Multinomial Sequence Based Estimation using Contiguous Subsequences of Length Three[⋆]

B. John Oommen[1] and Sang-Woon Kim[2]

[1] *Chancellor's Professor*, School of Computer Science, Carleton University, Ottawa, Canada: K1S 5B6. e-mail address: oommen@scs.carleton.ca.

[2] Dept. of Computer Engineering, Myongji University, Yongin, 17058 South Korea. e-mail address: kimsw@mju.ac.kr

Abstract. The Maximum Likelihood (ML) and Bayesian estimation paradigms work within the model that the data, from which the parameters are to be estimated, is treated as a *set* rather than as a *sequence*. The pioneering paper that dealt with the field of sequence-based estimation [2] involved utilizing both the information in the observations *and in their sequence of appearance*. The results of [2] introduced the concepts of Sequence Based Estimation (SBE) for the Binomial distribution, where the authors derived the corresponding MLE results when the samples are taken two-at-a-time, and then extended these for the cases when they are processed three-at-a-time, four-at-a-time etc. These results were generalized for the multinomial "two-at-a-time" scenario in [3]. This paper[3] now further generalizes the results found in [3] for the multinomial case and for subsequences of length 3. The strategy used in [3] (and also here) involves a novel phenomenon called "Occlusion" that has not been reported in the field of estimation. The phenomenon can be described as follows: By occluding (hiding or concealing) certain observations, we map the estimation problem onto a lower-dimensional space, i.e., onto a binomial space. Once these occluded SBEs have been computed, the overall Multinomial SBE (MSBE) can be obtained by combining these lower-dimensional estimates. In each case, we formally prove and experimentally demonstrate the convergence of the corresponding estimates.

Keywords: *Estimation using Sequential Information, Sequence Based Estimation, Estimation of multinomials, Fused Estimation Methods, Sequential Information.*

[⋆] The first author is a *Fellow: IEEE* and *Fellow: IAPR.* The work was done while he was visiting at Myongji University, Yongin, Korea. He also holds an *Adjunct Professorship* with the Department of Information and Communication Technology, University of Agder, Grimstad, Norway. The work was partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada and a grant from the National Research Foundation of Korea. This work was also generously supported by the National Research Foundation of Korea funded by the Korean Government (NRF-2012R1A1A2041661).

[3] *This paper is dedicated to the memory of Dr. Mohamed Kamel, who was a close friend of the first author.*

1 Introduction

The theory of estimation has been studied for hundreds of years [5,6,7], and it has been the backbone for the learning (training) phase of statistical pattern recognition systems [1,8,9]. Traditionally, the ML and Bayesian estimation paradigms work within the model that the data, from which the parameters are to be estimated, is known, and that it is treated as a *set*. The position that we respectfully submit is that traditional ML and Bayesian methods ignore and discard[4] valuable *sequence*-based information. The goal of this paper is to "extract" and "utilize" the information contained in the observations when they are perceived *both as a set* and in *their sequence of appearance*. Put in a nutshell, this paper deals with the relatively new field of sequence-based estimation in which the goal is to estimate the parameters of a distribution by maximally "squeezing" out the *set*-based and *sequence*-based information latent in the observations.

The Maximum Likelihood (ML) and Bayesian estimation paradigms work within the model that the data, from which the parameters are to be estimated, is treated as a *set* rather than as a *sequence*. The pioneering paper that dealt with the field of Sequence-Based Estimation (SBE) [2] involved utilizing both the information in the observations *and in their sequence of appearance*. The question that this entails 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** $x_i x_{i+1} \ldots x_{i+j}$ occurs $n_{i,i+1,\ldots i+j}$ times?". Our position, which we proved in [2] for binomial random variables[5], is that even though $\mathcal{X}$ is generated by an i.i.d. process, there is information in these pieces of sequential data which can be "maximally" utilized to yield the so-called family of SBEs.

