To the Graduate Council:

I am submitting herewith a thesis written by Tamela Lynne Willett entitled "The Mahalanobis Distance and Probability Metrics." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Mathematics.

Jan Rosinski

Jan Rosinski, Major Professor

We have read this thesis
and recommend its acceptance:

Don Hinton

Accepted for the Council:

Lawrence Dewminkel

Vice Provost
and Dean of The Graduate School

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Master's degree at The University of Tennessee, Knoxville, I agree that the Library shall make it available to borrowers under rules of the Library. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of the source is made.

Permission for extensive quotation from or reproduction of this thesis may be granted by my major professor, or in his absence, by the Head of Interlibrary Services when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Signature Pamela L. Willett

Date February 18, 1988

THE MAHALANOBIS DISTANCE AND PROBABILITY METRICS

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Tamela Lynne Willett

March 1988

ACKNOWLEDGEMENTS

The author wishes to express her sincere appreciation to Dr. Jan Rosinski for his guidance during the preparation of this thesis and also to Cindi Blair for her typing of the thesis.

She also wishes to thank her friends for their patience and understanding and her family for their love and support.

ABSTRACT

The purpose of this thesis is to examine two different concepts of distance between random variables which are the Mahalanobis distance and probability metrics. The thesis is divided into two parts. The first part deals with the Mahalanobis distance and its properties, distribution and moments. The first part concludes with a useful theorem which establishes the equivalence of the moments of different orders. The second part deals with probability metrics and reviews several well known probability metrics. This part concludes with a search for comparisons between the Mahalanobis distance and the Kantorovich-Rubinstein metric.

TABLE OF CONTENTS

CHAPTER PAGE

INTRODUCTION 1

1. THE MAHALANOBIS DISTANCE 3

1.1. The Concept of Mahalanobis Distance 3

1.2. Properties of the Mahalanobis Distance. 5

1.3. Moments of the Interclass Mahalanobis Distance. . . .10

1.4. Equivalence of Moments. 16

2. PROBABILITY METRICS. 22

2.1. Introduction. 22

2.2. Examples of Metrics 22

2.3. Moments of Mahalanobis distance and Probability Metrics 37

BIBLIOGRAPHY 42

VITA 44

INTRODUCTION

The Mahalanobis distance is used to measure the distance between two normal random vectors in Euclidean space. This notion is commonly utilized in the field of pattern recognition, as is described in the first section of Part 1. On the other hand, in the field of theoretical and applied probability, there exists a quite well developed theory of probability metrics. The purpose of this thesis is to examine these two different concepts of distances and to look for possible connections.

The thesis is divided into two parts. The first deals with the Mahalanobis distance. In the first section, the Mahalanobis distance is introduced and applications to pattern recognition are mentioned. The second section contains properties of the Mahalanobis distance including its distribution. In the third section, the moments of the interclass Mahalanobis distance are computed. The final section of the first part states and proves a theorem which shows the equivalence of the moments of the interclass Mahalanobis distance. This fact, which was not available in the literature on this subject, seems to be useful and interesting. It may also be noted that it cannot be directly derived from Gonzalez's and Wagner's paper [6] .

Part 2 reviews probability metrics. The first section is an introduction to probability metrics and states the definition of a probability metric. The second section lists, as compiled by Zolotarev [7], several well known probability metrics. The proofs of the metrics are also included in this section. Finally the last section looks

for comparisions between the Mahalanobis distance and the Kantorovich-Rubinstein metric.

1. THE MAHALANOBIS DISTANCE

1.1. The Concept of Mahalanobis Distance

The Mahalanobis distance, the main emphasis of this paper, is a concept often used in pattern recognition. Pattern recognition is a topic used in many fields including medicine, psychology, biology, engineering, electronics, nuclear power, and mechanics [1], [2], [3], [4], [5] . Basically pattern recognition is taking a sample of unknown status and recognizing the sample as a member of the class to which it belongs [2] . For example, a person with no medical problems has a "normal" blood pressure, temperature, heart rate, and pulse. Another person's medical factors can be compared to the "normal" factors in order to determine whether or not there may be medical problems [4] . The "normal" factors are established through examining several people without medical problems and then coming up with one set "normal" pattern. Other cases are then compared to this "normal" pattern, and a decision is made as to whether or not the case matches the "normal" pattern [2] .

One concept needed with pattern recognition is a way in order to determine whether the sample belongs to the established pattern. One such method used is the concept called the Mahalanobis distance [1] .

Consider two d dimensional Gaussian vectors $\vec{x}$ and $\vec{y}$ with corresponding mean vectors $\vec{m}_x$ and $\vec{m}_y$ and corresponding covariance matrices C_x and C_y . The interclass Mahalanobis distance between any member of $\vec{x}$ and $\vec{m}_x$ is then defined to be:

$$R(\vec{x}, \vec{m}_x) = (\vec{x} - \vec{m}_x)^T C_x^{-1} (\vec{x} - \vec{m}_x)$$

and similarly the interclass Mahalanobis distance between any member of $\vec{y}$ and $\vec{m}_y$ is defined to be:

$$R(\vec{y}, \vec{m}_y) = (\vec{y} - \vec{m}_y)^T C_y^{-1} (\vec{y} - \vec{m}_y) .$$

We will be concerned with the interclass Mahalanobis distance between members of $\vec{x}$ and the mean $\vec{m}_y$, which is defined by:

$$R(\vec{x}, \vec{m}_y) = (\vec{x} - \vec{m}_y)^T C_y^{-1} (\vec{x} - \vec{m}_y)$$

and similarly the interclass Mahalanobis distance between members of $\vec{y}$ and the mean $\vec{m}_x$ is defined by:

$$R(\vec{y}, \vec{m}_x) = (\vec{y} - \vec{m}_x)^T C_x^{-1} (\vec{y} - \vec{m}_x) \qquad [6] .$$

The Mahalanobis distance is a way of classifying a pattern for the following reason. We know that $\{\vec{x}: R(\vec{x}, \vec{m}_y) \leq r^2\}$ is an elipsoid centered at $\vec{m}_y$ and whose shape and orientation is determined by C_y. If we assume that $\vec{x} \stackrel{d}{=} \vec{y}$ (where $\stackrel{d}{=}$ means equal in distribution), then it will be shown in this paper that $R(\vec{x}, \vec{m}_y)$ has a χ^2 distribution. Then, using Chi-squared tables, we can calculate r so that $P\{R(\vec{x}, \vec{m}_y) \leq r^2\} = 1 - \alpha$. This gives a level α test for testing the null hypothesis that a vector has the same pattern as the "normal" pattern against the alternative hypothesis that they do not share the same pattern. If the sample vector $\vec{x}$ has a value of $R(\vec{x}, \vec{m}_y) \leq r^2$, that is, the vector lies inside the elipsoid, then we accept the null hypothesis. If the value of $R(\vec{x}, \vec{m}_y) > r^2$ then we reject the null hypothesis. The power of the test is the probability of rejecting the null hypothesis when the alternative is true. However, when the distribution of $\vec{x}$ is different from the distribution of $\vec{y}$, the distribution

of $R(\vec{x}, \vec{m}_y)$ is not any of the standard distributions (except of the case $C_x = C_y$, see Corollary 1) and $P\{R(\vec{x}, \vec{m}_y) > r^2\}$ cannot be found in statistical tables. Several probability inequalities involving moments exist to help estimate the above probability. It is therefore useful to have the moments of the interclass Mahalanobis distance. These moments will be calculated in the following sections.

1.2. Properties of the Mahalanobis Distance

In order to find the moments of the interclass Mahalanobis distance a transformation of the normal vectors will be needed. The following lemma is a well known fact, but for completeness it will be stated and proved.

Lemma 1. If $\vec{x}$ is a d-dimensional Gaussian vector with mean vector $\vec{m}_x$ and covariance matrix C_x, then for any real transformation matrix A,

(1) $\vec{r} = A\vec{x}$ is a d-dimensional Gaussian vector;

(2) $\vec{m}_r = A\vec{m}_x$;

(3) $C_r = AC_x A^T$.

