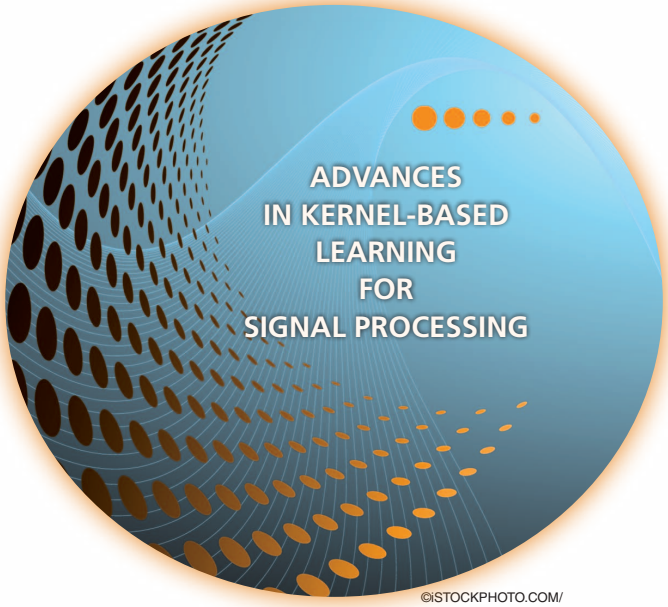
Zaid Harchaoui, Francis Bach, Olivier Cappé, and Éric Moulines

# Kernel-Based Methods for Hypothesis Testing




©iSTOCKPHOTO.COM/
ALEKSANDAR VELASEVIC

## A unified view

Kernel-based methods provide a rich and elegant framework for developing nonparametric detection procedures for signal processing. Several recently proposed procedures can be simply described using basic concepts of reproducing kernel Hilbert space (RKHS) embeddings of probability distributions, mainly mean elements and covariance operators. We propose a unified view of these tools and draw relationships with information divergences between distributions.

### INTRODUCTION AND CONTEXT

Testing hypotheses of signals is one of the key topics in statistical signal processing [1]. Popular examples include testing for equality of signals/homogeneity, as in speaker verification [2]–[4] or change detection [5], [6]. Testing for a change-point in a signal is an important problem that arises in many applications [5]; detecting potential changes can be either the final goal, as in surveillance and monitoring applications, or an intermediate step that is required to allow further processing an interpretation. In multimedia signal processing, unsupervised temporal segmentation can rely on change detection to segment the signal into coherent sections either based on higher-lever semantic concepts, for instance, by detecting cuts in video shot, or based on low-level signal properties, e.g., when a signal is segmented into sections on which it can be considered as stationary.

The most classical approaches for statistical detection of changes are parametric in nature, meaning that they rely on strong assumptions on the distributional properties of the observed signals [5], [7]. Procedures such as the classical CuSum statistic or Hotelling's $T^2$ test assume that the data is (possibly multivariate) Gaussian, whereas the $\chi^2$ and mutual information statistics apply to finite-valued data. These test statistics are widely used due to their simplicity and strong optimality guarantees in scenarios where the underlying distributional assumptions are satisfied. On the other hand, there is also a need for alternative methods, which could possibly be less efficient in some specific scenarios but more robust in the sense of providing reliable results over larger classes of data distributions. These methods are generally known as robust test statistics and usually rely on so-called nonparametric statistical concepts, where the term *nonparametric* refers to the possibility of obtaining performance guarantees that do

*Digital Object Identifier 10.1109/MSP.2013.2253631*
*Date of publication: 12 June 2013*

not depend on an assumed data distribution. For univariate data, rank-based statistics used in the well-known Mann-Whitney/Wilcoxon test [8] are largely recognized as a relevant robust alternative for detecting changes that affect the mean level of a signal.

However, when considering higher-dimensional data, there is no such natural candidate for building robust nonparametric change-detection statistics. Another challenging situation is the case of structured data, meaning data whose mere representation as a vector or as a series of scalar values would result in an important loss of information. Typical examples include graphs, structured text (e.g., hypertext with XML or HTML markup) but also histograms of (possibly redundant) features, which is the dominant paradigm, in particular, in computer vision [9] or natural language processing systems [10]. During the last two decades, kernel-based methods have been popular for supervised classification or regression problems [10]–[12]. Recently, kernel-based methods were designed for hypothesis testing problems, allowing the ability to work with high-dimensional and structured data, as soon as a positive semidefinite similarity measure (the so-called kernel) can be defined [13]–[17].

**KERNEL-BASED METHODS CAN POTENTIALLY BE APPLIED TO ANY KIND OF DATA, RANGING FROM DATA LIVING IN STANDARD EUCLIDEAN SPACES TO DATA CONSISTING OF HISTOGRAMS, CHAINS, TREES, OR GRAPHS.**

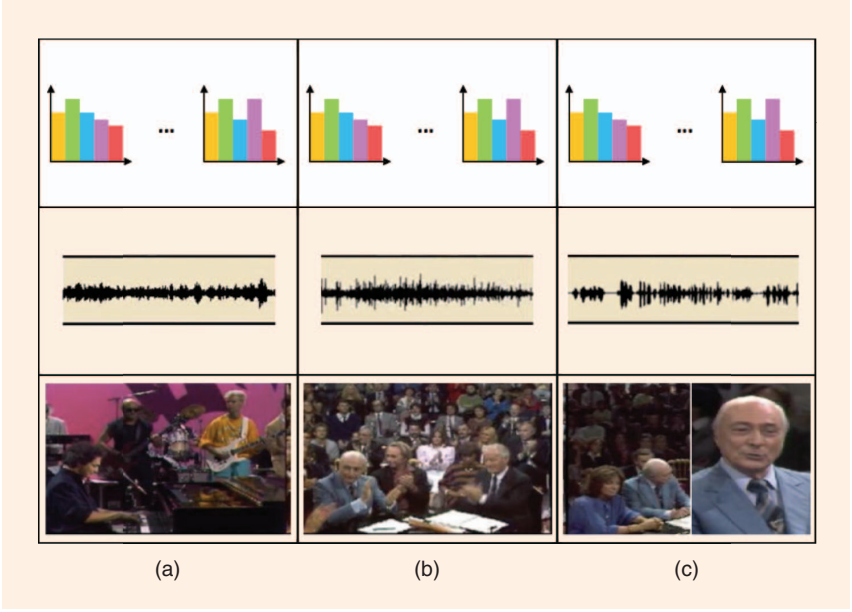
### *TEMPORAL SEGMENTATION OF AUDIOVISUAL CONTENT*

To illustrate the technical part of this tutorial, we first describe the example of temporal segmentation of multimedia signals [18]. Temporal segmentation is an important preliminary task for archiving audio or audiovisual content in databases while allowing for content-based retrieval of the data. Through this example, we would like to highlight the potential of kernel-based methods for change detection.

Temporal segmentation of audiovisual recordings involves two modalities [19]: video and audio. Temporal segmentation of videos is usually synonym of shot segmentation or scene segmentation, that is, detecting abrupt changes in the video content. State-of-the-art approaches use first-order derivatives of the color-histogram signal [20]–[22], leading to high-detection rates for abrupt changes. Such approaches are more difficult to apply when looking for changes in the semantic content of the video. On the other hand, most state-of-the-art approaches for temporal segmentation of audio streams, usually referred to as *audio diarization*, rely on supervised learning methods, and therefore require a significant amount of previously annotated training data [23], [22]. In contrast, kernel-based hypothesis tests for change detection can potentially be applied to a wide range of audiovisual documents without the need to assemble training data.

The data set used in [18] consists of both the audio and video recordings of the popular French 1980s talk shows (*Le Grand Echiquier*) of roughly three hours each. The goal is to blindly perform a temporal segmentation of the corresponding signals into "semantically homogeneous" segments, corresponding to different categories of content, such as "movie," "music," and "interview," among others (see Figure 1). Audio tracks are extracted from MPEG video files, converted to mono, downsampled to 16 kHz. The first 12 Mel-frequency cepstral coefficients (MFCCs), as well as first- and second-order coefficients and the 0th order cepstral coefficient are extracted every 10 ms. A "bag-of-words"-type representation as histograms [9] is then built over windows of size 33, eventually giving a signal of histograms of 128-dimensional audio features. A similar pipeline is adopted for the video track, starting from scale-invariant feature transform (SIFT) features [9], and yielding a signal of 2,176-dimensional feature vectors. Thanks to the preprocessing pipeline, the temporal correlation between the features is negligible and the segmentation task can be tackled by looking for changes in the probability distribution of the histogram features.

