1

SIMPLE STATISTICS FOR COMPLEX FEATURE SPACES

George Nagy[1] and Xiaoli Zhang[2]

[1] DocLab, Rensselaer Polytechnic Institute, Troy, NY 12180 USA
nagy@ecse.rpi.edu
[2] DocLab, Rensselaer Polytechnic Institute, Troy, NY 12180 USA
zhangxl@rpi.edu

Summary. We study the constraints that govern the distribution of symbolic patterns (letters, numerals and other glyphs used for communication) and natural patterns in high-dimensional feature spaces with a view to gaining insight into the complexity of classification tasks. Pattern vectors from several data sets of printed and hand-printed digits are standardized to identity covariance matrix variables via principal component analysis, shifting to zero mean and scaling. The probability density of the radius of the set of patterns (their distance from the origin) is computed and shown to predict accurately the observed average radius for a wide range of features and dimensionality. We predict further that the class centroids of symbolic patterns will form the vertices of a regular simplex (i.e., a d-dimensional tetrahedron). The observed pairwise distances of the 45 class centroids in ten-class problems are shown to be almost equal to the value predicted from the average radius of the class centroids. The class-conditional distributions of the patterns are compared using two measures of divergence. The difference between the distributions of the same class with different feature sets is found to be larger than the difference between the distributions of different classes with the same feature set. This suggests that the correlation among features of patterns of one class can predict the correlation among features of patterns in another class. The amount of within-source consistency in a data set is quantified using an entropy measure that takes into account small-sample effects. The statistical dependence between the features of same-source patterns of different classes is measured by mutual information applied to the discrete distributions resulting from quantization of the style assignments. If these observations are supported by further studies of symbolic and natural patterns with diverse data sets, they may eventually lead to improved classification methods for same-source ensembles of symbolic patterns.

1.1 INTRODUCTION

Better understanding of the disposition of patterns in feature space may help predict the difficulty and complexity of diverse classification tasks. To further this goal, we compute simple statistics of collections of patterns in several domains of numeral recognition. We focus on metrics that scale well with the number of samples and with the number of features. Unlike the metric properties in [1], these metrics describe only global aspects of the class and style distributions, and neglect fine geometric details of the class boundaries.We compute metrics only on the feature space, as opposed to the data space of bitmaps.

Statistical classification algorithms are often based on the assumption of multimodal mixtures of multivariate Gaussian distributions in feature space. Such distributions can be uniquely specified by their first and second order statistics. For d features (i.e., a d-dimensional feature space), there are $O(d)$ first-order statistics (mean vectors) and $O(d^2)$ second-order statistics (covariance matrices). With values of d in the 10-256 range and tens of thousands of samples per class, complete higher-order statistics cannot be estimated reliably. Furthermore, there is a dearth of parametric multivariate distributions that can be specified in terms of arbitrary frequencies of triples of variables, because there is no convenient structure, analogous to the covariance matrix, for estimating and specifying $O(n^3)$ dependences among n variables. Therefore we confine our attention to metrics that can be expressed in terms of only first and second order population, class, and subclass statistics, i.e., conditional means and covariance matrices.

We apply the proposed metrics to a set of 30,000 printed digits of six fonts, scanned at 300 dpi (Fig. 1.1), and to two sets of hand-printed digits (Fig. 1.2), SD3 (42,698 samples) and SD7 (11,495) samples), from the National Institute of Standards and Technology (NIST). For most of the analysis, SD3 and SD7 were merged to assure stable estimates of the distributions of the entire sample set and of each class. The features are localized, directional, blurred feature vectors [2], with 64 and 100 dimensions respectively. These time-tested features are based on eight chain-coded, directional edge-detectors applied in each of a set of rectangular overlapping zones superimposed on the size-normalized bitmaps of the patterns.

In order to remove effects of the arbitrary means, variances, and statistical correlation of these features, the printed and hand-printed feature data are separately subjected to Principal Components Analysis (PCA). The original features are projected on the eigenvectors, then shifted and scaled, resulting in 64-dimensional and 100-dimensional distributions with zero means and identity covariance matrices. The order of the eigenvalues is retained: in experiments on lower-dimensional feature spaces, we select the PCA features with the largest eigenvalues.

The above preprocessing scheme is illustrated in Fig. 1.3. It allows comparing data with continuous-valued features from different application areas.

Fig. 1.1. Samples from machine printed numeral database (originals printed at 6pt, scanned at 300 dpi)

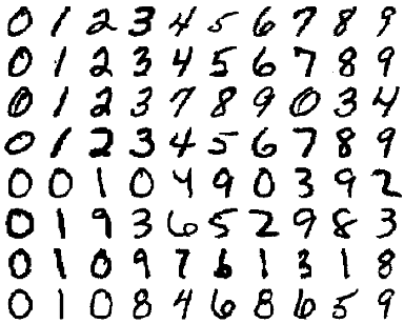


Fig. 1.2. Samples from handwritten numeral database. Each row corresponds to a different writer. The top four writers are from SD3 and the bottom four from SD7

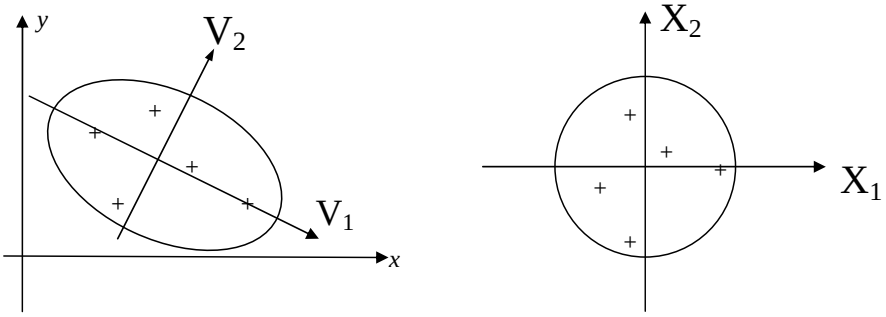


Fig. 1.3. Normalization of feature space. The feature vectors (shown as +) are projected onto the eigenvectors (V_1, V_2) of the overall covariance matrix. Then they are standardized to unit variance, and translated to have mean zero. Hence the resulting feature vectors (X_1, X_2) have mean zero and identity covariance matrix.

The preprocessing does not require labeled data and is applied to the entire data set, regardless of any subsequent separation into training and test data. In the standard configuration, all features have zero mean and unit variance, and they are uncorrelated. Differences between data sets are revealed by the class- and subclass-conditional distributions.

The means and covariance matrices of the class distributions are obtained by straightforward maximum likelihood estimation. The corresponding style labels are the font labels for the printed data. Each font has the same number of samples. The hand-printed digits were clustered into 3 clusters per class by the Matlab K-means routine with Euclidian metric. (Expecting 3 styles in each class is simply an act of faith: in hand printing, stylistic variations

form a continuum. We restricted the number of clusters to 3 to ensure that each cluster has enough samples for stable estimates. Given a fixed number of clusters, it would probably be better not to assign them to classes uniformly.) Here the style labels are the arbitrary cluster labels. The best of ten runs with different random initializations was retained. Table 1.7 in Section 1.5 shows the sizes of the resulting clusters for each class.

In Section 1.2, we examine the surprisingly predictable configuration of the class centroids. In Section 1.3, we observe relative concentrations and volumes of samples, expressed either in terms of the determinants of covariance matrices, or the average distance of patterns from their centroid. Section 1.4 is an attempt to discover systematic departures from the symmetries imposed by the Gaussian assumption. Sections 1.5 and 1.6 posit multiple sources of patterns that give rise to correlations across patterns that we call style. We ignore throughout statistical dependence between the labels of adjacent patterns, known in character and speech recognition as language context.