Proof. (1) Since $\vec{x}$ is a d-dimensional Gaussian vector, $\exists$ a $d \times d$ matrix C, constants $\mu_1, \ldots, \mu_d$ and independent random variables $z_1, \ldots, z_d$ $(z_i \sim N(0, 1))$ such that:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{d1} \\ \vdots & & \vdots \\ c_{d1} & \cdots & c_{dd} \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_d \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix}.$$

So $\vec{r} = A\vec{x}$

$$= \begin{pmatrix} a_{11} & \cdots & a_{1d} \\ & & \\ a_{d1} & \cdots & a_{dd} \end{pmatrix} \begin{pmatrix} c_{11} & \cdots & c_{1d} \\ & & \\ c_{d1} & \cdots & c_{dd} \end{pmatrix} \begin{pmatrix} z_1 \\ \\ z_d \end{pmatrix} + \begin{pmatrix} a_{11} & \cdots & a_{1d} \\ & & \\ a_{d1} & \cdots & a_{dd} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \\ \mu_d \end{pmatrix}$$

and hence $\vec{r} = A\vec{x}$ is a d-dimensional Gaussian vector.

(2) $m_r = E(\vec{r}) = E(A\vec{x})$

$$= E\left[\begin{pmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & & \vdots \\ a_{d1} & \cdots & a_{dd} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \right] = E \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1d}x_d \\ \vdots \\ a_{d1}x_1 + a_{d2}x_2 + \cdots + a_{dd}x_d \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}\mu_1 + a_{12}\mu_2 + \cdots + a_{1d}\mu_d \\ \vdots \\ a_{d1}\mu_1 + a_{d2}\mu_2 + \cdots + a_{dd}\mu_d \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & & \vdots \\ a_{d1} & \cdots & a_{dd} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} = A\vec{m}_x$$

(3) $C_r = [\mathrm{Cov}(r_i, r_j)]_{i,j=1,\ldots,d}$

$= [\mathrm{Cov}(\vec{e}_i^T \vec{r}\ , \vec{e}_j^T \vec{r})]_{i,j=1,\ldots,d}$ where $\vec{e}_j = \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} \rightarrow j^{th}$ spot

$= [\mathrm{Cov}(\vec{e}_i^T A\vec{x}\ , \vec{e}_j^T A\vec{x})]_{i,j=1,\ldots,d}$

$= [\mathrm{Cov}((A^T\vec{e}_i)^T\vec{x}\ , (A^T\vec{e}_j)^T\vec{x})]_{i,j=1,\ldots,d}$.

Now $\mathrm{Cov}(a^T\vec{x}, b^T\vec{x}) = \mathrm{Cov}(\Sigma a_i x_i\ , \Sigma b_j x_j)$

$$= \sum_{i,j} a_i b_j \mathrm{Cov}(x_i, x_j) = a^T C_x b$$

for any two 1 by d matrices $a^T = (a_1 \ldots a_d)$ and $b^T = (b_1 \ldots b_d)$.

Therefore $[\mathrm{Cov}((A^T \vec{e}_i)^T \vec{x}\ ,\ (A^T \vec{e}_j)^T \vec{x})]_{i,j=1,\ldots,d}$

$$= [(A^T \vec{e}_i)^T C_x A^T \vec{e}_j]_{i,j=1,\ldots,d}$$

$$= [\vec{e}_i^T (AC_x A^T) \vec{e}_j^T]_{i,j=1,\ldots,d}$$

$$= AC_x A^T \quad .$$

Lemma 2. $R(A\vec{x}, A\vec{m}_y) = R(\vec{x}, \vec{m}_y)$ for any real invertible transformation matrix A .

Proof. $R(A\vec{x}, A\vec{m}_y) = (A\vec{x} - A\vec{m}_y)^T (AC_y A^T)^{-1} (A\vec{x} - A\vec{m}_y)$
since from Lemma 1 $\mathrm{Cov}(A\vec{y}) = AC_y A^T$

$$= [A(\vec{x} - \vec{m}_y)]^T (C_y A^T)^{-1} A^{-1}[A(\vec{x} - \vec{m}_y)]$$

$$= (\vec{x} - \vec{m}_y)^T A^T (A^T)^{-1} C_y^{-1} A^{-1} A(\vec{x} - \vec{m}_y)$$

$$= (\vec{x} - \vec{m}_y)^T C_y^{-1} (\vec{x} - \vec{m}_y)$$

$$= R(\vec{x}, \vec{m}_y) \quad .$$

Note that since C_x and C_y are positive definite and symmetric, A may be chosen so that $C_r = AC_x A^T = I$ and $C_s = AC_y A^T = D$ where D is a diagonal matrix with elements $\gamma(i)$, the eigenvalues of $C_x^{-1} C_y$, on the diagonal. This fact can be found in [5, p. 34] .

Theorem 1. Every interclass Mahalanobis distance can be written as a sum of squares of d independent normal random variables. Specifically:

(1) $R(\vec{x}, \vec{m}_y) = \sum_{i=1}^{d} (r_i - m_{si})^2 \gamma^{-1}(i)$ where

$r_i - m_{si} \sim N(m_{ri} - m_{si}, 1)$;

(2) $R(\vec{y}, \vec{m}_x) = \sum_{i=1}^{d} (s_i - m_{ri})^2$ where

$(s_i - m_{ri})/\sqrt{\gamma(i)} \sim N(m_{si} - m_{ri}/\sqrt{\gamma(i)}, 1)$.

Proof. (1) From Lemma 2,

$$R(\vec{x}, \vec{m}_y) = R(\vec{r}, \vec{m}_s)$$

$$= (\vec{r} - \vec{m}_s)^T D^{-1}(\vec{r} - \vec{m}_s)$$

$$= \sum_{i=1}^{d} (r_i - m_{si})^2 \gamma^{-1}(i) .$$

Now since $\vec{r}$ is a Gaussian random vector with mean vector $\vec{m}_r$ and $C_r = I$ we get $r_i - m_{ri}$ are independent and $(r_i - m_{ri}) \sim N(0, 1)$

$$\Rightarrow (r_i - m_{ri} + m_{ri} - m_{si}) \sim N(m_{ri} - m_{si}, 1)$$

$$\Rightarrow (r_i - m_{si}) \sim N(m_{ri} - m_{si}, 1) .$$

(2) From Lemma 2

$$R(\vec{y}, \vec{m}_x) = R(\vec{s}, \vec{m}_r)$$

$$= (\vec{s} - \vec{m}_r)^T I^{-1}(\vec{s} - \vec{m}_r)$$

$$= \sum_{i=1}^{d} (s_i - m_{ri})^2 .$$

Now since $\vec{s}$ is a Gaussian random vector with mean vector $\vec{m}_s$ and $C_s = D$ we get $\frac{s_i - m_{si}}{\sqrt{\gamma(i)}}$ are independent and $(s_i - m_{si})/\sqrt{\gamma(i)} \sim N(0, 1)$

$$\Rightarrow (s_i - m_{si} + m_{si} - m_{ri})/\sqrt{\gamma(i)} \sim N(m_{si} - m_{ri}/\sqrt{\gamma(i)}, 1)$$

$$\Rightarrow (s_i - m_{ri})/\sqrt{\gamma(i)} \sim N(m_{si} - m_{ri}/\sqrt{\gamma(i)}, 1) .$$

Notice that since $r_i - m_{si} \sim N(m_{ri} - m_{si}, 1)$ then $r_i - m_{si} - m_{ri} + m_{si} = U_i$ where U_i are independent and $U_i \sim N(0, 1)$ and hence $r_i - m_{si} = U_i + m_{ri} - m_{si}$ where U_i are independent and $U_i \sim N(0, 1)$. Similarly $s_i - m_{ri}/\sqrt{\gamma(i)} = V_i + (m_{si} - m_{ri})/\sqrt{\gamma(i)}$ where $V_i \sim N(0, 1)$.

Corollary 1. In the following special cases, the Mahalanobis distance has the following distribution:

(1) $R(\vec{x}, \vec{m}_x)$ has an χ^2 distribution with d degrees of freedom;

(2) $R(\vec{x}, \vec{m}_y)$ has a non-central χ^2 distribution if $C_x = C_y$.