A simple approach for temporal segmentation is to reduce it to a sequence of tests for homogeneity between two parts of a sliding window over the signal to perform change detection. Such approaches are attractive because of their scalability, as they scale linearly in the length of the signals, in contrast to retrospective




**[FIG1] Temporal segmentation of TV archives: (a) music, (b) applause, and (c) speech.**

approaches taking the signal as whole and typically scale quadratically in the length of the signals [6], [24]. The main characteristic of the data we consider lies in its high-dimensional and structured nature. Each data-point is a 2,048-dimensional feature vector. Therefore, classical parametric multivariate test statistics cannot be applied [7]. Typically, these methods will have a low detection rate (high Type II error, low power) because there are too few samples to estimate the high-dimensional quantities appearing in the test statistic. For instance, the Hotelling $T^2$ test statistic involves the inverse of an estimate of the covariance matrix, and this inverse will be severely ill conditioned in high-dimensional settings.

A promising nonparametric alternative to parametric approaches is offered by kernel-based methods. In contrast to parametric approaches, kernel-based methods leverage the underlying "smoothness" of the data and rely on a positive semidefinite kernel to measure the similarity between observations possibly living in high-dimensional spaces [11], [10]. Indeed, kernel-based methods hinge upon the Hilbertian structure of the so-called RKHS, the natural function space associated with these nonparametric approaches [25], [26]. These methods work only on dot-products between feature maps of the observations in the RKHS associated with the kernel; in many situations, these dot-products may be computed directly without the need to explicitly compute the high-dimensional feature map, an operation often referred to as the "kernel trick." Hence, kernel-based methods can potentially be applied to any kind of data, ranging from data living in standard Euclidean spaces to data consisting of histograms, chains, trees, or graphs [27], [10]. Although we chose change detection in multimedia signals as our introductory example, kernel-based methods can be applied to a wide range of hypothesis testing problems beyond change detection. Promising results of kernel-based hypothesis tests were obtained in [18]. Yet, the potential of kernel-based approaches for hypothesis testing problems in signal processing remains to be fully explored.

We need a machinery of concepts tailored for kernel-based approaches to explore kernel-based methods for hypothesis testing. We shall see that these concepts are known as mean elements and covariance operators [28], [13], and that most kernel-based test statistics can be expressed in a simple manner using these concepts. But first let us recall the basics of statistical hypothesis testing and detection.

## HYPOTHESIS TESTING AND THE TWO TYPES OF ERROR

In this section, we recall the basic statistical groundwork for designing hypothesis tests (detectors) suitable for various signal processing applications. The approaches follow directly from the theory of hypothesis testing [1], [29]. We start from a simple example: testing for homogeneity in distribution. When data can only take a finite number of values, the problem is equivalent to comparing empirical probability masses, and $\chi^2$-tests based on the $\chi^2$-distance are typically used. Note that the kernel-based tests will extend both the discrete data setting and the continuous Gaussian setting that we present below.

Assume that we observe two samples, that is, two series of data-points $\mathrm{x}_{1,1}, \ldots, \mathrm{x}_{1,n_1}$ and $\mathrm{x}_{2,1}, \ldots, \mathrm{x}_{2,n_2}$ in $\mathbb{R}^d$, whose probability distribution function (PDF) are $\mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, respectively. The notation $\mathcal{N}(\mu_1, \Sigma)$ denotes here a Gaussian PDF with mean $\mu_1$ and covariance matrix $\Sigma_1$. We must therefore determine if $\mu_1 = \mu_2$, that is, if the two samples come from the same PDF ("homogeneity"), or if $\mu_1 \neq \mu_2$. So we have to choose between two competing hypotheses, and face the following decision problem:

$$\text{Decide between } \begin{array}{ll} \mathbf{H}_0: & \mu_1 = \mu_2 \\ \mathbf{H}_A: & \mu_1 \neq \mu_2. \end{array}$$

The hypothesis $\mathbf{H}_0$ is referred to as the *null hypothesis*, and $\mathbf{H}_A$ is the *alternative hypothesis*. The goal of hypothesis testing and detection is to build a statistical decision rule to answer the above problem. The decision rule can make two types of error. If we decide $\mathbf{H}_A$, but $\mathbf{H}_0$ is true, we make a Type I error (false-alarm rate). On the other hand, if we decide $\mathbf{H}_0$, but $\mathbf{H}_A$ is true, we make a Type II error (missed-detection rate):

$$P_{\mathrm{FA}} = \mathbb{P}(\text{decide } \mathbf{H}_A \,|\, \mathbf{H}_0 \text{ is true}) = \alpha = \text{Type I error},$$
$$P_{\mathrm{D}} = \mathbb{P}(\text{decide } \mathbf{H}_A \,|\, \mathbf{H}_A \text{ is true}) = \pi = 1 - \text{Type II error}.$$

Consider a test statistic $T_n$, then the decision rule will rely on a *critical region* $R(\alpha)$, where $\alpha$ is the Type I error. The decision rule writes as

- if $T_n \in R(\alpha)$, decide $\mathbf{H}_0$,
- if $T_n \notin R(\alpha)$, decide $\mathbf{H}_A$.

Clearly, the Type I error can be decreased by enlarging the acceptance region $R(\alpha)$ at the expense of the Type II error. It is not possible to reduce both error probabilities simultaneously. A typical approach, known as the Neyman-Pearson approach, is to hold one probability error fixed, the Type I error, while minimizing the other. In other words, assume that a prescribed false-alarm rate, or Type I error, is given. Then, over all critical regions $R(\alpha) \subset \mathbb{R}$, we chose the one that maximizes the probability of detection (minimizes the Type II error).

On our working problem, the classical test statistic is the so-called Hotelling $T^2$ test statistic [30], [29], defined as

$$T_n = \frac{n_1 n_2}{n_1 + n_2} (\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\Sigma}_W^{-1} (\hat{\mu}_2 - \hat{\mu}_1),$$

where $\hat{\mu}_2$ and $\hat{\mu}_1$ are the empirical mean vectors of the first and second sample, respectively, while $\hat{\Sigma}_W$ is the within-sample pooled covariance matrix

$$\hat{\Sigma}_W = \frac{n_1}{n_1 + n_2} \hat{\Sigma}_1 + \frac{n_2}{n_1 + n_2} \hat{\Sigma}_2.$$

There are two main ways to determine the critical region for a test statistic: 1) large-sample distribution under the null hypothesis and 2) sampling-based approaches such as the bootstrap [29]. For instance, the large-sample distribution of Hotelling's $T^2$ test statistic, that is, the distribution when the number of samples grows to infinity, is the $\chi^2_d$ distribution with $d$ degrees of freedom

$$T_n \xrightarrow{\mathcal{D}} \chi^2_d.$$

Therefore, for any false-alarm rate $\alpha$, the critical region of the Hotelling's $T^2$ test can be determined by computing the $(1-\alpha)$-quantile $c_{1-\alpha}$ of its large-sample distribution, that is, the $(1-\alpha)$-quantile of the $\chi^2_d$ distribution with $d$ degrees of freedom, which are well known and can easily be computed with arbitrary precision with any statistical software.

**ONE CAN USE GEOMETRICAL INTUITION TO BUILD KERNEL-BASED METHODS, BY DRAWING INSPIRATION FROM CLASSICAL MULTIVARIATE STATISTICS METHODS WORKING IN FINITE-DIMENSIONAL EUCLIDEAN SPACES.**

Another take on this is to use resampling techniques such as the bootstrap, which consists of drawing a large number of "pseudosamples," where the two samples are shuffled. Then the critical region is computed by looking at the $(1-\alpha)$-quantile of the empirical cumulative distribution function (CDF) of all these computed values of the test statistic on the pseudosamples [31].

To design kernel-based hypothesis tests, the major challenge is to derive the large-sample distribution under the null hypothesis of the kernel-based test statistic to compute a critical value [8]. Several kernel-based hypothesis algorithms were proposed recently, with successful applications in signal processing. Some of these algorithms were studied as hypothesis testing procedures, and large-sample distributions were derived. Some others were presented using other arguments. We present them here under the same unique framework, using simple concepts known as *mean elements* and *covariance operators* [28], [13]. We show that the underlying structure of the RKHS on which these test statistics rely is simply the eigenvector basis of particular covariance operators, depending on the kernel-based test statistic considered. We also show that, under the alternative hypothesis, these test statistics correspond to consistent estimators of well-known information divergences between probability distributions [32].

## RKHS EMBEDDINGS

We present here the notions of mean element and covariance operators in RKHS. Let us start by recalling the main ideas of kernel-based approaches and RKHSs.

