Taylor & Francis
Taylor & Francis Group

The American Mathematical Monthly

ISSN: 0002-9890 (Print) 1930-0972 (Online) Journal homepage: https://www.tandfonline.com/loi/uamm20

Singular Value Analysis of Cryptograms

Cleve Moler & Donald Morrison

To cite this article: Cleve Moler & Donald Morrison (1983) Singular Value Analysis of Cryptograms, The American Mathematical Monthly, 90:2, 78-87, DOI: 10.1080/00029890.1983.11971162

To link to this article: https://doi.org/10.1080/00029890.1983.11971162

newsletter culminated in the creation of our very successful, very well-received newsletter, *FOCUS*. Its editor, Marcia P. Sward, calls Ed the "Father of *FOCUS*".

Let me conclude by emphasizing other service to the Association. Ed was elected chairman of the Texas Section, but he left for Michigan before he took office. He was a visiting lecturer for the MAA, and served a five-year term as editor of the Mathematical Notes Section of the MONTHLY. As chairman of the Committee on Publications since 1971, he guided the unprecedented growth of our journals and our series of books and monographs with skill, determination, and enthusiasm.

In all his activities, Ed Beckenbach enlisted the cooperation of his colleagues by his skill at negotiation, his unfailing courtesy and consideration toward others, and his common sense and good humor. But Ed's cooperative and accommodating spirit at the committee table completely disappeared in another of his roles. On the tennis court Ed was a hard contender, a tough adversary who showed 'em no mercy. Captain of the tennis team at Rice University back in the twenties, and later the coach of the team, he spanned six decades with his favorite sport. Even recently Ed and his wife Alice competed in national tournaments of "superseniors"; in more peaceful moments they indulged their lifelong hobby, tending their hillside acre of plants ranging from apricots to orchids. Incidentally, Alice surely holds a record for faithful attendance at meetings of mathematicians, having started at the age of eleven as daughter of a one-time president of the Association.

Ed Beckenbach clearly served the mathematical community very well. We are all indebted to him for his preeminent leadership. Ed Beckenbach is indeed a most worthy recipient of the Award for Distinguished Service to Mathematics.

* * *

Edwin F. Beckenbach died on September 5, 1982.

SINGULAR VALUE ANALYSIS OF CRYPTOGRAMS

CLEVE MOLER AND DONALD MORRISON*

Department of Computer Science, University of New Mexico, Albuquerque, NM 87131

1. Singular Value Analysis and Cryptanalysis. The singular value decomposition is a matrix factorization which can produce approximations to large arrays. Cryptanalysis is the task of breaking coded messages. In this paper, we present an unusual merger of the two in which the singular value decomposition may aid the cryptanalyst in discovering vowels and consonants in messages coded in certain variations of simple substitution ciphers.

Texts in many languages, including English, have the property that vowels are frequently followed by consonants, and consonants are frequently followed by vowels. We say a text is a

---

Cleve B. Moler received his Ph.D. at Stanford under George Forsythe. He has taught both mathematics and computer science at Stanford, the University of Michigan, and the University of New Mexico. He recently succeeded Morrison as Chairman of the Department of Computer Science at New Mexico. His academic and research interests include: numerical analysis, mathematical software, and scientific computing.

Don Morrison earned his Ph.D. in mathematics at the University of Wisconsin in 1950, under C. C. MacDuffee. He taught mathematics at Tulane, 1950–55, and computer science at the University of New Mexico, 1967 to the present. He was a staff member, division supervisor and department manager at Sandia Corporation, 1955–71. He has one wife, three children and numerous grandchildren. His present research interests include cryptology, data structures, and design of algorithms.

*This work was partially supported by a grant from the National Science Foundation.

"vowel-follows-consonant" text, or more briefly, a "vfc" text if the proportion of vowels following vowels is less than the proportion of vowels following consonants. That is, if

$$\frac{\text{number of vowel-vowel pairs}}{\text{number of vowels}} < \frac{\text{number of consonant-vowel pairs}}{\text{number of consonants}}. \tag{1}$$

Some languages produce better vfc texts than others and some deviations by individual letters can be expected. In English text, the letter h is often preceded by other consonants to form a single sound, as in ch, gh, ph, sh, and, especially th. The letters l, n, m, and r are often followed by consonants. Nevertheless, English is a predominantly vfc language. In Hawaiian, every consonant is followed by a vowel, so there are no consonant-consonant pairs. In the romaji transliteration of Japanese, the only consonant-consonant pairs are ch, sh, ts, and n, followed by a consonant and several double consonants. Russian has many sounds which, in transliteration, are of the consonant-consonant or vowel-vowel form, such as sh, ch, ts, ya, and ye, but in the Cyrillic alphabet, these are represented as single letters. Thus, Russian is also a predominantly vfc language.