1.2 AVERAGE RADIUS OF THE PATTERNS

We first show that after the standardization described in Section 1.1, most of the patterns occupy a relatively thin spherical shell centered on the origin. Assume that the individual features are Gaussian. Consider the probability density function (pdf) of the distance from the origin, R_i , of a single sample X_i, with features $x_{i,j}$. R_i can be expressed as the sum of the squares of the d feature values of sample X_i:

$$R_i^2 = \sum_{j=1}^{n} x_{i,j}^2 \tag{1.1}$$

The pdf $f_{R^2}(r^2)$ has mean $\mu_{R^2} = d$ and variance $\sigma_{R^2}^2 = 2d$, because the sum of the squares of d samples from an *i.i.d.*, unit-variance Gaussian distribution is Chi-square with d degrees of freedom. PCA guarantees only uncorrelated rather than independent features, but uncorrelated Gaussian variables are independent. The sum is over the features (dimensions), not the samples.

$$f_R(r) = \frac{r^{d-1}e^{-r^2/2}}{2^{\frac{d-2}{2}}\Gamma(d/2)}, \text{ with mean } \mu_R = \frac{\sqrt{2\pi}1\cdot 3\cdot 5\ldots(d-1)}{(d/2-1)!2^{d/2}}, \text{for } d \text{ even} \tag{1.2}$$

The pdf of R, obtained by a transformation of variable from the χ^2 distribution, is plotted in Fig. 1.4 for several even values of d. For d=8, μ_R=2.74, which is in good agreement with the observed average values of 2.75 and 2.78, respectively, over all the samples of the two data sets. (Lower-case r is the instantiated value of the random variable R. The formula for odd values of d

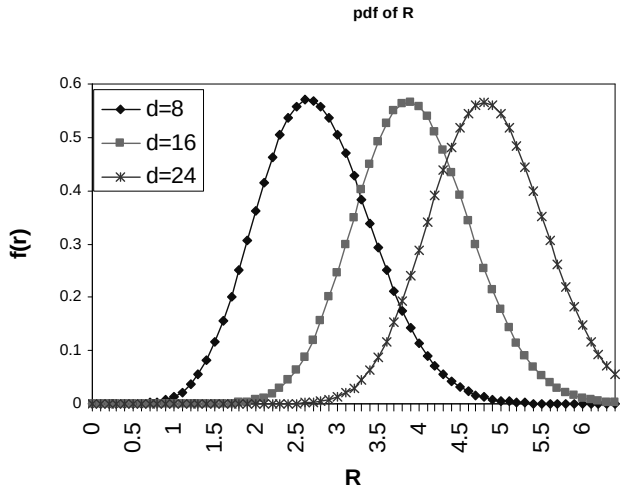


Fig. 1.4. Probability density function of radius of samples. As explained in the text, the pdf is related to the Chi-square density. The average distance of the samples from the origin increases with dimensionality, but the spread of the radii about their average value increases only slowly. The combined effect means that, at high dimensions, most of the samples are located in a thin spherical shell.

Table 1.1. Predicted and observed values of the average radius R of the samples. The observed average values of R agree well with the values predicted by the pdfs of Fig. 1.4.

	Theory		Experiment			
			Machine print		Hand print	
d	μ_R	σ_R	Average sample radius	Standard deviation	Average sample radius	Standard deviation
8	2.74	0.70	2.75	0.65	2.78	0.51
16	3.94	0.70	3.94	0.67	3.94	0.71
24	4.84	0.70	4.84	0.76	4.82	0.88
50	7.03	0.71	6.98	1.13	6.96	1.23

is slightly more complicated because the gamma function is not reduced to a factorial.)

From the expressions for μ_R and μ_{R^2} , we can see that when $d \to \infty$, $\mu_{R^2}/(\mu_R)^2 \to 1$, therefore the thickness-to-radius ratio σ_R/μ_R of the shell containing the samples converges to zero. The Central Limit Theorem also justifies this asymptotic result. However, the Gaussian assumption on the features is necessary for computing the variance of the radius (i.e., the thickness of the shell) for finite dimensions. Table 1.1 shows expected value μ_R, the sample average radius $\langle R \rangle$ of the samples, and the observed standard deviation $\sqrt{\langle R^2 \rangle - \langle R \rangle^2}$ as a function of d for both sets of samples. The unexpected increase in the variance with dimensionality is puzzling. We will next make

use of the predicted and observed regularity of the average radius to predict the configuration of the class centroids.

Before considering the class centroids, we expatiate on the difference between symbolic and natural patterns. Symbolic patterns are interpreted according to some alphabet intended for communicating messages. Examples are printed and hand-printed digits and letters of alphabetic scripts (Roman, Cyrillic, Hangul), shorthand alphabets, glyphs designed specifically for ease of machine reading (OCR fonts and Apple-Newton Graffiti), and the phoneme repertory of various languages. Communication symbols have either evolved, or were engineered, to maintain high separation between classes. We have no reason to believe that natural objects exhibit this property.

Given finite resources for producing each symbol (size and stroke-width limitations for print [3], limited ability to manipulate a stylus for hand print [4, 5, 6], energy budget and a fixed articulatory musculature for phonemes [7]), we would expect the distance between any pair of classes to be approximately the same. (If it weren't, then it would be possible to modify the symbols to further separate the closest pair of classes at the cost of reducing the separation between distant pairs.) "Appropriate" features would maintain this equidistance property. The ten digits in a variety of scripts suggest that in a given alphabet most pairs are, in fact, roughly equally distinguishable (Fig. 1.5). Exceptions may occur for very high frequency symbols, such as 0, 1, and 2 (according to Benson's Law, these digits account for 60% of all the leading digits in numerical fields [8, 9]), *e* in written English, and schwa in many spoken languages. An information-theoretic justification based on maximal entropy would, of course, also have to take into account linguistic context.

Fig. 1.5. Digits of different scripts.

If most of the samples were confined to a thin spherical shell, as argued above, then we would expect that the class centroids will be nearly equidistant from the origin. (The radii R_c of the class centroids are slightly smaller than the average sample radius $\langle R \rangle$ because the spread of the samples is orthogonal to the radii of the class centroids.) Further, if the c class means are equidistant from each other and are at the same radius $\langle R_c \rangle$ from the origin in d-dimensional space, then they must form a c-dimensional regular tetrahedron with edges of length e_d:

$$e_d = \frac{\langle R \rangle \sin\left(\arccos d^{-1}\right)}{\sin\left(\pi - 1/2(\arccos d^{-1})\right)}, \text{where } e_d \underset{d\to\infty}{\longrightarrow} \sqrt{2}\langle R_c \rangle \tag{1.3}$$

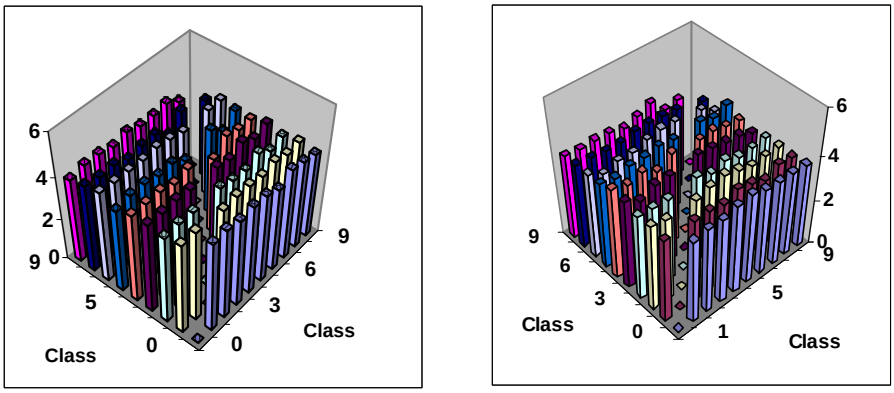


Fig. 1.6. The distance between class centroids is uniform. Machine-printed data on the left, hand-printed on the right. The distance between same-class pairs is zero.

Fig. 1.6 is a plot of the inter-centroid distances in 50-dimensional space. There appears to be little variation between them. Fig. 1.7 shows the ratio of the difference between the largest and smallest interclass distances (i.e., the Euclidian distances between the 45 pairs of class centroids), divided by the median interclass distance. This ratio is a very rigorous measure of uniformity. We see that for d=50, the largest deviation from the median value is less than 20% in both hand-printed and machine-printed data.

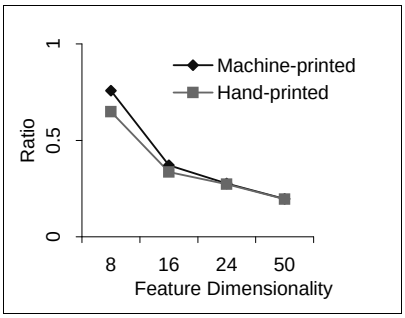


Fig. 1.7. Ratio of range of distances between class centroids to their median distance. The range is the difference between the maximum and the minimum separation of the 45 pairs of class centroids. The ratio of this range over the median interclass distance is plotted as a function of feature-space dimensionality. A low ratio means that the class centroids are located near the vertices of a regular simplex.

