No. 1778

PROJECTION PURSUIT TYPE APPLICATIONS OF MULTIVARIATE EMPIRICAL CHARACTERISTIC FUNCTIONS

by

Sucharita Ghosh*) *and* Frits H. Ruymgaart*)

Cornell University and Kath. Un. Nijmega

---

*)Part of this work has been done while the authors were visiting the Department of Statistics, Un. North Carolina, Chapel Hill, and while the second author was visiting the Department of Statistics, Texas A & M Un., College Station.

ABSTRACT

In this note we extend univariate tests for normality and symmetry based on empirical characteristic functions to the multivariate case, using a projection pursuit procedure. We allow the weighting over the directions in the unit sphere to depend on the cloud of data points. The proposed tests are of the Cramér-von Mises type.

AMS 1980 *Subject Classifications*. Primary 62G10, secondary 60G99.

Key words and phrases. Multivariate empirical characteristic function, projection pursuit, testing for multivariate normality, testing for multivariate symmetry.

SUCHARITA GHOSH
DEPT. OF ECONOMIC
AND SOCIAL STATISTICS
358 IVES HALL
CORNELL UNIVERSITY
ITHACA, N.Y. 14851-0952
U.S.A.

F.H. RUYMGAART
DEPT. OF MATHEMATICS
KATH. UNIVERSITEIT
TOERNOOIVELD
6525 ED NIJMEGEN
THE NETHERLANDS

1. INTRODUCTION AND BASIC NOTATION

In this note we extend some univariate test procedures based on empirical characteristic functions (e.c.f.'s) to higher dimensions by a projection pursuit method. Tests for normality and symmetry seem to be particularly well suited for an expedient use of the e.c.f. For testing univariate normality we refer to e.g. Murota & Takeuchi (1981); for the multivariate case and a survey of both univariate and multivariate procedures see Csörgő (1986). Tests for univariate symmetry are given in Feuerverger & Mureika (1972) and Heathcote (1972). An early application of projection pursuit is the union-intersection principle in Roy (1953). The idea was further established by Kruskal (1969, 1972) and Friedman & Tukey (1974); see Huber (1985) for a survey. Other examples of application of the idea of projection pursuit may be found in Beran & Millar (1986), Buhrman & Ruymgaart (1981), Hall (1988), and Ruymgaart (1981).

Throughout $X_1,\ldots,X_n$ will denote i.i.d. random vectors in $\mathbb{R}^d$ with common d-variate distribution function (d.f.) F with density f. The unit sphere in $\mathbb{R}^d$ will be denoted by

$$(1.1)\quad \Theta = \{\theta \in \mathbb{R}^d:\ \|\nu\| = 1\},$$

where $\|\cdot\|$ is the norm in $\mathbb{R}^d$. Vectors will be considered as rows with $*$ used to indicate the transpose. Let us introduce the one-dimensional projections

$$(1.2)\quad X_{j,\theta} = X_j\,\theta^*,\quad j \in \{1,\ldots,n\},\ \theta \in \Theta.$$

These random variables are i.i.d. with common d.f. F_θ and density f_θ, say. The empirical d.f. $\hat{F}_{n,\theta}$ of the $X_{1,\theta},\ldots,X_{n,\theta}$ are defined in the usual way.

The multivariate c.f. and its empirical analogue are given by

(1.3) $C(t) = \int \exp(i\ t\ x^*)dF(x),\ \hat{C}_n(t) = \int \exp(i\ t\ x^*)d\hat{F}_n(x),\ t \in \mathbb{R}^d,$

respectively, where i denotes the imaginary unit. Let us also introduce

(1.4) $C_\theta(\rho) = \int \exp(i\rho x)dF_\theta(x),\ \hat{C}_{n,\theta}(\rho) = \int \exp(i\rho x)d\hat{F}_{n,\theta}(x),\ \rho \in \mathbb{R}.$

Note that we have, of course,

(1.5) $C(t) = E\ \hat{C}_n(t),\ t \in \mathbb{R}^d;\ C_\theta(\rho) = E\ \hat{C}_{n,\theta}(\rho),\ \rho \in \mathbb{R}.$

Since each $t \in \mathbb{R}^d$ with $t \neq 0$ may be written as

(1.6) $t = \rho\theta,\ \rho = \|t\|,\ \theta = t/\|t\| \in \Theta,$

we have, moreover, the relations

(1.7) $C(t) = C_\theta(\rho),\ \hat{C}_n(t) = \hat{C}_{n,\theta}(\rho).$

Although $\hat{C}_{n,\theta}(\rho) = \hat{C}_{n,-\theta}(-\rho)$ it is often convenient to consider $\hat{C}_{n,\theta}(\rho)$ for all $\rho \in \mathbb{R}$ and $\theta \in \Theta$.

