JOURNAL OF MULTIVARIATE ANALYSIS 46, 131–138 (1993)

ML Characterization of the Multivariate Normal Distribution

W. STADJE

University of Osnabrück, Osnabrück, Germany

It is a well-known result (which can be traced back to Gauss) that the only translation family of probability densities on $\mathbb{R}$ for which the arithmetic mean is a maximum likelihood estimate of the translation parameter originates from the normal density. We generalize this characterization of the normal density to multivariate translation families. © 1993 Academic Press, Inc.

1. INTRODUCTION

The following characterization of the normal density $\varphi(x)=(2\pi)^{-1/2}\exp(-x^2/2)$, $x\in\mathbb{R}$, is well-known: Let f be a density on $\mathbb{R}$; let $X_1, ..., X_n$ be an independent sample from a distribution belonging to the translation family $f(\cdot-\theta)$, $\theta\in\mathbb{R}$. If $\bar{X}=n^{-1}(X_1+\cdots+X_n)$ is a maximum likelihood (ML) estimate of θ for $n=2, 3$, i.e., if

$$\prod_{i=1}^{n} f\left(x_i-n^{-1}\sum_{j=1}^{n} x_j\right)\geqslant\prod_{i=1}^{n} f(x_i-\theta) \qquad \text{for all} \quad x_1, ..., x_n, \theta\in\mathbb{R}, \tag{1.1}$$

then $f(x)=\sigma^{-1}\varphi(x/\sigma)$, $x\in\mathbb{R}$, for some $\sigma>0$. This property of φ can be traced back to Gauss [3], who derived it, in the context of least-squares, under the assumption that f is differentiable. The ML characterization apparently provided the first justification for the use of the normal density φ. Teicher [8] proved the result only assuming that f is lower semicontinuous at 0. Generalizing Teicher's theorem, Findeisen [2] showed that measurability of f is the only condition needed. Further extensions can be found in Stadje [6, 7]: The normal translation family on $\mathbb{R}$ can be characterized even in the class of all probability measures (not only the absolutely continuous ones) by the property that the arithmetic mean is a ML estimate, using the ML principle of Scholz [5]. Further, altering f on a Lebesgue null set (i.e., assuming (1.1) only on the complement of such a

Received August 13, 1992; revised December 15, 1992.

AMS 1980 subject classifications: primary 62E10; secondary 62F04.

Key words and phrases: multivariate normal distribution, translation family, maximum likelihood, characterization.

683 46 1-10

0047-259X/93 $5.00

Copyright © 1993 by Academic Press, Inc.
All rights of reproduction in any form reserved.

null set) does not invalidate the conclusion that $f(x) = \sigma^{-1}\varphi(x/\sigma)$ for some $\sigma > 0$.

The aim of this paper is to generalize the above characterization to the multivariate normal density. Thus let f now be a Borel-measurable non-negative function on $\mathbb{R}^d$ for some $d \in \mathbb{N}$ and let $\lambda^d(f>0)>0$, where λ^d denotes the d-dimensional Lebesgue measure. We present a proof of the following statement.

THEOREM. *Assume that for samples $x^{(1)}, ..., x^{(n)} \in \mathbb{R}^d$ of sizes $n = 2, 3, 4$ the arithmetic mean $\bar{x} = n^{-1}(x^{(1)} + \cdots + x^{(n)})$ is a ML estimate of the parameter $\theta \in \mathbb{R}^d$ of the translation family $(f(\cdot - \theta))_{\theta \in \mathbb{R}^d}$, i.e.,*

$$\prod_{i=1}^{n} f(x^{(i)} - \bar{x}) \geq \prod_{i=1}^{n} f(x^{(i)} - \theta) \qquad \text{for all} \quad \theta \in \mathbb{R}^d. \tag{1.2}$$

Then $f(x) = c \exp(-x'Ax)$, $x \in \mathbb{R}^d$, for some $c > 0$ and some non-negative definite $(d \times d)$-matrix A.

