

# On the Theory and Applications of Sequence Based Estimation of Independent Binomial Random Variables\*

B. John Oommen<sup>1,\*\*</sup>, Sang-Woon Kim<sup>2</sup>, and Geir Horn<sup>3</sup>

<sup>1</sup> Professor and Fellow of the IEEE, School of Computer Science, Carleton University, Ottawa, Canada : K1S 5B6

oommen@scs.carleton.ca

<sup>2</sup> Senior Member, IEEE, Dept. of Computer Science and Engineering, Myongji University, Yongin, 449-728 Korea

kimsw@mju.ac.kr

<sup>3</sup> SIMULA Research Laboratory, Martin Linges Vei 15-25, Fornebu, Norway

geirho@simula.no

**Abstract.** We re-visit the age-old problem of estimating the parameters of a distribution from its observations. Traditionally, scientists and statisticians have attempted to obtain strong estimates by “extracting” the information contained in the observations taken *as a set*. However, generally speaking, the information contained in the *sequence* in which the observations have appeared, has been ignored - i.e., except to consider dependence information as in the case of Markov models and n-gram statistics. In this paper, we present results which, to the best of our knowledge, are the first reported results, which consider how estimation can be enhanced by utilizing both the information in the observations *and in their sequence of appearance*. The strategy, known as Sequence Based Estimation (SBE) works as follows. We first quickly allude to the results pertaining to computing the Maximum Likelihood Estimates (MLE) of the data when the samples are taken individually. We then derive the corresponding MLE results when the samples are taken two-at-a-time, and then extend these for the cases when they are processed three-at-a-time, four-at-a-time etc. In each case, we also experimentally demonstrate the convergence of the corresponding estimates. We then suggest various avenues for future research, including those by which these estimates can be fused to yield a superior overall cumulative estimate of the parameter

---

\* The work of the first author was done while visiting at Myongji University, Yongin, Korea. The first author was partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada, a grant from the Korean Research Foundation, and a grant from the SIMULA Research Laboratory in Norway. This work was generously supported by the Korea Research Foundation Grant funded by the Korea Government(MOEHRD-KRF-2005-D00004).

\*\* The first author dedicates this paper to the memory of Mr. Sigurd Bratlie and Mrs. Rakel Bratlie, whose lives were instrumental in changing his life - in every sense of the word. “Thanks, Brother and Sister Bratlie. *What do I have that I did not receive?*”

of the distribution. We believe that our new estimates have great potential for practitioners, especially when the cardinality of the observation set is *small*.

## 1 Introduction

Estimation is a fundamental issue that concerns every statistical problem. Typically, the practitioner is given a set of observations involving the random variable, and his task is to estimate the parameters which govern the generation of these observations. Since, by definition, the problem involves random variables, decisions or predictions related to the problem are in some way dependent on the practitioner obtaining reliable estimates on the parameters that characterize the underlying random variable. Thus, if a problem can be modelled using a random variable which is binomially (or multinomially) distributed, the underlying statistical problem involves estimating the binomial (or multinomial) parameter(s) of the underlying distribution.

The theory of estimation has been studied for hundreds of years [1,3,8]. It is also easy to see that the learning (training) phase of a statistical pattern recognition system is, indeed, based on estimation theory [4,7,18]. Estimation methods generally fall into various categories, including the Maximum Likelihood Estimates (MLE) and the Bayesian family of estimates [1,3,4] which are well-known for having good computational and statistical properties.

To explain the contribution of this paper, we consider the strategy used for developing the MLE of the parameter of a distribution,  $f_X(\theta)$ , whose parameter to be estimated is  $\theta$ . The input to the estimation process is the set of points  $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ , which are assumed to be generated independently and identically as per the distribution,  $f_X(\theta)$ . The process involves deriving the likelihood function, i.e., the likelihood of the distribution,  $f_X(\theta)$ , generating the sample points  $\mathcal{X}$  given  $\theta$ , which is then maximized (by traditional optimization or calculus methods) to yield the estimate,  $\hat{\theta}$ . The general characteristic sought for is that the estimate  $\hat{\theta}_{MLE}$  converges to the true (unknown)  $\theta$  with probability one, or in a mean square sense. Bayesian and MLE estimates generally possess this desirable phenomenon.

Suppose now that the user received  $\mathcal{X}$  as a sequence of data points as in a typical real-life (or real-time) application such as those obtained in a data-mining application involving sequences, or in data involving radio or television news items<sup>1</sup>. The question which we have asked ourselves is the following: “Is there any information in the fact that in  $\mathcal{X}$ ,  $x_i$  specifically precedes  $x_{i+1}$  ?”. Or in a more general case, “Is there any information in the fact that in  $\mathcal{X}$ , the **sequence**  $x_i x_{i+1} \dots x_{i+j}$  occurs  $n_{i,i+1,\dots,i+j}$  times?”. Our position is that even though  $\mathcal{X}$  is generated by an i.i.d. process, there is information in these pieces of sequential information, and we propose here a method by which these pieces

---

<sup>1</sup> We are currently investigating how we can utilize SBEs to yield a superior classification scheme for a real-life problem involving news files from the CBC.

of information can be “maximally” utilized. The estimates that we propose are referred to as the Sequence Based Estimators (SBE).

As far as we know, there are no available results which utilize sequential information in obtaining such estimates<sup>2</sup>. Indeed, even the results we have here are only for the case when the distribution is Binomial. Although some preliminary results for the multinomial case are available, these are merely alluded to. The paper also leads to a host of open problems.

Once we have obtained the SBE estimates based on the occurrence and sequential information, the next question is that of combining all these estimates together to yield a single meaningful estimate. We propose to achieve this by using techniques from the theory of fusion - excellent studies of which are found in [9] and [10]. The paper also includes some specific applications of SBEs.

The paper is organized as follows. Section 2 lists the SBE estimation results obtained when sequential information is used to estimate the parameter of the Binomial distribution, and the sequences are processed two-at-a-time. This is followed in Section 3 by cases when the data is analyzed in sequences three-at-a-time, four-at-a-time respectively. Section 4 discusses the open problems that are currently unsolved, namely those involving the *fusing* of the individual SBEs and those which we have encountered in classifying artificial and real-life data. Section 5 concludes the paper.

