TABLE I

	$P(c)$	$P(e)$
Random Selection	97.17%	2.83%
Method Described	98.33%	1.67%

and $\eta_r > 0$ then it is guaranteed that the value of the objective function is unaltered, thus proving that (x, y) belongs to Λ_p.

The above analysis has implicitly assumed that a_s and b_r are nonzero. If they are zero, the corresponding equations become simpler and a direct solution can be obtained, namely,

$$\epsilon_r = -\epsilon_s = x_r'[e^{-\Delta I_1/a_r} - 1] \tag{26}$$

$$\eta_s = -\eta_r = y_s'[e^{-\Delta I_2/b_r} - 1] \tag{27}$$

which also satisfy all the required constraints.

Finally the case is analyzed where there is no index s such that $a_s < \gamma \cdot b_s$. In this case it is always possible to find an index s' such that $a_{s'} > \gamma \cdot b_{s'}$ and $x_{s'}' \geq \gamma \cdot y_{s'}'$ in which no further constraint should be imposed upon $\epsilon_{s'}$ and $\eta_{s'}$. These parameters can be substituted for ϵ_s and η_s in the previous analysis, leading to an identical result.

IV. An Example

In order to illustrate how the use of I-divergence may lead to effective results, a simple case is presented, namely the classification of the unconstrained alphanumeric characters 8 and B. The data set available consists of digitized versions of such characters in binary 24×16 arrays format denoted by a vector of the forms $(X_1, X_2, \cdots, X_N)$. The approximation employed is

$$f_i(X) = f_i(X_1, X_2, \cdots, X_N)$$

$$= \sum_{l=1}^{N/n} p_i(X_{j_{(l-1)n+1}}, \cdots, X_{j_{ln}}) \qquad i \in \{1, 2\}$$

where

$(X_{j_1}, X_{j_2}, \cdots, X_{j_N})$ is a permutation of $(X_1, \cdots, X_N)$.

The exact solution of this approximation generation problem can be obtained by solving an associated 0-1 linear program but its size together with the required knowledge of all n-order distributions make it computationally impracticable even for moderate values of N and n, requiring instead the use of heuristical approaches. One such approach consists of two phases. In the first one, the set of binary features is reduced by selecting the best N' ($N' \ll N$ and N' divides n) according to some specified criterion. For the sake of simplicity, the N' best single features [10] were selected. In the next phase, the solution of the original 0-1 linear program is approximated by a sequence of "maximum weight perfect matchings" which were solved by the Edmond's algorithm [11], although it is acknowledged that such a technique requires n to be a power of 2.

With the approximations thus generated and the data-set available (300 samples/character), a series of 30 simulation runs were carried out where the probability of correct classification was assessed by the "leave Q out" method [9]. Table I shows the generated results ($N = 384$, $N' = 32$, $n = 4$, $Q = 50$) together with those obtained when randomly selected approximations of the same kind are employed. Statistical analysis has revealed that the results are significantly different at a confidence level of 5 percent.

V. Conclusions

It has been shown that there is an upper bound for error probability $P^*(e)$ which is a nondecreasing function on the I-divergences between original and approximating distributions. As a direct consequence, the minimization of I-divergences when a structured family of approximations is employed often leads to a lower value of the upper bound of $P^*(e)$. It is acknowledged that once $P^*(e)$ is a piecewise constant function, the reduction of I-divergences can leave the upper bound for $P^*(e)$ unaffected but is never any possibility of increasing it.

Finally, it is important to bear in mind that if the knowledge about the distributions involved is only given by lower order distributions, the use of I-divergences seems to be one of the most effective means of approximation generation.

References

[1] P. M. Lewis, "Approximating probability distributions to reduce storage requirements," *Inform. Contr.*, vol. 2, pp. 214-225, 1959.

[2] D. T. Brown, "A note on approximations to discrete probability distributions," *Inform. Contr.*, vol. 2, pp. 386-392, 1959.

[3] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 3, pp. 462-467, 1968.

[4] D. G. Lainiotis and S. K. Park, "Probability error bounds," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-1, pp. 175-178, Apr. 1971.

