Statistical Papers **47**, 631-641 (2006)

Statistical Papers

Note

Statistical properties of distance estimators

Dechang Chen[1], Michael Fries[2], Xiuzhen Cheng[3]

[1] Division of Epidemiology and Biostatistics, Uniformed Service University of the Health Sciences, Bethesda, MD 20814; (e-mail: dchen@usuhs.mil)
[2] School of Computer Science, Telecommunications and Information Systems, DePaul University, Chicago, IL 60604; (e-mail: mfries@cs.deapul.edu)
[3] Department of Computer Science, The George Washington University, Washington, DC 20052; (e-mail: cheng@seas.gwu.edu)

Received: April 29, 2004; revised version: July 5, 2004

Abstract Estimating the distance between two points is of fundamental concern. This paper investigates some statistical properties of three estimators of the distance between two points on a plane. The results of several theoretical comparisons of the performance of the estimators assuming a large sample size are given. Also given is the comparison of the performance of the estimators using simulation when the sample size is small. These comparisons suggest that the estimator of choice is not the most "natural" estimator in this situation. Although the discussion is given in the framework of the plane, the results are readily extended to high dimensional spaces.

Key words Metric - MLE - UMVUE

1 Introduction

What is the distance between two points? This seemingly simple question can take many forms and has been the subject of a good deal of work. Perhaps the simplest example is the (Euclidean) distance between two points (vectors) in a p-dimensional Euclidean space $p \geq 2$. Let $\mathbf{a}' = (a_1, a_2, \ldots, a_p)$ and $\mathbf{b}' = (b_1, b_2, \ldots, b_p)$ be two points with $'$ denoting the transpose operation. Then the Euclidean distance between these two points is $d(\mathbf{a}, \mathbf{b}) = (\sum_{k=1}^{p}(a_k - b_k)^2)^{1/2}$. The Euclidean distance can be generalized in many different ways. One example of generalizations is the metric ([1]), which has wide applications in various areas such as differential equations, signal processing and control theories. Another example is the dissimilarity measure ([2]), which is fundamental to clustering techniques and recently has found important applications in microarray data analysis. We note that the Euclidean distance and its generalizations are often applied to compute the distance between two objects when the numerical characteristics (e.g., coordinates) of the objects are completely known. If such precise numerical information is not available, how can we proceed to obtain the distance? This paper will provide an answer for the case concerning the distance between two points on a plane. To do so, statistical properties of estimators of the distance need to be examined. This paper is motivated by the following situation.

Orthopaedic surgeons, before implanting metal stems into the bones of patients, need to match as closely as possible bone geometry and the geometry of the metal stem. At the time of this investigation, this was done by comparing x-rays of the bone to the stem templates. Before CT-scan measurements will be commonly used, it is important to understand sources and magnitudes of errors made when this technique is used to measure distances. Suppose A and B are two different points on a CT-scan image of bone cross-section. Also suppose $\{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n\}$ and $\{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n\}$ represent coordinates of the independent measurements of A and B in the source coordinate system. The question is how these measurements should be used to determine the distance d_{AB} between A and B.

In order to approach the problem in a statistical way, we may assume A, B, $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$ and $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n$ all lie in a two dimensional Euclidean plane. In this paper, we study properties of three estimators of d_{AB}. The first estimator d_1 is natural and is obtained by taking the average of distances between pairs of measurements. This is the estimator that is sometimes used in orthopaedic experiments. The second estimator d_2 is also natural and is just the distance of average locations. The third estimator d_3, defined in Section 3, is less natural. Section 2 of this paper shows that d_2 outperforms d_1 for large sample sizes under the normality assumption. Section 3 defines d_3 and shows that under the normality assumption, $(d_3)^2$ is the uniform minimum variance unbiased estimator (UMVUE) of $(d_{AB})^2$ while d_2 and $(d_2)^2$ are the maximum likelihood estimator (MLE) of d_{AB} and

$(d_{AB})^2$, respectively. Comparison between d_2 and d_3 for large sample sizes is also provided. In Section 4, some simulation studies are presented to compare the performance of three estimators for small samples. The conclusion is given in Section 5.