Proof. By Lemma 2

$$R(\vec{x}, \vec{m}_x) = R(\vec{r}, \vec{m}_r)$$

$$= (\vec{r} - \vec{m}_r)^T I^{-1} (\vec{r} - \vec{m}_r)$$

$$= \sum_{i=1}^{d} (r_i - m_{ri})^2 .$$

Now $r_i - m_{ri} \sim N(0, 1)$ so $\sum_{i=1}^{d} (r_i - m_{ri})^2$ is χ^2 with d degree of freedom.

(2) By Theorem 1,

$$R(\vec{x}, \vec{m}_y) = \sum_{i=1}^{d} (r_i - m_{si})^2 \gamma^{-1}(i) .$$

By hypothesis, $C_x = C_y$ and thus $C_r = C_s = I$ and hence

$$R(\vec{x}, \vec{m}_y) = \sum_{i=1}^{d} (r_i - m_{si})^2 .$$

Also by Theorem 1, we know $r_i - m_{si} \sim N(m_{ri} - m_{si}, 1)$ and so $R(\vec{x}, \vec{m}_y)$ has a non-central χ^2 distribution with non-central parameter $\theta^2 = \sum_{i=1}^{d} (m_{ri} - m_{si})^2$.

1.3. Moments of the Interclass Mahalanobis Distance

We see from the previous section that the distribution for the interclass Mahalanobis distance can only be found in the special case where $C_x = C_y$. This is not very often the case, so we will examine the moments in order to find out more. In this section the first moment as well as the general moments of the interclass Mahalanobis distance will be found.

A. First Moments

Theorem 2. The first moment of $R(\vec{x}, \vec{m}_y)$ is

$$E[R(\vec{x}, \vec{m}_y)] = \sum_{i=1}^{d} \gamma^{-1}(i)[1 + (m_{ri} - m_{si})^2]$$

and the first moment of $R(\vec{y}, \vec{m}_x)$ is

$$E[R(\vec{y}, \vec{m}_x)] = \sum_{i=1}^{d} [\gamma(i) + (m_{si} - m_{ri})^2] .$$

Proof. $E[R(\vec{x}, \vec{m}_y)] = E[\sum_{i=1}^{d} (r_i - m_{si})^2 \gamma^{-1}(i)]$ (from Theorem 1)

$$= \sum_{i=1}^{d} \gamma^{-1}(i) E(r_i - m_{si})^2$$

$$= \sum_{i=1}^{d} \gamma^{-1}(i) E\{[U_i + (m_{ri} - m_{si})]^2\}$$

where $U_i \sim N(0, 1)$ by Lemma 3

$$= \sum_{i=1}^{d} \gamma^{-1}(i) E[U_i^2 + 2U_i(m_{ri} - m_{si}) + (m_{ri} - m_{si})^2]$$

$$= \sum_{i=1}^{d} \gamma^{-1}(i) [1 + (m_{ri} - m_{si})^2]$$

and

$$E[R(\vec{y}, \vec{m}_x)] = E[\sum_{i=1}^{d} (s_i - m_{ri})^2] \quad \text{(from Theorem 1)}$$

$$= \sum_{i=1}^{d} E(s_i - m_{ri})^2$$

$$= \sum_{i=1}^{d} \gamma(i) E[V_i + (m_{si} - m_{ri})/\sqrt{\gamma(i)}]^2$$

where $V_i \sim N(0, 1)$ by Lemma 3

$$= \sum_{i=1}^{d} \gamma(i) E[V_i^2 + 2V_i(m_{si} - m_{ri})/\sqrt{\gamma(i)} + (m_{si} - m_{ri})^2/\gamma(i)]$$

$$= \sum_{i=1}^{d} \gamma(i) \left[1 + \frac{(m_{si} - m_{ri})^2}{\gamma(i)}\right] = \sum_{i=1}^{d} \gamma(i) + (m_{si} - m_{ri})^2 .$$

B. General Moments

Theorem 3.

$$\text{(i)} \quad E[R^n(\vec{x}, \vec{m}_y)] = \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i) \sum_{j=0}^{e_i} \frac{(2e_i)!(m_{ri} - m_{si})^{2j}}{(2j)!\, 2^{e_i-j} (e_i - j)!} \quad \text{where } 0 \le j \le e_i$$

$$\text{(ii)} \quad E[R^n(\vec{y}, \vec{m}_x)] = \sum_{e_i+\ldots+e_d=n} \frac{n!}{e_i!\ldots e_d!} \prod_{i=1}^{d} \gamma^{e_i}(i) \sum_{j=0}^{e_i} \frac{(2e_i)!}{(2j)!\, 2^{e_i-j} (e_i - j)!} \frac{(m_{si} - m_{ri})^{2j}}{[\gamma(i)]^j}$$

where $0 \le j \le e_i$.

Proof. (i) $R^n(\vec{x}, \vec{m}_y) = [\sum_{i=1}^{d} (r_i - m_{si})^2 \gamma^{-1}(i)]^n$ (by Theorem 1)

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} [(r_i - m_{si})^2/\gamma(i)]^{e_i}$$

(by multinomial theorem) .

Thus

$$E[R^n(\vec{x}, \vec{m_y})] = E[\sum_{e_i+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} [(r_i - m_{si})^2/\gamma(i)]^{e_i}]$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} E[(U_i + (m_{ri} - m_{si}))^{2e_i} \gamma^{-e_i}]$$

where $U_i \sim N(0, 1)$ by Lemma 3

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i)\, E[U_i^{2e_i} + 2e_i U_i^{2e_i-1} (m_{ri} - m_{si})$$

$$+ \frac{2e_i(2e_i - 1)}{2 \cdot 1} U_i^{2e_i-2} (m_{ri} - m_{si})^2$$

$$+ \ldots + 2e_i\, U_i (m_{ri} - m_{si})^{2e_i-1} + (m_{ri} - m_{si})^{2e_i}]$$

by the Binomial theorem

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i)$$

$$[\frac{(2e_i)!}{2^{e_i} \cdot (e_i)!} + 0 + \frac{(2e_i)(2e_i - 1)(2e_i - 2)!}{2 \cdot 2^{e_i-1} \cdot (e_i - 1)!} (m_{ri} - m_{si})^2$$

$$+ 0 + \ldots + 0 + (m_{ri} - m_{si})^{2e_i}]$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i) [\frac{(2e_i)!}{2^{e_i}(e_i)!} + \frac{(2e_i)!}{2^{e_i-1} \cdot 2(e_i - 1)}$$

$$(m_{ri} - m_{si})^2 + \frac{(2e_i)!}{4! \cdot 2^{e_i-2} \cdot (e_i - 2)!} (m_{ri} - m_{si})^4$$

$$+ \ldots + (m_{ri} - m_{si})^{2e_i}]$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i) \sum_{j=0}^{e_i} \frac{(2e_i)!}{(2_j)!2^{e_i-j}(e_i - j)!}$$

$$(m_{ri} - m_{si})^{2j} \qquad (\text{where } 0 \le j \le e_i) .$$

(ii) Similarly we find that

$$R^n(\vec{y}, \vec{m}_x) = [\sum_{i=1}^{d} (s_i - m_{ri})^2]^n$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} [(s_i - m_{ri})^2]^{e_i} .$$