**Contributions of the Paper:** The contributions of this paper are:

1. This paper lists the first reported results for obtaining the maximum likelihood estimates (called the Sequence Based Estimates (SBEs)) of the parameter of a binomial distribution when the data is processed both as a *set* of observations and as a *sequence* by which the samples occur in the set.
2. The paper contains the formal results<sup>3</sup> and verification for the cases when the sequence is processed in pairs, and in subsequences of length 3 and 4.
3. The paper lists a few potential strategies by which SBE estimators can be fused to yield a superior estimate utilizing the MLE and the SBEs.
4. The paper lists a few potential schemes by which the MLE and SBE estimators can be used in pattern classification and other applications.

To the best of our knowledge, all of these are novel to the field of estimation, learning and classification.

Throughout this paper we assume that we are estimating the parameters of a binomial distribution. The binomial distribution is characterized by two parameters, namely, the *number* of Bernoulli trials, and the parameter characterizing

---

<sup>2</sup> The question of utilizing and estimating sequential information is not entirely new. It has long been used in syntactic pattern recognition, in estimating the bigram and  $n$ -gram probabilities of streams of data and grammatical inference, and in the learning problem associated with modelling channels using Hidden Markov Models [2,4]. But all of these methods further emphasize the dependence between the occurrences. We show that such information can be gleaned even if the occurrences are *independent*.

<sup>3</sup> The paper lists at least 17 results. But as the proofs of many of the theorems are quite similar, the details of the proofs are merely alluded to in the interest of brevity.

each Bernoulli trial. In this regard, we assume that the number of observations is the number of trials. Thus, all we have to do is to estimate the *Bernoulli* parameter for each trial. Thus, in terms of notation, if  $X$  is a binomially distributed random variable, which takes on the value of either ‘1’ or ‘2’<sup>4</sup>, we assume that  $X$  obeys the distribution  $S$ , where  $S = [s_1, s_2]^T$ , where,  $s_1 + s_2 = 1$ , and

$$\begin{aligned} X = '1' & \text{ with probability } s_1 \\ & = '2' \text{ with probability } s_2, \end{aligned}$$

Then, the following elementary result about the MLE of  $S$  is given below.

**Result 1.** Let  $X$  be a binomially distributed random variable, obeying the distribution  $S$ , where  $S = [s_1, s_2]^T$ . If  $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$  is a realization of a sequence of occurrences of  $X$ , where each  $x_i$  is either ‘1’ or ‘2’, the MLE of  $s_i$  is  $\hat{s}_i = \frac{n_i}{N}$ , where  $n_i$  is the number of occurrences of ‘i’ in  $\mathcal{X}$ .  $\square$

**Notation 1:** To be consistent, we introduce the following notation.

- $X$  is a binomially distributed random variable, obeying the distribution  $S$ .
- $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$  is a realization of a sequence of occurrences of  $X$ , where each  $x_i$  is either ‘1’ or ‘2’.
- Let  $\langle j_1 j_2 \dots, j_k \rangle$  be the sequence examined in the set  $\mathcal{X}$ , where each  $j_m$ , ( $1 \leq m \leq k$ ), is either a ‘1’ or ‘2’. Then, the SBE for  $s_i$  obtained by examining the sequence  $\langle j_1 j_2 \dots, j_k \rangle$  will be given by  $\hat{s}_i |_{\langle j_1 j_2 \dots, j_k \rangle}$ .  $\square$

**Example of Notation 1:** The SBE of  $s_1$  obtained by examining all occurrences of the sequence ‘ $\langle 121 \rangle$ ’ will be given by  $\hat{s}_1 |_{\langle 121 \rangle}$ , and the SBE of  $s_2$  obtained by examining all occurrences of the sequence ‘ $\langle 2122 \rangle$ ’ will be given by  $\hat{s}_2 |_{\langle 2122 \rangle}$ . Observe, trivially, that

$$\hat{s}_2 |_{\langle j_1 j_2 \dots, j_k \rangle} = 1 - \hat{s}_1 |_{\langle j_1 j_2 \dots, j_k \rangle}.$$

We shall now derive the explicit form of  $\hat{s}_i |_{\langle j_1 j_2 \dots, j_k \rangle}$  for various instantiations of sequences  $\langle j_1 j_2 \dots, j_k \rangle$ . It is well known that the MLE converges with probability 1 and in the mean square sense to the true underlying parameter. Thus, all the estimates given in the following Sections/subsections converge (w. p. 1, and in the mean square sense) to the true underlying value of the parameter as the number of samples increases.

## 2 SBEs Using Pair-Wise Sequential Information

In this Section, we consider (analytically and experimentally) the estimation of the binomial parameter when we analyze the sequence of information by processing it in pairs. All the proofs of the results in this paper are either merely sketched or omitted in the interest of brevity, but can be found in [11].

**Theorem 1.** *Using Notation 1,  $\hat{s}_1 |_{\langle 11 \rangle}$ , the SBE of  $s_1$  obtained by examining the occurrences of ‘ $\langle 11 \rangle$ ’ is:*

<sup>4</sup> We depart from the traditional notation of the random variable taking values of ‘0’ and ‘1’, so that the notation is consistent when we deal with vectors.

$$\widehat{s}_1 |_{\langle 11 \rangle} = \sqrt{\frac{n_{11}}{N-1}} \quad (1)$$

where  $n_{11}$  is the number of occurrences of ' $\langle 11 \rangle$ ' in  $\mathcal{X}$ .

*Proof.* The number of sequences of length two<sup>5</sup> in  $\mathcal{X}$  is  $N - 1$ . Of these, we observe  $n_{11}$  which have the value ' $\langle 11 \rangle$ '. Consider now a random variable  $\xi_{11}$  which yields the outcome of either obtaining two consecutive 1's or not.  $\xi_{11}$  is a Bernoulli random variable whose distribution is :

$$\begin{aligned} \xi_{11} &= \langle 11 \rangle \text{ with probability } s_1^2 \\ &\neq \langle 11 \rangle \text{ with probability } 1 - s_1^2. \end{aligned}$$

The MLE of the Bernoulli parameter of  $\xi_{11}$  is  $\frac{n_{11}}{N-1}$ , and thus,

$$\widehat{s}_1^2 = \frac{n_{11}}{N-1}$$

whence  $\widehat{s}_1 |_{\langle 11 \rangle} = \sqrt{\frac{n_{11}}{N-1}}$  and the result follows.  $\square$

**Theorem 2.** Using Notation 1,  $\widehat{s}_1 |_{\langle 22 \rangle}$ , the SBE of  $s_1$  obtained by examining the occurrences of ' $\langle 22 \rangle$ ' is:

$$\widehat{s}_1 |_{\langle 22 \rangle} = 1 - \sqrt{\frac{n_{22}}{N-1}} \quad (2)$$

where  $n_{22}$  is the number of occurrences of ' $\langle 22 \rangle$ ' in  $\mathcal{X}$ .