2. The Cryptanalyst's Problem. A simple substitution cryptogram is a coded message in which the individual letters have been replaced by a permutation of themselves. When faced with such a message, a cryptanalyst might first count the occurrences of individual letters in the message and compare the frequencies with the known frequencies in typical uncoded text. However, a fairly long message is required before this technique has much chance of success.

Another aid in breaking the code is a partitioning of the alphabet of the cryptogram into two subsets, representing the vowels and the consonants. In order that such a partition be plausible, it ought to satisfy the vfc rule (1). This is the main task which we wish to consider here—and is what we call the cryptanalyst's problem. A trial and error solution, which tries all possible partitions until one satisfying the vfc rule is found, is clearly prohibitive.

Let n be the number of letters in the alphabet. For any text, the digram frequency matrix is the n-by-n array A with a_{ij} = the number of occurrences of the ith letter followed by the jth letter. Blanks and punctuation, if present, are ignored and the first letter of the text is assumed to follow the last letter. In general, the matrix is not symmetric, but for each i

$$\sum_j a_{ij} = \sum_j a_{ji} = f_i$$

where f_i = the number of occurrences of the ith letter.

For any proposed partitioning of the alphabet into vowels and consonants, two column vectors, v and c, can be defined by

$v_i = 1$ if the ith letter is a vowel, 0 otherwise,

$c_i = 1$ if the ith letter is a consonant, 0 otherwise.

Note that $v + c$ is a vector with all 1's and that the inner product v^Tc is zero. Also note that the value of the quadratic form v^TAv is the number of vowel-vowel pairs in the text.

Using A, v and c, the vfc rule (1) can be stated

$$\frac{v^TAv}{v^TA(v+c)} < \frac{c^TAv}{c^TA(v+c)}.$$

Cross-multiplying and cancelling the common term, we obtain

$$(v^TAv)(c^TAc) - (v^TAc)(c^TAv) < 0. \tag{2}$$

The cryptanalyst's problem, then, is: Given A, find a partitioning v and c so that (2) holds.

3. The Singular Value Decomposition. The singular value decomposition, or SVD, is a matrix factorization which numerical analysts use in a wide variety of ways. Although its primary uses are

in the analysis of systems of simultaneous linear equations and in the computation of pseudoinverses, we will use it here to obtain "simple" approximations to the digram frequency matrix. For this purpose, we express the SVD as a sum of rank one matrices of the form

$$A = \sigma_1 x_1 y_1^T + \sigma_2 x_2 y_2^T + \cdots + \sigma_n x_n y_n^T \tag{3}$$

where

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$$

$$x_i^T x_j = \delta_{ij} \text{ (the Kronecker delta)}$$

$$y_i^T y_j = \delta_{ij}.$$

The coefficients σ_j are known as the *singular values* and x_j, and y_j are the *left* and *right* singular vectors, respectively. They can also be characterized in terms of the solutions to the symmetric eigenvalue problem:

$$\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

It is not hard to see that the *rank* of a matrix is the number of nonzero singular values. (In fact, this is a particularly useful way to define rank.)

There is a fast, reliable algorithm for computing the SVD [3], [6]. For a 26-by-26 matrix, the computation of the singular values and corresponding left and right vectors takes only a few seconds on a medium speed modern computer.

The normalization of the vectors x_j and y_j insures that all of the rank one matrices $x_j y_j^T$ have the same norm (that is, the sum of the squares of the elements). Consequently, the numerical importance of each term in the sum (3) can be measured by the size of the coefficient σ_j. If the σ_j decrease fairly rapidly as j increases, then an accurate approximation to A can be obtained by truncating the series after only a few terms. Truncating the series after k nonzero terms provides a rank k approximation to A. Such approximations are related to those obtained by factor analysis and have a wide variety of applications. One unusual application is to digital image processing [1].