If the MLE and any SBE of the parameter θ converge to the *same true, unknown, value*, what then is the advantage of having multiple estimates? The answer lies simply in the fact that although the traditional MLE and the SBEs converge *asymptotically* to the same value, they all have *completely different* values. This is all the more true because the information used in procuring each of these estimates is "orthogonal". Further, since the convergence properties of MLEs is asymptotic, one can glean and effectively utilize other information when the number of samples examined is "small".

The consequences of invoking SBEs are potentially many. If we are able to obtain reliable estimates of the parameters under investigation by utilizing the set-based *and* sequence-based information, this could potentially have advantages in all the fields where estimation is used.

The pioneering paper concerning SBEs [2] introduced its theory, experimental results and applications for the Binomial distribution, where the authors derived

[4] This information is, of course, traditionally used when we want to consider *dependence* information, as in the case of Markov models and n-gram statistics.

[5] The papers [2] and [3] explain the application of SBEs, and also about how we can fuse them to yield superior estimates. These aspects are not included here in the interest of space.

the corresponding MLE results when the samples are taken two-at-a-time, and then extended these for the cases when they are processed three-at-a-time, four-at-a-time etc. These results were generalized for the *multinomial* "two-at-a-time" scenario in [3].

This paper now further generalizes the latter results (those found in [3]) for the multinomial case and for subsequences of length 3. The results of the case when we deal with subsequences of length greater than 3 are currently being compiled. To the best of our knowledge, apart from our previous results of [2] and [4], all of these are novel to the field of estimation, learning and classification.

In the interest of space and brevity, the proofs of the theoretical results presented here are omitted. They are found in [4]. However, we add that all the theoretical results have been experimentally verified

2 On Obtaining MSBEs Using Occluded SBEs

Informally speaking, the question of designing SBEs for multinomial random variables is, perhaps, "two orders of magnitude" more complex than that of designing them for binomial random variables[6] The reason for this is quite simple: For a vector of dimension d, there are $\binom{d}{2}$ possible pairs of binomial events, and it is no trivial task to generalize the expressions for the binomial SBEs (from [2]) to yield the corresponding multinomial SBE (MSBE). This, we believe, is the hurdle that we have encountered in this present paper, and its solution is the novel contribution.

How then have we proposed the solution to the problem even though we encounter $\binom{d}{2}$ possible pairs? Indeed, rather than consider the problem of computing the MSBE as a problem in its own right, we have shown how we can map this problem into a *linear* set of *Binomial* SBE (BSBE) problems. This is, as we shall see, achieved by effectively occluding (erasing, hiding or concealing) all the observations in the sequence other than the ones that are concerned in the specific binomial experiment. One can now procure corresponding BSBEs from these occluded sequences. The final MSBE result is now computed by effectively processing a sufficient set of such BSBEs, and combining them by means of a normalizing constraint. The details of all these aspects will be explained in the subsequent sections.

2.1 Notation: MSBEs Using Pairs and Subsequences

Before we proceed with the theoretical and experimental results, it is necessary for us to formalize the notation that will be used[7].

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

[6] The contents of this section is quite identical to the corresponding section in [3]. This is unavoidable because the notation is quite cumbersome. Besides, the fundamental theory of using "occlusion" is identical in both the papers. Unfortunately, it is futile to omit these concepts and to refer the reader to [3] - it will render the present paper to be quite incomprehensible.

[7] We apologize for this cumbersome notation, but this is unavoidable considering the complexity of the problem and the ensuing analysis.

