A SHARP FORM OF THE CRAMÉR–WOLD THEOREM

JUAN ANTONIO CUESTA-ALBERTOS, RICARDO FRAIMAN, AND THOMAS RANSFORD

Abstract. The Cramér–Wold theorem states that a Borel probability measure P on $\mathbb{R}^d$ is uniquely determined by its one-dimensional projections. We prove a sharp form of this result, addressing the problem of how large a subset of these projections is really needed to determine P. We also consider extensions of our results to measures on a separable Hilbert space. As an application of these ideas, we derive a simple, universally consistent goodness-of fit-test for data taking values in a Hilbert space.

1. Introduction

Let P be a Borel probability measure on $\mathbb{R}^d$, where $d \geq 2$. The Cramér–Wold theorem [2, p.291] states that P is uniquely determined by its one-dimensional projections. This paper addresses the problem of how large a subset of these projections is really needed to determine P.

In the case $d = 2$, Rényi [10, Theorem 1] proved that, provided that P is supported on a bounded subset of $\mathbb{R}^2$, it is determined by any infinite set of its one-dimensional projections. Gilbert [3, Theorem 1] subsequently extended this result by showing that the same conclusion holds if we merely assume that P has finite moments satisfying the Carleman condition.

When $d \geq 3$, this is no longer true: not every infinite set of one-dimensional projections suffices to determine P, even when P is compactly supported. For example, if $d = 3$, then all probability measures supported on the z-axis have the same image (a point mass at the origin) under projection onto any line in the xy plane.

So how large a set of one-dimensional projections is needed to determine P in general? We give a rather precise answer this question in §3, by formulating and proving a sharp form of the Cramér–Wold theorem, valid for all $d \geq 2$. When $d = 2$, it reduces to the theorem of Gilbert, mentioned above.

In §4, we extend our results to the case of a separable, infinite-dimensional Hilbert space.

Finally, in §5, we present an application of these ideas to derive a universally consistent Kolmogorov–Smirnov goodness-of-fit test for data taking values in a Hilbert space. We emphasize that our aim in this section is just to give a idea about how the results in the previous sections can be applied to obtain sound statistical procedures; we do not try here to optimize them.

2. Preliminaries

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

2000 *Mathematics Subject Classification.* Primary 60E05; Secondary 28C20, 60B11, 62H15.

Key words and phrases. Probability measures, projections, Cramér-Wold theorem, Hilbert spaces, Kolmogorov–Smirnov goodness-of-fit test.

First author partially supported by the Spanish Ministerio de Ciencia y Tecnología, grant BFM2002-04430-C02-02.

Second author partially supported by Instituto de Cooperación Iberoamericana, Programa de Cooperación Interuniversitaria AL-E 2003.

Third author partially supported by grants from NSERC and the Canada research chairs program.

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 a closed subspace L of $\mathcal{H}$, we denote by $\pi_L:\mathcal{H}\to L$ the orthogonal projection of $\mathcal{H}$ onto L. Also, given any subset S of $\mathcal{H}$, we write $S^\perp$ for set of vectors orthogonal to S.

Let P be a Borel probability measure on $\mathcal{H}$. Its characteristic function $\phi_P:\mathcal{H}\to\mathbb{C}$ is given by

$$\phi_P(x):=\int e^{i\langle x,y\rangle}\,dP(y)\qquad(x\in\mathcal{H}).$$

It is well known that P is uniquely determined by its characteristic function [6, Proposition 7.4.1]. Given a closed subspace L of $\mathcal{H}$, we denote by P_L the projection of P onto L, namely the probability measure on L given by

$$P_L(B):=P\big(\pi_L^{-1}(B)\big)\qquad(\text{Borel } B\subset L).$$

A simple calculation shows that $\phi_{P_L}(x)=\phi_P(x)$ for all $x\in L$.

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

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

where $\langle x\rangle$ denotes the one-dimensional subspace spanned by x. The set $\mathcal{E}(P,Q)$ will play a central role in what follows. It is obvious that $\mathcal{E}(P,Q)$ is a cone, i.e. a union of one-dimensional subspaces of $\mathcal{H}$. The following proposition gives a simple characterization of $\mathcal{E}(P,Q)$ in terms of characteristic functions.