[5] J. W. Van Ness, "Dimensionality and classification with independent coordinates," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, pp. 560-564, July 1977.

[6] D. Kazakos and T. Cotsidas, "A decision theory approach to the approximation of discrete probability densities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, no. 1, pp. 61-67, Jan. 1980.

[7] W. S. Meisel, *Computer Oriented Approach to Pattern Recognition (Mathematics in Science and Engineering Series)*, vol. 83. New York: Academic, 1972.

[8] J. M. Mendel and K. S. Fu, *Adaptive, Learning and Pattern Recognition Systems (Mathematics in Science and Engineering Series)*, vol. 66. New York: Academic, 1970.

[9] C. T. Toussaint, "Bibliography on estimation of misclassification," in *Machine Recognition of Patterns*, A. K. Agrawala, Ed. New York: IEEE Press, 1977.

[10] J. D. Elashoff, R. M. Elashoff, and G. E. Goldman, "On the choice of variables in classification problems with dichotomous variables," *Biometrika*, vol. 54, pp. 668-670, Dec. 1967.

[11] N. Christofides, *Graph Theory–An Algorithmic Approach*. New York: Academic, 1975.

Predicting the Required Number of Training Samples

H. M. KALAYEH AND D. A. LANDGREBE

Abstract–**In this paper a criterion which measures the quality of the estimate of the covariance matrix of a multivariate normal distribution is developed. Based on this criterion, the necessary number of training samples is predicted. Experimental results which are used as a guide for determining the number of training samples are included.**

Manuscript received April 26, 1982; revised July 9, 1982. This work was supported in part by NASA under Contract NSG-5414.

H. M. Kalayeh is with Object Recognition Systems, Inc., Princeton, NJ 08540.

D. A. Landgrebe is with the Engineering Experiment Station, Purdue University, West Lafayette, IN 47906.

0162-8828/83/1100-0664$01.00

Index Terms—**Multivariate normal distribution, parameter estimation, training samples, transformed divergence.**

I. Introduction

In practice, the number of training samples for a pattern classifier is frequently limited because it is expensive to collect many training samples. A typical application in which this is the case is the field of remote sensing, and we will use this application to illustrate the technique.

In remote sensing, the reflected and emitted electromagnetic energy of each pixel of a scene in several important wavelength bands is measured by a multispectral remote sensor system mounted on board an aircraft or spacecraft. The output of the sensor system is used to form a point in a q-dimensional space [6]. A commonly used pattern classification algorithm in this application is the maximum likelihood Gaussian scheme. In this instance, the classes are each characterized as a Gaussian distribution in q-space and these distributions in turn are specified by estimates of the means and covariances of each. However, we know that the performance of the estimators is dependent on the number of training samples. In the case of limited training samples, the estimates of the first- and second-order statistics cannot accurately depict all the information which is contained in the data. In particular, the estimate of the covariance matrix may be poor. As a result of this poor estimation, later analysis of the data (for example, classification accuracy and statistical distance measures) will be degraded. See [1] for more details. Therefore, it is important to predict how many samples will be needed in order that the performance of the estimators be statistically reasonable. In the following, a criterion is developed to measure the performance of the estimate of the covariance matrix; then the number of required samples is predicted.

II. Prediction Criterion

Let $X_1, X_2, \cdots, X_N$ be q-dimensional random sample vectors which are drawn from a normally distributed population with parameters $\theta = (M, \Sigma)$, where M is the true mean vector and Σ the true covariance matrix. In practice, M and Σ are not available, so they must be estimated from the observed data. The maximum likelihood estimates of M and Σ are

$$\hat{M} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (X_i - \hat{M})(X_i - \hat{M})^T. \tag{2}$$

For more detail, see [2].

