

SOME STATISTICAL ASPECTS OF THE ANALYSIS OF GENOMIC SEQUENCES

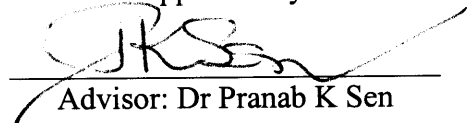
by

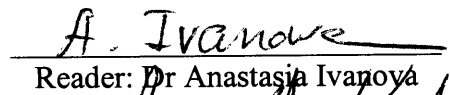
Lily Wang

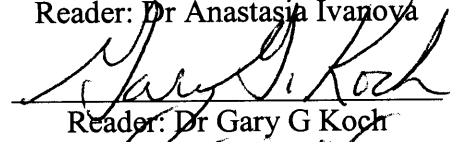
A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the
Department of Biostatistics, School of Public Health.

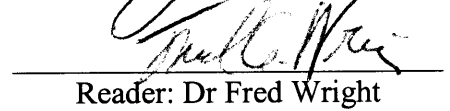
Chapel Hill
2004

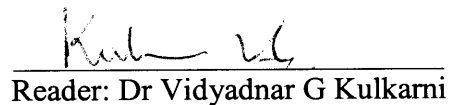
Approved by:


Advisor: Dr Pranab K Sen


Reader: Dr Anastasia Ivanova


Reader: Dr Gary G Koch


Reader: Dr Fred Wright


Reader: Dr Vidyadnar G Kulkarni

© 2004
Lily Wang
ALL RIGHTS RESERVED

ABSTRACT

LILY WANG. Some Statistical Aspects of the Analysis of Genomic Sequences
(Under the direction of Dr Pranab K Sen)

During the courses of evolution, biological sequences undergo continuous changes through mutations such as substitutions, insertions, or deletions. Along the lengthy stretches of the sequences, some regions that do not change as much as the rest are called the "functional domains". Their resistance to change often suggests these regions serve critical functions; therefore, similarities of the sequences often suggest likeness in structure and function, in addition to relationships in phylogeny. In the first part of the dissertation, we discuss the Headruns problem. The alignment score which measures similarity between two sequences will depends on how the sequences were aligned. Here we consider the case when we are given two sequences A_1, A_2, \dots , and B_1, B_2, \dots with letters independent and identically distributed in a fixed alignment, that is, the sequences has be aligned already and no more shifts is needed. The alignment of two sequences can be done by moving one sequence on top of the other sequence until certain number of letters, say 20, are found to be the same, then chopping off sections of the sequences in which there is no corresponding letters from the other sequence. To remove bias, the first 20, or the number that was used in the matching criteria of the aligned segment is deleted. In this case, the local alignment score R_n corresponds to the longest run of heads when a coin is flipped repeatedly. We study the asymptotic behavior of the

local alignment score by extending the Chen-Stein theorem to situations with very general assumptions: (1) the matching probabilities are possibly different at each position; (2) in addition to the first case, the underlying positions along the sequences are assumed to have first order Markovian dependent structure.

In the second part of the dissertation, we study the profile scores. Profile analysis uses multiple alignment profile of a family of related sequences to search a database for more examples of the family. Given a multiple alignment, as we are searching along another long sequence, we construct a distance measure between the profile and the sequence called the profile score. The scores are a set of dependent stationary triangular arrays. Unlike previous studies, to study the asymptotic behavior of the profile scores as well as their maximum, we put the profile in a random context and accommodate the possibility of gaps in the profile with small probability. First, we show that the tail behavior of the profile scores can be well modeled using normal mixture distribution. Next, we find the analytic formula for normalizing constants in extreme value theory for normal mixture distribution. We then derive the distribution for the maximum of the profile scores when means and variances of the scores are known by applying the Chen-Stein theorem. When the profile is random, the means and variances are unknown, we also derive distribution for the maximum of the profile scores in this case. Finally, we apply these results to the Immunoglobulin profile Ig and demonstrate numerical applications of the theories.

ACKNOWLEDGMENTS

I would like to thank my dissertation advisor, Dr Pranab K Sen for supervising me on this research. His intellectual wisdom, expert guidance and infinite patience are much appreciated. Many thanks to my committee members Dr Anastasia Ivanova, Dr Fred Wright, Dr Gary Koch, and Dr Vidyadnar Kulkarni for help discussions and suggestions. Especially, I would like to thank Dr Gary Koch, who has also been my academic advisor during graduate school, for providing me the opportunity to work at the Biometric Consulting Lab and to gain valuable working experiences. His insightful advices, generous support and valuable guidance are much appreciated. Many thanks to my friends at the BLC, who have made the consulting lab such a wonderful place to work.

I would like to thank my parents for their unlimited love, understanding and faith in me. Finally, my sincere appreciation goes to two of my dearest friends, brother Kai and fiancée Xi, for their love, encouragement, help and accompany during this difficult time of my life.

CONTENTS

LIST OF TABLES	ix
1 LITERATURE REVIEW	1
1.1 Introduction	1
1.2 Molecular Biology	2
1.2.1 Structure of DNA and RNA	2
1.2.2 Structure of Protein	3
1.2.3 Central Dogma of Molecular Biology	4
1.3 The Human Genome Project	5
1.4 Similarity Scores from Sequence Comparisons	11
1.4.1 Pairwise Sequence Alignment	11
1.4.2 Profile Scores	14
1.4.3 Phase Transition and Asymptotic Growth of Scores	16
1.4.4 Distribution for the Maximum	19
1.4.4.1 The Chen-Stein Theorem and Poisson Approximation	19
1.4.4.2 Ungapped Pairwise Alignment	20

1.4.4.3	Gapped Pairwise Alignment	22
1.4.4.4	Profile Analysis	26
1.5	Synopsis of Research	29
2	PAIRWISE ALIGNMENT – ANALYSIS OF HEADRUNS	33
2.1	Introduction	33
2.2	Nonhomogeneous Matching Probabilities, Independent Positions	35
2.3	Nonhomogeneous Matching Probabilities, Markovian Dependent Positions	42
3	MULTIPLE ALIGNMENT – PROFILE ANALYSIS	47
3.1	Introduction	47
3.2	Central Limit and Large Deviation Theorems for Triangular Arrays	52
3.2.1	Cramer’s Theorem	53
3.2.2	Application to Profile Analysis	54
3.3	Extreme Value Theory for Normal Mixture Distribution	57
3.3.1	Motivation	57
3.3.2	the Normal Mixture Model is of Exponential type in the Tails	61
3.3.3	The Normalizing Constants from Normal Mixture Distribution $F_\epsilon(x)$	64
3.3.3.1	Properties of the solution η_n	65
3.3.3.2	Case I: when c is large ($c \gg 1$)	67
3.3.3.3	Case II: when $c \approx 1$ (local contamination)	73

3.4	Order of $\bar{X}_n - \mu\sqrt{m}$ and $\frac{s_n}{\sigma} - 1$	76
3.4.1	Order of $\bar{X}_n - \mu\sqrt{m}$	76
3.4.2	order of $\frac{s_n}{\sigma} - 1$	84
3.5	Distribution for the Maximum of Profile Scores	99
3.5.1	When μ and σ are known	106
3.5.2	When μ and σ are unknown	110
4	NUMERICAL EXAMPLE	124
4.1	Description of the Data	124
4.2	Analysis Strategies	125
4.3	Results	129
5	SUMMARY AND FUTURE WORK	138
6	REFERENCES	142

LIST OF TABLES

3.1 Comparison of estimates from analytic formula and true x for n=200	121
3.2 Comparison of estimates from analytic formula and true x for n=400	122
4.1 Profile statistics for Ig Profile	134
4.2 Sequences found statistically most similar to immunoglobulin domain	136

LIST OF FIGURES

3.1 Multiple alignment profile with no insertion or deletions	47
3.2 Multiple alignment profile with some insertion or deletions	60
3.3 Largest characteristic observation: $h(x)$ vs. x	123
4.1 The immunoglobulin (Ig) profile	131
4.2 Protein sequences in FASTA format	135
4.3 Q-Q plot of empirical scores vs. quantiles from extreme value distribution	137

CHAPTER 1

LITERATURE REVIEW

1.1 Introduction

During the course of evolution, biological sequences undergo continuous changes through mutations such as substitutions, insertions or deletions. Along the lengthy stretches of sequences, some regions that do not change as much as the rest are called the "functional domains". Their resistance to change often suggests these regions serve critical functions; therefore, the similarities of sequences often suggest likeness in structure and function, in addition to relationships in phylogeny. In the post human genome era, as massive amount of genetic data are being generated, one important task is to organize and analyze these data to extract important information. One way to accomplish this is by using the methods of sequence alignment. We review relevant statistical and biological background for sequence analysis in this chapter.

In the first section, concepts in biology such as the structures of DNA, protein and the Central Dogma of Molecular Biology, are reviewed. The next section accounts for the history of the Human Genome Project and points out various ways sequence analysis may

help us to understand the genome. Next, in section three, recent works on the statistical properties of alignment scores are reviewed. Relevant theories for the asymptotic growth of alignment scores, the Chen-Stein theorem as well as its applications to pairwise and profile analysis are discussed. Finally, we outline the research in the next few chapters.

1.2 Molecular Biology

In this section, we review some basic genetic concepts, such as the structures of DNA, RNA, protein, and the central dogma of molecular biology. More detailed accounts can be found in standard genetic texts such as Klug and Cummings (2000).

1.2.1 Structure of DNA and RNA

In all organisms except certain virus, DNA is the genetic material that is physically transmitted from parent to offspring. In a DNA molecule, two long polynucleotide chains are coiled around a central axis to form a double helix in an anti-parallel fashion, where the C-5'-to-C-3' orientations run in opposite directions. Within a single chain, each nucleotide has three essential components: a nitrogenous base, a pentose sugar (5 carbon sugar), and a phosphate group. The name of DNA derives from its sugar component, a deoxyribose sugar, or a sugar molecule with H instead of OH at the 2' position. In addition, a phosphate group is attached to the 5' position and a base is attached to the 1' position of sugar molecule. The bases adenine (A), guanine (G), cytosine (C), and thymine (T) are classified into two types according to their chemical structures: A and G are purines, or nine-member double-ring; C, T are pyrimidines, or six-member single-ring. These bases are flat struc-

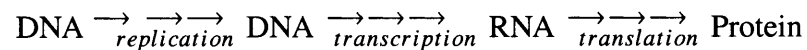
tures, lying perpendicular to the axis, and stacked on one another. Between the two chains, the bases are paired with each other by hydrogen bonding, a weak electrostatic attraction that usually forms between a hydrogen atom and a strongly electronegative atom such as oxygen or nitrogen. In DNA, only A = T and G = C are allowed. In contrast, a RNA molecule is a single stranded molecule, with bases A, G, C, and U (uracil). The base U appears only in RNA and pairs with A from DNA during the process of transcription.

1.2.2 Structure of Protein

While the genetic information is stored in DNA, the ultimate cellular activities are influenced through DNA encoded proteins. A protein is one or a few polypeptide chains made up of amino acids. Each amino acid has a carboxyl group, an amino group, and an R (radical) group, all bound to a central carbon atom. The 20 amino acids can be classified into four groups according to chemical structure of the R group: nonpolar, polar, negatively charged, and positively charged. During dehydration reaction, the amino acids are linked together by peptide bond, hence the name polypeptide chain. There are four levels of protein structures: primary, secondary, tertiary and quaternary. The primary structure is specified through the sequence of amino acids; the secondary structure refers to the local conformations, two most common configurations are α helix and β -pleated sheet; in contrast, tertiary structure describes the three dimensional conformation of the entire chain in space; and finally, quaternary structure defines the conformation of the various chains in relation to one another for proteins composed of more than one polypeptide chains, many of the enzymes, including DNA and RNA polymerases, show quaternary structure.

1.2.3 Central Dogma of Molecular Biology

The process that describes the information flow involving DNA, RNA and proteins within cells (except in certain retrovirus) is the central dogma of molecular genetics: DNA makes RNA, which in turn makes proteins. It can be shown more clearly in the following diagram:



First, the process of replication is initiated by unwinding of DNA helix, helped by proteins DnaA, DnaB and DnaC. Then a short segment of RNA, complementary to DnaA, is first synthesized to initiate DNA synthesis. Under the semi-conservative model, as the strands start to unwind, each of the two original DNA strand serves as a template for DNA synthesis. As DNA polymerase adds each nucleotide, it also proofreads and excises any mismatched nucleotide to increase the fidelity of the process. At this stage, genetic exchange between any two double stranded DNA molecules from homologous chromosomes often occurs, resulting in genetic recombination.

Next, transcription is initiated at upstream DNA region called the promoter, which contains specific DNA sequences such as TATA box. Similar to DNA replication, here, an RNA polymerase binds the promoter region and directs synthesis of a strand of RNA complementary to the DNA template; with the exception that A is paired with U instead of T. In eukaryotic cells such as animal cells, some internal DNA sequences called the introns or as they are often called "the junk DNA" are removed in the process of splicing. The remaining DNA, the exons ("ex" for expressed) proceed to be translated into proteins.

In the last step, translation occurs where each group of three ribonucleotides, called the codons, specifies one amino acid. These codons have the properties of unambiguous, degenerate, and non-overlapping: each triplet specifies only a single amino acid; but a given amino acid can be specified by more than one triplet; and once translation begins, any single ribonucleotide at a specific location is part of only one triplet. Initiation of translation involves assembly of large, small units of ribosome, which serve as workbench for the translation process for a few enzymes, the mRNA transcript and a tRNA molecule. The process then proceeds as tRNA attaches amino acids one at a time to the mRNA template, after which an enzyme called peptidyl transferase catalyzes the formation of peptide bonds and links each amino acid to the preceding one. Termination of the process occurs when the stop codons (UAG, UAA or UGA) are reached, after which tRNA and polypeptide chain are released. The polypeptide then folds, assumes a three dimensional conformation and becomes a functional protein.

1.3 The Human Genome Project

Although controversial from the beginning, the completion of a high-quality comprehensive sequence of human genome was considered one of the greatest achievements in the history of biology. The journey for completing this herculean task and the many debates along the way were described in the February issue of Nature magazine in 2001. (Robert, Service, Enserink, Vogel, Marx, Helmuth, Marshall, 2001) The goals of the Human Genome Project were to determine sequences of the 3 billion base pairs that make up human DNA; to identify all 30,000 genes in human DNA; to store this information in

databases; to improve tools for data analysis; and to address the ethical, legal and social issues. This landmark event has lead us into the new genomic era and will significantly advance our understandings of the complex human biological systems as well as roles of genetic factors in human diseases.

