Journal of Multivariate Analysis **63**, 313–319 (1997)
Article No. MV971705

On a Conjecture Concerning a Theorem of Cramér and Wold*

Guenther Walther

Stanford University

A conjecture concerning the Cramér–Wold device is answered in the negative by giving a Fourier-free, probabilistic proof using only elementary techniques. It is also shown how a geometric idea allows one to interpret the Cramér–Wold device as a special case of a more general concept.

1. INTRODUCTION

A fundamental and widely used theorem states that for checking convergence in distribution of multivariate random variables it is enough to check convergence for all one-dimensional projections. More specifically, the so-called "Cramér–Wold device," due to Cramér and Wold [3] where the technique was initiated, establishes the following two assertions:

(I) A probability measure on Euclidean space is uniquely determined by the values it gives to halfspaces.

(II) In Euclidean d-space, a sequence of random variables X_n converges in distribution to a random variable X if and only if $\langle a, X_n \rangle$ converges in distribution to $\langle a, X \rangle$ for each $a \in \mathbf{R}^d$.

Both theorems, (I) and the stronger (II), although they are very simple in their statements, have been conjectured to require Fourier analysis for their proofs; see, e.g., p. 396 of Billingsley [2] for the first and p. 49 of Billingsley [1] for the second part.

This note gives probabilistic proofs of the two theorems and thus answers this conjecture to the negative. The main argument of the proof is a simple probabilistic idea that goes back to the early stages of probability theory. Also, a geometric idea that belongs to the standard repertoire in

Received February 19, 1997.

AMS 1991 subject classification numbers: 60B10, 60E10.

Key words and phrases: Cramér–Wold device, determining class, Fourier analysis, halfspace, polar set.

* This research was partially supported by NSF Grant DMS-92-24868 at U.C. Berkeley.

0047-259X/97 $25.00

convex geometry shows how the Cramér–Wold device can be interpreted as a special case of a more general concept.

2. SOME FACTS ABOUT DETERMINING CLASSES

The setting throughout will be the Euclidean d-space $\mathbf{R}^d$ equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and the Euclidean norm $|\cdot|$. Write $\mathcal{M}^d$ for the set of probability measures on $\mathbf{R}^d$ and bM^d for the set of bounded, measurable, and real-valued functions on $\mathbf{R}^d$. Recall that a set $\mathcal{D} \subset \mathrm{bM}^d$ is called a determining class if $\mathrm{P}, \mathrm{Q} \in \mathcal{M}^d$ and $\int f \, d\mathrm{P} = \int f \, d\mathrm{Q}$ for all $f \in \mathcal{D}$ imply $\mathrm{P} = \mathrm{Q}$. A basic observation now is that the following lemma has a simple probabilistic proof that does not require Fourier analysis:

LEMMA 1. *Let $f \in \mathrm{bM}^d$ be nonnegative and satisfy $0 < \int f(x) \, dx < \infty$. Then $\{f((a - \cdot)/b),\ a \in \mathbf{R}^d,\ b > 0\}$ is a determining class.*

Proof. Let $\mathrm{P}, \mathrm{Q} \in \mathcal{M}^d$ and assume

$$\int f\left(\frac{a-x}{b}\right) P(dx) = \int f\left(\frac{a-x}{b}\right) Q(dx) \qquad \text{for all} \quad a \in \mathbf{R}^d, \quad b > 0. \quad (1)$$

Normalize f so that $\int f(x) \, dx = 1$ and define $F \in \mathcal{M}^d$ via its density f. Let X_P, X_Q, and X_F be independent random variables in $\mathbf{R}^d$ with distribution P, Q, and F, respectively. Then $X_P + \varepsilon X_F$ has density $\varepsilon^{-d} \int f((\cdot - x)/\varepsilon)$ $P(dx)$, so (1) shows that $\mathcal{L}(X_P + \varepsilon X_F) = \mathcal{L}(X_Q + \varepsilon X_F)$ for all $\varepsilon > 0$. Now let h be any continuous function in bM^d. Then bounded convergence gives

$$\mathbb{E}h(X_P) = \lim_{\varepsilon \downarrow 0} \mathbb{E}h(X_P + \varepsilon X_F) = \lim_{\varepsilon \downarrow 0} \mathbb{E}h(X_Q + \varepsilon X_F) = \mathbb{E}h(X_Q).$$

$P = Q$ follows. ∎

The idea of determining a probability measure by its convolutions with an appropriate class of measures goes back at least to Liapounoff [4] and Lindeberg [5], who employed convolutions in their proofs of the Central Limit Theorem to make use of the resulting smoothness properties.

