HIDDEN MARKOV MODELS FOR ENGLISH

Robert L. Cave
Lee P. Neuwirth

In 1913 A. A. Markov wrote a paper [1], in which he analyzed "chains" of vowels and consonants in Pushkin's Poem Eugene Onegin. The ideas and applications in this paper continue some of his earlier work [2, 3, 4, 5]. But more significantly, the concepts introduced in these papers have grown in applicability and have proved so important that later authors coined the phrase "Markov Chain" to describe the mathematical situation which is now very well known. The special situation Markov examined is particularly interesting to us, and we have tried, in the spirit of Markov's application, to examine the relation of Markov chains to the English language. The present form of written English is the result of a long complex process. Fascinating as this evolutionary process is, it is possible to ignore it completely, take a narrow view of the language, and recover some overt properties as well as try to understand the manner in which letters are put together. Instead of examining sentence structure or the etymology of words we may view language as a sequence of symbols from a 27-letter alphabet (Space is the 27th letter.) It is from this myopic viewpoint that we try to analyze such sequences.

Such efforts have been made before, but our method and results are new. Our results are, we believe, not surprising in the sense that they are subject to "natural" interpretation. For example, we find that separating the letters of the alphabet into vowels and consonants, as Markov did for his analysis, is proper in a very strong statistical sense for English. We are further able, roughly speaking, to "refine" the original separation into two classes by making more classes. We have succeeded in analyzing a separation for up to twelve classes. (These classes are not disjoint as will be seen.)

The text chosen for analysis was from the Brown University Sample of Present Day English. We have included word space as a twenty-seventh letter but have eliminated all case, punctuation and hyphenations.

All of this work was done at the Communications Research Division of the Institute for Defense Analyses.

II. The Type of Model

In order to analyze his text, Markov reduced the Russian alphabet to just two symbols, vowel and consonant, and explored the chains of symbols which resulted. We are more interested in the chains of English letters themselves, so that we must provide in our model means for generating letters. We could look upon the sequence of

letters themselves as a Markov chain of order one (or higher order), and this has been done frequently in the past. The resulting model is usually referred to as digraphic English and it requires 702 parameters to specify it. We would like to reduce the number of parameters and also learn more about the language itself, so we will not use this well-known model.

A particularly convenient model which provides what we need, does not destroy Markov's original intent, and is efficient as well, is the following:

We suppose the existence of a Markov chain of order one with two states. Let us call these states V and C, and the transition matrix A. In state V we produce each English letter (including word-space) with probability $P_V(a)$, $P_V(b)$, ..., $P_V(z)$, $P_V(\#)$.[1] In state C we produce the letters with probability $P_C(a)$, $P_C(b)$, ..., $P_C(z)$, $P_C(\#)$. Now of course we could make P_V zero for consonants, and non-zero and equal to the appropriate probability for vowels. Similarly we could make P_C zero for vowels and an appropriate value for consonants. If we did this then we would have a decent model for the generation of English which preserves Markov's division. On the other hand, let us consider for a moment our situation. We have a 2×2

[1] # means word-space.

transition matrix A, and a 2×27 matrix B. A contains the probabilities of the four sequences vowel-vowel, vowel-consonant, consonant-vowel, consonant-consonant. The matrix B contains the probabilities for each letter in case we are in state V, or in case we are in state C.

Now, we may ask, why are these two matrices, which after all completely determine our statistical model, the best ones to take? Perhaps some other pair of matrices would be better. Of course we have not said what "better" means, so let us agree now on that point. We say that one model, X, is better than another, Y, if the probability of producing some long sequence of English text is higher with X than with Y. So we may now rephrase our question more precisely: Given a long sequence of text what is the best model of the type we are considering? That is, what pair of stochastic matrices A, B will maximize the probability of observing the text under consideration? Later in this paper using the methods of [7] we propose an answer to this question.

Now it is clear how to generalize Markov's dichotomy; we have only to make our underlying Markov chains have more states!

So, to summarize, we will take as our general model an S state Markov chain of order 1, and for each state of that chain we will have a probability distribution on our

alphabet, i.e., a letter is produced with a probability dependent upon the underlying state. The states themselves are not assumed to be observable, and indeed we can only determine probabilistically which state we are in at a given time. Since this is the case we call these models Hidden Markov Models.[1]

Now for each $S = 2, 3, \ldots, 12$ we have found (we believe) that $S \times S$ matrix, and that set of S probability distributions which maximize the probability of observing some particular long stretch of English text.[7]

III. The Models Themselves

In the tables following, we present the raw results of the calculations. There is given for each $S = 2, 3, \ldots, 12$, a model, that is, an $S \times S$ transition matrix, and an $S \times 27$ output matrix giving the output distributions in each state. The stationary probabilities which are the probabilities of being in each state after a long period of time are also shown.

IV. Discussion of the Models

What is displayed in Tables I-1 — I-11 is merely a collection of parameters for a sequence of statistical models of English. What is important to keep in mind in

[1] They have been called probabilistic functions of Markov chains, but we find this a bit unwieldy.

TABLE I-1. HIDDEN MARKOV MODEL*

S = 2

Transition Probabilities

1	.275	.725
2	.780	.220

Output Probabilities

	1	2
A	—	.133
B	.022	—
C	.063	—
D	.056	—
E	—	.218
F	.037	—
G	.015	.010
H	.074	—
I	—	.150
J	—	—
K	.009	—
L	.060	—
M	.041	—
N	.140	—
O	—	.136
P	.030	.001
Q	.001	—
R	.087	—
S	.105	—
T	.157	.019
U	—	.045
V	.016	—
W	.020	—
X	.002	—
Y	.004	.018
Z	.001	—
#	.060	.269

Stationary Probabilities

.52	.48

* # denotes "Word Space".

