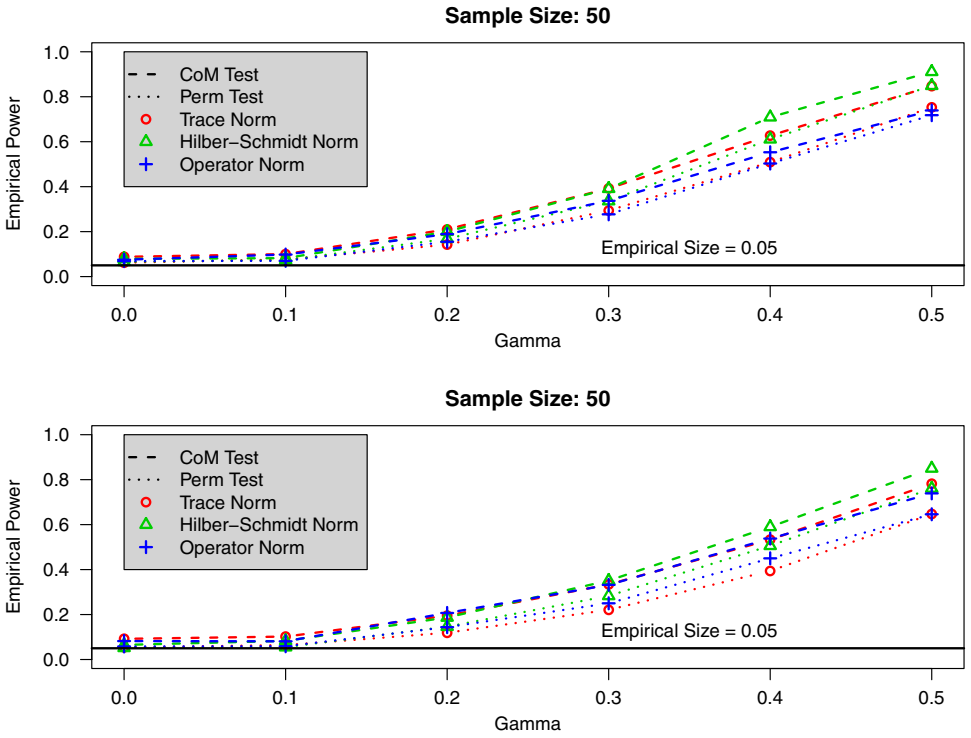

Figure 5: Similar to Figs. 3 and 4 except white noise with variance $c^2 = 100$ is added to each functional data observation. In the top plot, the eigenvalues of Σ decay as $O(k^{-4})$ as in Fig. 3; in the bottom plot, the eigenvalues of Σ decay as $O(k^{-2})$ as in Fig. 4

Defining $f^s = f(s)$, the integral kernel can be written as (Isserlis, 1918)

$$\begin{aligned} \mathrm{E} f^s f^t f^u f^v &= \mathrm{E} f^s f^t \mathrm{E} f^u f^v + \mathrm{E} f^s f^u \mathrm{E} f^t f^v + \mathrm{E} f^s f^v \mathrm{E} f^t f^u \\ &= c_f(s,t)c_f(u,v) + c_f(s,u)c_f(t,v) + c_f(s,v)c_f(t,u). \end{aligned}$$

Hence, we have that $\mathrm{E} f^s f^t f^u f^v - \Sigma_{s,t}\Sigma_{u,v} = \Sigma_{s,u}\Sigma_{t,v} + \Sigma_{s,v}\Sigma_{t,u}$ and that the operator $\mathrm{E} f^{\otimes 4} - \Sigma^{\otimes 2}$, which can be thought of as an Hilbert-Schmidt operator on the space $Op(L^2)$, can be represented by the integral kernel $c_f(s,u)c_f(t,v) + c_f(s,v)c_f(t,u)$. These two terms are merely relabeled versions of $\Sigma^{\otimes 2}$. Consequently, using the subadditivity of the norm, $\left\|\mathrm{E} f^{\otimes 4} - \Sigma^{\otimes 2}\right\|_p \leq \left\|\Sigma^{\otimes 2}\right\|_p + \left\|\Sigma^{\otimes 2}\right\|_p = 2\left\|\Sigma^{\otimes 2}\right\|_p$. For example, for the Hilbert-Schmidt norm,