The performance of an estimator is measured by properties, such as whether it provides 1) and unbiased estimate, 2) a consistent estimate, 3) an efficient estimate, and 4) a sufficient estimate. Now, let us study the properties of maximum likelihood estimates of M and Σ. From [2] we have

$$E[\hat{M}] = M \tag{3}$$

$$\operatorname{cov}[\hat{M}] = \frac{1}{N}\Sigma \tag{4}$$

$$E[\hat{\Sigma}] = \frac{N-1}{N}\Sigma. \tag{5}$$

Thus, by definition, $\hat{M}$ is an unbiased estimate of M, but $\hat{\Sigma}$ is not an unbiased estimate of Σ. However, if

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N} (X_i - \hat{M})(X_i - \hat{M})^T, \tag{6}$$

then $E[\hat{\Sigma}] = \Sigma$ which is unbiased. The density function of $\hat{M}$ and $\hat{\Sigma}$ are

$$p(\hat{M}) = \frac{1}{(2\pi)^{\frac{q}{2}}\left|\frac{1}{N}\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}(\hat{M} - M)^T N\Sigma^{-1}(\hat{M} - M)\right\} \tag{7}$$

$$p(\hat{\Sigma}) = \frac{(N-1)^q\,|\hat{\Sigma}|^{(N-q-2)/2} \exp\{-\frac{1}{2}(N-1)\operatorname{tr}\Sigma^{-1}\hat{\Sigma}\}}{2^{(N-1)q/2}\,\pi^{q(q-1)/4}\,|\Sigma|^{(N-1)/2}\prod_{i=1}^{q}\Gamma[\frac{1}{2}(N-i)]}. \tag{8}$$

That is, $\hat{M} \sim N(M, 1/N\Sigma)$, a normal distribution and $\hat{\Sigma} \sim W(\Sigma, N)$, a Wishart distribution. For more details of other properties of these estimators, see [2], [3] and for various properties of the Wishart distribution see [4].

Although the distribution of $\hat{\Sigma}$ is complex, the performance of the estimates of the covariance matrix which are of interest can be measured by the variance of the diagonal components of $\hat{\Sigma}$, as follows

$$\hat{\sigma}_{kk} = \frac{1}{N-1}\sum_{i=1}^{N} (X_{i_k} - \hat{m}_k)^2, \qquad k = 1, 2, \cdots, q. \tag{9}$$

In [3] it is shown that $(N-1)\,\hat{\sigma}_{kk}/\sigma_{kk}$ has a chi-square distribution with $(N-1)$ degree of freedom, and

$$E[\hat{\sigma}_{kk}] = \sigma_{kk} \tag{10}$$

$$E\left[\frac{\hat{\sigma}_{kk}}{\sigma_{kk}}\right] = 1 \tag{11}$$

$$\operatorname{var}[\hat{\sigma}_{kk}] = \frac{2\sigma_{kk}^2}{N-1} \tag{12}$$

$$\operatorname{var}\left[\frac{\hat{\sigma}_{kk}}{\sigma_{kk}}\right] = \frac{2}{N-1}. \tag{13}$$

In a similar manner, and in order to facilitate the evaluation of the covariance matrix one can work in a new space via the following transformation:

$$Y = \Lambda^{-1/2}\Phi^T(X - M)$$

where Φ and Λ are, respectively, the eigenvector matrix and the eigenvalue matrix of Σ.

This transformation leads to:

1) choose the mean M as origin,
2) transform the covariance matrix into the unity matrix.

In effect, we have

$$YY^T = \Lambda^{-1/2}\Phi^T(X - M)(X - M)^T\Phi\Lambda^{-1/2}$$

and

$$\operatorname{cov}(Y) = \Lambda^{-1/2}\Phi^T\Sigma\Phi\Lambda^{-1/2}.$$

Since

$$\Phi^T\Sigma\Phi = \Lambda;$$

therefore,

$$\operatorname{cov}(Y) = I.$$

In practice Φ and Λ are the eigenvector matrix and the eigenvalue of $\hat{\Sigma}$.

Hence,

$$Y = \hat{\Lambda}^{-1/2}\hat{\Phi}^T(X - \hat{M})$$

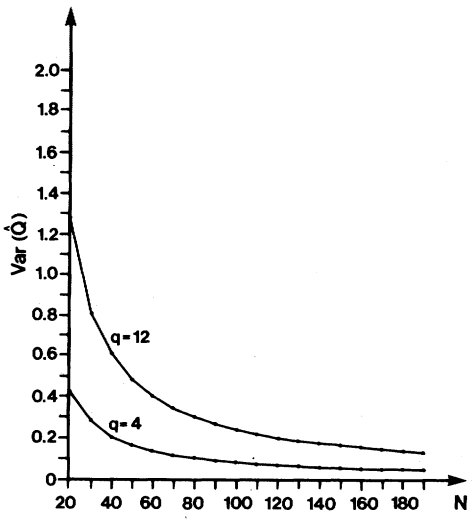

Fig. 1. Variance of $\hat{Q}$ as a function of number of training samples N.