TABLE I-2

S = 3

Transition Probabilities

1	.045	.479	.476
2	—	.322	.678
3	.864	—	.136

Output Probabilities

	1	2	3
A	—	—	.163
B	.031	.001	—
C	.078	.020	—
D	.051	.043	—
E	—	.173	.157
F	.048	.009	—
G	.018	.024	—
H	—	.153	—
I	—	.024	.169
J	.001	—	—
K	.009	.006	—
L	.056	.045	—
M	.039	.029	—
N	.200	.007	—
O	—	.022	.153
P	.035	.014	—
Q	.002	—	—
R	.087	.057	—
S	.113	.056	—
T	.178	.110	—
U	—	—	.055
V	.020	.006	—
W	.029	—	—
X	.002	—	—
Y	—	.042	—
Z	.001	—	—
#	.004	.160	.303

Stationary Probabilities

.36	.25	.39

TABLE I-3

S = 4

Transition Probabilities

1	—	—	.265	.735
2	.577	.327	—	.097
3	.105	.621	.093	.181
4	—	—	.899	.101

Output Probabilities

	1	2	3	4
A	.007	—	.217	—
B	—	.003	.003	.039
C	.006	.017	—	.086
D	—	.083	—	.029
E	—	.166	.210	—
F	—	.027	—	.033
G	—	.023	—	.044
H	—	—	.124	.037
I	.036	—	.172	—
J	—	—	—	.004
K	.001	.008	—	.003
L	—	.071	—	.073
M	—	.021	—	.057
N	.011	.155	—	.037
O	—	—	.217	—
P	.004	.009	—	.054
Q	—	—	.001	—
R	—	.095	—	.095
S	—	.133	—	.092
T	—	.110	—	.229
U	.006	.013	.050	—
V	—	—	—	.043
W	—	.016	—	.042
X	—	.005	—	—
Y	—	.043	.006	.001
Z	—	.001	—	.001
#	.928	—	—	—

Stationary Probabilities

.19	.27	.30	.24

TABLE I-4

S = 5

Transition Probabilities

1	.118	.026	.854	—	—
2	.014	.161	—	—	.824
3	.141	.135	.127	.520	.075
4	.197	.396	.026	.015	.364
5	.732	.004	.263	—	—

Output Probabilities

	1	2	3	4	5
A	—	—	.236	—	—
B	.036	—	—	.006	—
C	.114	—	—	.016	.005
D	.053	.182	—	.020	—
E	—	.260	.237	—	—
F	.040	.003	—	.043	—
G	.029	.054	—	.007	.006
H	.048	—	.088	—	—
I	—	—	.150	.004	.072
J	.004	—	—	—	—
K	.012	.007	—	.001	—
L	.047	.047	—	.137	—
M	.052	—	—	.034	—
N	.034	.004	—	.326	—
O	—	—	.216	—	—
P	.077	.004	—	.011	—
Q	.002	—	—	—	—
R	.093	—	—	.213	—
S	.082	.192	—	.108	—
T	.192	.167	—	.022	—
U	—	—	.068	.003	.002
V	.024	.001	—	.005	—
W	.038	—	—	.007	—
X	—	—	—	.016	—
Y	.012	.073	.001	.010	—
Z	.002	—	—	—	—
#	—	—	—	—	.912

Stationary Probabilities

.238	.129	.293	.155	.185

TABLE I-5

S = 6

Transition Probabilities

1	.009	.052	.215	.364	—	.357
2	.752	.021	.140	.007	—	.077
3	—	.611	.002	.021	.364	—
4	—	—	—	.171	—	.828
5	—	.724	.022	.238	.002	.011
6	—	.276	.664	—	.059	—

Output Probabilities

	1	2	3	4	5	6
A	—	.268	—	—	.004	.007
B	.001	—	.059	—	—	—
C	.047	—	.077	—	—	.008
D	.034	—	.046	.186	—	.002
E	—	.241	—	.357	.080	—
F	.044	—	.058	.001	—	—
G	—	.002	.047	.046	—	—
H	—	—	—	.023	.435	—
I	.016	.163	—	—	.125	.042
J	—	—	.016	—	—	—
K	.005	—	.005	.012	—	—
L	.091	—	.035	.035	.077	—
M	.039	—	.043	.006	.035	—
N	.265	—	.030	.002	.002	—
O	—	.239	—	—	.063	—
P	.012	—	.083	—	.002	—
Q	—	—	.002	—	—	—
R	.200	—	.042	—	.123	—
S	.103	—	.077	.134	—	.008
T	.077	—	.274	.108	.026	—
U	.013	.081	—	—	.015	.006
V	.008	—	.033	—	—	—
W	.007	—	.061	—	.008	—
X	.009	—	—	—	—	—
Y	.019	.001	.003	.084	—	—
Z	.001	—	.001	—	—	—
#	—	—	—	—	—	.924

Stationary Probabilities

.186	.244	.195	.112	.082	.179

TABLE I-6

S = 7

Transition Probabilities

1	.004	—	—	—	—	.995	—
2	.268	—	.447	—	—	—	.283
3	—	—	—	.029	—	—	.970
4	—	.814	.011	.173	—	—	—
5	.061	.362	.160	.377	.014	—	.024
6	—	.139	.029	.261	—	—	.569
7	.037	.052	.110	.017	.742	—	.039

Output Probabilities

	1	2	3	4	5	6	7
A	—	.006	—	—	.002	—	.271
B	.029	—	.062	—	.005	—	—
C	.099	.007	.073	—	.036	—	—
D	—	—	.062	.194	.031	—	—
E	—	—	—	.363	—	.058	.248
F	.022	—	.070	.001	.044	—	—
G	.081	—	.027	.044	.001	—	—
H	—	—	.052	—	—	.450	—
I	—	.058	—	—	.016	.111	.163
J	—	—	.023	—	—	—	—
K	—	—	.008	.012	.005	—	—
L	—	—	.053	.045	.096	.054	—
M	—	—	.071	.007	.039	.021	—
N	—	—	.045	.004	.265	.005	—
O	—	—	—	—	—	.125	.227
P	.090	—	.073	—	.011	.003	—
Q	—	—	.002	—	—	—	—
R	—	—	.079	—	.198	.120	—
S	.066	.007	.072	.133	.107	—	—
T	.556	—	.099	.108	.081	.034	—
U	—	.007	—	—	.012	.007	.086
V	—	—	.050	—	.006	—	—
W	.054	—	.060	—	.008	.005	—
X	—	—	—	—	.009	—	—
Y	—	—	.007	.084	.016	.001	.001
Z	—	—	.002	—	—	—	—
#	—	.911	—	—	—	—	—

Stationary Probabilities

.069	.181	.140	.115	.182	.069	.241

TABLE I-7

S = 8

Transition Probabilities

1	.003	.006	.165	.674	—	.051	.098	—
2	.002	.050	.151	.036	.074	.088	.595	—
3	.085	.097	.089	.707	.002	.017	—	—
4	—	.278	—	—	.314	.406	—	—
5	.184	.136	—	—	—	—	—	.678
6	.077	.911	—	—	.010	—	—	—
7	.081	.008	.366	.319	.121	.095	.007	—
8	.348	.529	—	.093	.001	.027	—	—

