MAHALANOBIS' DISTANCE BEYOND NORMAL DISTRIBUTIONS

JOAKIM EKSTRÖM

ABSTRACT. Based on the reasoning expressed by Mahalanobis in his original article, the present article extends the Mahalanobis distance beyond the set of normal distributions. Sufficient conditions for existence and uniqueness are studied, and some properties derived. Since many statistical methods use the Mahalanobis distance as e vehicle, e.g. the method of least squares and the chi-square hypothesis test, extending the Mahalanobis distance beyond normal distributions yields a high ratio of output to input, because those methods are then instantly generalized beyond the normal distributions as well. Mahalanobis' idea also has a certain conceptual beauty, mapping random variables into a frame of reference which ensures that apples are compared to apples.

1. Introduction

Distances have been used in statistics for centuries. Carl Friedrich Gauss (1809) proposed using sums of squares, a squared Euclidean distance, for the purpose of fitting Kepler orbits to observations of heavenly bodies. Karl Pearson (1900) proposed a weighted Euclidean distance, which he denoted χ, for the construction of a hypothesis test. Both Gauss' and Pearson's methods presume statistically independent and normally distributed observational errors, and therefore their distances are nowadays recognized as special cases of the Mahalanobis distance.

Proposed by Mahalanobis (1936), the Mahalanobis distance is a distance that accounts for probability distribution and equals the Euclidean distance under the standard normal distribution. The rationale for the latter fact is largely historical, but both Gauss (1809) and Pearson (1900) discussed the fact that the density function of a standard normally distributed random variable equals (a normalizing constant times) the composition $g \circ ||\cdot||$, where $||\cdot||$ is the Euclidean norm and g the Gaussian function $e^{-x^2/2}$.

Mahalanobis, who was trained as a physicist, explained his proposed distance by making parallels with Galilean transformations. The Galilean transformation is a standard method in physics that maps all coordinates into a frame of reference and thus ensures that apples are compared to apples, using the common expression. The Galilean transformation is conceptually beautiful and has the property that exceedingly complex problems often become trivially simple when evaluated within the frame of reference. The well-known Mahalanobis distance does precisely this, with the standard normal distribution as frame of reference. Indeed, the Mahalanobis distance is simply the composition of a Galilean transformation and the Euclidean distance. Notice in particular that given a random variable $U \sim \mathrm{N}(\mu, \Sigma)$, the transformation $T(x) = \Sigma^{-1/2}(x - \mu)$ transforms U into a standard normally distributed random variable and the Mahalanobis distance equals

$$d(x, y) = ||T(x) - T(y)||,$$

which is easy to verify.

Using this simple little observation, the present article aims to study Mahalanobis' idea beyond normal distributions. The approach (*ansatz*) is to firstly allow arbitrary distributions and then, secondly, to impose conditions which imply existence and uniqueness of the distance. A transformation which maps a random variable with distribution $\mathcal{F}$ to a random variable with distribution $\mathcal{G}$ is conveniently denoted $T : \mathcal{F} \mapsto \mathcal{G}$. Also, the calligraphy capital $\mathcal{N}$ denotes the multivariate standard normal distribution, of a dimension which is clear from the context. The *ansatz-definition* is the following.

Definition 1. The Mahalanobis distance under the distribution $\mathcal{F}$ is defined

$$d(x, y) = ||T(x) - T(y)||, \tag{1}$$

where $T : \mathcal{F} \mapsto \mathcal{N}$, and $||\cdot||$ denotes the Euclidean norm.

The transformation of Definition 1 is occasionally referred to as the Mahalanobis transformation. Suppose that the Mahalanobis transformations in Definition 1 are required to be affine, then it follows that the definition is identical to Mahalanobis' original definition (1936). This article aims to find weaker conditions that allow for distributions beyond the normal distributions.

Of great interest is of course existence and uniqueness of the Mahalanobis distance. The Mahalanobis distance exists if the right hand side of Expression (1) can be computed, i.e. if a transformation $T : \mathcal{F} \mapsto \mathcal{N}$ exists. The distance is unique, moreover, if for any two transformations $T_1, T_2 : \mathcal{F} \mapsto \mathcal{N}$ the right hand sides of Expression (1) are equal. Conditions implying existence are discussed in Section 2 and conditions implying uniqueness are discussed in Section 3. Section 4 discusses some properties as well as some applications of the distance.

2. Existence

The aim of the present section is to show sufficient conditions for existence of a Mahalanobis distance, i.e. conditions on the distribution $\mathcal{F}$ sufficient for existence of a transformation T that maps a such distributed random variable into a standard normally distributed random variable, $T : \mathcal{F} \mapsto \mathcal{N}$. The conditions are derived constructively by providing such a transformation, named the *conditional distribution transformation* and denoted ϕ.

First a few notes on notational conventions. If $(x, y) \in \mathbb{S} \times \mathbb{T}$ then π_1 denotes projection onto the first factor, i.e. $\pi_1 : (x, y) \mapsto x$. If $(x_1, \ldots, x_p) \in \mathbb{R}^p$, then π_k denotes projection onto the product of the first k factors, i.e. $\pi_k : (x_1, \ldots, x_p) \mapsto (x_1, \ldots, x_k)$. If $A \subset \mathbb{S} \times \mathbb{T}$ and $x \in \mathbb{S}$, then A_x denotes the section of A at x, i.e. $\{y \in \mathbb{T} : (x, y) \in A\}$. If f is a function defined on a product space, $f : (x, y) \mapsto z$, say, then f_x denotes the section of f at a fixed x, so if $(x, y) \in A \subset \mathbb{S} \times \mathbb{T}$ and f is defined on A then f_x is defined on $A_x \subset \mathbb{T}$, $f_x : y \mapsto f(x, y)$. If g is defined on $\mathbb{R}^k$ and $x \in \mathbb{R}^p$, the notation $g(x)$ is allowed for convenience and should be read as $g \circ \pi_k(x)$, i.e. the composition with projection is implicitly understood. The calligraphy capitals $\mathcal{N}$ and $\mathcal{U}$ denote the standard normal distribution and the standard uniform distribution respectively. While there are infinite-dimensional Gaussian spaces, normal distributions are in this article presumed to be finite-dimensional, i.e. the normally distributed random variables take values on (a space which can be identified with) $\mathbb{R}^p$. Normal distributions are further presumed to be non-degenerate.

