BULLETIN OF THE BRAZILIAN MATHEMATICAL SOCIETY

Bull Braz Math Soc, New Series 37(4), 1-25

Random projections and goodness-of-fit tests in infinite-dimensional spaces

Juan Antonio Cuesta-Albertos*, Ricardo Fraiman and Thomas Ransford**

Abstract. In this paper we provide conditions under which a distribution is determined by just one randomly chosen projection. Then we apply our results to construct goodness-of-fit tests for the one and two-sample problems. We include some simulations as well as the application of our results to a real data set. Our results are valid for every separable Hilbert space.

Keywords: Cramér-Wold theorem, random projections, Hilbert spaces, goodness-of-fit tests, Kolmogorov-Smirnov projected test, single null hypothesis, two samples.

Mathematical subject classification: Primary: 62H15; Secondary: 28C20, 60B11, 60E05.

1 Introduction

Recent advances in technology allow significantly more data to be recorded over a period of time, leading to samples composed of trajectories which are measured on each of a number of individuals. Such data are common in different fields, including health sciences, engineering, physical sciences, chemometrics, finance and social sciences. They are often referred to as functional data or longitudinal data (this last term being preferred in health and social sciences). In this context, the data can be considered as independent, identically distributed realizations of a stochastic process taking values in a Hilbert space. For instance, we might have a random sample $\{X_1(t), \ldots, X_n(t) : t \in T\}$ of trajectories with values in the Hilbert space $L^2(T)$, where T is an interval in $\mathbb{R}$.

Received 26 April 2006.

*Partially supported by the Spanish Ministerio de Ciencia y Tecnología, grant MTM2005-08519-C02-02.

**Partially supported by grants from NSERC and the Canada research chairs program.

Concerning the mathematical procedures to handle these data, it happens that, on the one hand, there are many problems whose solution is known from a theoretical point of view, while its implementation is difficult in practice if the dimension of the space which contains the data is infinite (or, simply, large). Of course, on the other hand, there are some problems whose solution is known for finite-dimensional data but it is unknown for functional data.

These kind of problems appear not only in Statistics but in many other fields in Mathematics. A way to circumvent them has been to employ randomly chosen projections. Broadly speaking, this procedure can be described as follows. Let us assume that we have to deal with a problem related to d-dimensional objects. The random projection method consists of choosing, at random, a subspace of dimension k (where k is low when compared to d), solve the problem in the k-dimensional subspace and, then, translate the solution to the original (d-dimensional) space.

Many of these applications are based on the fact that random projections approximately preserve pairwise distances with high probability (see, for instance, [21], Section 1.2 or Lemma 2.2 in [6] for two precise formulations of this statement).

We do not try to be exhaustive, but some applications of these ideas, can be seen, for instance, in [21], where they are employed to obtain approximate algorithms in problems of high computational complexity; or in [12] where the authors propose the use of randomly chosen projections as a tool to identify images and, then, detect copyright violations on images posted on the Internet.

It is curious that, though Statistics lies at the heart of the random projection method, this idea has seldom been applied to statistical problems. We are only aware of some results in which random projections have been used to estimate mixtures of distributions [8, 22], but even these papers have not been written from a purely statistical point of view but rather from the perspective of learning theory.

On the other hand, in [5] a generalization of the Cramér-Wold theorem was proved. Some results in this paper (Corollary 3.2 and Theorem 4.1) state that, under suitable conditions, a randomly chosen projection determines a distribution. We consider that these results could provide the basis to start the statistical analysis we refer to in previous paragraph and with this idea, we describe in this paper how they can be applied to obtain goodness-of-fit tests to a single distribution or to test whether two independent samples come from the same distribution.

Perhaps it is worth stressing that the proposed tests will be based on just a single (randomly chosen) one-dimensional projection. This is exactly contrary to

Projection Pursuit paradigm which, if applied to the above mentioned problems, would dictate consideration of every possible one-dimensional projection. Note that this renders Projection Pursuit extremely sensitive to the dimension and, then, it runs directly into the problems we mentioned in the beginning.

We remark that the results in [5] are valid in any separable, even infinite-dimensional, Hilbert space, and are thus applicable to the analysis of stochastic processes.

The organization of this paper is as follows. In Section 2 we present some of the results in [5]. To be precise, in fact we present a slight generalization of them which allows us to write them in a slightly sharper and perhaps more friendly way. Then, in Section 3 we show how these results can be applied to construct statistical tests. In Section 4, we present some simulations to show how the procedure behaves in practice. We conclude with the application, in Section 5, of the proposed method to a real data set. The data consist of the spectrograms of 95 healthy women and those of 121 women who suffered from ovarian cancer. They have been downloaded from http://ncifdaproteomics.com/OvarianCD_PostQAQC.zip. All computations, including the simulations, have been carried out with MatLab. Original codes are available from the first-named author upon request.

2 Basic results on random projections

We begin by establishing some notation, as well as a few basic elementary results.

Let $\mathcal{H}$ be a real, separable Hilbert space (finite- or infinite-dimensional). We write $\langle \cdot\,, \cdot \rangle$ for the inner product on $\mathcal{H}$, and $\| \cdot \|$ for the corresponding norm. Given $x \in \mathcal{H}$, we denote by $P_{\langle x \rangle}$ the marginal of P onto the one-dimensional subspace generated by x. Namely, if π_x denotes the orthogonal projection of $\mathcal{H}$ on the one-dimensional subspace spanned by x, and B is a Borel set in this subspace, then,

$$P_{\langle x \rangle}(B) := P[\pi_x^{-1}(B)].$$

Given two Borel probability measures P, Q on $\mathcal{H}$, we define

$$\mathcal{E}(P, Q) := \left\{ x \in \mathcal{H} \colon P_{\langle x \rangle} = Q_{\langle x \rangle} \right\}.$$

The set $\mathcal{E}(P, Q)$ is closed and, hence, Borel-measurable. This set will play a central role in what follows. Obviously, the very well known Cramér–Wold theorem for $\mathcal{H}$ can be stated in terms of $\mathcal{E}(P, Q)$.

Proposition 2.1. *If* $\mathcal{E}(P, Q) = \mathcal{H}$, *then* $P = Q$.

It is well known (see [17, Theorem 1]) that a compactly supported Borel probability measure on $\mathbb{R}^2$ is determined by its marginals onto any infinite set of lines. This result was generalized in [7, Theorem 1] by replacing the compactness condition by some hypothesis on the moments of the distribution, but it is not possible to generalize this result to $\mathbb{R}^d$ when $d \geq 3$. In some sense, the goal of the main result in [5] was to formulate the 'correct' condition in high-dimensional spaces. We will employ the following definition.

Definition 2.2. *Let P be a Borel distribution on the separable Hilbert space $\mathcal{H}$. We will say that P is determined by its moments if for every $n \in \mathbb{N}$, it happens that $\int \|x\|^n P(dx) < \infty$, and if Q is another Borel distribution on $\mathcal{H}$ such that*

$$\int \langle x, y\rangle^n P(dy) = \int \langle x, y\rangle^n Q(dy), \quad \text{for every } x \in \mathcal{H} \text{ and } n \in \mathbb{N},$$

then $P = Q$.