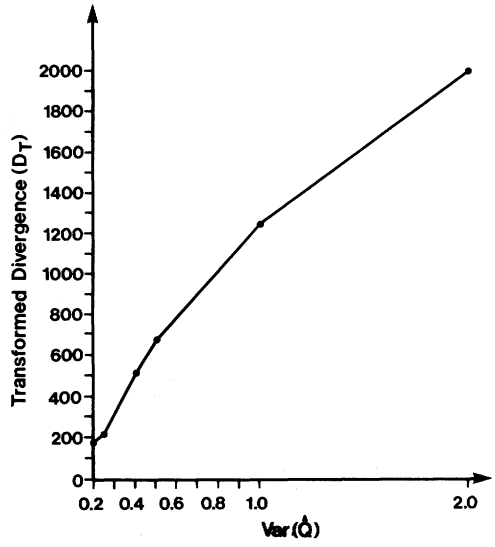

Fig. 2. The average transformed divergence as a function of variance of $\hat{Q}$.

and $\text{cov}(Y) = \hat{\Gamma}$ where the diagonal elements are denoted as $\hat{\gamma}_{kk}$. Because of the orthonormal transformation, the features in the new space are independent; therefore, $(N-1)\hat{\gamma}_{kk}$ has chi-square distribution with $(N-1)$ degrees of freedom. For brevity, let

$$(N-1)\hat{\gamma}_{kk} \sim \chi^2(N-1) \tag{14}$$

and

$$\hat{Q} = [\hat{\gamma}_{11} + \cdots + \hat{\gamma}_{qq}] \tag{15}$$

then

$$(N-1)\hat{Q} \sim \chi^2(q(N-1)) \tag{16}$$

$$E[(N-1)\hat{Q}] = q(N-1) \tag{17}$$

$$E[\hat{Q}] = q \tag{18}$$

$$\text{var}[(N-1)\hat{Q}] = 2q(N-1) \tag{19}$$

$$\text{var}(\hat{Q}) = \frac{2q}{N-1}. \tag{20}$$

A logical choice for our prediction criterion is $\text{var}(\hat{Q})$ because it measures the dispersion of the estimate of the covariance matrix.

To see how to apply the criterion, suppose it is desired that $\text{var}(\hat{Q}) \leqslant \alpha$. Therefore, from (20)

$$N \geqslant 1 + \frac{2q}{\alpha}. \tag{21}$$

Note that the minimum value of N is $q+1$, because if N is less than $q+1$, then the covariance matrix will be singular. So,

$$\text{var}(\hat{Q})_{\max} = \frac{2q}{N_{\min} - 1} = 2. \tag{22}$$

A plot of the $\text{var}(\hat{Q})$ as a function of N with q as a parameter is shown in Fig. 1. Now, if for example $\alpha = 0.2$, then $N \geqslant 1 + 10q$.

The next question to be addressed is how does one choose a reasonable value for α. To answer this question, let us consider the following. As shown in Fig. 1, if $N > 1 + 10q$, then $\text{var}(\hat{Q})$ is decreasing very slowly and its slope $(-\text{var}(\hat{Q})/N)$ is small, less than $-0.02/q$ because from (20), if $N > 1 + mq$, then the slope will be less than $-2/m^2 q$. This suggests that if $N = 1 + 10q$, then the statistical distance between the true probability density and the estimated one may be close to zero because the estimates of the mean vector and covariance matrix are very close to the true ones $(\text{var}(\hat{Q}) = 0.2)$. The transformed divergence [5], [6] is a useful statistical distance measure and is given by

$$D_T = 2000[1 - \exp(-D/8)] \tag{23}$$

where

$$D = \tfrac{1}{2}\,\text{tr}\,(\Sigma - \hat{\Sigma})(\hat{\Sigma}^{-1} - \Sigma^{-1}) + \tfrac{1}{2}\,\text{tr}\,(\Sigma^{-1} + \hat{\Sigma}^{-1})(M - \hat{M})(M - \hat{M})^T. \tag{24}$$

We will use it to experimentally measure the quality of the estimates of the parameters and also a guide to choosing α or N. The following procedure provides a practical means for doing so.