Suppose that an arbitrary distribution $\mathcal{F}$ has a density function $f : \mathbb{R}^p \to \mathbb{R}$. For $k = 1, \ldots, p$, let $F_k : \mathbb{R}^k \to \mathbb{R}$ be defined recursively by $F_p = f$ and

$$F_{k-1}(x_1, \ldots, x_{k-1}) = \int_{\mathbb{R}} F_k(x_1, \ldots, x_{k-1}, t)dt. \tag{2}$$

Note in particular that $F_0 = \int_{\mathbb{R}^p} f d\lambda = 1$, where λ denotes the Lebesgue measure of a dimensionality that generally is understood from the context. Furthermore, F_k is non-negative and, by the Fubini theorem, finite at almost every point of its domain. For $k = 1, \ldots, p$, let the component $\phi_k : \mathbb{R}^k \to \mathbb{R}$ be defined by

$$\phi_k(x_1, \ldots, x_k) = \frac{\int_{-\infty}^{x_k} F_k(x_1, \ldots, x_{k-1}, t)dt}{F_{k-1}(x_1, \ldots, x_{k-1})}, \tag{3}$$

for all $(x_1, \ldots, x_{k-1})$ such that F_{k-1} is positive and finite and defined $\phi_k = 0$ otherwise. The conditional distribution transformation $\phi : \mathbb{R}^p \to \mathbb{I}^p$ is then defined via its components, i.e. $\phi = (\phi_1, \ldots, \phi_p)$, the blackboard bold $\mathbb{I}$ denotes the unit interval and $\mathbb{I}^p$, consequently, the unit cube.

While the conditional distribution transformation in the present article serves a theoretical purpose, it lends itself suitable also for practical purposes. Given only the density function, a machine can easily compute values of the conditional distribution transformation through numerical integration.

Showing existence in the univariate case, i.e. $p = 1$, is particularly simple. Note that in this case the conditional distribution transformation reduces to the distribution function. This suggests the mapping $\mathcal{F} \mapsto \mathcal{U} \mapsto \mathcal{N}$ which is achieved through composition with the inverse standard normal distribution function, Φ^{-1}. Thus, if only the density function exists the composition $\Phi^{-1} \circ \phi$ can be used for the purpose of Definition 1. Consequently, a sufficient condition for existence of a Mahalanobis distance in the univariate case is simply that the distribution is absolutely continuous. Furthermore, if the density function f is continuous on an open subset S of its support, denoted $\operatorname{supp}(f)$, the conditional distribution transformation is continuously differentiable on S, $\phi \in C^1(S)$, increasing on S, and thus by the inverse function theorem also injective on S.

In the general multivariate case the mapping $\mathcal{F} \mapsto \mathcal{U} \mapsto \mathcal{N}$ is also achieved through the composition $\Phi^{-1} \circ \phi$, however stronger conditions are required and the mathematical details are more technical. The following four lemmas contain properties of the conditional distribution transformation; for enhanced readability the proofs are in an appendix.

Let $(f > t)$ denote the set $\{x : f(x) > t\}$, and let $\operatorname{Int}(A)$ and $\operatorname{Cl}(A) = \bar{A}$ denote the interior and closure of A, respectively. The set A is a continuity set if its boundary, ∂A, has probability zero. If f is a density function, then $(f > 0)$ is a continuity set if $\int_{\partial(f>0)} f d\lambda = 0$, and if so then $\operatorname{Int}(f > 0)$ and $\operatorname{Cl}(f > 0) = \operatorname{supp}(f)$ both have probability one.

Lemma 1. *Suppose f is a density function, then the corresponding conditional distribution transformation is injective at almost every point of* $\operatorname{Int}(f > 0)$.

Lemma 2. *Suppose f is a density function and* $(f > 0)$ *a continuity set, then the image of* $\operatorname{Int}(f > 0)$ *under the corresponding conditional distribution transformation,* $\phi(\operatorname{Int}(f > 0))$, *has Lebesgue measure one, and consequently contains almost every point of the image of the space, $\phi(\mathbb{R}^p) \subset \mathbb{I}^p$. Furthermore, the inverse, ϕ^{-1}, is well-defined at almost every point of $\mathbb{I}^p$.*

Lemma 3. *Suppose f is a density function and $(f > 0)$ a continuity set, then the corresponding conditional distribution transformation maps sets of probability zero to sets of Lebesgue measure zero.*

If $x \in \mathbb{R}^k$ and $f_x(t) = f(x, t)$, the section f_x is locally dominated if there is an integrable function g such that $|f_y| \leq g$ a.e. for all y in some ball $B_r(x)$ of x, $r > 0$.

Lemma 4. *Suppose f is a density function that is continuous on an open subset S of $\operatorname{supp}(f)$ such that $\lambda(\operatorname{supp}(f) \setminus S) = 0$. If $\lambda((\partial S)_x) = 0$ for all $x \in \pi_k(S)$, $k = 1, \ldots, p-1$, then ϕ is continuous on S. Furthermore, if $f \in C^1(S)$ and the sections $(Df)_x$ are locally dominated for all $x \in \pi_k(S)$, $k = 1, \ldots, p-1$, then $\phi \in C^1(S)$.*

Remark 1. The condition $\lambda((\partial S)_x) = 0$ for all $x \in \pi_k(S)$, $k = 1, \ldots, p-1$, while looking technical, is in practice often not very difficult to verify. For example, if S is convex then the condition follows immediately, which is easily seen from a standard contradiction argument.