Some conditions to ensure that a distribution is determined by its moments have been proposed in the literature. For instance, it is very well known that if the moment generating function of P is finite on a neighborhood of the origin, then P is determined by its moments. A more general condition, the so-called Carleman condition, is provided, for instance in [18, p.19] for the case in which $\mathcal{H}$ is finite dimensional but it is easily extended to cover also the general case as follows.

Proposition 2.3 (Carleman condition). *Let P be a Borel distribution on the separable Hilbert space $\mathcal{H}$. Assume that the absolute moments $m_n := \int \|x\|^n P(dx)$ are finite and satisfy $\sum_{n\geq 1} m_n^{-1/n} = \infty$. Then, P is determined by its moments.*

In the finite-dimensional case, the key result in [5] on the determination of a distribution by its one-dimensional marginals is Theorem 3.1. It relies on the following definition: a polynomial p on $\mathbb{R}^d$ is called *homogeneous of degree m* if $p(tx) = t^m p(x)$ for all $t \in \mathbb{R}$ and all $x \in \mathbb{R}^d$. A subset S of $\mathbb{R}^d$ is called a *projective hypersurface* if there exists a homogeneous polynomial p on $\mathbb{R}^d$, not identically zero, such that $S = \{x \in \mathbb{R}^d : p(x) = 0\}$.

Most of the proof of Theorem 3.1 in [5] consists of proving the following result.

Proposition 2.4. *Let P, Q be Borel probability measures on $\mathbb{R}^d$, where $d \geq 2$. Assume that:*

- $\int \|x\|^n P(dx) < \infty$, *for every $n \in \mathbb{N}$;*

- *the set* $\mathcal{E}(P, Q)$ *is not contained in any projective hypersurface in* $\mathbb{R}^d$.

Then $\int \langle x, y\rangle^n P(dy) = \int \langle x, y\rangle^n Q(dy)$, *for every* $n \in \mathbb{N}$ *and* $x \in \mathcal{H}$.

From this point, it is trivial to finish the proof of Theorem 3.1 in [5] and also to obtain the following slight generalization of this theorem.

Theorem 2.5. *Let* P, Q *be Borel probability measures on* $\mathbb{R}^d$, *where* $d \geq 2$. *Assume that:*

- P *is determined by its moments;*
- *the set* $\mathcal{E}(P, Q)$ *is not contained in any projective hypersurface in* $\mathbb{R}^d$.

Then $P = Q$.

Remark. In [5] some counterexamples and results are included which show that the conditions in this result are sharp but we do not include them here.

Particularly important, from the point of view of applications, is the following corollary which corresponds to Corollary 3.2 in [5].

Corollary 2.6. *Let* P, Q *be Borel probability measures on* $\mathbb{R}^d$, *where* $d \geq 2$. *Assume that:*

- P *is determined by its moments;*
- *the set* $\mathcal{E}(P, Q)$ *is of positive Lebesgue measure in* $\mathbb{R}^d$.

Then $P = Q$.

Proof. This is an immediate consequence of Theorem 2.5, because every projective hypersurface is of Lebesgue measure zero in $\mathbb{R}^d$. □

To extend Corollary 2.6 to infinite-dimensional spaces, we need to find a substitute for Lebesgue measure, which no longer has any sense in this setting. The substitute will be a non-degenerate gaussian measure. For the sake of completeness, we state here the definition. For more details on gaussian measures, see e.g. [11, §7.5 and §7.6].

Definition 2.7. *Let* $\mathcal{H}$ *be a separable Hilbert space. A Borel probability measure* μ *on* $\mathcal{H}$ *is called* gaussian *if each of its one-dimensional marginals is gaussian. It is* non-degenerate *if, in addition, each of its one-dimensional marginals is non-degenerate.*

The following result is the infinite-dimensional generalization of Corollary 2.6. The main technical difficulty in this result relies on the fact that it is not obvious that if a infinite-dimensional distribution, P, is determined by its moments, and we consider a finite-dimensional projection of P, then this projection is also determined by its moments. However, if we assume that P satisfies Carleman's condition, then it is obvious that their marginals also do, and, in consequence, they are determined by their moments. This was a keystone in the proof of Theorem 4.1 in [5]. Here, we will employ Proposition 2.4 to circumvent this difficulty.

Theorem 2.8. *Let $\mathcal{H}$ be a separable Hilbert space, and let μ be a non-degenerate gaussian measure on $\mathcal{H}$. Let P, Q be Borel probability measures on $\mathcal{H}$. Assume that:*

- *P is determined by its moments;*
- *the set $\mathcal{E}(P, Q)$ is of positive μ-measure.*

Then $P = Q$.

Proof. The first part follows the same steps as the proof of Theorem 4.1 in [5]. To this end, take an orthonormal basis of eigenvectors $\{e_k\}_{k\geq 1}$ of the covariance operator of μ. Then μ is the product measure of its marginals on F_k (the finite dimensional subspace generated by $\{e_1, \ldots, e_k\}$) and on $F_k^{\perp}$ (the orthogonal complement of F_k). Let us denote by P_k and Q_k the marginal distributions of P and Q on F_k respectively.

Obviously P_k satisfies first hypothesis in Proposition 2.4.

If we employ Fubini's theorem and carry out the same computations as in the proof of Theorem 4.1 in [5], we find that the k-dimensional Lebesgue measure of $\mathcal{E}(P_k, Q_k)$ is strictly positive. Thus, by Proposition 2.4, for every $n \in \mathbb{N}$, and $x \in F_k$, we have

$$\int \langle x, y\rangle^n P_k(dy) = \int \langle x, y\rangle^n Q_k(dy). \tag{1}$$

Now, if n is an even integer, then for every k and every $y \in F_k$ we have $\|y\|^n = c_{n,k} \int \langle x, y\rangle^n \sigma_k(dx)$, where σ_k denotes Lebesgue measure on the unit sphere of F_k, and $c_{n,k}$ is a positive constant depending only on n, k. Thus, integrating (1) with respect to σ_k, and using Fubini, we obtain

$$\int \|y\|^n \, P_k(dy) = \int \|y\|^n \, Q_k(dy).$$

In other words, writing π_k for the orthogonal projection of $\mathcal{H}$ onto F_k,

$$\int \|\pi_k(y)\|^n \, P(dy) = \int \|\pi_k(y)\|^n \, Q(dy).$$

Letting $k \to \infty$ and using the monotone convergence theorem, we deduce that

$$\int \|y\|^n \, P(dy) = \int \|y\|^n \, Q(dy).$$

As the left-hand side is finite, so is the right-hand side (for all even n and hence for all n).