We note that if f is assumed to be positive everywhere and twice differentiable, this assertion follows from the results of Campbell [1].

2. AUXILIARY RESULTS

We will have to consider the logarithm of f, and thus we have to make sure that $f(x) > 0$ for all $x \in \mathbb{R}^d$. This is verified in the following lemma. A similar result for $d = 1$ has been proved by Findeisen [2, Section 2]. The proof is simpler even in the one-dimensional case. We use the relation

$$f(x)\, f(-x) \geq f(\theta)\, f(\theta + 2x) \qquad \text{for all} \qquad x, \theta \in \mathbb{R}^d \tag{2.1}$$

which follows from (1.2) by setting $n = 2$, $x^{(1)} = -x^{(2)} = x$ and $\theta = -x - \tilde{\theta}$. In the sequel let $\mathbf{0} = (0, ..., 0)$ be the zero vector in $\mathbb{R}^d$.

LEMMA 1. *Let $f\colon \mathbb{R}^d \to \mathbb{R}_+$ satisfy (1.2) for $n = 2, 3$. Then $f(x) = 0$ λ^d-almost everywhere or $f(x) > 0$ for all $x \in \mathbb{R}^d$.*

Proof. By (1.2), $f(\mathbf{0})^3 \geq f(\theta)^3$ for all $\theta \in \mathbb{R}^d$, so that $f(\mathbf{0}) = \max\{f(\theta) \mid \theta \in \mathbb{R}^d\}$. We may thus assume that $f(\mathbf{0}) > 0$ and (after possibly changing from f to $f/f(\mathbf{0})$) that $0 \leq f \leq 1$.

We proceed by induction on d. First let $d = 1$. Suppose we can find a sequence $x^{(k)} \in \mathbb{R}$ such that $0 \neq x^{(k)} \to 0$ and

$$f(x^{(k)})\, f(-x^{(k)}) = 0 \qquad \text{for all} \quad k \in \mathbb{N}. \tag{2.2}$$

Without restriction of generality we can choose $x^{(k)} > 0$. From (2.1) it follows that

$$f(\theta)\, f(\theta + 2x^{(k)}) = 0 \qquad \text{for all} \quad \theta \in \mathbb{R} \qquad \text{and} \qquad k \in \mathbb{N}. \tag{2.3}$$

Fix $\theta_0 \in \mathbb{R}$ and consider the intervals $I_k = [\theta_0, \theta_0 + 3x^{(k)}]$. Then for every $x \in [\theta_0, \theta_0 + x^{(k)}]$ we have x, $x + 2x^{(k)} \in I_k$ and, by (2.3), $f(x) = 0$ or $f(x + 2x^{(k)}) = 0$. Thus, any I_k contains a Borel-measurable subset B_k of measure at least $\lambda^1(I_k)/3$ satisfying $f \mid B_k \equiv 0$. Consequently, the set $\{f > 0\}$ has a Lebesgue density of at most 2/3 at every point $\theta_0 \in \mathbb{R}$. By Lebesgue's density theorem, $\{f > 0\}$ is a Lebesgue null set.

If no sequence $x^{(k)}$ as above exists, there is an $\varepsilon > 0$ such that $f(x) > 0$ for $|x| < \varepsilon$. Then note that $0 \leqslant f \leqslant 1$ implies that, for any $x^{(1)}, x^{(2)}, x^{(3)} \in \mathbb{R}$,

$$f\left(x^{(1)} - \frac{1}{3}(x^{(1)} + x^{(2)} + x^{(3)})\right) \geqslant \prod_{i=1}^{3} f\left(x^{(i)} - \frac{1}{3}(x^{(1)} + x^{(2)} + x^{(3)})\right)$$

$$\geqslant f(x^{(1)})\, f(x^{(2)})\, f(x^{(3)}). \tag{2.4}$$

Setting $x^{(1)} = x$ and $x^{(2)} = x^{(3)} = -x$ in (2.4) we obtain

$$f(4x/3) \geqslant f(x)\, f(-x)^2. \tag{2.5}$$

Thus, $f(x) > 0$ for all $|x| < \varepsilon$ entails $f(x) > 0$ for all $|x| < 4\varepsilon/3$. Iterating this argument yields $f(x) > 0$ for all $x \in \mathbb{R}$.