The following theorem shows sufficient conditions for the conditional distribution transformation to map random variables the way it is designed to, $\phi : \mathcal{F} \mapsto \mathcal{U}$.

Theorem 5. *Suppose $f \in C^1(S)$ is a density function and S an open subset of $\operatorname{supp}(f)$ such that $\lambda(\operatorname{supp}(f) \setminus S) = 0$, and the sections $(Df)_x$ are locally dominated and $\lambda((\partial S)_x) = 0$ for all $x \in \pi_k(S)$, $k = 1, \ldots, p-1$. Then ϕ maps a such distributed random variable into one uniformly distributed on the unit cube.*

Proof. The proof uses the change-of-variables theorem (Rudin, 1987). By Lemmas 1, 3 and 4, ϕ is continuous, differentiable and almost everywhere injective on S, and ϕ maps sets of measure zero to sets of measure zero. This completes the verification of the conditions for the change-of-variables theorem.

Since ϕ_1 is a constant function of $x_2, \ldots, x_p$, the Jacobian determinant of ϕ equals $\prod_{k=1}^{p} \partial\phi_k / \partial x_k$. Moreover, ϕ_k is a quotient where the denominator is constant as a function of x_k and the partial derivative of the numerator is F_k. Thus,

$$J_\phi = \prod_{k=1}^{p} \frac{F_k}{F_{k-1}} = \frac{F_p}{F_0} = f.$$

Since f is positive, the absolute value of the Jacobian determinant of ϕ also equals f.

Let the random variable U have density function f and let V be a random variable uniformly distributed on the unit cube, $\mathbb{I}^p$. Note that by Lemmas 2 and 3 it holds that

$\mathbb{1}_{\phi(S)} = \mathbb{1}_{\mathbb{I}^p}$ a.e., where $\mathbb{1}$ denotes the indicator function. For an arbitrary Borel set B,

$$P(U \in B) = \int_{\mathbb{R}^p} \mathbb{1}_B f d\lambda = \int_S (\mathbb{1}_{\phi(B)} \circ \phi)|J_\phi| d\lambda = \int_{\mathbb{I}^p} \mathbb{1}_{\phi(B)} d\lambda = P(V \in \phi(B)).$$

The third equality above is, of course, the change-of-variables theorem. Hence it follows that, for any Borel set B,

$$P(\phi(U) \in B) = P(U \in \phi^{-1}(B)) = P(V \in \phi(\phi^{-1}(B))) = P(V \in B),$$

and thus the random variable $\phi(U)$ is uniformly distributed on the unit cube. □

In the multidimensional case, let $\Phi^{-1} : \mathbb{I}^p \to \mathbb{R}^p$ be the function which transforms each coordinate by the inverse standard normal distribution function. Then $\Phi^{-1} : \mathcal{U} \mapsto \mathcal{N}$, and consequently composition yields the transformation $\Phi^{-1} \circ \phi : \mathcal{F} \mapsto \mathcal{N}$, also in the general multivariate case. This transformation can of course be used for the purpose of Definition 1, and thus the hypothesis of Theorem 5 gives sufficient conditions for existence of a Mahalanobis distance. The following corollary is the main result of the section.

Corollary 6. *Under the hypothesis of Theorem 5 a Mahalanobis distance exists. In particular, the composition $\Phi^{-1} \circ \phi$ can be used as a transformation which transforms the random variable into a standard normally distributed one.*

3. Uniqueness

The present section aims to show conditions under which a change of transformation $\mathcal{F} \mapsto \mathcal{N}$ does not change the Mahalanobis distance as given by Expression (1), i.e. conditions on $\mathcal{F}$ and T under which the following diagram commutes in the sense that the Mahalanobis distance is unaltered.

$$\begin{tikzcd}
\mathcal{F} \arrow[r, "\phi"] \arrow[dr, "T"'] & \mathcal{U} \arrow[d, "\Phi^{-1}"] \\
 & \mathcal{N}
\end{tikzcd}$$

It suffices to show that the composition $T \circ \phi^{-1} \circ \Phi : \mathcal{N} \mapsto \mathcal{N}$ is an isometry under the Euclidean metric. Note that the inverse of the conditional distribution transformation ϕ exists at almost every point of $\mathbb{I}^p$ by Lemma 2. The following are useful general facts about isometries.

Lemma 7. *A function that is not continuous and injective is not an isometry.*

Lemma 8. *A transformation $G : \mathcal{N} \mapsto \mathcal{N}$ is an isometry if and only if it is orthogonal.*

The univariate case is a simple and important special case.

Lemma 9. *If the standard uniform distribution, $\mathcal{U}$, is univariate and $G : \mathcal{U} \mapsto \mathcal{U}$ is monotonic, then either $G(x) = x$ or $G(x) = 1 - x$.*

The following lemma pre-emptively clarifies what might potentially appear as contradictory, and it is also a univariate partial converse to Lemma 7.

Lemma 10. *If the standard uniform distribution, $\mathcal{U}$, is univariate and $G : \mathcal{U} \mapsto \mathcal{U}$ is injective almost everywhere and continuous, then G is injective and satisfies either $G(x) = x$ or $G(x) = 1 - x$.*

The following theorem yields necessary and sufficient conditions for uniqueness of the Mahalanobis distance in the univariate case. A property holds with probability one if the set of points where the property does not hold has probability zero.

Theorem 11. *If the univariate distribution $\mathcal{F}$ is absolutely continuous, then the Mahalanobis distance is unique if and only if transformations $T : \mathcal{F} \mapsto \mathcal{N}$ are injective with probability one and continuous.*

Proof. By Lemma 2, ϕ^{-1} is well-defined on a set, denoted M, which contains almost every point of $\mathbb{I}$. Thus $\phi^{-1}|_M : \mathcal{U} \mapsto \mathcal{F}$. Let $G = \Phi \circ T \circ \phi^{-1}$, then $G|_M : \mathcal{U} \mapsto \mathcal{U}$. The transformation $\phi^{-1}|_M$ is monotonic and Φ is continuous, injective and monotonic.