4. Rank One Approximation. A rough approximation to a digram frequency matrix can be obtained by taking only the first term in its SVD:

$$A \approx A_1 = \sigma_1 x_1 y_1^T.$$

Let e be the vector of all 1's and let f be the vector with f_i = the number of occurrences of the ith letter in the text. Then $Ae = A^T e = f$ and so

$$\sigma_1 (y_1^T e) x_1 \approx \sigma_1 (x_1^T e) y_1 \approx f.$$

Consequently, if we were to assume that the digram frequency matrix were only rank one, we would conclude that it was symmetric and that each row and column was proportional to the frequency vector f.

Of course, in practice, A is not rank one. Nevertheless, the first left and right singular vectors tend to be approximately equal and reflect the frequencies of the letters in the text.

5. Rank Two Approximation. The rank two approximation obtained from the first two terms of the SVD is the simplest approximation which takes into account the correlation between pairs of letters in the text. The second left and right singular vectors contain the key to the solution of the cryptanalyst's problem.

Assume that A is approximated by a matrix of rank two, so that

$$A \approx A_2 = \sigma_1 x_1 y_1^T + \sigma_2 x_2 y_2^T. \tag{4}$$

We propose to use the signs of the components of x_2 and y_2 to partition the alphabet as follows:

$$
\begin{aligned}
v_i &= \begin{cases} 1 & \text{if } x_{i2} > 0 \text{ and } y_{i2} < 0 \\ 0 & \text{otherwise,} \end{cases} \\
c_i &= \begin{cases} 1 & \text{if } x_{i2} < 0 \text{ and } y_{i2} > 0 \\ 0 & \text{otherwise,} \end{cases} \\
n_i &= \begin{cases} 1 & \text{if sign } (x_{i2}) = \text{sign } (y_{i2}) \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{5}
$$

The third category—the "neuter" letters—are the ones that cannot be classified as either vowels or consonants. It turns out in practice that few letters fall into this category. In English text, for example, the letter h is usually neuter. It might be possible to consider a finer partition involving "left vowel, right consonant" and so on, but we have not pursued this idea.

Using these definitions, we obtain a solution to the cryptanalyst's problem as follows.

THEOREM. *Let $A = A_2$ be a nonnegative rank 2 matrix with the* SVD *expansion in* (4). *Let v and c be defined by* (5). *Then the* vfc *rule,*

$$D = (v^T A v)(c^T A c) - (v^T A c)(c^T A v) < 0 \tag{6}$$

is satisfied.

Proof. Since A is a nonnegative matrix, it follows from the Perron-Frobenius theorem [4] that x_1 and y_1 have nonnegative components. Placing (4) in (2) and expanding produces eight terms. The two terms involving only subscript 1 cancel, so do the two terms involving only subscript 2:

$$
\begin{aligned}
D = \sigma_1\sigma_2\big(&v^T z_1 y_1^T v\; c^T z_2 y_2^T c + v^T z_2 y_2^T v\; c_1^T z\; y_1^T c \\
&- v^T z_1 y_1^T c\; c^T z_2 y_2^T v - v^T z_2 y_2^T c\; c^T z_1 y_1^T v\big).
\end{aligned}
$$

Of all the different inner products appearing in this expression, only two—namely, $y_2^T v$ and $c^T z_2$—are negative. Consequently, all four terms in the parentheses are negative and D is negative.

6. Effects of Encipherment. When a message M is encoded by a simple substitution cipher c, (where c is a permutation of the integers 1 to n) each occurrence of the ith letter u_i is replaced by the $c(i)$th letter, $u_{c(i)}$. The resulting cryptogram is called M_c. If A is the digram frequency matrix of M, and C is the permutation matrix $C = (\delta_{c(i),j})$ which has in its ith row, the $c(i)$th row of the identity matrix, then it is not difficult to show that the digram frequency matrix of M_c is CAC^T. This matrix has in row $c(i)$, column $c(j)$, the frequency of the digram $u_i u_j$ in M, which it represents. It follows that the singular values of CAC^T are the same as those of A, and the jth left and right singular vectors are, respectively, Cx_j and Cy_j. These have the same coefficients as x_j and y_j, but they are permuted by C; the coefficients appearing in row i in x_j and y_j appear in row $c(i)$ in Cx_j and Cy_j. If the scheme described above, applied to A, classifies the letter u_i as a vowel in M, then applied to CAC^T, it will classify $u_{c(i)}$, the encoding of u_i, as a vowel in M_c.