Output Probabilities

	1	2	3	4	5	6	7	8
A	.002	.286	—	.013	—	—	—	—
B	.001	—	—	—	.061	.049	.002	—
C	—	—	—	.005	.109	.086	.039	—
D	—	—	.263	—	.002	.063	.010	—
E	.798	.212	—	—	—	—	—	.035
F	—	—	.020	—	.015	.075	.046	—
G	—	—	.057	—	.071	.014	—	—
H	—	—	.001	—	—	.064	—	.492
I	—	.178	—	.061	—	—	.019	.106
J	—	—	—	—	—	.026	—	—
K	—	—	.022	—	.002	.004	—	—
L	—	—	.077	—	—	.055	.113	.074
M	—	—	.016	—	—	.089	.043	.020
N	.007	—	.016	—	—	.051	.324	.006
O	.088	.234	—	—	—	—	—	.074
P	—	—	—	—	.086	.076	.013	.002
Q	—	—	—	—	—	.005	—	—
R	—	—	—	—	—	.103	.242	.133
S	—	—	.230	.003	.067	.077	.076	—
T	.024	—	.215	—	.544	—	.024	.034
U	—	.088	—	.007	—	—	.018	.010
V	—	—	—	—	—	.076	—	—
W	—	—	—	—	.039	.071	.011	.006
X	—	—	—	.001	—	—	.011	—
Y	.076	—	.078	—	—	.004	.001	.001
Z	—	—	—	—	—	.004	—	—
#	—	—	—	.905	—	—	—	—

Stationary Probabilities

.070	.226	.108	.183	.093	.115	.142	.063

TABLE I-8

S = 9

Transition Probabilities

1	.002	.026	.042	—	—	.310	.045	.401	.100
2	—	—	.017	—	.182	—	.798	—	.002
3	—	—	—	.740	.196	—	.063	—	—
4	—	—	—	.007	.514	.001	.476	—	—
5	.096	.199	—	—	—	.481	.199	—	.022
6	—	.395	.245	.062	—	—	.296	—	—
7	.643	.053	.016	—	—	.030	.032	.211	.011
8	—	.027	—	.009	—	.561	.194	.001	.205
9	—	—	.003	—	—	.821	—	.174	—

Output Probabilities

	1	2	3	4	5	6	7	8	9
A	—	—	—	—	—	—	.303	—	.001
B	.014	.081	.017	—	—	—	—	—	—
C	.052	.117	.081	—	—	—	—	.015	—
D	.008	.070	.004	—	—	—	—	.237	—
E	—	—	—	—	.659	—	.208	—	.439
F	.044	.070	.015	—	—	—	—	.027	—
G	—	.026	.022	—	—	—	—	.079	—
H	—	—	—	.602	—	.004	—	—	.100
I	.018	—	—	.040	—	.075	.182	—	—
J	—	.022	—	—	—	—	—	—	—
K	—	.001	.001	—	—	—	—	.030	—
L	.105	.049	—	.052	—	—	—	.103	—
M	.035	.090	—	.012	—	—	—	.023	—
N	.318	.047	—	.002	—	.003	—	—	.020
O	—	—	—	.002	.316	—	.209	—	—
P	.010	.090	.108	.005	—	—	—	.001	—
Q	—	.005	—	—	—	—	—	—	—
R	.227	.057	—	.204	—	—	—	.009	—
S	.109	.105	.068	—	—	—	—	.083	.262
T	.015	—	.619	.060	—	—	—	.347	—
U	.015	—	—	.005	—	.005	.094	—	—
V	—	.069	—	—	—	—	—	—	—
W	.010	.073	.054	.009	—	—	—	—	—
X	.010	.005	—	—	—	—	—	—	—
Y	—	.008	—	.001	.024	—	—	.038	.175
Z	—	.004	—	—	—	—	—	—	—
#	—	—	—	—	—	.910	—	—	—

Stationary Probabilities

.150	.114	.057	.055	.060	.183	.223	.115	.043

TABLE I-9

S = 10

Transition Probabilities

1	—	.073	—	—	.721	.014	—	.003	.073	.113
2	.375	—	—	—	—	—	—	.152	—	.471
3	—	.008	.011	—	.006	—	—	.006	.966	—
4	—	.034	.166	—	.001	—	.005	.060	.579	.150
5	—	.101	.060	.403	.010	.045	—	.067	.310	—
6	.136	—	.126	—	—	.028	.707	—	—	—
7	.163	.011	.319	—	—	—	—	.286	.072	.147
8	—	—	—	.745	.173	.080	—	—	—	—
9	.241	.386	—	—	—	.271	.043	.054	—	.002
10	.072	.122	.014	.151	.532	.062	—	.044	—	—

Output Probabilities

	1	2	3	4	5	6	7	8	9	10
A	.318	—	.003	—	—	—	.043	.463	.001	—
B	—	.062	—	—	.007	.067	—	—	—	—
C	—	.100	—	.027	.037	.118	—	—	—	—
D	—	.073	—	.230	.010	.002	—	—	—	—
E	.113	—	.640	—	—	—	—	.244	—	.439
F	—	.081	—	.004	.061	.012	—	—	—	—
G	.004	.026	—	.067	—	.034	—	—	—	—
H	—	.010	.002	—	—	—	.625	—	.003	.001
I	.101	—	—	—	—	—	.002	.274	.071	.193
J	—	.023	—	—	—	—	—	—	—	—
K	—	.002	—	.029	—	—	—	—	—	—
L	—	.055	—	.083	.109	—	.083	—	—	—
M	—	.100	—	.017	.042	—	.003	—	—	—
N	—	.049	.003	—	.348	—	.002	—	—	—
O	.445	—	.078	—	—	—	.005	.016	—	.122
P	—	.091	—	—	.011	.094	.004	—	—	—
Q	—	.006	—	—	—	—	—	—	—	—
R	—	.080	—	—	.254	—	.146	—	—	—
S	—	.058	.117	.163	.077	.128	—	—	.003	—
T	—	.037	.028	.301	.005	.493	.073	—	—	—
U	.017	—	—	—	.007	—	—	.001	—	.240
V	—	.052	—	.019	—	—	—	—	—	—
W	—	.076	—	—	.011	.047	.006	—	—	—
X	—	—	—	—	.016	—	—	—	—	—
Y	—	.009	.124	.050	—	—	—	—	—	.002
Z	—	.002	—	.002	—	—	—	—	—	—
#	—	—	—	—	—	—	—	—	.919	—

Stationary Probabilities

.110	.108	.058	.119	.141	.070	.058	.065	.181	.091

TABLE I-10

S = 11

Transition Probabilities

1	—	.049	—	—	.068	—	—	.105	.535	—	.240
2	—	.040	.008	—	—	.089	.437	.001	—	.049	.372
3	—	.036	—	.698	.221	.001	—	—	.042	—	—
4	.013	.285	—	.007	.501	—	—	—	.190	.001	—
5	—	.033	.007	—	.011	.166	.159	.012	—	.608	—
6	—	—	—	—	.019	—	—	.015	—	.965	—
7	.010	.054	.029	—	.072	.107	.001	.078	—	.276	.370
8	—	.313	—	—	—	—	—	.008	.675	—	.002
9	—	.025	.004	—	.004	—	.665	.116	.076	.005	.101
10	—	.178	.253	.044	—	—	—	.409	.114	—	—
11	.192	.105	.002	—	.218	.124	—	.013	.076	.236	.029