2 Two Natural Estimators

Let μ_A and μ_B denote the coordinates of A and B respectively. For convenient comparisons, we assume that $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$ is a random sample from a bivariate normal distribution $N(\mu_A, \Sigma_1)$, and $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n$ an independent random sample from a bivariate normal distribution $N(\mu_B, \Sigma_2)$, where μ_A and μ_B are two unknown mean vectors and Σ_1 and Σ_2 are two unknown variance-covariance matrices (assumed to be positive definite). The above assumption will be used through out this paper. Our main concern is how one can use $\mathbf{U}_i$'s and $\mathbf{V}_i$'s to obtain a good estimator of the distance between μ_A and μ_B, i.e. d_{AB}.

Let $||\mathbf{x}||$ denote the usual norm of a vector $\mathbf{x}$ in a 2-dimensional Euclidean plane, i.e., $||\mathbf{x}|| = \sqrt{x_1^2 + x_2^2}$ for $\mathbf{x}' = (x_1, x_2)$. Then the distance between μ_A and μ_B is just $d_{AB} = ||\mu_A$ - $\mu_B||$. There are two intuitive ways for us to proceed to obtain estimators of d_{AB}. One intuitive way is to take the arithmetic average of the distances $||\mathbf{U}_i - \mathbf{V}_i||$, i.e., $\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{U}_i - \mathbf{V}_i||$, denoted $d_1(n)$. Another way is to look at unbiased estimators of $\mu_A - \mu_B$. Such an unbiased estimator is the difference of the averages of the two samples, given as $\bar{\mathbf{U}} - \bar{\mathbf{V}}$, where $\bar{\mathbf{U}} = \frac{1}{n}\sum_1^n \mathbf{U}_i$ and $\bar{\mathbf{V}} = \frac{1}{n}\sum_1^n \mathbf{V}_i$. Thus one may think of using $||\bar{\mathbf{U}} - \bar{\mathbf{V}}||$, denoted $d_2(n)$, to estimate d_{AB}. Now a natural question arises: which of $d_1(n)$ and $d_2(n)$ provides a better estimate? The following lemma suggests that $d_2(n)$ should be preferred to $d_1(n)$ if n is large.

Lemma 1 *For any finite n, both $d_1(n)$ and $d_2(n)$ are biased estimators of d_{AB}. For $n \to \infty$, $d_1(n)$ overestimates d_{AB} while $d_2(n)$ converges to d_{AB}.*

PROOF. Let $\mathbf{X} = \mathbf{U}_1 - \mathbf{V}_1$. Then $\mathbf{X} \sim N(\mu, \Sigma)$, where $\mu = \mu_A - \mu_B$ and $\Sigma = \Sigma_1 + \Sigma_2$ is positive definite. Clearly, $P(a + \mathbf{b}'\mathbf{X} < 0) > 0$ for any fixed number a and vector $\mathbf{b} = (b_1, b_2)'$ with $\mathbf{b}'\mathbf{X} \neq 0$. Therefore, $P(||\mathbf{X}|| = a + \mathbf{b}'\mathbf{X}) = 1 - P(||\mathbf{X}|| \neq a + \mathbf{b}'\mathbf{X}) \leq 1 - P(a + \mathbf{b}'\mathbf{X} < 0) < 1$. In other words, $P(||\mathbf{X}|| = a + \mathbf{b}'\mathbf{X}) \neq 1$ for any a and vector $\mathbf{b}$. This and the fact that the function $||x||$ is convex show that the strict form of Jensen's Inequality applies: $Ed_1(n) = E||\mathbf{U}_1 - \mathbf{V}_1|| = E||\mathbf{X}|| > ||E\mathbf{X}|| = ||\mu_A - \mu_B|| = d_{AB}$, where E refers to the expectation operation. Similarly, $Ed_2(n) = E||\bar{\mathbf{U}} - \bar{\mathbf{V}}|| > ||E\bar{\mathbf{U}} - E\bar{\mathbf{V}}|| = ||\mu_A - \mu_B|| = d_{AB}$. The above proves the first statement of the lemma.