Under the assumptions, T is monotonic and therefore $G|_M$ is also monotonic. By Lemmas 9 and 7, it follows that $G|_M$ is continuous and therefore it can be naturally extended to $\mathbb{I}$ by continuity. In fact, for all x, $T(\phi^{-1}(\{\phi(x)\})) = \{T(x)\}$, i.e. the function $T \circ \phi^{-1}$ is well defined for all $y \in \mathbb{I}$. By Lemma 9 and symmetry of Φ^{-1}, it holds

$$||\Phi^{-1} \circ G(x) - \Phi^{-1} \circ G(y)|| = ||\Phi^{-1}(x) - \Phi^{-1}(y)||,$$

and therefore the Mahalanobis distance exists and is equal for all T.

Conversely, if T is not injective with probability one, then nor is $G|_M$. Consequently, by Lemma 7, $G|_M$ is not an isometry, and nor its extension G. If T has a discontinuity at x then so has $||T(\cdot)||$, but then it cannot equal $||\Phi^{-1} \circ \phi(\cdot)||$ because the latter is continuous, and hence the Mahalanobis distance is not unique. This shows that the conditions are necessary, and the proof is thus complete. □

Note that since the univariate normal distribution is absolutely continuous and Mahalanobis' originally proposed transformation satisfy the hypothesis of Theorem 11, it follows that the present definition agrees with Mahalanobis' original definition. In particular, substitution of T for $\Phi^{-1} \circ \phi$ in Expression (1) explicitly yields

$$d(x, y) = |\Phi^{-1} \circ \phi(x) - \Phi^{-1} \circ \phi(y)|,$$

and since ϕ in this univariate case reduces to the distribution function the agreement with Mahalanobis' original definition is obvious.

Theorem 12 (Necessary Mahalanobis uniqueness conditions)**.** *Suppose the distribution $\mathcal{F}$ has a density function, then the Mahalanobis distance is unique only if transformations*

$T : \mathcal{F} \mapsto \mathcal{N}$ are injective with probability one and continuous. The conditions are sufficient if $\mathcal{F}$ is univariate.

Proof. Let p be the dimension of $\mathcal{N}$, and define $G(0) = 0$ and $G = (h, \mathrm{id}) : \mathbb{R} \times \mathbb{RP}^{p-1} \to \mathbb{R} \times \mathbb{RP}^{p-1}$ on $\mathbb{R}^p \setminus \{0\}$, where $\mathbb{RP}^{p-1}$ is the real projective space and id the identity map. Then G maps $\mathcal{N} \mapsto \mathcal{N}$ if and only if $h : \mathbb{R} \to \mathbb{R}$ maps a univariate standard normal random variable to a univariate standard normal random variable.

For any $T : \mathcal{F} \to \mathcal{N}$, define $\tilde{T} = G \circ T : \mathcal{F} \mapsto \mathcal{N}$ and denote $d(x, y) = ||T(x) - T(y)||$ and $\tilde{d}(x, y) = ||\tilde{T}(x) - \tilde{T}(y)||$. If $T(x) = (r_x, v_x) \in \mathbb{R} \times \mathbb{RP}^{p-1}$, then

$$\begin{aligned}|\tilde{d}(x, y) - d(x, y)| &= |\, ||(r_x, v_x) - (r_y, v_y)|| - ||(h(r_x), v_x) - (h(r_y), v_y)||\, | \\ &= |\, |r_x - r_y| - |h(r_x) - h(r_y)|\, |\, ||(1, v_x - v_y)|| \geq |\, |r_x - r_y| - |h(r_x) - h(r_y)|\, |\,,\end{aligned}$$

since $\alpha v_x = v_x$ for all non-zero scalars α. Note that the choice of binary operation and norm on $\mathbb{RP}^{p-1}$ is immaterial since a norm, or seminorm, is non-negative by definition. By Theorem 11, $|r_x - r_y| = |h(r_x) - h(r_y)|$ if and only if $h : \mathcal{N} \mapsto \mathcal{N}$ is injective with probability one and continuous, and thus $d = \tilde{d}$ only if Mahalanobis transformations are subject to these conditions. Hence the conditions are necessary for uniqueness of the Mahalanobis distance. That the conditions are sufficient in the univariate case is the statement of Theorem 11. □

In the multivariate case the conditions of Theorem 12 are generally not sufficient for uniqueness, in fact they are generally not even sufficient for existence of a Mahalanobis distance (cf. Theorem 5). Nevertheless, the theorem is of theoretical interest and useful in many instances. Since the conditions are sufficient for uniqueness in the univariate case, they are the strongest conditions that are necessary for uniqueness given an arbitrary distribution with density function.

In the general multivariate case agreement between Definition 1 with transformation $\Phi^{-1} \circ \phi$ and Mahalanobis' original definition is shown first. The agreement theorem uses the following lemma.

Lemma 13. *Suppose $\mathcal{F}$ is a normal distribution, then each component of the conditional distribution transformation satisfies an expression of form*

$$\phi_k(x) = \Phi(a'x + b),$$

where $a \in \mathbb{R}^k$ and $b \in \mathbb{R}$ are some constants, and Φ is the univariate standard normal distribution function.

Theorem 14. *Suppose $\mathcal{F}$ is the normal distribution with mean μ and variance Σ and $T(x) = \Sigma^{-1/2}(x - \mu)$, then the composition $T \circ \phi^{-1} \circ \Phi$ is an isometry.*

Proof. By construction the composition preserves Gaussian measure and by Lemma 13 it follows that it is linear. Thus the composition is orthogonal, and hence by Lemma 8 an isometry. □

Corollary 15. *Suppose $\mathcal{F}$ is a normal distribution, then the distance*

$$d(x, y) = ||\Phi^{-1} \circ \phi(x) - \Phi^{-1} \circ \phi(y)||$$

agrees with the conventional definition of the Mahalanobis distance.