- X is a multinomially distributed random variable, obeying the distribution S.
- $\mathcal{X} = \{x_1, x_2, \ldots, x_J\}$ is a realization of a sequence of occurrences of X, where each $x_i \in \mathcal{D}$.
- An index $a \in \mathcal{D}$ is said to be the unconstrained variable in any computation if all the other estimates $\{s_i\}$ are specified in terms of s_a, where $i \neq a$. It will soon be clear that in any computation there can only be *a single* unconstrained variable.
- $\mathcal{X}^{ab} = \{x_1, x_2, \ldots, x_{N_{ab}}\}$ is called the *Occluded* sequence of $\mathcal{X}$ (with N_{ab} items) with respect to a and b, if it is obtained from $\mathcal{X}$ by deleting the occurrences of all the elements except a and b. Whenever we refer to the sequence $\mathcal{X}^{ab} = \{x_1, x_2, \ldots, x_{N_{ab}}\}$, we always imply that the first variable (in this case a) is the unconstrained variable.
- Let $< j_1 j_2 \ldots, j_k >$ be the subsequence[8] examined in the *Occluded* sequence $\mathcal{X}^{ab}$, where each $j_m, (1 \leq m \leq k)$, is either a or b. Then[9]:
  - The BSBE, for s_a obtained by examining in $\mathcal{X}^{ab}$ the subsequence $< j_1 j_2 \ldots, j_k >$ will be given by $\widehat{q}_a\Big|^{ab}_{<j_1 j_2 \ldots, j_k>}$, where, as before, the first variable (in this case a) is the unconstrained variable.
  - Similarly, the BSBE, for s_b obtained by examining in $\mathcal{X}^{ab}$ the subsequence $< j_1 j_2 \ldots, j_k >$ will be given by $\widehat{q}_b\Big|^{ab}_{<j_1 j_2 \ldots, j_k>}$, where the first variable (in this case a) is the unconstrained variable.
- Consider the sequence $\mathcal{X}$ in which the index a is the unconstrained variable. Let $< j_1 j_2 \ldots, j_k >$ be the subsequence examined in the sequence $\mathcal{X}$, where each $j_m, (1 \leq m \leq k)$, is either a or '$*$', where each '$*$' is the *same* variable, say $c \in (\mathcal{D} - \{a\})$. Then:
  - The MSBE for s_a (where a is the unconstrained variable) obtained by examining in $\mathcal{X}$ the sequence $< j_1 j_2 \ldots, j_k >$ will be given by $\widehat{s}_a\Big|^{a}_{<j_1 j_2 \ldots, j_k>}$ where each j_i that is not a is replaced by a '$*$', and where each '$*$' is the *same* variable, say $c \in (\mathcal{D} - \{a\})$.
  - For any constrained variable b, the MSBE for s_b obtained by examining in $\mathcal{X}$ the sequence $< j_1 j_2 \ldots, j_k >$ will be given by $\widehat{s}_b\Big|^{ab}_{<j_1 j_2 \ldots, j_k>}$, where a is the unconstrained variable.
- Trivially, for all a and b:

$$\textstyle\sum_{b \neq a} \widehat{s}_b\Big|^{ab}_{<j_1 j_2 \ldots, j_k>} = 1 \text{ - } \widehat{s}_a\Big|^{a}_{<j_1 j_2 \ldots, j_k>}. \qquad \square$$

A detailed example of Notation 1 is found in [3].

[8] For the present, we consider non-overlapping subsequences. We shall later extend this to overlapping sequences when we report the experimental results.

[9] The reader must take pains to differentiate between the q's and the s's, because the former refer to the BSBEs and the latter to the MSBEs.

For any given a and b, if a is the unconstrained variable, we shall now derive the explicit form of $\widehat{q}_a\Big|^{ab}_{<j_1j_2\ldots,j_k>}$, $\widehat{q}_b\Big|^{ab}_{<j_1j_2\ldots,j_k>}$, $\widehat{s}_a\Big|^{a}_{<j_1j_2\ldots,j_k>}$, and $\widehat{s}_b\Big|^{ab}_{<j_1j_2\ldots,j_k>}$ for various subsequences $< j_1j_2\ldots,j_k >$.

By virtue of the Weak Law of Large Numbers, 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 converge (w. p. 1, and in the mean square sense) to the true underlying value of the parameter.

2.2 The Fundamental Theorem of Fusing Occluded Estimates

Our first task is to formulate how we can compute the MSBEs by utilizing information gleaned by the *Binomial* SBEs (BSBEs) obtained from the set of $\binom{d}{2}$ occluded sequences. The theoretical basis for this is the following: Consider an occluded sequence, $\mathcal{X}^{ab}$, extracted from the original sequence, $\mathcal{X}$, by removing all the variables except a and b. In the sequence being examined, we choose one variable, say a to be the unconstrained variable. We shall first attempt to obtain BSBEs of the relative proportions of s_a and s_b, from $\mathcal{X}^{ab}$. Thereafter, we utilize the set of these relative proportions to compute the MSBEs of all the variables.