The Human Genome Project was almost completed by two teams of researchers under one of the most hotly contested races in recent scientific history. The public project by the International Human Genome Mapping Consortium was an international thirteen-year effort that officially started in 1990. Funded mainly by the US department of Energy and the National Institute of Health in collaboration of Britain's Wellcome Trust, the entire project costed more than 3 billion US dollars. Collaborators from all over the world, including Japan, France, Germany and China also made important contributions. The private project was funded by the company Celera Genomics, which was formed in 1998. The first working draft of the entire human genome was completed in 2000 and in February issues of Nature and Science magazines, the public and private teams presented descriptions of their strategies used to obtain sequencing information of the DNA molecules that constitute the genome as well as the initial analysis of the human genome. (International Human Genome Sequencing Consortium, 2001a,b and The Celera Genomics Sequencing Team, 2001)

The private team used the whole genome shotgun sequencing method. First, the entire 3-billion-base genome was shredded into zillions of fragments. Then, these fragments were cloned into plasmids and sequenced on both strands. Once the sequences were obtained, they were aligned so that identical sequences were overlapping. The contiguous pieces were then assembled into finished sequence using one of the world's fastest super-

computer. In contrast, the public team used the hierarchical shotgun sequencing method. First, genomic DNA was fragmented and a library was constructed by inserting the pieces into BAC cloning vectors. The genomic DNA fragments in the library were then organized into a physical map and individual BAC clones were selected and sequenced by the random shotgun strategy. Finally, the clone sequences were assembled to reconstruct the sequence of the genome. Because human genome consisted large amount of repeated sequences (>50%), the hierarchical method where only local sequencing is used, was considered more accurate since long-range misassembly was eliminated and the risk of short-range misassembly was reduced, although at a higher cost, since a map of clone was needed initially. Birney et al. (2001) detailed a comparison of the quality of the two sequences. Bentley et al. (2001), Tilfore et al. (2001), Montgomery et al. (2001), Bruls et al. (2001) and Riethman et al. (2001) give detailed account of techniques used for sequencing as well as descriptions of contents of the individual chromosomes.

The initial analysis of the genome sequences provided some panoramic views of the human genetic landscape: as the first vertebrate genome to be sequenced, the human genome is about 3.2 gigabases (Gb), of that 2.95 Gb is euchromatic or gene rich; it is 30 times larger than worm and fly genomes, and 250 times larger than yeast genome. To give an idea about the size of human genome, if we were to compile it in books, we would need 200 volumes the size of Manhattan telephone book (at 1000 pages each) to hold it all; if we were to read it out loud, it would take 9.5 years on a reading rate of 600 bases/minute. (US Department of Energy Genome Program) The estimated number of human genes is 25,000-35,000 (Ewing, 2000 and Roest, 2000). The protein coding regions of the genes, called the

exons, are separated in the genome by the non-coding regions called the introns. In human genome, the exons only account for about 3% of the entire sequence, repeat sequence of various types form over half of the DNA. (Birney, 2001 and Baltimore, 2001).

To take advantage of this book of life, the important goals in the post-genome era will be to map positions of genes and coding sequence variations, and to annotate this map by understanding the functions of genes and their interplay with proteins and the environment to create complex and dynamic biological systems. This in turn will advance our understanding of the role of genetic factors in human health and disease by helping us identify genes responsible for human Mendelian diseases as well as good health. In addition, this insight will help us develop genome based diagnostic methods for the prediction of susceptibility to disease, the prediction of drug response, the early detection of illness and the accurate molecular classification of disease as well as serve as an engine for pharmaceutical discovery.

In achieving our goals, sequence similarity comparisons will serve as an important tool. Already, biologists have been trying to discover new genes in their favorite systems by carrying out some data mining exercises. The February issue of Nature magazine in 2001 contains these initial findings from the broad topics such as cancer, addiction, gene expression, immunology, evolutionary genomics, to the more specific topics, such as membrane trafficking, cytoskeleton, cell cycle, and circadian clock. (Futreal, Nestler, Tupler, Fahrner, Li, Bock, Pollard, Murray, Clayton, 2001)

There are a number of ways we can use sequence alignments to help us understand the genome. First, we can determine positions of genes by aligning mRNAs with ge-

omic sequences. For example, Sagane (1998) aligned mRNA of the membrane-bound metalloprotease-disintegrin ADAM23 to the draft genome to find that the gene consists of at least 23 exons. When mRNA species align differently to a genomic sequence, this suggests an alternative splicing has taken place. The International Human Genome Sequencing Consortium (2001) has shown 60% of human genes have multiple splicing variants. This is an important feature of the human genome.

Second, we can also search for proteins belonging to a particular family to discover new genes. Wolfsberg (2001) searched for proteins from human gene ADAM23 which maps to 2q33 (Poindexter 1999), against proteins from draft human sequence, and found that aside from matching itself, the best matching region was to a peptide from chromosome 20. No ADAM family has previously been mapped to this chromosome. Further analysis showed that this protein consists of structures characteristic of the ADAM family and confirmed it is a functional gene rather than a pseudogene. As another example, Nestler (2001) searched for G-protein receptor kinases in the human genome for their research.

Third, comparing homologous regions of genome sequences from different species will also help us identify genomic elements. It has been estimated that over 40% of the predicted human proteins share similarity with fruitfly or worm proteins. Clayton (2001) compared *Drosophila*'s period clock protein with human genome and found three known relatives and a possible fourth on Chromosome 7. Also, Wolfsberg (2001) compared mouse *Lmx1b* gene with human genome and found the human *LMX1B* gene which maps to 9q34. Further analysis confirmed the homology: the inactivation of *Lmx1b* gene in mouse lead to a phenotype that is very similar to human nail patella syndrome (NPS), an autosomal reces-

sive disorder characterized by limb and kidney defects which has been linked to LMX1B gene. In addition, the International Human Genome Sequencing Consortium (2001), the Mouse Genome Sequencing Consortium (2002) and Aparicio (2002) have also discovered many protein-coding sequences by comparing available vertebrate genome sequences. This is very important in that mutants of homologous genes from other species such as mouse will facilitate the characterization of functions and mutational processes in human genes. Moreover, determining the sequence difference between species will provide us with insight into the distinct anatomical, physiological and developmental features of different organisms and help define the genetic basis for speciation. (Sidow, 2002).

Finally, sequence alignment methods will also help us study the sequence variations within species. Individual humans differ from each other by about one base per thousand, and the most common class of variation is the SNP markers. Studies on “single nucleotide polymorphisms” (SNP) will help us uncover the genetic basis of many diseases. The total number of SNPs in the public database (dbSNP) (Smigielski, 2000) now exceeds 2.5 million, representing 1.5 million unique loci. The databases include information on flanking regions around the SNPs, so by sequence similarity comparison, we can localize variations within the human genome. This information will provide information about our personal responses to medicines and help pharmaceutical companies to develop drugs targeted to specific sites in the body and to specific populations. These drugs are promised to be more powerful and have fewer side effects. In the future, drugs might one day be tailor-made for individuals and adapted to each person’s own genetic make up. Instead of the standard trial-and-error method of matching patients with the right drugs, doctors will be able to

analyze a patient's genetic profile and prescribe the best available drug therapy from the beginning with the appropriate drug dosage.

In summary, the Human Genome project has provided us with unprecedented wealth of biological data. There are many undiscovered treasure in the current data set. Sequence comparison methods will be an important tool to help us translate genomic information into comprehensive understanding of biological systems and therapeutic advances.

1.4 Similarity Scores from Sequence Comparisons

1.4.1 Pairwise Sequence Alignment

There are many different ways to align two random sequences. The followings are a few overlapping categories as discussed in Ewens and Grant (2001): for global alignment, the entire length of the sequences are aligned; in comparison, for local alignment, only some parts of the entire sequences are aligned. In addition, there is fixed (known) alignment, where the alignment of two sequences with the same length is fixed in advance and alignment (unknown) with shift, where gaps are added and shifts are allowed to produce the alignment with highest score. Finally, for exact matching, we require all letters from both sequences to match exactly. In comparison, for approximate matching, a fraction of mismatches is allowed for alignment.

The score functions can be defined mathematically for each type of alignment. Let A_1, A_2, \dots, A_n and B_1, B_2, \dots, B_n be two sequences, the score function S_n for global fixed

alignment is defined by

$$S_n = \sum_{i=1}^n s(A_i, B_i)$$

where $s(a, b)$ is a non-negative finite-valued function on pairs of letters. In the simplest case, $s(a, b)$ can be indicator function for matching. If shifts are allowed such that $A_1, A_2, \dots, A_n \rightarrow A_1^*, A_2^*, \dots, A_L^*$ and $B_1, B_2, \dots, B_n \rightarrow B_1^*, B_2^*, \dots, B_L^*$ are two sequences with gaps added, then

$$S = \max \left\{ \sum_{i=1}^L s(A_i^*, B_i^*) : \text{all alignments} \right\}$$

Although the early studies focused on global alignment (Needleman and Wunsch, 1970; Sellers, 1974; Sankoff and Kruskal, 1983), it has been recognized that as biological sequences undergo evolution, only isolated regions of the sequences will remain similar. Therefore, local alignment which locates the best matching segments of two given sequences has been the focus of research in recent years. Waterman (1995) gives comprehensive review of various local alignment score statistics. For fixed local alignment where exact matching is required, a simple score function is

$$R_n \equiv \max\{m : A_{i+k} = B_{i+k} \text{ for } k = 1 \text{ to } m, 0 \leq i \leq n - m\}$$

If shifts are allowed, the score function then becomes

$$H_n \equiv \max\{m : A_{i+k} = B_{i+k} \text{ for } k = 1 \text{ to } m, 0 \leq i, j \leq n - m\}$$

In addition, if we also allow for removal of l single letters, the score function for approximate matching is

$$H_n^*(l) \equiv \max\{m : A_{i+k} = B_{i+k} \text{ for } k = 1 \text{ to } m, \text{ except for at most } l \text{ letters}\}$$

On the other hand, if we allow for a fraction α of mismatches, the score function is

$$R_n^\alpha = \max \left\{ t : \alpha t \leq \sum_{1 \leq k \leq t} C_{i+k}, 0 \leq i \leq n - t \right\}$$

where $C_i = I(X_i = Y_i)$.

One limitation of these scores is their inability to allow for mismatches that vary in degree. A more general scoring scheme assigns score $s(A, B)$ to matching between letters A and B according to some substitution matrix and only requires that $E\{s(A, B)\} < 0$, and $s^* = \max s(a, b) > 0$. The first condition ensures that the expected score for the segment has negative score, while the second condition guarantees some positive scores are possible. These conditions are necessary: if the expected score were positive, then the local alignment score could always be increased by increasing the matching segment length and this violates the idea of seeking local alignments; if all the scores were negative, then the maximum score would always be a single pair of residues, but that is not what we are interested in.

Karlin & Altschul (1990) and Altschul (1991) discuss an important result that describes the composition of the high-scoring segments in ungapped alignments. This theorem gives us a rational way to assign scores which are "optimal" for distinguishing biologically relevant patterns. The theorem says that when gaps are not allowed, for sequences generated from the i.i.d. models, pair of letters (i,j) in the best matching interval tends to occur with "target frequency"

$$q_{ij} = p_i p_j e^{\lambda_u s_{ij}}$$

where p_i is the "background" frequency for residue i , s_{ij} is score for matching residues i and j , and λ_u (u denotes ungapped) is the parameter to be calculated. Writing $s_{ij} = \log(q_{ij}/p_i p_j)/\lambda_u$, we can see that scores can be chosen, with arbitrary scale, for any desired set of q_{ij} , and any substitution matrix is implicitly a "log-odds" matrix with a specific target distribution for aligned pairs of residues. Dayhoff et al. (1978) and Henikoff et al. (1992) constructed the popular PAM and BLOSUM matrices by estimating the target frequencies for amino acids from properly aligned, but not strongly related, proteins, and assigned the scores with the explicit use of this logodds formula. Also, since $\sum_{i,j} q_{ij} = 1$, λ_u is calculated to be the unique positive solution to this equation. Notice that multiplying all scores in the substitution matrix by a constant c will not affect the relative scores of the local alignments, but will have the effect of dividing λ by c . Therefore, λ can be viewed as a natural scale for the scoring matrix. Waterman (1995) contains a heuristics for the proof and Arratia et al. (1988) gives the complete proof for this theorem. Altschul et al. (1997) conjectured corresponding result would hold for gapped alignment scoring system as well, if the gap costs used are sufficiently large.

1.4.2 Profile Scores

Another type of scoring matrix that compares a set of sequences already in a multiple alignment and a simple sequence is called the position-specific matrix, the profile, or the motif. Database searches using profiles are increasingly being popular since they are often more sensitive at detecting weak relationships than searches that use a simple sequence as

query. As discussed in Altschul et al. (1997), one reason for this is that each column of the profile gives a more accurate estimation of the probability which amino acids occur.

Tatusov et al. (1994) describes various ways of constructing the position-specific weight matrix. The most intuitive way is the average method proposed by Gribskov et al. (1987). Here, the score for amino acid j in column k is $W_{jk} = \sum_{i=1}^{20} c_{ik}s_{ij}/N$ where c_{ik} is the number of times residue i occurs within column k , and s_{ij} is the score for aligning residue i and j from a substitution matrix. Several other methods were also discussed in Tatusov et al. (1997): the Bayesian prediction method, the data-dependent pseudocount method, and the Dirichlet mixture method. In addition, the authors compared relative discriminating power of the weight matrices among themselves and with simple sequence as query in database searches. To evaluate the performances of different scoring schemes, five protein superfamilies that have been studied in detail where a canonical list of true family members could be produced was used. The measure of power for a given matrix was taken to be the cutoff at which the number of false positive, E , equals the number of false negative, F . Clearly, the matrix with greater discriminating power would have lower value of E , with E ideally being zero. The results show that position-specific weight matrices, especially the Dirichlet mixture method, significantly outperform using simple sequence as query in database searches.