From this point onward, the simple path is to establish, and settle with, that Definition 1 with transformation $\Phi^{-1} \circ \phi$ is a generalization of Mahalanobis' original definition, as follows by Corollary 15. The more satisfactory path, though, in the sense that it is more true to Mahalanobis' Galilean transformation reasoning, is a multivariate result corresponding to Theorem 11. The following theorem by Linnik & Eidlin (1968) implies that the composition $T \circ \phi^{-1} \circ \Phi$ is an isometry.

Theorem 16 (Linnik & Eidlin)**.** *There exists no non-linear transformation G of a standard normal random vector into a standard normal random scalar that is complex analytic (real on $\mathbb{R}^p$) and satisfies the growth condition*

$$\log\left(\max_{z \in D_r} |G(z)|\right) = O((\log r)^2),$$

where $D_r \subset \mathbb{C}^p$ is the closed ball with the zero vector as centerpoint and radius r.

Thus if the distribution $\mathcal{F}$ has a density function which is complex analytic, it follows that the conditional distribution transformation is complex analytic and therefore also its inverse by the inverse mapping theorem. Thus, by restricting distributions to those which have complex analytic density functions, and transformations T to complex analytic ones, the composition $T \circ \phi^{-1} \circ \Phi$ is complex analytic and preserves Gaussian measure. Note in particular that normal distributions have complex analytic density functions. However, in which situations the growth condition of Theorem 16 is violated is something which remains to be investigated.

4. Properties and examples of applications

Both Gauss (1809) and Pearson (1900) consider independent normally distributed random variables and propose sums of squares for the purpose of their respective methods. The following property shows how the Mahalanobis distance decomposes into a sum of squares for a joint distribution of independent random variables. The notation $d_{\mathcal{F}}$ means Mahalanobis distance under distribution $\mathcal{F}$.

Theorem 17 (The Pythagorean property)**.** *Suppose $X_1 \sim \mathcal{F}$, $X_2 \sim \mathcal{G}$ are statistically independent, and $(X_1, X_2) \sim \mathcal{H}$, then*

$$d_{\mathcal{H}}(x,y)^2 = d_{\mathcal{F}}(x_1,y_1)^2 + d_{\mathcal{G}}(x_2,y_2)^2,$$

where $x = (x_1, x_2)$, $y = (y_1, y_2)$.

Proof. If $T_1 : \mathcal{F} \mapsto \mathcal{N}$ and $T_2 : \mathcal{G} \mapsto \mathcal{N}$, then $T = (T_1, T_2) : \mathcal{H} \mapsto \mathcal{N}$ by statistical independence. Hence $T(x) = (T_1(x_1), T_2(x_2))$, and by the Pythagorean property of the Euclidean norm it then follows that

$$||T(x) - T(y)||^2 = ||T_1(x_1) - T_1(y_1)||^2 + ||T_2(x_2) - T_2(y_2)||^2,$$

which proves the theorem. □

Consequently if, for example, a sample consists of n independent observations, the squared Mahalanobis distance between the sample point and some other point decomposes into a sum of n squared Mahalanobis distances. Hence the Pythagorean property simplifies this common statistical independence situation considerably.

If, furthermore, the observations are univariate and normally distributed, then it is easily verified that the squared Mahalanobis distance between the sample point and the mean point equals Pearson's chi-square statistic. The result is not a coincidence; Pearson's statistic was defined as the squared chi-distance between the sample point and the mean point (Pearson, 1900), and the chi-distance is, of course, a now obsolete special case of the Mahalanobis distance. Conversely, Pearson's hypothesis test readily generalizes beyond normal distributions using the Mahalanobis distance as a vehicle, see, e.g., Ekström (2011a).

Another example of an application are loss functions for the purpose of model fitting. If observations are independent and normally distributed with mean zero, then the loss function proposed by Gauss (1809, §179), the now called weighted least squares loss function, is the Mahalanobis distance between the residual sample point and the zero point. If, moreover, the observations have equal variance then the Mahalanobis distance between the residual sample point and the zero point reduces to (a constant times) the least squares loss function. Hence Gauss' method readily extends to generally distributed, possibly interdependent, observations using the Mahalanobis distance as a vehicle, see, e.g., Ekström (2011b).

Mahalanobis balls under a distribution $\mathcal{F}$ are sets $B_r^{\mathcal{F}}(x) = \{y : d_{\mathcal{F}}(x,y) < r\}$, r being the radius and x the center point. When the distribution is clear from the context, the superindex $\mathcal{F}$ is often omitted. If T is the transformation used for the Mahalanobis distance, then the Mahalanobis ball can be expressed as a preimage of the Euclidean ball. The following theorem holds.

Theorem 18. *Suppose T is the transformation used for the Mahalanobis distance and let $B_r(x)$ denote the Mahalanobis ball and $E_r(x)$ the Euclidean ball, then*

$$B_r(x) = T^{-1}(E_r(T(x))).$$

Proof. Simply notice,

$$B_r(x) = \{y : ||T(y) - T(x)|| < r\} = \{y : T(y) \in E_r(T(x))\} = T^{-1}(E_r(T(x))),$$

which shows the statement. □

Since the Mahalanobis distance equals the Euclidean distance under the standard normal distribution, $\mathcal{N}$, Theorem 18 can be written $B_r^{\mathcal{F}}(x) = T^{-1}(B_r^{\mathcal{N}}(T(x)))$. This fact is a special case of the homogeneity property of Mahalanobis balls, that balls are preserved under suitable transformations of random variables.