Proposition 2.1. *With the above notation,*

$$\mathcal{E}(P,Q)=\{x\in\mathcal{H}:\phi_P(tx)=\phi_Q(tx)\ \text{for all}\ t\in\mathbb{R}\}. \tag{1}$$

Proof. If $x\in\mathcal{H}$, then $\phi_{P_{\langle x\rangle}}(tx)=\phi_P(tx)$ $(t\in\mathbb{R})$. The result thus follows from the uniqueness theorem for characteristic functions. □

Corollary 2.2. *$\mathcal{E}(P,Q)$ is closed in $\mathcal{H}$.*

Proof. This follows easily from (1), using the dominated convergence theorem. □

Another consequence is the Cramér–Wold theorem for $\mathcal{H}$.

Corollary 2.3. *If $\mathcal{E}(P,Q)=\mathcal{H}$, then $P=Q$.*

Proof. If $\mathcal{E}(P,Q)=\mathcal{H}$, then from (1) we get $\phi_P=\phi_Q$, and hence $P=Q$. □

Remarks. (i) Combining the two corollaries, we see that if $\mathcal{E}(P,Q)$ is dense in $\mathcal{H}$, then $P=Q$.

(ii) The Cramér–Wold theorem can viewed as a simple form of uniqueness theorem for the Radon transform, applied to measures rather than functions.

3. A sharp Cramér–Wold theorem in $\mathbb{R}^d$

As mentioned in the introduction, a compactly supported Borel probability measure on $\mathbb{R}^2$ is determined by its projections onto any infinite set of lines, but the same is no longer true in $\mathbb{R}^d$ when $d\ge 3$. We begin this section by formulating the 'correct' condition.

A polynomial p on $\mathbb{R}^d$ is called *homogeneous of degree m* if $p(tx)=t^mp(x)$ 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 $S=\{x\in\mathbb{R}^d:p(x)=0\}$.

The following result is a sharp form the Cramér–Wold theorem for $\mathbb{R}^d$.

Theorem 3.1. *Let P, Q be Borel probability measures on $\mathbb{R}^d$, where $d \geq 2$. Assume that:*

- *the absolute moments $m_n := \int \|x\|^n \, dP(x)$ are finite and satisfy $\sum_{n\geq 1} m_n^{-1/n} = \infty$;*
- *the set $\mathcal{E}(P, Q)$ is not contained in any projective hypersurface in $\mathbb{R}^d$.*

Then $P = Q$.

Remarks. (i) The condition $\sum_{n\geq 1} m_n^{-1/n} = \infty$ is known as the Carleman condition. A probability measure satisfying this condition is uniquely determined by its moments [11, p.19]. If P has a finite moment generating function in a neighbourhood of the origin, then it automatically satisfies the Carleman condition (but not conversely).

(ii) The Carleman condition is imposed only on P, not on Q. Thus, in the language of [1, §4.1], the theorem is a 'strong determination' result. This will be important for the statistical application in §5.

(iii) The condition that $\mathcal{E}(P, Q)$ not be contained in any projective hypersurface is equivalent to asking that $\mathcal{E}(P, Q)$ be dense in $\mathbb{R}^d$ with respect to the projective Zariski topology (see [9, p.50 and p.81]) (compare this with the remark at the end of §2).

(iv) When $d = 2$, the condition on $\mathcal{E}(P, Q)$ is equivalent to demanding that it contain an infinite number of lines, and thus, in this case, Theorem 3.1 reduces to Gilbert's theorem mentioned in the Introduction.

(v) Both conditions in Theorem 3.1 are sharp, in a sense to be made precise at the end of the section.