Theorem 1. *For every pair of indices, a and b, let $\mathcal{X}^{ab}$ be the Occluded sequence, extracted from the original sequence, $\mathcal{X}$, by removing all the variables except a and b. If we consider a to be the unconstrained variable, we define $q_a = \frac{s_a}{s_a+s_b}$ and $q_b = \frac{s_b}{s_a+s_b}$, where $q_a + q_b = 1$. Now let $\widehat{q}_a\Big|^{ab}_{\pi(a,b)} \neq 0$ and $\widehat{q}_b\Big|^{ab}_{\pi(a,b)} = 1 - \widehat{q}_a\Big|^{ab}_{\pi(a,b)}$ be the BSBEs of q_a and q_b respectively based on the occurrence[10] of any specific subsequence $\pi(a,b)$. Then, if c is a dummy variable[11] representing any of the variables, the MSBEs of s_a and s_b obtained by examining the occurrences[12] of $\pi(a,b)$ in every $\mathcal{X}^{ab}$ are:*

$$\widehat{s}_a\Big|^{a}_{\pi(a,b)} = \frac{1}{\sum_{\forall c}\rho_c}, \qquad \textit{and} \qquad \widehat{s}_b\Big|^{ab}_{\pi(a,b)} = \frac{\widehat{q}_b\Big|^{ab}_{\pi(a,b)}}{\sum_{\forall c}\rho_c}, \tag{1}$$

where $\rho_a = 1$ and $\forall c \neq a, \rho_c = \frac{\widehat{q}_c\Big|^{ac}_{\pi(a,c)}}{\widehat{q}_a\Big|^{ac}_{\pi(a,c)}}$.

Proof. The proof of the result is omitted due to space considerations. It is in [4]. An example clarifying its use is also found in [3] and [4]. □

[10] The issue of how BSBEs are obtained for specific instantiations of $\pi(a,b)$ is discussed in the subsequent sections.

[11] The fact that c is a dummy variable will not be repeated in the future invocations of this result.

[12] This, of course, makes sense only if $\forall c, \widehat{q}_a\Big|^{ac}_{\pi(a,c)} \neq 0$. This condition will not be explicitly stated in the future.

In [3], we had derived the explicit expressions for the MSBEs when the subsequences $\pi(a,b)$ are of length 2. We shall now generalize this for the case when the subsequences are of length greater than 2.

2.3 Computational Issues

In all the theoretical results that we shall prove, we shall deal with non-overlapping subsequences. Thus, the number of *non-overlapping* sequences of length two in $\mathcal{X}^{ab}$ is $\frac{N_{ab}}{2}$, and the number of *non-overlapping* sequences of length three in $\mathcal{X}^{ab}$ is $\frac{N_{ab}}{3}$ etc. In any sequence $\mathcal{X}^{ab}$, consider the contiguous sequences of length two (i.e., *aa*, *ab*, *ba* and *bb*). Since the elements of $\mathcal{X}$ are drawn independently and identically, the fact that two adjacent elements x_p and x_{p+1} in $\mathcal{X}^{ab}$ are a, is independent of the event that x_{p+1} and x_{p+2} can also assume the value of a. The pairwise event is thus, effectively, one of "drawing with replacement", and we can thus consider $N_{ab}-1$ consecutive pairs in $\mathcal{X}^{ab}$. Observe that it would be statistically advantageous (since the number of occurrences obtained would be almost doubled) if all the overlapping $N_{ab}-1$ subsequences of length 2 were considered, and where n_{aa}, n_{ab}, n_{ba} and n_{bb} were the number of occurrences of *aa*, *ab*, *ba* and *bb* respectively in these $N_{ab}-1$ subsequences. Similarly, it would be advantageous to consider the overlapping $N_{ab}-2$ subsequences of length 3 were considered etc. Indeed, we shall utilize *these* quantities in the experimental verification of our theoretical results.