Theorem 19 (The homogeneity property)**.** *Presuming distributions and transformations are such that Mahalanobis distances exist and are unique, suppose $T : \mathcal{G} \mapsto \mathcal{F}$. Then,*

$$B_r^{\mathcal{G}}(x) = T^{-1}(B_r^{\mathcal{F}}(T(x))).$$

Proof. Let $\tilde{T} : \mathcal{F} \mapsto \mathcal{N}$ be the transformation used for $d_{\mathcal{F}}$, then $\tilde{T} \circ T : \mathcal{G} \mapsto \mathcal{N}$ is a transformation for $d_{\mathcal{G}}$. The theorem then follows by Theorem 18 and uniqueness. □

In Pearson (1900), Mahalanobis balls are instrumental in the definition of p-values and, consequently, acceptance regions. Determination of acceptance regions is an example of an application where the homogeneity property of Mahalanobis balls comes in handy. For example, if T is a linear and injective transformation and $B_r(0)$ is an acceptance region for the random variable U, then $T(B_r(0))$ is an acceptance region for the random variable $T(U)$.

5. Concluding remarks

Many statistical methods use the Mahalanobis distance as a vehicle. Notable examples are the methods of Gauss (1809) and Pearson (1900), i.e. the method of least squares and the chi-square hypothesis test. As a consequence, extending the Mahalanobis distance beyond normal distributions yields an extraordinarily high ratio of output to input; all methods that use Mahalanobis' distance are immediately generalized beyond the set of normal distributions. For an overview of methods that use the Mahalanobis distance as a vehicle see, e.g., Mardia et al. (1979).

The conceptual beauty of Mahalanobis' Galilean transformation reasoning is immense. While it at first may seem like an impenetrable problem comparing values of random variables of different distributions, the difficulties are resolved entirely by simply mapping them into a frame of reference, which ensures that apples are compared to apples. Mahalanobis' idea is indeed a hitherto underappreciated egg of Columbus.

Acknowledgements

This work was supported by the Jan Wallander and Tom Hedelius Research Foundation, project P2008-0102:1, and the Swedish Research Council, project 435-210-565.

References

Billingsley, P. (1986). *Probability and Measure, 2nd ed.* New York: John Wiley & Sons.

Ekström, J. (2011a). On Pearson-verification and the chi-square test. UCLA Statistics Preprint.

Ekström, J. (2011b). On the determination of most probable subsets. UCLA Statistics Preprint.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*. Hamburg: F. Perthes und I. H. Besser. English translation by C. H. Davis, 1858.

Linnik, Y. V., & Eidlin, V. L. (1968). Remark on analytic transformations of normal vectors. *Theory Probab. Appl.*, *13*, 707–710.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Inst. Sci., India*, *2*, 49–55.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

Pearson, K. (1900). On the criterion that a system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag. Ser. 5*, *50*, 157–175.

Rudin, W. (1987). *Real and Complex Analysis, third ed.* Singapore: McGraw-Hill.

Appendix

Proof of Lemma 1. Let $\mathrm{Int}(f > 0)$ be denoted S. Since $F_1(x)$ is strictly positive on $\pi_1(S)$ and $\int_{\mathbb{R}} F_1(t)dt = 1 < \infty$, ϕ_1 is strictly increasing and hence injective on $\pi_1(S)$. The lemma is shown by induction, with induction hypothesis: if $(\phi_1, \ldots, \phi_k)$ is injective almost everywhere on $\pi_k(S)$ then $(\phi_1, \ldots, \phi_k, \phi_{k+1})$ is injective almost everywhere on $\pi_{k+1}(S)$.

Assume that $(\phi_1, \ldots, \phi_k)$ is injective almost everywhere on $\pi_k(S)$, and let $x \in \pi_k(S)$ and $(x, x_{k+1}) \in \pi_{k+1}(S)$. For a given $x \in \pi_k(S)$, $F_{k+1}(x, x_{k+1})$ is strictly positive on $(\pi_{k+1}(S))_x = \{x_{k+1} \in \mathbb{R} : (x, x_{k+1}) \in \pi_{k+1}(S)\}$. Because $\pi_k(S)$ is open, $F_k(x) > 0$, and by the Fubini theorem $F_k(x) < \infty$ at almost every x. Thus $h(x_{k+1}) = \int_{-\infty}^{x_{k+1}} F_{k+1}(x, t)dt$ is injective on $(\pi_{k+1}(S))_x$ given almost every x. At the same x's ϕ_{k+1} equals a constant times h and hence ϕ_{k+1} is injective on $(\pi_{k+1}(S))_x$ given almost every x. Consequently, $(\phi_1, \ldots, \phi_k, \phi_{k+1})$ is injective at almost every x.

The induction argument proves the lemma. □

Proof of Lemma 2. By the proof of Lemma 1, ϕ is injective on $\mathrm{Int}(f > 0) = S$ except if $F_k(x) = \infty$, some $k = 1, \ldots, p-1$. Let $N \subset S$ be the set of such points, and note that $\lambda(N) = 0$ by Lemma 1. Note also that $\phi(S \setminus N) \subset \mathrm{Int}(\mathbb{I}^p)$, while by definition $\phi(S \cap N) \subset \partial\mathbb{I}^p$, and thus $\phi|_{S\setminus N} : S \setminus N \to \phi(S \setminus N)$ is a bijection. By construction, if ϕ is bijective on a set A, then $(\pi_k\phi)$ is bijective on $\pi_k(A)$.