Hence

$$E[R^n(\vec{y}, \vec{m}_x)] = E[\sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} [(s_i - m_{ri})^2]^{e_i}]$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} E[[\sqrt{\gamma(i)}(V_i + \frac{m_{si} - m_{ri}}{\sqrt{\gamma(i)}})]^{2e_i}] \; V_i \sim N(0, 1)$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma(i)^{e_i} E[(V_i + (m_{si} - m_{ri})/\sqrt{\gamma(i)})^{2e_i}]$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma(i)^{e_i}$$

$$E[V_i^{2e_i} - 2e_i V_i^{2e_i-1} (m_{si} - m_{ri})/\sqrt{\gamma(i)}$$

$$+ \frac{2e_i(2e_i - 1)}{2 \cdot 1} V_i^{2e_i-2} (m_{si} - m_{ri})^2/(\sqrt{\gamma(i)})^2$$

$$+ \ldots + 2e_i \, V_i (m_{si} - m_{ri})^{2e_i - 1} / \sqrt{\gamma(i)}^{2e_i - 1} + \frac{(m_{si} - m_{ri})^{2e_i}}{(\sqrt{\gamma(i)})^{2e_i}}$$

$$= \sum_{e_1 + \ldots + e_d = n} \frac{n!}{e_1! \ldots e_d!} \prod_{i=1}^{d} \gamma(i)^{e_i} \left[\frac{(2e_i)!}{2^{e_i} (e_i)!} + 0 \right.$$

$$+ \frac{(2e_i)(2e_i - 1)(2e_i - 2)!}{2 \cdot 2^{e_i - 1} (e_i - 1)!} \frac{(ms_i - m_{ri})^2}{(\sqrt{\gamma(i)})^2}$$

$$+ 0 + \ldots + 0 + \left. \frac{(m_{si} - m_{ri})^{2e_i}}{(\sqrt{\gamma(i)})^{2e_i}} \right]$$

$$= \sum_{e_1 + \ldots + e_d = n} \frac{n!}{e_1! \ldots e_d!} \prod_{i=1}^{d} \gamma(i)^{e_i} \left[\frac{(2e_i)!}{2^{e_i} (e_i)!} \right.$$

$$+ \frac{(2e_i)!}{2^{e_i - 1} \cdot 2 \cdot (e_i - 1)!} \frac{(m_{si} - m_{ri})^2}{(\sqrt{\gamma(i)})^2}$$

$$\left. + \frac{(2e_i)!}{4! \, 2^{e_i - 2} (e_i - 2)!} \frac{(m_{si} - m_{ri})^4}{(\sqrt{\gamma(i)})^4} + \ldots + \frac{(m_{si} - m_{ri})^{2e_i}}{(\sqrt{\gamma(i)})^{2e_i}} \right]$$

$$= \sum_{e_1 + \ldots + e_d = n} \frac{n!}{e_1! \ldots e_d!} \prod_{i=1}^{d} \gamma(i)^{e_i} \sum_{j=0}^{e_i} \frac{(2e_i)!}{(2j)! \, 2^{e_i - j} (e_i - j)!}$$

$$\frac{(m_{si} - m_{ri})^{2j}}{[\gamma(i)]^j} \, .$$

An alternative form of the above moments is found in the paper "Moments of the interclass Mahalanobis distance" by R.C. Gonzalez and C.G. Wagner [6] , but for the purpose of this paper it is more efficient to use the moments calculated in the above section.

1.4. Equivalence of Moments

In this section we will establish the relationship between all of the moments of the interclass Mahalanobis distance.

Theorem 4. All moments of the interclass Mahalanobis distance are comparable to the first moment in the following manner:

$$E[R(\vec{x}, \vec{m}_y)] \leq [E(R^n(\vec{x}, \vec{m}_y))]^{1/n} \leq 2(n!)^{1/n} E[R(\vec{x}, \vec{m}_y)] \quad .$$

Proof. We know by Theorem 2 that $E(R(\vec{x}, \vec{m}_y)) = \sum_{i=1}^{d} \gamma^{-1}(i)[1 + (m_{ri} - m_{si})^2]$ and hence, by the Multinomial Theorem,

$$[E(R(\vec{x}, \vec{m}_y))]^n =$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} [\gamma^{-1}(i)[1 + (m_{ri} - m_{si})^2]]^{e_i}$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i)\,[1 + (m_{ri} - m_{si})^2]^{e_i}$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i)[1 + e_i(m_{ri} - m_{si})^2$$

$$+ \frac{e_i(e_i - 1)}{1 \cdot 2} (m_{ri} - m_{si})^4 + \ldots + e_i(m_{ri} - m_{si})^{2e_i - 2}$$

$$+ (m_{ri} - m_{si})^{2e_i}]$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i) \sum_{j=0}^{e_i} \frac{(e_i)!}{(j)!(e_i - j)!} (m_{ri} - m_{si})^{2j}$$

Using Theorem 3 we get

$$E(R^n(\vec{x}, \vec{m}_y)) = \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i)$$

$$\sum_{j=0}^{e_i} \frac{(2e_i)!}{(2j)!\, 2^{e_i-j} (e_i - j)!} (m_{ri} - m_{si})^{2j}$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i)$$

$$\sum_{j=0}^{e_i} \frac{(2e_i)(2e_i - 1)(2e_i - 2) \ldots (2e_i - (2e_i - 1))}{(2j)(2j - 1)(2j - 2) \ldots (2j - (2j - 1))\, 2^{e_i-j} (e_i - j)!}$$

$$(m_{ri} - m_{si})^{2j}$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i) \sum_{j=0}^{e_i}$$

$$\frac{2^{e_i} e_i(2e_i - 1)(e_i - 1)(2e_i - 3)\ldots(e_i - (e_i - 1))(2e_i - (2e_i - 1))}{2^j j(2j - 1)(j - 1)(2j - 3)\ldots(j - (j - 1))(2j - (2j - 1))2^{e_i-j} (e_i - j)!} (m_{ri} - m_{si})^{2j}$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i) \sum_{j=0}^{e_i}$$

$$\frac{(e_i)!\ (2e_i - 1)(2e_i - 3)\ldots(2e_i - (2e_i - 1))}{(j)!(2j - 1)(2j - 3)\ldots(2j - (2j - 1))(e_i - j)!} (m_{ri} - m_{si})^{2j}$$

$$= \sum_{e_1+\ldots+e_d=n} \frac{n!}{e_1!\ldots e_d!} \prod_{i=1}^{d} \gamma^{-e_i}(i) \sum_{j=0}^{e_i}$$

$$a(e_i,j) \frac{(e_i)!}{(j)!(e_i - j)!} (m_{ri} - m_{si})^{2j} \quad .$$

We now wish to examine bounds for $a(e_i,j)$, $0 \le j \le e_i$ and $0 \le e_i \le n$. We have

$$\frac{(2e_i)!}{(2j)!2^{e_i-j}(e_i - j)!} = a(e_i,j) \frac{(e_i)!}{(j)!(e_i - j)!}$$

If $e_i = 0$, then $j = 0$ and $a(0, 0) = 1$. Consider $j = 0$ and allow e_i to vary. Then

$$\frac{(2e_i)!}{1 \cdot 2^{e_i}(e_i)!} = a(e_i, 0) \frac{(e_i)!}{0!(e_i)!}$$

and since $0 \le e_i \le n$ we have that

$$1 \le a(e_i, 0) \le (2n)!/2^n(n)! \ .$$

Now consider $e_i > 0$, $j > 0$. Then we know

$$a(e_i,j) = \frac{(2e_i - 1)(2e_i - 3)\ldots(2e_i - (2e_i - 1))}{(2j - 1)(2j - 3) \ \ldots \ (2j - (2j - 1))} .$$

We know that, since $0 < e_i < n$ and $0 < j \leq e_i$

$$1 \leq (2e_i - 1)(2e_i - 2)\ldots(2e_i - (2e_i - 1))$$

$$\leq (2n - 1)(2n - 3) \ \ldots \ 1 \leq 2^n n!$$

and

$$1 \leq (2j - 1)(2j - 3)\ldots(2j - (2j - 1))$$

$$\leq (2n - 1)(2n - 3) \ \ldots \ 1 \leq 2^n n! \ .$$

The last inequality follows from the fact that

$$(2n - 1)(2n - 3) \ \ldots \ 1 = 2^n(n - \tfrac{1}{2})(n - \tfrac{3}{2}) \ \ldots \ \tfrac{1}{2}$$

$$\leq 2^n \cdot n(n - 1) \ \ldots \ 1 = 2^n n! \ .$$

Therefore, $1/(2^n n!) \leq a(e_i,j) \leq 2^n n!$ for $e_i > 0$, $j > 0$, and

$$1 \leq a(e_i, 0) \leq (2n)!/2^n(n)! \ .$$

Now

$$\frac{(2n)!}{2^n(n)!} = \frac{2^n\, n!(2n-1)(2n-3)\ldots 1}{2^n(n)!} \leq 2^n\, n!$$

and obviously

$$1 \geq 1/(2^n\, n!)$$

so we get that

$$1/(2^n\, n!) \leq a(e_i,j) \leq 2^n\, n! \quad \text{for all} \quad 0 \leq e_i < n \quad \text{and} \quad 0 \leq j \leq e_i\ .$$