Simple substitution ciphers are, of course, very simple ciphers, and no self-respecting cryptanalyst regards them as a challenge. A more sophisticated cipher is the k-alphabetic cipher (k is a positive integer). In this cipher, k permutations, $c_1, c_2, \ldots, c_k$ are used in cyclic fashion to encode the letters in M to produce the cryptogram M_c. A letter is said to be in position p, in M ($1 \le p < k$), if it is the mth letter of M and m is congruent p modulo k. If the letter u_i occurs in position p, it is encoded as $u_{c_p(i)}$. Thus, the encoding of a letter depends, not only on what letter it is, but also on its position in M.

The digram frequency matrix of a cryptogram encoded by a k-alphabet cipher is of little use in decoding it, since the various occurrences of a digram in M_c do not represent the same digram in M, if they do not occur in the same position. Cryptanalysts have some clever ways to deduce the probable value of the cycle length k. See, for example, Gaines [2] or Sinkov [5]. Armed with this information, a cryptanalyst can calculate k digram frequency matrices, $A_p = (a_{ijp})$ where a_{ijp} is

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	0	1	2	5	0	1	4	0	2	0	1	9	0	15	0	1	0	10	5	36	1	8	0	0	1	0
B	1	0	0	0	5	0	0	0	1	0	0	1	0	0	1	0	0	2	0	0	2	0	0	0	1	0
C	12	0	0	0	4	0	0	2	1	0	0	0	0	0	7	0	0	4	0	1	0	0	0	0	0	0
D	2	0	1	6	14	0	0	3	13	0	0	0	0	0	4	1	0	0	4	4	1	1	4	0	0	0
E	16	3	8	26	3	5	2	7	6	0	0	4	5	10	5	4	1	22	9	12	2	4	8	0	3	0
F	3	0	0	1	0	1	0	0	5	0	0	0	0	0	10	0	0	3	0	3	1	0	0	0	0	0
G	5	1	0	0	5	0	1	4	0	0	0	1	0	0	3	1	0	6	0	0	0	0	1	0	0	0
H	24	0	0	0	32	2	0	0	7	0	0	1	0	0	8	0	0	0	0	5	1	0	0	0	0	0
I	0	1	8	1	3	0	2	0	0	0	0	2	0	16	9	0	0	2	9	8	0	7	0	0	0	0
J	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
L	3	0	0	4	6	0	0	1	6	0	0	8	2	3	3	0	0	1	0	1	0	1	1	0	2	0
M	2	1	0	0	7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
N	10	0	5	9	4	1	9	0	2	0	0	3	2	4	12	0	0	0	4	8	1	1	0	0	2	0
O	1	3	1	3	0	6	1	1	0	0	0	1	4	20	2	5	0	17	3	13	7	2	3	0	0	0
P	0	0	0	0	5	0	0	0	0	0	0	4	0	0	4	0	0	2	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
R	8	0	0	1	26	4	3	0	1	0	1	3	0	1	6	2	0	0	5	12	3	0	3	0	0	0
S	4	2	2	0	10	2	1	6	1	0	1	0	0	1	4	0	0	1	0	8	1	0	0	0	0	0
T	4	1	4	1	11	5	1	47	18	0	0	3	0	2	11	1	0	2	0	9	0	0	5	0	1	0
U	1	0	0	0	0	0	3	0	0	0	0	2	0	3	0	0	0	5	5	2	0	0	0	0	0	0
V	2	0	0	0	17	0	0	0	3	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
W	2	1	0	0	11	0	0	8	1	0	0	0	0	1	2	0	0	0	0	1	0	0	1	0	0	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	2	0	0	1	1	0	1	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	1	0	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 1. The digram frequency matrix for Lincoln's Gettysburg address.