Proof of Theorem 3.1. By hypothesis, the absolute moments of P are finite. We begin by showing that the same is true of Q. Fix $n \geq 0$, and set

$$F := \Big\{ x \in \mathbb{R}^d : \int |\langle x, y\rangle|^n \, dQ(y) < \infty \Big\}.$$

Since $|\langle x, y\rangle|^n$ is a convex function of x, it is easy to see that F is a subspace of $\mathbb{R}^d$. Further, if $x \in \mathcal{E}(P, Q)$, then

$$\int |\langle x, y\rangle|^n \, dQ(y) = \int |t|^n \, dQ_{\langle x\rangle}(t) = \int |t|^n \, dP_{\langle x\rangle}(t) = \int |\langle x, y\rangle|^n \, dP(y) < \infty. \tag{2}$$

It follows that $\mathcal{E}(P, Q) \subset F$. If F were a proper subspace of $\mathbb{R}^d$, then we could find a non-zero $z \in F^\perp$, and so $\mathcal{E}(P, Q)$ would be in the zero set of the linear polynomial $p(x) := \langle x, z\rangle$, contrary to hypothesis. Therefore $F = \mathbb{R}^d$. Hence, writing $e_1, \dots, e_d$ for the standard unit vector basis of $\mathbb{R}^d$, we have

$$\int \|y\|^n \, dQ(y) = \int \Big(\sum_{j=1}^d |\langle e_j, y\rangle|^2 \Big)^{n/2} dQ(y) \leq d^{n/2} \sum_{j=1}^d \int |\langle e_j, y\rangle|^n \, dQ(y) < \infty,$$

as claimed.

Now fix $n \geq 0$ once again, and consider

$$p(x) := \int \langle x, y\rangle^n \, dP(y) - \int \langle x, y\rangle^n \, dQ(y) \qquad (x \in \mathbb{R}^d).$$

Clearly p is a homogeneous polynomial, and a similar calculation to (2) shows that $p(x) = 0$ for all $x \in \mathcal{E}(P, Q)$. By our assumption about $\mathcal{E}(P, Q)$, this is possible only if $p(x) = 0$ for all $x \in \mathbb{R}^d$. Moreover this holds for every $n \geq 0$. Thus P and Q have exactly the same moments. As P satisfies the Carleman condition, it is uniquely determined by its moments, and so we conclude that $P = Q$, as desired. □

Corollary 3.2. *Let P, Q be Borel probability measures on $\mathbb{R}^d$, where $d \geq 2$. Assume that:*

- *the absolute moments $m_n := \int \|x\|^n \, dP(x)$ are finite and satisfy $\sum_{n\geq 1} m_n^{-1/n} = \infty$;*
- *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 3.1, because every projective hypersurface is of Lebesgue measure zero in $\mathbb{R}^d$. □

Several authors have also considered determination of probability measures on $\mathbb{R}^d$ by their projections onto hyperplanes (see [1] and the references cited therein). Of course, when $d = 2$, the hyperplane projections are just the one-dimensional projections. However, if $d \geq 3$, then the one-dimensional projections are in some sense 'finer'. Indeed, by the original Cramér–Wold theorem, $P_L = Q_L \iff L \subset \mathcal{E}(P, Q)$. Using this remark, we can give a simple proof of the following result, which was already known (see e.g. [1, Theorem 4.11]).

Corollary 3.3. *Let P, Q be Borel probability measures on $\mathbb{R}^d$, where $d \geq 2$. Assume that:*

- *the absolute moments $m_n := \int \|x\|^n \, dP(x)$ are finite and satisfy $\sum_{n\geq 1} m_n^{-1/n} = \infty$;*
- *$P_L = Q_L$ for infinitely many hyperplanes L in $\mathbb{R}^d$.*

Then $P = Q$.

Proof. Again, this is an immediate consequence of Theorem 3.1, because a projective hypersurface in $\mathbb{R}^d$ can contain at most finitely many hyperplanes. □

In view of this result, it is tempting to conjecture that the hypersurface condition in Theorem 3.1 can be replaced by the weaker assumption that $\mathcal{E}(P, Q)$ is not contained in any finite union of hyperplanes. The following very simple example shows that this conjecture is false.

Example 3.4. *There exist probability measures P, Q on $\mathbb{R}^3$ such that:*

- *the moment generating functions of P and Q are finite everywhere,*
- *the set $\mathcal{E}(P, Q)$ is not contained in any finite union of hyperplanes,*

but $P \neq Q$.

Proof. Let X, Y be independent standard normal random variables. Let P and Q be the distributions of three-dimensional random vectors $(X, Y, 0)$ and $(X, -X, Y)$ respectively. Evidently, the moment generating functions of P, Q are finite everywhere.