Consider a set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, say, for instance, visual features of a series of images. The data points live in an input space, which is a subspace of $\mathbb{R}^d$ but has some structure.

Kernel-based methods work as follows. As soon as one can define a dot-product $k(\mathbf{x}, \mathbf{x}')$ between two data points $\mathbf{x}$ and $\mathbf{x}'$, which can be interpreted as a similarity measure between $\mathbf{x}$ and $\mathbf{x}'$, one can devise a whole spectrum of statistical methods, kernel-based methods, working directly on the dot-products $k(\mathbf{x}, \mathbf{x}')$ instead of the raw data points $\mathbf{x}$ and $\mathbf{x}'$. The only requirement is to be given as an input a symmetric matrix $\mathbf{K} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{1 \le i,j \le n}$, called the Gram matrix or kernel matrix, which should be positive semidefinite. Popular examples of kernel-based methods are kernel principal component analysis (KPCA) [33], kernel ridge regression (KRR) [10], and support vector machines (SVMs) [11].

**[TABLE 1] EXAMPLES OF KERNELS.**

| KERNEL | EXPRESSION |
|---|---|
| LINEAR | $k(x, y) = x^T y$ |
| GAUSSIAN | $k(x, y) = \exp(-\|x-y\|^2/\sigma^2)$ |
| $\chi^2$-KERNEL | $k(x, y) = \exp\left(-\frac{1}{\sigma^2}\sum_{\ell=1}^{d}\frac{(p_\ell - q_\ell)^2}{p_\ell + q_\ell}\right)$ |

The requirement on the Gram matrix is satisfied as soon as the kernel $k(\cdot, \cdot)$ is symmetric, i.e., $k(x, y) = k(y, x)$ for all $x, y \in \mathcal{X}$, and positive semidefinite, that is,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j k(x_i, x_j) \ge 0,$$

for all $m \in \mathbb{N}^*$, for all $x_1, \ldots, x_m \in \mathcal{X}$, and for all $c_1, \ldots, c_m \in \mathbb{R}$.

The simplest kernel is the linear kernel, defined for all $x, y \in \mathcal{X}$ by $k(x, y) = x^T y$. It turns out that a positive semidefinite (psd) kernel can always be interpreted as dot-product in a Hilbert space $\mathcal{H}$, (the RKHS). Thus, for any kernel $k(\cdot, \cdot)$ acting on $\mathcal{X}$, there exists a feature map $[\phi : \mathcal{X} \to \mathcal{H}]$ such that, for all $x, y \in \mathcal{X}$

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}.$$

Here, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the dot-product defined in the RKHS, which may be an infinite-dimensional Hilbert space. This remarkable property has important consequences. Indeed, one can use "geometrical" intuition to build kernel-based methods, by drawing inspiration from classical statistical methods working in finite-dimensional Euclidean spaces.

Starting from a classical multivariate-statistics method, hinging upon computations that can be written as dot-products $\langle x, y \rangle_{\mathbb{R}^p} = x^T y$, one can immediately design its kernel-based counterpart by replacing all $\langle x, y \rangle_{\mathbb{R}^p} = x^T y$ by $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = k(x, y)$. However, this is just an heuristic to derive new approaches, and a sound interpretation has to be developed to check whether the "kernelized" counterpart of the classical multivariate method is actually meaningful. Table 1 summarizes some usual kernels. For measuring similarity of histograms, the so-called $\chi^2$-kernel applies the Gaussian kernel on top of the $\chi^2$-divergence between histograms, a well-known divergence between histograms that is more sensitive to differences in the tails than the $L_2$-divergence.

## MEAN ELEMENT AND COVARIANCE OPERATORS

A natural question is how a probability distribution $\mathbb{P}$ is represented in an RKHS $\mathcal{H}$ [34]. We show now that infinite-dimensional counterparts of two fundamental multivariate statistics concepts, mainly the mean vector and the covariance matrix, are particularly appropriate for this purpose. These

RKHS-counterparts of the mean vector and the covariance matrix are called the *mean element* and the *covariance operator,* respectively [35], [28], [13]; see Figure 2. The different names emphasize that RKHSs might be infinite-dimensional, and that one should be careful not to make hasty conclusions by simply translating finite-dimensional intuitions to RKHSs.

Consider a random variable $X$ taking values in $\mathcal{X}$ and a probability distribution $\mathbb{P}$. The mean element $\mu_{\mathbb{P}}$ associated with $X$ is the unique element of the RKHS $\mathcal{H}$, such that, for all $f \in \mathcal{H}$,

$$\langle \mu_{\mathbb{P}}, f \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[f(X)].$$

Similarly, the covariance operator $\Sigma_{\mathbb{P}}$ associated with $X$ is the only operator $[\Sigma_{\mathbb{P}} : \mathcal{H} \otimes \mathcal{H} \to \mathbb{R}]$ such that, for all $f, g \in \mathcal{H}$,

$$\begin{aligned}\langle f, \Sigma_{\mathbb{P}} g \rangle_{\mathcal{H}} &= \mathrm{Cov}_{\mathbb{P}}(f(X), g(X)) \\ &= \mathbb{E}_{\mathbb{P}}[f(X)g(X)] - \langle \mu_{\mathbb{P}}, f \rangle_{\mathcal{H}} \langle \mu_{\mathbb{P}}, g \rangle_{\mathcal{H}}.\end{aligned}$$

Note that so far we have only focused on population quantities. Their empirical counterparts are as expected. Consider a sample $X_1, \ldots, X_n$ drawn independent and identically distributed (i.i.d.) from $\mathbb{P}$. Then one can estimate $\mu_{\mathbb{P}}$ by the empirical mean element, defined as the unique element in the RKHS $\mathcal{H}$, such that, for all $f \in \mathcal{H}$,

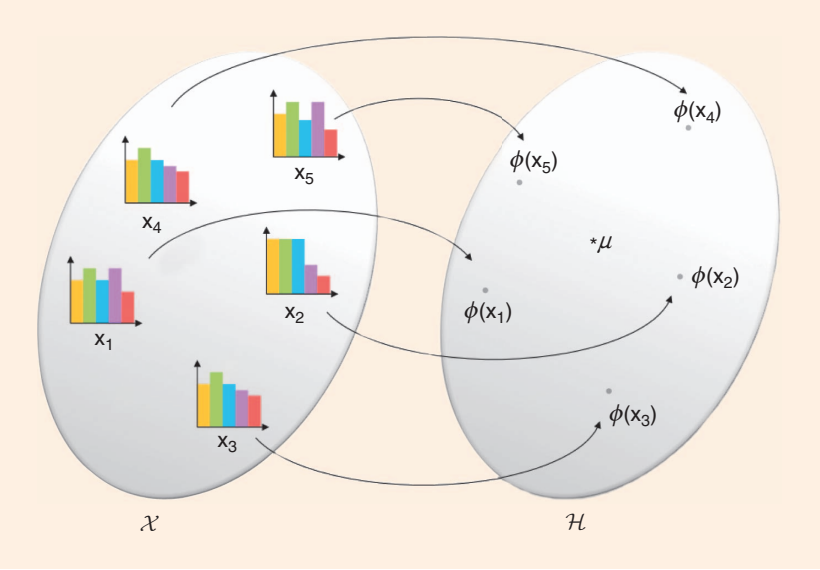
$$\langle \hat{\mu}, f \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Figure 3 provides a simple illustration of the concept of "mean element" in one dimension. Consider, say, five realizations of a uniform random variable $X$ in $\mathcal{X} = [-5, 5]$ and a Gaussian radial-basis function (RBF) kernel $k(x, y) = \exp(-2(x-y)^2)$. Then each point $x$ is embedded into a "Gaussian bump" $k(x, \cdot)$ in the RKHS, plotted with different colors for each data point in Figure 3. Then the empirical mean element is a function, corresponding to the black curve, aggregating all the Gaussian bumps corresponding to the five data points.