1.4.3 Phase Transition and Asymptotic Growth of Scores

A critical phenomena for pairwise local alignment scores is that as $n \rightarrow \infty$, there is a phase transition between their linear growth in n , when the penalty parameters are small, and their logarithmic growth in n , when the penalties are large. This phenomena was first studied by Arratia and Waterman (1985), Arratia et al. (1987), and Arratia et al. (1988). The more recent results are given by Arratia and Waterman (1994) and Dembo et al. (1994), where phase transition was shown for i.i.d. or Markov sequences. For simplicity, for simple alignment scoring functions, where aligned letters x and y score

$$s(x, y) = \begin{cases} 1 & \text{if } x = y \\ -\mu & \text{if } x \neq y \end{cases}$$

and deleted letters score $-\delta$ per letter. Let S_n be the global alignment score over all possible "alignments" between two sequences, and let the limiting score per letter be $\alpha(\mu, \delta) = \lim_{n \rightarrow \infty} \frac{ES_n}{n}$. Kingman's theorem (Durrett, 1991) for subadditivity implies this limit exists almost surely. When $\mu = \infty$ and $\delta = 0$, α is the Chvatal-Sankoff constant. (Chvatal and Sankoff, 1975). However, in all other cases, much unknown still remains, only the bound $0.7615 \leq \alpha \leq 0.8575$ is known. Arratia and Waterman showed that the set of values (μ, δ) for which $\alpha = 0$ forms a line in the parameter space separating the negative region for α from the positive, and this line is the location of the phase transition between logarithmic and linear growth for local alignment score.

To prove the phase transition phenomena, Arratia and Waterman (1994) first showed

that the large deviations for S_n have probability which is exponentially small as $t \rightarrow \infty$, i.e.

$$\Pr(S_k \geq ES_k + \epsilon k) \leq e^{-\frac{\epsilon^2 k}{2c^2}}$$

where $c = \min(2 + 4\delta, 2 + 2\mu)$. This proof is based on the Azuma-Hoeffding inequality (Williams, 1991) for martingales with bounded increments. Then, it was shown that the phase transition result follows rigorously from the large deviation result. Finally, Arratia and Waterman was able to extend this result to local alignment scores with general scoring scheme with symmetric scoring where $s(x, y) = s(y, x)$ and subadditive gap weights $w(k) \leq w(k) + w(l)$.

For asymptotic growth of local alignment scores, Erdos and Renyi (1970) first presented the result for i.i.d. sequences with restricted scoring $p \in (0, 1)$:

$$\Pr\left(\lim_{n \rightarrow \infty} \frac{R_n}{\log_{1/p} n} = 1\right) = 1$$

where R_n is defined as before. Waterman(1995) gives an easy heuristic: in the case of fixed alignment, the longest matching segment between two sequences corresponds to the longest headrun. Now, a headrun of length m has probability p^m , there are about n possible headruns so

$$E(\# \text{ of headruns with length } m) \approx np^m$$

If the largest run is unique, its length R_n should satisfy $1 = np^{R_n}$, which has solution $R_n = \log_{1/p} n$. In addition, Waterman (1995) pointed out that in the case of alignment with shifts, allowing shifts gives approximately n^2 choices for (i, j) , the starting position of a match run. This suggests that H_n grows like $\log_{1/p} n^2$. These heuristics turn out to

be correct, Karlin et al.(1983) and Arratia and Waterman(1985) independently proved for i.i.d. sequences

$$\Pr\left(\lim_{n \rightarrow \infty} \frac{H_n}{\log_{1/p} n} = 2\right) = 1$$

In a series of papers, Arratia and Waterman (1985,1989) showed the $\log n^2$ asymptotic growth for H_n also holds for Markovian sequences and two sequences with different underlying distributions. More specifically, if $A_1, A_2, \dots, B_1, B_2, \dots$ are two independent Markov chains on a finite alphabet A . Assume the chains are irreducible, aperiodic, and have transition probabilities $(p_{ij}), i, j \in A$. Let $\lambda \in (0, 1)$ be the largest eigenvalue of the substochastic matrix $\{(p_{ij}^2)\}, i, j \in A$, then

$$\Pr\left(\lim_{n \rightarrow \infty} \frac{H_n}{\log_{1/\lambda} n} = 2\right) = 1$$

Also, if A_1, A_2, \dots is distributed as ξ and B_1, B_2, \dots is distributed as ν with all letters independent and $p = \Pr(A_1 = B_1) \in (0, 1)$. Then, there is a constant $C(\xi, \nu) \in [1, 2]$ such that

$$\Pr\left(\lim_{n \rightarrow \infty} \frac{H_n}{\log_{1/p} n} = C(\xi, \nu)\right) = 1$$

Finally, Arratia et al. (1988) opened up the important topic of matching with scores. They proved the $\log n^2$ asymptotic growth for H_n for general scores which only required $E\{s(A, B)\} < 0$, and $s^* = \max s(a, b) > 0$.

1.4.4 Distribution for the Maximum

1.4.4.1 The Chen-Stein Theorem and Poisson Approximation

In the studies of sequence alignment, one of the basic assumptions is that truly homologous biological sequences contain segments with close matches. Often the interest is focused on evaluating statistical significance of segments with exact or close matches between two sequences by computing tail probabilities. Before the 1990s, the standard way was using the Bonferroni inequalities introduced by Watson (1954) to approximate distribution of counts of weakly dependent rare events. See Karlin and Ost (1987), Karlin et al. (1983) for examples. However, the Bonferroni inequality method often involves tedious and technically demanding computation of moments of large order. In 1990, Arratia et al. pioneered the use of methods developed by Chen(1975) and Stein (1986) to establish a Poisson approximation for dependent events in the context of sequence matching scores. Another good reference that is accessible to the general audiences on the Chen-Stein method of Poisson approximation is Arratia et al. (1989). We next state this important theorem.

Theorem 1.1 (*Chen-Stein*) *Let I be an index set, and for each $i \in I$, and let X_i be an indicator random variable. The total number of occurrence of events is*

$$W = \sum_{i \in I} X_i$$

For each $i \in I$, let J_i be the set of dependence of i , and assume that

X_i is independent of $\{X_j\}$, $j \notin J_i$.

Let Z be a Poisson random variable with $E(Z) = E(W) = \lambda$. Then,

$$\|W - Z\| \leq 2(b_1 + b_2) \frac{1 - e^{-\lambda}}{\lambda} \leq 2(b_1 + b_2)$$

where

$$b_1 \equiv \sum_{i \in I} \sum_{j \in J_i} E(X_i)E(X_j)$$

and

$$b_2 \equiv \sum_{i \in I} \sum_{i \neq j \in J_i} E(X_i X_j)$$

in particular,

$$|\Pr(W = 0) - e^{-\lambda}| \leq (b_1 + b_2) \frac{1 - e^{-\lambda}}{\lambda} \quad (1.1)$$

1.4.4.2 Ungapped Pairwise Alignment

We next apply the Chen-Stein method of Poisson approximation to compute the probability of occurrences for long runs of heads or matches in the case of restricted scoring. As pointed out by Arratia et al. (1990), there are two distinct issues: the expected number of events (λ) must be approximated, and the dependence among the events being counted must be controlled.

In the analysis of headruns, that is, exact matching segments for two given sequences,

A_1, \dots, A_n and B_1, \dots, B_m , we let

$$\begin{aligned} D_i &= I(A_i = B_i) \\ X_1 &= \prod_{i=1}^t D_i \\ X_i &= (1 - D_{i-1}) \prod_{j=0}^{t-1} D_{i+j}, \quad i \geq 2 \end{aligned}$$

Here, the X_i 's are a set of dependent Bernoulli variables, each of which models the event that a "clump" or headrun of length t begins at position i . The index set for the X_i 's is $I = \{1, 2, \dots, n - t + 1\}$ and we assume each X_i is independent of $\{X_j\}$ where $j \notin J_i$. Let R_n be the longest headruns and let $W = \sum_{i \in I} X_i$, which is the total number of headruns longer than length t . Arratia (1990) has shown that for $t = \log_{1/p}[n(1 - p)] + c$,

$$|\Pr\{W = 0\} - e^{-\lambda}| = |\Pr\{R_n < t\} - e^{-\lambda}| < O\left(\frac{\log(n)}{n}\right)$$

where

$$\lambda = p^t + (n - t)(1 - p)p^t$$

In addition, Arratia et al. (1990) derived distribution for the length of longest consecutive run, contained within the first n tosses, in which the fraction of heads $\geq \alpha$, and the distribution for the length of the longest "quality α " matching consecutive segment common to the two sequences $A_1 \dots A_m$ and $B_1 \dots B_n$.

For the general scoring scheme where only $E\{s(A, B)\} < 0$, and $s^* = \max s(a, b) > 0$ are required, Dembo et al. (1994) studied the distribution for the Maximal Segment Pair (MSP), as defined in Altschul (1991). This is the pair of equal length segments that, when

aligned, have the greatest aggregate score. Given two independent random sequences with letter probabilities $\tilde{p} = \{p_1, \dots, p_r\}$ of length m and n , let λ be the unique positive solution to $\sum_{i,j} p_i p_j e^{\lambda s_{i,j}} = 1$. Applying the Chen-Stein theorem, the number of distinct local alignments with score at least x is approximately Poisson distributed with mean $K m n e^{-\lambda x}$, so for large x , the MSP score M has distribution

$$\Pr\{M \geq x\} \approx 1 - e^{-K m n e^{-\lambda x}} \approx K m n e^{-\lambda x}$$

or

$$\Pr\left\{M \geq \frac{\log(nm)}{\lambda} + x\right\} \leq K e^{-\lambda x} \quad (1.2)$$

where K is an explicit calculable parameter depending on s_{ij} and \tilde{p} (Karlin et al., 1990). Therefore, an alignment of segments from the two sequences would have an MSP score statistically significant at the α level, if M exceeds $(\log nm)/\lambda^* + x_0$ where x_0 is determined so that $K^* e^{-\lambda^* x_0} = \alpha$. For database searches using a simple sequence, the number of distinct "locally optimal" MSPs with score at least S is expected to occur by chance approximately $K N e^{-\lambda S}$ times where N is the product of query length and the database length in residues.

1.4.4.3 Gapped Pairwise Alignment

Although computational experiments strongly suggest equation (1.2) remain valid for the gapped case (Waterman and Vingron, 1994ab, Mott, 1992), the proof for the theorem became much less tractable mathematically. Neuhauser (1994) obtained an approximation for restricted scoring, in the case that a fixed number of gaps is allowed but aligned letters

are required to match. The form of the approximation under general scoring was worked out only recently by Siegmund and Yakir (2000).

In their notation, given two finite sequences $x = \{x_1 x_2 \dots x_m\}$ and $y = \{y_1 y_2 \dots y_n\}$ with $x_i, y_j \in A$. Assume that x and y are independently distributed with $P_o(x_i = \alpha) = \mu_\alpha$ for all i , and $P_o(y_j = \beta) = v_\beta$ for all j . Assume $E_o K(x, y) < 0$ and $P_o \{K(x, y) > 0\} > 0$ for general scoring. A candidate alignment is defined as $z = \{(i_t, j_t) : 1 \leq t \leq k\}$ for some $1 \leq i_1 < i_2 < \dots < i_k \leq m$ and $1 \leq j_1 < j_2 < \dots < j_k \leq n$ such that x_{i_t} and y_{j_t} are aligned for all $t = 1, \dots, k$. Now, we associate a score $S_z = S_z(x, y)$ with each candidate alignment z . To calculate $P_o(\max_z S_z \geq b)$ where b is the observed value of the score for the best alignment and P_o is the null probability for large values of b, m and n , Siegmund and Yakir (2000) rewrote this probability as a penalized log-likelihood ratio

$$P_o \left\{ \max_z [l_z - g(z)] \geq a \right\}$$

where $l_z = \theta^* \sum_{t=1}^k K(x_{i_t}, y_{j_t})$ and θ^* is the solution to $\psi(\theta) = \log E_o e^{\theta K(x, y)} = 0$. They then approximated the probability under two scenarios:

1. assume that all unaligned letters lie in at most a fixed number j of gaps.
2. the maximum number of gaps is not fixed, but each gap is assessed a cost Δ , "gap open" cost in addition to the cost δ "gap extension" cost.

In the first case, $g(z) = \theta^* \delta l$ and in the second case $g(z) = \theta^* (\delta l + \Delta j)$ where l is the total number of unaligned letters. In the second case, the approximation obtained can be turned into a Poisson approximation, therefore showing that the conjectured form

of approximation, the same as in the ungapped case, is correct, at least when the technical hypothesis concerning Δ is satisfied. Therefore, the theorem from Siegmund and Yakir (2000) can be viewed as a theoretical justification for the practice of applying extreme value theorem from ungapped alignment to gapped case. Siegmund and Yakir (2000) pointed out that the condition for Δ serves the purpose of keeping the problem well within the so-called "logarithmic domain". In the linear region, however, the theorem may not be true, since if the cost Δ of initiating a new gap is fixed while m and n become large, there may be cases when the score under the null hypothesis can be improved asymptotically by candidate alignment with a very large number of gaps.

Unlike in the ungapped case, the statistical parameters λ and K in the gapped case however, are no longer supplied by theory. They must be estimated using comparisons of either simulated or real but unrelated sequences. Letting $\lambda = \log \frac{1}{p}$, $x = c$, Waterman et al. (1994) rewrote equation (1.2) as

$$\Pr \{M \leq t = \log_{1/p}(nm) + c\} \approx e^{-\gamma mnp^t}$$

The goal then is to estimate parameters γ and p . The most intuitive to do this is by direct estimation. First, for a given t , calculate the empirical distribution function of optimal local alignment scores for many pairs of statistically independent sequences by computing the fraction of alignments with scores less than t . Taking a $\log(-\log(\text{data}))$ transformation and plotting against t gives a straight line, and estimating the intercept and slope would give estimates for γ and p .

