



Notes

A characterization of multivariate normality through univariate projections

Yongzhao Shao^{a,b,*}, Ming Zhou^b

^a Division of Biostatistics, New York University SOM, 650 First Avenue, New York, NY 10016, USA

^b Department of Statistics, Iowa State University, Ames, IA 50011, USA

ARTICLE INFO

Article history:

Received 11 September 2009

Available online 15 June 2010

AMS 2010 subject classifications:

primary 62E10

62E15

secondary 62H10

Keywords:

Goodness of fit

Linear combination of components

Marginal distribution

Multivariate normal distribution

Non-normality

ABSTRACT

This paper introduces a new characterization of multivariate normality of a random vector based on univariate normality of linear combinations of its components.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

As is well known, the multivariate normal distribution is central to multivariate analysis. Therefore, characterizations and assessments of multivariate normality have attracted sustained interest from researchers as demonstrated in the monographs and papers by [1,12,11,10] and others.

Commonly used assessments of multivariate normality or non-normality of a random vector include a variety of approaches based on linear combinations of variates. In particular, many types of univariate-based plots are both easy to make and simple to use for detecting skewness, outliers, and other departures from multivariate normality [11]. In addition, there exist many formal tests for multivariate normality of a random vector based on examination of selected linear combinations of its components [11,9]. Indeed, as pointed out by Anderson [1, p. 23], “One of the reasons that the study of normal multivariate distributions is so useful is that marginal distributions and conditional distributions derived from multivariate normal distributions are also normal distributions. Moreover, linear combinations of multivariate normal variates are again normally distributed.”

One the other hand, it is well known that a non-normal random vector may have some normally distributed linear combinations of its components [12,9]. This does raise a serious question concerning the effectiveness of the common statistical practice for assessing multivariate normality by examining a few linear combinations of components. After all, only a few of the infinitely many linear combinations can be plotted or tested in practice. Therefore, it is of theoretical interest to characterize or measure the size of the set of normally distributed linear combinations. Probabilistically, one

* Corresponding author at: Division of Biostatistics, New York University SOM, 650 First Avenue, NY 10016, USA.

E-mail addresses: shaoy01@nyu.edu, shaoy01@nyumc.org (Y. Shao), zhouming@iastate.edu (M. Zhou).

might ask: how large is the chance that a randomly selected linear combination of components from a non-normal random vector is normally distributed? Indeed, this problem has attracted the attention of many researchers for a long time [6,8,7]. Remarkably, Hamedani and Tata [8] proved that a bivariate random variable is normally distributed if it has a infinite collection of distinct linear combinations of its components that are normally distributed. In particular, this result implies that a non-normal bivariate random vector can only have finitely many normally distributed linear combinations of its components. However, this characterization of bivariate normality cannot be extended to the multivariate case in a straightforward way [7]. The main objective of this paper is to introduce a new characterization of multivariate normality through univariate projections that holds in all dimensions. We show that, for any non-normal random vector, the set of normally distributed linear combinations of its components is negligible among all possible linear combinations. In particular, in any dimensions, the probability is zero that a randomly selected linear combination of components of a non-normal random vector is normally distributed. This finding includes the existing bivariate result of Hamedani and Tata [8] as a corollary (see Remark 2 in Section 2 for more details). Given the prominent role of normal distributions in multivariate statistical analysis [1,12,11], the finding of this paper bears a certain significance for the assessment of multivariate normality, and thus might be of interest to many researchers.

In the next section, we establish a new characterization of multivariate normality for a random vector by assessing the normality of linear combinations of its components. The linear combinations will also be called projections on the vector of coefficients. Section 3 contains concluding remarks.

2. Main results

The first subsection introduces the basic notation, the multivariate normal distribution, the normal directions, and a few lemmas. The proofs of these lemmas are rather elementary, but are included for completeness. The new characterization of multivariate normality can be found in the second subsection.