The average values of R_c for the ten classes are 2.51 and 2.75, respectively, for the two data sets (d=50). From these values we can predict an inter-class

separation e_d of 3.58 and 3.92, whereas the observed average values are 3.75 and 4.10 respectively. Although the observed pairwise distances are slightly larger than predicted (because the class centroids are not all at exactly the same distance from the origin), it is clear that the pairwise distances are quite uniform in high dimensions. This confirms our tetrahedral assumption. Note that while the convergence of the sample configuration to a thin shell in high dimensions is a universal law (the Central Limit Theorem) given random feature perturbations, the equidistance property of the class means is a consequence of the type of classification problem that we have posed.

It is impossible to place more than (d+1) equidistant points in d-dimensional space. We therefore cannot expect the tetrahedral conjecture to be satisfied for $d < 9$. With increasing d, the class separation grows and its variance among the class-pairs decreases. However, the rate of increase in the separation of the classes tapers off as more features are added, in conformance with the Hughes phenomenon [10].

Table 1.2. Comparison of predicted and observed values of sample radii.

			Machine-printed				Hand-printed			
d	μ_R	σ_R	$\langle R_c \rangle$	$\langle r_c \rangle$	$\sqrt{\langle R_c \rangle^2 + \langle r_c \rangle^2}$	$\langle R \rangle$	$\langle R_c \rangle$	$\langle r_c \rangle$	$\sqrt{\langle R_c \rangle^2 + \langle r_c \rangle^2}$	$\langle R \rangle$
8	2.74	0.70	2.31	1.44	2.73	2.75	1.99	1.90	2.75	2.78
16	3.94	0.70	2.58	2.95	3.92	3.94	2.31	3.16	3.91	3.94
24	4.84	0.70	2.65	4.02	4.81	4.84	2.41	4.15	4.80	4.82
50	7.03	0.71	2.75	6.39	6.96	6.98	2.51	6.49	6.96	6.96

where

d	*dimensionality*
μ_R	theoretical mean radius
σ_R	theoretical stardard deviation of radius
$\langle R_c \rangle$	average centroid radius of each class
$\langle r_c \rangle$	average radius of samples of each class about class centroid
$\langle R \rangle$	average radius of all the samples about the grand centroid

Since the radius of the samples about their class centroid is also the sum of independent variables, each class distribution is also a thin shell, with an average radius of r_c about the class centroid. The radius vector R_c to each class centroid is orthogonal to the subspace spanned by the samples in the remaining dimensions, and therefore obeys the Pythagorean equality $R^2 = r_c^2 + R_c^2$. The pairwise distance e_d is given by (3). Therefore in standardized feature space, any of three single parameters can describe the configuration: (1) the average distance $\langle R_c \rangle$ of the class centroids from the grand (overall) centroid, or (2) the average distance $\langle r_c \rangle$ of the samples from their own class centroid, or (3) average separation $\langle e_c \rangle$ of the class centroids. These parameters depend on the features used. Once we know any one of them, we can compute the average separation of the class centroids and the average overlap of the

class shells. These relations hold up surprisingly well across a broad range of dimensionality for both hand-printed and machine-printed data (Table 1.2). Fig. 1.8 is an attempt to illustrate the putative d-dimensional configuration.

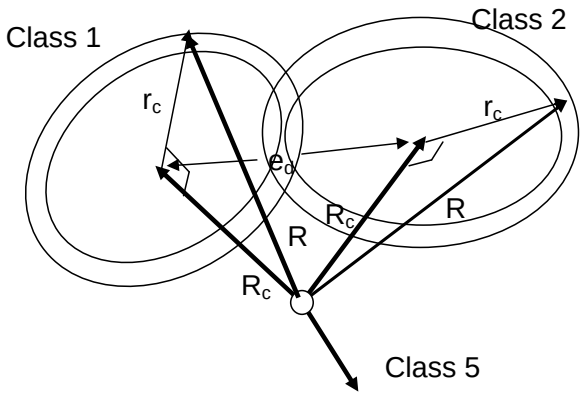


Fig. 1.8. Class configurations in feature space. The radii of the samples from their class centroids are orthogonal to the radii of the class centroids from the origin. The pairwise distance between class centroids can be computed from either.

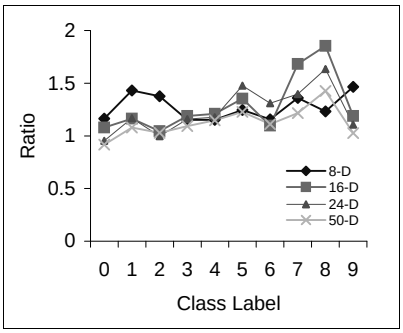


Fig. 1.9a. Ratio of range of distances between cluster centroids to their median distance (machine-printed digits). The range is the difference between the maximum and the minimum separation of the 3 pairs of cluster centroids. The ratios are high relative to those of Figure 1.6 because separation between cluster centroids varies much more than between class centroids.

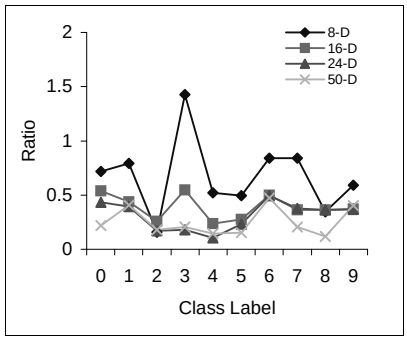


Fig. 1.9b. Ratio of range of distances between cluster centroids to their median distance (hand-printed digits). The separation of the cluster centroids is not even approximately constant as in the machine-printed data.

The disposition of the sub-class centroids about the class centroids is not tetrahedral, as can be observed from Fig. 1.9a and 1.9b. This is not surprising, because for communication purposes there is no real premium in being able to recognize style. Nevertheless, there are secondary problems like font and writer recognition where it is desired to discriminate styles rather than classes. We have found that the principal components that discriminate between classes are also most effective in separating styles [11].

Our observations are summarized in Table 1.3, which shows the parameters that characterize our data sets in low-dimensional and in high-dimensional feature space. Table 1.3 indicates clearly that the separation of the class centroids is higher, and the class distributions are more compact, for machine print than for handprint.

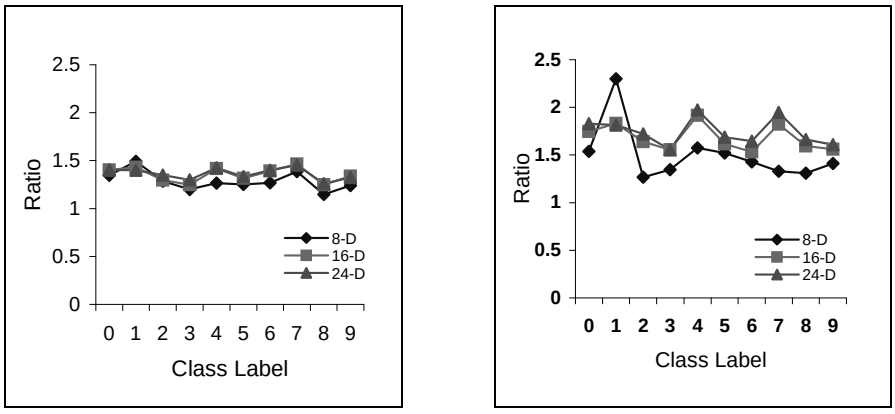
Table 1.3. Parameters for two data sets in 8-d and in 50-d.

Data	d	$\langle R \rangle$	$\langle R_c \rangle$	$\langle r_c \rangle$	$\langle e \rangle$
MP	8	2.7	2.3	1.4	3.5
HP	8	2.7	2.1	1.9	3.0
MP	50	7.0	2.8	6.4	4.7
HP	50	7.0	2.5	6.5	3.7

1.3 CLASS DISTRIBUTIONS

The determinant of the covariance matrix is a measure of the volume of a distribution. If the Gaussian assumption held, then the square root of the determinant would be proportional to the volume of the feature space that holds all the samples within one standard deviation of the mean. For a 1-d spherical distribution, 50% of the samples are within 0.6 standard deviations from the mean. For 8-d, 16-d, and 24-d, the corresponding values are 2.7, 3.9, and 4.8 standard deviations. We have computed the determinants of classes and subclasses in both datasets. We also recorded the average distance $\langle r_c \rangle$ and mean square distance $\langle r_c^2 \rangle$ of the samples from their class centroid.