The direct estimation method however, usually requires large number of sequence

alignments, say more than 1000, to derive the parameters. The more powerful method introduced by Waterman and Vingron (1994) is the declumping estimation method. First, from a few sequence comparisons, calculate $H_{(1)}, H_{(2)}, \dots, H_{(N)}$. In the example of this paper, only about 10 sequence comparisons where approximately 300 suboptimal solutions for each pair was collected, was suffice to produce accurate estimation. The idea is that the mean $\lambda_g(t) = E[W(t)] = \gamma mnp^t$ where

$$W(t) = \text{number of alignment clumps of score greater than or equal to } t$$

can be estimated as the average number of $H_{(i)}$ exceeding a threshold t . Proper transformation and plotting this average against t would then again give accurate estimates for γ and p .

Another approach is the maximum likelihood estimation method. First, with given distribution function and therefore density function for M , the likelihood can be written and solved for γ and p . Mott (1992) fitted an extreme value distribution by MLE to the scores from a database search using a Smith-Waterman algorithm. Other early attempt of approximating statistical parameters using scores from database searches includes Smith et al. (1985), Coulson et al. (1987) and Collins et al. (1990). All of the methods noted above provides virtually identical results, while the declumping methods is notably faster to implement.

1.4.4.4 Profile Analysis

Goldstein and Waterman (1994) studied the statistical distribution for the comparison of an ungapped multiple alignment profile with a simple sequence. In their notation, suppose we have aligned sequences

$$l_1 = l_{11}l_{12}...l_{1m}$$

$$l_2 = l_{21}l_{22}...l_{2m}$$

...

$$l_{N1} = l_{N1}l_{N2}...l_{Nm}$$

and we wish to compare it to a sequence $L_1, L_2, \dots, L_{n+m-1}$ with letters identically distributed over an alphabet A with size d . We summarize the statistics of the letters at position i , $1 \leq i \leq m$, in order to form a 'weighted average sequence': we form the quantity $\mathbf{f}_i = (f_{i1}, \dots, f_{id})$ where $f_{i\alpha} = \sum_{k=1}^N \frac{I\{l_{ki}=\alpha\}}{N}$, which is the fraction of α 's in column i . For example, for DNA sequences, $f_i = (f_{iA}, f_{iC}, f_{iG}, f_{iT})$. Next, the profile table is formed by $P = \{P_i\}_{\{i: 1 \leq i \leq m\}}$, where the weighted averages $P_i(l) = \sum_k s(l, k) f_{ik}$. This is essentially the average method first proposed by Gribskov et al. (1987). Here, each P_i measures the similarity between the letter l and the profile statistics at position i and $s(l, k)$ is score for aligning l, k from substitution matrix. Then the scores constructed for aligning a profile P with a sequence $l = l_1l_2...l_{j+m-1}$ is

$$X_j = \sum_{i=1}^m P_i(l_{i+j-1}) \quad j = 1, 2, \dots, n$$

To determine where in a sequence the entire profile best fits, we are interested in the maximum standardized profile score over all sets of m consecutive letters:

$$M_n = \max_{1 \leq j \leq n} Y_j$$

where

$$Y_j = \frac{X_j - E X_j}{\sqrt{Var X_j}}$$

Note that if all L_i 's are i.i.d., then the X_i 's are identically distributed. Goldstein and Waterman (1994) applied the Central Limit and large deviations theorem to show that the Y_j s can be approximated by a set of m -dependent normals as the size of the profile $m \rightarrow \infty$. They then showed that under appropriate conditions, the Chen-Stain theorem can be applied to approximate the maximum by the extreme value distribution and provide an error bounds that assess the quality of the approximation. We next state this important theorem:

Theorem 1.2 *Let L be a sequence, $L_1, L_2, \dots, L_{n+m-1}$ composed of independent and identically distributed letters over an alphabet A . Suppose that the profile tables P satisfy the following conditions:*

$$\sup_{n, 1 \leq i \leq m} ||P_i|| = k < \infty \quad (1.3)$$

and there exist $C > 0$ such that

$$Var P_i \geq C \text{ for all } n \text{ and } 1 \leq i \leq m \quad (1.4)$$

and the maximum column correlation is bounded strictly by 1:

$$\rho = \sup_{1 \leq i \leq \infty} \{|\rho_\delta| : 1 \leq n < \infty \text{ and } 1 \leq \delta < m\} < 1 \quad (1.5)$$

Let M_n be the maximum profile score,

$$M_n = \max_{1 \leq j \leq n} Y_j$$

and for given x , let

$$\lambda_n = n \Pr(Y_1 > u_n) \text{ with } u_n = \frac{x}{a_n} + c_n$$

where

$$a_n = \sqrt{2 \log n} \text{ and } c_n = \sqrt{2 \log n} \left\{ 1 - \frac{\frac{1}{2}(\log 4\pi + \log \log n)}{2 \log n} \right\}$$

With $0 \leq \rho < 1$, suppose that $m \asymp n^k$ where $k \in (0, \frac{1-\rho}{1+\rho})$, then

$$|\Pr\{a_n(M_n - c_n) \leq x\} - e^{-\lambda_n}| = o(n^{-\gamma}) \text{ for every } \gamma \in (0, \frac{1-\rho}{1+\rho} - k)$$

To prove this theorem, the authors constructed dependent binomial variables $B_j = I(Y_j > u_n)$ for $j \in I = \{1, 2, \dots, n\}$ and their sum $W_n = \sum_{i \in I} B_j$. Then, they showed $\lambda_n \rightarrow e^{-x}$ since each Y_i behaves like a normal random variable in the tail, and that b_1 and b_2 converges to 0 at $o(n^{-\gamma})$ rate.

Until now, when gaps are allowed, there is no corresponding analytic theory to estimate statistical significance of a profile compared with a simple sequence, under the general scoring scheme. The most widely used BLAST estimates λ_{pg} (λ for gapped profile) by first constructing amino acid scores within each column of a profile to the same "scale" (i.e. with the same ungapped lamda λ_u) as those for a standard amino acid substitution matrix. More specifically, for ungapped alignments, any substitution matrix takes the form $s_{ij} = \log \frac{q_{ij}}{p_i p_j} / \lambda_u$. For a profile, each column has its own unique set of amino acid target frequencies q_i , a scores for this column then may be constructed to the same "scale" by

using the formula $s_i = (\log \frac{q_i}{p_i})/\lambda_u$. BLAST then uses the same position-independent gap costs, and applies the same λ_u corresponding to the standard substitution matrix. Altschul et al. (1997) provides simulation experiments to show this practice provides accurate approximation for λ_{pg} .

1.5 Synopsis of Research

The Human Genome project has provided us with unprecedented wealth of biological data. Sequence comparison methods will be an important tool to help us translate genomic information into comprehensive understanding of biological systems. Assessing statistical significance and determining how high a score may be expected to occur by chance have always been a central question. Most recent works on sequence analysis as reviewed in this chapter have focused on i.i.d. sequences. However, biological sequences rarely satisfy the i.i.d. assumption. Therefore, in this research, we extend the current theories to account for the non-homogeneities, dependencies and the gapping that are inherent in biological sequences.

In chapter two, we study the headruns problem, or approximation for statistical distribution of fixed exact local alignment scores, under a general setup where non-homogenous matching probability and Markov dependency among the positions on the sequences are allowed. More specifically, we apply the Chen-Stein method of Poisson approximation to study alignment scores under these general assumptions: (1) the matching probabilities along the sequences are possibly different at each position, (2) in addition to the first case,

the underlying positions along the sequences are assumed to have first order Markovian dependent structure. In each of these situations, we will find an approximation formula for the expected number of events λ_n , calculate b_1 , b_2 indicating the dependency among the events being counted, and show that the error bounds converge to zero asymptotically.

In chapter three, we study scores derived from comparing multiple alignment profiles with sequences in protein databases. We derive asymptotic distribution for the maximum of profile scores for the cases (1) when means and variances of the scores are known (2) when mean and variance of the scores are unknown. Unlike previous studies (Goldstein et al., 1994), we put the profile in a random context with distributional properties. We study profile scores while assuming sequences in the profiles are random and they are compared to a simple i.i.d. sequence.

First, without any gaps in the profile, the tails of the profile scores can be approximated by standard normal distributions. The proof for this approximation is an application of the Central Limit and Large Deviation theorems to triangular arrays of variables. Next, accommodating the possibility of a few gaps in the profiles, we model the profile scores in the tails with normal mixture distribution rather than the standard normal distribution. This allows more flexibility and better approximation in that parameters from the mixture distribution can then be estimated to provide a better fit to data.

Next, to show the maximum of profiles scores can be approximated by a modified extreme value distribution of the third type, we will need to know the normalizing constants corresponding to extreme value theory for normal mixture distribution. Both analytical formula and numerical approximations are discussed in section 3.3. The accuracy of the

analytic formula is evaluated in table 3.1 and 3.2. Letting $B_j = I(Y_j > u_n)$ and $W_n = \sum_{j=1}^n B_j$, where $u_n = \frac{y}{a_n} + c_n$, a_n, c_n are normalizing constants, we next apply the Chen-Stein theorem and calculate b_1 and b_2 . In the case when the means and variances of the profile scores are not known, the proof is more complicated in that asymptotic orders of $\frac{s_n}{\sigma}$ and $\bar{X}_n - \mu\sqrt{m}$ needs to be estimated. Note that when the means and variances for the profile scores are known, the set of X_{nj} s is a set of m -dependent random variables. When means and variances of the profile scores are not known, however, the set of X_{nj} s are more dependent than m -dependent, since the same profile is involved with all the X_{nj} s. These dependencies need to be taken correctly to calculate the orders for $\frac{s_n}{\sigma}$ and $\bar{X}_n - \mu\sqrt{m}$. It is then shown that $\frac{s_n}{\sigma} \sim O_p\left(\sqrt{\frac{m}{n}}\right)$ and $\bar{X}_n - \mu\sqrt{m} \sim O_p\left(\sqrt{\frac{m}{n}}\right)$ where m is number of column for the profile and n is the number of residues for the simple sequence. Finally, we show that the maximum of normalized profile scores converges to a modified extreme value distribution:

$$\Pr\{a_n(M_n - c_n) \leq y\} \rightarrow \exp\left(-e^{-y} \left\{1 + \frac{1-\varepsilon}{c\varepsilon} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon}\right)^{c^2-1} e^{-y(c^2-1)}\right\}\right)$$

where $M_n = \max_{1 \leq j \leq n} Y_j$, $Y_j = \frac{X_{nj} - \bar{X}_n}{s_n}$ and $s_n^2 = \frac{1}{n[1+(c^2-1)\varepsilon]} \sum_{j=1}^n (X_{nj} - \bar{X}_n)^2$

Next in chapter four, we demonstrate the utility of our theory by applying it to a real data set. The Ig profile for Immunoglobulin domain (accession number PF00047) and the UniProt protein database in FASTA format are first downloaded from Washington University at St. Louis and European Bioinformatics Institute websites. To estimate parameters c and ε on the right side of above equation, we applied the Maximum Likelihood method

introduced by Mott (1992). Initially, without knowing c and ε , we only know $U_n = \frac{X_{nj} - \bar{X}_n}{\sqrt{\sum_{j=1}^n (X_{nj} - \bar{X}_n)^2 / n}}$, so we re-write the distribution function above in terms of U_n and maximize the log-likelihood with respect to the parameters to get the maximum likelihood estimates \hat{c} and $\hat{\varepsilon}$. Then, we calculate M_n based on \hat{c} and $\hat{\varepsilon}$. According to the theorem, the j th largest normalized maximum score should be approximately equal to the $j/(N+1)$ th quantile from the extreme value distribution $G(y) = \exp\left(-e^{-y} \left\{1 + \left(\frac{\sqrt{4\pi \log n \varepsilon}}{n \varepsilon e^y}\right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon}\right\}\right)$. Since the sequences in the database have different lengths, we compared to the quantiles from $e^{-e^{-y}}$ instead. The normalized maximum scores were ordered and plotted against the quantiles from the extreme value distribution. The Q-Q plot in Figure 4.4 shows very good fit for our approximation.

Finally, in chapter five, we summarize this research and point to directions for future research. In both pairwise alignment and profile analysis, studies that model the heterogeneities and dependencies inherent in biological sequences are needed. We discuss possible extensions in detail and outline some numerical methods that can be applied to assess the quality of approximations in these cases.

CHAPTER 2

PAIRWISE ALIGNMENT – ANALYSIS OF HEADRUNS

2.1 Introduction

The alignment score which measures the similarity between two sequences will depend on how the sequences were aligned. Here we consider the case when we are given two sequences $A_1, A_2, \dots, A_n, B_1, B_2, \dots, B_n$ in a fixed alignment, that is, the sequences have been aligned already and no more shifts are needed. The alignment of two sequences can be done by moving one sequence on top of the other sequence until certain number of letters, say 20, are found to be the same, then chopping off sections of the sequences in which there is no corresponding letters from the other sequence. To remove bias, the first 20, or whatever number that was used in the matching criteria of the aligned segment is also often deleted. We consider the local alignment score $H(A, B) = \max \{S(I, J) : I \subset A, J \subset B\}$ which is the score for the best matching region along the two sequences. For exact matching with no insertions or deletions allowed, $H(A, B)$ becomes $R_n = \max\{m : A_{i+r} = B_{i+r} \text{ for } r = 1 \text{ to } m, 0 \leq i \leq n - m\}$, which is the longest perfect matching subsequence between **A** and **B**. R_n behaves like longest

run of heads when flipping a coin n times with probability $p = \Pr(A_i = B_i)$ for heads each time. We are interested in approximating the probability $\Pr(R_n < t)$ and find the associated error bounds of the approximation.

The previous works discussed in literature review have shown that the growth of R_n is proportional to $\log(n)$ so that $\lim R_n / \log(n) \rightarrow 1$ with probability 1. Also, as described before, Poisson approximation via Chen-Stein Theorem provides a way to approximate $\Pr(R_n < t)$ with explicit error bounds: for $t = \log_{1/p}[n(1 - p)] + c$,

$$|\Pr\{W = 0\} - e^{-\lambda}| = |\Pr\{R_n < t\} - e^{-\lambda}| < O\left(\frac{\log n}{n}\right)$$

where

$$\lambda = p^t + (n - t)(1 - p)p^t, \quad p \equiv \Pr(A_i = B_i)$$

In the previous analysis of headruns, the matching probability $p = \Pr(D_i = 1) = \Pr(A_i = B_i)$ was assumed to be independent of position i , and the positions along the sequences were assumed to be independent. Our objective is to extend Chen-Stein Theorem to situations with more general assumptions, more specifically, we study the headruns problem under the scenarios of:

- (1) the matching probabilities are different at each position, that is we assume $p_i = \Pr(A_i = B_i)$.
- (2) In addition to the first case, the underlying positions along the sequences are assumed to have Markov dependent structure. That is, our parameter set also includes $\alpha_i = \Pr(A_{i+1} = B_{i+1} \mid A_i = B_i)$.