Let $x = (x_1, x_2, x_3)$ be a unit vector in $\mathbb{R}^3$. Then $P_{\langle x\rangle}$ and $Q_{\langle x\rangle}$ are centered gaussian distributions with variances $x_1^2 + x_2^2$ and $(x_1 - x_2)^2 + x_3^2$ respectively. Thus

$$(x_1, x_2, x_3) \in \mathcal{E}(P, Q) \iff x_1^2 + x_2^2 = (x_1 - x_2)^2 + x_3^2 \iff 2x_1x_2 = x_3^2.$$

As $\mathcal{E}(P, Q)$ is a cone, it follows that

$$\mathcal{E}(P, Q) = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : 2x_1x_2 = x_3^2\}.$$

It is an elementary exercise to check that this set is not contained in any finite union of hyperplanes. Finally, $P \neq Q$ because $\mathcal{E}(P, Q) \neq \mathbb{R}^3$. □

Remark. The degeneracy in the preceding example is not important. Indeed, if we consider four independent standard one-dimensional normal random variables X, Y, Z, W, and take P, Q to be the distributions of the vectors $(X, W+Y, Z)$ and $(X, W-X, Y+Z)$, then we are in exactly the same situation as in the example, and P and Q are both non-degenerate gaussian distributions.

By employing some harmonic analysis, we can go rather further and show that the hypersurface condition in Theorem 3.1 is sharp, in a sense made precise by the following theorem.

Theorem 3.5. *Let S be a projective hypersurface in $\mathbb{R}^d$. Then there exist Borel probability measures P, Q on $\mathbb{R}^d$ such that*

- *both P and Q are supported on bounded subsets of $\mathbb{R}^d$,*
- *$\mathcal{E}(P, Q) = S$,*

but P and Q are mutually singular.

Proof. The proof depends on an auxiliary function, $f : \mathbb{C}^d \to \mathbb{C}$, defined by

$$f(z) := \prod_{j=1}^{d} \Big(\frac{\sin z_j - z_j}{z_j^3}\Big) \qquad (z := (z_1, \dots, z_d) \in \mathbb{C}^d).$$

It is elementary to check that f has the following properties:

(i) f is an even entire function which is real-valued on $\mathbb{R}^d$;
(ii) $|f(z)| \le \text{const.} \exp(\sum_1^d |z_j|)$ on $\mathbb{C}^d$;
(iii) $|f(x)| \le \text{const.}/(1 + \|x\|^2)$ on $\mathbb{R}^d$;
(iv) $f(0) \neq 0$.

By definition of projective hypersurface, $S = \{x \in \mathbb{R}^d : p(x) = 0\}$, where p is a homogeneous polynomial, not identically zero. Define $g : \mathbb{R}^d \to \mathbb{R}$ by $g(x) := p(x)^2 f(x)^N$, where N is a positive integer, chosen large enough so that $g \in L^2(\mathbb{R}^d)$ (this is possible, by (iii) above). Let $h = \widehat{g}$, the Fourier transform of g. By Plancherel's theorem $h \in L^2(\mathbb{R}^d)$, and h is real-valued since g is even and real-valued. Moreover, since g is the restriction to $\mathbb{R}^d$ of an entire function of exponential type (namely $p(z)^2 f(z)^N$), the Paley–Wiener theorem [14, Theorem 4.9] tells us that h is supported on a compact subset of $\mathbb{R}^d$.

Define finite positive Borel measures on $\mathbb{R}^d$ by

$$P(B) := \int_B h^+(x)\,dx \quad \text{and} \quad Q(B) := \int_B h^-(x)\,dx \qquad (\text{Borel } B \subset \mathbb{R}^d).$$

Clearly P and Q are compactly supported and mutually singular. Also, using the Fourier inversion theorem, their characteristic functions satisfy

$$\phi_P(x) - \phi_Q(x) = cg(x) = cp(x)^2 f(x)^N \qquad (x \in \mathbb{R}^d), \tag{3}$$

where c is a non-zero constant. In particular,

$$P(\mathbb{R}^d) - Q(\mathbb{R}^d) = \phi_P(0) - \phi_Q(0) = cp(0)^2 f(0)^N = 0,$$

so, multiplying by a constant if necessary, we can arrange that P, Q are both probability measures. Also, it follows from (1) and (3) that

$$\mathcal{E}(P, Q) = \{x \in \mathbb{R}^d : p(tx)^2 f(tx)^N = 0 \text{ for all } t \in \mathbb{R}\}.$$

As p is homogeneous and $f(0) \neq 0$, we deduce that $\mathcal{E}(P, Q) = \{x \in \mathbb{R}^d : p(x) = 0\} = S$. □

Finally, for the record, we state a theorem showing that the first condition in Theorem 3.1 (the Carleman condition) is also sharp.