Similarly, the covariance operator is estimated by $\hat{\Sigma}$, the empirical covariance operator, defined as the unique operator, such that, for all $f, g \in \mathcal{H}$

$$\langle f, \hat{\Sigma} g \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \langle \hat{\mu}, f \rangle_{\mathcal{H}})(g(X_i) - \langle \hat{\mu}, g \rangle_{\mathcal{H}}).$$

This is where one should refrain from drawing hasty conclusions by relying on multivariate Gaussian intuitions. In the classical Gaussian multivariate case where $\mathcal{X} = \mathbb{R}^d$, the mean vector and the covariance are necessary and sufficient to characterize the probability distribution. In the infinite-dimensional case, that is, when $\dim(\mathcal{H}) = \infty$, interesting phenomena arise. In particular, the mean element




**[FIG2] A schematic view of kernel embedding and mean element.**

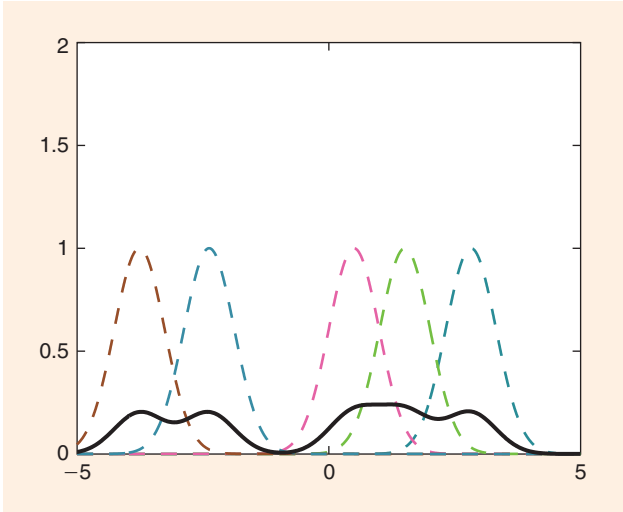
plays a more central role. Let us now look at the function $m(\cdot)$ mapping any probability distribution $\mathbb{P}$ to its corresponding mean element $\mu_{\mathbb{P}}$. Note that $m(\cdot)$ depends on the kernel $k(\cdot, \cdot)$ associated with $\mathcal{H}$. For a large class of kernels, the function $m$ is injective. Consider two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$. If for all $f \in \mathcal{H}$

$$\langle \mu_{\mathbb{P}}, f \rangle_{\mathcal{H}} = \langle \mu_{\mathbb{Q}}, f \rangle_{\mathcal{H}},$$

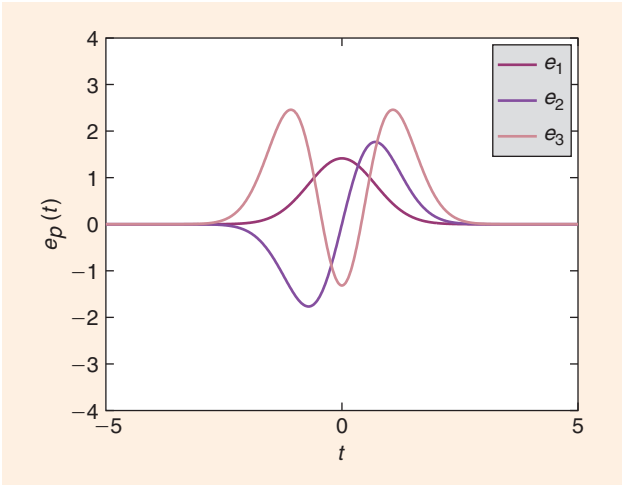
then

$$\mathbb{P} \equiv \mathbb{Q},$$

as soon as the RKHS $\mathcal{H}$ associated with the kernel $k(\cdot, \cdot)$ is dense in $L^2(\mathbb{P})$ for all probability distributions $\mathbb{P}$ [36], [37], [13],




**[FIG3] Kernel embeddings $k(x_1, \cdot), \ldots, k(x_5, \cdot)$, respectively, of five one-dimensional data points $x_1, \ldots, x_5$ (colored Gaussian bumps), and the corresponding empirical mean element $\hat{\mu} = 1/5 \sum_{i=1}^{5} k(x_i, \cdot)$ (black curve) with a Gaussian RBF kernel.**



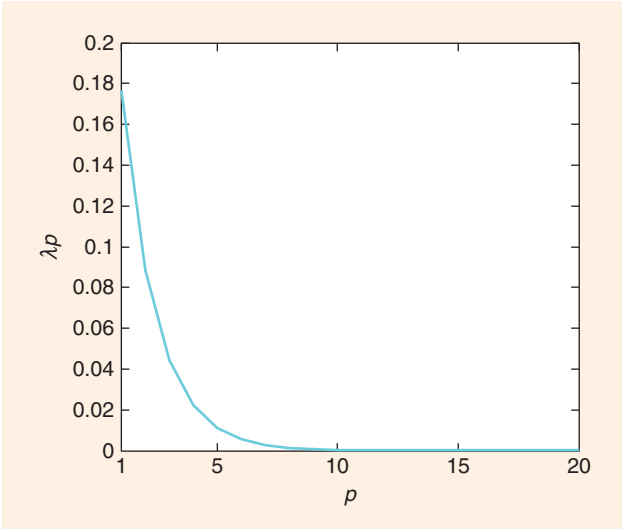

**[FIG4] The first three eigenfunctions $e_1(\cdot), e_2(\cdot)$, and $e_3(\cdot)$ of the covariance operator corresponding to a marginal probability density function $p(x) = 1/\sqrt{2\pi}\exp(-x^2/2)$ and a Gaussian RBF kernel $k(x, y) = \exp(-2(x-y)^2)$.**

[38], [39]. It is worthwhile to note that equality of covariance operators $\Sigma_{\mathbb{P}} = \Sigma_{\mathbb{Q}}$ is implied by $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$, in contrast to the Gaussian multivariate case.

Another interesting fact is related to the spectrum of the covariance operator $\Sigma_{\mathbb{P}}$. The covariance operator is self-adjoint, positive, and for any complete orthonormal basis $\{\psi_p\}_{p\geq 1}$ of $\mathcal{H}$, the sum $\sum_{p=1}^{\infty}\langle \psi_p, \Sigma_{\mathbb{P}}\psi_p\rangle_{\mathcal{H}}$ is finite and independent of the basis $\{\psi_p\}_{p\geq 1}$ of $\mathcal{H}$. The trace of $\Sigma_{\mathbb{P}}$ is then defined as

$$\mathrm{Tr}(\Sigma_{\mathbb{P}}) = \sum_{p=1}^{\infty}\langle e_p, \Sigma_{\mathbb{P}} e_p\rangle_{\mathcal{H}}.$$

The covariance operator is also Hilbert-Schmidt, that is $\sum_{p=1}^{\infty}\lambda_p^2 < \infty$, where $\{\lambda_p\}_{p\geq 1}$ is the (infinite) sequence of




**[FIG5] The spectrum of the covariance operator corresponding to a marginal probability density function $p(x) = 1/\sqrt{2\pi}\exp(-x^2/2)$ and a Gaussian RBF kernel $k(x, y) = \exp(-2(x-y)^2)$.**

eigenvalues of $\Sigma_{\mathbb{P}}$. It is more convenient to work in the Hilbertian basis $\{e_p\}_{p\geq 1}$ of the eigenfunctions of $\Sigma_{\mathbb{P}}$

$$\Sigma_{\mathbb{P}} = \sum_{p=1}^{\infty}\lambda_p (e_p \otimes e_p).$$

Examples of eigenfunctions are depicted in Figure 4. Consider a normal random variable $X$ in $\mathcal{X} = \mathbb{R}$, and a Gaussian RBF kernel $k(x, y) = \exp(-2(x-y)^2)$. Then, the eigenfunctions of the corresponding covariance operator can be expressed analytically using Hermite polynomials [40]. The first three eigenfunctions are illustrated in Figure 4. The sequence of eigenvalues $\{\lambda_p\}_{p\geq 1}$ can also be expressed analytically, and its decay is polynomial; see Figure 5.

The fact that $\Sigma_{\mathbb{P}}$ is of bounded trace-norm corresponds to $\sum_{p=1}^{\infty}\lambda_p < \infty$. In other words, there is no such thing as an isotropic probability distribution in $\mathcal{H}$ when $\dim(\mathcal{H}) = \infty$. In particular, $\Sigma_{\mathbb{P}} = \mathrm{I}_{\mathcal{H}}$ is impossible in infinite-dimensional RKHSs. Probability distributions in $\mathcal{H}$ are indeed highly "anisotropic," that is, the sequence of eigenvalues is decreasing $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq \cdots$ and must satisfy $\sum_{p=1}^{\infty}\lambda_p < \infty$.