As an aside, note that Lemma 1 can be sharpened with the use of Fourier analysis and the additional assumption that $\int e^{i\langle t, x \rangle} f(x) \, dx \neq 0$ for all t: Requiring (1) only for $b = 1$ gives $P * F = Q * F$. The characteristic functions of these convolutions factor, so dividing by the nonzero characteristic function of F and using the uniqueness theorem of characteristic functions shows $P = Q$. The resulting determining class apparently is much smaller than the one required in Lemma 1, but in the case of interest here this is only seemingly so:

For fixed $u \neq 0$ the function $f(x) := 1(\langle x, u \rangle \leqslant 1)$ is the indicator of a halfspace and one readily checks that $\{f(a - \cdot),\ a \in \mathbf{R}^d\} = \{f((a - \cdot)/b),\ a \in \mathbf{R}^d,\ b > 0\}$, so nothing is lost by forgoing Fourier analysis in this context.

Of course the above function f is not integrable, so Lemma 1 does not apply. But an application of Fubini's theorem to the result of Lemma 1 gives

LEMMA 2. *Let* $f(x, u) \in \mathrm{bM}^{d+p}$, $\mu_1, ..., \mu_m \in \mathscr{M}^p$, *and* $a_1, ..., a_m \in \mathbf{R}$ *such that* $F(x) := \sum_{k=1}^{m} a_k \int f(x, u)\, \mu_k(du)$ *is nonnegative and satisfies* $0 < \int F(x)\, dx < \infty$. *Then* $\{f((a - \cdot)/b, u),\ a \in \mathbf{R}^d,\ u \in \mathbf{R}^p,\ b > 0\}$ *is a determining class.*

PROOF OF (I)

Set $f(x, u) := 1(\langle x, u \rangle \leqslant 1)$. Then Lemma 2 leads one to consider functions of the form

$$f_\mu^*(\cdot) := \int_{\langle \cdot, u \rangle \leqslant 1} \mu(du). \tag{2}$$

Denote by Φ and ϕ the distribution function and the density function, respectively, of the standard normal distribution on $\mathbf{R}$, and set $\Phi_\sigma(\cdot) = \Phi(\cdot/\sigma)$. We will show in a moment:

(L) There exists a linear combination $g(t) := \sum_{k=1}^{d+1} a_k \Phi_{\sigma_k}(t) + a_{d+2}$ with

- $g(0) = 0$,
- $g(t)$ is strictly increasing for $t \geqslant 0$,
- $g(t) = O(t^{d+1})$ as $t \downarrow 0$.

Now apply Lemma 2 with $F(x) := \sum_{k=1}^{d+2} a_k f_{\mu_k}^*(x)$, where $\mu_k = N(0, \sigma_k^2 I_d)$ for $1 \leqslant k \leqslant d+1$ and $\mu_{d+2} = \delta_0$. Here I_d and δ_0 denote the $d \times d$ identity matrix and point mass at 0, respectively. Observe that $f_{\delta_0}^* \equiv 1$, and as projections of standard normal distributions are standard normal (which can be shown without Fourier analysis),

$$f_{N(0,\sigma^2 I_d)}^*(x) = \Phi_\sigma(1/|x|) \qquad (\text{if } x = 0 \text{ interpret } \Phi_\sigma(1/|x|) = 1).$$

Hence $F(x) = g(1/|x|) = O(1/|x|^{d+1})$ as $|x| \to \infty$. Together with the properties of g one sees that F is nonnegative and satisfies $0 < \int F(x)\,dx < \infty$.

The Cramér–Wold theorem (I) now follows from Lemma 2 and the fact that $f((a-\cdot)/b, u)$ is the indicator of a closed halfspace or $\mathbf{R}^d$ for all $a, u \in \mathbf{R}^d$, $b > 0$.