Returning to (1), it says that for all $n \in \mathbb{N}$, and all $x \in \mathcal{H}$,

$$\int \langle \pi_k(x), \pi_k(y)\rangle^n \, P(dy) = \int \langle \pi_k(x), \pi_k(y)\rangle^n \, Q(dy).$$

The modulus of the integrand is bounded by the function $y \mapsto \|x\|^n \|y\|^n$, which is both P- and Q-integrable. So we can let $k \to \infty$ and use the dominated convergence theorem to obtain

$$\int \langle x, y\rangle^n \, P(dy) = \int \langle x, y\rangle^n \, Q(dy).$$

Since P is determined by its moments, we conclude that $P = Q$. □

3 Application: Goodness-of-fit tests

Goodness-of-fit tests of Kolmogorov–Smirnov type are the most widely used tests to decide whether it is reasonable to assume that some one-dimensional data come from a given distribution. The problem is the following: Given i.i.d. real random variables $X_1, \ldots, X_n$ on a probability space $(\Omega, \mathcal{A}, \nu)$, can we accept that their underlying common distribution is a given P_0? Thus, in terms of a statistical test-of-hypothesis problem, the null hypothesis H_0 is that the true underlying distribution P is equal to P_0, while the alternative hypothesis H_A is that $P \neq P_0$.

To carry out this test, Kolmogorov [9] suggested using the statistic

$$D_n := \sup_{t \in \mathbb{R}} |F_n(t) - F_0(t)|, \tag{2}$$

where F_0 is the distribution function of P_0, and F_n is the empirical distribution function, defined by

$$F_n(t) := \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,t]}(X_i) \qquad (t \in \mathbb{R}),$$

rejecting the null hypothesis when D_n is large.

If F_0 is continuous, and the null hypothesis holds, then the statistic D_n has the important property of being distribution-free, i.e. its distribution does not depend on the true underlying distribution P_0, but only on n. This distribution was tabulated by Smirnov [20] and Massey [14, 15], and is available in most statistical packages. Kolmogorov [9] also found the asymptotic distribution of $\sqrt{n}D_n$ when H_0 holds. This distribution coincides with that of the maximum of a Brownian bridge. Its explicit expression is

$$\lim_{n\to\infty} \nu(\sqrt{n}D_n \le t) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2t^2} \qquad (t > 0).$$

Later on, Smirnov [19] and Kolmogorov [10] treated the two-sample problem with similar techniques. Here, we have two independent random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, taken from the distributions P and Q respectively, and the problem is to decide whether it is reasonable to assume that $P = Q$. Thus, the null hypothesis H_0 is now $P = Q$, while the alternative hypothesis H_A is $P \neq Q$. Denoting by F_n and G_m the respective empirical distributions obtained from each sample, the proposed statistic for this problem was

$$D_{n,m} := \sup_{t\in\mathbb{R}} \big|F_n(t) - G_m(t)\big|.$$

The properties of $D_{n,m}$ are very similar to those of D_n. In particular, under the null hypothesis, if P (and hence Q) is continuous, then $D_{n,m}$ is distribution-free. Moreover,

$$\lim_{\min(n,m)\to\infty} \nu\left(\sqrt{\frac{mn}{m+n}}D_{n,m} \le t\right) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2t^2} \qquad (t > 0).$$

Turning now to higher dimensions, to the best of our knowledge there are still no satisfactory extensions of the Kolmogorov–Smirnov tests, even for two-dimensional data. All proposals fail on at least one of the following two counts: (i) being independent of a reference basis on the space, i.e. equivariant with respect to orthogonal transformations, and/or (ii) being distribution-free. One of the main problems in constructing a distribution-free test in higher dimensions is to define appropriate correlates of the rank statistics in order to obtain the analogue of F_n, the empirical distribution function. (Recall that, given distinct real numbers $x_1, \ldots, x_n$, the rank R_i of x_i is the place that x_i occupies in the ordered vector $x^{(1)} < \ldots < x^{(n)}$ obtained by ordering the original vector, i.e. $x_i = x^{(R_i)}$.)

The results in this section will provide goodness-of-fit tests for random elements taking values in a separable Hilbert space $\mathcal{H}$. In particular, this will provide goodness-of-fit tests for stochastic processes. As far as we know, this is the first such proposal in this setting. The problem that we shall analyze is the following: Let P_X denote the common probability law of the random elements $X_1, \ldots, X_n$ in $\mathcal{H}$. Given a probability measure P_0 on $\mathcal{H}$, provide a procedure to decide when the data call into question the null hypothesis $H_0\colon P_X = P_0$ in favor of the alternative $H_A\colon P_X \neq P_0$.

The procedure we propose consists of (i) to choose a random direction h in $\mathcal{H}$, according to a non-degenerate gaussian law μ on $\mathcal{H}$, and then (ii) to apply the standard Kolmogorov–Smirnov test to the orthogonal projections of the data onto the one-dimensional subspace spanned by h. Thus, according to (2), we compute the statistic

$$D_n(h) := \sup_{t\in\mathbb{R}} \left|F_n^h(t) - F_0^h(t)\right|, \tag{3}$$

where now

$$F_0^h(t) := P_0\left\{x \in \mathcal{H}\colon \langle x, h\rangle \leq t\right\} \quad \text{and}$$

$$F_n^h(t) := \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(\langle X_i, h\rangle) \qquad (t \in \mathbb{R}),$$

and reject the null hypothesis when $D_n(h)$ is large enough.

The properties of the proposed procedure are summarized in the following theorem. We shall say that P is *continuous* if each of its one-dimensional projections is continuous. This is equivalent to demanding that every closed affine hyperplane in $\mathcal{H}$ be of P-measure zero.

Theorem 3.1. *Let $\{X_n\}_{n\geq 1}$ be a sequence of independent, identically distributed random elements, defined on the probability space $(\Omega, \mathcal{A}, \nu)$, and taking values in a separable Hilbert space $\mathcal{H}$. Let P_0 be a probability measure on $\mathcal{H}$. Given $h \in \mathcal{H}$ and $n \geq 1$, define $D_n(h)$ as in* (3)*.*

(a) *Suppose that the common distribution of $\{X_n\}_{n\geq 1}$ is P_0. Suppose also that P_0 is continuous. Then, for all $h \in \mathcal{H} \setminus \{0\}$ and all $n \geq 1$, the statistic $D_n(h)$ has the same distribution as D_n. In particular, this distribution is independent of h, and*

$$\lim_{n\to\infty} \nu\left(\sqrt{n}D_n(h) \leq t\right) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2t^2} \qquad (t > 0).$$

(b) *Suppose that the common distribution of* $\{X_n\}_{n\geq 1}$ *is* $Q \neq P_0$. *Suppose also that* P_0 *is determined by its moments. Then, given any non-degenerate gaussian measure* μ *on* $\mathcal{H}$, *for* μ*-almost all* $h \in \mathcal{H}$ *we have*

$$\nu\left(\liminf_{n\to\infty} D_n(h) > 0\right) = 1.$$

Part (a) of the theorem tells us how, given a level α, we can find $c_{\alpha,n}$ (independent of h) such that, under the null hypothesis,

$$\nu(D_n(h) > c_{\alpha,n}) = \alpha,$$

thereby providing an α-level conditional test. Part (b) of the theorem says that the test is consistent against every possible alternative.

Proof Theorem 3.1.

(a) If the common distribution of $\{X_n\}_{n\geq 1}$ is P_0, then the common distribution function of the real random variables $\{\langle X_n, h\rangle\}_{n\geq 1}$ is just F_0^h, which is continuous. Also, the empirical distribution function of $\langle X_1, h\rangle, \ldots, \langle X_n, h\rangle$ is exactly F_n^h. Therefore this part follows by the standard properties of the one-dimensional Kolmogorov–Smirnov test.

