Note that the complexity of the system being analyzed is a function of the conceptual system. Remarks in the case of the former hold as well for the latter. Depending on the conceptual apparatus used, many things may become redundant which otherwise may not be so. As in the case of the system being analyzed, it is possible to delete too much or too little. My attempt in this paper has been to develop a conceptual system in which unnecessary redundancy was hopefully eliminated. However, whether coherency between the conceptual system and the system being analyzed was achieved for certain is not known. The issue of congruence in the measurement of system complexity and in information theory in general is an area requiring considerable further research.

ACKNOWLEDGMENT

For his many valuable comments, my warm thanks to Professor Robert W. Rempfer, and the usual absolution.

REFERENCES

[1] F. Attneave, *Applications of Information Theory to Psychology*. Holt, Rinehart and Winston, 1959.
[2] G. Birkhoff, *Hydrodynamics, A Study in Logic, Fact and Similitude*. Princeton, N.J.: Princeton University Press, 2nd ed., 1960.
[3] B. Chandrasekaran and A. K. Jain, "Independence, measurement complexity and classification performance," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-5, no. 2, pp. 240–244, March 1975.
[4] R. C. Conant, "Detecting subsystems of a complex system," *IEEE Transactions on Systems, Man and Cybernetics*, pp. 550–553, Sept. 1972.
[5] C. W. J. Granger and M. Hatanaka, *Spectral Analysis of Economic Time Series*. Princeton University Press, 1964.
[6] J. A. Gukhman, *Introduction to Theory of Similarity*. Academic Press (transl.), 1965.
[7] H. A. Hayek, "The theory of complex phenomena," in *The Critical Approach to Science and Philosophy, Essays in Honour of Karl R. Popper*, M. Bunge, Ed., pp. 332–349, London, 1964.
[8] J. G. Kemeny, "Two measures of complexity," *The Journal of Philosophy*, pp. 722–733, Nov. 1955.
[9] T. C. Koopmans and O. Reiersol, "The identification of structural characteristics," *The Annals of Mathematical Statistics*, vol. 21, pp. 165–181, 1950.
[10] E. Kuh, "The validity of cross-sectionally estimated behaviour equations in time series applications," *Econometrica*, vol. 27, pp. 197–214, 1959.
[11] D. H. Krantz, P. D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement*, vol. I, Academic Press, 1971.
[12] W. J. McGill, "Isomorphism in statistical analysis," pp. 56–62 in H. Quastler [17].
[13] J. R. Meyer and E. Kuh, "Correlations and regression estimates when data are ratios," *Econometrica*, vol. 23, pp. 400–416, 1955.
[14] G. A. Miller, "Note on the bias of information estimates," pp. 95–100 in H. Quastler [17].
[15] J. Pietarinen, "Quantitative tools for evaluating scientific systematizations," J. Hintikka and P. Suppes, Eds., in *Information and Inference*, pp. 123–147, D. Reidel Publishing Co., 1970.
[16] A. Phillips, *Technology and Market Structure*. Cambridge: D. C. Heath, 1971.
[17] H. Quastler, Ed., *Information Theory in Psychology*, Proceedings of a Conference on Information Flow, Free Press, 1955.
[18] C. R. Rao, *Advanced Statistical Methods in Biometric Research*. Wiley, 1952.
[19] D. Sahal, "Generalized Poisson and related models of technological innovation," *Technological Forecasting and Social Change*, vol. 6, no. 4, pp. 403–436, 1974.
[20] ——, "A generalized logistic model for technological forecasting," *Technological Forecasting and Social Change*, vol. 7, no. 1, pp. 81–95, 1975.
[21] ——, "Cross impact analysis and prediction of technological developments: A case study of farm tractors," *IEEE Transactions on Engineering Management*, vol. EM-22, no. 2, pp. 76–79, 1975.
[22] ——, "A reformulation of technological progress function," *Technological Forecasting and Social Change*, vol. 8, no. 1, 1975.
[23] ——, "A spectral analytic study of some postulates concerning engineering design and related processes," *Technological Forecasting and Social Change*, vol. 8, no. 3, 1975.
[24] ——, "A stochastic learning model of machine design changes," Report 74-12, Systems Science, Portland State University, 1974.
[25] ——, "On the conception and measurement of trade-off in the evolution of engineering systems, a case study of aircraft design process," *Technological Forecasting and Social Change*, vol. 8, 1975 (in press).
[26] ——, "On the conception and measurement of technology, a case study of aircraft design process," *Technological Forecasting and Social Change*, vol. 8, 1975 (in press).
[27] ——, "Another measure of technology: The generalized Mahalanobis distance, a discriminant analytic framework for technology assessment," *Technological Forecasting and Social Change*, vol. 9, 1976 (in press).
[28] ——, "Evolving parameter models of technology assessment, *Technology Assessment*, 1976 (in press).
[29] ——, "Technological systems: An exploration in the dynamical aspects of theory of hierarchical structures," *General Systems Yearbook*, vol. xx, 1975.
[30] ——, "From dimensional analytic theory to a stochastic systems theory," submitted for publication.
[31] C. E. Shanon and W. Weaver, *The Mathematical Theory of Communication*. Urbana: The University of Illinois Press, 1964.
[32] H. A. Simon, "The architecture of complexity," *Proceedings of the American Philosophical Society*, vol. 106, no. 6, 1962.
[33] J. Von Neumann, *The Theory of Self-replicating Systems*. Urbana: University of Illinois Press, 1966.
[34] P. Zarembka, "Transformation of variables in econometrics," P. Zarembka, Ed., in *Frontiers in Econometrics*, Academic Press, 1974.