1) Assume that the true probability density of the data is normal with mean vector M and covariance matrix Σ (M and Σ are chosen to be 12×1 and 12×12 matrices, respectively).

2) Based on the true parameters of the distribution, N_i data points are randomly generated.

3) The parameters of the distribution are estimated based on the N_i randomly generated samples and then, using transformed divergence, the statistical distance between the true probability density and the estimated one is computed.

4) Step 3 is repeated five times and the average transformed divergence is calculated.

5) The average transformed divergence for different values of var $(\hat{Q})$ is computed and shown in Fig. 2.

The result in Fig. 2 shows almost a linear relationship between D_T and var $(\hat{Q})$. This implies that when var $(\hat{Q})$ = var $(\hat{Q})_{\max} = 2$, then $D_T = (D_T)_{\max} = 2000$. This indicates that the quality of the estimates of the parameters (mean vector and covariance matrix) is very poor. However, if var $(\hat{Q}) = 0.2$, then $D_T = 175$, which suggests that the estimated probability density is very close to the true one. In practice, however, the true parameters of the distribution are not available and neither is the transformed divergence. As mentioned earlier, a logical choice for our prediction criterion is var $(\hat{Q})$ because it measures the dispersion of the estimate.

We have found that $D_T = 500$, or equivalently, $\alpha = 0.4$ is a logical threshold to decide whether the estimates of the parameters are good or not. This choice implies that the number of training samples should not be less than $1 + 5q$. However, we believe by using information given in Table I, one should be

TABLE I
DISTANCE BETWEEN THE TRUE DISTRIBUTION AND ESTIMATED ONE AS A FUNCTION OF var ($\hat{Q}$) OR NUMBER OF TRAINING SAMPLES

var ($\hat{Q}$)	D_T	D	N
1.00	1250	7.85	$1 + 2q$
0.50	675	3.40	$1 + 4q$
0.40	500	2.30	$1 + 5q$
0.25	210	0.80	$1 + 8q$
0.20	175	0.70	$1 + 10q$

able to establish an upperbound on var ($\hat{Q}$) and consequently estimate the required number of training samples.

III. CONCLUSION

The main purpose of this paper was to develop a criterion to measure the dispersion of the estimate of the covariance matrix of a multivariate normal distribution and, based on this criterion, to be able to predict the necessary number of training samples. To accomplish this, the variance of $\hat{Q} = \text{tr}\,(\hat{I} = \hat{\Lambda}^{-1/2}\hat{\Phi}^T\hat{\Sigma}\hat{\phi}\hat{\Lambda}^{-1/2})$ was chosen as the predictor criterion. It was theoretically shown that variance of $\hat{Q}$ is equal to $2q/N - 1$ with maximum value of 2. Also, the divergence between the true distribution and the estimated one for different values of variance of $\hat{Q}$ was experimentally computed and used to establish an upperbound on the variance of $\hat{Q}$. It was suggested that the required training samples should be about five times the number of features.

REFERENCES

[1] M. A. Muasher and D. A. Landgrebe, "Multistage classification of multispectral earth observational data: The design approach," School Elec. Eng., Tech. Rep. TR-EE 81-41, and Lab. Applications of Remote Sensing, LARS Tech. Rep. 101381, Purdue Univ., West Lafayette, IN, Dec. 1981.
[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972.
[3] J. P. Bickel and A. K. Docksum, *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco, CA: Holden-Day, 1977.
[4] T. W. Anderson, *Introduction to Multivariate Statistical Analysis.* New York: Wiley, 1958.
[5] P. H. Swain and R. C. King, "Two effective feature selection criteria for multispectral remote sensing," Lab. Applications of Remote Sensing, Purdue Univ., West Lafayette, IN, LARS Tech. Rep. 042673, Apr. 1973.
[6] P. H. Swain and S. M. Davis, Eds., *Remote Sensing: The Quantitative Approach.* New York: McGraw-Hill, 1978.