2.2 Nonhomogeneous Matching Probabilities, Independent Positions

First, we review the Chen-Stein Theorem:

Theorem 2.1 (*Chen-Stein*) *Let I be an index set, and for each $i \in I$, and let X_i be an indicator random variable. The total number of occurrence of events is*

$$W = \sum_{i \in I} X_i$$

For each $i \in I$, let J_i be the set of dependence of i , and assume that

$$X_i \text{ is independent of } \{X_j\}, j \notin J_i.$$

Let Z be a Poisson random variable with $E(Z) = E(W) = \lambda$. Then,

$$\|W - Z\| \leq 2(b_1 + b_2) \frac{1 - e^{-\lambda}}{\lambda} \leq 2(b_1 + b_2)$$

where

$$b_1 \equiv \sum_{i \in I} \sum_{j \in J_i} E(X_i)E(X_j)$$

and

$$b_2 \equiv \sum_{i \in I} \sum_{i \neq j \in J_i} E(X_i X_j)$$

in particular,

$$|\Pr(W = 0) - e^{-\lambda}| \leq (b_1 + b_2) \frac{1 - e^{-\lambda}}{\lambda} \quad (2.1)$$

We next state the theorem for homogenous matching probabilities. We show the following proof from Waterman (1995) for completeness.

Theorem 2.2 *Let A_1, A_2, \dots and B_1, B_2, \dots be two identically distributed sequences with fixed local alignment score*

$$R_n = \max\{m : A_{i+r} = B_{i+r} \text{ for } r = 1 \text{ to } m, 0 \leq i \leq n - m\}$$

and $\Pr(A_i = B_i) = p$ for all i , then applying the Chen-Stein Theorem, we get for $t = \log_{1/p} [n(1 - p)] + c$,

$$|\Pr(R_n \geq t) - (1 - e^{-\lambda})| = O\left(\frac{\log n}{n}\right)$$

where $\lambda = p^t + (n - t)(1 - p)p^t$

Proof. Let $D_i = I(A_i = B_i)$ or $\Pr(D_i = 1) = 1 - \Pr(D_i = 0) = p$. Define X_i to be the event that a headrun of length t begins at position i . Then

$$X_1 = \prod_{i=1}^t D_i$$

$$X_i = (1 - D_{i-1}) \prod_{j=0}^{t-1} D_{i+j}, \quad i \geq 2$$

The index set is $I = \{1, 2, \dots, n - t + 1\}$ and the dependence set is $J_i = \{j \in I : |i - j| \leq t\}$.

This gives

$$\begin{aligned} \lambda &= E(W) = E\left(\sum_{i=1}^{n-t+1} X_i\right) = \sum_{i=1}^{n-t+1} E(X_i) \\ &= p^t + (n - t)(1 - p)p^t \end{aligned}$$

Due to the factor of $(1 - D_i)$ for declumping, X_i and X_j can not be 1 for $i \neq j$ where $i, j \in J_i$. Therefore

$$b_2 = \sum_{i \in I} \sum_{i \neq j \in J_i} E(X_i X_j) = 0$$

It remains to calculate b_1 where

$$\begin{aligned}
b_1 &= \sum_{i \in I} \sum_{j \in J_i} E(X_i)E(X_j) \\
&= p^t \sum_{j \in J_1} E(X_j) + \sum_{i=2}^{n-t+1} (1-p)p^t \sum_{j \in J_i} E(X_j) \\
&< p^t [2t(1-p)p^t + p^t] + (n-t)(2t+1) [(1-p)p^t]^2 \\
&= (2t+1)(1-p)^2 p^{2t} \left\{ n-t + \frac{1}{(1-p)^2(2t+1)} + \frac{2t+1-1}{(1-p)(2t+1)} \right\} \\
&= (2t+1)(1-p)^2 p^{2t} \left\{ n-t + \frac{1}{(1-p)^2(2t+1)} + \frac{1}{1-p} + \frac{-1}{(1-p)(2t+1)} \right\} \\
&= (2t+1)(1-p)^2 p^{2t} \left\{ n-t + \frac{1}{1-p} + \frac{1-(1-p)}{(1-p)^2(2t+1)} \right\} \\
&= (2t+1)(1-p)^2 p^{2t} \left\{ n-t + \frac{1}{1-p} + \frac{p}{(1-p)^2(2t+1)} \right\}
\end{aligned}$$

Applying the Chen-Stein Theorem,

$$|\Pr(W = 0) - e^{-\lambda}| \leq b_1 \min \{1, 1/\lambda\}$$

Now $W = 0$ iff there are no headruns $R_n \geq t$ or

$$\{W = 0\} = \{R_n < t\}$$

and

$$|\Pr(R_n < t) - e^{-\lambda}| \leq b_1 \min \{1, 1/\lambda\}$$

This equation can be restated as

$$|\Pr(R_n \geq t) - (1 - e^{-\lambda})| \leq b_1 \min \{1, 1/\lambda\}$$

For interesting probabilities, we want $\lambda = \lambda(t)$ to be bounded away from 0 and ∞ . The term of the mean λ to worry about is np^t , so having $t - \log_{1/p} n$ bounded is required. For

$t = \log_{1/p} [n(1-p)] + c$, we obtain

$$\begin{aligned}\lambda &= p^t + (n-t)(1-p)p^t \\ &= \frac{p^c}{n(1-p)} + (1 - \frac{t}{n})p^c \approx p^c\end{aligned}$$

and

$$\begin{aligned}& |\Pr(R_n < \log_{1/p} [n(1-p)] + c) - e^{-\lambda}| \\ & < \frac{2t+1}{n} p^{2c} \left\{ 1 - \frac{t}{n} + \frac{1}{n(1-p)} + \frac{p}{n(1-p)^2(2t+1)} \right\} \\ & = O\left(\frac{\log n}{n}\right)\end{aligned}$$

■

However, biological sequences rarely follow i.i.d. distributions. We relax the assumptions on the underlying distributions of the sequences next. Before we state the theorem where non-homogeneity for the matching probabilities is allowed, we calculate expectations/moment generating functions for some familiar distributions in the following lemma.

Lemma 2.1

(1) Let $X \sim \text{Beta}(\alpha, \beta)$, such that $f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$, then $E(X) = \frac{\alpha}{\alpha+\beta}$

(2) Let $X \sim \text{Normal}(0, \sigma^2)$, where $-\infty < X < u$, and $f(x|\sigma) = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}}{\int_{-\infty}^u \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} dx}$,
then $E(e^{sx}) = \frac{F_o(u/\sigma - \sigma s)}{F_o(u/\sigma)} e^{\frac{\sigma^2 s^2}{2}}$ where F_o is the d.f. for standard normal distribution.

Proof.

(1) The proof can be found in any standard statistics text such as Casella and Berger (1990).

(2) We start with

$$f(x|\sigma^2) = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}}{\int_{-\infty}^u \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} dx}$$

let $w = x/\sigma$, then

$$\begin{aligned} & \int_{-\infty}^u \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} dx \\ &= \int_{-\infty}^{u/\sigma} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw \\ &= F_o\left(\frac{u}{\sigma}\right) \end{aligned}$$

Now,

$$\begin{aligned} E(e^{sx}) &= \frac{\int_{-\infty}^u \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} e^{sx} dx}{\int_{-\infty}^u \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} dx} \\ &= \frac{\int_{-\infty}^u \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} e^{sx} dx}{F_o\left(\frac{u}{\sigma}\right)} \end{aligned}$$

let $y = x - u$, then

$$\begin{aligned} F_o\left(\frac{u}{\sigma}\right) E(e^{sx}) &= \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y+u)^2}{2\sigma^2} + s(y+u)} dy \\ &= \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2} \{(y+u)^2 - 2\sigma^2 s(y+u)\}} dy \\ &= \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2} (y+u-\sigma^2 s)^2} dy \times e^{\frac{\sigma^2 s^2}{2}} \end{aligned}$$

let $z = \frac{y+u-\sigma^2 s}{\sigma}$, then

$$\begin{aligned} F_o\left(\frac{u}{\sigma}\right) E(e^{sx}) &= \int_{-\infty}^{\frac{u}{\sigma} - \sigma s} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \times e^{\frac{\sigma^2 s^2}{2}} \\ &= e^{\frac{\sigma^2 s^2}{2}} F_o\left(\frac{u}{\sigma} - \sigma s\right) \end{aligned}$$

therefore,

$$E(e^{sx}) = \frac{F_o(u/\sigma - \sigma s)}{F_o(u/\sigma)} e^{\frac{\sigma^2 s^2}{2}}$$

■

In the next theorem, we apply the Chen-Stein Theorem to approximate P values for fixed alignment scores R_n when some variation of the matching probabilities is allowed.

Theorem 2.3 *Let A_1, A_2, \dots and B_1, B_2, \dots be two sequences with fixed local alignment score*

$$R_n = \max\{m : A_{i+r} = B_{i+r} \text{ for } r = 1 \text{ to } m, 0 \leq i \leq n - m\}$$

and $\Pr(A_k = B_k) = p_k$ such that $p_k \sim F(\cdot)$ with $E(p_k) = p$ for all k . Assume $\tilde{p} = (p_1, \dots, p_n)$ is a vector of i.i.d. matching probabilities, then for $t = \log_{1/p} [n(1 - p)] + c$,

$$|\Pr(R_n \geq t) - (1 - e^{-\lambda})| = O\left(\frac{\log n}{n}\right)$$

where $\lambda = p^t + (n - t)(1 - p)p^t$. For example,

(1) *if $p_k \sim \text{i.i.d. Beta}(\alpha, \beta)$, $\beta = O(\alpha)$, then $p = E(p_k) = \frac{\alpha}{\alpha + \beta}$*

(2) *if $p_k = \mu e^{\delta_k}$, where $\delta_k \sim \text{i.i.d. truncated Normal}(0, \sigma^2)$, $-\infty < \delta_k < u$, then*

$$p = E(p_k) = \mu \frac{F_o(u/\sigma - \sigma)}{F_o(u/\sigma)} e^{\frac{\sigma^2}{2}} \text{ where } F_o \text{ is the d.f. for standard normal distribution.}$$

Proof. Let

$$D_k = I(A_k = B_k) \text{ with } \Pr(D_k = 1) = p_k \text{ and } \tilde{p} = (p_1, p_2, \dots, p_n)$$

$$X_1 = \prod_{k=1}^t D_k$$

$$X_i = (1 - D_{i-1}) \prod_{k=i}^{i+t-1} D_k, \quad i \geq 2$$

then the index set is $I = \{1, 2, \dots, n + t - 1\}$, and the dependent set is $J_i = \{j \in I : |i - j| \leq t\}$.

Now, let $W = \sum_{i=1}^{n-t+1} X_i$, then

$$E(X_1|\tilde{p}) = \prod_{k=1}^t p_k$$

$$E(X_i|\tilde{p}) = (1 - p_{i-1}) \prod_{k=i}^{i+t-1} p_k, \quad i \geq 2$$

$$E(W|\tilde{p}) = \sum_{i=1}^{n-t+1} E(X_i|\tilde{p})$$

$$= \prod_{k=1}^t p_k + (1 - p_1) \prod_{k=2}^{t+1} p_k + (1 - p_2) \prod_{k=3}^{t+2} p_k + \dots + (1 - p_{n-t}) \prod_{k=n-t+1}^n p_k$$

Assume $\tilde{p} = (p_1, \dots, p_n)$ is a vector of i.i.d. matching probabilities, then as before,

$$EX_1 = E[E(X_1|\tilde{p})] = p^t$$

$$EX_i = E[E(X_i|\tilde{p})] = (1 - p)p^t \quad i \geq 2$$

and

$$\lambda = E[E(W|\tilde{p})]$$

$$= p^t + (n - t)(1 - p)p^t$$

$$b_1 = \sum_{i \in I} \sum_{j \in J_i} E(X_i)E(X_j)$$

$$< (2t + 1) [(1 - p)p^t]^2 \left\{ n - t + \frac{1}{1 - p} + \frac{p}{(1 - p)^2(2t + 1)} \right\}$$

and

$$b_2 = \sum_{i \in I} \sum_{i \neq j \in J_i} E(X_i X_j) = 0$$

Therefore, again,

$$\begin{aligned} & |\Pr(W = 0) - e^{-\lambda}| \\ &= |\Pr(R_n < \log_{1/p} [n(1-p)] + c) - e^{-\lambda}| \\ &< \frac{2t+1}{n} p^{2c} \left\{ 1 - \frac{t}{n} + \frac{1}{n(1-p)} + \frac{p}{n(1-p)^2(2t+1)} \right\} \\ &= O\left(\frac{\log n}{n}\right) \end{aligned}$$

For the examples, (1) follows directly. For (2), let $p_k = \mu e^{\delta_k}$ where $\delta_k \sim i.i.d. \text{Normal}(0, \sigma^2)$, since $0 < p_k < 1$, we have

$$-\infty < \log p_k < 0$$

$$-\infty < \log p + \delta_k < 0$$

$$-\infty < \delta_k < -\log p = u$$

So δ_k follow the truncated normal distribution, that is, $\delta_k \sim N(0, \sigma^2)$ and $-\infty < \delta_k < u$, therefore, by the lemma above, $p = E(p_k) = \mu \frac{F_o(u/\sigma - \sigma)}{F_o(u/\sigma)} e^{\frac{\sigma^2}{2}}$ where F_o is the d.f. for standard normal distribution. ■

2.3 Nonhomogeneous Matching Probabilities, Markovian Dependent Positions

We now investigate the case when in addition to non-homogeneous matching probabilities, the positions along the sequences also exhibit Markovian dependency. We show that

the Chen-Stein Theorem can still be applied to approximate the tail probabilities for R_n .