2.1. Notation and lemmas

Let \mathbb{R}^p ($p \geq 1$) be the p -dimensional Euclidean space. The inner product of two vectors $\mathbf{x} = (x_1, \dots, x_p)^T$ and $\mathbf{y} = (y_1, \dots, y_p)^T \in \mathbb{R}^p$ is denoted as $\mathbf{x}^T \mathbf{y} = \sum_{i=1}^p x_i y_i$. We use $\mathbb{S} = \{\mathbf{u} \in \mathbb{R}^p \mid \mathbf{u}^T \mathbf{u} = 1\}$ to denote the unit sphere, m the Lebesgue measure in \mathbb{R}^p , i.e., $m(A)$ denotes the Lebesgue measure of a measurable set A . Also, let \mathcal{I} denote the uniform measure on the unit sphere \mathbb{S} , and \mathbb{N} the set of natural numbers.

We will say that a random vector $\mathbf{X} = (X_1, \dots, X_p)$ has a multivariate normal distribution if the support of \mathbf{X} is the entire space \mathbb{R}^p and there exist a p -vector μ and a symmetric, positive-definite $p \times p$ matrix Σ , such that the probability density function of \mathbf{X} can be expressed as

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right),$$

where $|\Sigma|$ is the determinant of Σ . The vector μ is the expected value and the matrix Σ is the covariance matrix of \mathbf{X} . If a random vector has a p -variate normal distribution, by the above definition, it must have a density function and a non-singular covariance matrix. As is well known, given independent and identically distributed random observations, the mean vector μ and the covariance matrix Σ can be consistently estimated by their sample counterparts, i.e., the sample mean \bar{X} and sample covariance matrix S_n^2 , respectively. Moreover, given the existence of a p -variate Lebesgue density of \mathbf{X} , the sample covariance matrix S_n^2 is non-singular almost surely [5,4]. Therefore, it is not essential to know the mean vector μ and the covariance matrix Σ . Indeed, without loss of generality, both the mean vector μ and the covariance matrix Σ are commonly assumed unknown in statistics and many other applications. Throughout this paper, we consider a given random vector $\mathbf{X} = (X_1, \dots, X_p)^T \in \mathbb{R}^p$ possessing a density function $f(\mathbf{x})$ relative to the Lebesgue measure m . In particular, we call a vector $\mathbf{u} = (u_1, \dots, u_p)^T \in \mathbb{S}$ a normal direction of \mathbf{X} (or $f_{\mathbf{X}}$) if its one-dimensional projection on \mathbf{u} , $\mathbf{u}^T \mathbf{X}$, has a univariate normal distribution.

Note that when $\mathbf{u}^T \mathbf{X}$ is normally distributed, its moment generating function exists. Then we can denote its mean and variance by $\mu_{\mathbf{u}}$ and $\sigma_{\mathbf{u}}^2$, respectively. Therefore $\mathbf{u}^T \mathbf{X}$ is normally distributed if and only if $E\{\exp(t\mathbf{u}^T \mathbf{X})\} = \exp(\mu_{\mathbf{u}} t + t^2 \sigma_{\mathbf{u}}^2 / 2)$, $\sigma_{\mathbf{u}}^2 > 0$, for all $t \in \mathbb{R}$. Or equivalently, in terms of the density f of \mathbf{X} ,

$$\int_{\mathbb{R}^p} \exp(t\mathbf{u}^T \mathbf{x}) f(\mathbf{x}) \, d\mathbf{x} = \exp(\mu_{\mathbf{u}} t + t^2 \sigma_{\mathbf{u}}^2 / 2), \quad \sigma_{\mathbf{u}}^2 > 0, \text{ for all } t \in \mathbb{R}. \tag{1}$$

Let \mathbb{G} be the set of lines in \mathbb{R}^p that lie on normal directions of \mathbf{X} and pass through the origin, that is,

$$\mathbb{G} = \{\mathbf{u} \in \mathbb{R}^p \mid \mathbf{u}^T \mathbf{X} \text{ is normally distributed}\}. \tag{2}$$