*Proof.* The proof is similar to the proof of Theorem 1, except that we first solve for  $\widehat{s}_2 |_{\langle 22 \rangle}$  and then obtain  $\widehat{s}_1 |_{\langle 22 \rangle}$ . The details are omitted.  $\square$

**Theorem 3.** Using Notation 1,  $\widehat{s}_1 |_{\langle 12 \rangle}$  and  $\widehat{s}_1 |_{\langle 21 \rangle}$ , the SBEs of  $s_1$  obtained by examining the occurrences of ' $\langle 12 \rangle$ ' and ' $\langle 21 \rangle$ ', respectively, can be obtained if and only if the roots of the quadratic equation given below are :

1.  $\widehat{s}_1 |_{\langle 12 \rangle}$  is the real root of  $\lambda^2 - \lambda + \frac{n_{12}}{N-1} = 0$  whose value is closest to  $\widehat{s}_1$ .
2.  $\widehat{s}_1 |_{\langle 21 \rangle}$  is the real root of  $\lambda^2 - \lambda + \frac{n_{21}}{N-1} = 0$  whose value is closest to  $\widehat{s}_1$ .

*Proof.* The proof of the result is found in [11].  $\square$

A simple study of the patterns<sup>6</sup> that can occur will demonstrate that  $n_{21}$  differs from  $n_{12}$  by at most unity. Thus, the corresponding estimates  $\widehat{s}_1 |_{\langle 21 \rangle}$  and  $\widehat{s}_1 |_{\langle 12 \rangle}$  are almost the same. The ensemble estimates are, however, different.

**Example:** Let us suppose that  $\mathcal{X}$  is the set

{1, 2, 1, 2, 1, 1, 1, 2, 2, 1, 1} where the elements occur in the specified order.

Then :  $n_1 = 7, n_{11} = 3, n_{12} = 3, n_{21} = 3$ , and  $n_{22} = 1$ . Thus, the SBEs are:

$$\widehat{s}_1 |_{\langle 1 \rangle} = \frac{7}{11} = 0.6364.$$

<sup>5</sup> The number of *distinct* sequences of length two is  $\frac{N}{2}$ . But since the elements of  $\mathcal{X}$  are drawn independently and identically, there are  $N - 1$  consecutive pairs ("drawn with replacement") to be considered. More details of this are found in [11].

<sup>6</sup> The proverb "What goes up must come down !" is applicable here.

$$\hat{s}_1 |_{\langle 11 \rangle} = \sqrt{\frac{3}{10}} = 0.5477.$$

$\hat{s}_1 |_{\langle 12 \rangle} = \text{Root}(\lambda^2 - \lambda + \frac{3}{10} = 0)$ . In this case, the roots of the quadratic are complex. Hence the quantity  $n_{12}$  can provide us no information about  $s_1$ . Similarly, the quantity  $n_{12}$  also leads to complex roots and so can provide us no information about  $s_1$ .

$$\text{Finally, } \hat{s}_2 |_{\langle 22 \rangle} = \sqrt{\frac{1}{10}} = 0.3162, \text{ and hence, } \hat{s}_1 |_{\langle 22 \rangle} = 0.6838. \quad \square$$

**Experimental Results:** We present the results of our simulations<sup>7</sup> on synthetic data for the case when the sequence is processed in pairs. The SBE process for the estimation of the parameters for binomial random variables was extensively tested for numerous distributions, but in the interest of brevity, we merely cite one specific example. Also, to make the comparison meaningful, we have followed the “traditional” MLE computation (i.e., the one which does not utilize the sequential information) using the identical data stream. In each case, the estimation algorithms were presented with random occurrences of the variables for  $N = 1, 953, 125$  (i.e.,  $5^9$ ) time instances.

In the case of the SBE, the true underlying value of  $s_1$  was computed using each of the estimates,  $\hat{s}_1 |_{\langle 11 \rangle}$ ,  $\hat{s}_1 |_{\langle 12 \rangle}$ ,  $\hat{s}_1 |_{\langle 21 \rangle}$  and  $\hat{s}_1 |_{\langle 22 \rangle}$ , and the results are tabulated in Table 1. This table reports the values of the estimates as time progresses. However, to demonstrate the true convergence properties of the estimates, we have also reported the values of the ensemble averages of the estimates in Table 1, taken over an ensemble of 100 experiments, which are given in the second line of each row. The convergence of every single estimate is remarkable.

The reader should observe that the MLE and SBE taken for a *single* experiment are much more sporadic. This can be observed from Table 1. It is here that we believe that the SBE will find its niche, namely to enhance the MLE estimate using the information gleaned from the various SBEs.

### 3 SBEs Using Subsequences of Length Three and Four

We first consider the case when subsequences of length 3 are processed. The following results, whose proofs are found in [11], are true.

**Theorem 4.** *Using Notation 1,  $\hat{s}_1 |_{\langle 111 \rangle}$ , the SBE of  $s_1$  obtained by examining the occurrences of ‘ $\langle 111 \rangle$ ’ is:*

$$\hat{s}_1 |_{\langle 111 \rangle} = \sqrt[3]{\frac{n_{111}}{N-2}} \quad (3)$$

where  $n_{111}$  is the number of occurrences of ‘ $\langle 111 \rangle$ ’ in  $\mathcal{X}$ . □

**Theorem 5.** *Using Notation 1,  $\hat{s}_1 |_{\langle 222 \rangle}$ , the SBE of  $s_1$  obtained by examining the occurrences of ‘ $\langle 222 \rangle$ ’ is:*

<sup>7</sup> In the tables, values of *unity* or *zero* represent the cases when the roots are complex or when the number of occurrences of the event concerned are zero.