By the strong law of large numbers, it is seen that as n $\to \infty$,

$$d_2(n) = ||\bar{\mathbf{U}} - \bar{\mathbf{V}}|| \xrightarrow{a.s.} ||E\mathbf{U}_1 - E\mathbf{V}_1|| = ||\mu_A - \mu_B||,$$

and

$$d_1(n) = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{U}_i - \mathbf{V}_i|| \xrightarrow{a.s.} E||\mathbf{U}_1 - \mathbf{V}_1|| > ||\mu_A - \mu_B||.$$

This completes the proof of the second statement.

■

3 A Third Estimator

To obtain another estimator of d_{AB}, we begin with a UMVUE of $(d_{AB})^2$. Intuitively the square root of a UMVUE of $(d_{AB})^2$ should provide a good estimator of d_{AB}.

Let $\mathbf{X}_i = \mathbf{U}_i - \mathbf{V}_i$. Then $\mathbf{X}_i \sim N(\mu, \Sigma)$, where $\mu = \mu_A - \mu_B$ and $\Sigma = \Sigma_1 + \Sigma_2$. Let $\mathbf{X}_i' = (X_{i1}, X_{i2})$, $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, for $j = 1, 2$, and matrix $S = (s_{lj})$ with $s_{lj} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{kl} - \bar{X}_l)(X_{kj} - \bar{X}_j)$ for $l, j = 1, 2$. s_{jj} is the sample variance of the sample $X_{1j}, X_{2j}, \ldots, X_{nj}$ and will also be denoted by S_j^2. Set $d_3(n) = (\bar{\mathbf{X}}'\bar{\mathbf{X}} - \frac{1}{n}tr(S))^{1/2}$, where $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$ and $tr(S)$ denotes the trace of the matrix S.

Lemma 2 *$(d_3(n))^2$ is a UMVUE of $(d_{AB})^2$.*

Proof. Let $\mu' = (\mu_1, \mu_2)$. From results on univariate normal distributions it is well known that $\bar{x}_j^2 - \frac{1}{n}S_j^2$ is a UMVUE of μ_j^2 for $j = 1, 2$ [See, for example, Problem 2.5 of Chapter 2, [3]]. Then since the sum of UMVUEs is also a UMVUE [See, for example, 5a.2.(e) of [4]], it follows that $\bar{x}_1^2 - \frac{S_1^2}{n} + \bar{x}_2^2 - \frac{S_2^2}{n} = \bar{\mathbf{X}}'\bar{\mathbf{X}} - \frac{1}{n}tr(S)$ is the UMVUE of $\mu'\mu = (d_{AB})^2$.

■

Thus $(d_3(n))^2$ is the best estimator of $(d_{AB})^2$. Clearly, $\bar{\mathbf{X}}$ is the MLE of μ. By the invariance property of MLEs, $(d_2(n))^2 = \bar{\mathbf{X}}'\bar{\mathbf{X}}$ is the MLE of $\mu'\mu = (d_{AB})^2$ (and $d_2(n)$ is the MLE of d_{AB}). We note that the bias of $(d_2(n))^2$ in estimating $(d_{AB})^2$ is $E(d_2(n))^2 - (d_{AB})^2 = E[(d_3(n))^2 + \frac{S_1^2}{n} +$

$\frac{S_2^2}{n}] - (d_{AB})^2 = \frac{1}{n}(ES_1^2 + ES_2^2) = \frac{1}{n}(\sigma_1^2 + \sigma_2^2)$, where σ_1^2 and σ_2^2 are the upper left hand and lower right hand elements of Σ, respectively.

Lemma 2 implies that $d_3(n)$ could be used as a good estimator of d_{AB}. The following comparison shows that d_2 and d_3 are very close to each other when n is large.