the frequency with which the digram $u_i u_j$ occurs in positions p, $p+1$ modulo k. All of these occurrences are encoded into the same digram $u_{c_p(i)}u_{c_{p+1}(i)}$ in the cryptogram M_c. By an argument similar to that given above for the simple substitution, if A_p is the digram frequency matrix for digrams occurring in position p, $p+1$ in M, then $C_p A_p C_{p+1}^T$ is the corresponding digram frequency matrix for the cryptogram M_c. It has the same coefficients as A_p, but its rows are permuted by C_p, and its columns by C_{p+1}. Its singular values are the same as those of A_p, and its jth left and right singular vectors are, respectively, $C_p x_{pj}$ and $C_{p+1} y_{pj}$, where x_{pj} and y_{pj} are those of A_p. If the classification scheme described above, applied to x_{pj} and $y_{(p-1)j}$ classifies u_i as a vowel, then it will also, applied to $C_p x_{pj}$ and $C_p y_{(p+1)j}$, classify its encoding $u_{c_p(i)}$ as a vowel.

If this is done for each p, then the cryptanalyst has, for each p, a classification of the letters in position p of the cryptogram, into vowels and consonants. The classification is, of course, only as reliable as it would have been if it had been applied to the original message M, or, more particularly, to those subsequences of M consisting of the digrams occurring in a particular position p, $p+1$. If these are representative samples of the digrams occurring in the language of M, then the classification is as reliable as it is when applied to equally representative plaintexts, as we have done in sections 7 and 8.

1	81.7256
2	53.5189
3	45.1604
4	31.1073
5	19.9258
6	18.8196
7	16.8777
8	12.1174
9	10.1647
10	9.4667
11	7.6957
12	6.2491
13	4.7882
14	2.7120
15	2.1048
16	1.7556
17	1.6395
18	0.9360
19	0.8129
20	0.4996
21	0.3232
22	0.2069
23	0.0148
24	0.0000
25	0.0
26	0.0

FIG. 2. The singular values.

7. Experimental Results. Fig. 1 is the digram frequency matrix for Lincoln's Gettysburg Address, a text of 1,148 characters. Notice, for example, that "th," with 47 occurrences, is the most frequent pair, that "q" occurs only once, and that "j," "x" and "z" do not occur at all.

Fig. 2 gives the singular values of the matrix in Fig. 1. Since three letters are missing, the matrix has rank 23 at most, and 3 of the singular values should be zero. The subroutine finds two exact zero values and one value, the size of roundoff error on the computer. (Exact zeros are printed as 0.0, while numbers less than 10^{-5} are printed as 0.00000.)

It certainly cannot be claimed that the singular values decrease rapidly. In fact, the rank two approximation only vaguely resembles the original matrix. Nevertheless, useful information can be obtained from the first two pairs of singular vectors.

Fig. 3 shows the first singular vectors x_1 and y_1, together with the frequency vector f. It can be seen that the components of the two singular vectors are roughly equal, and are roughly proportional to the components of the frequency vector. Thus, even though the matrix is not particularly well approximated by the first term of its SVD, the singular vectors still retain the properties predicted by the rank one theory.

A	0.3275	0.3221	102.
B	0.0470	0.0441	14.
C	0.1200	0.1135	31.
D	0.2011	0.2259	58.
E	0.4394	0.4517	165.
F	0.0875	0.1065	27.
G	0.0966	0.0799	28.
H	0.3481	0.3378	80.
I	0.1830	0.2339	68.
J	0.0000	0.0000	0.
K	0.0099	0.0097	3.
L	0.1203	0.1218	42.
M	0.0607	0.0468	13.
N	0.2165	0.2435	77.
O	0.2387	0.2563	93.
P	0.0522	0.0564	15.
Q	0.0007	0.0054	1.
R	0.2954	0.2493	79.
S	0.1683	0.1391	44.
T	0.4453	0.4367	126.
U	0.0532	0.0597	21.
V	0.1167	0.0817	24.
W	0.1210	0.1044	28.
X	0.0	0.0	0.
Y	0.0339	0.0344	10.
Z	0.0	0.0	0.

FIG. 3. The first right and left singular vectors and the letter frequency vector.