Using these bounds we obtain

$$1/(2^n\, n!)[ER(\vec{x}, \vec{m}_y)]^n \leq E(R^n(\vec{x}, \vec{m}_y)) \leq 2^n n![ER(\vec{x}, \vec{m}_y)]^n\ .$$

The left-side inequality is not sharp since, by Holder's inequality, we know that

$$[ER(\vec{x}, \vec{m}_y)]^n \leq E(R^n(\vec{x}, \vec{m}_y))$$

and hence we obtain

$$[ER(\vec{x}, \vec{m}_y)]^n \leq E(R^n(\vec{x}, \vec{m}_y)) \leq 2^n\, n![ER(\vec{x}, \vec{m}_y)]^n$$

and the result follows.

Remark. The usefulness of this theorem can be seen in two cases:

1. While the first moment is relatively simple to calculate, the difficulty of calculating the higher order moments grows rapidly with each order. Using the above theorem we are able to get bounds on the higher order moments.

2. The above theorem shows that the convergence in all moments of the interclass Mahalanobis distances are equivalent.

2. PROBABILITY METRICS

2.1. Introduction

In Probability Theory the concept of distance of probability distributions is formalized in terms of probability metrics. The purpose of this section is to examine the Mahalanobis distance to see if it is comparable to a probability metric.

All probability metrics have the same characteristics expressed in the following definition.

Definition. A functional $d(X, Y)$ is said to be a probability metric if it possesses the following three properties:

1. $P(X = Y) = 1 \Rightarrow d(X, Y) = 0$
2. $d(X, Y) = d(Y, X)$
3. $d(X, Y) \leq d(X, Z) + d(Z, Y)$.

2.2. Examples of Metrics

In this section, several known probability metrics will be shown to satisfy the definition. Several probability metrics have been studied. Among them the best known are [7]:

1. Lévy Metric:

$$d(X, Y) = \inf \{\varepsilon > 0 : F_X(x - \varepsilon) - \varepsilon \leq F_Y(x) \leq F_X(x + \varepsilon) + \varepsilon,\ x \in \mathbb{R}\}$$

where F_X and F_Y are the distribution functions of X and Y respectively.

2. Uniform Metric:

$$d(X, Y) = \sup\{|F_X(x) - F_Y(x)| : x \in \mathbb{R}\} \ ;$$

this metric is also sometimes called the Kolmogorov's Metric.

3. Distance of total variation:

$$d(X, Y) = \frac{1}{2}\int_{-\infty}^{\infty} |f_X(x) - f_Y(y)|dx$$

where f_X and f_Y are the densities of X and Y, respectively.

4. L_p-Metric:

$$d_p(X, Y) = [\int |F_X(x) - F_Y(x)|^p \, dx]^{\min(1, 1/p)} \quad p > 0 \ .$$

5. Ky Fan Metric:

$$d(\vec{X}, \vec{Y}) = \inf\{\varepsilon: P(d'(\vec{X}, \vec{Y}) > \varepsilon) < \varepsilon\}$$

where $d'(\vec{X}, \vec{Y}) = \sqrt{\Sigma(X_i - Y_i)^2}$, $\vec{X} = (X_1, \ldots, X_d)$, $\vec{Y} = (Y_1, \ldots, Y_d)$.

6. Kantorovich-Rubinstein Metric:

$$d^2(\vec{X}, \vec{Y}) = \inf\{E \, ||\vec{X}_1 - \vec{Y}_1||^2 : \vec{X}_1 \stackrel{d}{=} \vec{X} \ , \ \vec{Y}_1 \stackrel{d}{=} \vec{Y}\}.$$

We shall now prove that all of the above mentioned metrics satisfy the definition stated in the chapter introduction.

Theorem 5. The functional called the Levy metric which is defined as follows,

$$d(X, Y) = \inf\{\varepsilon > 0: F_X(x - \varepsilon) - \varepsilon \leq F_Y(x) \leq F_X(x + \varepsilon) + \varepsilon\,,\ x \in \mathbb{R}\}$$

where F_X and F_Y are distribution functions of X and Y respectively

satisfies the definition of a metric.

Proof. (i) $P(X = Y) = 1 \Rightarrow d(X, Y) = 0$.

If $P(X = Y) = 1$, then $F_X = F_Y$ so we have

$$d(X, X) = \inf\{\varepsilon > 0: F_X(x - \varepsilon) - \varepsilon \leq F_X(x) \leq F_X(x + \varepsilon) + \varepsilon,\ x \in \mathbb{R}\} = 0.$$

(ii) $d(X, Y) = d(Y, X)$.

We know $d(X, Y) = \inf\{\varepsilon > 0: F_X(x - \varepsilon) - \varepsilon \leq F_Y(x) \leq F_X(x + \varepsilon) + \varepsilon,\ x \in \mathbb{R}\}$. Thus if $F_X(x - \varepsilon') - \varepsilon' \leq F_Y(x) \leq F(x + \varepsilon') + \varepsilon'$ then $\varepsilon' \geq d(X, Y)$. Now $d(Y, X) = \inf\{\varepsilon > 0: F_Y(x - \varepsilon) - \varepsilon \leq F_X(x) \leq F_Y(x + \varepsilon) + \varepsilon,\ x \in \mathbb{R}\}$. Hence, there exists $\varepsilon' > 0$ such that

$$F_Y(x - \varepsilon') - \varepsilon' \leq F_X(x) \leq F_Y(x + \varepsilon') + \varepsilon' \quad \text{for all } x \in \mathbb{R}$$

$\Rightarrow$ (1) $F_Y(x - \varepsilon') - \varepsilon' \leq F_X(x)\ \forall x$ and (2) $F_X(x) \leq F_Y(x + \varepsilon') + \varepsilon'\ \forall\, x$.

Let $x = x + \varepsilon'$. Then

$$F_Y(x + \varepsilon' - \varepsilon') - \varepsilon' \leq F_X(x + \varepsilon') \qquad \text{(from (1))}$$

$$\Rightarrow F_Y(x) \leq F_X(x + \varepsilon') + \varepsilon'$$

Now let $x = x - \varepsilon'$. Then

$$F_X(x - \varepsilon') \leq F_Y(x - \varepsilon' + \varepsilon') + \varepsilon' \qquad \text{(from (2))}$$

$$\Rightarrow F_X(x - \varepsilon') - \varepsilon' \leq F_Y(x) \quad .$$

Therefore, if $\varepsilon' > 0$ satisfies $F_Y(x - \varepsilon') - \varepsilon' \leq F_X(x) \leq F_Y(x + \varepsilon') + \varepsilon'$, $\forall x \in \mathbb{R}$ then $F_X(x - \varepsilon') - \varepsilon' \leq F_Y(x) \leq F(x + \varepsilon') + \varepsilon'$

$$\Rightarrow \varepsilon' \geq d(X, Y) \quad .$$

Taking the infinum of all ε' ,

$$\Rightarrow d(Y, X) \geq d(X, Y) \ .$$

Similarly, we can get $d(X, Y) \geq d(Y, X)$ by switching $d(X, Y)$ to $d(Y, X)$ in the above steps and making the necessary changes. Thus, since $d(X, Y) \geq d(Y, X)$ and $d(Y, X) \geq d(X, Y)$ we know $d(X, Y) = d(Y, X)$.

(iii) $d(X, Y) \leq d(X, Z) + d(Z, Y)$.