We assume $\mathbf{0} \in \mathbb{G}$. Also, denote \mathbb{U} as the set of normal directions of \mathbf{X} ; then $\mathbb{U} = \mathbb{G} \cap \mathbb{S}$. Since a univariate normal distribution is completely determined by its moments [3, p. 389], \mathbb{U} can be written in terms of moment equations. Let ϕ be the density of the standard normal distribution; then

$$\mathbb{U} = \left\{ \mathbf{u} \in \mathbb{S} \mid \int_{\mathbb{R}^p} (\mathbf{u}^T \mathbf{x})^n f(\mathbf{x}) \, d\mathbf{x} - \int_{\mathbb{R}} t^n \frac{1}{\sigma_{\mathbf{u}}} \phi\left(\frac{t - \mu_{\mathbf{u}}}{\sigma_{\mathbf{u}}}\right) dt = 0, \text{ for all } n \in \mathbb{N} \right\}. \tag{3}$$

With the above notation, it is well known that \mathbf{X} is normally distributed if and only if $\mathbb{G} = \mathbb{R}^p$ or $\mathbb{U} = \mathbb{S}$. Next we are going to show that \mathbf{X} is normally distributed as long as \mathbb{G} has positive Lebesgue measure. In the first lemma, we will show that \mathbb{G} is a closed set and thus Lebesgue measurable.

Lemma 1. *The set $\mathbb{G} = \{\mathbf{u} \in \mathbb{R}^p \mid \mathbf{u}^T \mathbf{X} \text{ is normally distributed}\}$ is closed if \mathbf{X} has a density in \mathbb{R}^p .*

Proof. It suffices to show that the set \mathbb{G} contains all its limiting points. If a non-zero sequence $\{\mathbf{u}_n\}_{n \geq 1} \subset \mathbb{G}$ converges to $\mathbf{u}_0 \neq \mathbf{0}$, then $\mathbf{u}_n^T \mathbf{X}$ converges to $\mathbf{u}_0^T \mathbf{X}$ in distribution, where $\mathbf{u}_0^T \mathbf{X}$ is non-degenerate because \mathbf{X} has a Lebesgue density by assumption. Let $\alpha_n = E(\mathbf{u}_n^T \mathbf{X})$, $\beta_n^2 = \text{Var}(\mathbf{u}_n^T \mathbf{X})$; then $\beta_n^{-1}(\mathbf{u}_n^T \mathbf{X} - \alpha_n)$ has a standard normal distribution. By the convergence of types theorem [3, p. 193], there exist real numbers $\beta > 0$ and α such that $\lim_{n \rightarrow \infty} \alpha_n = \alpha$, $\lim_{n \rightarrow \infty} \beta_n = \beta$ and $\mathbf{u}_0^T \mathbf{X}$ has a normal distribution. Thus $\mathbf{u}_0 \in \mathbb{G}$ and \mathbb{G} is a closed set in \mathbb{R}^p . \square

Before proving that \mathbf{X} is normally distributed, it is necessary to show that all moments of \mathbf{X} exist, which is true if \mathbb{G} has positive Lebesgue measure, i.e. $m(\mathbb{G}) > 0$, as asserted by the next lemma.

Lemma 2. *For a random vector \mathbf{X} with a Lebesgue density in \mathbb{R}^p , all moments of \mathbf{X} exist if the set $\mathbb{G} = \{\mathbf{u} \in \mathbb{R}^p \mid \mathbf{u}^T \mathbf{X} \text{ is normally distributed}\}$ has positive Lebesgue measure.*