**Table 1.** A table of the value of the MLE,  $\hat{s}_1$ , and the SBEs  $\hat{s}_1 |_{\langle 11 \rangle}$ ,  $\hat{s}_1 |_{\langle 22 \rangle}$ ,  $\hat{s}_1 |_{\langle 12 \rangle}$ , and  $\hat{s}_1 |_{\langle 21 \rangle}$ , at time ‘ $N$ ’, where the latter SBEs were estimated by using the results of Theorems 1, 2, and 3 respectively. The values of the second line of each row are the ensemble averages of the corresponding estimates, taken over an ensemble of 100 experiments.

$N$	$\hat{s}_1$	$\hat{s}_1  _{\langle 11 \rangle}$	$\hat{s}_1  _{\langle 22 \rangle}$	$\hat{s}_1  _{\langle 12 \rangle}$	$\hat{s}_1  _{\langle 21 \rangle}$
$5^1$ (5)	0.8000	0.8660	1.0000	1.0000	0.5000
	0.7300	0.6686	0.8539	0.6000	0.6150
$5^2$ (25)	0.8000	0.8165	0.7959	0.8536	0.7887
	0.7212	0.7240	0.7267	0.6744	0.6816
$5^3$ (125)	0.7920	0.8032	0.7800	0.8111	0.8111
	0.7210	0.7215	0.7213	0.7121	0.7132
$5^4$ (625)	0.7456	0.7489	0.7375	0.7563	0.7532
	0.7248	0.7237	0.7282	0.7205	0.7206
$5^5$ (3,125)	0.7200	0.7226	0.7143	0.7277	0.7277
	0.7244	0.7240	0.7254	0.7231	0.7231
$5^6$ (15,625)	0.7199	0.7210	0.7171	0.7234	0.7233
	0.7246	0.7245	0.7249	0.7243	0.7243
$5^7$ (78,125)	0.7245	0.7244	0.7248	0.7242	0.7241
	0.7249	0.7248	0.7249	0.7248	0.7248
$5^8$ (390,625)	0.7252	0.7253	0.7250	0.7255	0.7255
	0.7250	0.7250	0.7249	0.7250	0.7250
$5^9$ (1,953,125)	0.7244	0.7243	0.7245	0.7242	0.7242
	0.7250	0.7250	0.7250	0.7250	0.7250

$$\hat{s}_1 |_{\langle 222 \rangle} = 1 - \sqrt[3]{\frac{n_{222}}{N-2}} \quad (4)$$

where  $n_{222}$  is the number of occurrences of ‘ $\langle 222 \rangle$ ’ in  $\mathcal{X}$ .  $\square$

**Theorem 6.** Using Notation 1, the SBEs of  $s_1$  obtained by examining the occurrences of subsequences which contain a single ‘2’ such as ‘ $\langle 211 \rangle$ ’, ‘ $\langle 121 \rangle$ ’, and ‘ $\langle 112 \rangle$ ’, can be obtained as the real root of the cubic equation given below :

1.  $\hat{s}_1 |_{\langle 211 \rangle}$  is the real root of  $\lambda^3 - \lambda^2 + \frac{n_{211}}{N-2} = 0$  whose value is closest to  $\hat{s}_1$ .
2.  $\hat{s}_1 |_{\langle 121 \rangle}$  is the real root of  $\lambda^3 - \lambda^2 + \frac{n_{121}}{N-2} = 0$  whose value is closest to  $\hat{s}_1$ .
3.  $\hat{s}_1 |_{\langle 112 \rangle}$  is the real root of  $\lambda^3 - \lambda^2 + \frac{n_{112}}{N-2} = 0$  whose value is closest to  $\hat{s}_1$ .  $\square$

**Theorem 7.** Using Notation 1, the SBEs of  $s_1$  obtained by examining the occurrences of subsequences which contain two ‘2’s such as ‘ $\langle 122 \rangle$ ’, ‘ $\langle 212 \rangle$ ’, and ‘ $\langle 221 \rangle$ ’, can be obtained as the real root of the cubic equation given below :

1.  $\hat{s}_2 |_{\langle 122 \rangle}$  is the real root of  $\lambda^3 - \lambda^2 + \frac{n_{122}}{N-2} = 0$  whose value is closest to  $\hat{s}_2$ , whence the estimate  $\hat{s}_1 |_{\langle 122 \rangle}$  can be obtained as  $1 - \hat{s}_2 |_{\langle 122 \rangle}$ .

2.  $\widehat{s}_2 |_{\langle 212 \rangle}$  is the real root of  $\lambda^3 - \lambda^2 + \frac{n_{212}}{N-2} = 0$  whose value is closest to  $\widehat{s}_2$ , whence the estimate  $\widehat{s}_1 |_{\langle 212 \rangle}$  can be obtained as  $1 - \widehat{s}_2 |_{\langle 212 \rangle}$ .
3.  $\widehat{s}_2 |_{\langle 221 \rangle}$  is the real root of  $\lambda^3 - \lambda^2 + \frac{n_{221}}{N-2} = 0$  whose value is closest to  $\widehat{s}_2$ , whence the estimate  $\widehat{s}_1 |_{\langle 221 \rangle}$  can be obtained as  $1 - \widehat{s}_2 |_{\langle 221 \rangle}$ .  $\square$

**Experimental Results:** We now present the results of our simulations on synthetic data for the cases studied in the previous sub-section, namely for the case when the sequence is processed in subsequences of length three. To make the comparison (with the pairwise computation) meaningful, we again report the result when the true value of  $s_1$  is 0.725.

In the case of the SBE, the true underlying value of  $s_1$  was computed using each of the estimates, SBEs  $\widehat{s}_1 |_{\langle 111 \rangle}$ ,  $\widehat{s}_1 |_{\langle 222 \rangle}$ ,  $\widehat{s}_1 |_{\langle 211 \rangle}$ ,  $\widehat{s}_1 |_{\langle 121 \rangle}$ ,  $\widehat{s}_1 |_{\langle 112 \rangle}$ ,  $\widehat{s}_1 |_{\langle 122 \rangle}$ ,  $\widehat{s}_1 |_{\langle 212 \rangle}$ , and  $\widehat{s}_1 |_{\langle 221 \rangle}$ , and the results are tabulated in Table 2 as a function of the number of samples processed.