Scale Preserving Smoothing of Polygons

R. L. KASHYAP AND B. J. OOMMEN

Abstract—**A smoother version of a polygon ξ is defined as a polygon which approximates ξ according to a given criterion and which simultaneously has no more edges than ξ itself. In this paper, a scale preserving smoothing algorithm is presented. The input to the algorithm is a polygon ξ and the output is its smoothed version ξ_ϵ. ξ_ϵ, which contains all the scale information that ξ contains, is called the linear minimum perimeter polygon (LMPP) of ξ within a tolerance of ϵ. Using the quantity ϵ the degree of with ξ_ϵ approximates ξ can be controlled. From the LMPP a representation for a polygon approximating ξ can be procured, which is invariant to scale and translation changes. Examples of smoothing maps and characters have been presented.**

Index Terms—**Cartography, minimum perimeter polygon, polygonal representation, scale preserving smoothing, smoothing.**

Manuscript received July 19, 1982; revised December 10, 1982. This work was supported in part by the CIDMAC Project of Purdue University.

R. L. Kashyap is with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.
B. J. Oommen is with the School of Computer Science, Carleton University, Ottawa, Ont., Canada K1S 5B6.

I. INTRODUCTION

Over the past two decades considerable research has gone into the study of the automatic recognition of shape. In this context polygons have played a major role especially since the outer boundary of an object without holes can be approximated as a polygon. The advantages of such representations can be found, for example, in [5, ch. VII]. Consider a typical image processing environment in which the picture of an object to be recognized, silhouetted against a background, is given by a two-dimensional pixel array. Using any boundary tracking algorithm a polygonal representation for the shape of the object can be obtained. It is not uncommon that such a representation involves a polygon having many hundreds of edges [3], [5], [8]. Since the time required for processing the polygon is dependent on the number of edges it possesses, this polygon is usually approximated using a smoothing technique, such as the split and merge technique or the linear scan technique. These techniques and their variants have been well described in [5, pp. 161–184].

Useful as these techniques are, none of these techniques preserves all the scale information contained in the unsmoothed boundary. To clarify this assertion, let ξ and τ be two unsmoothed polygons with τ being a scaled version of ξ, the scaling factor being $k > 0$. Let ξ^* and τ^* be the corresponding smoothed version, the smoothing being performed using any of the algorithms known in the literature. Even though τ is a scaled version of ξ, none of the currently available techniques can guarantee that ξ^* is a scaled version of τ^*.

In this paper we propose a smoothing scheme which can indeed guarantee the preservation of scale information. The input to the scheme is a polygon ξ and its output is ξ_ϵ, the smoothed version of ξ, referred to as linear minimum perimeter polygon (LMPP) of ξ within the tolerance ϵ. The quantity ϵ, $0 \leqslant \epsilon \leqslant 1$, is termed as the tolerance factor. The value $\epsilon = 0$ yields ξ_ϵ identical to ξ and as ϵ increases ξ_ϵ approximates ξ more and more crudely. Within reasonable limits of ϵ, ξ_ϵ indeed preserves the scale information in ξ and yields ϵ as a *single* control parameter by which the smoothing can be controlled.

A natural consequence of this technique is a representation for a smoothed version of a shape, which is invariant to changes in scaling and the translation of the coordinate system in which the *original* shape is drawn.

In the next section we shall describe the minimum perimeter polygon (MPP) of a polygon. We then proceed to define the linear minimum perimeter polygon (LMPP) and demonstrate its properties. We conclude the paper with examples of the use of the LMPP in smoothing maps and characters.

II. THE MINIMUM PERIMETER POLYGON

Let ξ be any polygon specified as an ordered sequence of points in the plane as below.

$$\xi' = \{P_i / i = 1, \cdots, N\}. \tag{1}$$

Let δ_i be a prespecified circular or polygonal constraint domain in the neighborhood of P_i. Let ξ' be any polygon specified by