Since $d(X, Y) = \inf\{\varepsilon > 0: F_X(x - \varepsilon) - \varepsilon \leq F_Y(x) \leq F_X(x + \varepsilon) + \varepsilon\}$ then we know if $F_X(x - \varepsilon') - \varepsilon' \leq F_Y(x) \leq F_X(x + \varepsilon') + \varepsilon'$ then $\varepsilon' \geq d(X, Y)$. Now $d(X, Z) = \inf\{\varepsilon > 0: F_X(x - \varepsilon) - \varepsilon \leq F_Z(x) \leq F_X(x + \varepsilon) + \varepsilon\}$ where F_Z is the distribution function of Z .

$\Rightarrow \exists\, \varepsilon_1 > 0$

such that

(1) $F_X(x - \varepsilon_1) - \varepsilon_1 \leq F_Z(x) \;\forall\, x$

and

(2) $F_Z(x) \leq F_X(x + \varepsilon_1) + \varepsilon_1 \;\forall\, x$.

Also $d(Z, Y) = \inf\{\varepsilon > 0: F_Z(x - \varepsilon) - \varepsilon \leq F_Y(x) \leq F_Z(x + \varepsilon) + \varepsilon\}$

$\Rightarrow \exists\, \varepsilon_2 > 0$

such that

(3) $F_Z(x - \varepsilon_2) - \varepsilon_2 \leq F_Y(x) \;\forall\, x$

and

(4) $F_Y(x) \leq F_Z(x + \varepsilon_2) + \varepsilon_2 \;\forall\, x$.

Let $x = x + \varepsilon_2$. Then from (3)

$$F_Z(x + \varepsilon_2 - \varepsilon_2) - \varepsilon_2 \leq F_Y(x + \varepsilon_2)$$

$$\Rightarrow F_Z(x) \leq F_Y(x + \varepsilon_2) + \varepsilon_2$$

$$\Rightarrow F_X(x - \varepsilon_1) - \varepsilon_1 \leq F_Y(x + \varepsilon_2) + \varepsilon_2 \;\forall\, x$$

(from (1))

$$\Rightarrow F_X(x - \varepsilon_1 - \varepsilon_2) - \varepsilon_1 \leq F_Y(x + \varepsilon_2 - \varepsilon_2) + \varepsilon_2$$

$(x = x - \varepsilon_2)$

$$\Rightarrow F_X(x - (\varepsilon_1 + \varepsilon_2)) - (\varepsilon_1 + \varepsilon_2) \leq F_Y(x) \ .$$

Now let $x = x - \varepsilon_2$. Then from (4)

$$F_Y(x - \varepsilon_2) \leq F_Z(x + \varepsilon_2 - \varepsilon_2) + \varepsilon_2$$

$$\Rightarrow F_Y(x - \varepsilon_2) - \varepsilon_2 \leq F_Z(x)$$

$$\Rightarrow F_Y(x - \varepsilon_2) - \varepsilon_2 \leq F_X(x + \varepsilon_1) + \varepsilon_1 \ \forall x$$

(from (2))

$$\Rightarrow F_Y(x + \varepsilon_2 - \varepsilon_2) - \varepsilon_2 \leq F_X(x + \varepsilon_2 + \varepsilon_1) + \varepsilon_1$$

$(x = x + \varepsilon_2)$

$$\Rightarrow F_Y(x) \leq F_X(x + (\varepsilon_1 + \varepsilon_2)) + (\varepsilon_1 + \varepsilon_2) \ .$$

Therefore, if ε_1 and $\varepsilon_2 > 0$ satisfy (1 - 4) then

$$F_X(x - (\varepsilon_1 + \varepsilon_2)) - (\varepsilon_1 + \varepsilon_2) \leq F_Y(x)$$

$$\leq F_X(x + (\varepsilon_1 + \varepsilon_2)) + (\varepsilon_1 + \varepsilon_2)$$

$$\Rightarrow d(X, Y) \leq \varepsilon_1 + \varepsilon_2$$

and taking the infinum of ε_1 and ε_2,

$$\Rightarrow d(X, Y) \leq d(X, Z) + d(Z, Y) \ .$$

Theorem 6. The functional called the Uniform Metric which is defined as follows:

$$d(X, Y) = \sup\{|F_X(x) - F_Y(x)| : x \in \mathbb{R}\}$$

satisfies the definition of a metric.

Proof. (i) $P(X = Y) = 1 \Rightarrow d(X, Y) = 0$.

If $P(X = Y) = 1$ then $F_X = F_Y$ so we have

$$d(X, X) = \sup\{|F_X(x) - F_X(x)| : x \in \mathbb{R}\} = 0 \quad .$$

(ii) $d(X, Y) = d(Y, X)$.

We know that

$$|F_X(x) - F_Y(x)| = |F_Y(x) - F_X(x)| \; \forall \, x$$

and thus

$$\sup\{|F_X(x) - F_Y(x)| : x \in \mathbb{R}\} = \sup\{|F_Y(x) - F_X(x)| : x \in \mathbb{R}\} \quad .$$

Thus $d(X, Y) = d(Y, X)$.

(iii) $d(X, Y) \leq d(X, Z) + d(Z, Y)$.

By the triangle inequality, for absolute values, we know

$$|F_X(x) - F_Y(x))| \leq |F_X(x) - F_Z(x)| + |F_Z(x) - F_Y(x)|$$

$$\Rightarrow \sup\{|F_X(x) - F_Y(x)| : x \in \mathbb{R}\}$$

$$\leq \sup\{|F_X(x) - F_Z(x)| + |F_Z(x) - F_Y(x)| : x \in \mathbb{R}\}$$

$$\Rightarrow \sup\{|F_X(x) - F_Y(x)| : x \in \mathbb{R}\}$$

$$\leq \sup\{|F_X(x) - F_Z(x)| : x \in \mathbb{R}\}$$

$$+ \sup\{|F_Z(x) - F_Y(x)| : x \in \mathbb{R}\}$$

$$\Rightarrow d(X, Y) \leq d(X, Y) + d(Z, Y) .$$

Theorem 7. The functional called the Distance of Total Variation, which is defined as follows:

$$d(X, Y) = \frac{1}{2} \int_{-\infty}^{\infty} |f_X(x) - f_Y(x)| dx$$

where f_X and f_Y are the densities of X and Y respectively, satisfies the definition of a metric.

Proof. (i) $P(X = Y) = 1 \Rightarrow d(X, Y) = 0$.

If $P(X = Y) = 1$ then $f_X = f_Y$ so we have

$$d(X, X) = \frac{1}{2} \int_{-\infty}^{\infty} |f_X(x) - f_X(x)| dx = \frac{1}{2} \int_{-\infty}^{\infty} 0 \ dx = 0 .$$

(ii) $d(X, Y) = d(Y, X)$.

We know that

$$|f_X(x) - f_Y(x)| = |f_Y(x) - f_X(x)| \quad \forall x$$

and thus

$$\frac{1}{2} \int_{-\infty}^{\infty} |f_X(x) - f_Y(x)| dx = \frac{1}{2} \int_{-\infty}^{\infty} |f_Y(x) - f_X(x)| dx .$$

Therefore, $d(X, Y) = d(Y, X)$.

(iii) $d(X, Y) \leq d(X, Z) + d(Z, Y)$.

By the triangle inequality, for absolute values, we know

$$|f_X(x) - f_Y(x)| \leq |f_X(x) - f_Z(x)| + |f_Z(x) - f_Y(x)| \ , \ \forall\ x$$

$$\Rightarrow \int_{-\infty}^{\infty} |f_X(x) - f_Y(x)|dx$$

$$\leq \int_{-\infty}^{\infty} [|f_X(x) - f_Z(x)| + |f_Z(x) - f_Y(x)|]\ dx$$

$$\Rightarrow \int_{-\infty}^{\infty} |f_X(x) - f_Y(x)|dx$$

$$\leq \int_{-\infty}^{\infty} |f_X(x) - f_Z(x)|dx + \int_{-\infty}^{\infty} |f_Z(x) - f_Y(x)|dx$$

$$\Rightarrow d(X, Y) \leq d(X, Z) + d(Z, Y) \ .$$

Lemma 3. For arbitrary real numbers x, y, z and $0 < p < 1$

$$|x - y|^p \leq |x - z|^p + |z - y|^p \ .$$

Proof. Let $a = x - z$ and $b = z - y$ then $x - y = a + b$.