Effect of Intraclass Correlation on Confidence Coefficients of Confidence Sets Based on Chi-Square Statistics

REKHA BASU, J. P. BASU, AND T. O. LEWIS

***Abstract*—This paper investigates how the presence of simple equicorrelation in a multivariate normal sample affects the confidence coefficients of confidence sets based on chi-square statistics and constructed on the assumption of independence of the sample for the mean when the dispersion matrix is known and the scalar σ^2 in $\sigma^2\Sigma$ when Σ is known.**

I. INTRODUCTION

In remote sensing data analysis, as in other areas of statistical data analysis, confidence interval procedures or significance testing procedures are usually derived on the assumption that the observations in the sample are independently and identically distributed normal vectors. In fact, even when the observations are identically distributed normal vectors, they are at most equicorrelated, but rarely independent. Dr. William Coberly [1] analyzed some remote sensing data from Earth Resources Technology Satellite (ERTS-1, renamed Landsat) to determine the validity of the assumption of independence of observation vectors in a sample. He discovered that the observations were significantly correlated. Thus, in many instances, it would be more rational to assume the sample to be equicorrelated or simply equicorrelated, that is, all pairs of observations have the same covariance, rather than to assume the sample to be independent.

Walsh [2] has shown how the presence of intraclass correlation in univariate normal samples affects the confidence coefficients of some confidence intervals (or equivalently the significance level of some tests of significance). Basu, Odell, and Lewis [3] have shown how the presence of simple equicorrelation in multivariate normal samples affects the confidence coefficients of the confidence sets based on a T^2-statistic for the mean of a single population and the difference of means of two populations. In this correspondence, it is shown how the presence of intraclass correlation (simple equicorrelation) in multivariate normal samples affects the confidence coefficients of the confidence sets based on a chi-square statistic for the mean of a population with known dispersion matrix and the scalar σ^2 in $\sigma^2\Sigma$ of the dispersion matrix when Σ is known.

The above procedures are often used for deciding whether a sample has come from a prescribed population. In remote sensing data analysis, photointerpreters label areas as belonging to different crops. The samples of observations coming from those areas are used in training the classifier for automatic

Manuscript received June 13, 1975; revised December 17, 1975.
R. Basu is with the University of Houston, Houston, TX 77004.
J. P. Basu is with the Lockheed Electronics Co., Inc., Houston, TX 77058.
T. O. Lewis is with Texas Tech University, Lubbock, TX 79409.

classification of the data into different crop classes. Following the above procedures, new samples are either combined with the existing training samples or kept separately as training samples for unencountered classes.

II. Basic Concepts

The $p \times 1$ random vectors $X_1, X_2, \cdots, X_n$ are said to be *equicorrelated* [3] if

1) $D(X_i) = E[(X_i - EX_i)(X_i - EX_i)^T] = \Sigma$, a symmetric positive definite matrix, for all i $(1 \leq i \leq n)$;
2) $\text{cov}\,(X_i, X_j) = E[(X_i - EX_i)(X_j - EX_j)^T] = C$, a symmetric semidefinite matrix, for all $i \neq j$.