Theorem 2.4 *Let A_1, A_2, \dots and B_1, B_2, \dots be two sequences with the score function for fixed alignment*

$$R_n = \max\{m : A_{i+k} = B_{i+k} \text{ for } k = 1 \text{ to } m, 0 \leq i \leq n - m\}$$

Let $D_k = I(A_k = B_k)$ be Markov dependent with $\Pr(D_k = 1) = p_k$, $\Pr(D_{k+1} = 1 | D_k = 1) = \alpha_k$, so that $\tilde{\alpha} = (\alpha_1, \dots, \alpha_n)$ and $\tilde{p} = (p_1, \dots, p_n)$ are transition and matching probabilities respectively. Now, suppose $\tilde{\alpha} = (\alpha_1, \dots, \alpha_n)$ is a vector of i.i.d. probabilities with common c.d.f. $F(\cdot)$ and common mean α , suppose $\tilde{p} = (p_1, \dots, p_n)$ is a vector of probabilities with common mean p , such that $\frac{p(1-\alpha)}{1-p} \leq 1$, then for $t = \log_{1/\alpha} np(1-\alpha) + c$,

$$|\Pr(R_n < t) - e^{-\lambda}| = O\left(\frac{\log n}{n}\right)$$

where $\lambda = p\alpha^{t-1} \{1 + (n-t)(1-\alpha)\}$

Proof. let $D_k = I(A_k = B_k)$ with $p_k = \Pr(D_k = 1)$, $\alpha_k = \Pr(D_{k+1} = 1 | D_k = 1)$, $\beta_k = \Pr(D_{k+1} = 1 | D_k = 0)$, then

$$\alpha_k p_k + \beta_k (1 - p_k) = p_{k+1}$$

$$\beta_k = \frac{p_{k+1} - \alpha_k p_k}{1 - p_k}$$

Let $X_1 = \prod_{k=1}^t D_k$, $X_i = (1 - D_{i-1}) \prod_{k=i}^{i+t-1} D_k$, then

$$\begin{aligned}
E(X_1|\tilde{\alpha}) &= p_1 \prod_{k=1}^{t-1} \alpha_k \\
E(X_i|\tilde{\alpha}) &= (1 - p_{i-1}) \beta_{i-1} \prod_{k=i}^{i+t-2} \alpha_k \\
&= (1 - p_{i-1}) \frac{p_i - \alpha_{i-1} p_{i-1}}{1 - p_{i-1}} \prod_{k=i}^{i+t-2} \alpha_k \\
&= p_i \prod_{k=i}^{i+t-2} \alpha_k - p_{i-1} \prod_{k=i-1}^{i+t-2} \alpha_k \\
E(X_{n-t+1}|\tilde{\alpha}, \tilde{p}) &= p_{n-t+1} \prod_{k=n-t+1}^{n-1} \alpha_k - p_{n-t} \prod_{k=n-t}^{n-1} \alpha_k
\end{aligned}$$

Since $\tilde{\alpha} = (\alpha_1, \dots, \alpha_n)$ is a vector of i.i.d. transition probabilities, assuming p_i is independent of $\{\alpha_k : k \geq i\}$, and $E(\alpha_k) = \alpha$, $E(p_k) = p$ for all k , then

$$\begin{aligned}
E(X_1) &= E[E(X_1|\tilde{\alpha}, \tilde{p})] = p\alpha^{t-1} \\
E(X_i) &= E\{E(X_i|\tilde{\alpha}, \tilde{p})\} = p\alpha^{t-1} - p\alpha^t = p\alpha^{t-1}(1 - \alpha)
\end{aligned}$$

Now,

$$\begin{aligned}
\lambda &= E(W) = \sum_{i=1}^{n-t+1} EX_i \\
&= p\alpha^{t-1} + (n-t)p\alpha^{t-1}(1 - \alpha) \\
&= p\alpha^{t-1} \{1 + (n-t)(1 - \alpha)\}
\end{aligned}$$

To bound $\lambda(t)$ away from 0 and ∞ , the term to worry is $n\alpha^{t-1}p(1-\alpha)$, so take $t = \log_{1/\alpha} np(1-\alpha) + c$, then

$$\begin{aligned}\alpha^{t-1} &= \frac{\alpha^c}{np(1-\alpha)} \\ \lambda &= \frac{\alpha^c p}{np(1-\alpha)} \{1 + (n-t)(1-\alpha)\} \\ &= \alpha^c \left\{ \frac{1}{n(1-\alpha)} + \left(1 - \frac{t}{n}\right) \right\} \\ &\approx \alpha^c\end{aligned}$$

To calculate b_2 , note that in the expression for X_i , because of the factor $(1 - D_{i-1})$ for declumping, that is, by requiring mismatch at $i - 1$ th position, we count only the longest headrun among the group of headruns that occur next to each other, therefore, X_i and X_j can not be 1 for $i \neq j$ where $i, j \in J_i$. This implies that

$$b_2 = \sum_{i \in I} \sum_{i \neq j \in J_i} E(X_i X_j) = 0$$

We now calculate b_1

$$\begin{aligned}b_1 &= \sum_{i \in I} \sum_{j \in J_i} E(X_i) E(X_j) \\ &= p\alpha^{t-1} \sum_{j \in J_1} E(X_j) + \sum_{i=2}^{n-t+1} p\alpha^{t-1}(1-\alpha) \sum_{j \in J_i} E(X_j) \\ &= p\alpha^{t-1} \left\{ 2t(1-\alpha)p\alpha^{t-1} + p\alpha^{t-1} \right\} + (n-t)(2t+1) \left\{ (1-\alpha)p\alpha^{t-1} \right\}^2 \\ &= (2t+1)(1-\alpha)^2 p^2 \alpha^{2t-2} \left\{ (n-t) + \frac{1}{(2t+1)(1-\alpha)^2} + \frac{2t+1-1}{(2t+1)(1-\alpha)} \right\}\end{aligned}$$

$$\begin{aligned}
&= (2t+1)(1-\alpha)^2 p^2 \alpha^{2t-2} \left\{ \begin{aligned} &(n-t) + \frac{1}{(1-\alpha)^2(2t+1)} \\ &+ \frac{1}{1-\alpha} - \frac{1}{(2t+1)(1-\alpha)} \end{aligned} \right\} \\
&= (2t+1)(1-\alpha)^2 p^2 \alpha^{2t-2} \left\{ (n-t) + \frac{1}{1-\alpha} + \frac{1-(1-\alpha)}{(2t+1)(1-\alpha)^2} \right\} \\
&= (2t+1)(1-\alpha)^2 p^2 \alpha^{2t-2} \left\{ (n-t) + \frac{1}{1-\alpha} + \frac{\alpha}{(2t+1)(1-\alpha)^2} \right\}
\end{aligned}$$

Therefore, for $t = \log_{1/\alpha} np(1-\alpha) + c$,

$$\begin{aligned}
&|\Pr(W=0) - e^{-\lambda}| \\
&= |\Pr(R_n < \log_{1/\alpha} np(1-\alpha) + c) - e^{-\lambda}| \\
&\leq (2t+1)(1-\alpha)^2 p^2 \frac{\alpha^{2c}}{n^2 p^2 (1-\alpha)^2} \left\{ (n-t) + \frac{1}{1-\alpha} + \frac{\alpha}{(2t+1)(1-\alpha)^2} \right\} \\
&= (2t+1) \frac{\alpha^{2c}}{n} \left\{ \left(1 - \frac{t}{n}\right) + \frac{1}{n(1-\alpha)} + \frac{\alpha}{n(2t+1)(1-\alpha)^2} \right\} \\
&= O\left(\frac{\log n}{n}\right)
\end{aligned}$$

■

CHAPTER 3

MULTIPLE ALIGNMENT – PROFILE ANALYSIS

3.1 Introduction

We discuss the profile model here in more detail. Given a set of multiple alignment profile \mathbf{l}

$$\begin{aligned}l_1 &= l_{11}l_{12}...l_{1m} \\l_2 &= l_{21}l_{22}...l_{2m} \\&\dots \\l_{N1} &= l_{N1}l_{N2}...l_{Nm}\end{aligned}$$

Figure 3.1

and a long sequence $L_1, L_2, \dots, L_{n+m-1}$, the objective of profile analysis is to find the region along \mathbf{L} where the profile \mathbf{l} fits best and assign it a score according to certain scheme. Let Λ be the set of alphabet, in the simple case of DNA sequences, $\Lambda = \{ A, C, G, T \}$. In the case of protein sequences, Λ consists of twenty amino acids. We assume for the long sequence \mathbf{L} ,

(1) the letters $L_1, L_2, \dots, L_{n+m-1}$ are *i.i.d.*

(2) $(n_A, n_C, n_G, n_T) \sim \text{multinomial}(n + m - 1, \pi)$ where $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$

and the profile \mathbf{l} is random:

(3) for column $i, i = 1, \dots, m, \mathbf{n}_i = (n_{iA}, n_{iC}, n_{iG}, n_{iT}) \sim \text{multinomial}(N, \pi_i)$ where

$$\pi_i = (\pi_{iA}, \pi_{iC}, \pi_{iG}, \pi_{iT}) \text{ and } \frac{\sum_{i=1}^m \pi_i}{m} \longrightarrow \bar{\pi} \text{ with rate } \left| \sum_{i=1}^m \pi_i / m - \bar{\pi} \right| = O\left(\frac{1}{\sqrt{m}}\right).$$

(4) the sequences in the profiles are mutually independent, and each is independent of \mathbf{L} .

The statistical question is to assess whether the alignment score from the best matching region is significant enough to indicate association between the given sequence and the multiple alignment, that is, whether the sequence \mathbf{L} belongs to the protein family \mathbf{l} .

In the process of moving the profile along sequence \mathbf{L} , for each match we assign a profile score

$$X_{nj} = \frac{1}{\sqrt{m}} \sum_{i=1}^m P_i(L_{i+j-1}) \quad j = 1, 2, \dots, n \quad (3.1)$$

Here,

$$P_i(L) = \frac{1}{N} \sum_{\alpha} s(L, \alpha) n_{i\alpha} \quad (3.2)$$

is a weighted average that measures the similarity between the letter L and the profile statistics at position i , $s(L, \alpha)$ is score for aligning L and α from substitution matrix, and

$n_{i\alpha} = \sum_{k=1}^N I\{l_{ki} = \alpha\}$ is the number of α 's in column i . In the simplest case, for DNA

sequences for example, we use score $s(L, \alpha) = I(L = \alpha) = \begin{cases} 1 & \text{if } L = \alpha \\ 0 & \text{if } L \neq \alpha \end{cases}$. Then the

column score becomes

$$\begin{aligned}
P_i(L_j) &= P_{ij} \\
&= \frac{1}{N} \{I(L_j = A) n_{iA} + I(L_j = C) n_{iC} + I(L_j = G) n_{iG} + I(L_j = T) n_{iT}\} \\
&= \frac{1}{N} \sum_{\alpha \in \Lambda} I(L_j = \alpha) n_{i\alpha}
\end{aligned}$$

Their means and variances can be worked out easily: since for each column i , $\mathbf{n}_i = (n_{iA}, n_{iC}, n_{iG}, n_{iT}) \sim \text{multinomial}(N, \pi_i)$ where $\pi_i = (\pi_{iA}, \pi_{iC}, \pi_{iG}, \pi_{iT})$, we have

$$\begin{aligned}
E(n_{i\alpha}) &= N\pi_{i\alpha} \\
\text{Var}(n_{i\alpha}) &= N\pi_{i\alpha}(1 - \pi_{i\alpha}) \\
E(n_{i\alpha}^2) &= N^2\pi_{i\alpha}^2 + N\pi_{i\alpha}(1 - \pi_{i\alpha}) \\
\text{Cov}(n_{i\alpha}, n_{i\beta}) &= -N\pi_{i\alpha}\pi_{i\beta} \\
E(n_{i\alpha}n_{i\beta}) &= -N\pi_{i\alpha}\pi_{i\beta} + N^2\pi_{i\alpha}\pi_{i\beta}
\end{aligned}$$

So for the column score

$$\begin{aligned}
E(P_{ij}) &= \frac{1}{N} \sum_{\alpha} E[I(L_j = \alpha) n_{i\alpha}] \\
&= \sum_{\alpha} \Pr(L_j = \alpha) \pi_{i\alpha} \\
&= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} = \mu_i < \infty
\end{aligned}$$

Moreover,

$$\begin{aligned}
EP_{ij}^2 &= \frac{1}{N^2} E \left\{ \sum_{\alpha \in \Lambda} I(L_j = \alpha) n_{i\alpha} \sum_{\alpha \in \Lambda} I(L_j = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^2} E \left\{ \sum_{\alpha \in \Lambda} I(L = \alpha)^2 n_{i\alpha}^2 + \sum_{\alpha, \beta \in \Lambda} I(L = \alpha) I(L = \beta) n_{i\alpha} n_{i\beta} \right\} \\
&= \frac{1}{N^2} \left\{ \sum_{\alpha \in \Lambda} \Pr(L_j = \alpha) E(n_{i\alpha}^2) \right\} \\
&= \frac{1}{N^2} \left\{ \sum_{\alpha \in \Lambda} \pi_\alpha [N^2 \pi_{i\alpha}^2 + N \pi_{i\alpha} (1 - \pi_{i\alpha})] \right\} \\
&= \sum_{\alpha \in \Lambda} \pi_\alpha \pi_{i\alpha}^2 + \frac{1}{N} \sum_{\alpha \in \Lambda} \pi_{i\alpha} (1 - \pi_{i\alpha}) \pi_\alpha \\
(E P_{ij})^2 &= \left(\sum_{\alpha \in \Lambda} \pi_\alpha \pi_{i\alpha} \right)^2 = \sum_{\alpha \in \Lambda} \pi_\alpha^2 \pi_{i\alpha}^2 + \sum_{\alpha, \beta \in \Lambda} \pi_\alpha \pi_\beta \pi_{i\alpha} \pi_{i\beta}
\end{aligned}$$