Output Probabilities

	1	2	3	4	5	6	7	8	9	10	11
A	—	.596	—	—	—	—	—	—	.035	—	—
B	—	—	.077	—	—	—	.015	.063	—	—	.001
C	—	—	.041	—	—	—	.043	.110	—	—	.071
D	—	—	.002	—	—	.333	.011	.085	—	—	.081
E	—	—	—	—	.854	—	—	—	.340	—	—
F	—	—	.014	—	—	—	.058	.081	—	—	.012
G	—	—	.024	—	—	.030	—	.031	—	—	.067
H	.147	—	—	.700	—	.068	—	.021	—	—	—
I	.679	.359	—	—	—	—	.003	—	.033	—	—
J	—	—	—	—	—	—	—	.025	—	—	—
K	—	—	—	—	—	.001	—	.001	—	—	.030
L	—	—	—	.047	—	—	.098	.054	—	—	.126
M	.068	—	—	—	—	.002	.041	.081	—	—	.027
N	—	—	—	—	—	.013	.319	.038	—	—	.036
O	—	.002	—	.003	.116	—	—	—	.443	—	—
P	—	—	.115	—	—	—	.013	.105	—	—	.001
Q	—	—	—	—	—	—	—	.006	—	—	—
R	—	—	—	.212	—	—	.248	.073	—	—	.008
S	—	.041	.027	—	—	.361	.080	.095	—	—	.065
T	.039	—	.635	—	—	—	.009	—	—	—	.413
U	.065	—	—	.003	—	—	.028	—	.147	—	—
V	—	—	—	—	—	—	—	.037	—	—	.035
W	—	—	.059	.009	—	—	.012	.084	—	—	—
X	—	—	—	—	—	—	.015	—	—	—	—
Y	—	—	—	.023	.029	.188	—	—	—	—	.015
Z	—	—	—	—	—	—	—	—	—	—	.004
#	—	—	—	—	—	—	—	—	—	1.00	—

Stationary Probabilities

.024	.106	.049	.042	.071	.051	.144	.101	.129	.166	.116

TABLE I-11

S = 12

Transition Probabilities

1	—	—	—	—	—	.022	—	.015	—	—	.959	.001
2	—	—	.383	—	—	.229	.061	.051	—	.243	—	.029
3	.116	.016	.010	—	.035	.016	.647	.078	—	—	.057	.020
4	—	.036	—	—	.262	.023	.003	.379	—	.213	.080	—
5	.120	—	.467	—	—	—	.340	—	.012	—	—	.058
6	.009	.297	—	—	.350	.014	—	.024	—	.302	—	—
7	.115	.240	—	—	.058	.045	—	.190	—	.063	.284	—
8	.252	—	.041	—	—	—	—	—	.035	—	.667	.003
9	.075	.014	—	—	.062	.166	.004	.063	.005	.001	.551	.054
10	—	—	.108	—	.135	.064	.073	—	.495	—	.117	.005
11	—	—	—	.018	.097	.417	—	—	—	.177	—	.288
12	—	—	—	.669	—	—	—	—	—	.293	—	.036

Output Probabilities

	1	2	3	4	5	6	7	8	9	10	11	12
1	—	—	—	—	.456	—	—	—	—	.218	—	—
2	—	—	.027	—	—	.090	.001	—	—	—	—	.001
3	—	—	.079	—	—	.070	.044	—	—	—	—	.140
4	.320	—	.015	—	—	.084	.100	—	—	—	—	.001
5	—	.442	—	—	.034	—	—	.960	—	.213	—	—
6	.005	—	.003	—	—	.082	—	—	.139	—	—	.012
7	—	—	—	—	—	.027	.074	—	—	.002	—	.049
8	.013	.001	—	.718	—	.043	—	—	—	—	—	—
9	—	.458	.029	—	.241	—	—	—	—	.058	—	—
10	—	—	—	—	—	.022	—	—	—	—	—	.001
11	—	—	—	—	—	.001	.035	—	.002	—	—	—
12	—	—	.131	.031	—	.049	.155	.008	.046	—	—	—
13	.005	—	—	—	—	.101	.014	—	.105	—	—	—
14	.019	—	.320	.001	—	.039	.015	—	.360	—	—	—
15	—	.007	—	.046	.066	—	—	—	—	.506	—	—
16	—	—	.005	—	—	.085	.003	—	.021	—	.002	.105
17	—	—	—	—	—	.006	—	—	—	—	—	—
18	.004	—	.236	.151	—	.097	.009	—	.246	—	—	—
19	.411	—	.118	—	—	.076	.064	—	.007	—	.003	.101
20	—	—	.026	.042	—	—	.423	.029	—	—	—	.528
21	—	.090	—	.004	.197	—	—	—	.006	—	—	—
22	—	—	—	—	—	.035	.040	—	—	—	—	—
23	—	—	.001	—	—	.073	—	—	.028	—	—	.058
24	—	—	.001	—	—	—	—	—	.035	—	—	—
25	.220	—	—	.003	.003	.008	.018	.001	—	—	—	—
26	—	—	.002	—	—	.002	—	—	—	—	—	—
27	—	—	—	—	—	—	—	—	—	—	.994	—

Stationary Probabilities

.051	.060	.083	.046	.095	.111	.098	.053	.059	.113	.167	.064

the following discussion is as we have said that in each case we have obtained that set of parameters which maximizes the probability of observing a certain long (15,000 letters) sequence of English text.

A cursory examination of the first of the models (2-state) suggests the naturalness (for English) of Markov's choice of a dichotomy of his study. It is clear from this model that the division of English into vowels and consonants is natural in a statistical sense. Although y is more comfortable in this model as a vowel, later models have led us to reject y as a vowel, and from now on let us agree to call only a, e, i, o, u vowels. On the other hand, such a simple division is no longer possible in the bigger models, and an understanding of the meaning of various states requires some more careful examination.

We notice in the 7 state model (and others) that a particular state (state 7 in the 7 state model) is a "vowel" state and we see from the transition matrix that state 3 is a kind of "pre-vowel" state; on the other hand, state 4 is usually followed by state 2 which is dominated by #. Thus state 4 might be thought of as a "final-letter" state.