Again, the reader should observe that the MLE and SBE taken for a *single* experiment are not as smooth - especially when the number of samples processed is small. This can be observed from Table 2. In practice, this is augmented by the fact that the SBEs sometimes lead to complex solutions or to unrealistic solutions when the number of samples processed is small. But fortunately, things “average” out as time proceeds.

**Table 2.** A table of the value of the MLE,  $\widehat{s}_1$ , and the SBEs  $\widehat{s}_1 |_{\langle 111 \rangle}$ ,  $\widehat{s}_1 |_{\langle 222 \rangle}$ ,  $\widehat{s}_1 |_{\langle 211 \rangle}$ ,  $\widehat{s}_1 |_{\langle 121 \rangle}$ ,  $\widehat{s}_1 |_{\langle 112 \rangle}$ ,  $\widehat{s}_1 |_{\langle 122 \rangle}$ ,  $\widehat{s}_1 |_{\langle 212 \rangle}$ , and  $\widehat{s}_1 |_{\langle 221 \rangle}$ , at time ‘ $N$ ’, where the latter SBEs were estimated by using the results of Theorems 4, 5, 6, and 7, respectively. The values of the second line on each row mean the ensemble averages of the corresponding estimates, taken over an ensemble of 100 experiments.

$N$	$\widehat{s}_1$	$\widehat{s}_1  _{\langle 111 \rangle}$	$\widehat{s}_1  _{\langle 222 \rangle}$	$\widehat{s}_1  _{\langle 211 \rangle}$	$\widehat{s}_1  _{\langle 121 \rangle}$	$\widehat{s}_1  _{\langle 112 \rangle}$	$\widehat{s}_1  _{\langle 122 \rangle}$	$\widehat{s}_1  _{\langle 212 \rangle}$	$\widehat{s}_1  _{\langle 221 \rangle}$
$5^1$	0.8000	0.8736	1.0000	0	1.0000	1.0000	1.0000	1.0000	1.0000
	0.7300	0.5154	0.9497	0.4700	0.5300	0.4900	0.9100	0.9100	0.9500
$5^2$	0.8000	0.8050	1.0000	0	0.8903	0.7921	0.7610	1.0000	0.7610
	0.7212	0.7173	0.8359	0.4769	0.5724	0.4629	0.7237	0.8299	0.7189
$5^3$	0.7920	0.7958	0.7989	0.7083	0.8556	0.7083	0.7703	0.9052	0.7703
	0.7210	0.7199	0.7530	0.4534	0.4906	0.4351	0.7161	0.7305	0.7158
$5^4$	0.7456	0.7521	0.7178	0.7697	0.7627	0.7697	0.7510	0.7459	0.7510
	0.7248	0.7226	0.7316	0.4653	0.4474	0.4723	0.7273	0.7253	0.7271
$5^5$	0.7200	0.7229	0.7102	0.7260	0.7503	0.7260	0.7173	0.7282	0.7173
	0.7244	0.7236	0.7251	0.5607	0.4780	0.5607	0.7258	0.7247	0.7258
$5^6$	0.7199	0.7231	0.7133	0.7443	0.7318	0.7447	0.7200	0.7137	0.7200
	0.7246	0.7244	0.7247	0.7076	0.6717	0.7076	0.7251	0.7250	0.7251
$5^7$	0.7245	0.7253	0.7232	0.7341	0.7192	0.7342	0.7260	0.7199	0.7260
	0.7249	0.7248	0.7245	0.7246	0.7234	0.7246	0.7251	0.7248	0.7251
$5^8$	0.7252	0.7256	0.7254	0.7296	0.7274	0.7296	0.7247	0.7238	0.7247
	0.7250	0.7250	0.7249	0.7251	0.7250	0.7251	0.7250	0.7249	0.7250
$5^9$	0.7244	0.7243	0.7245	0.7239	0.7233	0.7239	0.7245	0.7243	0.7245
	0.7250	0.7250	0.7249	0.7252	0.7249	0.7252	0.7250	0.7249	0.7250

We now extend the previous cases to consider the scenario when the sequential information is processed in subsequences of length four.



**Theorem 8.** Using Notation 1,  $\widehat{s}_1 |_{\langle 1111 \rangle}$ , the SBE of  $s_1$  obtained by examining the occurrences of ' $\langle 1111 \rangle$ ' is:

$$\widehat{s}_1 |_{\langle 1111 \rangle} = \sqrt[4]{\frac{n_{1111}}{N-3}} \quad (5)$$

where  $n_{1111}$  is the number of occurrences of ' $\langle 1111 \rangle$ ' in  $\mathcal{X}$ .  $\square$

**Theorem 9.** Using Notation 1,  $\widehat{s}_1 |_{\langle 2222 \rangle}$ , the SBE of  $s_1$  obtained by examining the occurrences of ' $\langle 2222 \rangle$ ' is:

$$\widehat{s}_1 |_{\langle 2222 \rangle} = 1 - \sqrt[4]{\frac{n_{2222}}{N-2}} \quad (6)$$

where  $n_{2222}$  is the number of occurrences of ' $\langle 2222 \rangle$ ' in  $\mathcal{X}$ .  $\square$

To simplify matters, we deal with the rest of the cases that involve four-at-a-time subsequences, by sub-dividing them into the cases when the subsequences contain *one* '2', *two* '2's, and *three* '2's, respectively. In each case, we shall deal with all the corresponding subsequence patterns in a single theorem.