If $X_1, \cdots, X_n$ are equicorrelated random vectors, the dispersion matrix V of their joint distribution is given by

$$V = \begin{bmatrix} \Sigma & C & \cdots & C \\ C & \Sigma & \cdots & C \\ \vdots & \vdots & & \vdots \\ C & C & \cdots & \Sigma \end{bmatrix} = I_n \otimes (\Sigma - C) + E_n \otimes C, \quad (1)$$

where $A \otimes B$ denotes the Kronecker product [4] of the matrices A and B, I_n is the $n \times n$ identity matrix, and E_n is the $n \times n$ matrix, all of whose elements are 1.

The random vectors $X_1, \cdots, X_n$ are said to be simply *equicorrelated* if they are equicorrelated and

$$C = \text{cov}\,(X_i, X_j) = \rho\Sigma, \quad (2)$$

where ρ is a scalar. When $X_1, \cdots, X_n$ are simply equicorrelated, the dispersion matrix V of their joint distribution is given by

$$V = [(1 - \rho)I_N + \rho E_N] \otimes \Sigma. \quad (3)$$

A sample $X_1, \cdots, X_n$ from a multivariate population is said to have *intraclass correlation* if $X_1, \cdots, X_n$ are equicorrelated or simply equicorrelated. The common covariance matrix C of all pairs of X_i and X_j $(i \neq j)$ is referred to as the *intraclass correlation matrix*. The scalar ρ is referred to as the *coefficient of simple equicorrelation.*

III. Sampling Distribution

Independent Sample

Let $X_1, \cdots, X_n$ be an independent sample from an $N_p(\mu, \Sigma)$ population. They are identically distributed as $N_p(\mu, \Sigma)$, and the dispersion matrix V of their joint distribution is given by $V = I_n \otimes \Sigma$.

Let the $n_p \times 1$ random vectors X and Z be defined as follows:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_{n-1} \\ X_n \end{bmatrix} \qquad Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_{n-1} \\ Z_n \end{bmatrix} = (B \otimes I_p)X, \quad (4)$$

where B is the $n \times n$ orthogonal Helmert matrix given by

$$B = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{2.3}} & \frac{1}{\sqrt{2.3}} & \frac{-2}{\sqrt{2.3}} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & & \cdots & \frac{1}{\sqrt{n(n-1)}} & \frac{-(n-1)}{\sqrt{n(n-1)}} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & & \cdots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \end{bmatrix}. \quad (5)$$

Since $BB^T = I_n$ and

$$\begin{aligned} D(Z) &= (B \otimes I_p)(D(X)(B^T \otimes I_p) \\ &= (B \otimes I_p)(I_n \otimes \Sigma)(B^T \otimes I_p) \\ &= (BI_nB^T) \otimes (I_p\Sigma I_p) = I_n \otimes \Sigma, \end{aligned}$$

it is easy to see that $Z_1, \cdots, Z_n$ are independently distributed as follows:

$$Z_n = (X_1 + \cdots + X_n)/n = \sqrt{n}\,\bar{X}, \quad (6)$$

$$Z_n \sim N_p(\sqrt{n}\,\mu, \Sigma) \quad (7)$$

and

$$Z_i = \sum_{j=1}^{i} \frac{(X_1 + \cdots + X_{i-1})}{\sqrt{i(i+1)}} - (i-1)\frac{X_i}{\sqrt{i(i+1)}} \quad (8)$$

$$Z_i \sim N_p(0, \Sigma), \qquad i = 1, \cdots, n-1, \quad (9)$$

where the symbol $\sim$ means "is distributed as."

Let us define

$$Q_1 = n(\bar{X} - \mu)^T\Sigma^{-1}(\bar{X} - \mu) \quad (10)$$

and

$$\begin{aligned} Q_2 &= \sum_{i=1}^{n} (X_i - \bar{X})^T\Sigma^{-1}(X_i - \bar{X}) \\ &= \text{tr}\,\Sigma^{-1}\left[\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T\right]. \end{aligned} \quad (11)$$

Then it may be noted that Q_1 is the Mahalanobis distance between $\sqrt{n}\,\bar{X}$ and $\sqrt{n}\,\mu$. It is well known [4, p. 54] that if $Y \sim N_p(0, \Sigma)$ and Σ is nonsingular, then

$$Y^T\Sigma^{-1}Y \sim \chi^2(p) \quad (12)$$

and hence,

$$Q_1 \sim \chi^2(p) \quad (13)$$

where $\chi^2(p)$ denotes a chi-square distribution with p degrees of freedom. Since B is orthogonal, then

$$\begin{aligned} Q_2 &= \text{tr}\,\Sigma^{-1}\left\{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T\right\} \\ &= \text{tr}\,\Sigma^{-1}\left\{\sum_{i=1}^{n} X_iX_i^T - n\bar{X}\bar{X}^T\right\} \\ &= \text{tr}\,\Sigma^{-1}\left(\sum_{i=1}^{n-1} Z_iZ_i^T\right) \\ &= \sum_{i=1}^{n-1} Z_i^T\Sigma^{-1}Z_i. \end{aligned}$$