This kind of analysis could be continued but let us say a bit more precisely what it is we need to do in order to understand our models. We would like to

"associate" somehow the underlying states with definable properties of English text letters. A little thought on this makes one realize that such an association might be made in one of two ways. On the one hand we can say that when a certain property p is satisfied by some letter, then we must be in a particular state s at that time, ($p \Longrightarrow s$). On the other hand we can say that when we are in some state, then whatever letter is produced must have a particular property ($s \Longrightarrow p$). Now the fact of the matter is that both of these possibilities (which we call Implicative Associations) occur in our models, as well as their conjunction. But, we hasten to add, one must broaden the idea of association so that a set of states rather than a single state may be associated with some observable property. Now of course the negation of a property is again a property, as well as the conjunction, and disjunction of properties. These logical operations among properties are mirrored by the appropriate operations among the subsets of states which correspond to them, e.g., suppose σ and τ are subsets of the set of states which are associated with vowels and initial letters, respectively, then $\sigma \cap \tau$ will be associated with initial letters which are vowels. The association between $\sigma \cap \tau$ and initial vowels might not be of the implicative type if

σ and τ are not implicative in the same direction. To stem the confusion which must now be arising in the reader's mind, we may consider for illustrative purposes the 11 state model:

In this model, the following are examples of subsets of states of Implicative type with the implications going in the directions indicated to the properties indicated:

Properties		Subsets
Word Space	<==>	{10}
Vowel	<==	{2,5,9}
Initial Letter	==>	{2,3,4,8,9}
Final Letter	==	5,6,7,11}
Vowel Preceder	<==	{4,9}
Vowel Successor	==>	{7}
Consonant Successor	<==	{1}
J	==>	{8}
Non-final silent h or i before e or o	==>	{1}

This list is not meant to be exhaustive, but it contains the most obvious set of properties we have been able to isolate. Except for the last, the associations listed

may be verified by examining the matrices for the 11 state model. Other implications, such as the last may be verified by computing the probability of being in each state at each text position [7], and then examining particular text configurations. We have asserted implications whenever they are true roughly 95% of the time. In Table II are displayed lists for each of the 11 models constructed.

We may look at the manner in which new states are utilized as we increase the size of our model. The very first division which appears is the vowel-consonant one, (with # favored 9 to 2 as a vowel), and this division persists through the largest model. Eventually four states are required for the vowels and this allows different distributions for initial and final vowels as we shall see in a moment.

The second property extracted by this technique is the word space, and following this, word beginnings and word endings appear in the list of implicative properties. Since the latter two each include several states we can note that the intersection of the initial (or final) states with the vowel states gives states properly interpreted as initial and final vowel states, although these are not implicative properties since initial and final letters are implicative in one direction, while vowels are implicative in the other.

TABLE II

2 States

			Remarks
Vowel	⟶	(2)	

3 States

Vowel or Space	⟵	(3)	The only exit from 2 is to 3
Space	⟶	(2,3)	The only entry to 3 is from 1
Consonant	⟵	(1)	

4 States

Vowel or h	⟵	(3)
Space	⟶	(1)
Consonant	⟵	(4)
Final letter	⟶	(2,3)
Initial letter	⟶	(3,4)

5 States

Vowel or h	⟵	(3)
Space	⟶	(5)
Consonant	⟵	(1,4)
Final Letter	⟶	(2,3,4)
Initial Letter	⟶	(1,3)

6 States

Vowel	⟵	(2)
Space	⟶	(6)
Consonant	⟵	(3)
Final Letter	⟶	(1,2,4)
Initial Letter	⟶	(2,3,5)
Vowel Follower	⟵	(1)
Non-Vowel Follower (Consonant or #)	⟵	(5)

7 States

Vowel	<—	(7)
Space	—>	(2)
Consonant	<—	(1,3)
Final Letter	—>	(4,5,6,7)
Initial Letter	—>	(1,3,7)
Vowel Follower	<—	(5)
Vowel Preceder	<—	(3)
Consonant Follower	<—	(6)

State 6 is entered only from 1, t dominates 1, and h, 6.

8 States

Vowel	<—	(2)
Space	—>	(4)
Consonant	<—	(3,5,6)
Final Letter	—>	(1,2,3,7,8)
Initial Letter	—>	(2,5,6)
Vowel Follower	<—	(7)
Vowel Preceder	<—	(6)
Consonant Follower	<—	(8)

State 8 is entered only from 5. t dominates 5, and h, 8. State 1 produces a vowel with probability ~.888.

9 States

Vowel	<—	(5,7)
Space	—>	(6)
Consonant	<—	(2,3,8)
Final Letter	—>	(1,5,7,8,9)
Initial Letter	—>	(2,3,4,7)
Vowel Follower	<—	(1)
Vowel Preceder	<—	(2,4)

The transition 3 —> 4 is strong. t dominates 3, h dominates 4.

10 States

Vowel	<—	(1,8,10)
Space	—>	(9)
Consonant	<—	(2,4,6)
Final Letter	—>	(1,3,4,5,7)
Initial Letter	—>	(1,2,6,7,8)
Vowel Follower	<—	(5)
Vowel Preceder	<—	(2)
Final Letter	<—	(3)

State 7 tends strongly to produce vowel preceders. The transition 6 —> 7 is very strong, t dominates 6, h dominates 7.

11 States

Vowel	<—	{2,5,9}
Space	<—>	{10}
Consonant	<—	{3,4,6,7,8,11}
Final Letter	—>	{2,5,6,7,11}
Initial Letter	—>	{2,3,4,8,9}
Vowel Follower	<—	{7}
Vowel Preceder	<—	{4,8}
Final Letter	<—	{6}
Post-Consonant	<—	{1}

The transition 3 ——> 4 is strong. t dominates 3, h dominates 4, i dominates 1, and i is entered 90% of the time from 3 (t dominated).

More detailed analysis has shown that i in ie and io combination arises from state 1.

12 States

Vowel	<—	{2,5,8,10}
Space	<—>	{11}
Consonant	<—	{1,6,7,9,12}
Final Letter	—>	{1,3,4,7,8,9,10}
Initial Letter	—>	{4,5,6,10,12}
Vowel Follower	<—	{3,9}
Vowel Preceder	<—	{6}
Final Letter	<—	{1}
Consonant Follower	<—	{4,2}
E	<—	{8}

The transition 12 ——> 4 is strong. t dominates 12, h dominates 4.

A vowel following state is next to appear although the most common digraph (pair of adjacent letters) th, begins to influence the models at about this point (6 states) and a consonant follower state containing h appears. As with the vowel states almost all the implicative properties which present themselves persist through 12 states.