Lemma 3 *$d_3(n) = d_2(n) + O(\frac{1}{n})$ a.s.*

PROOF. Using the previous notations, we see that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ now constitute a random sample from the distribution with mean μ and variance-covariance matrix Σ. It is well known that $S_j^2 \to \sigma_j^2$ a.s. as $n \to \infty$ for $j = 1, 2$. So $\frac{1}{n}(S_1^2 + S_2^2) = O(\frac{1}{n})$ a.s. Therefore $(d_3(n))^2 = (d_2(n))^2 - \frac{1}{n}(S_1^2 + S_2^2) = (d_2(n))^2 + O(\frac{1}{n})$ a.s. By the Taylor expansion, $((d_2(n))^2 + O(\frac{1}{n}))^{1/2} = d_2(n) + \frac{1}{2\sqrt{\xi}}O(\frac{1}{n})$, where ξ is some statistic taking values between $(d_2(n))^2$ and $(d_2(n))^2 + O(\frac{1}{n})$. Since $d_2(n) \to ||\mu||(\neq 0)$ a.s. as $n \to \infty$, it follows that $\frac{1}{2\sqrt{\xi}} = O(1)$ a.s. Thus $d_3(n) = d_2(n) + O(\frac{1}{n})$ a.s.

■

Note that the proof of Lemma 3 does not need the normality assumption imposed on $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n$ and $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n$.

4 Simulation Study

Most of the above analysis focuses on the comparisons of the three estimators $d_1(n)$, $d_2(n)$, and $d_3(n)$ for large n. For small values of n, the comparison may be done through simulation. Results of simulations under various forms of variance-covariance matrices show: a) For any n, the mean squared error (MSE) of $d_1(n)$ is larger than that of $d_2(n)$ or $d_3(n)$. b) The difference between the MSEs of $d_2(n)$ and $d_3(n)$ decreases as n increases, and such a difference usually becomes negligible for $n \geq 20$. c) The MSEs of the three estimators tend to stabilize after $n = 500$.

While c) shows the convergence rate of the estimators, both a) and b) indicate that for small n values, $d_2(n)$ or $d_3(n)$ should be used instead of $d_1(n)$. Figures 1 and 2 provide a typical plot for comparing the three estimators with small n and large n, respectively. In the plots, the two bivariate normal distributions $N(\mu_A, \Sigma_1)$ and $N(\mu_B, \Sigma_2)$ are such that $\mu'_A = (0,0)$, $\mu'_B = (10,0)$,

$$\Sigma_1 = \begin{pmatrix} 10 & -1 \\ -1 & 10 \end{pmatrix}, \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 10 & -1 \\ -1 & 10 \end{pmatrix}.$$

From the figures, it is seen that the performance of $d_1(n)$ is the worst and $d_2(n)$ and $d_3(n)$ have similar performance.

5 Conclusion

This paper investigates some statistical properties of three estimators d_1, d_2, and d_3 of distance d_{AB} between two points A and B on a plane. Although this presentation focuses on the distance between two points on a plane, the results are readily extended to three or higher dimensional cases. Motivated by the CT-scan measurement problem, the topic on examining statistical properties of various distance estimators may find many impor-