We must now show $|a + b| \leq |a|^p + |b|^p$.

Case 1. a and b have opposite signs.

If a and b have opposite signs then

$$|a + b| \leq |a|$$

or

$$|a + b| \leq |b| \ .$$

If $|a + b| \leq |a|$ then

$$|a + b|^p \leq |a|^p$$

$$\Rightarrow |a + b|^p \leq |a|^p + |b|^p .$$

Likewise, if $|a + b| \leq |b|$ then

$$|a + b|^p \leq |b|^p$$

$$= |a + b|^p \leq |b|^p + |a|^p .$$

Case 2. a and b have the same signs.

If a and b are the same signs then we can assume they are both positive (if both are negative then we can factor out the negative).

We will show

$$|a + b|^p - |a|^p \leq |b|^p$$

or, since a and b are positive we must show

$$(a + b)^p - a^p \leq b^p .$$

Now

$$(a + b)^p - a^p = \int_a^{a+b} px^{p-1}\, dx$$

(by Fundamental Theorem of Calculus)

$$= \int_0^b p(u + a)^{p-1}\, du$$

$$\leq \int_0^b pu^{p-1}\, du$$

$$= b^p$$

where the inequality follows from the fact that the exponent $p - 1$ is negative.

Theorem 8. The functional called the L_p-metric, which is defined as follows,

$$d_p(X, Y) = [\int |F_X(x) - F_Y(x)|^p\, dx]^{\min(1,\, 1/p)} \qquad p > 0$$

satisfies the definition of a metric.

Proof. (i) $P(X = Y) = 1 \Rightarrow d(X, Y) = 0$.

If $P(X = Y) = 1$ then $F_X(x) = F_Y(x)$ so we have

$$d_p(X, X) = [\int |F_X(x) - F_X(x)|^p\, dx]^{\min(1,\, 1/p)} = 0 .$$

(ii) $d(X, Y) = d(Y, X)$.

Since $|F_X(x) - F_Y(x)| = |F_Y(x) - F_X(x)| \ \forall x$ we have that

$$[\int |F_X(x) - F_Y(x)|^p \, dx]^{\min(1, 1/p)}$$

$$= [\int |F_Y(x) - F_X(x)|^p \, dx]^{\min(1, 1/p)} .$$

(iii) $d(X, Y) \leq d(X, Z) + d(Z, Y)$.

Case 1. $p \geq 1$.

By Minkowski's inequality,

$$[\int_{-\infty}^{\infty} |F_X(x) - F_Y(x)|^p \, dx]^{1/p}$$

$$\leq [\int_{-\infty}^{\infty} |F_X(x) - F_Z(x)|^p \, dx]^{1/p} + [\int_{-\infty}^{\infty} |F_Z(x) - F_Y(x)|^p dx]^{1/p}$$

$$\Rightarrow d(X, Y) \leq d(X, Z) + d(Z, Y) .$$

Case 2. $0 < p < 1$.

By Lemma 3,

$$|F_X(x) - F_Y(x)|^p \leq |F_X(x) - F_Z(x)|^p + |F_Z(x) - F_Y(x)|^p .$$

Thus

$$\int_{-\infty}^{\infty} |F_X(x) - F_Y(x)|^p \, dx$$

$$\leq \int_{-\infty}^{\infty} |F_X(x) - F_Z(x)|^p \, dx + \int_{-\infty}^{\infty} |F_Z(x) - F_Y(x)|^p \, dx .$$

Lemma 4. $P[\xi + \eta > a + b] \leq P[\xi > a] + P[\eta > b]$ for any random variables ξ and η and reals a, b .

Proof. It is enough to show

$$\{\xi + \eta > a + b\} \subset \{\xi > a\} \cup \{\eta > b\}$$

and then the result will follow. Now

$$\{\xi + \eta > a + b\} \subset \{\xi > a\} \cup \{\eta > b\}$$

if and only if

$$\{\xi + \eta > a + b\}^c \supset [\{\xi > a\} \cup \{\eta > b\}]^c$$

$$\Rightarrow \{\xi + \eta \leq a + b\} \supset \{\xi \leq a\} \cap \{\eta \leq b\} \quad .$$

Now if $\xi \leq a$ and $\eta \leq b$ then $\xi + \eta \leq a + b$ so

$$\{\xi + \eta > a + b\}^c \supset [\{\xi > a\} \cup \{\eta > b\}]^c$$

$$\Rightarrow \{\xi + \eta > a + b\} \supset \{\xi > a\} \cup \{\eta > b\} \quad .$$

Thus

$$P\{\xi + \eta > a + b\} \leq P\{\{\xi > a\} \cup \{\eta > b\}\}$$

$$\leq P\{\xi > a\} + P\{\eta > b\} \quad .$$

Theorem 9. The functional called the Ky Fan metric, which is defined as follows:

$$d(\vec{X}, \vec{Y}) = \inf\{\varepsilon: P(d'(\vec{X}, \vec{Y}) > \varepsilon) < \varepsilon\}$$

where

$$d'(\vec{X}, \vec{Y}) = \sqrt{\Sigma(X_i - Y_i)^2}$$

satisfies the definition of a metric.

Proof. (i) $P(\vec{X} = \vec{Y}) = 1 \Rightarrow d(\vec{X}, \vec{Y}) = 0$.

If $P(\vec{X}, \vec{Y}) = 1$ then $d'(\vec{X}, \vec{Y}) = 0$ and

$$\inf\{\varepsilon: P(0 > \varepsilon) < \varepsilon\} = 0 \quad .$$

(ii) $d(\vec{X}, \vec{Y}) = d(\vec{Y}, \vec{X})$.

$$\sqrt{\Sigma(X_i - Y_i)^2} = \sqrt{\Sigma(Y_i - X_i)^2}$$

and thus

$$d'(\vec{X}, \vec{Y}) = d'(\vec{Y}, \vec{X}) \quad .$$

Therefore,

$$\inf\{\varepsilon: P(d'(\vec{X}, \vec{Y}) > \varepsilon) < \varepsilon\}$$

$$= \inf\{\varepsilon: P(d'(\vec{Y}, \vec{X}) > \varepsilon) < \varepsilon\}$$

and thus

$$d(\vec{X}, \vec{Y}) = d(\vec{Y}, \vec{X}) .$$

(iii) $d(\vec{X}, \vec{Y}) \leq d(\vec{X}, \vec{Z}) + d(\vec{Z}, \vec{Y})$.

Define $d(\vec{X}, \vec{Z}) = a$ and $d(\vec{Z}, \vec{Y}) = b$; thus

$$d(\vec{X}, \vec{Z}) + d(\vec{Z}, \vec{Y}) = a + b .$$

We must show $d(\vec{X}, \vec{Y}) \leq a + b$. Take a $\delta > 0$, then by the definition of $d(\vec{Z}, \vec{Y})$ and $d(\vec{X}, \vec{Z})$

(a) $P[d'(\vec{X}, \vec{Z}) > a + \delta] < a + \delta$

(b) $P[d'(\vec{Z}, \vec{Y}) > b + \delta] < b + \delta$.

By the triangle inequality for the Euclidean metric

$$d'(\vec{X}, \vec{Y}) \leq d'(\vec{X}, \vec{Z}) + d'(\vec{Z}, \vec{Y}) .$$

Therefore,

$$P[d'(\vec{X}, \vec{Y}) > a + b + 2\delta] \leq P[d'(\vec{X}, \vec{Z}) + d'(\vec{Z}, \vec{Y}) > a + b + 2\delta]$$

$$\leq P[d'(\vec{X}, \vec{Z}) > a + \delta] + P[d'(\vec{Z}, \vec{Y}) > b + \delta] \quad \text{(by Lemma 4)}$$

$$< a + \delta + b + \delta = a + b + 2\delta .$$

Thus

$$d(\vec{X}, \vec{Y}) \leq a + b + 2\delta \ .$$

Then by letting $\delta \to 0$ we get

$$d(\vec{X}, \vec{Y}) \leq a + b$$

$$\Rightarrow d(\vec{X}, \vec{Y}) \leq d(\vec{X}, \vec{Z}) + d(\vec{Z}, \vec{Y}) \ .$$

A proof of property 3 for the Kantorovich-Rubinstein metric is very difficult. It is known that it uses the Kolmogorov's Extension Theorem and because of its difficulty, we will omit it. Proofs of properties 1 and 2 are trivial.