As may be guessed, a vowel precedent state next shows itself, and this is essentially the last new state type. What does happen when more states are added is that the space state becomes implicative in both directions, as does a final letter state. This is what the model is willing to "spend" its parameters on, and it is interesting to note that rather than reflect new properties, the new model has "chosen" to strengthen the properties governing its behavior in smaller versions. There is, despite the remarks above concerning the persistence of properties, a trend toward states dominated by single letters, and the "e" state in the 12 state model is the strongest example of this, although the remarks alongside the models point out other instances. The reader may notice how h becomes more and more powerful in one state.

IV. The Letters of the Alphabet

The letters themselves may be examined from the point of view we have been developing here. To fix ideas

let us concentrate on the 12 state model. As one might expect, q is to be found in the pre-vowel state 6. If each of the other letters is inspected in this way, the table on the following page results. We have marked only those properties opposite a letter which are practically assured. Thus, for example, if a letter can occur in other than a final letter state, an x is not marked in the final letter column opposite that letter.

Each letter has its own character, appearing in each of the 12 states with some (possibly zero) probability. These 12 numbers present a unique profile of each letter, and similarities in these profiles will naturally suggest similarities in usage. Thus, e.g., l and r have similar profiles. It is instructive to plot these profiles for the 2-state model. This is done in Figure 1, where the x-axis corresponds to state 1, and the y-axis to state 2.

The crowding along the axes is alleviated when more states are introduced, and by the time one has 11 states the separation of letters is quite good.

V. The Information in English

Some time ago Shannon [6] investigated the entropy of English and found that in a certain sense the information rate of English is very close to one bit per letter.

TABLE III

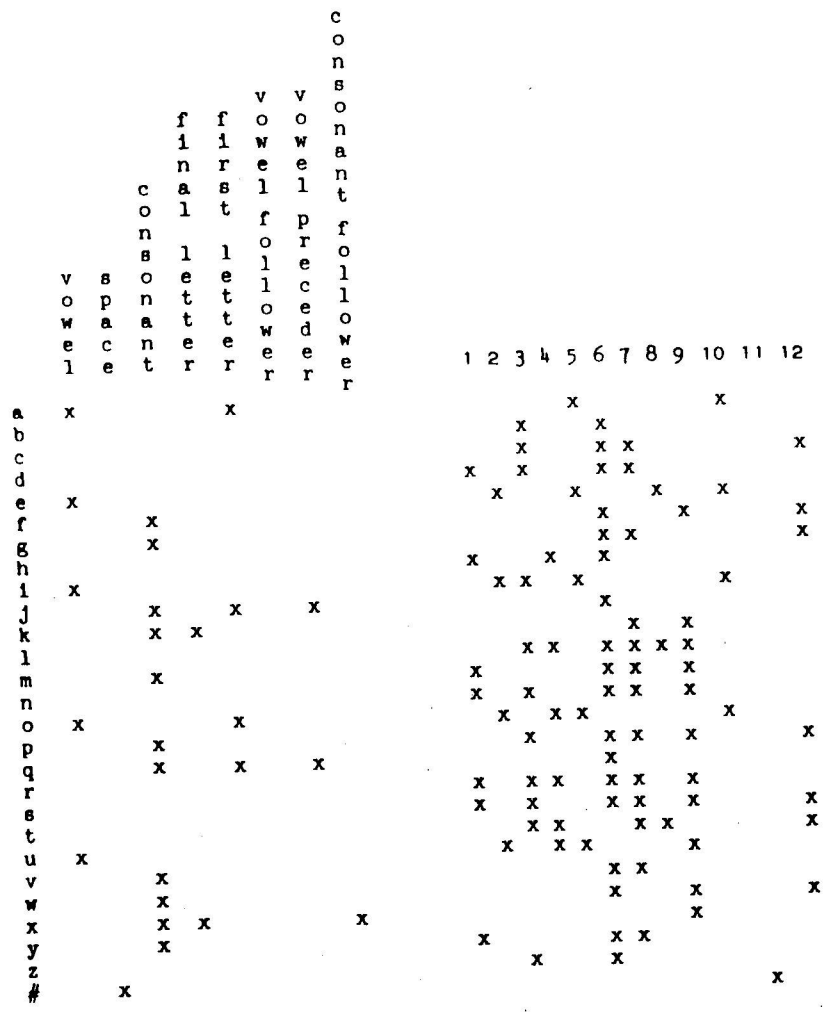
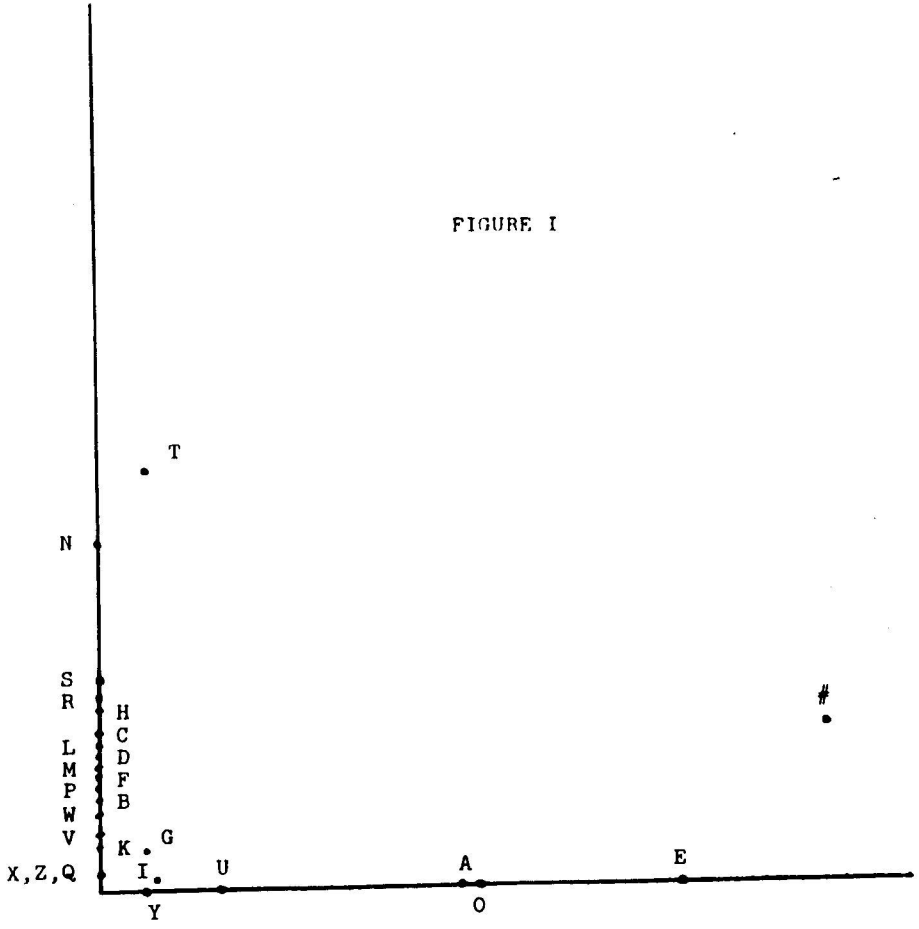