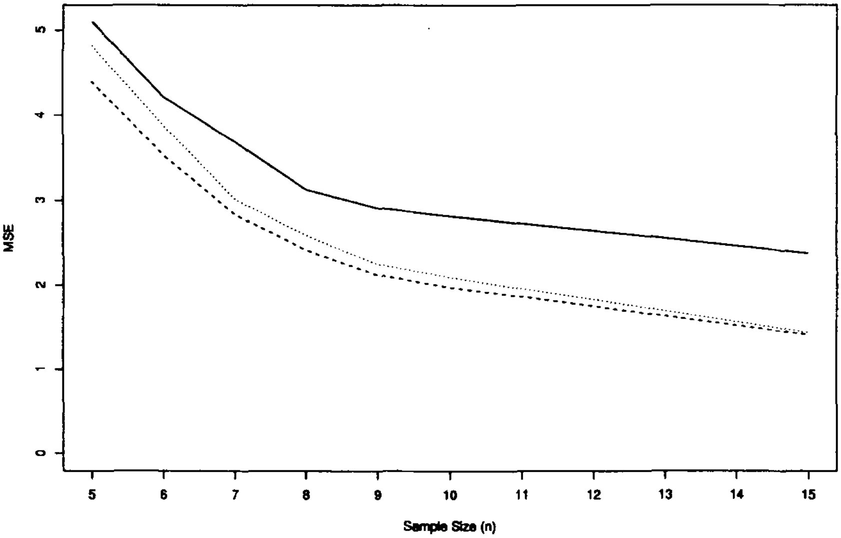

Fig. 1 MSEs of the three estimators against small sample size. The solid curve is the MSE of $d_1(n)$, dashed curve is the MSE of $d_2(n)$, and the dotted curve is the MSE of $d_3(n)$.

tant practical applications in other fields. For example, in wireless sensor network, a current hot area of computer science, range estimation for location discovery requires the distance computation between two reference points [see, for example, [5]]. When the reference points do not have deterministic location information, which is a common case, distance estimators such as those discussed in this paper may need to be applied for better performance.

Based on the work in this paper, we now make suggestions to the use of these three estimators. For large sample size n, Lemma 1 and Lemma 3 suggest that d_1 be discarded and either d_2 or d_3 be used to estimate d_{AB}.

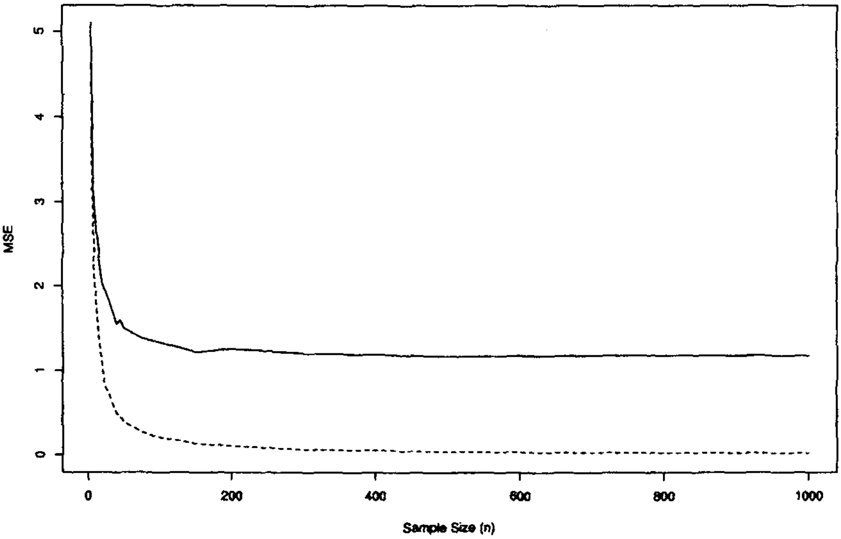

Fig. 2 MSEs of the three estimators against large sample size. The solid curve is the MSE of $d_1(n)$, dashed curve is the MSE of $d_2(n)$, and the dotted curve is the MSE of $d_3(n)$.

When n is small, we suggest the use of d_2 or d_3. This recommendation is based on Lemma 2 and simulation results described in Section 4.

Acknowledgments

The authors are grateful to the Editor and referees for many valuable comments.

References

1. Bryant, V. (1996), *Metric Spaces,* Cambridge University Press.

2. Hastie, T., Tibshirani, R., and Friedman, J.(2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

3. Lehmann, E. L, and Casella, G. (1998), *Theory of Point Estimation,* 2nd edition, Springer, New York.

4. Rao, C. R. (1965), *Linear Statistical Inference and Its Application,* Wiley, New York.

5. Ray, S., Ungrangsi, R., DePellegrini, F., Trachtenberg, A., and Starobinski, D. (2003), " Robust Location Detection in Emergency Sensor Networks," IEEE INFOCOM 2003.