Since for fixed θ the density f_θ is univariate normal or univariate symmetric under the respective hypotheses of multivariate normality and symmetry, we may first test the univariate hypotheses for various θ using well-known statistics based on the $\hat{C}_{n,\theta}$ in (1.4) and next try to combine these statistics in some way to obtain a procedure for the multivariate problem. More specifically, assuming that $T_n(\theta)$ is a standardized test statistic for one of these univariate problems, one might expect that

(1.8) $T_{n,\upsilon} = \int_{\theta\in\Theta} T_n(\theta)d\upsilon(\theta),$

where υ is a finite measure on Θ, is a useful test statistic for the multivariate extension. Simple statistics are obtained by concentrating μ on a finite number of points.

The choice of the measure υ in (1.8) will have an important impact on the power of the test. In this measure we may reflect our ideas which directions θ are interesting by letting υ concentrate mass around such directions. In testing situations interesting directions typically depend on the type of alternative. If we don't have sufficient information about the alternative we might decide on choosing υ to be the uniform measure on Θ or on a finite subset of Θ.

Another possibility is to allow the measure υ to be random and to depend on the cloud of data points. Let us describe the construction of such a measure in the simplest case where we know that the underlying multivariate density f has a center at the origin, say. Let us consider

$$(1.9) \quad \theta_j = X_j / \|X_j\|, \; j \in \{1,\ldots,n\}.$$

Then $\theta_1,\ldots,\theta_n$ are i.i.d. random vectors in the sphere Θ with a common density φ that depends on f. Directions θ with relatively high values $\varphi(\theta)$ are directions where data points have a tendency to cluster around and are therefore interesting. Since in general f and hence φ are unknown, we need to use an estimator $\hat{\varphi}_n$ of φ. Let us here use the naive estimator introduced in Ruymgaart (1988); in terms of the original sample elements this estimator is based on the relative number of X_i in small cones with vertices at the origin. These considerations lead to test statistics of the type

$$(1.10) \quad \hat{T}_n = \int_{\theta\in\Theta} T_n(\theta)\, \hat{\varphi}_n(\theta)d\theta,$$

which will in general have the same limiting distribution as $\int_{\theta\in\Theta} T_n(\theta)\varphi(\theta)d\theta$ due to the consistency of $\hat{\varphi}_n$.

In Section 2 we will briefly consider a Cramér-von Mises type test for testing multivariate normality. Although no projection pursuit method is used in Csörgő (1986) the main tool in that paper remains most useful for our purpose as well. The problem of testing multivariate symmetry will be considered in Section 3.

2. TESTING NORMALITY

In this section we will consider the problem of testing the hypothesis that the X_i have a d-variate $N(\mu,\Sigma)$ - distribution with arbitrary μ and Σ. We will use the notation

$$(2.1)\quad \bar{X} = \frac{1}{n}\Sigma_{j=1}^{n} X_j,\ \hat{S} = \frac{1}{n-1}\Sigma_{j=1}^{n}(X_j-\bar{X})^*(X_j-\bar{X}),$$

for the sample mean and sample covariance matrix respectively.

Extending the idea in Murota & Takeuchi (1981) of studentizing the empirical characteristic function to the multivariate case, Csörgő (1988) introduces the stochastic process

$$(2.2)\quad Z_n(t) = n^{1/2}\{|\hat{C}_n(t\,\hat{S}^{-1/2})|^2 - \exp(-t\,t^*)\},\quad t\in[-c,c]^d,$$

for some $c\in(0,\infty)$. It has been shown in Csörgő (1988) that under the hypothesis of normality these processes converge weakly to a zero mean Gaussian process Z in the space of continuous functions $\mathbb{C}([-c,c]^d)$ endowed with the supremum metric and with covariance function

$$(2.3)\quad K(t_1,t_2)\ 4\ \exp(-t_1t_1^* - t_2t_2^*)\{\cosh(t_1t_2^*) - 1 - \tfrac{1}{2}(t_1t_2^*)^2\};$$
$$t_1,t_2\in[-c,c]^d\ .$$

For a given direction $\theta\in\Theta$ and $a\in\mathbb{R}$ let us first consider

$$(2.4)\quad S_n(\theta) = Z_n(a\,\theta)$$

as a test statistic. Let us note that $T_n(\theta)$ is the standardized empirical characteristic function of the random variables

(2.5) $\hat{X}_{j,\theta} = X_j \hat{S}^{-1/2} \theta^*$,

evaluated at t=a. For d=1 this statistic coincides with the one proposed in Murota & Takeuchi (1981); these authors, moreover, recommended certain choices for the number a.