Proof. Let m be the Lebesgue measure in \mathbb{R}^p . Since $m(\mathbb{G}) > 0$, there exists a basis of \mathbb{R}^p , $\{\mathbf{u}_1, \dots, \mathbf{u}_p\} \subset \mathbb{G}$. Otherwise there exists $\{\mathbf{u}_1, \dots, \mathbf{u}_r\} \subset \mathbb{G}$ with $r < p$, such that any element in \mathbb{G} is a linear combination of $\mathbf{u}_1, \dots, \mathbf{u}_r$. Then \mathbb{G} would be a subset of the linear vector space spanned by $\mathbf{u}_1, \dots, \mathbf{u}_r$, which has Lebesgue measure 0 in \mathbb{R}^p . Consequently $m(\mathbb{G}) = 0$, which is a contradiction. Now we can assume that $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ can be chosen as a basis in \mathbb{R}^p . Let $\mathbf{Y} = (Y_1, \dots, Y_p)^T = (\mathbf{u}_1, \dots, \mathbf{u}_p)^T \mathbf{X}$, $i = 1, \dots, p$; then $E|Y_i|^m < \infty$ for all $m \in \mathbb{N}$ because $Y_i = \mathbf{u}_i^T \mathbf{X}$ is normal. Moreover, $\mathbf{X} = \{(\mathbf{u}_1, \dots, \mathbf{u}_p)^T\}^{-1} \mathbf{Y}$, that is, each X_i is a linear combination of normal random variables. Thus for each i , $E|X_i|^m < \infty$ for all $m \in \mathbb{N}$, or equivalently, $E\{|X_1|^{r_1} \cdots |X_p|^{r_p}\} < \infty$ for all $r_1, \dots, r_p \in \mathbb{N}$. \square

Remark 1. It is clear that $m(\mathbb{G}) = 0$ if and only if $\Pi(\mathbb{U}) = 0$, where m and Π are the Lebesgue measures in \mathbb{R}^p and on the unit sphere \mathbb{S} , respectively.

When all moments of \mathbf{X} exist, let $\mathbf{W} = (W_1, \dots, W_p)^T$ be a normal random vector having the same mean and covariance matrix as \mathbf{X} and define the following moment equations:

$$g_n(\mathbf{u}) = E\{(\mathbf{u}^T \mathbf{X})^n\} - E\{(\mathbf{u}^T \mathbf{W})^n\}, \quad \mathbf{u} = (u_1, \dots, u_p)^T \in \mathbb{R}^p, \quad n \in \mathbb{N}. \tag{4}$$

Let \mathbb{H}_n be the set of solutions to the above moment equations $g_n = 0$, that is,

$$\mathbb{H}_n = \{\mathbf{u} \in \mathbb{R}^p \mid g_n(\mathbf{u}) = 0\}, \quad n \in \mathbb{N}. \tag{5}$$

Lemma 3. *Using the notation in (2), (4), (5), if all moments of \mathbf{X} exist, then $\mathbb{G} = \bigcap_{n \geq 1} \mathbb{H}_n$. Moreover, for each n , either $m(\mathbb{H}_n) = 0$ or $\mathbb{H}_n = \mathbb{R}^p$.*

Proof. $\mathbb{G} = \bigcap_{n \geq 1} \mathbb{H}_n$ follows from the fact that a univariate normal distribution is determined by its moments. When all moments of \mathbf{X} exist, $g_n(\mathbf{u})$ is a homogeneous multivariate polynomial in u_1, \dots, u_p with degrees at most n . If g_n is the zero function, then $\mathbb{H}_n = \mathbb{R}^p$. If g_n is not the zero function, then for any fixed $(u_1, \dots, u_{p-1})^T$, there are at most n values of u_p such that $(u_1, \dots, u_p) \in \mathbb{H}_n$ by the fundamental theorem of algebra (i.e. a polynomial of degree n has at most n solutions). Thus $m(\mathbb{H}_n) = 0$, because we define $\mathbb{H}_n(u_1, \dots, u_{p-1}) = \{u_p \in \mathbb{R} \mid (u_1, \dots, u_p)^T \in \mathbb{H}_n\}$, which is a finite set in this case. Let m_1 be the Lebesgue measure in \mathbb{R} ; then the Lebesgue measure m in \mathbb{R}^p is the product measure $m_1^p = m_1 \times \cdots \times m_1$. By Tonelli's theorem [2, p. 152], $m(\mathbb{H}_n) = \int_{\mathbb{R}^{p-1}} m_1 \{ \mathbb{H}_n(u_1, \dots, u_{p-1}) \} dm_1^{p-1} = 0$. \square