$$\left\|\mathrm{E} f^{\otimes 4} - \Sigma^{\otimes 2}\right\|_{HS}^2 = \iiiint \{c_f(s,u)c_f(t,v) + c_f(s,v)c_f(t,u)\}^2 \, dsdtdudv$$

$$= 2\|\Sigma\|_{HS}^4 + 2\iiiint c_f(s,u)c_f(s,v)c_f(t,v)c_f(t,u) \times dsdtdudv \leq 4\|\Sigma\|_{HS}^4 .$$

Lemma 5.1 of Horváth and Kokoszka (2012) gives an explicit form of a covariance operator of Σ in terms of the eigenfunctions of Σ for Gaussian data in the Hilbert-Schmidt setting.

Given λ_i, the eigenvalues of Σ, the spectrum of $\Sigma^{\otimes 2}$ is $\{\lambda_i\lambda_j\}_{i,j=1}^{\infty}$. Hence, for any of the p-Schatten norms, $\|\Sigma \otimes \Sigma\|_p = \|\Sigma\|_p^2$. Note that in the above calculations, the weak variance depends on the unknown Σ. In practice, this can be replaced by the empirical estimate $\hat{\Sigma}$.

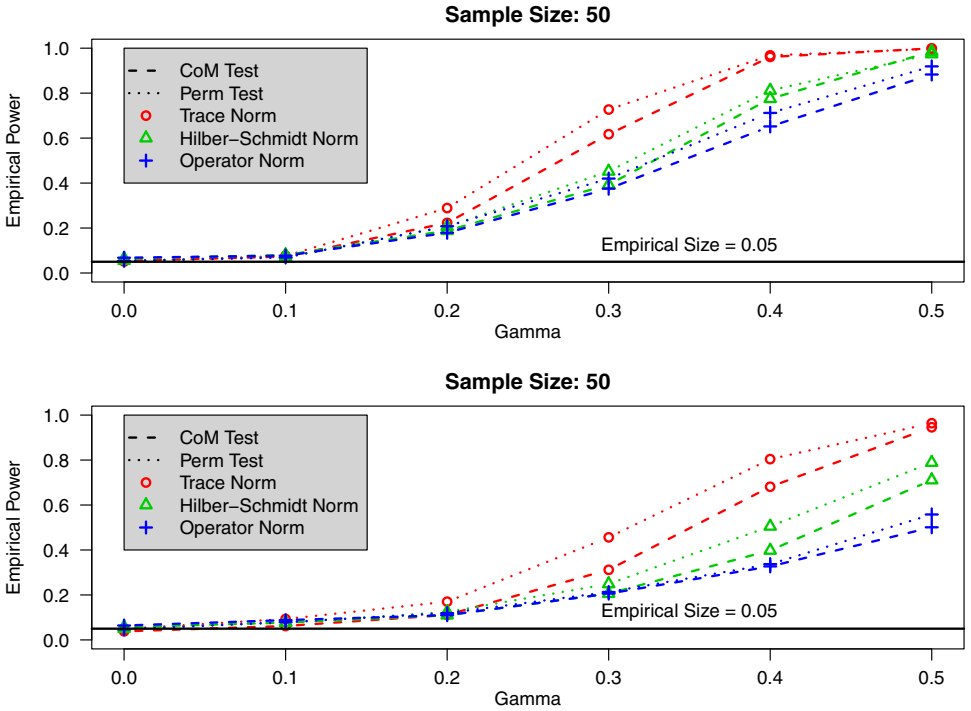

Figure 6: A repetition of the experiments from Figs. 3 and 4 but with data simulated from a multivariate t-distribution with 6 degrees of freedom. In the top plot, the eigenvalues of Σ decay as $O(k^{-4})$ as in Fig. 3; in the bottom plot, the eigenvalues of Σ decay as $O(k^{-2})$ as in Fig. 4