Theorem 3.6. *Let C be a proper closed cone in $\mathbb{R}^d$, and let $(M_n)_{n\geq 0}$ be a positive sequence satisfying*

$$M_0 = 1, \qquad M_n^2 \leq M_{n-1}M_{n+1}\ (n \geq 1) \qquad \text{and} \qquad \sum_{n\geq 1} M_n^{-1/n} < \infty.$$

Then there exist Borel probability measures P and Q on $\mathbb{R}^d$ such that

- *both $\int \|x\|^n\, dP(x) \leq M_n$ and $\int \|x\|^n\, dQ(x) \leq M_n$, for all $n \geq 0$,*
- *the set $\mathcal{E}(P,Q)$ contains C,*

but P and Q are mutually singular.

Proof. This is just a slight restatement of [1, Theorem 5.4]. □

4. Extensions to infinite dimensions

In this section, we shall show that both the Corollaries 3.2 and 3.3 have rather natural extensions to infinite dimensions. In the case of Corollary 3.2, since Lebesgue measure no longer makes sense in infinite dimensions, we shall use gaussian measures instead.

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 projections is gaussian. It is *non-degenerate* if, in addition, each of its one-dimensional projections is non-degenerate. If μ is gaussian, then its characteristic function has the form

$$\phi_\mu(x) = \exp\bigl(i\langle a, x\rangle - \tfrac{1}{2}\langle Sx, x\rangle\bigr) \qquad (x \in \mathcal{H}), \tag{4}$$

where $a \in \mathcal{H}$ (the mean of μ) and S is a positive, trace-class operator on $\mathcal{H}$ (the covariance operator of μ). For more details, see e.g. [6, §7.5 and §7.6].

The following result is the infinite-dimensional generalization of Corollary 3.2.

Theorem 4.1. *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:*

- *the absolute moments $m_n := \int \|x\|^n\, dP(x)$ are finite and satisfy $\sum_{n\geq 1} m_n^{-1/n} = \infty$;*
- *the set $\mathcal{E}(P,Q)$ is of positive μ-measure.*

Then $P = Q$.

Proof. Let S be the covariance operator of μ. By the spectral theorem, S has an orthonormal basis of eigenvectors $(e_n)_{n\geq 1}$. For each $n \geq 1$, let F_n be the linear span of $\{e_1, \dots, e_n\}$, and let μ_n and ν_n be the projections of μ onto F_n and $F_n^\perp$ respectively. Then μ_n, ν_n are non-degenerate gaussian measures on F_n and $F_n^\perp$ respectively, and $\mu = \mu_n \otimes \nu_n$, their product measure (this is simply a restatement of the familiar fact that uncorrelated gaussian random variables are independent).

Fix $n \geq 1$. By Fubini's theorem,

$$\mu\bigl(\mathcal{E}(P,Q)\bigr) = \int_{F_n^\perp} \mu_n\bigl(\mathcal{E}(P,Q)^x\bigr)\, d\nu_n(x),$$

where $\mathcal{E}(P,Q)^x$ denotes the x-section of $\mathcal{E}(P,Q)$, i.e. the set of $y \in F_n$ such that $x + y \in \mathcal{E}(P,Q)$. Since $\mu(\mathcal{E}(P,Q)) > 0$, there exists $x \in F_n^\perp$ such that $\mu_n(\mathcal{E}(P,Q)^x) > 0$. As ν_n is non-degenerate, we can suppose that $x \neq 0$. As μ_n is non-degenerate, it is absolutely continuous with respect to Lebesgue measure on F_n, and so $\mathcal{E}(P,Q)^x$ is a set of positive n-dimensional Lebesgue measure.