If the class distributions were spherical like the overall distribution, then the expected squared radius $\mu_{r_c^2}$ of the samples could be computed from $|\Sigma_c|$, the determinant of their covariance matrix, and vice versa. As seen in Section 1.2, the expected squared radius of a zero-mean d-variant spherical distribution with components of variance σ^2 is $\mu_{r_c^2} = d\sigma^2$. Also $\sigma^2 = |\Sigma|^{1/d}$. Therefore $\sqrt{\mu_{r_c^2}} = \sqrt{d}|\Sigma_c|^{1/2d}$. Fig. 1.10 plots $\sqrt{\langle r_c^2 \rangle}/\sqrt{d}\langle|\Sigma_c|\rangle^{1/2d}$ against d for both data sets. (The average is taken over all samples and classes and that of the determinants over the classes.) The ratio is greater than unity, therefore the class distributions must be somewhat flattened. A more detailed



(a) (b)

Fig. 1.10. The Ratio of mean-square radius of the samples (distance from their class centroid) to value predicted under the spherical assumption from the determinant of the class covariance matrix. (a)Machine-printed samples (b) Hand-printed samples .

examination shows that the above ratio is similar for all classes, with less than 5% difference for d=50. The flattening increases with the dimensionality. We may imagine the class distributions as saucers of approximately the same size located at the vertices of a tetrahedron (Fig. 1.8).

Feature space, like the physical Universe, is very sparsely populated. We observed three volumes spanned by all the samples, by the class distributions, and by the sub-class (Class-style) distributions. Each distribution is assumed to have hyper-ellipsoidal equiprobability contours. The volume of the hyper-ellipsoid at a given probability density is proportional to the square root of the determinant of the corresponding covariance matrix. We compared three measures: the volume of all the samples, the sum of the volume of samples from each class and the sum of the volume of samples from each sub-class represented by the square root of the grand covariance matrix, the sum of the square roots of determinants of the class-conditional covariance matrix, and the sums of the square roots of the determinants of the subclass-conditional covariance matrices respectively. The values in Table 1.4 are the ratios of these different measures in 8-dimensional feature space. Ratios of the same order of magnitude are obtained by computing the volumes of hyper-spheres according to the average radii. We conclude that there is a lot of empty space between classes, but much less between sub-classes.

Table 1.4. The ratios of the square roots of determinants of different covariance matrices.

Data	$\sqrt{\lvert\Sigma_g\rvert}/\sum_{c=1}^{C}\sqrt{\lvert\Sigma_c\rvert}$	$\sum_{c=1}^{C}\sqrt{\lvert\Sigma_c\rvert}/\sum_{c=1}^{C}\sum_{k=1}^{K}\sqrt{\lvert\Sigma_{c,k}\rvert}$
MP	8.7	1.4
HP	283	3.5

where	Σ_g	grand covariance matrix
	Σ_c	class-conditional covariance matrix
	$\Sigma_{c,k}$	class-style-conditional covariance matrix

The grand covariance matrix is the sum of the covariance matrices of the class mean vectors and of all the class-conditional covariance matrices. Similarly, for each class, the class-conditional covariance matrix is the sum of the covariance matrix of the subclass mean vectors and of all the subclass-conditional covariance matrices. These relationships are schematically illustrated in Fig. 1.11. This figure seeks to portray the relationship of equiprobability density contours of 3 classes and 12 sub-classes in two dimensions.

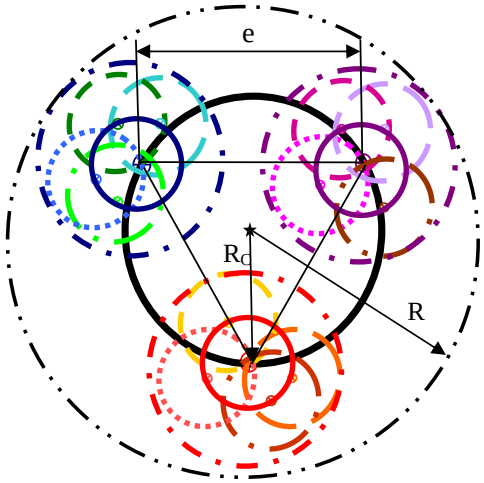


Fig. 1.11. Diagram of class and subclass geometry. The large solid circle passes through the three class centroids, located at the apexes of an equilateral triangle. The small solid circles pass through the centroids of the four subclasses of each class. The dotted and dashed circles represent equiprobability contours corresponding to the overall, class, and sub-class covariance matrices.

Classes are of course most clearly distinguished from one another by their mean vectors. But are their covariance matrices also highly class-dependent? To answer this question, we compared pairs of estimated class-conditional

covariance matrices under the assumption that they specify the feature dependences completely, i.e., that they induce Gaussian feature densities. For these comparisons we used two common similarity measures for probability densities, specialized to Gaussian densities: the Bhattacharyya [12] and the Kullback-Leibler divergence [10].

The ***Bhattacharyya (Bhatt)*** distance between two distributions $p(X)$ and $q(X)$ is defined as

$$D_B(p(X);q(X)) = -\ln \int_{\varphi} [p(X)q(X)]^{\frac{1}{2}} dX \tag{1.4}$$

where φ is the feature space containing all samples.

Between two Gaussian distributions, $p(X) = N(X;\mu_p,\Sigma_p), q(X) = N(X;\mu_q,\Sigma_q)$, it is:

$$DB_N(\mu_p,\Sigma_p;\mu_q,\Sigma_q) = \frac{1}{8}(\mu_p-\mu_q)^T[\frac{\Sigma_p+\Sigma_q}{2}]^{-1}(\mu_p-\mu_q) + \frac{1}{2}\ln\frac{|\frac{\Sigma_p+\Sigma_q}{2}|}{\sqrt{|\Sigma_p||\Sigma_q|}} \tag{1.5}$$

The ***Kullback-Leibler(KL)*** divergence for two distributions $p(X)$ and $q(X)$ is:

$$KL(p(X);q(X)) = \int_{\varphi} p(X)\ln\frac{p(X)}{q(X)}dX \tag{1.6}$$

and for two Gaussian distributions, $p(X) = N(X;\mu_p,\Sigma_p), q(X) = N(X;\mu_q,\Sigma_q)$, it is:

$$KL_N(\mu_p,\Sigma_p;\mu_q,\Sigma_q) = \frac{1}{2}\ln\frac{|\Sigma_q|}{|\Sigma_p|} + \frac{1}{2}Tr(\Sigma_q^{-1}\Sigma_p) + \frac{1}{2}(\mu_p-\mu_q)^T\Sigma_q^{-1}(\mu_p-\mu_q) - \frac{d}{2} \tag{1.7}$$

Table 1.5. Divergence by class and by feature set.

Configration	Feature pairs	Machine-print				Hand-print			
		KL divergence		***Bhatt divergence***		***KL divergence***		***Bhatt divergence***	
		mean	std dev	mean	std dev	mean	std dev	mean	std dev
Different class	**1-8**	8.30	7.72	1.00	0.25	4.60	2.40	0.70	0.22
Same features	**9-16**	3.80	2.38	0.60	0.19	2.30	1.30	0.40	0.13
	17-24	3.20	2.12	0.50	0.15	1.60	0.91	0.30	0.10
	43-50	1.20	0.82	0.20	0.12	0.60	0.39	0.10	0.07
Same class	**1-8 9-16**	35.70	10.99	1.80	0.22	10.50	5.51	0.90	0.26
Different features	**1-8 17-24**	50.20	18.22	2.10	0.30	10.50	4.26	1.00	0.21
	1-8 43-50	54.20	23.10	2.20	0.25	10.80	3.72	0.10	0.20
Different class	**1-8 9-16**	43.50	26.13	1.90	0.45	10.10	6.02	0.90	0.29
Different features	**1-8 17-24**	56.90	35.28	2.10	0.50	11.80	7.82	1.00	0.36
	1-8 43-50	57.50	36.21	2.20	0.55	12.30	7.29	1.00	0.37

Table 1.5 shows the KL and Bhattacharyya distances (computed from the sample means and sample covariance matrices) between class-conditional distributions of selected sets of features for both hand-printed and machine-printed data. Comparing the low values of the distances in the top part of the table to the high values in the two bottom parts, it appears that the covariance matrix depends more on the feature set than on the class. With a given feature set, all of the classes will have similar covariance matrices. Therefore a good estimate of the covariances can be obtained by pooling samples from all classes, while estimating the variances separately for each class. This may explain the relative success of classifiers based on an average covariance matrix and on covariance matrix regularization [13].