3 MSBEs Using Three-at-a-Time Sequential Information

3.1 Theoretical Results

The following analytic results are true when the sequential information is processed three-at-a-time.

Theorem 2. *Let* $q_a = \frac{s_a}{s_a+s_b}$ *and* $q_b = \frac{s_b}{s_a+s_b}$*, where* $q_a + q_b = 1$*. Then,* $\widehat{q_a}\Big|_{<aaa>}^{ab}$ *and* $\widehat{q_b}\Big|_{<aaa>}^{ab}$*, the BSBEs of* q_a *and* q_b *obtained by examining the occurrences of* $<aaa>$ *in* $\mathcal{X}^{ab}$ *are:*

$$\widehat{q_a}\Big|_{<aaa>}^{ab} = \sqrt[3]{\frac{n_{aaa}}{N_{ab}/3}}, \qquad \textit{and} \qquad \widehat{q_b}\Big|_{<aaa>}^{ab} = 1 - \sqrt[3]{\frac{n_{aaa}}{N_{ab}/3}}, \tag{2}$$

where n_{aaa} *is the number of occurrences of* $<aaa>$ *from among the* $\frac{N_{ab}}{3}$ *non-overlapping subsequences of length* 3 *in* $\mathcal{X}^{ab}$*. Consequently,*

$$\widehat{s}_a\Big|_{<aaa>}^{a} = \frac{1}{\sum_{\forall c}\rho_c}, \qquad \textit{and} \qquad \widehat{s}_b\Big|_{<aaa>}^{ab} = \frac{\widehat{q_b}\Big|_{<aaa>}^{ab}}{\sum_{\forall c}\rho_c}, \tag{3}$$

where $\rho_a = 1$ *and* $\forall c \neq a, \rho_c = \frac{1-\sqrt[3]{\frac{n_{aaa}}{N_{ac}/3}}}{\sqrt[3]{\frac{n_{aaa}}{N_{ac}/3}}}$.

Proof. The proof of the result is found in [4]. □

Theorem 3. *Let $q_a = \frac{s_a}{s_a+s_b}$ and $q_b = \frac{s_b}{s_a+s_b}$, where $q_a + q_b = 1$. Then, $\widehat{q}_a\Big|^{ab}_{<bbb>}$ and $\widehat{q}_b\Big|^{ab}_{<bbb>}$, the BSBEs of q_a and q_b obtained by examining the occurrences of $< bbb >$ in $\mathcal{X}^{ab}$ are:*

$$\widehat{q}_a\Big|^{ab}_{<bbb>} = 1 - \sqrt[3]{\frac{n_{bbb}}{N_{ab}/3}}, \qquad \textit{and} \qquad \widehat{q}_b\Big|^{ab}_{<bbb>} = \sqrt[3]{\frac{n_{bbb}}{N_{ab}/3}}, \tag{4}$$

where n_{bbb} is the number of occurrences of $< bbb >$ from among the $\frac{N_{ab}}{3}$ non-overlapping subsequences of length 3 in $\mathcal{X}^{ab}$. Consequently,

$$\widehat{s}_a\Big|^{a}_{<bbb>} = \frac{1}{\sum_{\forall c}\rho_c}, \qquad \textit{and} \qquad \widehat{s}_b\Big|^{ab}_{<bbb>} = \frac{\widehat{q}_b\Big|^{ab}_{<bbb>}}{\sum_{\forall c}\rho_c}, \tag{5}$$

where $\rho_a = 1$ and $\forall c \neq a, \rho_c = \frac{\sqrt[3]{\frac{n_{ccc}}{N_{ac}/3}}}{1-\sqrt[3]{\frac{n_{ccc}}{N_{ac}/3}}}$.

Proof. The details are thus omitted. It is found in [4]. □

To simplify matters, we deal with the rest of the cases that involve three-at-a-time subsequences, by sub-dividing them into the cases when the subsequences contain *one* b, or *two* b's, which are then dealt with in a single theorem.