2.2. A new characterization of multivariate normality

If the set $\mathbb{G} = \{\mathbf{u} \in \mathbb{R}^p \mid \mathbf{u}^T \mathbf{X} \text{ is normally distributed}\}$ has positive Lebesgue measure, then, by Lemma 2, all moments of \mathbf{X} exist, and then $\mathbb{G} = \mathbb{R}^p$ by Lemma 3. On the other hand, if \mathbb{G} has zero measure, then clearly \mathbf{X} cannot be normally distributed. This yields the following theorem.

Theorem 1. *A random vector $\mathbf{X} \in \mathbb{R}^p$ with a Lebesgue density f is not normally distributed if and only if the set of normal directions, $\mathbb{U} = \{\mathbf{u} \in \mathbb{S} \mid \mathbf{u}^T \mathbf{X} \text{ is normally distributed}\}$, has measure 0, i.e., $\Pi(\mathbb{U}) = 0$.*

One might think that a set with Lebesgue measure zero is not necessarily small. For example, the set of rational numbers has Lebesgue measure zero but is dense in \mathbb{R}^p . However, \mathbb{G} here is a nowhere dense set. In particular, in the bivariate case, if \mathbf{X} is not normally distributed, \mathbb{U} not only has measure zero, but also is a finite set, as claimed by the next corollary.

Corollary 1. *If a bivariate random vector \mathbf{X} (or its density) is not normal, then \mathbf{X} has at most finitely many normal directions, i.e., $\mathbb{U} = \{\mathbf{u} \in \mathbb{S} \mid \mathbf{u}^T \mathbf{X} \text{ is normally distributed}\}$ is a finite set.*

Proof. Suppose \mathbb{U} has two or more points; the same arguments as in Lemma 2 yield that \mathbf{X} has finite moments of all orders and \mathbb{U} satisfies all the moment equations $g_n = 0$ by Lemma 3. However, if g_n is not the zero function, $g_n(\mathbf{u})$ is essentially a univariate polynomial (due to the homogeneity of g_n), which has finitely many solutions on the unit circle. Thus \mathbb{U} is a finite set if \mathbf{X} is not normal. \square

Remark 2. A result equivalent to the above corollary for the bivariate case was established previously by Hamedani and Tata [8] and also claimed as part of the results in [6]. While Ferguson [6] did not give a proof, Hamedani and Tata [8] proved the fact using characteristic functions. In particular, Theorem 3 of Hamedani and Tata [8] asserts that, given $\{(a_k, b_k), k = 1, 2, \dots\}$, a countable distinct sequence in \mathbb{R}^2 , such that for each k , $a_k X_1 + b_k X_2$ is a normal random variable, then $\mathbf{X} = (X_1, X_2)^T$ is a bivariate normal random variable. To see that this fact directly follows from the above corollary, it suffices to take $\mathbf{u}_k = (u_{1k}, u_{2k})^T$ where $u_{1k} = a_k / \sqrt{a_k^2 + b_k^2}$ and $u_{2k} = b_k / \sqrt{a_k^2 + b_k^2}$. Then $a_k X_1 + b_k X_2$ is a normal random variable if and only if $\mathbf{u}_k = (u_{1k}, u_{2k})^T$ is a normal direction of $\mathbf{X} = (X_1, X_2)^T$. However, the above result as stated in [8] for the bivariate case does not hold in three or higher dimensions as pointed out in [7]. Thus, Theorem 1 of this paper, which holds for any dimension $p \geq 2$, provides a non-straightforward generalization to the existing result for the bivariate case.