**Theorem 10.** Using Notation 1, the SBEs of  $s_1$  obtained by examining the occurrences of subsequences which contain a single '2', can be obtained by the real roots (if any) of the quartic equations given below :

1.  $\widehat{s}_1 |_{\langle 2111 \rangle}$  is the real root of  $\lambda^4 - \lambda^3 + \frac{n_{2111}}{N-3} = 0$  whose value is closest to  $\widehat{s}_1$ .
2.  $\widehat{s}_1 |_{\langle 1211 \rangle}$  is the real root of  $\lambda^4 - \lambda^3 + \frac{n_{1211}}{N-3} = 0$  whose value is closest to  $\widehat{s}_1$ .
3.  $\widehat{s}_1 |_{\langle 1121 \rangle}$  is the real root of  $\lambda^4 - \lambda^3 + \frac{n_{1121}}{N-3} = 0$  whose value is closest to  $\widehat{s}_1$ .
4.  $\widehat{s}_1 |_{\langle 1112 \rangle}$  is the real root of  $\lambda^4 - \lambda^3 + \frac{n_{1112}}{N-3} = 0$  whose value is closest to  $\widehat{s}_1$ .  $\square$

**Theorem 11.** Using Notation 1, the SBEs of  $s_1$  obtained by examining the occurrences of subsequences which contain two '2's, can be obtained by the real roots (if any) of the quadratic (not quartic !!!) equations given below :

1.  $\widehat{s}_1 |_{\langle 1122 \rangle}$  is the real root of  $\lambda^2 - \lambda + \sqrt{\frac{n_{1122}}{N-3}} = 0$  with value closest to  $\widehat{s}_1$ .
2.  $\widehat{s}_1 |_{\langle 1212 \rangle}$  is the real root of  $\lambda^2 - \lambda + \sqrt{\frac{n_{1212}}{N-3}} = 0$  with value closest to  $\widehat{s}_1$ .
3.  $\widehat{s}_1 |_{\langle 1221 \rangle}$  is the real root of  $\lambda^2 - \lambda + \sqrt{\frac{n_{1221}}{N-3}} = 0$  with value closest to  $\widehat{s}_1$ .
4.  $\widehat{s}_1 |_{\langle 2112 \rangle}$  is the real root of  $\lambda^2 - \lambda + \sqrt{\frac{n_{2112}}{N-3}} = 0$  with value closest to  $\widehat{s}_1$ .
5.  $\widehat{s}_1 |_{\langle 2121 \rangle}$  is the real root of  $\lambda^2 - \lambda + \sqrt{\frac{n_{2121}}{N-3}} = 0$  with value closest to  $\widehat{s}_1$ .
6.  $\widehat{s}_1 |_{\langle 2211 \rangle}$  is the real root of  $\lambda^2 - \lambda + \sqrt{\frac{n_{2211}}{N-3}} = 0$  with value closest to  $\widehat{s}_1$ .  $\square$

**Theorem 12.** Using Notation 1, the SBEs of  $s_1$  obtained by examining the occurrences of subsequences which contain three '2's, can be obtained by determining the real roots (if any) of the quartic equations given below and then subtracting their value from unity as below:

1.  $\widehat{s}_1 |_{\langle 1222 \rangle}$  is the quantity  $[1 - \text{Root}(\lambda^4 - \lambda^3 + \frac{n_{1222}}{N-3} = 0)]$
2.  $\widehat{s}_1 |_{\langle 2122 \rangle}$  is the quantity  $[1 - \text{Root}(\lambda^4 - \lambda^3 + \frac{n_{2122}}{N-3} = 0)]$
3.  $\widehat{s}_1 |_{\langle 2212 \rangle}$  is the quantity  $[1 - \text{Root}(\lambda^4 - \lambda^3 + \frac{n_{2212}}{N-3} = 0)]$
4.  $\widehat{s}_1 |_{\langle 2221 \rangle}$  is the quantity  $[1 - \text{Root}(\lambda^4 - \lambda^3 + \frac{n_{2221}}{N-3} = 0)]$

where each of the above estimates is the value closest to  $\widehat{s}_1$ . □

**Experimental Results:** The simulation results for the the case when the sequence is processed in subsequences of length four is presented below. The experimental settings are identical to the ones used in the case of processing it in pairs and in subsequences of length three, namely, when  $s_1$  is 0.725, and  $N = 1, 953, 125$  (i.e,  $5^9$ ) time instances.

Table 3 lists the values of the SBEs, computed using each of the estimates,  $\widehat{s}_1 |_{\langle 1111 \rangle}$ ,  $\widehat{s}_1 |_{\langle 2222 \rangle}$ ,  $\widehat{s}_1 |_{\langle 2111 \rangle}$ ,  $\widehat{s}_1 |_{\langle 1122 \rangle}$ , and  $\widehat{s}_1 |_{\langle 1222 \rangle}$ , and their ensemble averages. The other cases when the subsequences with *one* ‘2’, *two* ‘2’s, and *three* ‘2’s (the other cases listed in Theorems 10, 11 and 12) are identical to the ones reported and so omit them here for ease of readability.

Again, we observe that the convergence of every single estimate is remarkable. For example, the traditional MLE,  $\widehat{s}_1$ , had the ensemble average of 0.7210 when only  $N = 125$  symbols were processed. This value became 0.7248 when  $N = 625$  symbols were processed, which converged to 0.7250 when  $N = 5^9$ . By way of comparison, for the same case, the SBE,  $\widehat{s}_1 |_{\langle 1222 \rangle}$ , had the ensemble average of 0.7444 when only  $N = 125$  symbols were processed. It had the value 0.7319 after  $N = 625$  symbols were processed, and as in the case of  $\widehat{s}_1$  became increasingly closer to the true value as  $N$  increased. In this case, when  $N = 5^9$ , the value of  $\widehat{s}_1 |_{\langle 1222 \rangle}$ , was also exactly 0.7250. This was also true for the other SBEs.

In this case, the solutions to the equations were often complex initially (i.e., for small values of ‘N’). But as time proceeded, the number of occurrences of the outcomes was more reasonable, and the solution obtained converged as expected.

## 4 Open Issues and Potential Applications of SBEs

As mentioned earlier, we believe that there are a host of open problems which concern the family of SBEs. We shall highlight them in the following subsections.

**Higher Order SBEs:** Till now, we have considered how we can obtain effective SBEs by considering subsequences of lengths 2, 3 and 4 respectively. There is no reason why we cannot consider subsequences of even longer length. Without much ado, we list (without proof) the form the SBEs would take for a few simple cases when subsequences of length 5 are analyzed. Indeed, using Notation 1, we can state that:

1.  $\widehat{s}_1 |_{\langle 11111 \rangle}$ , the SBE of  $s_1$  obtained by examining the occurrences of ‘ $\langle 11111 \rangle$ ’ is :  $\widehat{s}_1 |_{\langle 11111 \rangle} = \sqrt[5]{\frac{n_{11111}}{N-4}}$ . □

**Table 3.** A table of the values of the MLE,  $\hat{s}_1$ , and the SBEs  $\hat{s}_1 |_{\langle 1111 \rangle}$ ,  $\hat{s}_1 |_{\langle 2222 \rangle}$ ,  $\hat{s}_1 |_{\langle 2111 \rangle}$ ,  $\hat{s}_1 |_{\langle 1112 \rangle}$ , and  $\hat{s}_1 |_{\langle 1222 \rangle}$ , at time ‘ $N$ ’, where the latter SBEs were estimated by using the results of Theorems 8, 9, 10, 11 and 12 respectively. The other cases when the subsequences with *one* ‘2’, *two* ‘2’s, and *three* ‘2’s (the other cases listed in Theorems 10, 11 and 12) are identical.