As $\mathcal{E}(P,Q)$ is a cone, it follows that $\mathcal{E}(P,Q)^{tx}$ is also of positive n-dimensional Lebesgue measure, for each $t \in \mathbb{R} \setminus \{0\}$. Therefore $\mathcal{E}(P,Q) \cap G$ is of positive $(n+1)$-dimensional Lebesgue measure, where G is the linear span of $\{e_1, \dots, e_n, x\}$. By Corollary 3.2, we deduce that $P_G = Q_G$. In particular, since $F_n \subset G$, we obtain $P_{F_n} = Q_{F_n}$. This implies that $\phi_P = \phi_Q$ on F_n. Finally, since $\cup_{n \geq 1} F_n$ is dense in $\mathcal{H}$ and ϕ_P, ϕ_Q are continuous, it follows that $\phi_P = \phi_Q$ on $\mathcal{H}$, and thus $P = Q$, as desired. $\square$

We now present the infinite-dimensional generalization of Corollary 3.3. In this context, hyperplane should be taken to mean closed subspace of codimension one.

Theorem 4.2. *Let P, Q be Borel probability measures on a separable Hilbert space $\mathcal{H}$. Assume that:*

- *the absolute moments $m_n := \int \|x\|^n \, dP(x)$ are finite and satisfy $\sum_{n \geq 1} m_n^{-1/n} = \infty$;*
- *$P_L = Q_L$ for infinitely many hyperplanes L in $\mathcal{H}$.*

Then $P = Q$.

For the proof, we need a simple lemma. Recall that, given a closed subspace F of a Hilbert space $\mathcal{H}$, we write $\pi_F : \mathcal{H} \to F$ for the orthogonal projection of $\mathcal{H}$ onto F.

Lemma 4.3. *Let $\mathcal{H}$ be a Hilbert space, and let $(L_k)_{k \geq 1}$ be distinct hyperplanes in $\mathcal{H}$. Then there exists a two-dimensional subspace F of $\mathcal{H}$ such that $(F \cap L_k)_{k \geq 1}$ are distinct hyperplanes in F.*

Proof. For each k, there exists $x_k \in \mathcal{H} \setminus \{0\}$ such that $L_k = x_k^\perp$. Given a closed subspace F of $\mathcal{H}$, the sets $F \cap L_k$ and $F \cap L_l$ are distinct hyperplanes in F if and only if the pair $\{\pi_F(x_k), \pi_F(x_l)\}$ is linearly independent. In particular, if F is two-dimensional, say $F =$ the linear span of $\{y, z\}$, then

$$F \cap L_k \neq F \cap L_l \iff \begin{vmatrix} \langle x_k, y \rangle & \langle x_k, z \rangle \\ \langle x_l, y \rangle & \langle x_l, z \rangle \end{vmatrix} \neq 0.$$

Given k, l with $k \neq l$, let U_{kl} denote the set of pairs $(y, z) \in \mathcal{H} \times \mathcal{H}$ for which the determinant on the right-hand side is non-zero. Then U_{kl} is a dense open subset of $\mathcal{H} \times \mathcal{H}$. By the Baire category theorem, it follows that $\cap_{k,l} U_{kl}$ is non-empty. Pick a pair (y, z) in this intersection, and let F be the linear span of $\{y, z\}$. Then F has the property stated in the lemma. $\square$

Proof of Theorem 4.2. Let $(L_k)_{k \geq 1}$ be a sequence of distinct hyperplanes such that $P_{L_k} = Q_{L_k}$ for all $k \geq 1$. Let F be a two-dimensional subspace as in the statement of the lemma. Pick an orthonormal basis $(e_n)_{n \geq 1}$ of $\mathcal{H}$ such that F is spanned by $\{e_1, e_2\}$. For each $n \geq 2$, let F_n be the linear span of $\{e_1, \dots, e_n\}$. Then $(F_n \cap L_k)_{k \geq 1}$ is a family of distinct hyperplanes in F_n, and $P_{F_n \cap L_k} = Q_{F_n \cap L_k}$ for all k. By Corollary 3.3, it follows that $P_{F_n} = Q_{F_n}$. Finally, just as in the proof of Theorem 4.1, we conclude that $P = Q$. $\square$

5. 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 $X_1, \dots, X_n$ 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 [4] suggested using the statistic

$$D_n := \sup_{t\in\mathbb{R}} |F_n(t) - F_0(t)|, \tag{5}$$

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 [13] and Massey [7, 8], and is available in most statistical packages. Kolmogorov [4] 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 [12] and Kolmogorov [5] treated the two-sample problem with similar techniques. Here, we have two independent random samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$, 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

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

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 $x_1, \dots, x_n$, the rank R_i of x_i is the place that x_i occupies in the ordered vector $x^{(1)} < \dots < x^{(n)}$ obtained by ordering the original vector, i.e. $x_i = x^{(R_i)}$.)