Now let $d \geqslant 2$ and assume the assertion holds for all $d' < d$. Obviously it is sufficient to show that $f \equiv 0$ λ^d-a.e. or $f(x) > 0$ for all x in some neighborhood of $\mathbf{0}$. Suppose that there is a sequence of points $x^{(k)} \in \mathbb{R}^d$ such that $x_i^{(k)} \neq 0$, $i = 1, ..., d$, and $f(x^{(k)})\, f(-x^{(k)}) = 0$ for all $k \in \mathbb{N}$. As above we conclude from (2.1) that

$$f(\theta + 2x^{(k)}) = 0 \qquad \text{for all} \quad \theta \in \mathbb{R}^d,\ k \in \mathbb{N}. \tag{2.6}$$

Any d-dimensional rectrangle of the form

$$I = \prod_{i=1}^{d} [\theta_i - 2\,|x_i^{(k)}|,\ \theta_i + 3\,|x_i^{(k)}|]$$

contains a Borel-measurable subset of measure at least $(1/5)^d \lambda^d(I)$ on which f is positive, because for any $x \in \prod_{i=1}^{d} [\theta_i, \theta_i + |x_i^{(k)}|]$ we have $f(x) = 0$ or $f(x + 2x^{(k)}) = 0$ by (2.6). Therefore

$$\lambda^d(I \cap \{f > 0\})/\lambda^d(I) \leqslant 1 - (1/5)^d < 1,$$

and Lebesgue's density theorem yields $\lambda^d(f > 0) = 0$.

If no sequence $x^{(k)}$ with the above properties exists, there is a neighborhood U of $\mathbf{0}$ such that $f(x)>0$ for all $x\in U$ having non-vanishing components. But if $x\in U$ has some components equal to 0, we still have $f(x)f(-x)\geqslant f(\theta)f(\theta+2x)$ for all $\theta\in\mathbb{R}^d$, and it is easy to find a $\theta\in U$ such that $\theta+2x$ is also in U and θ and $\theta+2x$ have non-vanishing components. Then $f(\theta)f(\theta+2x)>0$, implying that $f(x)>0$. It follows that $U\subset\{f>0\}$. The Lemma is proved.

The proof of the next Lemma is straightforward and therefore omitted.

LEMMA 2. *Any monotone function* $\alpha:\mathbb{Q}\setminus\{0\}\to\mathbb{R}$ *satisfying*

$$\alpha(u+v)-\alpha(u)=\alpha(u'+v)-\alpha(u')$$

for all u, u', $v\in\mathbb{Q}\setminus\{0\}$, $u\neq -v\neq u'$, *is affine.*

3. PROOF OF THE THEOREM

By Lemma 1 we can define a real-valued function g by setting $g=-\ln f$. Then (1.1) is clearly equivalent to the implication

$$\sum_{i=1}^{n} x^{(i)}=\mathbf{0}\Rightarrow\sum_{i=1}^{n} g(x^{(i)})\leqslant\sum_{i=1}^{n} g(x^{(i)}-\theta)\qquad\text{for all}\quad\theta\in\mathbb{R}^d,\tag{3.1}$$

where $n=2$, 3, 4. The function $h(x)=g(x)+g(-x)$ is even and satisfies also (3.1). Moreover, the function $x_1\mapsto h(x_1,...,x_d)$ is midconvex, i.e.,

$$h((x_1+x_1')/2, x_2, ..., x_d)\leqslant\frac{1}{2}(h(x_1, x_2, ..., x_d)+h(x_1', x_2, ..., x_d))\tag{3.2}$$

for all x_1, x_1', x_2, ..., $x_d\in\mathbb{R}$. To see (3.2), set $n=2$, $x^{(1)}=x=-x^{(2)}$ in (3.1) and add the inequalities

$$h(x)\leqslant g(x-\theta)+g(-x-\theta),\ h(-x)\leqslant g(-x+\theta)+g(x+\theta).$$

This yields $2h(x)\leqslant h(x+\theta)+h(x-\theta)$ for all $x\in\mathbb{R}^d$, which is tantamount to (3.2).