FIGURE I



This may be expressed by saying that, given a long sequence of English text, there are, on average, about two possibilities for the very next letter. Since entropy is usually measured in $\log_2$ the entropy of English is then about 1. It is possible, using the following formulas, to compare the information (or resolution of uncertainty) in our models with Shannon's estimate: We let $H(X)$ denote the uncertainty (entropy) in the Hidden Markov Chain, $H(Y)$ the uncertainty in the output of our model for English, $H(X|Y)$ the uncertainty in the Hidden Markov Chain given a sequence of text, $H(Y|X)$ the uncertainty in the text given the Hidden Markov Chain, and $H(X,Y)$ the uncertainty in the joint process. The following formulas are valid:

$$H(X|Y) = H(X, Y) - H(Y) \tag{1}$$

$$H(Y|X) = H(X, Y) - H(X) \ . \tag{2}$$

Subtracting gives

$$H(Y) = H(Y|X) + H(X) - H(X|Y) \ .$$

Now each term on the right may be calculated independently (the first two precisely, and the third estimated). When this is done we may compare the uncertainty in our

models with the uncertainty as Shannon estimates it. This is shown in Table III.

TABLE III

States	H(X)	H(Y|X)	H(X|Y)	H(Y)
2	.8254	3.2890	.1657	3.9486
3	.8875	3.1266	.6199	3.3942
4	1.0839	2.7060	.6028	3.1871
5	1.2629	2.5332	.6874	3.1087
6	1.2066	2.4868	.8757	2.8176
7	1.1851	2.4433	1.0312	2.5973
8	1.5306	2.2143	.9909	2.7540
9	1.5488	2.1497	.9757	2.7198
10	1.6750	2.0766	1.1791	2.5834
11	1.8025	1.8415	1.0082	2.6358
12	1.8416	1.8550	1.2138	2.4828
		Shannon's Estimate		1.0

We may note how the entropy for our models quickly falls below that for monographic and digraphic English (Table V). This is probably due to the extra structure we have imposed. H(X), of course, increases as does H(X|Y) and we will discuss this later.

In addition to this comparison, we may examine the uncertainty associated with each of the states of our Hidden Markov Chain. This uncertainty is of two sorts. There is uncertainty about the next state, and there is uncertainty about the output given a state. If we consider the ith state, then we let $H_1(X)$ denote the entropy of the former

and $H_1(Y|X)$ denote the entropy of the latter. These are displayed in the following tables.

These figures may be compared with entropy for forward and backward digraphic English. Specifically, we suppose English is sampled from a Markov chain of order 1 with 27 states. The states are the letters themselves and the transition probability $a_{\alpha\beta}$ is the frequency of the digraph α, β in English divided by the frequency of the letter α in the forward case and the letter β in the backward case. Thus, for example, the digraph ab is the b state followed by the a state in the backward case and the reverse in the forward case. The entropy in each of these cases may be calculated. Table V is the result.

The entropy of each letter is indicative of the uncertainty of the succeeding letter in the forward case and the preceding letter in the backward case. From this point of view (and as any school child knows) there is no information (resolution of uncertainty) in letters following q. On the other hand, the first letter of a word resolves the maximum amount of uncertainty since the entropy is maximum for letters following the space state. Thus the first letter of a word, in this sense carries more information than followers of each of the other letters.

TABLE IV

Hidden Markov Entropy

State	$H_1(X)$
1	1.221
2	.906
3	.574
Weighted Average	.8875 = H(X)

Hidden Markov Entropy

State	$H_1(Y\|X)$
1	3.532
2	3.617
3	2.447
Weighted Average	3.127 = H(Y|X)

Hidden Markov Entropy

State	$H_1(X)$
1	.834
2	1.311
3	1.535
4	.473
Weighted Average	1.084 = H(X)

Hidden Markov Entropy

State	$H_1(Y\|X)$
1	.529
2	3.520
3	2.537
4	3.685
Weighted Average	2.706 = H(Y|X)

Hidden Markov Entropy

State	$H_1(X)$
1	.698
2	.745
3	1.940
4	1.751
5	.869
Weighted Average	1.263 = H(X)

Hidden Markov Entropy

State	$H_1(Y\|X)$
1	3.813
2	2.724
3	2.466
4	2.960
5	.505
Weighted Average	2.533 = H(Y|X)

Hidden Markov Entropy

State	$H_1(X)$
1	1.826
2	1.164
3	1.106
4	.662
5	1.049
6	1.148
Weighted Average	1.207 = H(X)

Hidden Markov Entropy

State	$H_1(Y\|X)$
1	3.297
2	2.266
3	3.601
4	2.690
5	2.638
6	.536
Weighted Average	2.487 = H(Y|X)

Hidden Markov Entropy

State	$H_1(X)$
1	.039
2	1.544
3	.191
4	.756
5	1.950
6	1.515
7	1.357
Weighted Average	1.185 = H(X)

Hidden Markov Entropy

State	$H_1(Y\|X)$
1	2.171
2	.571
3	4.032
4	2.608
5	3.303
6	2.543
7	2.252
Weighted Average	2.443 = H(Y|X)

Hidden Markov Entropy

State	$H_1(X)$
1	1.435
2	1.860
3	1.418
4	1.576
5	1.223
6	.478
7	2.152
8	1.490
Weighted Average	1.531 = H(X)

Hidden Markov Entropy

State	$H_1(Y\|X)$
1	1.075
2	2.236
3	2.731
4	.604
5	2.229
6	3.907
7	2.938
8	2.447
Weighted Average	2.214 = H(Y|X)

Hidden Markov Entropy

State	$H_1(X)$
1	2.130
2	.829
3	1.035
4	1.073
5	1.887
6	1.797
7	1.594
8	1.615
9	.704
Weighted Average	1.546 = H(X)

Hidden Markov Entropy

State	$H_1(Y\|X)$
1	2.955
2	3.837
3	1.953
4	1.856
5	1.054
6	.513
7	2.246
8	2.736
9	1.932
Weighted Average	2.150 = H(Y|X)

Hidden Markov Entropy

State	$H_1(X)$
1	1.373
2	1.457
3	.278
4	1.779
5	2.173
6	1.280
7	2.221
8	1.049
9	1.983
10	2.080
Weighted Average	1.676 = H(X)