$N$	$\hat{s}_1$	$\hat{s}_1  _{\langle 1111 \rangle}$	$\hat{s}_1  _{\langle 2222 \rangle}$	$\hat{s}_1  _{\langle 1222 \rangle}$	$\hat{s}_1  _{\langle 1122 \rangle}$	$\hat{s}_1  _{\langle 2111 \rangle}$
$5^1$ (5)	0.8000 0.7300	0.8409 0.3645	1.0000 0.9916	1.0000 0.8800	1.0000 0.8100	1.0000 0.6600
$5^2$ (25)	0.8000 0.7212	0.7765 0.6867	1.0000 0.9428	1.0000 0.8354	0.6918 0.5944	0.6918 0.4880
$5^3$ (125)	0.7920 0.7210	0.7920 0.7178	1.0000 0.8832	0.7811 0.7444	0.7625 0.6577	0.7625 0.3152
$5^4$ (625)	0.7456 0.7248	0.7492 0.7223	0.7619 0.7468	0.6976 0.7319	0.7402 0.7255	0.7659 0.3755
$5^5$ (3,125)	0.7200 0.7244	0.7247 0.7236	0.7412 0.7249	0.6942 0.7261	0.7069 0.7250	0.7143 0.4050
$5^6$ (15,625)	0.7199 0.7246	0.7232 0.7244	0.7090 0.7244	0.7161 0.7250	0.7221 0.7246	0.7196 0.4922
$5^7$ (78,125)	0.7245 0.7249	0.7253 0.7249	0.7205 0.7245	0.7248 0.7245	0.7285 0.7251	0.7277 0.6649
$5^8$ (390,625)	0.7252 0.7250	0.7256 0.7250	0.7257 0.7247	0.7252 0.7250	0.7250 0.7250	0.7257 0.7249
$5^9$ (1,953,125)	0.7244 0.7250	0.7241 0.7250	0.7244 0.7248	0.7246 0.7250	0.7250 0.7251	0.7247 0.7249

2.  $\hat{s}_1 |_{\langle 22222 \rangle}$ , the SBE of  $s_1$  obtained by examining the occurrences of ‘(22222)’, is :  $\hat{s}_1 |_{\langle 22222 \rangle} = 1 - \sqrt[5]{\frac{n_{22222}}{N-4}}$ .  $\square$

We believe that deriving the expressions for other higher order SBEs is not expedient. Obtaining them would involve explicitly solving algebraic equations which are higher than of a *quintic* order, and it is well known that this is intractable (other than by resorting to numerical methods).

We conclude this section by stating that the question of how we can effectively *compute* the SBEs for orders of 5 and higher is still effectively open.

**Fusing the MLE and the SBEs to Yield a Superior Estimate:** One of the most interesting problems that still remains open involves the question of how the MLE and the SBEs can be *fused* to yield a superior estimate. Rather than discuss “specifics”, let us assume that we have obtained a set of estimates  $\Phi = [\phi_0, \phi_1, \phi_2, \dots, \phi_D]^T$ , where, for simplicity, we use the notation that  $\phi_0$  is the traditional MLE, and the other  $\phi_i$ ’s are the SBEs. Thus, for example, an instantiation of  $\Phi$  could be the 7-component vector:

$$\Phi = [\hat{s}_1, \hat{s}_1 |_{\langle 11 \rangle}, \hat{s}_1 |_{\langle 12 \rangle}, \hat{s}_1 |_{\langle 111 \rangle}, \hat{s}_1 |_{\langle 222 \rangle}, \hat{s}_1 |_{\langle 1111 \rangle}, \hat{s}_1 |_{\langle 1222 \rangle}]^T.$$

The aim of the fusing exercise is to combine the information in the components of  $\Phi$  to obtain an even more superior estimate  $\hat{\hat{s}}_1$ .

The first question that needs to be answered is the following: If the traditional MLE and all the SBEs converge to the *same true value*, what is the advantage of such a fusing process? The answer lies simply in the fact that although the traditional MLE and all the SBEs converge *asymptotically* to the same value,

they all have *completely different values*<sup>8</sup> when the number of samples examined is “small”. Thus, for example, when the number of samples examined is only 125 and the true value of  $s_1$  is 0.7250, the value of  $\Phi$  is:

$$\Phi = [0.7920, 0.8032, 0.7800, 0.8111, 0.7958, 0.7989, 0.7920, 0.7811]^T.$$

Observe that the traditional MLE is 0.7920 (quite distant from the true value of 0.7250), while other SBEs are closer to the true value. Thus, it would be advantageous to seek a scheme which uses these different “descriptors” of  $s_1$  to lead to a more accurate estimate of  $s_1$ . In all these estimates, we consciously discard elements of  $\Phi$  which are *unity* or *zero*, as these represent the cases when the solution of the underlying equation don’t lead to a realistic estimate.

We have designed four different fusion methods that use the  $D$  components of  $\Phi$ . The details of the methods are omitted here in the interest of brevity, but can be found in [11], but this entire avenue is open to further research.

**Classification Using the MLE and the SBE:** Another major possibility for further research involves combining the classifier decisions obtained by the MLEs and the various SBEs. In the study of pattern recognition, classifier combination has received considerable attention because of its potential to improve the performance of individual classification systems. The basic idea is to solve each classification problem by designing a *specific* classifier, and then to combine the classifiers in some way to achieve reduced classification error rates. Indeed, the choice of an appropriate fusion method can further improve on the performance of the combination. Various Classifier Fusion Schemes (CFS) have been proposed in the literature - excellent studies are found in [9,?]. We have designed a few different fusion classification methods for SBEs (omitted here in the interest of brevity, but included in [11]), but here too the ground is fertile, and we believe that a lot of research remains to be accomplished.