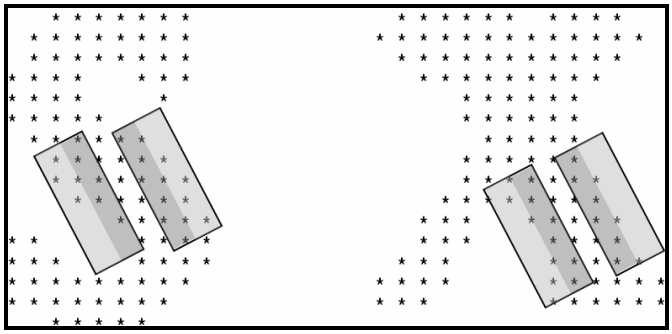


Fig. 1.12. The co-occurrence of two shapes, which leads to positive correlation between the corresponding features, depends more on the nature of the features than of the classes.

Fig. 1.12 suggests why the covariance matrices depend more on the feature set than on the individual classes. Two features are shown, both responses to directional edge detectors. Such complementary gradients are likely to be present or absent at the same time, regardless of the class. Similar arguments can be made for many other types of features.

1.4 DEPARTURES FROM THE GAUSSIAN MODEL

The patterns are not distributed symmetrically about their class means. The usual measure of asymmetry is the *Coefficient of Skew* of the distribution. It is zero for a Gaussian. We project the patterns of each class onto the vector from the grand centroid (Origin) to the class centroid. Fig. 1.13 shows the distribution of these projected patterns in each class. Table 1.6 lists the coefficient of skew of these patterns. The negative values indicate a long tail towards the grand centroid. The machine-printed data also exhibit negative, though smaller, coefficients of skew.

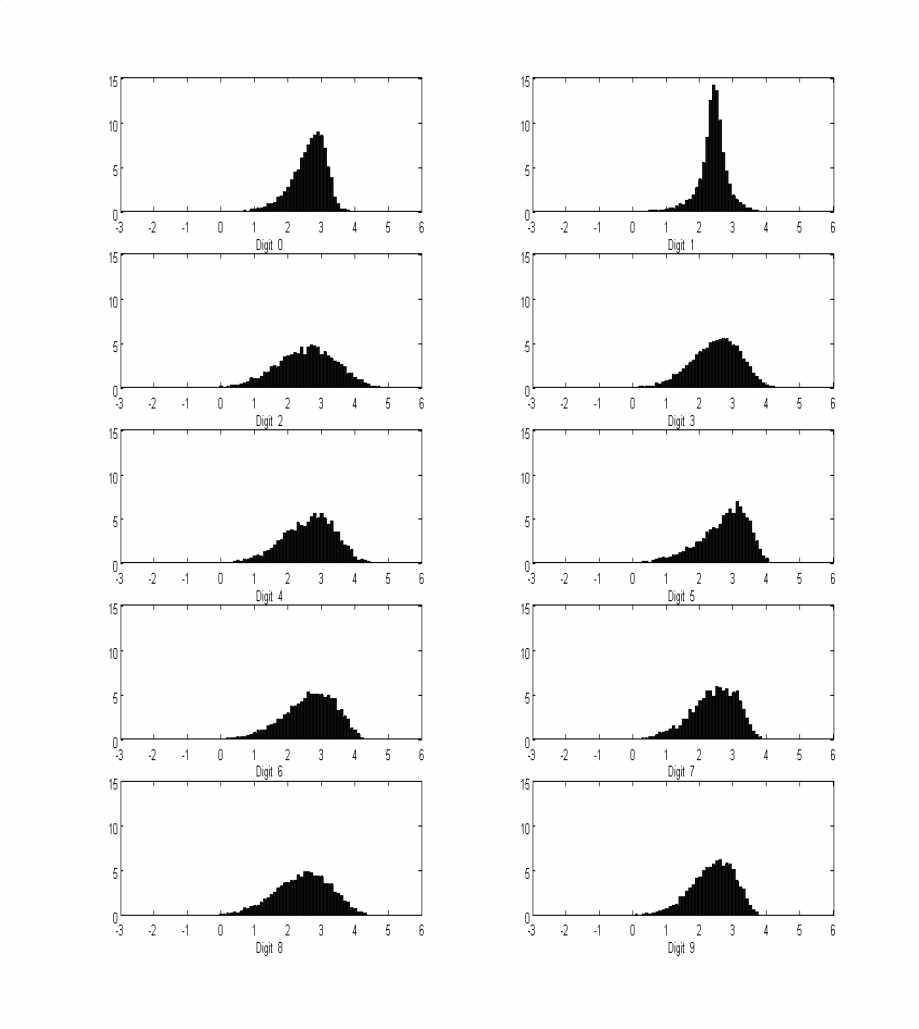


Fig. 1.13. Histogram of the distribution of the patterns of each class projected onto the vector from the grand centroid to the class centroid in 50-dimension feature space. The horizontal axes are labeled in units of standard deviation of the overall standardized sample distribution.

Table 1.6. Coefficients of skew for hand-printed data.

Class	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Features**
Skew	-0.46	-0.91	0.03	-0.83	-0.27	0.18	-0.47	-0.55	-0.76	-0.38	2
	-0.66	-0.67	-0.09	-0.56	-0.43	-0.58	-1.03	-0.80	-0.45	-1.08	8
	-0.78	-0.95	-0.29	-0.46	-0.65	-0.72	-0.59	-0.13	-0.30	-0.80	16
	-0.93	-0.79	-0.36	-0.28	-0.48	-0.68	-0.40	-0.19	-0.29	-0.89	24
	-0.87	-1.08	-0.32	-0.46	-0.45	-0.84	-0.53	-0.55	-0.32	-0.68	50