It is clear from (2.2) and (2.4) that under the hypothesis the processes S_n converge weakly to the zero mean Gaussian process $Z(a\cdot)$ in the space of continuous function $\mathbb{C}(\theta)$ endowed with the supremum norm, where this limiting process has covariance function

(2.6) $\sigma(\theta_1,\theta_2) = K(a\,\theta_1,\, a\,\theta_2);\ \theta_1,\theta_2 \in \Theta$.

For a given finite measure υ on Θ let us consider the covariance function in (2.6) as an integral operator on $L^2(\Theta,\upsilon)$. Let $\lambda_{1,\upsilon} \geq \lambda_{2,\upsilon} \geq \ldots \geq 0$ be the eigenvalues and $e_{1,\upsilon}$, $e_{2,\upsilon}$,... the set of corresponding eigenfunctions in $L^2(\Theta,\upsilon)$. Due to the studentization the covariance operator and its eigenvalues and eigenfunctions are independent of the underlying distribution when the hypothesis is fulfilled. Let us now propose a test of Cramér-von Mises type, based on $S_n^2(\theta) = T_n(\theta)$, $\theta \in \Theta$.

THEOREM 2.1. *For any finite measure υ on Θ a test for the hypothesis of multivariate normality is obtained by rejecting for large values of*

(2.7) $T_{n,\upsilon} = \int_{\theta\in\Theta} S_n^2(\theta)d\upsilon(\theta)$,

The asymptotic level of this test may be derived from

(2.8) $T_{n,\upsilon} \Rightarrow \Sigma_{k=1}^{\infty} \lambda_{k,\upsilon} Z_k^2$, *as* $n \to \infty$,

where $Z_1, Z_2, \ldots$, are i.i.d. standard normal random variables.

PROOF. Because the functional $\int_\Theta (\cdot)^2 d\,\upsilon$ is continuous on $\mathbb{C}(\Theta)$ the weak convergence of $S_n(\cdot)$ to $Z(a\cdot)$ entails at once that $T_{n,\upsilon} \Rightarrow \int_\Theta Z^2(a\cdot)d\,\upsilon$. The proof that the distribution of the latter random variable equals that of the one on the right in (2.8) follows the standard pattern using the properties of the integral operator σ on $L^2(\Theta,\upsilon)$. QED

In the present situation there is no information about the center of the underlying multivariate density, not even under the hypothesis. Let us therefore consider

$$(2.9)\quad \hat{\theta}_j = (X_j-\bar{X})S^{-\frac{1}{2}} \,/\, \|(X_j-\bar{X})S^{-\frac{1}{2}}\|,\ j \in \{1,\ldots,n\}.$$

Although no longer independent these random vectors in Θ still have the same distribution. Let $\hat{\varphi}_n$ be the estimator of the density of this distribution constructed in the same way as described in Section 1, but based on the $\hat{\theta}_j$ rather than the θ_j.

THEOREM 2.2. *A test for the hypothesis of multivariate normality is also obtained by rejecting for large values of*

$$(2.13)\quad \hat{\hat{T}}_n = \int_{\theta\in\Theta} S_n^2(\theta)\hat{\hat{\varphi}}_n(\theta)d\theta.$$

Under the hypothesis of multivariate normality the $\hat{\hat{T}}_n$ *have the limiting distribution on the right in* (2.8) *with* υ *the uniform measure on* Θ.

PROOF. It is easy to see that under the hypothesis, for large n, the distribution of the $\hat{\theta}_j$ approaches the uniform distribution on Θ with density u, say. Due to symmetry considerations one might even argue that, for each n, the distribution of the $\hat{\theta}_j$ is uniform when we sample from a $N(\mu,\ddagger)$-distribution. The dependence of the $\hat{\theta}_j$ is due to the random elements $\bar{X}$ and $S^{-\frac{1}{2}}$ they have in

common. Because these are close to μ and $\Sigma^{-\frac{1}{2}}$ respectively, it requires only a minor modification of the result in Ruymgaart (1988) to prove that

$$(2.11)\quad \sup_{\theta\in\Theta}|\hat{\hat{\varphi}}_n(\theta) - u(\theta)| \xrightarrow[a.s.]{} 0, \quad as\ n \to \infty.$$