A	-0.5097	0.1574
B	0.0412	-0.0386
C	0.0719	-0.0788
D	0.1436	-0.2163
E	-0.3306	0.5305
F	-0.0006	-0.0198
G	0.0521	-0.0623
H	0.3877	0.3293
I	-0.2070	0.1791
J	0.0000	-0.0000
K	0.0033	-0.0049
L	0.0298	-0.0853
M	0.0634	-0.0613
N	-0.0649	-0.3983
O	-0.3891	0.1070
P	0.0385	-0.0535
Q	-0.0010	-0.0062
R	0.1692	-0.3402
S	0.0801	-0.1462
T	0.3785	-0.3878
U	-0.0860	-0.0516
V	0.1884	-0.1405
W	0.1559	-0.0112
X	0.0	0.0
Y	-0.0045	-0.0215
Z	0.0	0.0

FIG. 4. The second right and left singular vectors. The sign patterns identify vowels and consonants.

Fig. 4 shows the second singular vectors x_2 and y_2. The alternating sign patterns predicted by the rank two theory are clearly evident. Figs. 5 and 6 give a graphical summary of the quantitative information in Fig. 4 by plotting each letter at a point in the two-dimensional plane determined by its components in x_2 and y_2. Thus, "a" is plotted at coordinates $(-0.5097, 0.1574)$, "b" at coordinates $(0.0412, -0.0386)$, and so on. The more frequent letters are in Fig. 5 and the less frequent letters in Fig. 6. The box in both figures has corners at $(\pm 0.1, \pm 0.1)$.

Since they fall in the same quadrant (the second), the letters "a," "e," "i" and "o" should all be classified as either vowels or consonants, and we have, of course, chosen to call them vowels.

The letters "h," "n," "u" and "y" must be called neuter because the corresponding signs in the two vectors agree. The letter "q" occurred only once. The classification of q as neuter is nevertheless interesting. Its one occurrence is in the word "equal." One might expect it to be classified as a consonant, since it occurs between two vowels. Note, however, that the algorithm does not recognize u as a vowel. It is classified as neuter. This is, no doubt, attributable to the very high frequency of the digram ou, which accounts for 7 of the 21 occurrences of u. The letter "h"

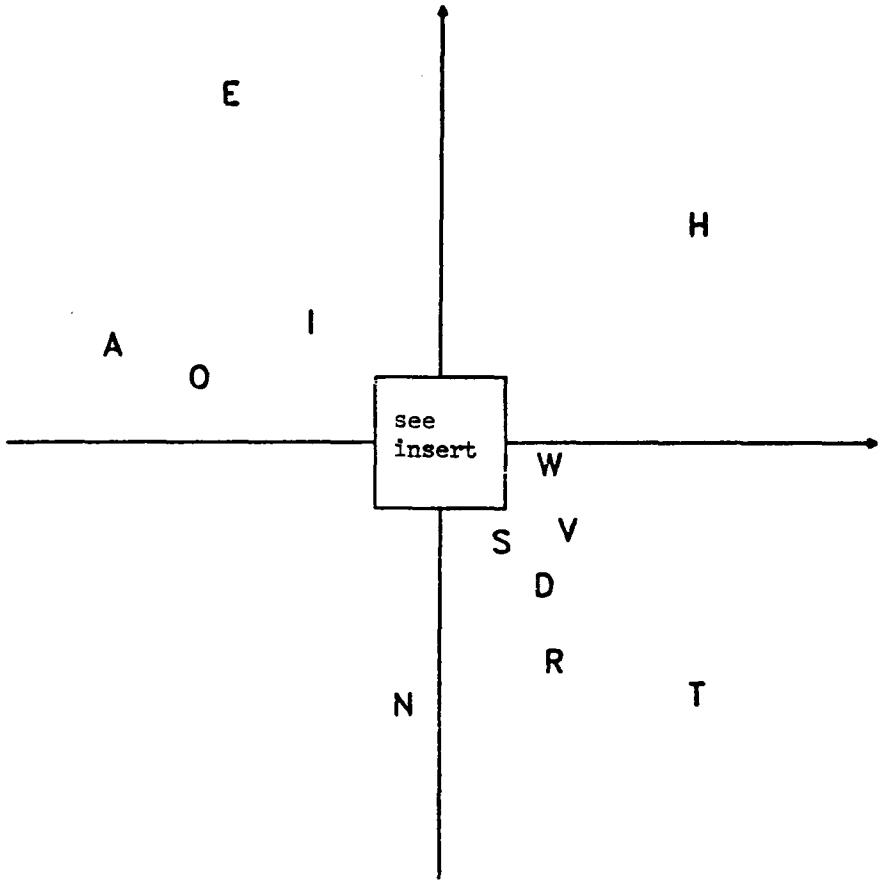

FIG. 5. The more frequent letters, plotted with coordinates from Fig. 4.

clearly shows a tendency to be followed by a vowel, and to be preceded by a consonant. The letter "n" shows a weak tendency to be followed by a consonant and a strong tendency to be preceded by a vowel. The other three neuter letters occur infrequently. The letter "j," "x," and "z" are seen not to occur at all. The remaining 14 letters are classified as consonants.