2.3. Moments of Mahalanobis distance and Probability Metrics

We now will examine the connection between the Mahalanobis distance and probability metrics.

Theorem 10. The square root of the first moment of the interclass Mahalanobis distance is not a probability metric.

Proof. Let $f(\vec{X}, \vec{Y}) = \sqrt{E[(\vec{X} - \vec{m}_Y)^T C_Y^{-1} (\vec{X} - \vec{m}_Y)]}$.

We wish to show that the functional $f(\vec{X}, \vec{Y})$ is not a probability metric. Now if $\vec{X} = \vec{Y}$ then by Corollary 1 we know that $(\vec{X} - \vec{m}_X)^T C_y^{-1} (\vec{X} - \vec{m}_Y) \sim \chi^2$. Thus $f(\vec{X}, \vec{X}) = \sqrt{E(K)}$ where $K \sim \chi^2$.

Now the expected value of a random variable that is χ^2 is equal to r where r is a positive integer that is the number of degrees

of freedom. Thus $f(\vec{X}, \vec{X}) \neq 0$ and hence it fails the first condition for a probability metric.

Since the square root of the first moment of the interclass Mahalanobis distance is not itself a metric, we will see if it is comparable to another metric which we can obtain an explicit form for with two normal random variables. The probability metric that we will examine is the Kantorovich-Rubinstein metric.

We will examine the Kantorovich-Rubinstein metric for two normal random variables X and Y, i.e., $d = 1$. In this case the Kantorovich-Rubinstein metric can be easily computed.

Proposition 1. If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ then the Kantorovich-Rubinstein metric can be expressed in the following simplified form:

$$d^2(X,Y) = \inf\{E|X_1 - Y_1|^2 : (X_1, Y_1) \quad \text{such that} \quad X_1 \stackrel{d}{=} X,\ Y_1 \stackrel{d}{=} Y\}$$

$$= (\sigma_1 - \sigma_2)^2 + (\mu_1 - \mu_2)^2 .$$

Proof. We have that

$$(1)\quad (X_1, Y_1) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)\ , \quad (2)\quad X \sim N(\mu_1, \sigma_1^2)$$

$$\Rightarrow X_1 \sim N(\mu_1, \sigma_1^2)$$

and

$$(3)\quad Y \sim N(\mu_2, \sigma_2^2) \Rightarrow Y_1 \sim N(\mu_2, \sigma_2^2) .$$

So

$$E|X_1 - Y_1|^2 = E(X_1 - Y_1)^2$$

$$= E(X_1^2 - 2X_1Y_1 + Y_1^2)$$

$$= E(X_1^2) - 2E(X_1Y_1) + E(Y_1^2)$$

$$= Var(X_1) + [E(X_1)]^2 - 2E(X_1Y_1) + Var(Y_1)$$

$$+ [E(Y_1)]^2$$

$$= \sigma_1^2 + \mu_1^2 - 2E(X_1Y_1) + \sigma_2^2 + \mu_2^2 \quad .$$

Now

$$\rho = Cor(X_1, Y_1) = \frac{Cov(X_1, Y_1)}{\sigma_1\sigma_2}$$

$$= \frac{E(X_1Y_1) - E(X_1)\,E(Y_1)}{\sigma_1\sigma_2}$$

$$= (E(X_1Y_1) - \mu_1\mu_2)/\sigma_1\sigma_2 \quad .$$

Therefore

$$\rho(\sigma_1\sigma_2) + \mu_1\mu_2 = E(X_1Y_1) \ .$$

Hence

$$E|X_1 - Y_1|^2 = \sigma_1^2 + \mu_1^2 - 2\rho\sigma_1\sigma_2 - 2\mu_1\mu_2 + \sigma_2^2 + \mu_2^2$$

$$= \sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2 + \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2 \quad .$$

Now $\sigma > 0$, $\sigma' > 0$ and $-1 \leq \rho \leq 1$ so $\inf E|X_1 - Y_1|^2$ is when

$$\rho = 1$$

$$\Rightarrow \inf E|X_1 - Y_1|^2 = \sigma_1^2 - 2\sigma_1\sigma_2 + \sigma_2^2 + \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2$$

$$= (\sigma_1 - \sigma_2)^2 + (\mu_1 - \mu_2)^2 \quad .$$

It is also easy to compute the first moment of the interclass Mahalanobis distance for two normal random variables X and Y.

Proposition 2. If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, then

$$E[R(X, \mu_2)] = \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{\sigma_2^2} \quad .$$

Proof. $R(X, \mu_2) = (X - \mu_2)(\sigma_2^2)^{-1}(X - \mu_2)$

$$= (X - \mu_2)^2/\sigma_2^2 \quad .$$

Thus

$$E[R(X, \mu_2)] = E[(X - \mu_2)^2/\sigma_2^2]$$

$$= \left[1/\sigma_2^2\right] E(X^2 - 2\mu_2 X + \mu_2^2)$$

$$= \left[1/\sigma_2^2\right](\sigma_1^2 + \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2)$$

$$= [\sigma_1^2 + (\mu_1 - \mu_2)^2]/\sigma_2^2 \quad .$$

Remark. From the two propositions above, we can see that there is no obvious relation between the Kantorovich-Rubinstein metric and the interclass Mahalanobis distance for $d = 1$. When $d \geq 2$, and we are concerned with normal random vectors, the following form of the Kantorovich-Rubinstein metric is available [8].

$$d^2(\vec{X}, \vec{Y}) = ||\vec{m}_x - \vec{m}_y||^2 + \mathrm{tr}\{C_x + C_y - 2(C_x C_y)^{1/2}\}, \quad (\mathrm{tr} \equiv \mathrm{trace}).$$

We can see that this can be viewed as a natural extension of the formula in Proposition 1 as follows: $(\mu_1 - \mu_2)^2$ is replaced by $||m_x - m_y||^2$ and $(\sigma_1 - \sigma_2)^2$ is replaced by $\mathrm{tr}(C_x^{1/2} - C_y^{1/2})^2$. We are also able to find the first moment of the interclass Mahalanobis distance using Theorem 2; however, these quantities do not seem to be easily comparable.

BIBLIOGRAPHY

BIBLIOGRAPHY

Gonzalez, R.C. and L.C. Howington, "Machine Recognition of Abnormal Behavior in Nuclear Reactors", IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-7, number 10, pp. 717-728.

Sebestyn, George S., Decision-Making Processes in Pattern Recognitions. New York: The Macmillan Company, 1962.

Batchelor, Bruce G. (editor), Pattern Recognition - Ideas in Practice. New York: Plenum Press, 1978.

Patrick, Edward A., Fundamentals of Pattern Recognition, Englewood, N.J.: Prentice-Hall, Inc., 1972.

Gonzalez, R.C. and C.G. Wagner, "Moments of the interclass Mahalanobis distance", IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-13, number 6, pp. 1135-1139.

Zolotarev, V.M., "Probability metrics". Theory of Probability and Its Applications, vol. 28, number 3, pp. 278-302.

Rachev, S.T., "The Monge-Kantorovich Mass Transference Problem and Its Stochastic Applications", Theory of Probability and Its Applications, vol. 29, number 4, pp. 647-676.

VITA

Tamela Willett was born on April 22, 1963 in St. Johns, Michigan. She graduated from St. Johns High School in 1981. The following August she entered Alma College in Alma, Michigan and received a Bachelor of Science degree in Mathematics in April, 1985.

The following September, she began study toward a Master of Science degree at The University of Tennessee, Knoxville, and worked as a Graduate Teaching Assistant. She completed the degree in March of 1988 and returned to Michigan.