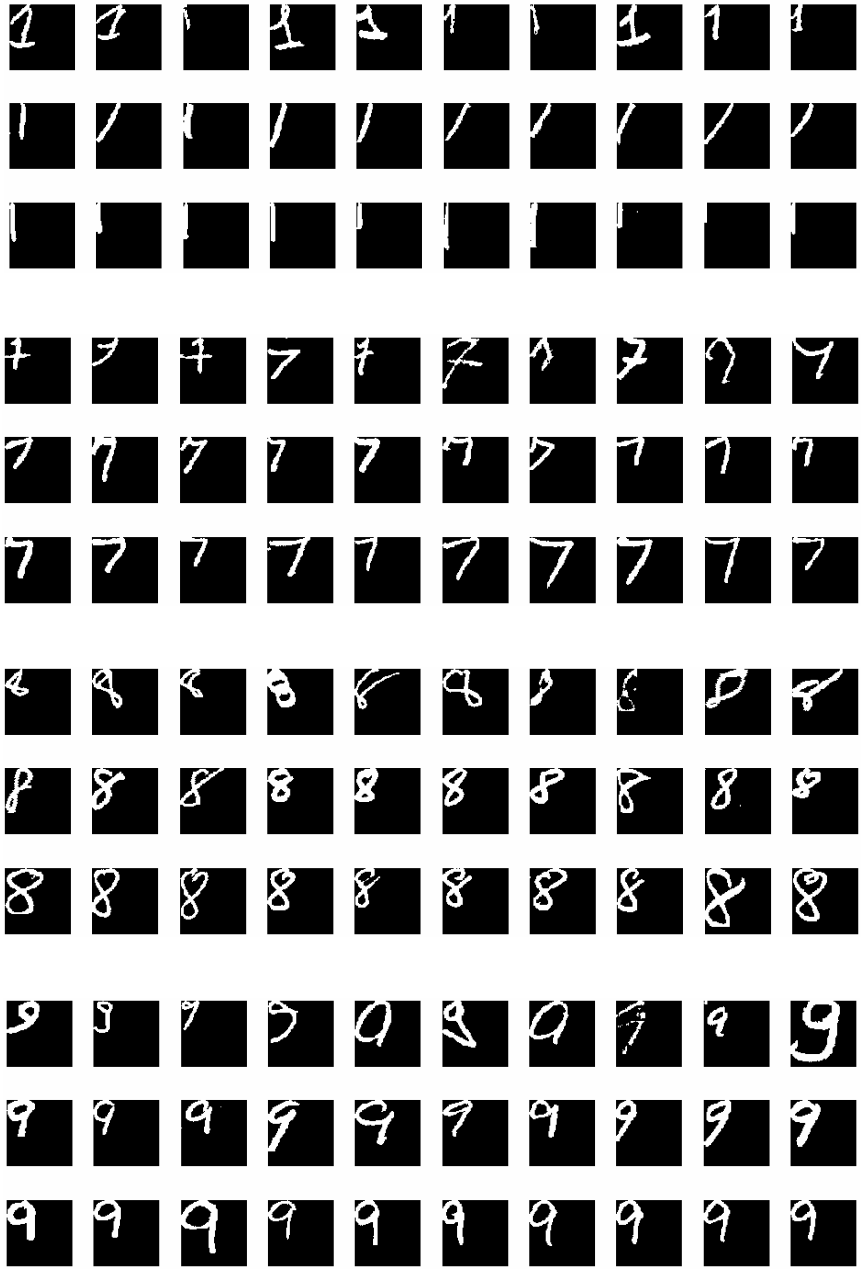


Fig. 1.14. For each class, from top to bottom: patterns nearest the grand centroid, near the class median sample, and farthest from the grand centroid. The patterns on the "outside" appear much more consistent. The patterns near the origin, and therefore near other classes, are more likely to be confused.

Fig. 1.14 shows some samples near the median and at the extreme values of the class distribution projected onto the vector from the origin to the class centroid. The variability of the patterns closer to the grand centroid is greater than that of the patterns far away. We have observed earlier that the error rate of a quadratic discriminant varies by a factor of five depending on which half of the patterns of each class is used for estimating the class-conditional covariance matrices [11]. These observations may eventually also lead to improved methods of regularization for estimating covariance matrices.

1.5 SINGLE CLASS STYLE

We now consider patterns labeled by source as well as by class. Instead of the distribution of all the samples of a class, we observe class- and source-conditional properties. We have already observed in Section 3 that the volume occupied by samples partitioned either by clustering or by font is much less than the volume spanned by all the samples.

In this section we use *entropy* as a more precise measure of source consistency. *Single-class style* is the shape consistency of a single class among the samples from each source. For handwriting, sources usually correspond to writers, so in the rest of this section we will refer specifically to writers rather than to generic sources. Our hand-written data sets contain about 100 isolated digits from each of 500 writers. They were partitioned, as mentioned, into three clusters for each class. How consistent are the writers? Do most of the digits of each writer tend to fall into a single cluster, indicating strong single-class style?

Note that style is not a property of a single writer, but of a whole group of writers. Even if a single writer always wrote a particular digit in the same way, without looking at the other writers we could not know whether it was the only way of writing this digit. If 90% of the writers always cross their sevens, then there is less style than if each writer is completely consistent, but half cross their sevens and half don't. The proposed measure reflects this consideration.

Quantification of single-class style

To quantify single-class style, we first cluster the N feature vectors of M writers of a single digit class c into K clusters. (We do not use a subscript to denote the class, because all of the calculations are performed separately for each class.) We record $N_{m,k}$, the number of samples for each writer and cluster. For writer m with N_m samples of the digit class under consideration, the *Writer-Class-Style Probability Vector* is

$$\mathbf{p}_m = (p_{m,1}, p_{m,2}, \ldots, p_{m,K}) \tag{1.8}$$

where $p_{m,k} = N_{m,k}/N_m$, for $m = 1, 2, \ldots, M$, $k = 1, 2, \ldots, K$.

We can now calculate the writer-class entropy for each writer:

$$H_m = -\sum_{k=1}^{K} p_{m,k} \log p_{m,k} \tag{1.9}$$

If all the samples of a digit class for a specific writer are assigned to the same cluster, then this writer's entropy is zero. That means that this writer has maximal single-class consistency. In contrast, the writer entropy reaches the maximum value of $\log_2 K$ for a writer who has K equally probable variations in writing the same digit. Such a writer does not have a stable single-class style for the observed digit. The *Average Writer-Class Entropy* is defined as:

$$H_{average} = \frac{1}{M} \sum_{m=1}^{M} H_m \tag{1.10}$$

Both the writer-class entropy H_m, and its average value $H_{average}$ over all writers depend on the underlying source probability distribution (i.e., the number of samples from all writers in each cluster), as well as on the amount of style. In order to eliminate the effects of the source distributions, we compute as a normalizing factor the *Class Entropy* H_c:

$$H_c = -\sum_{k=1}^{K} p_k \log_2 p_k \qquad \text{for } \; k = 1, 2, \ldots, K \tag{1.11}$$

where $p_k = N(k)/N$. The *class-style membership* $N(k)$ is the sum of the class-style assignments $N_{m,k}$ over all writers, and N is the number of samples of this digit class from all writers. Since the entropy function is convex, $H_{average}$ is less than or equal to H_c.

For an infinite number of samples per class per writer, the amount of single-class style is indicated by the difference between the average writer-class entropy and the class entropy. A large difference indicates strong single-class style for most writers.

Suppose that we cluster the samples of 100 writers into 3 clusters corresponding to their styles. The *Class-Style Probability Vector* $\boldsymbol{p} = (p_1, p_2, p_3) = (0.5, 0.3, 0.2)$, and the class entropy is

$$H_c = -(0.5 \log_2 0.5 + 0.3 \log_2 0.3 + 0.2 \log_2 0.2) = 1.49 \tag{1.12}$$

If each writer were perfectly consistent, then there would be 50 writers with every sample in Cluster 1, 30 writers with every sample in Cluster 2, and 20 writers with every sample in Cluster 3. The average writer-class entropy would be zero. In contrast, with no style, every writer would have a mixture of samples in the ratio 5:3:2, and $H_{average} = H_c$. We will compare the empirically computed average entropy $H_{average}$ with its maximum possible value, the class entropy H_c, as a measure of single-class consistency.

However, first we must compensate for small-sample effects. With a finite number of samples, even if the writers did not exhibit single-class style, the cluster assignments would not all be exactly proportional to the elements of the class-style probability vector because of sampling fluctuations. These sampling fluctuations decrease the average entropy and may result in a significant difference between the average entropy and the class entropy even in the complete absence of single-class style. To account for the finite sample size (10 samples per class per writer), we compute the *Expected Class Entropy* E[H] under the multinomial sampling distribution $P[n_1, n_2, \ldots, n_K; p_1, p_2, \ldots, p_K]$.

$$P[n_1, n_2, \ldots, n_K; p_1, p_2, \ldots, p_K] = \frac{n!}{\prod_{k=1}^{K} n_k!} \prod_{k=1}^{K} p_k^{n_k} \qquad p_k = \frac{N(k)}{N}, n = \sum n_k = \frac{N}{M} \tag{1.13}$$

where $N(k)$ and N are the cluster and class memberships defined earlier.

We consider all the cases of the partitioning of n samples among K clusters and obtain the class entropy for each case

$$H[n_1, n_2, \ldots, n_K] = -\sum_{k=1}^{K} \frac{n_k}{n} \log_2 \frac{n_k}{n} \tag{1.14}$$

The expected class entropy is then obtained by summing the product of the multinomial probability and the class entropy for every possible cluster assignment vector:

$$\mathrm{E}[H] = \sum_{n_1=0}^{n} \sum_{n_2=0}^{n-n_1} \cdots \sum_{n_K=0}^{n-n_1-\ldots-n_{K-1}} P[n_1, n_2, \ldots, n_K; p_1, p_2, \ldots, p_K] H[n_1, n_2, \ldots, n_K] \tag{1.15}$$

Table 1.7. Expected and Average Entropy for all classes of hand-printed digits. If there were no single-class style, the ratio of the Average Entropy to the Expected Entropy would be near one. (The entropies are normalized by dividing them by $log_2 3$, but this does not affect their ratio.)