Hidden Markov Entropy

State	$H_1(Y\|X)$
1	1.871
2	3.951
3	1.666
4	2.806
5	2.740
6	2.315
7	1.820
8	1.633
9	.459
10	1.880
Weighted Average	2.077 = H(Y|X)

Hidden Markov Entropy

State	$H_1(X)$
1	1.799
2	1.840
3	1.222
4	1.624
5	1.663
6	.253
7	2.410
8	.990
9	1.618
10	2.031
11	2.686
Weighted Average	1.803 = H(X)

Hidden Markov Entropy

State	$H_1(Y\|X)$
1	1.503
2	1.188
3	1.882
4	1.287
5	.705
6	2.054
7	2.933
8	3.805
9	1.792
10	.000
11	2.928
Weighted Average	1.842 = H(Y|X)

Hidden Markov Entropy		Hidden Markov Entropy	
State	$H_1(X)$	State	$H_1(Y\|X)$
1	.298	1	1.841
2	2.134	2	1.416
3	1.845	3	2.699
4	2.136	4	1.388
5	1.733	5	1.930
6	1.858	6	3.890
7	2.529	7	2.754
8	1.282	8	.281
9	2.098	9	2.538
10	2.177	10	1.722
11	1.921	11	.053
12	1.089	12	2.129
Weighted Average	1.842 = H(X)	Weighted Average	1.855 = H(Y|X)

TABLE V

Entropy of Monographic English 4.10

Entropy of Forward Digraphic English 3.36

A	3.73	P	3.23
B	2.96	Q	0.00
C	3.25	R	3.53
D	2.43	S	2.94
E	3.43	T	3.09
F	2.71	U	3.58
G	3.07	V	1.66
H	2.27	W	2.93
I	3.56	X	2.64
J	1.93	Y	1.39
K	2.56	Z	1.72
L	3.34	#	4.10
M	3.03		
N	3.38		
O	3.69		

Entropy of Backward Digraphic English 3.36

A	3.814	P	2.760
B	1.826	Q	2.155
C	2.836	R	3.263
D	2.873	S	3.466
E	4.030	T	3.1070
F	2.306	U	3.697
G	2.728	V	2.953
H	2.021	W	1.517
I	3.974	X	1.381
J	1.305	Y	3.413
K	3.2027	Z	.927
L	3.538	#	3.624
M	3.170		
N	2.774		
O	3.911		

This is also reflected in the high entropy for the initial letter state in the Hidden Markov models.

On the other hand, from the backward entropy figures we see that final letters have somewhat lower entropy, and this again is seen in the Hidden Markov model.

One may attempt, using these ideas, to analyze how the information in English is distributed. This is not so interesting in the case of digraph models, for in that case we must confine ourselves to statements about the information in letters following or preceding some specific letter, e.g., letters following e carry more information (resolve more uncertainty) than the letters following the letter v. On the other hand, the Hidden Markov models allow very different and broader sorts of statements. We may examine the trend in uncertainty in the letter produced as the number of states increases. We may track the source of the uncertainty in the letters produced. We may assert for example that while the vowel states have low entropy insofar as the letters produced, they have relatively high entropy insofar as the successor state is concerned. Further, in the 12 state model, for example, there is a post-vowel state with low entropy, thus the high uncertainty in letters following vowels (as seen

from the digraph figures) is traceable partly to the high uncertainty in the identity of the next state. From this we may infer that vowel states are flexible in that they allow great latitude in the types of letter distributions for following letters, viz., final letters, space, vowels themselves, consonants, pre-vowels. The vowels may then be viewed as pivotal letters "permitting" word endings and other types of letter distributions to produce what they may. Perhaps one may think of them as the oil of English.

Gross similarity of h to vowels in the 12 state model (where h dominates state 4) can be observed when we refer to Table IV where we see that in state 4 we have relatively high uncertainty about the next state, and low entropy for the letter produced, and this is a phenomenon shared by vowel states. When we have a model with a small number of states and this phenomenon is not possible we find h in a vowel state.

Let us return to the point raised earlier, that we may study the trend in uncertainty as the number of states increases.

As one might expect, as the number of states increases, the average uncertainty in the letter produced by a state (H(Y|X)) decreases uniformly, as the letters tend to

separate more strongly into the more refined types of states one encounters; thus, in general, states have less and less uncertainty in the letters they produce. However, as the number of states increases, the uncertainty in the next state following a state (H(X)) would be expected to increase since there are more states to go to. This is borne out, but the increase is far less than one might naively expect. For example, suppose the number of states is increased from 5 to 10. On the basis explained above, the entropy H(X) (since it is expressed in logarithms to the base 2) should increase by 1. It does not; from Table III we see that it increases by only .41. This indicates in some sense the fact that the models become more meaningful; increasing order is introduced into the models.

BIBLIOGRAPHY

[1] A. A. Markov, An example of statistical investigation in the text of "Eugene Onyegin" illustrating coupling of "tests" in chains, Proc. Acad. Sci. St. Petersburg, 6, VI Ser., Vol. 7, 1913, No. 3, pp. 153-162.

[2] A. A. Markov, Investigation of remarkable cases of dependent "tests", Proc. Acad. Sci. St. Petersburg, 6, VI Ser., Vol. 1, 1907, No. 3, pp. 61-80.

[3] A. A. Markov, Investigation of general cases of "tests", constrained to chains. St. Petersburg, 6, 1910, (Memoirs Acad. Sci. VIII ser., Vol. 25, No. 3).

[4] A. A. Markov, On a case of a "test", constrained to composed chains, Proc. Acad. Sci. St. Petersburg, 6, VI Ser., Vol. 5, 1911, No. 3, pp. 171-186.

[5] A. A. Markov, On "tests" constrained to chains not controlled by events, Proc. Acad. Sci. St. Petersburg, 6, VI Ser., Vol. 6, 1912, No. 8, pp. 551-572.

[6] C. E. Shannon, The Mathematical Theory of Communication, Bell System Technical Journal, July and October, 1948.

[7] Leonard E. Baum and Ted Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Stat., Vol. 37, No. 6, 1966, pp. 1554-1563.

MAXIMUM-LIKELIHOOD ANALYSIS FOR MULTIVARIATE OBSERVATIONS OF MARKOV SOURCES

Louis A. Liporace

ABSTRACT

This paper discusses a rather general statistical model for multivariate random processes. The model describes the observation vectors as noisy multivariate observations of a Markov chain. That is, each observed vector in turn is drawn from one of S continuous multivariate distributions. The sequence of these underlying distributions forms a Markov chain.