From (9) and the fact that $Z_1, \cdots, Z_{n-1}$ are independent, it follows that

$$Q_2 \sim \chi^2[(n-1)p]. \quad (14)$$

Simply Equicorrelated Sample

Let $X_1,\cdots,X_n$ be a simply equicorrelated sample from an $N_p(\mu,\Sigma)$ population, the dispersion matrix V of their joint distribution function being given by

$$V = [(1 - \rho)I_n + \rho E_n] \otimes \Sigma, \qquad -1/(n-1) < \rho < 1.$$

Then defining X and Z as in (4) and noting that BE_nB^T is a matrix whose (n,n)th element is n and all other elements are 0, it can be shown that

$$\begin{aligned} D(Z) &= E[(Z - EZ)(Z - EZ)^T] \\ &= (B \otimes I_p)E[(X - EX)(X - EX)^T](B^T \otimes I_p) \\ &= (B \otimes I_p)\{[(1 - \rho)I_n + \rho E_n] \otimes \Sigma\}(B^T \otimes I_p) \\ &= B[(1 - \rho)I_n + \rho E_n]B^T \otimes (I_p \Sigma I_p) \\ &= [(1 - \rho)BI_nB^T + \rho BE_nB^T] \otimes \Sigma \\ &= [(1 - \rho)I_n + \rho BE_nB^T] \otimes \Sigma \\ &= \begin{bmatrix} (1-\rho)I_{n-1} & 0 \\ 0 & 1 + (n-1)\rho \end{bmatrix} \otimes \Sigma. \end{aligned} \tag{15}$$

Thus the dispersion matrix of Z is a block diagonal matrix with off-diagonal blocks consisting of $p \times p$ null matrices. Therefore, the random vector Z being normally distributed, its component random vectors $Z_1,\cdots,Z_{n-1}$, Z_n are independently distributed.

From (6), (8), and (15), it follows that

$$Z_n \sim N_p(\sqrt{n}\,\mu, [1 + (n-1)\rho]\Sigma) \tag{16}$$

and

$$Z_i \sim N_p(0,(1 - \rho)\Sigma), \qquad i = 1,\cdots,n-1. \tag{17}$$

It may be of interest to note that if $\rho = -1/(n-1)$, then Z_n reduces to a degenerate random vector for which the probability mass is concentrated at the single point $Z_n = \sqrt{n}\,\mu$. The quadratic form $Q_1 = 0$ with probability one.

When $-1/(n-1) < \rho < 1$, it follows from (12), (16), and (17) that

$$Q_1/\{1 + (n-1)\rho\} \sim \chi^2(p) \tag{18}$$

and

$$Q_2/(1 - \rho) \sim \chi^2[(n-1)p]. \tag{19}$$

IV. Confidence Sets

Confidence sets are constructed on the basis of a sample $X_1,\cdots,X_n$ from a multivariate normal population. H_0 and H_1, respectively, will denote the following two hypotheses:

H_0 the observations are independent ($\rho = 0$),
H_1 the observations are simply equicorrelated and ρ ($\rho \neq 0$ and $-1/(n-1) < \rho < 1$) known.

Confidence Set for Mean, Dispersion Matrix Known

Let the population density be $N_p(\mu,\Sigma)$, the dispersion matrix Σ (nonsingular) being known. Then, when H_0 is true, that is, when the observations are independent, it is known [4, p. 55] that a confidence set with confidence coefficient ε for μ can be given by

$$I_0(\varepsilon) = \{m: Q_1(m,0) \leq \chi_\varepsilon^2(p)\} \tag{20}$$

where

$$Q_1(m,0) = n(\bar{X} - m)^T\Sigma^{-1}(\bar{X} - m)$$

and $\chi_\varepsilon^2(p)$ is the 100ε percent point of $\chi^2(p)$. Thus

$$P(\mu \varepsilon I_0(\varepsilon) \mid H_0) = \varepsilon. \tag{21}$$

TABLE I
Values of α

n \ ρ	p	0	0.1	0.2	0.3	0.4
6	5	0.99	0.925	0.82	0.70	0.59
		0.95	0.81	0.65	0.51	0.43
21	5	0.99	0.59	0.30		
		0.95	0.41			
31	5	0.99	0.43			
		0.95	0.27			