**Non Pattern Recognition Potential Applications of SBEs:** Apart from the obvious applications in Pattern Recognition (PR) (mentioned above), there are numerous situations where data arrives in a sequence, and where it is possible to assign a binary indicator variable to the arriving data. This section introduces two such example applications that are representative for wide classes of use. We note, in passing, that these methods can be used for even broader application areas when we seek methods to obtain the sequence estimates so as to improve the quality and convergence of lower order estimators. In particular, this may allow the use of estimation in real-time control applications where the convergence of MLEs are too slow compared to the time constant of the system being controlled.

**Network Transmission Quality:** The Internet is omnipresent today, and the Internet Protocol (IP) is the dominating protocol for long haul data communication. However, the IP itself does not provide any transmission guarantees, and

---

<sup>8</sup> This is reminiscent of the fairy tale of the seven blind men who each described an elephant with completely different descriptions. While each of the descriptors was, in itself, inaccurate, the composite picture was accurate!

packets may freely be delayed, delivered out of order, or even discarded in the case of network congestion. To remedy this situation most Internet applications requiring reliable and sequential delivery use the Transmission Control Protocol (TCP) on top of the IP, namely, the TCP/IP [14]. For applications which cannot accept neither long delays nor jitter (delay variation) nor out of order delivery the TCP is not the solution, and various other protocols and mechanisms to achieve the desired transmission quality in the Internet have been proposed [17]. Providing strict transmission quality guarantees per flow can be done, but this requires intelligence installed in every router along the transmission path [15]. Further, if a flow crosses multiple ISPs it will mix with other traffic between the ISPs, and the end-user's SLA with its ISP will not extend to the ISPs further down-stream. Currently, no end-to-end transmission guarantees for a flow in the Internet can be given beyond that of the TCP; this is an active research topic [5].

By categorizing a packet within the delay bound as a *Success* (say, with value '1') and a dropped packet as a *Failure* (with value '2'), the situation fits the SBE framework presented in this paper. The delay sensitive flow can be admitted if the failure rate,  $\hat{s}_2$ , estimated from the probing is below a certain limit. A side effect of using the SBE instead of the MLE would be the advantage of being able to gather statistics on the probability of a *sequence* of subsequent failures. More details of this proposition are found in [11].

**Arithmetic Coding:** Arithmetic coding [6,12] has the ability to serve two purposes: Lossless data compression [16] and the assignment of unique signatures for sets and databases. The fundamental idea is to encode a sequence of symbols taken from an *alphabet* as a sequence of bits, so that more frequently occurring symbols are assigned a lesser number of bits, where the assignment can be achieved either static or adaptive.

Given the encoded bit string the SBE is directly applicable. Using the SBE on the bit stream the decoder might be able to predict the incoming symbol before it has completely arrived. Such a look-ahead would be a result of combining the relative frequencies of '0' and '1' in the stream with the relative occurrence of the *sub-sequences*. The added benefit will be that the encoder can use the information of the symbol it is set to decode immediately, i.e. update the frequency table of the model *before* encoding the symbol. The details of this opportunity, including rendering it less vulnerable to transmission errors (see [11] needs to be further investigated, and is a topic for further research.

## 5 Conclusions

In this paper, we considered the age-old problem of estimating the parameters of a distribution from its observations. Unlike the method that is customarily employed, (which processes the information contained in the observations taken *as a set*), we demonstrate how the estimation can be enhanced by utilizing both the information in the observations *and in their sequence of appearance*. In this

regard, we have derived the corresponding MLE results when the samples are taken two-at-a-time, three-at-a-time, four-at-a-time etc. In each case, we also experimentally demonstrated the convergence of the corresponding estimates. We have visited the various strategies by which these estimates could be fused to yield superior overall cumulative estimates. Our results demonstrate that the strategy is very promising, and that it has potential applications in fused PR systems, and in the domains of enhancing Internet protocols and data encoding.

## References

1. P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume I. Prentice Hall, Second Edition, 2000.
2. H. Bunke. Structural and Syntactic Pattern Recognition. *Handbook of Pattern Recognition and Computer Vision*, World Scientific-25, 1993.
3. G. Casella and R. Berger. *Statistical Inference*. Brooks/Cole Pub. Co., Second Edition, 2001.
4. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, NY, Second Edition, 2000.
5. W.-C. Feng, D. D. Kandlur, D. Saha, and K. G. Shin, "Adaptive packet marking for maintaining end-to-end throughput in a differentiated-services internet," *IEEE/ACM Transactions on Networking*, vol. 7, pp. 685–697, Oct. 1999.
6. J. Glen and G. Langdon, "Arithmetic coding," *IBM Journal of Research and Development*, vol. 28, pp. 135–149, Mar. 1984.
7. R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, Cambridge, Massachusetts, 2001.
8. B. Jones, P. Garthwaite, and Ian Jolliffe. *Statistical Inference*. Oxford University Press, Second Edition, 2002.
9. J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, "On combining classifiers", *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 20, No. 3, pp. 226 - 239, Mar. 1998.
10. L. I. Kuncheva, "A theoretical study on six classifier fusion strategies", *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 24, no. 2, pp. 281 - 286, Feb. 2002.
11. B. J. Oommen, S.-W. Kim, and G. Horn, "On the Estimation of Binomial Random Variables Using Occurrence and Sequential Information". *Unabridged version of this paper*.
12. J. Rissanen and J. Glen G. Langdon, "Arithmetic coding," *IBM Journal of Research and Development*, vol. 23, pp. 149–162, Mar. 1979.
13. A. S. Tanenbaum, *Computer Networks*. New Jersey: Prentice Hall, 4th ed., 2003.
14. U. o. S. C. Information Science Institute, "Transmission control protocol (TCP)." <http://www.ietf.org/rfc/rfc0793.txt>, sep 1981. Clarifications, corrections, extensions and a textbook introduction can be found in [13].
15. P. P. White, "RSVP and integrated services in the internet: A tutorial," *IEEE Communications Magazine*, pp. 100–106, May 1997.
16. I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, pp. 520–540, Jun. 1987.
17. X. Xiao and L. M. Ni, "Internet QoS: A big picture," *IEEE Network*, vol. 13, pp. 8–18, Mar. 1999.
18. A. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, N.York, Second Edition, 2002.