Appendix C: Heavy Tails and Noisy Measurements

As often in practice functional data comes from noisy measurements, consider data of the form $Y_i = X_i + \varepsilon_i$ where X_i is a mean zero Gaussian process with covariance operator Σ and ε_i is Gaussian white noise with covariance $c^2 I$ for some $c^2 > 0$. Figure 5 repeats the previous power analysis for the two sample test but in the moderately noisy settings.

Secondly, heavier tailed data, specifically t-distributed data with 6 degrees of freedom, can also be handled by this method. Figure 6 repeats the earlier two sample power analysis but with the heavier tailed distribution in place of the Gaussian. Here, the coefficient of $(k+2)/(k+3)$ in Eq. 4.1 was replaced with simply 1 in order to achieve the correct empirical size. In general, given arbitrary data, one can simulate null data and adjust the tuning parameters to match the desired empirical size of the test.

Lastly, the empirical coverage of the concentration based confidence set is still comparable to the desired coverage in the heavy tailed case. Consider t-distributed data with six degrees of freedom; Nine operators were randomly generated and data was simulated from each. Figure 7 recreates the simulated confidence sets from Fig. 2, but with the t-distributed data.

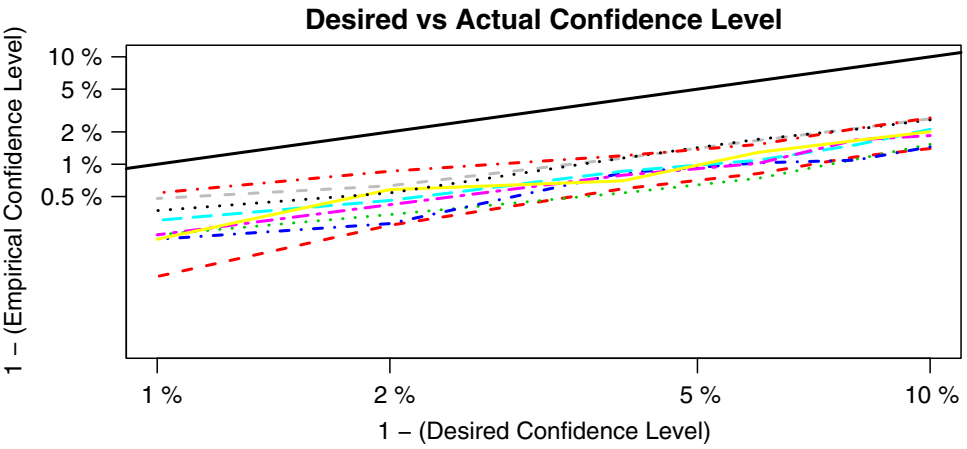

Figure 7: The empirical confidence level of the set from Eq. 3.1 for nine different operators given a sample size of 35 curves generated from a t-distributed process with 6 degrees of freedom. The black line is where the desired and empirical levels are equal. The desired level ranges from $\alpha = 1\%$ to $\alpha = 10\%$. 10,000 replications were used to produce these curves

To achieve these empirical coverages, the Gaussian weak variance, previously calculated to be $\sigma^2 = 2\|\Sigma\|_p^2$, is scaled by a factor of $\nu/(\nu-4)$ where ν is the degrees of freedom.

Adam B. Kashlak
John A. D. Aston
Richard Nickl
Statistical Laboratory,
University of Cambridge
Cambridge, UK
E-mail: j.aston@statslab.cam.ac.uk
E-mail: r.nickl@statslab.cam.ac.uk

Adam B. Kashlak
Mathematical and Statistical
Sciences,
University of Alberta,
Edmonton, Alberta, Canada
E-mail: kashlak@ualberta.ca

Paper received: 25 November 2017.