It remains to prove (L). We will choose the a_k in the linear combination $g(t)$ to eliminate the d coefficients of the t^n, $n = 1, ..., d$, in the Taylor series expansion about 0,

$$\Phi_\sigma(t) = 1/2 + \sum_{n=1}^{d} \frac{1}{n!\,\sigma^n}\,\Phi^{(n)}(0)\,t^n + O(t^{d+1}). \tag{3}$$

For simplicity of exposition we will not make use of the fact that $\Phi^{(2n)}(0) = 0$ for $n \geqslant 1$. Using pairwise different $\sigma_k > 0$ and setting $x := \sigma_{d+1}^{-1}$ in the polynomial interpolation formula

$$x^n = \sum_{k=1}^{d} \sigma_k^{-n} \prod_{i=1,\, i \neq k}^{d} \left(\frac{x - \sigma_i^{-1}}{\sigma_k^{-1} - \sigma_i^{-1}} \right), \qquad n = 0, ..., d-1,$$

one sees that $b_k := \prod_{i=1,\, i\neq k}^{d} (\sigma_{d+1}^{-1} - \sigma_i^{-1})/(\sigma_k^{-1} - \sigma_i^{-1})$, $k = 1, ..., d$, and $b_{d+1} = -1$ solve the system of d equations

$$\sum_{k=1}^{d+1} \sigma_k^{-n} b_k = 0, \qquad n = 0, ..., d-1. \tag{4}$$

Employing the increasing sequence $\sigma_k := 4^k$, one concludes that

$$a_k := -\sigma_k b_k = (-1)^{d+1-k}\, 4^k \prod_{i=1,\, i\neq k}^{d} \frac{|4^{-(d+1)} - 4^{-i}|}{|4^{-k} - 4^{-i}|}, \qquad k = 1, ..., d+1,$$

solve the system (4) for $n = 1, ..., d$. Hence the expansion (3) gives

$$g(t) = \sum_{k=1}^{d+1} a_k \Phi_{\sigma_k}(t) - \sum_{k=1}^{d+1} a_k/2 = O(t^{d+1}) \qquad \text{as} \quad t \to 0.$$

Further,

$$g'(t) = \sum_{k=1}^{d+1} \frac{a_k}{\sigma_k}\,\phi\left(\frac{t}{\sigma_k}\right),$$

and for $1 \leqslant k \leqslant d$,

$$\left|\frac{(a_{k+1}/\sigma_{k+1})\,\phi(t/\sigma_{k+1})}{(a_k/\sigma_k)\,\phi(t/\sigma_k)}\right|$$

$$= \frac{\prod_{i=1,\, i \neq k+1}^{d} |4^{-(d+1)} - 4^{-i}| \cdot \prod_{i=1,\, i \neq k}^{d} |4^{-(k+1)} - 4^{-(i+1)}| \cdot 4^{d-1}}{\prod_{i=1,\, i \neq k}^{d} |4^{-(d+1)} - 4^{-i}| \cdot \prod_{i=0,\, i \neq k}^{d-1} |4^{-(k+1)} - 4^{-(i+1)}|}$$

$$\times \frac{\phi(t/4\sigma_k)}{\phi(t/\sigma_k)}$$

$$= \frac{|4^{-(d+1)} - 4^{-k}| \cdot 4^{d-1}}{|4^{-(k+1)} - 4^{-1}|} \cdot \frac{(\phi(t/\sigma_k))^{1/16}}{\phi(t/\sigma_k)}$$

$$= \frac{|4^{-1} - 4^{d-k}| \cdot 4^k}{|1 - 4^k|} \cdot \left(\phi\left(\frac{t}{\sigma_k}\right)\right)^{-15/16}$$

$$\geqslant 1,$$

because $|4^{-1} - 4^{d-k}| \geqslant 3/4$ and $\phi(t/\sigma_k) \leqslant \phi(0) = 1/\sqrt{2\pi} \leqslant (3/4)^{16/15}$.

As the signs of the a_k, $k \geqslant 1$, are alternating with the sign of a_{d+1} being positive, it follows that $g'(t) > 0$ for $t > 0$. Clearly, $g(0) = 0$.