So for each column

$$\begin{aligned}
Var P_{ij} &= EP_{ij}^2 - (EP_{i,j})^2 \\
&= \sum_{\alpha \in \Lambda} \pi_\alpha (1 - \pi_\alpha) \pi_{i\alpha}^2 - \sum_{\alpha, \beta \in \Lambda} \pi_\alpha \pi_\beta \pi_{i\alpha} \pi_{i\beta} + \frac{1}{N} \sum_{\alpha \in \Lambda} \pi_{i\alpha} (1 - \pi_{i\alpha}) \pi_\alpha \\
&= \sigma_i^2 < \infty
\end{aligned}$$

and for the profile, let $\mu = \frac{1}{m} \sum_{i=1}^m \mu_i$, and $\sigma^2 = \frac{1}{m} \sum_{i=1}^m \sigma_i^2$,

$$X_{nj} = \frac{1}{\sqrt{m}} \sum_{i=1}^m P_i(L_{i+j-1})$$

$$EX_{nj} = \frac{1}{\sqrt{m}} \sum_{i=1}^m \mu_i = \sqrt{m} \frac{\sum_{i=1}^m \mu_i}{m} = \sqrt{m} \mu$$

$$\begin{aligned}
Var X_{nj} &= \frac{1}{m} \sum_{i=1}^m Var (P_{ij}) \\
&= \frac{1}{m} \sum_{i=1}^m \left\{ \sum_{\alpha \in \Lambda} \pi_{\alpha} (1 - \pi_{\alpha}) \pi_{i\alpha}^2 - \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta} \right. \\
&\quad \left. + \frac{1}{N} \sum_{\alpha \in \Lambda} \pi_{i\alpha} (1 - \pi_{i\alpha}) \pi_{\alpha} \right\} \\
&= \frac{1}{m} \sum_{i=1}^m \sigma_i^2 = \sigma^2 < \infty
\end{aligned}$$

So we have a set of profile scores $X_{n1} = \frac{1}{\sqrt{m}}\{P_{11} + P_{22} + \dots + P_{mm}\}$, $X_{n2} = \frac{1}{\sqrt{m}}\{P_{12} + P_{23} + \dots + P_{m,m+1}\}$, ..., X_{nn} . To determine the significance of the best alignment score, we are interested in the distribution of maximum standardized profile score:

$$M_n^o = \max_{1 \leq j \leq n} Y_{nj}^o$$

where

$$Y_{nj}^o = \frac{X_{nj} - E X_{nj}}{\sqrt{Var X_{nj}}}$$

When the profile is random, μ and σ are unknown, so we are more interested in

$$M_n = \max_{1 \leq j \leq n} Y_{nj} \tag{3.3}$$

where

$$Y_{nj} = \frac{X_{nj} - \bar{X}_n}{s_n}$$

Here, \bar{X}_n and s_n are estimators for the corresponding population parameter.

As was discussed in literature review, assuming the profile **I** is given fixed, Goldstein (1994) et. al. have approximated the Y_{nj} s as a set of standard normal variables in the tail, used corresponding extreme value theory for normal random variables to approximate the

distribution for M_n^o , as well as employed Chen-Stein Theorem to work out the error bounds for the approximation.

In contrast, we assume the underlying profile is random with distributional properties discussed above and accommodate the possibilities of a few gaps (insertions and deletions). We derive the corresponding theory for the maximum M_n . Some modifications to the Goldstein paper are involved: (1) in section 3.2, we first show that in this case the tails of the profile scores can still be approximated by normal distribution using large deviation theorem for triangular arrays, (2) next, in section 3.3.1, to accommodate insertions and deletions in the multiple alignment profile, we show that the tail behavior of Y_{nj} s can be well modeled using two component normal mixture model. (3) we then find the analytic formula for normalizing constants and work out the corresponding extreme value theory for normal mixture distribution. (4) Finally, in section 3.5.2, when the profile \mathbf{l} is random, the means and variances of the X_{nj} s are unknown, the distribution of the maximum M_n in this case is then derived.

3.2 Central Limit and Large Deviation Theorems for Triangular Arrays

In this section, we study central limit and large deviation theorems for triangular arrays of random variables. Crammer first developed the theory for ordinary arrays of random variables, and S.A. Book (1970) later modified the theorem to accommodate doubly indexed variables. We will show how these theories can be applied to profile analysis. In the profile model, because the scores for each profile column depends on N , which may be large, so to account for this dependency, we approximate the profile scores Y_{nj} in the tails

using large deviation theorem for triangular arrays.

3.2.1 Cramer's Theorem

Theorem 3.1 (Cramer, S.A.Book) *If $\{Z_{ni} : 1 \leq i \leq n, 1 \leq n \leq \infty\}$ is a triangular array of random variables such that*

- (1) Z_{1n}, \dots, Z_{nn} are independent for each n ;
- (2) $E(Z_{ni}) = 0$ for all i and n ;
- (3) $\sum_{i=1}^n E(Z_{ni}^2) = 1$ for all n ;
- (4) $E(|Z_{ni}|^{q_0}) = \beta_{q_0ni} < \infty$ for all i and n , for some $q_0 \geq 3$;
- (5) each Z_{ni} has d.f. $F_{ni}(z) = \alpha_{ni} F_{1ni}(z) + (1 - \alpha_{ni}) F_{2ni}(z)$, where $0 < \alpha_{ni} \leq 1$, $F_{1ni}(z)$ is absolutely continuous, and $F_{2ni}(z)$ has no absolutely continuous component;
- (6) each density $f_{1ni}(z) = F'_{1ni}(z)$ has finite total variation v_{1ni} on $(-\infty, \infty)$; and
- (7) if $\Omega_n = \{i : 1 \leq i \leq n, v_{1ni} \leq (\sqrt{3}/8)T_{q_0n}\}$, then every sequence $\{n_r : 1 \leq r < \infty\}$ of positive integers contains a subsequence $\{n_p : 1 \leq p < \infty\}$ such that either (A) $\lim_{p \rightarrow \infty} \frac{\sum_{i \in \Omega_{n_p}} \alpha_{n_p i}}{\log n_p} = \infty$; or (B) $\lim_{p \rightarrow \infty} \frac{T_{q_0n_p}^2 \sum_{i \in \Omega_{n_p}} \alpha_{n_p i} / v_{1n_p i}^2}{\log n_p} = \infty$

Then, if $F_n(z) = \Pr(S_n \leq z)$ is the distribution function of $S_n = \sum_{i=1}^n Z_{ni}$,

$$F_n(z) = \Phi(z) + \sum_{q=1}^{q_0-3} n^{-q/2} P_{qn}(-\Phi) + R_{q_0n}(z) \quad (3.4)$$

$$= \Phi(z) + \sum_{q=1}^{q_0-3} n^{-q/2} p_{3q-1}(z) e^{-z^2/2} + R_{q_0n}(z) \quad (3.5)$$

where $|R_{q_0n}(z)| < Q/T_{q_0n}^{q_0-2}$ for Q dependent on q_0 but independent of n and z .

Some notations worth mentioning: here, β_{qni} is the q th absolute moment of the random variable Z_{ni} , $B_{qn} = n^{-1} \sum_{i=1}^n \beta_{qni}$ is the average of the q th absolute moment for the entire row indexed by n ; $\rho_{qn} = n^{q/2} B_{qn}$ and $T_{qn} = \sqrt{n}/4\rho_{qn}^{3/q}$. The quantity $P_{qn}(-\Phi)$ denotes a certain linear combination of the first $3q$ derivatives of the normal probability d.f. $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt$; the quantity $p_{3q-1,n}(z) = e^{z^2/2} P_{qn}(-\Phi)$ is a polynomial of degree $3q - 1$ in z .

3.2.2 Application to Profile Analysis

To use Theorem 3.2.1 for profile models, let n be the length of the query sequence \mathbf{L} , we assume the number of columns in a profile $m = m(n)$, and the number of rows in the multiple alignment $N = N(n)$, both depend on n . As was discussed earlier, we have the column scores

$$P_i(L_j) = P_{ij} = \frac{1}{N} \sum_{\alpha \in \Lambda} I(L_j, \alpha) n_{i\alpha} \quad (3.6)$$

and as we move the profile \mathbf{I} along the single sequence \mathbf{L} , for each match, we have the profile scores

$$\begin{aligned} X_{nj} &= \frac{1}{\sqrt{m}} \sum_{i=1}^m P_i(L_{i+j-1}) \\ EX_{nj} &= \sqrt{m} \frac{\sum_{i=1}^m \mu_i}{m} = \sqrt{m} \mu \\ Var X_{nj} &= \frac{1}{m} \sum_{i=1}^m \sigma_i^2 = \sigma^2 < \infty \end{aligned} \quad (3.7)$$

and the standardized profile score

$$\begin{aligned}
Y_{nj}^o &= \frac{X_{nj} - EX_{nj}}{\sqrt{Var X_{nj}}} \\
&= \frac{\frac{1}{\sqrt{m}} \sum_{i=1}^m P_{i,i+j-1} - \frac{1}{\sqrt{m}} \sum_{i=1}^m \mu_i}{\sigma} \\
&= \sum_{i=1}^m \left(\frac{P_{i,i+j-1} - \mu_i}{\sqrt{m}\sigma} \right) \\
&= \sum_{i=1}^m Z_{mi}^j
\end{aligned} \tag{3.8}$$

where $Z_{mi}^j = \frac{P_{i,i+j-1} - \mu_i}{\sqrt{m}\sigma}$. We then have a set of doubly indexed arrays of random variables: $\{Z_{mi}^j : 1 \leq i \leq m, 1 \leq m \leq \infty\}$. We would like to approximate the sum Y_{nj} in the tails by standard normal random variable. Assuming Z_{mi}^j s are independent, then conditions (1) – (3) are satisfied:

$$\begin{aligned}
EZ_{mi}^j &= 0 \\
\sum_{i=1}^m E(Z_{mi}^j)^2 &= \sum_{i=1}^m Var Z_{mi}^j = Var \left(\sum_{i=1}^m Z_{mi}^j \right) = Var Y_{nj} = 1
\end{aligned}$$

condition (4) is also true since $Z_{mi}^j < \infty$.

We now state two definitions that are useful in verifying conditions (5) – (7).

Definition 3.1 A distribution function F is said to be absolutely continuous if there exist a non-negative function $f(x)$ such that

$$F(x) = \int_{-\infty}^x f(u) du$$

Definition 3.2 Let F be a function over (a, b) , let $\Delta : a = a_0 < a_1 < \dots < a_k = b$ be a partition of an interval (a, b) . Define

$$\|F\|_{\Delta} = \sum_{i=1}^k |F(a_i) - F(a_{i-1})|$$

F is said to be of bounded variation over (a, b) if $\sup_{\Delta} \|F\|_{\Delta} < \infty$.

For (5), assume Z_{mi} s have continuous and differentiable distribution function F_{mi} and take $f_{mi}(z) = \frac{d}{dx} F_{mi}(z)$. Z_{mi}^j s are then absolutely continuous and $\alpha_{mi} = 1$; (6) is satisfied since all absolutely continuous distributions are of bounded variation; finally, (7) is true for continuous distribution by taking the subsequence as the sequence itself.

Now, let $F_n^j(y) = \Pr(Y_{nj} \leq y)$ be the distribution function for $Y_{nj} = \sum_{i=1}^m Z_{mi}^j$. Take $q_0 = 4$, by Theorem 3.2.1, we then have

$$\begin{aligned} \rho_{4m} &= m^{4/2} B_{4m} = m^{4/2} E\left(\frac{\sum_{i=1}^m Z_{mi}^j}{m}\right)^4 = m E \sum_{i=1}^m \left(\frac{P_{i,i+j-1} - \mu_i}{\sqrt{m}\sigma}\right)^4 < \infty \\ T_{4m} &= \frac{\sqrt{m}}{4\rho_{4m}^{3/4}} \sim O(\sqrt{m}) \\ |R_{4m}| &< \frac{Q}{T_{4m}^2} \sim O\left(\frac{1}{m}\right) \end{aligned}$$

where Q is dependent on q_0 , but independent of m and y . Also, for $q_0 = 4$

$$\sum_{q=1}^{q_0-3} m^{-q/2} p_{3q-1}(y) e^{-y^2/2} \sim \frac{y^2 e^{-y^2/2}}{\sqrt{m}}$$

We have therefore showed under conditions (1)–(7),

$$\begin{aligned} F_n^j(y) &= \Pr(Y_{nj} \leq y) = \Phi(y) + O\left(\frac{y^2 e^{-y^2/2}}{\sqrt{m}}\right) + O\left(\frac{1}{m}\right) \\ 1 - F_n^j(y) &= \Pr(Y_{nj} > y) = 1 - \Phi(y) + O\left(\frac{y^2 e^{-y^2/2}}{\sqrt{m}}\right) + O\left(\frac{1}{m}\right) \end{aligned}$$

and since $1 - \Phi(y) = \frac{f_0(y)}{y} (1 + O(1/y^2))$,

$$\frac{\Pr(Y_{nj} > y)}{1 - \Phi(y)} = 1 + \frac{O\left(\frac{y^2 e^{-y^2/2}}{\sqrt{m}}\right) + O\left(\frac{1}{m}\right)}{\frac{e^{-y^2/2}}{y} \{1 + O(\frac{1}{y^2})\}} = 1 + O\left(\frac{y^3}{\sqrt{m}}\right) + O\left(\frac{y}{m}\right) = 1 + O\left(\frac{y^3}{\sqrt{m}}\right)$$

We summarize the results in the following Lemma:

Lemma 3.1 *For the profile model, let m be the number of columns in the profile, let Y_{nj} be defined as in (3.8), then for any y ,*

$$\Pr(Y_{nj} > y) - \{1 - \Phi(y)\} = O\left(\frac{y^2 e^{-y^2/2}}{\sqrt{m}}\right) + O\left(\frac{1}{m}\right) \rightarrow 0 \text{ as } n \rightarrow \infty$$

and

$$\frac{\Pr(Y_{nj} > y)}{1 - \Phi(y)} = 1 + O\left(\frac{y^3}{\sqrt{m}}\right)$$

3.3 Extreme Value Theory for Normal Mixture Distribution

In this section, we will first justify using normal mixture distribution to approximate profile scores, then we will develop extreme value theorem specific for the two component normal mixture distribution. More specifically, we will show that the normal mixture distribution in the tails is of exponential type and therefore the maximum random variable, properly normalized, follows the extreme value distribution of the third type $\exp(-e^{-t})$. Next, we will find analytic formula for the normalizing constants for the cases where (1) c is large (2) c is small (local contamination). Here, both constants are functions of the largest characteristic observations from the normal mixture model. Both analytical and numerical ways for approximating the largest observation are developed.