(b) By Theorem 2.8, if $Q \neq P_0$, then, for μ-almost all $h \in \mathcal{H}$, there exists $t_h \in \mathbb{R}$ such that

$$P_0\big\{x \in \mathcal{H}\colon \langle x, h\rangle \leq t_h\big\} \neq Q\big\{x \in \mathcal{H}\colon \langle x, h\rangle \leq t_h\big\}.$$

Let δ_h be the absolute value of the difference. Then, using the triangle inequality,

$$D_n(h) \geq \big|F_n^h(t_h) - F_0^h(t_h)\big| \geq \delta_h - \big|F_n^h(t_h) - G^h(t_h)\big|,$$

where $G^h(t) := Q\{x \in \mathcal{H}\colon \langle x, h\rangle \leq t\}$. By the strong law of large numbers, $F_n^h(t_h) \to G^h(t_h)$ ν-almost surely. The result follows. □

We remark that our aim is to provide a so-called 'universal' test, namely a test valid in any context, rather than trying to be optimal in a particular setting. In fact, in some of the simulations that we shall present later, we shall restrict the alternative to a particular parametric family, and it is well known that, against this restricted alternative, there are more powerful tests. The problem is that these tests are not, in general, consistent against every possible alternative, whereas our proposed procedure is. This point will be taken up again later.

In practice, for a given problem, instead of taking just one random direction, we can choose a finite set of directions $h_1, \ldots, h_k$ at random, and then consider as statistic $D_n^k := \max_{1\le i\le k} D_n(h_i)$, the maximum of the projected one-dimensional Kolmogorov–Smirnov statistics over the k directions. The asymptotic distribution of this statistic is easy to derive. A drawback of this approach is that we lose the distribution-free property, since the distribution of D_n^k will depend on the covariance function of the underlying distribution P_X.

On the other hand, if the sample size is large, then we can still obtain a distribution-free statistic as follows. Split the sample into k subsamples,

$$\left\{X_{m_1}, \ldots, X_{m_{n_i}}\right\}, \quad i = 1, \ldots, k\,,$$

select k independent directions $\{h_1, \ldots, h_k\}$ at random, then, for each $i = 1, \ldots, k$, compute the one-dimensional Kolmogorov–Smirnov statistic of the projection of the subsample $\left\{X_{m_1}, \ldots, X_{m_{n_i}}\right\}$ on the direction given by h_i, and, finally, compute the maximum of these quantities. The distribution of the statistic thereby obtained is just that of the maximum of k independent one-dimensional Kolmogorov–Smirnov random variables, and is therefore still distribution-free. However, it should be remarked that in general this procedure entails a loss of power, which is not good statistical behavior.

As we can see, the random projection method serves to reduce the problem so we can apply univariate goodness-of-fit tests to the projected data. Once the data are projected, we can use not only the Kolmogorov–Smirnov test, but any other univariate goodness-of-fit test. In particular, for mixtures of stochastic processes, a modified Kolmogorov–Smirnov test like that proposed by [13] can be applied. The test will remain distribution-free, as long as the univariate test applied to the projected data has this property.

The two-sample problem can be treated in a very similar way. Let us assume that our data are independent, identically distributed realizations $\{X_1, \ldots, X_n\}$, $\{Y_1, \ldots, Y_m\}$ of two random processes taking values in the separable Hilbert space $\mathcal{H}$. Let P_X and P_Y stand for the common probability laws of the random elements X_i and Y_j, respectively. A goodness-of-fit test for the two-sample problem in this context will be a procedure to decide between the null hypothesis $H_0\colon P_X = P_Y$ and the alternative $H_A\colon P_X \neq P_Y$, based on $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$.

As in the one-sample case, we propose the following procedure: first choose a random direction $h \in \mathcal{H}$, according to the gaussian measure μ, and then calculate the following statistic:

$$D_{n,m}(h) := \sup_{t\in\mathbb{R}} \left|F_n^h(t) - G_m^h(t)\right|,$$

where

$$F_n^h(t) := \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(\langle X_i, h\rangle) \quad \text{and} \quad G_m^h(t) := \frac{1}{m}\sum_{j=1}^{m} I_{(-\infty,t]}(\langle Y_j, h\rangle),$$

rejecting the null hypothesis if $D_{n,m}(h)$ is large enough. Under the null hypothesis, the asymptotic distribution of $(mn)^{1/2}(m+n)^{-1/2}D_{n,m}(h)$ as $\min(n, m) \to \infty$ is the same as for the one-sample problem.

The possibility of handling the maximum deviation on a finite set of directions can be treated similarly in this case to that of the one-sample problem.

4 Simulations

In this section we present some simulations to show how the proposed tests work in practice. We consider the one-sample and the two-sample problems and we also analyze how using more than one random projection can contribute to increasing the power of the procedure. In all the examples we take $\mathcal{H} = L^2[0, 1]$. In the two-sample case we consider only one random projection. There are two reasons for this. Firstly the effect of taking more than one projection is similar to that obtained in the one sample case. Secondly, the computational burden increases strongly if more than one projection is handled because in this case the projected test is not distribution-free and the rejection region must be computed via simulations (see Subsection 4.2).

As mentioned in the Introduction, our aim in this section is to give an idea about how the results in the previous sections can be applied to obtain sound statistical procedures. We have not tried here to optimize them.

4.1 One-sample case. One random projection

In this section we assume that we have a random sample $X_1, \ldots, X_n$ of trajectories in $L^2[0, 1]$ and we want to test the null hypothesis that its underlying distribution (the one which produced the data) is that of the standard Brownian motion W on $[0, 1]$. To symplify the computations, the random direction will be chosen taking μ also to be the distribution of standard Brownian motion.

The proposed procedure only requires us to consider the scalar products $\langle X_i, h\rangle$, and it happens that, under the null hypothesis, the distribution of these real random variables is $N(0, \sigma^2(h))$, where

$$\sigma^2(h) := \int_0^1\int_0^1 \min(s, t)h(s)h(t)\, ds\, dt.$$

Therefore, under the null hypothesis, our procedure is equivalent to the Kolmogorov–Smirnov goodness-of-fit test applied to determine if a one-dimensional sample comes from the $N(0, \sigma^2(h))$ distribution.

To analize the behavior of our test under the alternative, we generate samples from some rescaled and shifted Brownian processes $S(t) := sW(t) + f(t)$, where $s \neq 0$. In this case, the distribution of $\langle S, h\rangle$ is also normal, with variance $s^2\sigma^2(h)$, and mean

$$\mu(h) := \int_0^1 f(t)h(t)\,dt.$$

Therefore, in some sense, the quality of the proposed procedure depends on the difference between $\mu(h)$ and zero, on that between s and 1, and, of course, on the capacity of the Kolmogorov–Smirnov test to detect them.