If a simple substitution cryptogram were made from text such as the Gettysburg Address, the digram matrix A would be replaced by PAP^T for some unknown permutation matrix P. As shown in Section 6, the singular vectors of the transformed matrix would classify the encoding of each letter in the same way as x_2 and y_2 classified the letters of the plaintext.

8. Other Languages. The same experiment was performed on texts of approximately one thousand characters each, written in five other languages selected for their diversity. The classifications of letters resulting from these tests are tabulated below.

Language	*Vowels*	*Consonants*	*Neuter*	*Absent*
Hawaiian	AEIOU	HKLMNPW		BCDFGJQ RSTVXYZ
Japanese	AEIOU	BDGHKMNRSTWYZ		CFJLPQVX
German	AEFOPU	BHIKLNRV	CDGJMSTWZ	QXY
Spanish	AEO	CDFGLMNQRSXYZ	BHIJPTUV	KW
Finnish	AEIOUY	DHJKLMNPRSTV	B	CFGQWXZ

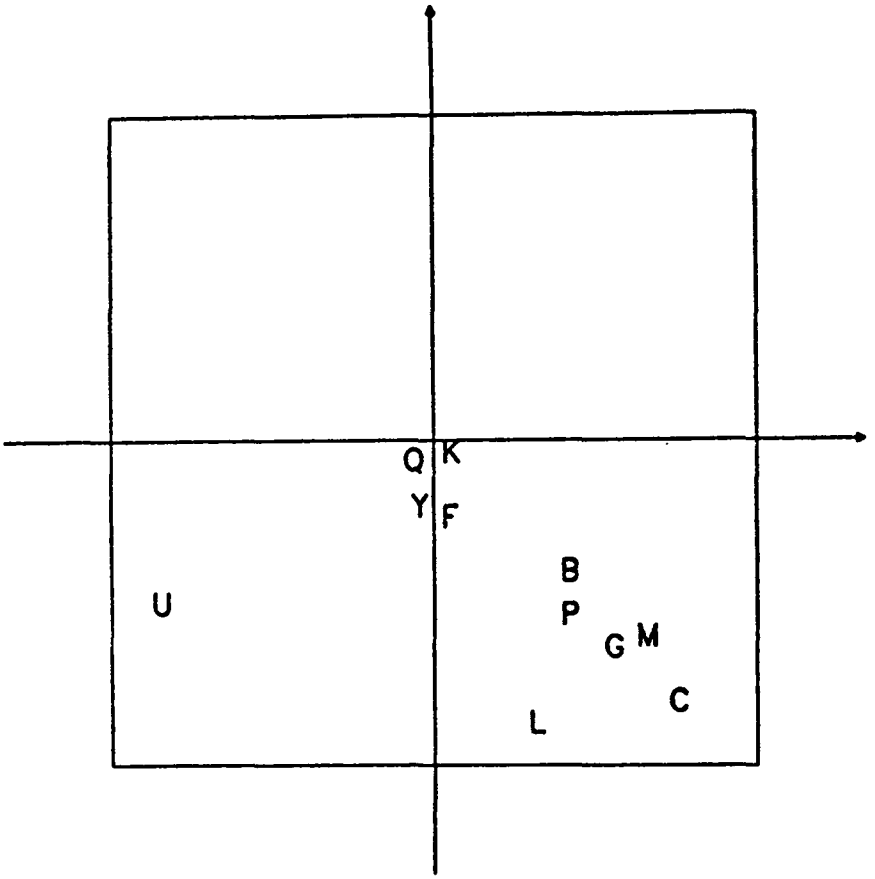

FIG. 6. The less frequent letters. The scale is 5 times that of Fig. 5.