Since $(f > 0)$ by assumption is a continuity set, it holds by the Fubini theorem at almost every $x \in \mathbb{R}^k$ that $\int_{\mathbb{R}^{p-k}} f_x(t)d\lambda(t) = \int_{\mathbb{R}^{p-k}} \mathbb{1}_{S_x}(t)f_x(t)d\lambda(t)$. Denoting $\phi_{k+1}(x, \mathbb{R}) = \{\int_{-\infty}^{y} F_{k+1}(x,t)dt/F_k(x) : y \in \mathbb{R}\}$, it holds for almost every $x \in \pi_k(S\setminus N)$ that $\lambda(\phi_{k+1}(x, S_x)) = \lambda(\phi_{k+1}(x, \mathbb{R})) = \lambda(\mathbb{I}) = 1$. Thus, using $z \in \pi_k(\phi(S \setminus N))$ and $(\pi_k\phi)^{-1}(z) = x$, it holds

$$\lambda(\pi_{k+1}(\phi(S \setminus N))) = \int_{\pi_k(\phi(S\setminus N))} \lambda(\phi_{k+1}(x, S_x))d\lambda(z) = \lambda(\pi_k(\phi(S \setminus N))).$$

Since $\lambda(\pi_1(\phi(S\setminus N))) = \lambda(\phi_1(\pi_1(S\setminus N))) = \lambda(\mathbb{I}) = 1$, induction yields that $\lambda(\phi(S\setminus N)) = 1$. Since, by definition of ϕ, $\lambda(\phi(\mathbb{R}^p)) \leq \lambda(\mathbb{I}^p) = 1$, the lemma follows. □

Proof of Lemma 3. Assume $\mathrm{Prob}(A) = 0$, let $S = \mathrm{Int}(f > 0)$ and $N \subset S$ be the set where ϕ is not injective. By Lemma 1 $\lambda(N) = 0$ and by Lemma 2 $\lambda(\phi((S \setminus N)^c)) = 0$. Since $(f > 0)$ by assumption is a continuity set, it follows that $\lambda(B) = 0$ where $B = A \cap (S \setminus N)$. By the Fubini theorem, $\lambda(B_x) = 0$ for almost every $x \in \mathbb{R}^{p-1}$. Since $\phi_p(x, t)$ is absolutely continuous in t for $x \in \pi_{p-1}(S \setminus N)$, $\lambda(\phi_p(x, B_x)) = 0$ for almost every $x \in \mathbb{R}^{p-1}$ by the Luzin N property. Letting $z \in \pi_{p-1}(\phi(S \setminus N))$ and $(\pi_{p-1}\phi)^{-1}(z) = x$,

$$\lambda(\phi(B)) = \int_{\pi_{p-1}(\phi(S\setminus N))} \lambda(\phi_p(x, B_x))d\lambda(z) = 0.$$

Since $\phi(A) \subset \phi(B) \cup \phi((S \setminus N)^c)$, by Lemma 2 the conclusion $\lambda(\phi(A)) = 0$ follows. □

The proof of Lemma 4 uses the following lemma.

Lemma 20. *Let $\mathbb{S}$ be a metric space, $\mathbb{T}$ a topological space, and let A be a subset of $\mathbb{S} \times \mathbb{T}$ such that $\pi_2(A) \subset \mathbb{T}$ is relatively compact. Let $x \in \pi_1(A) \subset \mathbb{S}$ and let μ be a regular measure on $\mathbb{T}$. If $\mu((\partial A)_x) = 0$, then $\lim_{y\to x} \mu(A_x \Delta A_y) = 0$ for each sequence $\{y_n\}$ which is included in $\pi_1(A)$ and converges to x.*

Proof. Let $\{y_n\}_{n=1}^{\infty}$ be any sequence included in $\pi_1(A)$ that converges to x. If there is none the statement holds trivially.

If $z \in \mathrm{Int}(A)_x$, then (x, z) is an interior point of A and hence there is a neighborhood $N_{(x,z)} \subset \mathrm{Int}(A)$. Thus, $d_{\mathbb{S}}(x, y) < r$, for some $r > 0$, implies that $z \in A_y$. As a result, $z \in \mathrm{Int}(A)_x$ implies $z \in \liminf_{y\to x} A_y$ and consequently also $\mathrm{Int}(A)_x \subset \limsup_{y\to x} A_y$.

For every convergent sequence $\{(y_n, z_n)\}$, where for each n, $(y_n, z_n) \in \{y_n\} \times A_{y_n} \subset A$, the point $\lim_{n\to\infty}(y_n, z_n)$ is a limit point of A. Consequently, $\limsup_{y\to x} A_y \subset (\bar{A})_x$. Intersecting both sides with A_x^c yields $\limsup_{y\to x}(A_y \setminus A_x) \subset (\partial A)_x$.

Note that $\mu(A_x \Delta A_y) = \mu(A_x \setminus A_y) + \mu(A_y \setminus A_x)$. With respect to the first term, it holds that

$$\begin{aligned}\lim_{y\to x} \mu(A_x \setminus A_y) &\leq \lim_{n\to\infty} \mu(\cup_{m=n}^{\infty} A_x \setminus A_{y_m}) = \mu(\limsup_{y\to x} A_x \setminus A_y) \\ &= \mu(A_x \cap \limsup_{y\to x} A_y^c) = \mu(A_x \cap (\limsup_{y\to x} A_y)^c) \leq \mu((\partial A)_x).\end{aligned}$$

The limit interchange equality holds because the sequence of unions is non-increasing and relatively compact, and μ is regular. The last inequality holds because $B \subset C \implies C^c \subset B^c$. With respect to the second term,

$$\lim_{y\to x} \mu(A_y \setminus A_x) \leq \lim_{n\to\infty} \mu(\cup_{m=n}^{\infty} A_{y_m} \setminus A_x) = \mu(\limsup_{y\to x} A_y \setminus A_x) \leq \mu((\partial A)_x).$$

Thus, under the hypothesis $\mu((\partial A)_x) = 0$ it follows that $\lim_{y\to x} \mu(A_x \Delta A_y) = 0$. □

Proof of Lemma 4. The proof uses Scheffé's theorem (see Billingsley, 1986). Assume $\pi_k(S)$ is relatively compact. Let $x \in \pi_k(S)$, $\{x_n\}$ a sequence in $\pi_k(S)$ converging to x and temporarily denote the sections $f_n(y) = f(x_n, y)$ and $f_0(y) = f(x, y)$. Since the sections f_n are locally dominated in a neighborhood of x, the sections are integrable and thus densities in the sense of Scheffé's theorem. It holds that $f_n \to f_0$ everywhere except on the set $S_x \Delta S_{x_n}$, and thus by Lemma 20 it follows that $f_n \to f_0$ a.e. By Scheffé's theorem, then, both the numerator and the denominator of the component ϕ_k are continuous at x. Since it holds for all $x \in \pi_k(S)$, $k = 1, \ldots, p-1$, and the all component denominators are positive on S it follows that all components are continuous and hence ϕ is continuous. If $\pi_k(S)$ is not relatively compact, the statement is shown by taking an increasing sequence of relatively compact sets with limit $\pi_k(S)$ and noting that the statement holds for every set in the sequence and consequently also for the union.