Theorem 4. *Let $q_a = \frac{s_a}{s_a+s_b}$ and $q_b = \frac{s_b}{s_a+s_b}$, where $q_a+q_b = 1$. Then, $\widehat{q}_a\Big|^{ab}_{<uvw>}$, the BSBE of q_a obtained by examining the occurrences of subsequences of length 3 of the form $< uvw >$ in $\mathcal{X}^{ab}$ of which only a* single *variable is b, can be computed as the real roots (if any) of the cubic equations given below for each such subsequence:*

1. *$\widehat{q}_a\Big|^{ab}_{<baa>}$ is the real root, λ_a, of $\lambda^3 - \lambda^2 + \frac{n_{baa}}{N_{ab}/3} = 0$ whose value is closest to $\widehat{q_a}$;*
2. *$\widehat{q}_a\Big|^{ab}_{<aba>}$ is the real root, λ_a, of $\lambda^3 - \lambda^2 + \frac{n_{aba}}{N_{ab}/3} = 0$ whose value is closest to $\widehat{q_a}$;*
3. *$\widehat{q}_a\Big|^{ab}_{<aab>}$ is the real root, λ_a, of $\lambda^3 - \lambda^2 + \frac{n_{aab}}{N_{ab}/3} = 0$ whose value is closest to $\widehat{q_a}$;*

where n_{baa}, n_{aba} and n_{aab} are the number of occurrences of $< baa >$, $< aba >$ and $< aab >$ respectively from among the $\frac{N_{ab}}{3}$ non-overlapping subsequences of length 3 in $\mathcal{X}^{ab}$. Similarly, $\widehat{q}_b\Big|^{ab}_{<uvw>} = \lambda_b = 1 - \lambda_a$. Finally, in each case,

$$\widehat{s}_a\Big|^{a}_{<uvw>} = \frac{1}{\sum_{\forall c}\rho_c}, \qquad \textit{and} \qquad \widehat{s}_b\Big|^{ab}_{<uvw>} = \frac{\widehat{q}_b\Big|^{ab}_{<uvw>}}{\sum_{\forall c}\rho_c}, \tag{6}$$

where $\rho_a = 1$ and $\forall c \neq a, \rho_c = \frac{\lambda_c}{\lambda_a}$.

Proof. The details of the proof are omitted here and are included in [4]. □

We now consider the scenario when the subsequence examined contains two b's. However, here, we first estimate the probability q_b using which infer the estimate of q_a. Indeed, the theorem mirrors the one above.

Theorem 5. *Let $q_a = \frac{s_a}{s_a+s_b}$ and $q_b = \frac{s_b}{s_a+s_b}$, where $q_a+q_b = 1$. Then, $\widehat{q_b}\Big|_{<uvw>}^{ab}$, the BSBE of q_b obtained by examining the occurrences of subsequences of length 3 of the form $< uvw >$ in $\mathcal{X}^{ab}$ of which* exactly two *variables are b's, can be computed as the real roots (if any) of the cubic equations given below for each such subsequence:*

1. *$\widehat{q_b}\Big|_{<abb>}^{ab}$ is the real root, λ_b, of $\lambda^3 - \lambda^2 + \frac{n_{abb}}{N_{ab}/3} = 0$ whose value is closest to $\widehat{q_b}$;*
2. *$\widehat{q_b}\Big|_{<bab>}^{ab}$ is the real root, λ_b, of $\lambda^3 - \lambda^2 + \frac{n_{bab}}{N_{ab}/3} = 0$ whose value is closest to $\widehat{q_b}$;*
3. *$\widehat{q_b}\Big|_{<bba>}^{ab}$ is the real root, λ_b, of $\lambda^3 - \lambda^2 + \frac{n_{bba}}{N_{ab}/3} = 0$ whose value is closest to $\widehat{q_b}$;*

where n_{abb}, n_{bab} and n_{bba} are the number of occurrences of $< abb >$, $< bab >$ and $< bba >$ respectively from among the $\frac{N_{ab}}{3}$ non-overlapping subsequences of length 3 in $\mathcal{X}^{ab}$. Similarly, $\widehat{q_a}\Big|_{<uvw>}^{ab} = \lambda_a = 1 - \lambda_b$. Finally, in each case,