To help understand why extensions to higher dimensions are of interest, we remark that recent advances in modern 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 $\{X_1(t), \dots, X_n(t) : t \in T\}$ of trajectories with values in the Hilbert space $L^2(T)$, where T is an interval in $\mathbb{R}$.

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 $X_1, \dots, X_n$ 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 $H_0 : P_X = P_0$ in favor of the alternative $H_A : P_X \neq P_0$.

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 projections of the data onto the one-dimensional subspace $\langle h \rangle$. Thus, according to (5), we compute the statistic

$$D_n(h) := \sup_{t \in \mathbb{R}} |F_n^h(t) - F_0^h(t)|, \tag{6}$$

where now

$$F_0^h(t) := P_0\{x \in \mathcal{H} : \langle x, h \rangle \leq t\} \quad \text{and} \quad F_n^h(t) := \frac{1}{n} \sum_{i=1}^n I_{(-\infty,t]}(\langle X_i, h \rangle) \qquad (t \in \mathbb{R}),$$

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

The properties of the proposed procedure are summarized in the following theorem. Recall that a probability measure P on a separable Hilbert space $\mathcal{H}$ is said to satisfy the Carleman condition if the absolute moments $m_n := \int \|x\|^n \, dP(x)$ are finite and satisfy $\sum_{n \geq 1} m_n^{-1/n} = \infty$. Also, 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 5.1. *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* (6)*.*

(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*

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

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

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

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 of Theorem 5.1. (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 $\langle X_1, h\rangle, \dots, \langle X_n, h\rangle$ is exactly F_n^h. Therefore this part follows by the standard properties of the one-dimensional Kolmogorov–Smirnov test.

(b) By Theorem 4.1, if $Q \neq P_0$, then, for μ-almost all $h \in \mathcal{H}$, there exists $t_h \in \mathbb{R}$ such that

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

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

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

where $G^h(t) := Q\{x \in \mathcal{H} : \langle x, h\rangle \leq t\}$. 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 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 $h_1, \dots, h_k$ at random, and then consider as statistic $D_n^k := \max_{1\leq i\leq k} D_n(h_i)$, 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, $\{X_{m_1}, \dots, X_{m_{n_i}}\}$, $i = 1, \dots, k$, select k independent directions $\{h_1, ..., h_k\}$ at random, then, for each $i = 1, \dots, k$, compute the one-dimensional Kolmogorov–Smirnov statistic of the projection of the subsample $\{X_{m_1}, \dots, X_{m_{n_i}}\}$ 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.

The two-sample problem can be treated in a very similar way. Let us assume that our data are independent, identically distributed realizations $\{X_1, \dots, X_n\}$, $\{Y_1, \dots, Y_m\}$ 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 $H_0 : P_X = P_Y$ and the alternative $H_A : P_X \neq P_Y$, based on $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$.

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:

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

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.

We conclude with an example to show how the test works in practice. We confine ourselves to the one-sample problem, the other one being similar.

In our example, we take $\mathcal{H} = L^2[0,1]$, and the distribution P_0 in the null hypothesis is that of the standard Brownian motion W on $[0,1]$. According to our procedure, we have to choose the random vector $h \in \mathcal{H}$ using a non-degenerate gaussian law μ on $\mathcal{H}$. To ease the computations, we also take μ to be the standard Brownian motion.

Now, we should generate a random sample $W_1, \ldots, W_n$ from the Brownian motion we are considering. However, according to the previous results, we only need to consider the scalar products $\langle W_i, h\rangle$, and it happens that 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.