4. PROOF OF (II) AND A GENERALIZATION

Part II of the Cramér–Wold theorem follows readily from Part I: $X_n \xrightarrow{\text{dist}} X$ implies

$$\langle a, X_n \rangle \xrightarrow{\text{dist}} \langle a, X \rangle \qquad \text{for each} \quad a \in \mathbf{R}^d \tag{5}$$

by the continuous mapping theorem. Conversely, suppose (5) holds. Let $\{e_1, ..., e_d\}$ be an orthonormal system in $\mathbf{R}^d$. For $\delta > 0$, $\bigcap_{i=1}^d \{x\colon \langle \delta e_i, x \rangle \leqslant 1\} \cap \bigcap_{i=1}^d \{x\colon \langle -\delta e_i, x)\rangle \leqslant 1\}$ is a cube centered at 0 with sidelength $2/\delta$. Hence a variation of Boole's inequality together with (5) shows that the sequence $\{\mathscr{L}(X_n)\}$ is uniformly tight. By Prohorov's theorem and the subsequence criterion for metric spaces it is therefore enough to show that any weakly convergent subsequence $\{\mathscr{L}(X_{n_k})\}$ converges to $\mathscr{L}(X)$. But this follows from the already proved implication (5) together with the uniqueness theorem (I).

There is a fundamental geometric concept involved in (2) that allows the Cramér–Wold theorem to be interpreted as a special case of a more general statement:

The *dual* (*polar*) *set* of a set $X \in \mathbf{R}^d$ is defined as

$$X^* := \{u \in \mathbf{R}^d : \langle x, u\rangle \leqslant 1 \text{ for all } x \in X\},$$

see e.g. Stoer and Witzgall [6]. If $x \in \mathbf{R}^d \setminus \{0\}$ then $\{x^*\} = \{u \in \mathbf{R}^d : \langle x, u\rangle \leqslant 1\}$ is a closed halfspace containing 0 in its interior; if $x = 0$ then $\{x^*\}$ is all of $\mathbf{R}^d$. This geometric concept leads one to define for a probability measure $\mu \in \mathscr{M}^d$ the *dual measure* μ^* via its density f^*_μ given in (2). One checks that f^*_μ is upper semicontinuous, hence is measurable. Thus f^*_μ is indeed the density of a σ-finite measure μ^*. μ^* is always an infinite measure. See Walther [7], where also statistical motivations are given for constructing such measures. μ^* can formally also be motivated as follows: For simplicity consider a one-dimensional setting and let F denote the distribution function of a probability measure. For real x write

$$\begin{aligned} F(x) &= \int 1_{(-\infty, x]}(u)\, F(du) \\ &= \int 1_{[u, \infty)}(x)\, F(du). \end{aligned} \tag{6}$$

Formally, (6) can be read as a mixture of uniform densities (albeit not of probability densities).

Now the Cramér–Wold theorem is a consequence of the following more general theorem about dual measures a proof of which can be found in Walther [8]:

THEOREM 1. *Let $X, X_1, X_2, ..., \in \mathbf{R}^d$ be a sequence of random variables with $\mathscr{L}(X) = F$, $\mathscr{L}(X_n) = F_n$, $n \geqslant 1$. Then the following are equivalent*:

(i) $F_n \xrightarrow{\text{weakly}} F$

(ii) $\langle a, X_n\rangle \xrightarrow{\text{dist}} \langle a, X\rangle$ *for all* $a \in \mathbf{R}^d$

(iii) $f^*_{F_n} \xrightarrow{\text{a.e.}} f^*_F$

(iv) $F^*_n \to F^*$ *in variation norm on compacts*

(v) $F^*_n \xrightarrow{\text{vaguely}} F^*$.

If F_n is the empirical measure of F, then (iii) *can be strengthened to uniform convergence F-almost surely.*

As a corollary one obtains the following identifiability property: $F^* = G^*$ iff $F = G$.

ACKNOWLEDGMENT

I thank Richard Olshen for his interest in the paper and for valuable suggestions.

REFERENCES

[1] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[2] Billingsley, P. (1986). *Probability and Measure*, 2nd ed. Wiley, New York.

[3] Cramér, H., and Wold, H. (1936). Some theorems on distribution functions. *J. London Math. Soc.* **11** 290–294.

[4] Liapounoff, A. M. (1900). Sur une proposition de la théorie des probabilités. *Bull. Acad. Impériale Sci. St. Pétersbourg* **13** 359–386.

[5] Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.* **15** 211–225.

[6] Stoer, J., and Witzgall, C. (1970). *Convexity and Optimization in Finite Dimensions*, Vol. 1. Springer-Verlag, Berlin/Heidelberg.

[7] Walther, G. (1995). Monte Carlo sampling in dual space for approximating the empirical halfspace distance. *Ann. Statist.*, in press.

[8] Walther, G. (1994). *Statistical Applications of Geometric Duality*. Ph.D. thesis, Dept. of Statistics, University of California, Berkeley.