The theorem follows if we can prove that

$$(2.12)\quad \left|\int_{\theta\in\Theta} S_n^2(\theta)\hat{\hat{\varphi}}_n(\theta)d\theta - \int_{\theta\in\Theta} S_n^2(\theta)u(\theta)d\theta\right| \leq$$

$$\leq \sup_{\theta\in\Theta}|\hat{\hat{\varphi}}_n(\theta)-u(\theta)|\int_{\theta\in\Theta} S_n^2(\theta)d\theta \xrightarrow[p]{} 0, \ as\ n \to \infty.$$

This is immediate from the previous theorem and (2.11). QED

3. TESTING SYMMETRY

First we need to define the kind of multivariate symmetry that we are going to consider here. A multivariate density f will be called (*multivariate*) *symmetric* about the origin if f_θ is (univariate) symmetric about zero for each $\theta \in \Theta$. In a similar way multivariate symmetry about an arbitrary point may be defined. According to this definition each multivariate normal is symmetric about its mean. We will focus on testing symmetry about the origin.

For testing univariate symmetry of f_θ for arbitrary $\theta \in \Theta$, Feuerverger & Mureika (1977) propose the test statistic

$$(3.1)\quad T_{n,\mu}(\theta) = n \int \hat{I}^2_{n,\theta}(\rho)d\mu(\rho),$$

where $\hat{I}_{n,\theta}(\rho) = (1/n)\ \Sigma_{i=1}^n \sin\rho\ X_{j,\theta}$ is the imaginary part of $\hat{C}_{n,\theta}(\rho)$ and μ a finite measure on $\mathbb{R}$ which is itself *symmetric* about zero. This choice is motivated by the circumstance that the imaginary part I_θ of C_θ satisfies

$$(3.2)\quad C_\theta(\rho) = \int \sin\rho x\ dF_\theta(x) = 0,\ p \in \mathbb{R},$$

when f_θ is symmetric about zero.

In the multivariate case it seems natural to choose a test statistic like (υ is a finite measure on Θ)

(3.3) $T_{n,\mu,\upsilon} = n \int_{\theta\in\Theta} \int_{\rho\in\mathbb{R}} \hat{I}^2_{n,\theta}(\rho)\, d\mu(\rho)\, d\upsilon(\theta)$,

and reject for large values. Even under the hypothesis of symmetry, however, the limiting distribution of such statistics depends in general on the underlying density f. In order to find this limiting distribution we may bypass the weak convergence of the processes $\{T_n(\theta),\ \theta \in \Theta\}$ in (3.1) that don't seem to be of particular interest, by appealing to the theory of U-statistics; see e.g. Serfling (1980).

Let us introduce the bounded symmetric kernel

(3.4) $h(x,y) = \int_\theta \int_\rho (\sin\rho x\theta^*)(\sin\rho y\theta^*)d\mu(\rho)d\upsilon(\theta),\ x,y \in \mathbb{R}^d$,

and consider the integral operator

(3.5) $A\, g(x) = \int_y h(x,y)g(y)dF(y),\ x \in \mathbb{R}^d$,

on $L^2(\mathbb{R}^d,F)$. Let

(3.6) $\lambda_1(F),\ \lambda_2(F),\ldots$

be the eigenvalues of this operator (note that these eigenvalues also depend on μ and υ).

THEOREM 3.1. *Under the hypothesis of multivariate symmetry we have*

(3.7) $T_{n,\mu,\upsilon} \Rightarrow \Sigma_{k=1}^{\infty} \lambda_t(F)Z_k^2$,

where $Z_1, Z_2, \ldots$ *are i.i.d. standard normal random variables.*

PROOF. First let us observe that we may write

$$(3.8)\quad T_n = n^{-1} \Sigma \Sigma_{j,k} h(X_j, X_k) =$$

$$= n^{-1} \Sigma \Sigma_{j \neq k} h(X_j, X_k) + n^{-1} \Sigma_j h(X_j, X_j)$$

$$= 2\, n^{-1} \binom{n}{2} \binom{n}{2}^{-1} \Sigma \Sigma_{j<k} h(X_j, X_k) + n^{-1} \Sigma_j h(X_j, X_j).$$