	Cluster Membership					$\frac{H_{average}}{\mathrm{E}[H]}$
Class c	N(1)	N(2)	N(3)	E[H]	$H_{average}$	
0	1179	1874	1391	0.89	0.52	0.58
1	2818	1393	625	0.77	0.39	0.51
2	1088	1459	1723	0.89	0.60	0.67
3	1100	1699	1701	0.89	0.58	0.65
4	1512	1822	769	0.85	0.45	0.54
5	912	1488	1312	0.87	0.48	0.55
6	950	1706	1547	0.88	0.47	0.54
7	811	1960	1710	0.85	0.49	0.57
8	1548	1344	1345	0.91	0.50	0.56
9	843	1672	1668	0.86	0.49	0.57

The expected entropy predicts the average entropy when there is no single-class style. We can judge how much single-class style is present in a data set according to the ratio of the average entropy to the expected entropy. If there were no single-class style, we would expect the average entropy to be close to the expected entropy. We have verified that when the writer identities of the samples are shuffled randomly, their ratio is over 0.98 for all classes. In contrast, the actual ratios are between 0.5 and 0.7, as seen from Table 1.7, which lists the average and expected entropy for K=3 and 400 writers in the NIST SD3 data set. We also see from the table that '1' has least entropy while '2' has most. This means that individual writers exhibit more variability in writing the digit '2' than '1'. We summarize this section with a diagram:

Writer statistics Cluster statistics

$$
\begin{array}{ll}
H_m & H_C \\
\Downarrow & \Downarrow \\
H_{average} & E[H] \\
\Downarrow \quad \Downarrow & \\
1-\dfrac{H_{average}}{E[H]} = \text{amount of single-pattern style} &
\end{array}
$$

1.6 MULTI-CLASS STYLE

Multi-class style derives from the correlation between the features of different samples, of the same or different class, from the same source. For now, consider only two patterns at a time, although the concept of multi-class style can be extended to an arbitrary number of patterns. We assume that all the samples have been clustered, as in the previous section. Now observe the cluster assignments of *pairs* of samples from the same writers. If the cluster assignments for samples of class i and class j of the same writer are statistically independent, then we can affirm that there is no multi-class style for class-pair (i, j).

Table 1.8 shows a toy example for the digits 5 and 6 with only ten writers and three clusters. There are exactly 10 samples per class per writer. It is clear that writers who favor Cluster 2 for digit 5 tend to favor Cluster 3 for digit 6, while writers whose 5 usually falls into Cluster 3 tend to write a 6 that often falls into Cluster 1. Finally, fives of Cluster 1 often happen to be associated with sixes of Cluster 2. The numbering of the clusters is completely arbitrary.

A convenient measure of nonlinear statistical dependence is the *Mutual Information* (MI) between two variables X and Y with joint discrete probability distribution $P_{X,Y}(x, y)$. (Consider X the cluster assignment for the digit 5, and Y the cluster assignment for digit 6. For each writer, X and Y can each take on one of three possible values.) The marginal distributions $P_X(x)$ and $P_Y(y)$ can be computed by summation. Then

$$MI_{X,Y} = \sum_{x,y} P_{X,Y}(x, y) \log_2 \frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)} \tag{1.16}$$

We assign to every writer a cluster assignment vector (X, Y), X=1,2, or 3; Y=1,2, or 3, according to where (in which cluster) most of that writer's digits 5 and digits 6 fall. We count the number of each of the nine possible (x, y) combinations. Dividing these sums by the number of writers yields estimates of the probabilities $P_{X,Y}(x, y)$ required to compute the mutual information. Note that here again we have dropped the class subscripts, because the mutual information for every pair of classes is computed independently.

Table 1.8 shows cluster probabilities for each writer and the resulting style assignments. As mentioned, in this example, the cluster assignments are based simply on the dominant cluster of each writer and class, i.e., the cluster with the largest number of samples. Ties are broken randomly. The marginal and joint probabilities of X and Y, and the terms of the MI, are displayed below. The value 0.97 for the MI in this example is the sum of the individual terms. The minimum value of MI, when X and Y are independent, is 0. The maximum value of MI is $min[H_X, H_Y]$, which is $\log_2 3 = 1.6$ when all the cluster assignments are equally probable.

It is possible to quantize the cluster assignments more finely. Even for only pairs of samples, the number of combinations of the values of X and Y grows quadratically. One soon reaches the point where there are not enough samples for accurate estimates of the joint probabilities.

Table 1.9 shows the observed values of the MI for every pair of classes in SD3. We generated 64 possible cluster assignments for each digit-pair by quantizing the cluster probabilities at two levels each, resulting in 8 assignments per writer per class. Many of these assignments never occur in our data. The quantization threshold was the corresponding cluster probability. More than one style can be assigned to a writer if he is inconsistent. We observe that the values on the diagonal, i.e., for same class pairs, deviate from their maximum possible value of $\log_2 8 = 3.0$ because the clusters don't contain the same number of samples. We have verified that when the data is shuffled to eliminate multi-class style, the values of the MI are always less than 0.1.

1.7 CONCLUSIONS

We studied the configuration of symbolic patterns in feature space. Our conjectures are based on relatively broad assumptions about sets of indepen-

Table 1.8. An example of two-pattern style for digits 5 and 6.

Writer specific cluster membership, probability, and assignment																		
	Number of samples falling in each cluster						**Writer-specific cluster probability**						**Writer-specific cluster assignments**					
Class	**5**			**6**			**5**			**6**			**5**			**6**		
Cluster	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Writer 1	1	8	1	2	0	8	0.1	0.8	0.1	0.2	0.0	0.8	0	1	0	0	0	1
Writer 2	7	2	1	1	6	3	0.7	0.2	0.1	0.1	0.6	0.3	1	0	0	0	1	0
Writer 3	0	1	9	8	1	1	0.0	0.1	0.9	0.8	0.1	0.1	0	0	1	1	0	0
Writer 4	2	1	7	7	3	0	0.2	0.1	0.7	0.7	0.3	0.0	0	0	1	1	0	0
Writer 5	1	9	0	1	2	7	0.1	0.9	0.0	0.1	0.2	0.7	0	1	0	0	0	1
Writer 6	6	3	1	0	10	0	0.6	0.3	0.1	0.0	1.0	0.0	1	0	0	0	1	0
Writer 7	3	7	0	1	0	9	0.3	0.7	0.0	0.1	0.0	0.9	0	1	0	0	0	1
Writer 8	8	1	1	2	8	0	0.8	0.1	0.1	0.2	0.8	0.0	1	0	0	0	1	0
Writer 9	1	0	9	0	10	0	0.1	0.0	0.9	0.0	1.0	0.0	0	0	1	0	1	0
Writer 10	10	0	0	9	1	0	1.0	0.0	0.0	0.9	0.1	0.0	1	0	0	1	0	0
	39	32	29	31	41	28	0.39	0.32	0.29	0.31	0.41	0.28	4	3	3	3	4	3

Marginal probability P(x), P(y) and joint probabilities P(x,y)				
Joint probability P(x,y)	**P(y)-Class 6**			**P(x) (Class 5)**
X-Class 5	**1**	**2**	**3**	
1	0.1	0.3	0.0	0.4
2	0.0	0.0	0.3	0.3
3	0.2	0.1	0.0	0.3
P(y) (Class 6)	0.3	0.4	0.3	1.0

Mutual Information MI = 0.97			
P(x,y)logP(x,y)/[P(x)P(y)]	**1**	**2**	**3**
1	-0.02	0.27	0
2	0	0	0.52
3	0.23	-0.03	0

Table 1.9. Observed MI for hand-printed class pairs.

Class / MI	0	1	2	3	4	5	6	7	8	9
0	2.29	0.49	0.26	0.34	0.41	0.26	0.52	0.49	0.34	0.49
1		1.93	0.31	0.44	0.63	0.29	0.52	0.60	0.44	0.69
2			2.31	0.25	0.21	0.13	0.24	0.30	0.18	0.24
3				2.31	0.37	0.17	0.30	0.44	0.28	0.39
4					2.21	0.24	0.42	0.50	0.37	0.61
5						2.32	0.24	0.25	0.23	0.22
6							2.25	0.53	0.34	0.50
7								2.19	0.33	0.65
8									2.37	0.35
9										2.24

dent samples of several classes generated by multiple sources (i.e., *source-conditional independence* between the observable attributes of the patterns). They are supported only by statistics collected on specific features extracted from scanned printed and hand-printed numerals. We now summarize our findings, subject to this caveat.