A Borel-measurable midconvex function is convex (Roberts and Varberg [4, Ch. VII]). It follows that for fixed x_2, ..., x_d the derivative $D_1h=\partial h/\partial x_1$ exists everywhere except at an at most countable set of points. D_1h is non-decreasing and has positive jumps at the points where it is not defined. Let $D(x_2^{(0)}, ..., x_d^{(0)})$ be the set of all rational multiples of the points x_1 for which

$D_1 h$ does not exist at $(x_1, x_2^{(0)}, ..., x_d^{(0)})$. Note that $0 \in D(x_2, ..., x_d) = D(-x_2, ..., -x_d)$, because h is even, and that every set $D(x_2, ..., x_d)$ is countable.

Now fix $x_2, ..., x_d$. We choose an $u_0 \in (0, \infty) \backslash D(x_2, ..., x_d)$. For rational numbers $p, q, r, s \neq 0$ satisfying $p+q+r+s=0$ we define

$$\begin{aligned} H(u) = {} & h(pu_0+u, x_2, ..., x_d) + h(-qu_0-u, x_2, ..., x_d) \\ & + h(ru_0+u, x_2, ..., x_d) + h(-su_0-u, x_2, ..., x_d). \end{aligned}$$

Since $pu_0, qu_0, ru_0, su_0 \in \mathbb{R} \backslash D(x_2, ..., x_d) = \mathbb{R} \backslash D(-x_2, ..., -x_d)$, the function H is differentiable at $u=0$. Furthermore, H has a minimum at $u=0$, because h satisfies (3.1) for $n=4$, is even and $pu_0+qu_0+ru_0+su_0=0$ (here we need the sample size 4). It follows that

$$\begin{aligned} 0 = H'(0) = {} & \frac{\partial}{\partial u} (h(pu_0+u, x_2, ..., x_d) + h(-qu_0-u, x_2, ..., x_d) \\ & + h(ru_0+u, x_2, ..., x_d) + h(-su_0-u, x_2, ..., x_d))|_{u=0} \\ = {} & D_1 h(pu_0, x_2, ..., x_d) - D_1 h(-qu_0, x_2, ..., x_d) \\ & + D_1 h(ru_0, x_2, ..., x_d) - D_1 h(-su_0, x_2, ..., x_d). \end{aligned} \tag{3.3}$$

By (3.3),

$$\begin{aligned} & D_1 h((u+v)u_0, x_2, ..., x_d) - D_1 h(uu_0, x_2, ..., x_d) \\ & = D_1 h((u'+v)u_0, x_2, ..., x_d) - D_1 h(u'u_0, x_2, ..., x_d) \end{aligned}$$

for all rational $u, u', v \in \mathbb{Q}$ such that $u, u' \neq 0$, $u \neq -v \neq u'$. Moreover, $D_1 h$ is monotone non-decreasing on $\mathbb{Q}u_0 \backslash \{0\}$. By Lemma 2, $D_1 h$ is linear on $\mathbb{Q}u_0 \backslash \{0\}$. But at any jump point u of $D_1 h$ the left-hand and the right-hand derivative of h (which exist as $h(\cdot, x_2, ..., x_d)$ is convex) lie between

$$\lim_{\substack{u' \uparrow u \\ u' \in D(x_2, ..., x_d)^c}} D_1 h(u', x_2, ..., x_d) \quad \text{and} \quad \lim_{\substack{u' \downarrow u \\ u' \in D(x_2, ..., x_d)^c}} D_1 h(u', x_2, ..., x_d),$$

and these limits coincide. Thus, h is everywhere partially differentiable with respect to x_1, the derivative being given by

$$D_1 h(x) = [D_1 h(1, x_2, ..., x_d) - D_1 h(0, x_2, ..., x_d)]\, x_1 + D_1 h(0, x_2, ..., x_d).$$