We will take $s = .5, .75, 1, 1.5, 2$. When $s \neq 1$, we will take $f(t) = 0$. When $s = 1$ we will consider $f(t) = \delta t$ for $\delta = 0, .25, .5, 1$ and $f(t) = \sin(\pi t)$. If $s = 1$ and $\delta = 0$, then the samples are generated by a standard Brownian motion and, therefore we are under the null hypothesis. This case is included to verify that the procedure provides the right level values under the null hypothesis.

Let us focus now on the case $s = 1$ and $f(t) = \delta t$, with $\delta \neq 0$. For this family of alternatives, the problem could also be handled by testing the null hypothesis H_0: 'the distribution of $S(1)$ is $N(0, 1)$', against H_A: 'the distribution of $S(1)$ is $N(\delta, 1)$ for some $\delta \neq 0$'. We can do this test performing the well-known Normal test for the mean of a normal distribution when the variance is known. Moreover, Prof. Barrio [2] kindly informed us that, by employing Girsanov's Theorem, it is possible to show that this test is uniformly most powerful against this family of alternatives, thus providing a gold standard for comparisons when handling this family of alternatives. However, this test would be useless in detecting alternatives such as the distribution of $S(1)$ is standard Normal.

Notice that the distribution under the null hypothesis is continuous. Thus, the projected test is distribution-free and the rejection region can be computed directly with the Kolmogorov-Smirnov test.

Concerning the simulations, we assume that the trajectories in the sample are observed on the equally spaced points $0 = t_0 < \ldots < t_{100} = 1$. This allows us to generate the standard Brownian motion from a discrete version which exploits the independent-increments property, i.e. we start at 0 at time zero, and define iteratively the value at the next time by adding an independent $N(0, 1/100)$ variable.

We have applied our procedure and the standard Normal test described above to 5000 random samples with sizes 30, 50, 100 and 200. The results are reported in Table 4.1.

Size			$s=1,\ f(t)=\delta t$ with $\delta=$				$s=1$ $f(t)=$	$f(t)=0$ and $s=$			
			0	.25	.5	1	$\sin(\pi t)$	.5	.75	1.5	2
30	rejections	KSp	.050	.176	.509	.934	.903	.444	.063	.234	.598
		$N1$	.048	.288	.780	.999	.052	0	.006	.186	.328
	average	KSp	.516	.364	.152	.024	.035	.087	.382	.239	.077
		$N1$	.496	.269	.052	0	.490	.704	.595	.373	.294
50	rejections	KSp	.052	.258	.726	.959	.938	.848	.099	.369	.861
		$N1$	.053	.437	.940	1	.052	0	.008	.184	.326
	average	KSp	.515	.294	.076	.016	.023	.026	.300	.154	.024
		$N1$	.498	.188	.012	0	.506	.707	.596	.372	.292
100	rejections	KSp	.049	.455	.911	.976	.957	.999	.264	.678	.996
		$N1$	.049	.697	1	1	.054	0	.009	.184	.329
	average	KSp	.504	.180	.029	.010	.017	.002	.166	.055	.002
		$N1$	.502	.075	0	0	.490	.703	.599	.380	.290
200	rejections	KSp	.046	.709	.965	.981	.970	1	.642	.966	1
		$N1$	.051	.939	1	1	.051	0	.009	.194	.332
	average	KSp	.497	.081	.015	.008	.012	0	.057	.008	0
		$N1$	.503	.013	0	0	.499	.706	.590	.368	.289

Table 4.1: Application of proposed procedure to the Brownian process $S(t)=sW(t)+f(t)$. The null hypothesis is the standard Brownian motion (i.e. $f(t)=0$ and $s=1$). As alternative hypotheses we take $s=1$ and $f(t)=\delta t,\ \delta=0.25, 0.5, 1$ or $f(t)=\sin(\pi t)$ and $f(t)=0$ and $s=.5, .75, 1.5, 2$. Samples sizes are 30, 50, 100 and 200. KSp denotes the proposed test while $N1$ stands for the standard Normal test applied at S(1). 'Rejections' denotes the proportion of rejections of the null hypothesis over 5000 trials. 'average' is the average of the p-values over these replications.

The column $\delta=0$ corresponds to the behavior under the null hypothesis of a test at the level $\alpha=0.05$. The remaining columns correspond to the behavior under different alternatives. We have chosen two parameters to measure the behavior of both tests: the 'rate of rejections' and the 'average p-value', which we now explain.

Recall that, for each random sample, the proposed procedure consists of selecting, at random, $h\in\mathcal{H}$, and then computing the probability that the Kolmogorov-Smirnov statistic D_n takes a value greater than the observed value of $D_n(h)$. We call this probability the *p-value*, and reject the null hypothesis if the p-value is less than 0.05. Otherwise we accept the null hypothesis. The Normal test works similarly. The average p-value is simply the mean of the observed p-values. An optimal procedure should provide averages close to 0.5 if the null hypothesis holds, and close to 0 under the alternative.

The rate of rejections is the proportion of times in which the procedure rejects the null hypothesis. Thus, this parameter should be close to 0.05 under the null hypothesis. Under the alternative, the bigger this parameter is, the better. This rate is an estimate of the power of the test under the simulated conditions.

We can summarize the results in Table 4.1 as follows. The proposed test performs well under the null hypothesis, $\delta = 0$. When we have a linear shift, the test performs clearly worse than the Normal test. The loss of power is around 40% in the worst case ($\delta = .25$ and sample size equal to 50), becoming less if we increase δ or the sample size. Roughly speaking, we can say that we have a loss of efficiency around 50% because we need to double the sample size in order to get a performance similar as to the Normal test.

On the other hand, if we change the alternative and we consider a sinusoidal shift, as expected, the Normal test becomes useless (with the same performance as under the null hypothesis) while the projected test works remarkably well.

Finally, in all the cases in which there is no shift but we change the variance, the random procedure clearly outperforms the Normal test. In those cases, the projected process is a zero-mean (real) random variable with variance $s^2\sigma^2(h)$. However, it happens that the Normal test depends on the absolute value of the sample mean of the values at $t = 1$. This value is compared with the corresponding one if the sample were produced from a standard normal distribution. When $s < 1$, it is expected the observed difference be less than the corresponding one under the null hypothesis. The expected difference increases with s and becomes larger than the target when $s > 1$. This explains the increase (decrease) observed in the rejections-rows (average-rows) for the Normal test. On the other hand, the projected test behaves more reasonably detecting more easily the alternative when the difference between s and 1 increases.

Let us pay some attention to the loss of power in the models with linear drift. It can be due to two facts: the loss of information caused by considering just an one-dimensional projection, and the loss of power due to employing the Kolmogorov-Smirnov test.

In order to separate these factors, we have done the following. First we have applied the Kolmogorov-Smirnov test to the sample composed of the values at $t = 1$ to test the null hypothesis that those values come from a standard Normal distribution. On the other hand, as previously stated, under the null hypothesis, the projections are a random sample taken from the $N(0, \sigma^2(h))$ distribution and we can check this applying the Normal test to the projections.