To show $\phi \in C^1(S)$ under the supplementary condition, note first that since $f|_{S^c} = 0$ almost everywhere it also holds that $(Df)|_{S^c} = 0$ almost everywhere, and consequently $Df = \mathbb{1}_S Df$ almost everywhere. Let $\{e_1, \ldots, e_k\}$ be the standard basis in $\mathbb{R}^k$ and consider the directional derivative $D_{e_i} F_k$ at $x \in \pi_k(S)$. Since $|D_{e_i} f| \leq |Df|$ the directional derivatives $(D_{e_i} f)_x$ are also locally dominated. By Lebesgue's dominated convergence theorem, it follows that

$$(D_{e_i} F_k)(x) = \int_{\mathbb{R}^{p-k}} (D_{e_i} f)(x, t) d\lambda(t).$$

The right hand side exists because the section is dominated. To show that the right hand side is continuous at x, note first that since $(D_{e_i} f)_x$ is (locally) dominated there is for every $\varepsilon > 0$ a compact $K \subset \mathbb{R}^{p-k}$ such that the integral on the right hand side above over K^c is less than ε for each x_n. On K, $(D_{e_i} f)(x_n, t) \to (D_{e_i} f)(x, t)$ for almost every t as $x_n \to x$ by continuity and Lemma 20, and the integral difference $|(D_{e_i} F_k)(x_n) - (D_{e_i} F_k)(x)|$ then goes to zero as $x_n \to x$ by Vitali's convergence theorem. Hence it follows that

$D_{e_i}F_k$ is continuous at x. Since all partial derivatives exist and are continuous at every x, $F_k \in C^1(S)$.

Let $g_k(x,y) = \int_{-\infty}^{y} F_k(x,t)dt$. Clearly $g_k(x,y)$ is, for fixed x, a differentiable function of y and it follows that $D_{e_k}g_k = F_k$ by the fundamental theorem of calculus. Differentiation along the other standard basis vectors yields

$$(D_{e_i}g_k)(x,y) = \int_{-\infty}^{y}\int_{\mathbb{R}^{p-k-1}} (D_{e_i}f)(x,t,s)d\lambda(s)dt,$$

by Lebesgue's dominated convergence theorem. With the argument of the preceding paragraph, it is easily shown that $D_{e_i}g_k$ exists and is continuous at (x,y). Hence all partial derivatives exist and are continuous, and $g_k \in C^1(S)$.

Since $\phi_k = g_k/F_{k-1}$ it follows by the so-called quotient rule that $\phi_k \in C^1(S)$ (the denominator is positive on S). This shows that $\phi \in C^1(S)$. □

Proof of Lemma 7. Suppose the function G is not injective at x, let $z = G(x)$ and $y \in G^{-1}(\{z\})$, $y \neq x$, then $d(x,y) > 0$ but $d(G(x), G(y)) = 0$ and therefore G is not an isometry. Secondly, suppose G is discontinuous at x, then there is an $\varepsilon > 0$ such that $d(G(x), G(y)) > \varepsilon$ for some y for which $d(x,y) < \varepsilon$, hence G is not an isometry. □

Proof of Lemma 8. The support of the standard normal density function is a vector space and its Mahalanobis distance induced by a norm. Since the transformation G by assumption is measure preserving, it must be surjective. By Lemma 7 it is also injective. By the Mazur-Ulam theorem, then, G is affine, and an affine transformation of a standard normal random variable is standard normal if and only if it is orthogonal. □

Proof of Lemma 9. A monotonic function has only jump discontinuities, however had G a jump discontinuity it were not measure preserving and hence G is continuous. A monotonic function is not injective on sets only if it is constant, however since G is measure preserving it cannot be constant on any set with positive measure, and consequently it is injective almost everywhere.

A monotonic function is, furthermore, differentiable at almost every point. If N is the set where G is not injective or differentiable, then $\lambda(G(N)) = 0$ since G is measure preserving. The change-of-variables theorem then yields that $|dG/dx| = 1$ at almost every point. Continuity, monotonicity and the measure preserving property yield that G is absolutely continuous, and integration then yields the result. □

Proof of Lemma 10. Under the assumptions, $G : \mathcal{U} \mapsto \mathcal{U}$ is monotonic and the result follows by Lemma 9. □

Proof of Lemma 13. By the change-of-variables theorem, it holds that

$$\int_{(-\infty,x_k)} F_k(x_1,\ldots,x_{k-1},t)dt = \int_{h^{-1}(-\infty,x_k)} (F_k \circ h)|J_h|dt,$$

under some conditions on the transformation h.

From the property of normally distributed random variables, that conditional distributions are themselves normal, there is a linear function h such that the integrand factorizes into

$$(F_k \circ h)|J_h| = F_{k-1}(x_1, \ldots, x_{k-1})\tilde{f}(t),$$

where $\tilde{f}$ is the density function of a univariate normally distributed random variable with some mean and variance (see e.g. Mardia et al., 1979, Theorem 3.2.3). Substitution into the expression for ϕ_k and cancelation yields a univariate normal distribution function, which equals $\Phi(a'x + b)$ for some constants $a \in \mathbb{R}^k$ and $b \in \mathbb{R}$. □

UCLA Department of Statistics, 8125 Mathematical Sciences Building, Box 951554, Los Angeles CA, 90095-1554

E-mail address: joakim.ekstrom@stat.ucla.edu