Next consider, for any given $x \in \mathbb{R}^d$, the three points $x^{(1)} = x$, $x^{(2)} = (0, -x_2, ..., -x_d)$, $x^{(3)} = (-x_1, 0, ..., 0)$. Then $x^{(1)} + x^{(2)} + x^{(3)} = 0$, so that

$$\begin{aligned} 0 &= D_1 h(x^{(1)}) + D_1 h(x^{(2)}) + D_1 h(x^{(3)}) \\ &= D_1 h(x) - D_1 h(0, x_2, ..., x_d) - D_1 h(x_1, 0, ..., 0), \end{aligned} \tag{3.4}$$

since $D_1 h(x) = -D_1 h(-x)$. The function $x_1 \mapsto D_1 h(x_1, 0, ..., 0)$ is linear so that $h(x_1, 0, ..., 0) = \alpha^2 x_1^2 + \beta x_1 + \gamma$; since $x_1 = 0$ is a maximum of $f(x_1, 0, ..., 0)$ and thus a minimum of $h(x_1, 0, ..., 0)$, it follows that $\beta = 0$. By assumption we have $f(\mathbf{0}) = 1$, so that also $\gamma = 0$. Hence we obtain $\alpha^2 = h(1, 0, ..., 0)$ and, by (3.4),

$$D_1 h(x) = 2h(1, 0, ..., 0)\, x_1 + D_1 h(0, x_2, ..., x_d). \tag{3.5}$$

Since $D_1 h$ is convex in x_1, we also have

$$D_1 h(1, 0, ..., 0) \geqslant 0. \tag{3.6}$$

By induction on d we can now prove the following assertion: For any even function $h: \mathbb{R}^d \to \mathbb{R}$ satisfying (3.1) (with g replaced by h) the function $h(x) - h(\mathbf{0})$ is a non-negative definite quadratic form. For $d = 1$ this follows immediately from (3.5) and (3.6).

Integrating (3.5) with respect to x_1 yields

$$h(x) = h(1, 0, ..., 0)\, x_1^2 + D_1 h(0, x_2, ..., x_d)\, x_1 + h(0, x_2, ..., x_d). \tag{3.7}$$

Now assume the assertion is true for $d - 1$, where $d \geqslant 2$. Obviously, the function $\bar{h}(x_2, ..., x_d) = h(0, x_2, ..., x_d) - h(\mathbf{0})$ is even and satisfies (3.1) so that the induction hypothesis can be applied to $\bar{h}$. Thus, $\bar{h}$ is a non-negative definite quadratic form in $x_2, ..., x_d$. By (3.7),

$$\begin{aligned} &h(x_1, ..., x_d) - h(\mathbf{0}) \\ &\quad = h(1, 0, ..., 0)\, x_1^2 + D_1 h(0, x_2, ..., x_d)\, x_1 + \bar{h}(x_2, ..., x_d). \end{aligned} \tag{3.8}$$

Setting $x_1 = x_2$ in (3.8) we obtain

$$\begin{aligned} D_1 h(0, x_2, ..., x_d) x_2 = {} & h(x_2, x_2, x_3, ..., x_d) - h(\mathbf{0}) \\ & - \bar{h}(x_2, ..., x_d) - h(1, 0, ..., 0) x_2^2. \end{aligned} \tag{3.9}$$