The results appear in Table 4.2, where we have denoted by KSp and Np the Kolmogorov-Smirnov and Normal tests applied to the projections, and $KS1$ and $N1$ the Kolmogorov-Smirnov and Normal tests applied to the values at $t = 1$.

		rejections				average			
Size	δ	KSp	Np	$KS1$	$N1$	KSp	Np	$KS1$	$N1$
30	0	.051	.052	.054	.050	.506	.499	.510	.500
	.25	.175	.221	.227	.282	.361	.324	.318	.276
	.5	.510	.626	.650	.779	.153	.111	.089	.051
	1	.939	.958	.997	1	.023	.017	.001	0
50	0	.049	.044	.050	.046	.508	.503	.515	.501
	.25	.252	.325	.318	.416	.301	.254	.249	.195
	.5	.721	.824	.870	.941	.077	.052	.028	.012
	1	.965	.971	1	1	.013	.011	0	0
100	0	.050	.053	.052	.052	.507	.505	.503	.494
	.25	.448	.562	.584	.708	.176	.134	.114	.074
	.5	.916	.946	.994	.999	.027	.019	.002	0
	1	.978	.981	1	1	.008	.007	0	0
200	0	.048	.048	.052	.053	.508	.499	.501	.497
	.25	.720	.814	.865	.941	.078	.055	.029	.013
	.5	.964	.971	1	1	.013	.011	0	0
	1	.984	.985	1	1	.007	.007	0	0

Table 4.2: Application of proposed procedure to the Brownian process $S(t) = W(t) + \delta t$. The null hypothesis is the standard Brownian motion (i.e. $\delta = 0$). As alternative hypotheses we take $\delta = 0.25, 0.5, 1$. Samples sizes are 30, 50, 100 and 200. KSp denotes the proposed test, Np stands for the standard Normal test applied to the projections, $KS1$ and $N1$ are the Kolmogorov-Smirnov and Normal tests applied to the sample of values at 1. 'Rejections' denotes for the proportion of rejections of the null hypothesis along 5000 trials. 'average' is the average of the p-values along those trials.

Differences observed between the values shown in Table 4.2 and the corresponding ones in Table 4.1 are due to the randomness of the experiments.

Roughly speaking again, the results obtained with Np are only a little bit worse than those obtained with $KS1$, thus suggesting that the loss of power due to taking a random projection are similar to that observed when using the Kolmogorov-Smirnov test against the optimal one for the family of alternatives under consideration.

We can summarize our point of view saying that the behavior of the projected test is quite encouraging, because the observed loss of power with respect to the optimal test in the case of linear shifts is outweighed by the gain we achieve in the remaining cases.

4.2 One-sample case. More than one random projection

The goal of this sub-section is to show how the consideration of more than one projection leads to an increase of the power of the test, or, in other words, to increase the number of rejections under the alternative. As in Subsection 4.1, under the null hypothesis we will assume that the distribution which produced the sample is of standard Brownian motion. We will analyze the power of the projected tests only for alternatives of a standard Brownian motion with a linear drift because as stated in Subsection 4.1, in this case we know the maximum attainable power.

We will use the statistic D_n^k introduced in Section 3 for this problem. Thus, given $k \in \mathbb{N}$ and a random sample $X_1, \ldots, X_n$ of trajectories, we have to choose at random k vectors $h_1, \ldots, h_k \in \mathcal{H}$ using a non-degenerate gaussian law μ on $\mathcal{H}$, then compute the projections of the sample on the k one-dimensional subspaces which these vectors span, compute the Kolmogorov-Smirnov statistic in all of them and then take the maximum of these values.

As in Subsection 4.1, we will take μ to be the distribution of standard Brownian motion. We have taken $k = 1, 2, 3, 5, 10$ and 25. As for the sample size, we have chosen the worst case obtained in that subsection. Thus, we have fixed the sample size at $n = 50$.

The statistic D_n^k is not distribution-free if $k > 1$, its distribution depending on the vectors chosen and on the distribution we are considering under the null hypothesis. However, it is easy to simulate this distribution because, once the vectors have been chosen, we only need fix $B \in \mathbb{N}$ and, then, repeat for $b = 1, \ldots, B$ the following: produce n trajectories $W_1^b, \ldots, W_n^b$ with standard Brownian–motion distribution and compute for each one the value of the statistic D_n^k. In this way we obtain a random sample, with size B, of the distribution of D_n^k under the null hypothesis. Let us denote these values by $D_n^{k,b}$, $b = 1, \ldots, B$. By the Glivenko-Cantelli theorem, if B is large enough, the empirical distribution based on these values is close to the distribution of D_n^k under the null hypothesis, and, in consequence, we can reject the null hypothesis if the value we obtained for this statistic lies in the upper 5% of the sorted sample.

This procedure is theoretically simple but computationally very expensive. To understand this, it is enough to say that whereas repeating one of the previous simulations 5000 times took about 12 minutes, to complete 1000 repetitions of this procedure on the same computer took about 16 hours. This is the reason that in this case we only include the values for 1000 trials.

The procedure to simulate the discrete approximation of the trajectories is the same as in Section 4.1. The results are reported in Table 4.3. They show

that taking more than one projection increases the power of the proposed test. However, this increase is not spectacular and a greater increase is obtained when going from $k = 1$ to $k = 2$ projections. Further down the table, the increases are not so big and, it seems that going from $k = 10$ to $k = 25$ even decreases the power.

The way to choose an optimal number of random directions remains an open problem. This limited simulation suggests that if we weigh the slim increase in rejections against the big increase in computational burden, it seems that it is not worth taking more than one projection. However, this issue deserves more research in order to figure out if this behavior depends on the specific characteristics of the particular problem considered in the simulation, or if it also appears in other situations.

	$\delta = 0$		$\delta = .25$		$\delta = .5$		$\delta = 1$	
	D_n^k	Np^k	D_n^k	Np^k	D_n^k	Np^k	D_n^k	Np^k
$k = 1$	.057	.057	.261	.317	.730	.849	.965	.969
$k = 2$	.054	.057	.270	.326	.777	.902	.998	.999
$k = 3$	.061	.054	.273	.325	.783	.903	.999	1
$k = 5$	.066	.059	.278	.321	.784	.899	.999	1
$k = 10$	.066	.059	.285	.316	.795	.896	.999	1
$k = 25$	.064	.060	.273	.320	.796	.881	.999	1
$N1$	.053		.437		.940		1	

Table 4.3: Application of proposed procedure to the Brownian process $S(t) = W(t) + \delta t$. The null hypothesis is the standard Brownian motion (i.e. $\delta = 0$). As alternative hypotheses we take $\delta = 0.25, 0.5, 1$. Samples size is fixed at 50. Columns D_n^k (respectively, Np^k) show the results obtained to apply the test based on D_n^k (the normal test) to the k projections for $k = 1, 2, 3, 5, 10$. Last row shows the results obtained with the standard normal test applied to the values $S(1)$. The table shows proportion of rejections along 1000 repetitions excepting for the Normal test where we reproduce the values shown in Table 4.1.

4.3 Two-sample. One random projection

In our next example we consider the two-sample problem. To this end, we have selected some diffusions. Those processes are used very often to represent asset prices because of their nice properties (see, for instance [16]). A diffusion is a solution of the stochastic differential equation

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{4}$$

where W_t, $t \in T$, is a standard Brownian motion and $T \subset \mathbb{R}$ is an interval. Here, we will take $T = [0, 1]$. For suitable selections of b (usually called drift) and σ (the volatility), the solution of the equation (4) belongs to $L_2(T)$ and, therefore, our theory is applicable.