Even though the mean-element map is injective, the covariance operator is still valuable to normalize test statistics. To illustrate this, let us consider a sample $\{X_1, \ldots, X_n\}$ drawn i.i.d. from $\mathbb{P}$ on $\mathcal{X}$, with $(\mu, \Sigma)$ as the pair of mean element and covariance operator, and $(\hat{\mu}, \hat{\Sigma})$ their empirical counterparts. Denote $\{\lambda_p\}_{p\geq 1}$ and $\{e_p\}_{p\geq 1}$ the sequence of eigenvalues and eigenfunctions, respectively, of the covariance operator $\Sigma$. Let us look at

$$\langle \hat{\mu} - \mu, e_p\rangle_{\mathcal{H}}, \quad p = 1, \ldots, \infty.$$

Simple calculations reveal that for $p = 1, \ldots, \infty$

$$\mathbb{E}[\langle \hat{\mu} - \mu, e_p\rangle_{\mathcal{H}}] = 0$$
$$\mathrm{Var}[\langle \hat{\mu} - \mu, e_p\rangle_{\mathcal{H}}] = \lambda_p.$$

Therefore, valuable information is carried in the spectrum of the covariance operator $\Sigma$, encoded in the eigenvalues $\{\lambda_p\}_{p\geq 1}$ and the eigenfunctions $\{e_p\}_{p\geq 1}$. To build a well-normalized test statistic based on these quantities, one needs to know the variance of $\langle \hat{\mu} - \mu, e_p\rangle_{\mathcal{H}}$. This variance is actually given by the eigenvalue $\mathrm{Var}[\langle \hat{\mu} - \mu, e_p\rangle_{\mathcal{H}}] = \lambda_p$. If we want to weigh in a fair manner the different test statistics, say, the quantities $\langle \hat{\mu} - \mu, e_p\rangle_{\mathcal{H}}$ and $\langle \hat{\mu} - \mu, e_q\rangle_{\mathcal{H}}$ along the eigenfunctions $e_p$ and $e_q$ with $p \neq q$, it is essential to rely on the corresponding variances $\lambda_p$ and $\lambda_q$. Thus, the spectrum of covariance operators is crucial to design normalized test statistics.

Equipped with this arsenal of tools, we now present kernel-based test statistics for detection problems arising in signal processing.

## KERNEL-BASED HYPOTHESIS TESTING

We start by focusing on our working example: testing for homogeneity. We present several kernel-based test statistics, and relate them to information divergence functionals. Consider two independent samples $\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n_1}$ and $\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,n_2}$ drawn respectively from probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$.

We shall denote by $p_1$ and $p_2$ the corresponding probability densities. Testing the homogeneity of the two samples $\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n_1}$ and $\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,n_2}$ corresponds to deciding between the two hypotheses:

$$\mathbf{H}_0: \ \mathbb{P}_1 = \mathbb{P}_2$$
$$\mathbf{H}_A: \ \mathbb{P}_1 \neq \mathbb{P}_2.$$

A first test statistic, called the kernel Fisher discriminant analysis (KFDA) test statistic [13], is inspired by the Hotelling $T^2$ test statistic that we discussed previously. The test statistic is related to the kernel-based methods called KFDA for binary classification [41], [11]. The test statistic writes as

$$T_{n_1,n_2} = \frac{\frac{n_1 n_2}{n_1+n_2} \left\langle \hat{\mu}_2 - \hat{\mu}_1, (\hat{\Sigma}_W + \gamma \mathrm{I})^{-1} (\hat{\mu}_2 - \hat{\mu}_1) \right\rangle_{\mathcal{H}} - d_1}{\sqrt{2}\, d_2},$$

where $(\hat{\mu}_1, \hat{\Sigma}_1)$ and $(\hat{\mu}_2, \hat{\Sigma}_2)$ et $\hat{\Sigma}_W$ are the empirical covariance operators, and $\gamma$ is a positive regularization parameter. The quantities $d_1 = d_{1,n_1,n_2;\gamma}(\hat{\Sigma}_W)$ and $d_2 = d_{2,n_1,n_2;\gamma}(\hat{\Sigma}_W)$ are normalization factors

$$d_{1,n_1,n_2;\gamma}(\hat{\Sigma}_W) := \mathrm{Tr}((\hat{\Sigma}_W + \gamma)^{-1} \hat{\Sigma}_W),$$
$$d_{2,n_1,n_2;\gamma}(\hat{\Sigma}_W) := [\mathrm{Tr}((\hat{\Sigma}_W + \gamma)^{-2} \hat{\Sigma}_W^2)]^{1/2}.$$

All quantities involved can be easily computed using the kernel trick; see [13] for details.

The KFDA test statistic can be calibrated using its large-sample distribution under the null hypothesis [13], [42]. There are two main asymptotic settings to study its large-sample distribution under the null: 1) $\gamma$ is held fixed as the sample size goes to infinity and 2) $\gamma \to 0$ as the sample size goes to infinity. For the sake of conciseness, we shall only focus on the setting 2), and refer to [42] for a thorough discussion.

**THERE ARE USUALLY TWO WAYS TO LOOK AT KERNEL-BASED TEST STATISTICS: A HEURISTIC APPROACH INSPIRED BY MULTIVARIATE STATISTICS, AND A SOUND ONE THAT RELATES IT TO A NONPARAMETRIC ESTIMATE OF AN INFORMATION DIVERGENCE.**

Under the null hypothesis, with mild conditions on the kernel and the spectrum of the covariance operator $\Sigma_W$, typically that the kernel is bounded and that $\sum_{p=1}^{\infty} \lambda_p^{1/2}(\Sigma_W) < \infty$, and assuming that

$$\gamma + d_2^{-1} d_1 \gamma n^{-1/2} \to 0,$$

we have

$$T_{n_1,n_2} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

as $n_1, n_2 \to \infty$. This result also allows us to understand the role of the normalization constants $d_1$ and $d_2$. Thanks to these normalization constants, the test statistic is well normalized so that its large-sample distribution under the null does not depend on the probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$.

Under the alternative hypothesis, the KFDA test statistic can also be shown to be related to a nonparametric version of the classical $\chi^2$ test-statistic for testing homogeneity of discrete distributions [13], [42]. The population version of $\| (\hat{\Sigma}_W + \gamma \mathrm{I})^{-1} (\hat{\mu}_2 - \hat{\mu}_1) \|_{\mathcal{H}}^2$, that is, the key quantity appearing in the KFDA test statistic, actually coincides with the $\chi^2$-divergence between $\mathbb{P}_1$ and $\mathbb{P}_2$.

Indeed, under mild assumptions, assuming that $\mathbb{P}_1$ and $\mathbb{P}_2$ are nonsingular, defining $\rho = n_1/(n_1+n_2)$, we have

$$\| \Sigma_W^{-1} (\mu_2 - \mu_1) \|_{\mathcal{H}}^2 = \frac{D_{\chi^2}(\mathbb{P}_1 \| \mathbb{P}_2)}{1 - \rho^2 D_{\chi^2}(\mathbb{P}_1 \| \mathbb{P}_2)},$$

where

$$D_{\chi^2}(\mathbb{P}_1 \| \mathbb{P}_2) = \int \frac{(p_2 - p_1)^2}{\rho p_1 + (1-\rho) p_2}.$$

As we emphasized before, there are usually two ways to look at kernel-based test statistics, a heuristic approach that inspired its design, and a sound approach that relates it to a nonparametric estimate of an information divergence. The heuristic way to look at the KFDA is to view it as a kernelized version of Hotelling's $T^2$ test statistic. The more sound way is to realize that the KFDA test statistic is actually related to a kernel-based nonparametric estimate of the $\chi^2$-divergence between the two probability distributions. Note that when the data are vectors and the kernel is linear, the test statistics reduces to the Hotelling's $T^2$ test statistic, while when the data take finitely many values, the test statistics is strongly related to the $\chi^2$ test statistic, which is the method of choice in this situation.

Following up along the same lines, one can consider other kernel-based test statistics for testing homogeneity, corresponding to different information divergences [32]. For instance, the so-called maximum mean discrepancy (MMD) test statistic [43], [44] writes as

$$T_{n_1,n_2}^{\mathrm{MMD}} = (n_1+n_2) \| \hat{\mu}_2 - \hat{\mu}_1 \|_{\mathcal{H}}^2.$$

The MMD can easily be computed using the kernel trick; see [43] and [44] for details. Note that it is clearly related to the KFDA test statistic, except for the normalization by the inverse of the covariance operator. We will see that the eigenvalues of the covariance operator will appear in the limiting distribution under the null because of the absence of normalization in the test statistic.