Suppose \mathbf{Y} is another random vector with Lebesgue density. If \mathbf{X} is not normally distributed, then $m(\mathbb{G}) = m(\{\mathbf{u} \in \mathbb{R}^p \mid \mathbf{u}^T \mathbf{X} \text{ is normally distributed}\}) = 0$ by Theorem 1. Thus $P(\mathbf{Y} \in \mathbb{G}) = 0$ or $P\{(\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} \mathbf{Y} \in \mathbb{U}\} = 0$, since the probability measure of \mathbf{Y} is dominated by m . Therefore we obtain the following corollary.

Corollary 2. *If a random vector \mathbf{X} is not normally distributed, then for any other random vector $\mathbf{Y} \in \mathbb{R}^p$ with a Lebesgue density, the probability of $(\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} \mathbf{Y}$ taking values of normal directions of \mathbf{X} is zero.*

Remark 3. Formal tests for multivariate normality of a random vector might be constructed on the basis of randomly selected linear combinations of its components. Suppose $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n$ is an independent random sample from an unknown density f . Then we can consider univariate data $\mathbf{X}^T \mathbf{X}_i$, $i = 1, \dots, n$, which can be viewed as projections of $\mathbf{X}_1, \dots, \mathbf{X}_n$ on \mathbf{X} . If f is not normal, then each $\mathbf{X}^T \mathbf{X}_i$, conditioned on \mathbf{X} , is not normally distributed almost surely, and thus can be tested using a consistent univariate test for normality such as the [11,9] test. By Corollary 2, such a univariate-based test would have power against any non-normal alternative density. Thus one may construct univariate tests for multivariate normality based on a randomly selected direction. Tests based such univariate projections might be found in [11] and others.

3. Concluding remarks

This paper establishes that a multivariate density is not normal if and only if its set of normal directions has Lebesgue measure zero. Consequently, the normal directions of a non-normal density are indeed quite rare. Note that this characterization of a non-normal multivariate density holds in any fixed dimension. Moreover, this new characterization is not an asymptotic result and thus its validity does not depend on typical assumptions such as a large sample size. The main finding of this paper may have some significance for the assessment of multivariate normality which is of great relevance in multivariate analysis.

Acknowledgments

The authors thank the editor and the reviewers for their valuable comments and suggestions. This research was partially supported by a research grant from the Stony Wold-Herbert Foundation (YS) and by a translational research grant NIH/NCI P30 CA 16087-24 (YS).

References

- [1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, 3rd edition, Wiley, New York, 2003.
- [2] K. Arthreya, S. Lahiri, Measure Theory and Probability Theory, Springer, New York, 2006.
- [3] P. Billingsley, Probability and Measure, 3rd edition, Wiley, New York, 1995.
- [4] R.L. Dykstra, Establishing the positive definiteness of the sample covariance matrix, The Annals of Mathematical Statistics 41 (1970) 2153–2154.
- [5] M.L. Eaton, M.D. Perlman, The non-singularity of generalized sample covariance matrices, The Annals of Statistics 1 (1973) 710–717.
- [6] T. Ferguson, On the determination of the joint distributions from the marginal distributions of linear combinations (abstract), Annals of Mathematical Statistics 30 (1959) 255.
- [7] G.G. Hamedani, Nonnormality of linear combinations of normal random variables, American Statistician 38 (1984) 295–296.
- [8] G.G. Hamedani, M.N. Tata, On the determination of the bivariate normal distribution from distributions of linear combinations of the variables, The American Mathematical Monthly 82 (1975) 913–915.
- [9] S.W. Looney, How to use tests for univariate normality to assess multivariate normality, American Statistician 49 (1995) 64–70.
- [10] F. Sinz, S. Gerwinn, M. Bethge, Characterization of the p -generalized normal distribution, Journal of Multivariate Analysis 100 (2009) 817–820.
- [11] H.C. Thode Jr., Testing for Normality, Marcel Dekker, Inc., New York, 2002.
- [12] Y.L. Tong, The Multivariate Normal Distribution, Springer-Verlag, New York, 1990.