In particular, the model (4) includes the standard Brownian motion and, if we take $b(t, X_t)$ constant, we obtain the Brownian motions with linear drift we treated in Subsections 4.1 and 4.2.

To perform the simulations, we obtained the values for X_t at the points $t_i = i/N,\ i = 0, \ldots, N$, where we chose $N = 100$. Thus, we fixed $X_0 = 0$, and for every $i \geq 1$, we define

$$X_{i+1} = X_i + b(t_i, X_{t_i})\frac{1}{N} + \sigma(t_i, X_{t_i})Z_i \tag{5}$$

where $\{Z_i,\ i = 1, \ldots, N\}$ are independent and identically distributed random variables with centered gaussian distribution with variance N^{-1}. However, in those cases in which $\sigma(t_i, 0) = 0$, we have employed formula (5) only for $i \geq 2$, replacing $\sigma(t_0, X_{t_0})$ by 1 to simulate X_{t_1}, because, otherwise we would have obtained a constant zero trajectory.

Simulations were performed fixing the sample size at 100. We performed 5000 trials for every combination of the selected drifts and volatilities. For the reasons we mentioned in Subsection 4.2, we selected just one random direction to project. The results appear in Table 4.4, where we present the proportion of rejections of the hypothesis that both samples come from the same distribution and, also, the mean of the p-values we obtained over the 5000 trials.

We consider the results obtained very promising, because in most cases we obtain quite high rejection rates.

We were surprised by the fact that values in the main diagonal in Table 4.4 are a bit biased. One would expect the proportion of rejections to be around .05, and all of them (except one which is .049) are above this quantity. Moreover, the mean of the p-values should be .500 and all of them are above this quantity. After analyzing the situation, we consider that the most probable explanation for it is the approximation which MatLab uses for the exact distribution of the two-sample Kolmogorov-Smirnov statistic. We consider that this does not affect substantially the results obtained.

5 An application: Analysis of spectrograms

Proteomics, broadly speaking, is a recently developed family of procedures allowing researchers to analyze proteins. Some of these techniques allows one to

		$\sigma(t, X_t) = 1$ and $b(t, X_t) =$					$b(t, X_t) = 0$ and $\sigma(t, X_t) =$				
		0	t	X_t	$\sin(\pi t)$	$\sin(\pi X_t)$	2	t	X_t	$\sin(\pi t)$	$\sin(\pi X_t)$
		(A.1)	(A.2)	(A.3)	(A.4)	(A.5)	(B.1)	(B.2)	(B.3)	(B.4)	(B.5)
(A.1)	reject	.058									
	aver	.518									
(A.2)	reject	.503	.058								
	aver	.176	.522								
(A.3)	reject	.346	.660	.049							
	aver	.181	.095	.520							
. (A.4)	reject	.887	.346	.886	.054						
	aver	.039	.247	.038	.526						
(A.5)	reject	.434	.755	.181	.923	.055					
	aver	.163	.062	.259	.025	.517					
(B.1)	reject	.871	.932	.204	.979	.676	.052				
	aver	.026	.015	.295	.058	.058	.511				
(B.2)	reject	.952	.997	.995	.997	.976	1	.050			
	aver	.012	.001	.001	.001	.005	0	.529			
(B.3)	reject	1	1	1	1	1	1	.920	.054		
	aver	0	0	0	0	0	0	.026	.525		
(B.4)	reject	.293	.830	.902	.984	.894	.997	.743	.995	.056	
	aver	.221	.039	.024	.005	.037	.002	.061	.001	.518	
(B.5)	reject	1	1	1	1	1	1	.994	854	.999	.053
	aver	0	0	0	0	0	0	.001	051	0	.527

Table 4.4: Application of the two sample procedure to two diffusions solving $dX_t = b(t, x_t)dt + \sigma(t, X_t)dW_t$ for several values of b and σ. The null hypothesis is that both distributions coincide. Sample size is fixed at 100. Table shows proportion of rejections along 5000 repetitions as well as the mean value of the obtained p-values. Key in rows refers to the corresponding case in columns: 'reject' stands for proportion of rejections and 'aver' for the mean of the p-values.

separate mixtures of complex molecules according to their rate mass/charge. A spectrogram is a curve showing the number of molecules (or fragments) found for every mass/charge ratio. Thus, molecules with the same mass/charge ratio are indistinguishable by a spectrogram. Taking into account that, in principle, the ratios can take every positive value, a spectrogram is a $\mathbb{N}$-valued map defined on $\mathbb{R}^+$.

The idea behind the application of the spectrograms is that, when a cancer starts to grow, its cells produce a different kind of proteins than those produced by healthy cells. Moreover, the amount of commonly produced proteins may be different. Figure 1 shows spectrograms of a healthy woman (right-hand graph) and of a woman suffering from ovarian-cancer (left-hand graph). In this case, the observation interval was [699.99, 12000]; i.e., the spectrograms, instead of being

defined on $\mathbb{R}^+$, are defined on this interval. However, in the measurement process, some discretization is required and, at the end, the obtained data belonged to the space $\mathbb{R}^{373401}$. Two modifications have been done on the original data before to start our analysis. First, they were normalized to take values between 0 and 1. The value 0 already was observed in every women. Thus, the modification consisted just in dividing every trajectory by its maximum. Then, in order to lower the computational requirements, we reduced the data to be defined on $\mathbb{R}^{14936}$ by adding together the proteins found in the first 25 coordinates, then the proteins found in the following 25 coordinates,... The last coordinate was rejected.

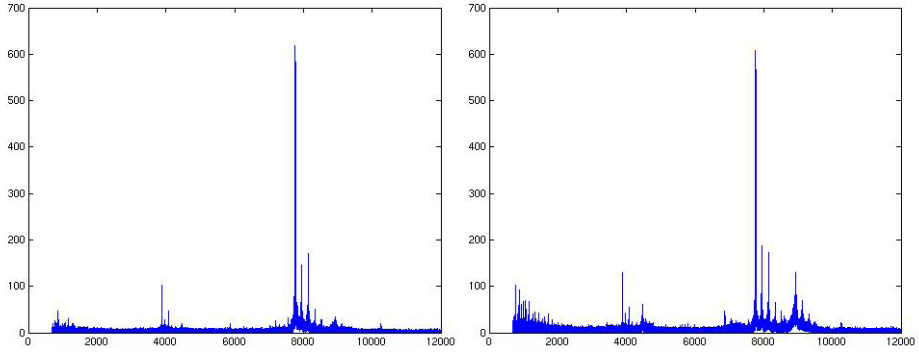

Figure 1: Spectrograms of a healthy woman(left-hand graph) and of a woman suffering from ovarian cancer (right-hand graph).

More details on spectrograms and statistically related aspects of their analysis, can be found in the December 2003 issue of *Chance*, (see [3]). The same data we are studying here were analyzed from the classification point of view in [1] and [4].