None of the discrepancies is very surprising. Several of them are associated with letters which occurred so infrequently that no significance can be attached to their classification. In the Finnish example, B occurred only once, at the beginning of a word of foreign origin. It followed a final N in the preceding word and preceded an E. In the German text, P occurred only once, and that occurrence was in the very Germanic trigram SPR. Occurring only between two consonants, it is not surprising that the analysis classified it as a vowel. The most common neighbor of F, on both sides, in the selected sample, is R. The classification of I is confused because of the very high frequency of the diagrams IE and EI. More generally, the German language does not share the aversion of the other languages to consecutive consonants, and consequently many German letters fall into the neuter category. The classification of Y as a vowel in the Finnish sample is not surprising; Y has a value in Finnish that is hardly distinguishable to a non-Finnish ear, from that of U. The neuter classification of I and U in the Spanish example is clearly attributable to the high frequency of vowel-vowel digrams in which U or I is the first letter, having a value equivalent to W or Y in English. The perfect performance of the algorithm in Japanese and Hawaiian is clearly the result of their rigorously observed exclusion of consonant-consonant digrams. Most of the Japanese Hiragana characters are transliterated as a consonant-vowel digram.

9. Conclusions. The second singular vectors in the singular value decomposition of the digram frequency matrix provide the cryptanalyst with a helpful and surprisingly reliable way to classify the letters in a cryptogram as vowels or consonants, if the encoding algorithm is simple

substitution or k-alphabetic substitution, with k known, and if the text if vfc text. Texts written in many natural languages, with certain exceptions, tend to be vfc texts. Near-exceptions are German (and, presumably, other germanic languages) in which the vfc character is somewhat diminished by the frequency of consonant-consonant digrams, and Spanish (and, presumably, other romance languages) in which certain vowel-vowel pairs are frequent. Despite these deviations from the vfc rule, the second singular vectors classify correctly most of the letters which occur with a frequency high enough to be statistically significant.

A computer program which uses the SVD as the starting point in an automated, heuristic approach to solving cryptograms is described by Schatz [7].

References

1. H. C. Andrews and C. L. Patterson, Outer product expansions and their uses in digital image processing, this MONTHLY, 82 (1975) 1–12.

2. H. F. Gaines, Cryptanalysis, Dover, New York, 1956.

3. G. E. Forsythe, M. A. Malcolm, and C. B. Moler, Computer Methods for Mathematical Computations, Prentice-Hall, Englewood Cliffs, NJ, 1977.

4. Richard S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1962.

5. A. Sinkov, Elementary Cryptanalysis, A Mathematical Approach, New Mathematical Library, vol. 22, Mathematical Association of America, Washington, D.C., 1968.

6. J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, LINPACK Users' Guide, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

7. Bruce R. Schatz, Automated analysis of cryptograms, Cryptologia, 1 (1977) 116–142.

WHAT IS THE GEOMETRY OF A SURFACE?

ROGER FENN

School of Mathematics and Physical Sciences, University of Sussex, Falmer, Brighton BN1 9QH, England

1. In this article we shall assume the definition of a geometry proposed by Klein in his Erlangen program, that is, a geometry is a group of transformations of space together with the propositions left invariant by this group. The space in question will be the two-dimensional plane. By a surface we shall mean a closed compact orientable surface of genus γ. That is a sphere with γ handles attached. The fact that two such surfaces are homeomorphic if and only if their genera are equal was probably known to Riemann. A modern proof can be found in Massey's book [5, Chapter 1].

The exact result which will be proved is the following:

THEOREM 1. *Let $\gamma > 1$. Then there is a group of hyperbolic translations of the hyperbolic plane such that the space of orbits under these translations is homeomorphic to a surface of genus γ.*

Moreover, there is a compact polygon in the plane whose translates under the group are distinct for distinct group elements and which form a network of nonoverlapping polygons covering the whole hyperbolic plane.

Our aim is to prove the above by means as elementary as possible, in particular without using deep results from the theory of functions.

In order to illustrate the general result, we consider firstly the case where $\gamma = 1$. Here the

---

Roger Fenn is a Lecturer in Mathematics at the University of Sussex, England. He received his Ph.D. in 1968 under the direction of John Reeve and is presently writing a book on Geometric Topology. He believes people should be at the same time renaissance polymaths and twentieth century specialists insofar as this is possible. He is married, with three children, and enjoys playing chess, singing madrigals, and walking.