For the sake of analyzing the behavior of our test under the alternative, we shall consider the shifted Brownian processes $S(t) := W(t) + \delta t$, where $\delta \neq 0$. In this case, the distribution of $\langle S, h\rangle$ is also normal, with the same variance as before, but with mean given by

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

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

Notice that, if we were to fix this family of alternatives, then the problem could also be handled by testing the null hypothesis H_0: 'the distribution of $W(1)$ is $N(0,1)$', against H_A: 'the distribution of $W(1)$ is $N(\delta, 1)$ for some $\delta \neq 0$'. We would just have to perform a well-known test based on the normal distribution. However, it should be recalled that the alternative we are actually considering is that the distribution of the empirical process is different from that of standard Brownian motion, and this includes many processes $X(t)$ such that the distribution of $X(1)$ is $N(0,1)$, for which the normal-test is useless. Similar remarks apply if we consider under the alternative a mean curve that takes value 0 at 1.

We summarize the results we have obtained in Table 5.1, in which we have applied our procedure to 1500 random samples with sizes 30, 50 and 200 from standard Brownian motion, which we

Table 5.1. *Application of proposed procedure to the Brownian process $W(t) + \delta t$. The null hypothesis is the standard Brownian motion (i.e. $\delta = 0$). As alternative hypotheses we take $\delta = 0.25, 0.5$ and 1. Samples sizes are 30, 50 and 200.*

		Slope δ			
Sample size		0	0.25	0.5	1
$n = 30$	Rate of correct decisions	.96	.06	.26	.72
	Average p-value	.59	.51	.30	.08
$n = 50$	Rate of correct decisions	.94	.15	.42	.90
	Average p-value	.47	.40	.19	.034
$n = 200$	Rate of correct decisions	.94	.47	.93	.99
	Average p-value	.50	.15	.02	.004

assume to be observed on the equally spaced points $0 = t_0 < \ldots < t_{50} = 1$. For the alternative hypothesis, we consider the shifted Brownian motion with slopes $\delta = 0.25, 0.5$ and 1. The discrete version of the Brownian motion is generated using 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 $N(0, 1/50)$ variable.

The first slope column corresponds to the behavior under the null hypothesis of a test at the level $\alpha = 0.05$. The remaining three columns correspond to the behavior under the alternative for different values of the slope parameter δ of the shifted Brownian processes. We have chosen two parameters to measure this behavior: 'rate of correct decisions' and 'average p-value', which we now explain.

Recall that, for each random sample, the procedure consists of selecting a random $h \in \mathcal{H}$, and then computing the probability that 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 '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 correct decisions' is the proportion of times for which the procedure correctly identifies the situation, i.e. the proportion of times in where it accepts H_0 when $\delta = 0$ and the proportion of times in where it rejects H_0 when $\delta \neq 0$. Thus, this parameter should be close to 0.95 under the null hypothesis. Under the alternative, the bigger this parameter is, the better.

We can summarize Table 5.1 as follows. The test performs well under the null hypothesis, $\delta = 0$. For the other values of δ, the performance of the test is good near the bottom right-hand corner of the table, poor near the top left-hand corner, and intermediate in between.

As mentioned in the Introduction, our aim in this section has been 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. Research into practical implementations of these ideas is still in progress.

Acknowledgments. Part of this research was carried out during a visit of the second author to the Universidad de Cantabria. The authors thank Fernando Etayo Gordejuela and Carlos Matrán Bea for helpful discussions.

References

[1] Béslisle, C., Massé, J.C. and Ransford, T. (1997). When is a probability measure determined by infinitely many projections? *Ann. Probab.* **25**, 767–786.

[2] Cramér, H. and Wold, H. (1936). Some theorems on distribution functions. *J. London Math. Soc.* **11** 290–295.

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

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

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

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

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

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

[9] Reid M. (1988) *Undergraduate Algebraic Geometry*, Cambridge University Press, Cambridge.

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

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

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

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

[14] Stein, E. M. and Weiss, G. (1971) *Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ.

Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Spain
E-mail address: `cuestaj@unican.es`

Departamento de Matemática y Ciencias, Universidad de San Andrés, Argentina
Current address: Centro de Matemática, Universidad de la República, Uruguay
E-mail address: `rfraiman@cmat.edu.uy`

Département de mathématiques et de statistique, Université Laval, Québec (QC), Canada G1K 7P4
E-mail address: `ransford@mat.ulaval.ca`