When H_1 is true, then from (18) it follows that a confidence set $I_1(\varepsilon)$ with coefficient ε can be similarly given by

$$I_1(\varepsilon) = \{m: Q_1(m,\rho) \leq \chi_\varepsilon^2(p)\}, \tag{22}$$

where

$$\begin{aligned} Q_1(m,\rho) &= n(\bar{X} - m)^T[\Sigma^{-1}/\{1 + (n-1)\rho\}](\bar{X} - m) \\ &= Q_1(m,0)/\{1 + (n-1)\rho\}. \end{aligned} \tag{23}$$

Evidently, $I_1(\varepsilon)$ can be rewritten as

$$I_1(\varepsilon) = \{m: Q_1(m,0) \leq [1 + (n-1)\rho]\chi_\varepsilon^2(p)\}. \tag{24}$$

Now

$$P(\mu \varepsilon I_1(\varepsilon) \mid H_1) = \varepsilon. \tag{25}$$

If H_1 is true and yet inadvertently $I_0(\varepsilon)$ is used as a confidence set for μ, then the confidence coefficient changes to some α, where

$$P(\mu \varepsilon I_0(\varepsilon) \mid H_1) = \alpha, \tag{26}$$

but

$$\alpha = P(\mu \varepsilon I_1(\alpha) \mid H_1),$$

where

$$I_1(\alpha) = \{m: Q_1(m,0) \leq [1 + (n-1)\rho]\chi_\alpha^2(p)\}. \tag{27}$$

It is evident from (26) and (27) that α satisfies the functional equation

$$\chi_\alpha^2(p) = \chi_\varepsilon^2(p)/\{1 + (n-1)\rho\}. \tag{28}$$

Evidently, if $-1/(n-1) < \rho < 0$, then $\alpha > \varepsilon$. However, if $\rho > 0$, then $\alpha < \varepsilon$. The effect of positive values of ρ on the true confidence coefficient α of the confidence set $I_0(\varepsilon)$ is given in Table I for $\varepsilon = 0.99$ and 0.95. If $\alpha < 0.2$, the entry in the table has been left blank.

Confidence Interval for Dispersion Scalar

Let the population density be $N_p(\mu,\alpha^2\Sigma)$, Σ being the known nonsingular matrix. When H_0 is true, it follows from (14) that a confidence interval for σ^2 with confidence coefficient ε can be given by

$$0 \leq \sigma^2 \leq Q_2/\chi_\varepsilon^2[(n-1)p]. \tag{29}$$

When H_1 is true, then using (19), a confidence interval for σ^2 with confidence coefficient ε can be given by

$$0 \leq \sigma^2 \leq Q_2/\{(1 - \rho)\chi_\varepsilon^2[(n-1)p]\}. \tag{30}$$

If H_1 is true and yet the confidence interval given by (29) is used, the confidence coefficient changes to α, where α can be shown to satisfy the functional equation

$$(1 - \rho)\chi_\alpha^2[(n-1)p] = \chi_\varepsilon^2[(n-1)p]. \tag{31}$$

TABLE II
VALUES OF α

$(n-1)p$ \ ρ	0	0.1	0.2	0.3	0.4	0.5
4	0.99	0.988	0.986	0.983	0.979	0.971
	0.95	0.941	0.930	0.918	0.900	0.872
16	0.99	0.982	0.966	0.941	0.890	0.790
	0.95	0.919	0.870	0.790	0.670	0.490
32	0.99	0.975	0.946	0.867	0.715	0.440
	0.95	0.900	0.810	0.640	0.380	0.170

Evidently, if $\rho < 0$, $\alpha > \varepsilon$, but if $\rho > 0$, $\alpha < \varepsilon$. The effect of positive values of ρ upon the confidence coefficient α is given in Table II for $\varepsilon = 0.99$ and 0.95.

V. Conclusion

This correspondence has demonstrated that the presence of intraclass correlation affects the confidence coefficients of the confidence sets obtained under the assumption of independence for the mean of a normal population with known dispersion matrix and the dispersion scalar σ^2 in $\sigma^2\Sigma$ with Σ known. Also, the following was demonstrated.

a) If the sample is simply equicorrelated with positive coefficient of simple equicorrelation, then the confidence coefficient decreases with sample size.

b) If the sample is simply equicorrelated with negative coefficient of simple equicorrelation $(-1/(n-1) < \rho < 0)$, then the confidence coefficient increases with sample size.