The MMD test statistic can be calibrated using its large-sample distribution under the null hypothesis. Under some mild conditions, such as $n_1, n_2 \to \infty$, one can prove that under the null hypothesis

$$T_{n_1,n_2}^{\mathrm{MMD}} \xrightarrow{\mathcal{D}} \sum_{p=1}^{\infty} \lambda_p \left[ \left( \frac{Y_p}{\sqrt{\rho}} - \frac{Z_p}{\sqrt{1-\rho}} \right)^2 - \frac{1}{\rho(1-\rho)} \right],$$

where the $\{\lambda_p\}_{p \geq 1}$ is the sequence of eigenvalues of $\Sigma_W$, $Y_1, \ldots, Y_p$ are independent normally distributed random

**[TABLE 2] THE RELATIONSHIPS BETWEEN KERNEL-BASED TEST STATISTICS AND INFORMATION DIVERGENCES.**

| TEST STATISTIC | INFORMATION DIVERGENCE |
|---|---|
| KFDA [13] | $D_{\chi^2}(\mathbb{P}_1 \,\|\, \mathbb{P}_2) = \int \frac{(p_2 - p_1)^2}{\rho p_1 + (1-\rho) p_2}$ |
| MMD [43] | $D_{L_2}(\mathbb{P}_1 \| \mathbb{P}_2) = \int (p_2 - p_1)^2$ |
| KDR [48] | $D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2) = \int p_1 \log\left(\frac{p_1}{p_2}\right)$ |

variables, and the same for $Z_1, \ldots, Z_p$. In other words, the MMD test statistic converges to an infinite linear combination of $\chi^2_1$ with one degree of freedom, with weights corresponding to the eigenvalues of the unknown covariance operator $\Sigma_W$.

The null distribution of the MMD test statistics can be made independent of the kernel when the kernel parameter (say, the bandwidth parameter in the Gaussian RBF kernel) shrinks to zero as the sample size goes to infinity, or by using a variant called the linear-time MMD; see [45] and [15], respectively, for more details.

Under the alternative hypothesis, the MMD test statistic can also be shown to be a nonparametric version of the $L^2$ test-statistic for testing homogeneity [45]. Consider the "square-root" convolution kernel $\kappa(\cdot, \cdot)$ (i.e., a kernel used for density estimation [46]) and a kernel $k(\cdot, \cdot)$ defined as $k(x, y) = \int \kappa(x, z)\kappa(y, z)dz$. Assuming that $\kappa(\cdot, \cdot)$ has a bandwidth parameter $h_n$ where $2n = n_1 + n_2$, then the key quantity $\| \hat{\mu}_2 - \hat{\mu}_1 \|^2_{\mathcal{H}}$ in the MMD test statistic is an empirical estimator of

$$D_{L_2}(\mathbb{P}_1 \| \mathbb{P}_2) = \int (p_2 - p_1)^2.$$

With appropriate renormalization, convergence in distribution can be established [45], [15], assuming that $h_n^d \to 0$ and $nh_n^d \to \infty$. It is worthwhile to note that the $L^2$-divergence test statistic, in contrast to the $\chi^2$-divergence one, is less sensitive to differences in the tails. Several interpretations of the MMD test statistic are discussed and reviewed in [38], [15], and [47], respectively.

Another test statistic is the kernel density-ratio test statistic (KDR) [48], related to a nonparametric estimate of the $f$-divergence between probability distributions (see also [49] for related work). Examples of $f$-divergences include, for instance, the Kullback-divergence. The KDR test statistic [48], [50], [16] relies on an estimator $\hat{\nu}_n$ of the density ratio $p_1/p_2$, found by minimizing a convex optimization objective [48]. The estimator $\hat{\nu}_n$ writes as

$$\hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i \phi(x_i),$$

where $\theta_1, \ldots, \theta_n \geq 0$ and $\{\mathbf{x}_i\}_{i=1,\ldots,n_1+n_2}$ denotes the pooled sample $\{\mathbf{x}_{1,i}\}_{i=1\ldots,n_1} \cup \{\mathbf{x}_{2,i}\}_{i=1\ldots,n_2}$. Then the KDR test writes as

$$T_n^{\mathrm{KDR}} = \frac{1}{n} \sum_{i=1}^{n} \log(\langle \hat{\nu}_n, \phi(x_i) \rangle_{\mathcal{H}}).$$

Details on computation of the KDR statistic are given in [48] and [16]. Under mild assumptions, such as $n \to \infty$, the KDR test statistic converges to the Kullback-divergence

$$T_n^{\mathrm{KDR}} \to D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2),$$

where

$$D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2) = \int p_1 \log\left(\frac{p_1}{p_2}\right).$$

All in all, several kernel-based test statistics can be understood as nonparametric estimates of well-known information divergences, such as the $\chi^2$-divergence, the $L_2$-divergence, or the Kullback-Leibler divergence, as summarized in Table 2.

Kernel-based test statistics related to the above information divergences were proposed, such as the KCD test statistic from [51]–[53]. The test statistic is in fact close to MMD and could be related to

$$T_{n_1,n_2} = \| \hat{\nu}_2 - \hat{\nu}_1 \|^2_{\mathcal{H}},$$

where $\hat{\nu}_1$ and $\hat{\nu}_2$ are, respectively, trimmed empirical mean elements

$$\hat{\nu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \alpha_{1,i} \phi(x_{1,i})$$

$$\hat{\nu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \alpha_{2,i} \phi(x_{2,i}),$$

with the weights $\{\alpha_{1,i}\}_{i=1}^{n_1}$ and $\{\alpha_{1,i}\}_{i=1}^{n_2}$ learned by training one-class SVMs on each sample independently; see [18] for more details. A statistical interpretation of one-class SVMs is given in [53] and [54].

Other kernel-based estimates of information divergences were proposed, which we do not cover here; see, e.g., [36] and [55].

## TESTS OF INDEPENDENCE

Kernel-based methods can also be used to design test statistics for testing independence. We only quickly review these methods, as the principles underlying the test are similar.

Consider two samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$ in two different measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, drawn i.i.d. under a joint probability distributions $\mathbb{P}_{X,Y}$. We shall denote by $p_{x,y}$ the corresponding probability density, and by $p_x$, $p_y$ the corresponding marginal probability densities. Testing the independence of the two samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$ corresponds to deciding between the two hypotheses

$$\mathbf{H}_0 : \ \mathbb{P}_X \perp \mathbb{P}_Y$$
$$\mathbf{H}_A : \ \mathbb{P}_X \not\perp \mathbb{P}_Y,$$

where $\mathbb{P}_X \perp \mathbb{P}_Y$ means that the random variables $X$ and $Y$ are independent of each other. Note that testing for homogeneity may be seen as testing for the independence of the observed variable and a binary-valued variable indicating from which sample it came.

Testing for independence may be naturally cast in the covariance operators framework by considering cross-covariance

operators. Given an RKHS $\mathcal{H}_\mathcal{X}$ on $\mathcal{X}$ and an RKHS $\mathcal{H}_\mathcal{Y}$ on $\mathcal{Y}$, then the cross-covariance operator $\Sigma_\mathbb{P}^{X,Y}: \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{X}$ is the only operator so that, for all $(f, g) \in \mathcal{H}_\mathcal{X} \times \mathcal{H}_\mathcal{Y}$,

$$\langle f, \Sigma_\mathbb{P}^{X,Y} g \rangle_{\mathcal{H}_\mathcal{X}} = \mathrm{Cov}_\mathbb{P}(f(X), g(Y)).$$

For infinite-dimensional kernels where the mean element function is injective, then the variables $X$ and $Y$ are independent if and only if $\Sigma_\mathbb{P}^{X,Y} = 0$. It is thus natural to define test statistics from the empirical estimate $\hat{\Sigma}_\mathbb{P}^{X,Y}$ of $\Sigma_\mathbb{P}^{X,Y}$ obtained by considering empirical covariances. This leads naturally to the Hilbert Schmidt independence criterion (HSIC) [17], [56], which is the squared Hilbert-Schmidt norm of $\hat{\Sigma}_\mathbb{P}^{X,Y}$. This criterion corresponds to comparing distributions through $D_{L_2}(p_x p_y \| p_{x,y})$ [57]. Alternatively, like for homogeneity testing, to consider tests that reduce to traditional tests in the Gaussian and discrete cases, "studentization" may be performed by considering the regularized cross-correlation operators

$$(\Sigma_\mathbb{P}^{X,X} + \gamma \mathrm{I})^{-1/2} \Sigma_\mathbb{P}^{X,Y} (\Sigma_\mathbb{P}^{Y,Y} + \gamma \mathrm{I})^{-1/2}$$

and its empirical counterpart. The largest singular value is the largest kernel canonical correlation [36], [58], while its Hilbert-Schmidt norm leads to comparing distributions with $D_{\chi^2}(p_x p_y \| p_{x,y})$ [59]. Such hypothesis tests for testing independence are useful for performing independent component analysis (ICA), with applications such as source separation [36], [60]. Indeed, most algorithms for ICA optimize a non-convex objective, and therefore require multiple restarts for optimization. Kernel independence tests are valuable for checking the quality of obtained solutions with different restarts [36], [60].