$$\widehat{s_a}\Big|_{<uvw>}^{a} = \frac{1}{\sum_{\forall c} \rho_c}, \quad \textit{and} \quad \widehat{s_b}\Big|_{<uvw>}^{ab} = \frac{\widehat{q_b}\Big|_{<uvw>}^{ab}}{\sum_{\forall c} \rho_c}, \tag{7}$$

where $\rho_a = 1$ and $\forall c \neq a, \rho_c = \frac{\lambda_c}{\lambda_a}$.

Proof. This proof is similar to the proof of Theorem 4 (by merely replacing a by b and vice versa) and is not included to avoid repetition. □

3.2 Experimental Results: Sequences of Length Three

To justify and experimentally verify the claims of Section 3.1, we now present the results of our simulations on synthetic data for the cases studied in that subsection, namely for the case when the sequence is processed in subsequences of length three. As in the case of sequences of length 2, by virtue of the arguments of Section 2.3, we evaluate the *approximated* versions of the respective equations by considering the $N-2$ overlapping sequences of length 3, and so the solutions are obtained by replacing the existing term, $N/3$, by $N-2$ in Theorems 2 to 5.

As in the case of using pairwise sequences of symbols, the MSBE process for the estimation of the parameters for multinomial random variables was extensively tested for numerous distributions, but we merely cite one specific example. The case we report is when $d = 5$ and the true value of $S = [0.33\ 0.25\ 0.18\ 0.14\ 0.10]^T$.

Here too, we have simultaneously tracked the progress of the "traditional" MLE computation using the identical data stream. Both the estimation methodologies were presented with random occurrences of the variables for $N = 390625$ (i.e, 5^8) time instances. As in [3,4], the criteria for the quality of the estimates were the values of E_{MLE}, the error of the MLE, and the error of the MSBE, E_{MSBE}, at time N.

In the case of the MSBE, the true underlying value of the estimates was computed using each of the estimates when the triples examined in every $\mathcal{X}^{ab}$ were $< aaa >$, $< bbb >$, $< baa >$, $< aba >$, $< aab >$, $< abb >$, $< bab >$ and $< bba >$. The results obtained are tabulated in [3,4] as a function of the number of samples processed. However, to demonstrate the true convergence properties of the estimates and to mitigate the sampling error, we report the values of the ensemble average of the errors in Table 1 taken over an ensemble of 100 experiments. From it one can observe the amazing convergence of every single estimate. For example, the traditional MLE, had the ensemble average error, E_{MLE}, of 0.1918 when only $N = 625$ symbols were processed. The error of the MSBE (when the subsequence examined was $< aaa >$) at that time was 0.2081. When $N = 390625$, the value of the E_{MLE} was exactly 0.1885, while the value of the E_{MSBE} was 0.1886 – demonstrating the power of the estimation strategy!

The same phenomenon can be observed for the other MSBEs, except that in some cases the estimates were much better for smaller values of N. One also observes that the error of the MLE and MSBE evaluated for a *single* experiment are not as smooth - especially when the number of samples processed is small[13]. But fortunately, things "average" out as time proceeds.

Table 1. A table of the *ensemble* averages (taken over 100 experiments) of the error of the MLE, E_{MLE}, and the error of the MSBE, E_{MSBE}, at time N, when the triples examined in every $\mathcal{X}^{ab}$ were $< aaa >$, $< bbb >$, $< baa >$, $< aba >$, $< aab >$, $< abb >$, $< bab >$ and $< bba >$. Here $d = 5$ and $S = [0.33\ 0.25\ 0.18\ 0.14\ 0.10]^T$. The latter MSBEs were estimated by using the approximated results of Theorems 2 to 5 respectively involving the $N_{ab} - 2$ overlapping subsequences of length 3 (approximated using the issues discussed in Section 2.3).