Now note that we can also apply the induction hypothesis to the function $\hat{h}(x_2, ..., x_d) = h(x_2, x_2, x_3, ..., x_d) - h(\mathbf{0})$ because $\hat{h}$ is even and clearly satisfies (3.1) with d replaced by $d - 1$. It follows that $\hat{h}$ is a quadratic form of the $d - 1$ variables $x_2, ..., x_d$. Thus (3.9) shows that $D_1 h(0, x_2, ..., x_d) x_2$ is a quadratic form of $x_2, ..., x_d$, say $\sum_{2 \leqslant i \leqslant j \leqslant d} b_{ij} x_i x_j$. Since $D_1 h(0, x_2, ..., x_d)$ remains bounded as $x_2 \to 0$ (for any fixed $x_3, ..., x_d$), the function $(1/x_2) \sum_{2 \leqslant i \leqslant j \leqslant d} b_{ij} x_i x_j$ remains bounded as $x_2 \to 0$, so that $b_{ij} = 0$ for $i > 2$ and $j > 2$. It follows that $D_1 h(0, x_2, ..., x_d)$ is a linear form of $x_2, ..., x_d$. From (3.8) we can now conclude that $h(x_1, ..., x_d) - h(\mathbf{0})$ is a quadratic form of $x_1, ..., x_d$. Note that h has a minimum at $\mathbf{0}$ so that $h(x) - h(\mathbf{0}) \geqslant 0$.

Finally we have to return to g. Clearly,

$$\begin{aligned} h(x) &= g(x_1, ..., x_d) + g(-x_1, ..., -x_d) \\ &\leqslant g(x_1 + \theta_1, x_2, ..., x_d) + g(-x_1 + \theta_1, -x_2, ..., -x_d) \\ &= g(x_1 + \theta_1, x_2, ..., x_d) + h(x_1 - \theta_1, x_2, ..., x_d) \\ &\quad - g(x_1 - \theta_1, x_2, ..., x_d) \end{aligned} \tag{3.10}$$

for all $\theta_1 \in \mathbb{R}$. Replacing θ_1 by $-\theta_1$ in (3.10) we get

$$\begin{aligned} h(x) &\leqslant g(x_1 - \theta_1, x_2, ..., x_d) + g(-x_1 - \theta_1, -x_2, ..., -x_d) \\ &= g(x_1 - \theta_1, x_2, ..., x_d) + h(x_1 + \theta_1, x_2, ..., x_d) \\ &\quad - g(x_1 + \theta_1, x_2, ..., x_d). \end{aligned} \tag{3.11}$$

By (3.10) and (3.11),

$$\begin{aligned} h(x_1, ..., x_d) &- h(x_1 - \theta_1, x_2, ..., x_d) \\ &\leqslant g(x_1 + \theta_1, x_2, ..., x_d) - g(x_1 - \theta_1, x_2, ..., x_d) \\ &\leqslant h(x_1 + \theta_1, x_2, ..., x_d) - h(x_1, ..., x_d). \end{aligned} \tag{3.12}$$

The inequalities (3.12) show that g is partially differentiable with respect to x_1 and that $\partial g/\partial x_1 = \frac{1}{2}(\partial h/\partial x_1)$. The same argument applies to the variables $x_2, ..., x_d$, so that g has continuous partial derivatives given by

$$\partial g/\partial x_i = \frac{1}{2} (\partial h/\partial x_i), \qquad i = 1, ..., d.$$

Hence $g = h/2$. The proof is complete.

References

[1] Campbell, L. L. (1970). Equivalence of Gauss's principle and minimum discrimination information estimation of probabilities. *Ann. Math. Statist.* **41** 1011–1015.

[2] Findeisen, P. (1982). Die Charakterisierung der Normalverteilung nach Gauß. *Metrika* **29** 55–64.

[3] Gauss, C. F. Theoria motus corporum coelestium. In *Werke*, liber II, Sectio III, 240–244.

[4] Roberts, A. W., and Varberg, D. E. (1973). *Convex Functions.* Academic Press, New York/London.

[5] SCHOLZ, F. W. (1980). Towards a unified definition of maximum likelihood. *Canad. J. Statist.* **8** 193–203.

[6] STADJE, W. (1987). An extension theorem for convex functions and an application to Teicher's characterization of the normal distribution. *Mathematika* **34** 155–159.

[7] STADJE, W. (1988). A generalized maximum likelihood characterization of the normal distribution. *Metrika* **35** 93–97.

[8] TEICHER, H. (1961). Maximum likelihood characterization of distributions. *Ann. Math. Statist.* **32** 1214–1222.