**KERNEL-BASED METHODS OFFER AN ELEGANT SET OF TOOLS TO TACKLE THE CHALLENGES ARISING IN NEW APPLICATIONS OF SIGNAL PROCESSING.**

### CALIBRATING KERNEL-BASED TESTS

We now briefly explain more precisely how to calibrate kernel-based test statistics, that is, how to compute the critical region for a prescribed level $\alpha$.

The first approach is to calibrate the test statistic using the limiting distribution under the null hypothesis. Consider a test statistic $T_n$, whose large-sample distribution under the null hypothesis is a random variable $V$. Then one can compute a critical value $c_{1-\alpha}$ that guarantees asymptotically a Type I error (false-alarm rate) of $\alpha$ by computing $c_{1-\alpha}$ such that

$$P(V > c_{1-\alpha} | \mathrm{H}_0 \text{ is true}) = \alpha.$$

Sometimes the limiting random variable $V$, the random variable to which the test statistic converges to under the null, depends on some unknown quantities. For instance, the large-sample distribution of MMD under the null depends on the eigenvalues $\{\lambda_p\}_{p \geq 1}$ of the covariance operator. Then, one can usually replace the unknown eigenvalues $\{\lambda_p(\Sigma_W)\}_{p \geq 1}$ by their statistically consistent estimates $\{\lambda_p(\hat{\Sigma}_W)\}_{p \geq 1}$ [44], and compute instead $c_{1-\alpha}$ such that

$$P(V(\{\lambda_p(\hat{\Sigma}_w)\}_{p \geq 1}) > c_{1-\alpha} | \mathrm{H}_0 \text{ is true}) = \alpha.$$

Other approaches approximate the distribution under the null distribution by moment-matching of a parametrized family of distributions, say, based on the first four moments, and compute the critical value from this approximate null distribution [44]. Such approaches usually yield good results in practice but lack statistical guarantees. They ignore higher-order moments and they therefore do not lead to statistically consistent procedures.

The second approach is to calibrate the test statistic using resampling techniques. Consider the case of testing for homogeneity. Assume that the two samples have same size for the sake of clarity. The test statistic is in fact a function of the two samples

$$T_n = g(\mathcal{S}_n),$$

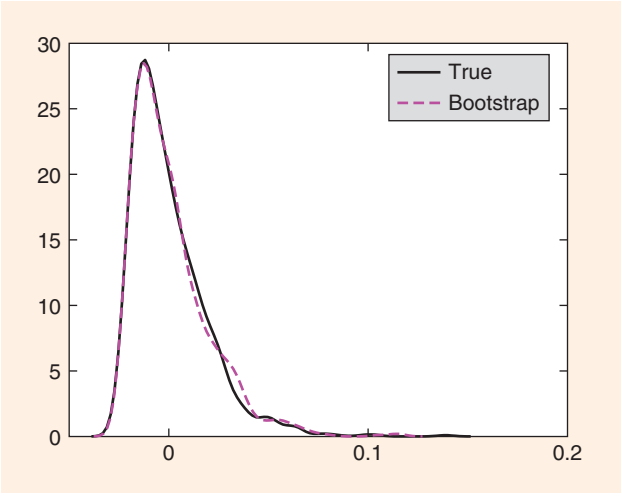
where $\mathcal{S}_n = \{\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n}\}; \{\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,n}\}$ Calibrating the test statistics corresponds to finding the quantile at level $(1-\alpha)$ of the cumulative distribution function $F$ of the test statistic $T_n$. The idea of sampling-based calibration is to estimate $c_{1-\alpha}(T_n; F)$ by $c_{1-\alpha}(T_n; \hat{F}_n)$, and to approximate $c_{1-\alpha}(T_n; \hat{F}_n)$ using simulations. Hence, we simulate $\mathcal{S}_n^* = \{\mathbf{x}_{1,1}^*, \ldots, \mathbf{x}_{1,n}^*\}; \{\mathbf{x}_{2,1}^*, \ldots, \mathbf{x}_{2,n}^*\}$ from $\hat{F}_n$ and then compute $T_n^* = g(\mathcal{S}_n^*)$. This constitutes one draw from the distribution of $T_n$. Thus, to simulate "ghost" bootstrap samples $\mathcal{S}_n^* \sim \hat{F}_n$, it suffices to draw $2n$ observations with a replacement from $\mathbf{x}_{1,i}, \ldots, \mathbf{x}_{1,n}; \mathbf{x}_{2,i}, \ldots, \mathbf{x}_{2,n}$. This can be summarized by the following diagram, paraphrased from [31]:

$$\begin{aligned} \text{Real world} \quad & F \to \mathcal{S}_n \to T_n = g(\mathcal{S}_n) \\ \text{Bootstrap world} \quad & \hat{F}_n \to \mathcal{S}_n^* \to T_n = g(\mathcal{S}_n^*), \end{aligned}$$

with

$$\begin{aligned} \text{Real sample} \quad & \mathcal{S}_n = (\{\mathbf{x}_{1,i}\}_{i=1}^n; \{\mathbf{x}_{2,i}\}_{i=1}^n) \\ \text{“Ghost” sample} \quad & \mathcal{S}_n^* = (\{\mathbf{x}_{1,i}^*\}_{i=1}^n; \{\mathbf{x}_{2,i}^*\}_{i=1}^n). \end{aligned}$$

We provide an illustration of the accuracy of the sampling-based calibration for MMD using a Gaussian RBF kernel with bandwidth set to 0.5. We consider two samples with 100 observations drawn for a normal distribution. We compare the true null distribution, which we can simulate using a large number of replications or compute using the limiting null distribution. We also compute the sampling-based distribution using sampling with replacement in these two samples. As Figure 6 shows, the sampling-based distribution leads to a rather accurate calibration of the test.




**[FIG6] A comparison of the true null distribution (black) and the sampling-based distribution for the MMD test statistic.**

## CONCLUSIONS

Kernel-based methods were used extensively and successfully for binary and multiway supervised classification problems in signal processing and machine learning. We showed how kernel-based methods can also be used for detection purpose, to build kernel-based hypothesis test statistics. These test statistics can, most of the time, be related to well-known information divergences between distributions. Therefore, kernel-based methods offer an attractive framework to build nonparametric detection procedures, applicable to a wide range of high-dimensional and structured data. Many detection problems were considered in the signal processing literature, and new detection problems arise with new applications. Kernel-based methods offer an elegant set of tools to tackle these new challenges.

## ACKNOWLEDGMENTS

The authors acknowledge the contribution of colleagues at Institut National de l'Audiovisuel (INA). This work was partially funded by a MSTIC project from the Université de Grenoble, European Research Council (SIERRA Project), French State Agency for Innovation OSEO under the Quaero program, and the PASCAL 2 Network of Excellence.

## AUTHORS

***Zaid Harchaoui*** (zaid.harchaoui@inria.fr) graduated from the Ecole Nationale Superieure des Mines, Saint-Etienne, in 2004, and received the Ph.D. degree in 2008 from Telecom ParisTech, Paris. He is now a researcher in the LEAR team of INRIA in the Laboratoire Jean Kuntzmann, Grenoble, France. His research interests include machine learning, statistics, optimization, kernel-based methods, audio, and computer vision. He coorganized the 2009 Neural Information Processing Systems Foundation Workshop on Temporal Segmentation in Vancouver, Canada, and the 2012 International Conference on Machine Learning Workshop on Kernel-Based Methods in Edinburgh, Scotland.

***Francis Bach*** (francis.bach@ens.fr) graduated from the Ecole Polytechnique, Palaiseau, France, in 1997. He received the Ph.D. degree in 2005 from the Computer Science Division at the University of California, Berkeley. He is the leading researcher of the Sierra project-team of INRIA in the Computer Science Department of the Ecole Normale Supérieure, Paris, France. His research interests include machine learning, statistics, optimization, graphical models, kernel methods, and statistical signal processing. He is currently the action editor of the *Journal of Machine Learning Research* and associate editor of *IEEE Transactions in Pattern Analysis and Machine Intelligence*.