Here $\binom{n}{2}^{-1} \Sigma \Sigma_{j<k} h(X_j, X_k)$ is a U-statistic with

$$(3.9)\quad E\, h(X_1, X_2) = \int_\theta \int_\rho E(\sin \rho X_1 \theta^*) E(\sin \rho X_2 \theta^*) d\mu(\rho) d\upsilon(\theta) = 0,$$

under the hypothesis; see also (3.2). In a similar way it follows that $h_1(x) = E\, h(x, X_1) = 0$ for all $x \in \mathbb{R}^d$, which entails that

$$(3.10)\quad \zeta_1 = \mathrm{Var}\, h_1(X_1) = E\, h_1^2(X_1) = 0.$$

Finally, we have

$$(3.11)\quad \zeta_2 = \mathrm{Var}\, h_2(X_1, X_2) = E\, h_2^2(X_1, X_2) = E\, h^2(X_1, X_2) > 0.$$

Application of serfling (1980, Theorem p. 194) yields

$$(3.12)\quad n \binom{n}{2}^{-1} \Sigma_{j<k} h(X_j, X_k) \Rightarrow \int_{k=1}^{\infty} \lambda_k(F)(Z_k^2 - 1).$$

According to the strong law of large numbers we have

$$(3.13)\quad n^{-1} \Sigma_j h(X_j, X_j) \xrightarrow{a.s.} \int_x h(x,x) dF(x), \text{ as } n \to \infty.$$

Combination of (3.8), (3.12) and (3.13) yields (3.7) because the operator A in (3.5) satisfies

$$(3.14)\quad \mathrm{tr}\, A = \int_x h(x,x) dF(x) = \Sigma_{k=1}^{\infty} \lambda_k(F),$$

and since $2\binom{n}{2} n^{-2} \to 1$, as $n \to \infty$. QED

When we restrict to alternatives that still have a center at the origin we might take for υ the random measure with density $\hat{\varphi}_n$ on Θ, based on the i.i.d. random vectors in (1.9). We can then prove that

$$(3.15)\quad \hat{T}_{n,\mu} = n \int_\theta \int_\rho \hat{I}^2_{n,\theta}(\rho)\, \hat{\varphi}_n(\theta)\, d\mu(\rho)d\theta$$

converges weakly to the distribution of the random variable on the right in (3.7) with υ the uniform measure on Θ.

It has already been observed that the limiting distribution depends on the underlying distribution function so that estimation or bootstrapping will be needed. The estimation may be carried out in a similar way as in Feuerverger & Mureika (1977, Section 4).

An extension is needed in the case where the center of symmetry is unknown and has to be estimated. If τ is the true unknown center of symmetry and $\hat{\tau}$ a consistent estimator, it can be shown that the modifications of the statistics in (3.1) and (3.15), obtained by replacing the X_j by the $X_j-\hat{\tau}$, have the same limiting distributions as the original statistics.

REFERENCES

[1] BERAN, R. & MILLAR, P.W. (1986). Confidence sets for a multivariate distribution. *Ann. Statist.* *14*, 431-443.

[2] BUHRMAN, H. & RUYMGAART, F.H. (1981). An application of linearization in nonparametric multivariate analysis. *Sankhyā* 43, A, 52-66.

[3] CSÖRGŐ, S. (1986). Testing for normality in arbitrary dimension. *Ann. Statist.* *14*, 708-723.

[4] FEUERVERGER, A. & MUREIKA, R.A. (1977). The empirical characteristic function and its applications. *Ann. Statist.* 5, 88-97.

[5] FRIEDMAN, J.H. & TUKEY, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* C-23, 881-889.

[6] HALL, P. (1988). Estimating the direction in which a data set is most interesting. *Probab. Theory Rel. Fields*, to appear.

[7] HEATHCOTE, C.R. (1972). A test of goodness of fit for symmetric random variables. *Austral. J. Statist. 14*, 172-181.

[8] HUBER, P. (1985). Projection pursuit, *Ann. Statist. 13*, 435-475.

[9] KRUSKAL, J.B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation". In *Statistical Computation* (R.C. Milton and J.A. Nelder, eds.), Acad. Press, New York.

[10] MUROTA, K. & TAKEUCHI, K. (1981). The studentized empirical characteristic function and its application to test for the shape of distribution. *Biometrika* 68, 55-65.

[11] ROY, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist. 24*, 220-238.

[12] RUYMGAART, F.H. (1981). A robust principal component analysis. *J. Multivar. Analysis 11*, 485-497.

[13] RUYMGAART, F.H. (1988). Strong uniform convergence of density estimators on spheres. *J. Statist. Pl. Inf.*, to appear.

[14] SERFLING, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.