Therefore, to be sure about the accuracy of the inference regarding the confidence coefficient, it is necessary to test the sample for independence or for the type of correlation—positive or negative. In case the sample is found to be positively correlated, it is advocated that test statistics appropriate for simply equicorrelated data be used. When the coefficient of positive simple equicorrelation is known or has been estimated, the test statistics appropriate for simply equicorrelated samples can be easily derived from the corresponding statistics for the independent samples.

References

[1] W. Coberly, personal communications. Memorandum from Mathematical Physics Branch of Mission Planning and Analysis Division, NASA/JSC, Houston, TX, May 1973.
[2] J. E. Walsh, "Concerning the effects of intraclass correlation on certain significance tests," *Annals of Mathematical Statistics*, vol. 18, pp. 88–96, 1947.
[3] J. P. Basu, P. L. Odell, and T. O. Lewis, "The effects of intraclass correlation on certain significance tests when sampling from multivariate normal population," *Communication in Statistics*, vol. 3, pp. 899–908, September 1974.
[4] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.

An Experiment with the Edited Nearest-Neighbor Rule

IVAN TOMEK

***Abstract*—A number of computer simulation experiments with the nearest-neighbor classification rule are described. They include classification by the usual k-NN rule, classification with k-NN on a design set edited once according to Wilson and classification with k-NN on a design set edited unlimited number of times by two methods described in the text. Results of experiments indicate that editing improves performance of the rule. This is not proved rigorously, but a possible approach to a proof is outlined.**

Manuscript received September 12, 1975; revised January 19, 1976.
The author is with the Department of Computer Science, Acadia University, Wolfville, N.S., Canada BOP 1XO.

Introduction

Let D be a set of samples from a d-dimensional Euclidian space whose members are selected as follows. Select $m = 1$ with probability p_1 and $m = 2$ with probability $p_2 = 1 - p_1$. Given m, select $x \in D$ from a population with density $q_m(x)$. In other words, D consists of samples from two classes whose underlying density functions are $q_m(x)$ and *a priori* probabilities p_m $(m = 1,2)$. D will be called the *design set* and its members *labeled prototypes.*

The k-nearest-neighbor (k-NN) rule is a method of classification which works as follows. Given a sample y to be assigned to one of the two classes, find the k nearest neighbors (kNN) of y in D using a chosen definition of distance. Assign y to class m (either 1 or 2) if the majority of its kNN belong to class m; break ties arbitrarily. (In practice, k is usually odd to avoid ties. This will also be assumed for simplicity in the following text unless specified otherwise.)

Much research has been devoted to the k-NN rule (for a list of references see, for example, Duda and Hart [1]). One of the most important results is that k-NN has asymptotically very good performance. Loosely speaking, for a very large design set, the expected probability of incorrect classification (error) P achievable with k-NN is bounded as follows:

$$P^* \leq P < 2P^*. \tag{1}$$

Here P^* is the optimal (minimal) error rate for the given underlying distributions $p_i, q_i(x)$ $(i = 1,2)$ (see [2]). In many situations the rule performs almost as well as the optimal classifier [3]. Furthermore, it has been shown by Wilson [3] that the following simple editing of D improves the performance of k-NN even further:

1) classify each sample $x(i) \in D$ by k-NN using samples $x \in D$, $x \neq x(i)$;
2) form a new design set D' containing exactly those samples from D which have been classified in accordance with their actual membership in Step 1).

k-NN classification with D replaced by D' reduces the expected error rate below that associated with D.

It is natural to ask what would similar editing of D' (leading to D''), D'' etc. do to the design set. Should we expect progressively better and better classification, or will editing distort the design set and result in deteriorating performance?

A satisfactory answer to this question has not been found yet. The reason is, basically, the following difficulty. Wilson has been able to show that if a design set is "very large" and consists of independently chosen samples, then, under quite general conditions, the asymptotic probability that a sample x classified by its k nearest neighbors in D' is assigned to class $C(1)$ is

$$\bar{P}_\infty(C(1)/x) = \frac{p_1 P_\infty(C(1)/x)}{p_1 P_\infty(C(1)/x) + p_2 P_\infty(C(2)/x)}. \tag{2}$$

Here $P_\infty(C(i)/x)$ is the asymptotic probability that sample x is assigned to class $C(i)$ by the k-NN rule with an "infinitely" large design set D.

Unfortunately Wilson's proof depends on the assumption that samples in D are chosen independently. Editing generates a design set in which samples are *not* independent since retention