The analysis we have carried out here has been the following. As stated in the introduction, the data are 216 spectrograms. 95 of them corresponding to healthy women and 121 to women suffering from ovarian cancer.

A possibility which we must reject from the outset is that the data contains so great a variability that, if we take two samples at random, then the null hypothesis of being produced from a common distribution is rejected. We mean that, it should not happen that, if we take two disjoint groups of healthy women, or two disjoint groups of cancer-suffering women, then the null hypothesis of being produced from a common distribution is rejected. Once this step is fixed, we can safely assume that the spectrogram from the healthy women (respectively, cancer-suffering women) are a random sample of an unknown distribution and

we can test the null hypothesis that both distributions coincide. According to the theory exposed above, this null hypothesis should be rejected.

In order to check these hypothesis, we did the following. We split the healthy women into two groups at random, the first one with 48 members and the second one with 47. These groups are called Healthy1 and Healthy2. Then we split, also at random, the ill women into a group with 61 women, called Cancer1, and a second one with 60, the group Cancer2. Finally, we chose, using the distribution of a standard Brownian motion, a vector $h \in L^2[0, 1]$ at random (to be precise, the approximation to such a vector employing the points $t_i = i/N$, $i = 0, \ldots, N$ where $N = 14936$), we have computed the projections and we have checked the following series of null hypotheses

H_0^1: the distributions which generated samples in groups Healthy1 and Healthy2 coincide

H_0^2: the distributions which generated samples in groups Healthy1 and Cancer1 coincide

H_0^3: the distributions which generated samples in groups Healthy1 and Cancer2 coincide

H_0^4: the distributions which generated samples in groups Healthy2 and Cancer1 coincide

H_0^5: the distributions which generated samples in groups Healthy2 and Cancer2 coincide

H_0^6: the distributions which generated samples in groups Cancer1 and Cancer2 coincide.

We repeated this procedure 1000 times. According to the theory, the hypotheses H_0^1 and H_0^6 should be rejected around 5% of the time, and the mean of the p-values should be around 0.500. Moreover, the remaining hypotheses should be rejected a high proportion of times and the mean of the p values should be very low.

These results are shown in Table 5.1, and, roughly speaking, they coincide with what we expect.

Acknowledgments. The first author gave a invited talk in the 9ª Escola Brasileira de Probabilidade, held at São Sebastião from 31th July to 6th August 2005, which was mainly based on the content of this paper. He wants to express here

	Healthy1	Healthy2	Cancer1
Healthy2	.048 (.525)		
Cancer1	.958 (.015)	.954 (.015)	
Cancer2	.951 (.013)	.952 (.013)	.045 (.512)

Table 5.1: Application of the two sample procedure to check, first, the homogeneity of the spectrograms from healthy (respectively cancer-suffering) women, and, then, whether the distributions which generated the spectrograms of healthy and cancer suffering women coincide. In every cell it is checked the hypothesis that the distribution which generated the spectrograms in the groups which determine it coincide. Table shows proportion of rejections as well as, in parenthesis, the mean of p-values over 1000 trials in which both populations were split, at random, in two groups.

his deep acknowledgment to the organizing committee of the school, especially to its chairman, Prof. Antonio Galves, for the warm hospitality he received there; and, also, to congratulate them for the nice atmosphere they achieved in these schools which must have been very benefical for the participating students.

All the authors want to thank to Prof. E. del Barrio (Universidad de Valladolid) for his kindness in providing us with the content of [2].

We also want to thank to an anonymous referee who pointed us the reference [13] and also pointed out some interesting comments on the paper.

References

[1] D. Banks and E. Petricoin, Finding cancer signals in mass spectrometry data. *Chance* **16**(4) (2003), 8–12 and 57.

[2] E del Barrio, Personal communication (2006).

[3] Chance, Issue devoted to Proteomics, **16** (2003), 4.

[4] J.A. Cuesta-Albertos and R. Fraiman, Impartial Trimmed k-means and Classification Rules for Functional Data. *Preprint.* (2005).

[5] J.A. Cuesta-Albertos, R. Fraiman and T. Ransford, A sharp form of the Cramér-Wold theorem. To appear in *J. Theoret. Probab.* (2006).

[6] S. Dasgupta and A. Gupta, An elementary proof of the Johnson-Lindenstrauss Lemma. (1999). Technical report TR-99-006. International Computer Science Institute, Berkeley.

[7] W.M. Gilbert, Projections of probability distributions. *Acta Math. Acad. Sci. Hungar.*, **6** (1955), 195–198.

[8] G. Gldófalvi, Random projection for non-Gaussian mixture models. Downloaded from http://www-cse.ucsd.edu/elkan/254spring02/gidofalvireport.pdf. (2002).

[9] A.N. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione. *Giorn. dell'Instituto Ital. degli Attuari.*, **4** (1933), 83–91.

[10] A.N. Kolmogorov, Confidence limits for an unknown distribution function. *Ann. Math. Stat.*, **12** (1941), 461–463.

[11] R.G. Laha and V.K. Rohatgi, *Probability Theory*, Wiley, New York (1979).

[12] H. Lejsek, F.H. Ásmundsson, B.P. Jónson and L. Amsaleg, Efficient and effective image copyright enforcement. *Publication Interne*, **1699** (2005), IRISA, Rennes.

[13] D.M. Mason and J.H. Schuenemeyer, A modified Kolmogorov–Smirnov test sensitive to tail alternatives. *Ann. Statist.*, **11** (1983), 933–946.

[14] F.J. Massey Jr., The distribution of the maximum deviation between two sample cumulative step functions. *Ann. Math. Stat.*, **22** (1951), 125–128.

[15] F.J. Massey Jr., Distribution table for the deviation between two sample cumulatives. *Ann. Math. Stat.*, **23** (1952), 435–441.

[16] R.C. Merton, *Continuous time finance.* Blackwell, Cambridge (1990).

[17] A. Rényi, On projections of probability distributions *Acta Math. Acad. Sci. Hungar.*, **3** (1952), 131–132.

[18] J.A. Shohat and J.D. Tamarkin, *The Problem of Moments*, Amer. Math. Soc., Providence, RI (1943).

[19] N.V. Smirnov, Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow Univ. 2*, **2** (1939), 3–16.

[20] N.V. Smirnov, Table for estimating the goodness-of-fit of empirical distributions. *Ann. Math. Stat.*, **19** (1948), 279–281.

[21] S.S. Vempala, *The Random Projection Method.* American Mathematical Society, Rhode Island (2004).

[22] S.S. Vempala and G. Wang, A spectral algorithm for learning mixture models. *J. Computer and System Sciences*, **68** (2004), Issue 4, 841–860.

Juan Antonio Cuesta-Albertos
Departamento de Matemáticas
Estadística y Computación
Universidad de Cantabria
SPAIN

E-mail: cuestaj@unican.es

Ricardo Fraiman
Departamento de Matemática y Ciencias
Universidad de San Andrés
ARGENTINA

E-mail: rfraiman@udesa.edu.ar

Thomas Ransford
Département de Mathématiques et de Statistique
Université Laval
Québec (QC)
CANADA G1K 7P4

E-mail: ransford@mat.ulaval.ca