N	E_{MLE}	E_{MSBE} $< aaa >$	E_{MSBE} $< bbb >$	E_{MSBE} $< baa >$	E_{MSBE} $< aba >$	E_{MSBE} $< aab >$	E_{MSBE} $< abb >$	E_{MSBE} $< bab >$	E_{MSBE} $< bba >$
5^2 (25)	0.1279	NaN	0.2955	NaN	NaN	NaN	NaN	NaN	NaN
5^3 (125)	0.1695	NaN	0.2305	0.2163	0.2014	0.2128	NaN	NaN	NaN
5^4 (625)	0.1875	0.1920	0.2110	0.2096	0.2025	0.2082	NaN	NaN	NaN
5^5 (3,125)	0.1886	0.1891	0.1973	0.1958	0.2027	0.1968	0.1925	NaN	0.1928
5^6 (15,625)	0.1883	0.1884	0.1937	0.1912	0.1984	0.1916	0.1905	0.1910	0.1905
5^7 (78,125)	0.1879	0.1880	0.1919	0.1881	0.1879	0.1880	0.1878	0.1882	0.1878
5^8 (390,625)	0.1879	0.1879	0.1895	0.1881	0.1882	0.1880	0.1883	0.1880	0.1883

[13] 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 too small.

The estimated ensemble values of E_{MLE} and $E_{MSBE}|_{<aaa>}$ with N are plotted in [4]. From it we can see that after an initial transient phase, the two curves are *almost undistinguishable.* This same true for the values of E_{MSBE} estimated using the subsequences $< bbb >$, $< baa >$ etc. It is important to mention that the approximated values (using the $N-2$ *overlapping* subsequences) also converge rapidly to the true values of S with a remarkable accuracy.

4 Conclusions

In this paper, we have considered the problem of achieving Sequence Based Estimation (SBE) for multinomial distributions. Unlike traditional estimates, which ignore and discard valuable *sequence*-based information, SBEs "extract" the information contained in the observations when perceived as a *sequence.* The pioneering work in SBEs was presented in [2], and concerned Binomial distributions. Since then, the analysis for multinomial distributions was left open. The first step in solving the SBE problem for multinomial distributions was made in [3]. The strategy that we developed there involved a novel and previously-unreported phenomenon called "Occlusion" where by hiding (or concealing) certain observations, we mapped the original estimation problem onto a lower-dimensional binomial space. We have also shown how these consequent occluded SBEs could be fused to yield overall Multinomial SBE (MSBE). The results in [3] achieved this by only investigating the information found in pairs of symbols in the occluded sequence. In this paper, we have further generalized these results when we considered contiguous subsequences of length 3 in the occluded sequence, which was then fused to yield the overall MSBE. The theoretical results have been experimentally verified. The analytic and experimental results for the cases when the subsequences are of lengths greater than 3 will soon be published.

References

1. K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.
2. B. J. Oommen, S-W. Kim and G. Horn, "On the Estimation of Independent Binomial Random Variables Using Occurrence and Sequential Information", *Pattern Recognition*, vol. 40, pp. 3263-3276, November 2007.
3. B. J. Oommen and S-W. Kim, "*Multinomial* Sequence Based Estimation: The Case of *Pairs* of Contiguous Occurrences." To appear in the *Proceedings of AI'16, the 2016 Canadian Artificial Intelligence Conference*, Victoria, Canada, May 2016. *This talk will be a Plenary/Keynote Talk at the Conference.*
4. B. J. Oommen and S-W. Kim, "Occlusion-based Estimation of Independent *Multinomial* Random Variables Using Occurrence *and* Sequential Information". To be submitted for Publication.
5. S. Ross. *Introduction to Probability Models.* Academic Press, second edition, 2002.
6. J. Shao. *Mathematical Statistics.* Springer Verlag, second edition, 2003.
7. R. Sprinthall. *Basic Statistical Analysis.* Allyn and Bacon, second edition, 2002.
8. F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, parameter estimation and state estimation: An Engineering Approach using MATLAB*, John Wiley and Sons, Ltd., England, 2004.
9. A. Webb. *Statistical Pattern Recognition.* John Wiley & Sons, N.York, Second Edition, 2002.