***Olivier Cappé*** (olivier.cappe@telecom-paristech.fr) received the M.Sc. degree in electrical engineering from the Ecole Supérieure d'Electricité, Paris, France, in 1990, and the Ph.D. degree in signal processing from the Ecole Nationale Superieure des Telecommunications (ENST), Paris, in 1993. He is now senior research scientist and director of Laboratoire Traitement et Communication de l'Information, a joint lab between Centre National de la Recherche Scientifique and Telecom ParisTech. His research interests are in statistical signal processing, computational statistics, and statistical learning. He was a member of the IEEE Signal Processing Society's Signal Processing Theory and Methods Technical Committee from 2005 to 2010, and associate editor of *IEEE Transactions on Signal Processing* from 2000 to 2003. He is currently an associate editor of *Journal of the Royal Statistical Society* (*Series B*).

***Éric Moulines*** (eric.moulines@telecom-paristech.fr) received the M.S. degree from Ecole Polytechnique, Palaiseau, France, in 1984, and the Ph.D. degree in signal processing from Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1990. Since 1990, he has been with ENST, where he is currently a professor. His teaching and research interests include statistical signal processing, speech processing, and mathematical and computational statistics, i.e., Markov chain Monte Carlo, population methods, and hidden Markov models. He is currently the editor-in-chief of *Bernoulli Journal*.

## REFERENCES

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP*, vol. 4, pp. 430–451, Jan. 2004.

[3] J. Keshet and S. Bengio, Eds. *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*. Hoboken, NJ: Wiley, 2008.

[4] L. R. Rabiner and R. W. Schäfer, "Introduction to digital signal processing," *Found. Trends Inform. Retrieval*, vol. 1, nos. 1–2, pp. 1–194, 2007.

[5] M. Basseville and N. Nikiforov, *The Detection of Abrupt Changes* (Information and System Sciences Series). Englewood Cliffs, NJ: Prentice Hall, 1993.

[6] P. Fearnhead, "Exact and efficient Bayesian inference for multiple changepoint problems," *Stat. Comput.*, vol. 16, no. 2, pp. 203–213, June 2006.

[7] J. Chen and A. K. Gupta, *Parametric Statistical Change-Point Analysis*. Cambridge, MA: Birkhäuser, 2000.

[8] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, 3rd ed. New York: Springer-Verlag, 2005.

[9] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, 2nd. ed. Englewood Cliffs, NJ: Prentice Hall, 2012.

[10] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[11] B. Schölkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[12] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (Adaptive Computation and Machine Learning). Cambridge, MA: MIT Press, 2005.

[13] Z. Harchaoui, F. Bach, and E. Moulines, "Testing for homogeneity with kernel Fisher discriminant analysis," in *Advances in Neural Information Processing Systems*, 2008.

[14] Z. Harchaoui, F. Bach, and E. Moulines, "Kernel change-point analysis," in *Advances in Neural Information Processing Systems*, 2009.

[15] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[16] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[17] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola, "A kernel statistical test of independence," in *Advances in Neural Information Processing Systems*, 2008.

[18] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappé, "A regularized kernel-based approach to unsupervised audio segmentation," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 1665–1668.

[19] O. Gillet, S. Essid, and G. Richard, "On the correlation of automatic audio and visual segmentation of music videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 347–355, 2007.

[20] R. Brunelli, O. Mich, and C. M. Modena, "A survey on the automatic indexing of video data," *J. Vis. Commun. Image Represent.*, vol. 10, no. 2, pp. 78–112, 1999.

[21] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," *Comput. Vis. Image Understand.*, vol. 114, no. 4, pp. 411–418, 2010.

[22] F. Vallet, "Structuration automatique de shows télévisés," Ph.D. dissertation, Telecom ParisTech, France, 2011.

[23] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 2005, vol. 5, pp. v/953–v/956.

[24] Z. Harchaoui and O. Cappé, "Retrospective multiple change-point estimation with kernels," in *Proc. IEEE Workshop Statistical Signal Processing (SSP)*, 2007, pp. 768–772.

[25] H. Wendland, *Scattered Data Approximation* (Cambridge Monographs on Applied and Computational Mathematics). Cambridge, U.K.: Cambridge Univ. Press, 2005.

[26] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[27] M. Hein and O. Bousquet, "Hilbertian metrics and positive-definite kernels on probability measures," in *Proc. AISTATS*, 2004, pp. 136–143.

[28] K. Fukumizu, F. Bach, and A. Gretton, "Statistical convergence of kernel canonical correlation analysis," *J. Mach. Learn. Res.*, vol. 8, no. 8, pp. 361–383, 2007.

[29] E. Lehmann, *Elements of Large-Sample Theory*. New York: Springer-Verlag, 1999.

[30] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley Interscience, 2003.

[31] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer Texts in Statistics). New York: Springer-Verlag, 2004.

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.

[33] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[34] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proc. Int. Conf. Machine Learning (ICML)*, 2009, p. 121.

[35] N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan, *Probability Distributions on Banach Spaces*. Amsterdam, The Netherlands: Reidel, 1987.

[36] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, July 2002.

[37] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," *J. Mach. Learn. Res.*, vol. 7, pp. 2651–2667, Dec. 2006.

[38] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *Proc. 21st Annu. Conf. Learning Theory (COLT)*, 2008, pp. 111–122.

[39] I. Steinwart and A. Christmann, *Support Vector Machines*. New York: Springer-Verlag, 2008.

[40] G. Blanchard, O. Bousquet, and L. Zwald, "Statistical properties of kernel principal component analysis," *Mach. Learn.*, vol. 66, no. 2–3, pp. 259–294, 2007.

[41] S. Mika, G. Raetsch, J. Weston, B. Schoelkopf, A. J. Smola, and K.-R. Mueller, "Constructing descriptive and discriminative non-linear features: Rayleigh coefficients in kernel feature spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 623–628, 2003.

[42] Z. Harchaoui, F. Bach, and E. Moulines, "Testing for homogeneity with kernel Fisher discriminant analysis." (2009). [Online]. Available: http://arxiv.org/pdf/0804.1026v1.pdf

[43] A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola, "A kernel method for the two-sample problem," in *Advances in Neural Information Processing Systems*, 2006.

[44] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur, "A fast, consistent kernel two-sample test," in *Advances in Neural Information Processing Systems*, 2009.

[45] N. H. Anderson, P. Hall, and D. M. Titterington, "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates," *J. Multivariate Anal.*, vol. 50, no. 1, pp. 41–54, 1994.

[46] L. Wasserman, *All of Nonparametric Statistics* (Springer Texts in Statistics). New York: Springer-Verlag, 2006.

[47] D. Sejdinovic, A. Gretton, B. K. Sriperumbudur, and K. Fukumizu, "Hypothesis testing using pairwise distances and associated kernels," in *Proc. Int. Conf. Machine Learning (ICML)*, 2012.

[48] T. Kanamori, T. Suzuki, and M. Sugiyama, "f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models," *IEEE Trans. Inform. Theory*, vol. 58, no. 2, pp. 708–720, 2012.

[49] X. L. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inform. Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[50] T. Kanamori, T. Suzuki, and M. Sugiyama, "Statistical analysis of kernel-based least-squares density-ratio estimation," *Mach. Learn.*, vol. 86, no. 3, pp. 335–367, 2012.

[51] M. Davy, F. Désobry, and C. Doncarli, "An online support vector machine for abnormal events detection," *Signal Process.*, vol. 1, no. 8, pp. 2009–2025, 2005.

[52] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2961–2974, Aug. 2005.

[53] M. Davy, F. Desobry, and S. Canu, "Estimation of minimum measure sets in reproducing kernel Hilbert spaces and applications," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2006, vol. 3, pp. 14–19.

[54] R. Vert and J.-P. Vert, "Consistency and convergence rates of one-class SVM and related algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 817–854, June 2006.

[55] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *J. Mach. Learn. Res.*, vol. 5, pp. 819–844, Apr. 2004.

[56] A. Gretton and L. Györfi, "Consistent nonparametric tests of independence," *J. Mach. Learn. Res.*, vol. 11, pp. 1391–1423, Apr. 2010.

[57] J. C. Principe, W. Liu, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*. Hoboken, NJ: Wiley, 2012.

[58] K. Fukumizu, F. R. Bach, and A. Gretton, "Statistical consistency of kernel canonical correlation analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 361–383, July 2007.

[59] A. Gretton, O. Bousquet, A. Smola, and S. Bernhard, "Measuring statistical dependence with Hilbert–Schmidt norms," in *Proc. 16th Int. Conf. Algorithmic Learning Theory (ALT)*, 2005, pp. 63–77.

[60] H. Shen, S. Jegelka, and A. Gretton, "Fast kernel-based independent component analysis," *IEEE Trans. Signal Processing*, vol. 57, no. 9, pp. 3498–3511, 2009.

[SP]