1. Standardizing the pattern vectors to zero mean, identity covariance variables facilitates comparing data sets, within or across domains, with different patterns, features and dimensionality.

2. In a standardized feature space, the average radius (distance from the origin) of the patterns depends only on the number of features. The mean of the radius can be predicted accurately based on the dimensionality alone. The mean increases faster with dimensionality than the standard deviation. Therefore in any standardized high-dimensional feature space, most of the samples will be contained in a thin spherical shell even if the sample density is highest at the origin. This phenomenon is a direct consequence of well-known statistical facts. The optimal decision boundaries will therefore intersect at the origin of the standardized feature space.

3. When the feature dimensionality exceeds the number of classes, the centroids of *symbolic* patterns are located at the vertices of a regular simplex. The observed distribution of pairwise distances is very peaked and its average value can be predicted accurately from the average radius of the class centroids. This can be explained only by the evolution of symbolic patterns towards maximum discriminability, and may not hold for *natural* patterns. The equidistance of class centroids may serve as an indication of the merit of a feature set, without resorting to actual classification.

4. The pairwise distances between subclass centroids obtained by clustering the feature vectors of each class, or by typeface designations, are *not* approximately equal, even in high dimensions, nor could one expect it. Some pairs of styles are likely to more similar than others if style labels reflect shape.

5. The average distance of the samples from their own class centroids is about 50% higher than predicted from the determinants of their covariance matrices under a spherical (identity covariance matrix) assumption. The class distributions are therefore somewhat "flattened." However, the ratio of the predicted distances to observed distances is fairly uniform across classes, suggesting that the distributions may be similar, except for scale.

6. The divergence between pattern distributions of the same class with different feature sets is significantly larger than between different classes with the same feature set. We expect the class-conditional correlation matrices to be quite similar in any given feature space, because they are determined more by the feature set than by the class. This holds for both *Kullback-Leibler* and *Bhattacharyya* distance, although these two measures of the similarity of two pdfs are not highly correlated. These observations bear on the *regularization* of covariance matrices in small-sample conditions.

7. The class-conditional distributions of symbolic features are asymmetric. In training a classifier, one may safely ignore samples on the "far" side of

their class centroids which exhibit less variation. Support vector machines, of course, do just that.

8. The amount of single-class style in a data set, i.e., within-source consistency, can be quantified by comparing the observed average entropy of style assignments to the expected entropy of an appropriate multinomial distribution. Single-class style can be exploited for classification through adaptation [14, 15, 16].

9. The amount of multi-class style, i.e., the statistical dependence between features extracted from samples of different classes from the same source, can be quantified by the mutual entropy of the style assignments. Multi-class style can be exploited through style-consistent classification [17, 18].

We intend to conduct similar measurements on natural patterns. We hope that a growing collection of such measurements on diverse multi-source collections of samples will provide insight into the intrinsic complexity of classification tasks.

Acknowledgement

We are indebted to Rensselaer graduate student Srinivas Andra's help in some of the calculations, and most grateful for many excellent suggestions by a conscientious and perspicacious referee.

References

1. T. K. Ho and M. Basu, "Complexity measures of supervised classification problems", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 289–300, 2002.
2. C.L. Liu, H. Sako, and H. Fujisawa, "Performance evaluation of pattern classifiers for handwritten character recognition", *International Journal on Document Analysis and Recognition*, vol. 4, no. 3, pp. 191–204, 2002.
3. R. McLean, *Typography*, Thames & Hudson, London, 1980.
4. R. Plamondon, "A kinematic theory of rapid human movements, Part I", *Biological Cybernetics*, vol. 72, no. 4, pp. 295–307, 1995.
5. R. Plamondon, "A kinematic theory of rapid human movements, Part II", *Biological Cybernetics*, vol. 72, no. 4, pp. 309–320, 1995.
6. R. Plamondon, "A kinematic theory of rapid human movements, Part III", *Biological Cybernetics*, vol. 78, pp. 133–145, 1999.
7. J. Greenberg, *Universals of Human Language*, vol. 2, Stanford University Press, 1978.
8. R. Raimi, "The first digit problem", *American Mathematical Monthly*, vol. 83, pp. 521–538, Aug.–Sept. 1976.
9. Mark. J. Nigrini, "I've got your number - CPA use of Benford's law of mathematics in discovering fraud: How a mathematical phenomenon can help CPAs uncover fraud and other irregularities", *Journal of Accountancy*, May 1999.
10. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, New York: John Wiley and Sons, 1973.

11. G. Nagy and S. Veeramachaneni, "A ptolemaic model for OCR", in *Procs. ICDAR-03*, Edinburgh, August 2003, pp. 1060–1064.
12. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.
13. G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Series in Probability and Mathematical Statistics, 1992.
14. G. Nagy and G. L. Shelton, "Self-corrective character recognition system", *IEEE Trans. on Information Theory*, vol. 12, no. 2, pp. 215–222, April 1966.
15. Sriharsha Veeramachaneni and George Nagy, "Adaptive classifiers for multi-source OCR", *International Journal on Document Analysis and Recognition*, vol. 6, no. 3, pp. 154–166, 2004.
16. G. Nagy, "Classifiers that improve with use", in *Procs. Conference on Pattern Recognition and Multimedia*, Tokyo, Februrary 2004, IEICE, pp. 79–86.
17. P. Sarkar and G. Nagy, "Style consistent classification of isogenous patterns", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 88–98, January 2005.
18. S. Veeramachaneni and G. Nagy, "Style context with second-order statistics", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 14–22, January 2005.

Glossary

C	The number of classes
$D_B(\cdot)$	Bhattacharyya distance between two distributions $p(X)$ and $q(X)$
$D_{BN}(\cdot)$	Bhattacharyya distance between two Gaussian distributions
$\mathrm{E}[H]$	Expected class entropy for samples of a class sampled from multinomial distribution
$H_{average}$	Average writer-class entropy
H_c	Class entropy
H_m	Writer-class entropy for writer m
$H[n_1, n_2, \ldots, n_K]$	Class entropy for a cluster assignment vector $[n_1, n_2, \ldots, n_K]$
K	Number of styles
$KL(\cdot)$	Kullback-Leibler divergence for two distributions $p(X)$ and $q(X)$
$KL_N(\cdot)$	Kullback-Leibler divergence for two Gaussian distributions
M	Number of writers
$MI_{X,Y}$	Mutual information between two variables x and y
N	Number of patterns of a class from all writers
$N(X; \mu, \Sigma)$	Normal distribution with mean μ and covariance matrix Σ
N_m	Number of patterns within a class for writer m
$N_{m,k}$	Number of patterns within a class for writer m and class-style k
$N(k)$	Class-style membership for class-style k
P	Class-style probability vector
$P_X(x)$	Marginal probability for variable X
$P_{X,Y}(x, y)$	Joint discrete probability distribution between variable X and Y
R_i	Distance of pattern X_i from the origin
R_c	Radius of centroid of class c
X_i	Instance of the i^{th} random d-dimensional singlet-pattern feature vector
c	Instance of a class label
d	Dimensionality of the singlet-pattern feature space
e_d	Length of edges of c-dimensional regular tetrahedron
k	Instance of a style label
m	Instance of a writer
n	Average number of patterns within a class for each writer
n_k	Number of patterns of class-style k for each writer
$[n_1, n_2, \ldots, n_K]$	Class-cluster assignment vector
$p(\cdot)$	Distribution of samples
p_k	Probability for class-style k
$\mathbf{p}_m$	Writer-class-style probability vector for writer m
$p_{m,k}$	Writer-class-style probability for writer m and class-style k
r_c	Radius of samples of class c from class centroid
r_c^2	Square radius of samples of class c from class centroid
$x_{i,j}$	The j^th feature component of a feature vector X_i
Σ_c	Class covariance matrix
$\Sigma_{c,k}$	Class-style covariance matrix
Σ_g	Grand covariance matrix
φ	The feature space of all samples
μ	Mean
σ	Standard Deviation
$<>$	Average operation
$\lvert \cdot \rvert$	Determinant of a matrix