3.3.1 Motivation

The two component normal mixture distribution has the form

$$F_\varepsilon(y) = (1 - \varepsilon)F_o(y) + \varepsilon F_o\left(\frac{y}{c}\right), \quad 0 < \varepsilon < 1, c > 1$$

Here, $F_o(y) = \int_{-\infty}^y \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt$ is c.d.f. of standard normal distribution, $0 < \varepsilon < 1$ indicates the amount of contamination, and $c^2 > 1$ scales the variance. We now justify using the normal mixture model for profile analysis.

Lemma 3.2 *Let \mathbf{l} be a multiple alignment profile such that conditions (1)-(4) in section 3.1 are satisfied, let $P_i(L)$, X_{nj} and Y_{nj}^o be as defined in (3.6), (3.7), and (3.8), if the profile has no gaps with probability $1 - \varepsilon$, and has large amount of gaps with probability ε , then*

$$\begin{aligned} \Pr(Y_{nj}^o > y) &= 1 - F_\varepsilon(y) + O\left(\frac{(\frac{y}{c})^2 e^{-(\frac{y}{c})^2/2}}{\sqrt{m}}\right) + O\left(\frac{1}{m}\right) \\ &\rightarrow 1 - F_\varepsilon(y) \text{ as } n \rightarrow \infty \end{aligned}$$

where $F_\varepsilon(y) = (1 - \varepsilon)F_o(y) + \varepsilon F_o(\frac{y}{c})$ and

$$\frac{\Pr(Y_{nj}^o > y)}{1 - F_\varepsilon(y)} = 1 + O\left(\frac{y^3}{\sqrt{m}}\right)$$

Moreover, under the normal mixture model, let $\mu = \frac{1}{m} \sum_{i=1}^m \mu_i$, and $\sigma^2 = \frac{1}{m} \sum_{i=1}^m \sigma_i^2$, then

$$\begin{aligned} E(X_{nj}) &= \sqrt{m}\mu \\ \text{Var}(X_{nj}) &= \sigma^2[1 + (c^2 - 1)\varepsilon] \end{aligned}$$

Proof. If the profile does not have gaps and looks like Figure 3.1 (case I), as discussed in introduction, we then have the column scores

$$\begin{aligned}
P_i(L) &= \frac{1}{N} \sum_{\alpha} I(L, \alpha) n_{i\alpha} \\
E(P_{ij}) &= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} = \mu_i \\
Var P_{i,j} &= \sum_{\alpha \in \Lambda} \pi_{\alpha} (1 - \pi_{\alpha}) \pi_{i\alpha}^2 - \sum_{\alpha, \beta \in \Lambda} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta} + \frac{1}{N} \sum_{\alpha \in \Lambda} \pi_{i\alpha} (1 - \pi_{i\alpha}) \pi_{\alpha} \\
&= \sigma_i^2 < \infty
\end{aligned}$$

and the profile score

$$\begin{aligned}
X_{nj} &= \frac{1}{\sqrt{m}} \sum_{i=1}^m P_{i(L_i+j-1)} \\
EX_{nj} &= \sqrt{m} \mu \\
Var X_{nj} &= \sigma^2
\end{aligned}$$

Now let $Y_{nj}^o = \frac{X_{nj} - \sqrt{m} \mu}{\sigma}$, it was shown in Lemma 3.1 that $\Pr(Y_{nj}^o > y) = 1 - F_o(y) + O(\frac{y^2 e^{-y^2/2}}{\sqrt{m}}) + O(\frac{1}{m})$. Suppose with small probability ε , the profile has lots of gaps and looks like (case II)

$$\begin{array}{cccccc}
- & l_{12} & l_{13} & l_{14} & - & - \\
l_{21} & l_{22} & l_{23} & l_{24} & l_{25} & l_{26} \\
- & l_{32} & l_{33} & l_{34} & - & l_{36} \\
- & l_{42} & l_{43} & l_{44} & - & l_{46} \\
l_{51} & l_{52} & l_{53} & l_{54} & - & l_{56} \\
l_{61} & l_{62} & l_{63} & l_{64} & l_{65} & l_{66}
\end{array}$$

Figure 3.2

then

$$\begin{aligned}
& \text{Var } X_{nj} \\
&= \frac{1}{m} \sum_{i=1}^m \left\{ \sum_{\alpha} \pi_{\alpha} (1 - \pi_{\alpha}) \pi_{i\alpha}^2 - \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta} + \frac{1}{N_i} \sum_{\alpha} \pi_{i\alpha} (1 - \pi_{i\alpha}) \pi_{\alpha} \right\} \\
&= \sigma'^2 \geq \sigma^2 \quad \text{where } N_i \leq N
\end{aligned}$$

Now let $C = \frac{\sigma'}{\sigma}$, $C \geq 1$, then under case II, $\Pr\left(\frac{X_{nj} - \sqrt{m}\mu}{\sigma'} > y\right) = 1 - F_o(y) + O\left(\frac{y^2 e^{-y^2/2}}{\sqrt{m}}\right) +$

$O\left(\frac{1}{m}\right)$, so

$$\begin{aligned}
\Pr\left(Y_{nj}^o > y \mid \text{case II}\right) &= \Pr\left\{\frac{X_{nj} - \sqrt{m}\mu}{\sigma} > y \mid \text{case II}\right\} \\
&= \Pr\left\{\frac{X_{nj} - \sqrt{m}\mu}{\sigma'/c} > y\right\} \\
&= \Pr\left\{\frac{X_{nj} - \sqrt{m}\mu}{\sigma'} > \frac{y}{c}\right\} \\
&= 1 - F_o\left(\frac{y}{c}\right) + O\left(\frac{\left(\frac{y}{c}\right)^2 e^{-\left(\frac{y}{c}\right)^2/2}}{\sqrt{m}}\right) + O\left(\frac{1}{m}\right)
\end{aligned}$$

the mixture model can then be justified:

$$\begin{aligned}
& \Pr\left(Y_{nj}^o > y\right) \\
&= \Pr(Y_{nj}^o > y \mid \text{case I}) \Pr(\text{case I}) + \Pr\left(Y_{nj}^o > y \mid \text{case II}\right) \Pr(\text{case II}) \\
&= \Pr(Y_{nj}^o > y \mid \text{case I}) (1 - \varepsilon) + \Pr\left(Y_{nj}^o > y \mid \text{case II}\right) \varepsilon \\
&= (1 - \varepsilon) [1 - F_o(y)] + \varepsilon [1 - F_o\left(\frac{y}{c}\right)] \\
&\quad + O\left(\frac{\left(\frac{y}{c}\right)^2 e^{-\left(\frac{y}{c}\right)^2/2}}{\sqrt{m}}\right) + O\left(\frac{1}{m}\right) \\
&= 1 - F_{\varepsilon}(y) + O\left(\frac{\left(\frac{y}{c}\right)^2 e^{-\left(\frac{y}{c}\right)^2/2}}{\sqrt{m}}\right) + O\left(\frac{1}{m}\right)
\end{aligned}$$

Since $Y_{nj}^o = \frac{X_{nj} - \sqrt{m}\mu}{\sigma}$, The mean and variance of X_{nj} can then be worked out under normal mixture model:

$$\begin{aligned}
E(X_{nj}) &= \sqrt{m}\mu \\
Var(X_{nj}) &= E(X_{nj} - \sqrt{m}\mu)^2 \\
&= \sigma^2 Var Y_{nj}^o \\
&= \sigma^2 \left\{ \int_{-\infty}^{\infty} y^2 d\{(1 - \varepsilon)F_o(y) + \varepsilon F_o(\frac{y}{c})\} \right\} \\
&= \sigma^2 \left\{ (1 - \varepsilon) \int_{-\infty}^{\infty} y^2 dF_o(y) + \varepsilon \int_{-\infty}^{\infty} y^2 dF_o(\frac{y}{c}) \right\} \\
&= \sigma^2 [(1 - \varepsilon) + \varepsilon c^2] \\
&= \sigma^2 [1 + (c^2 - 1)\varepsilon]
\end{aligned}$$

■

3.3.2 the Normal Mixture Model is of Exponential type in the Tails

The exponential type distribution is defined as follows:

Definition 3.3 Let X be a random variable with distribution function F with an infinite upper end point and such that $F^{(j)}(x)$, $j = 1, 2, \dots$, exists. We say that F is of the Exponential Type if for large x

$$-\frac{F^{(1)}(x)}{1 - F(x)} \simeq \frac{F^{(2)}(x)}{F^{(1)}(x)} \simeq \frac{F^{(3)}(x)}{F^{(2)}(x)} \simeq \dots \quad (3.9)$$

In the process of verifying normal mixture distribution is of exponential type, we use the Hermite polynomials: the Hermite polynomial of degree m is defined by

$$H_m(x) = m! \sum_{k=0}^{[m/2]} \frac{(-1)^k x^{m-2k}}{k! (m-2k)! 2^k}$$

where $[m/2]$ is the integral part of $m/2$.

The Hermite polynomials has the property that for standard normal distribution $f_0(x) = e^{-x^2/2}/\sqrt{2\pi}$,

$$\frac{d^m}{dx^m} f_0(x) = (-1)^m H_m(x) f_0(x)$$

The first five Hermite polynomials are:

$$H_0(x) = 1; H_1(x) = x; H_2(x) = x^2 - 1;$$

$$H_3(x) = x^3 - 3x; H_4(x) = x^4 - 6x^2 + 3; H_5(x) = x^5 - 10x^3 + 15x$$

Lemma 3.3 *The normal mixture distribution is of exponential type in the tails.*

Proof. We would like to show the normal mixture distribution of the form

$$F_\varepsilon(y) = (1 - \varepsilon)F_o(y) + \varepsilon F_o\left(\frac{y}{c}\right), \quad 0 < \varepsilon < 1, c > 1$$

satisfies equation (3.9). That is, we'd like to show

$$-\frac{F_\varepsilon^{(1)}(x)}{1 - F_\varepsilon(x)} \simeq \frac{F_\varepsilon^{(2)}(x)}{F_\varepsilon^{(1)}(x)} \simeq \frac{F_\varepsilon^{(3)}(x)}{F_\varepsilon^{(2)}(x)} \simeq \dots$$

First, since

$$\frac{d^m}{dx^m} f_0\left(\frac{x}{c}\right) = (-1)^m H_m\left(\frac{x}{c}\right) f_0\left(\frac{x}{c}\right) \frac{1}{c^m}$$

where $H_m(\frac{x}{c}) = m! \sum_{k=0}^{[m/2]} \frac{(-1)^k (x/c)^{m-2k}}{k! (m-2k)! 2^k}$, we have

$$\begin{aligned}
& \frac{(1-\varepsilon)f_o^{(m)}(x)}{\frac{\varepsilon}{c}f_o^{(m)}(x/c)} \\
&= \frac{c(1-\varepsilon)}{\varepsilon} \frac{H_m(x)f_o(x)}{H_m(x/c)f_o(x/c)\frac{1}{c^m}} \\
&= \frac{c(1-\varepsilon)}{\varepsilon} \frac{e^{-x^2/2}}{e^{-x^2/2c^2}} c^{2m} \\
&= \frac{c^{2m+1}(1-\varepsilon)}{\varepsilon} e^{-x^2(1/2-1/2c^2)}
\end{aligned}$$

since c and ε are constants and if $x \sim c\sqrt{2\log n}$ as will be needed in later theorems, we then have

$$e^{-\frac{x^2}{2}(1-\frac{1}{c^2})} \sim O\left(e^{-\frac{1}{2}(1-\frac{1}{c^2})c^2 2\log n}\right) = O\left(\frac{1}{n^{c^2-1}}\right)$$

Therefore,

$$\lim_{x \rightarrow \infty} \frac{(1-\varepsilon)f_o^{(m)}(x)}{\frac{\varepsilon}{c}f_o^{(m)}(x/c)} = 0. \quad (3.10)$$

$f_o^{(m)}(x)$ is negligible compared to $f_o^{(m)}(x/c)$ as $x \rightarrow \infty$. Now, let $\frac{d^n f(x)}{dx^n} = f^{(n)}(x)$, by equation (3.10),

$$F_\varepsilon^{(1)}(x) = f_\varepsilon(x) = (1-\varepsilon)f_o(x) + \frac{\varepsilon}{c}f_o\left(\frac{x}{c}\right) \approx \frac{\varepsilon}{c}f_o\left(\frac{x}{c}\right)$$

Similarly,

$$F_\varepsilon^{(2)}(x) = f_\varepsilon^{(1)}(x) = (1-\varepsilon)f_o^{(1)}(x) + \frac{\varepsilon}{c}f_o^{(1)}\left(\frac{x}{c}\right) \approx \frac{\varepsilon}{c}f_o^{(1)}\left(\frac{x}{c}\right)$$

...

$$F_\varepsilon^{(m+1)}(x) = f_\varepsilon^{(m)}(x) = (1-\varepsilon)f_o^{(m)}(x) + \frac{\varepsilon}{c}f_o^{(m)}\left(\frac{x}{c}\right) \approx \frac{\varepsilon}{c}f_o^{(m)}\left(\frac{x}{c}\right)$$

Therefore, as $x \rightarrow \infty$,

$$\frac{F_\varepsilon^{(m+1)}(x)}{F_\varepsilon^{(m)}(x)} \approx \frac{f_o^{(m)}(x/c)}{f_o^{(m-1)}(x/c)} = \frac{-H_m(x/c)\frac{1}{c^m}}{H_{m-1}(x/c)\frac{1}{c^{m-1}}} \approx \frac{-x^m}{x^{m-1}c^2} \approx \frac{-x}{c^2}$$

Also, since for large x , $F_o(x) \approx 1$, we have

$$F_\varepsilon(x) = (1 - \varepsilon)F_o(x) + \varepsilon F_o(x/c) \approx 1 - \varepsilon[1 - F_o(x/c)]$$

in addition, by Mill's ratio, as $x \rightarrow \infty$, $1 - F_o(x) = \frac{f_o(x)}{x}\{1 + O(x^{-2})\}$, so

$$1 - F_\varepsilon(x) \approx \varepsilon[1 - F_o(x/c)] \approx \frac{\varepsilon f_o(x/c)}{x/c}$$

and $\frac{-f_\varepsilon(x)}{1 - F_\varepsilon(x)} \approx \frac{-\frac{\varepsilon}{c}f_o(\frac{x}{c})}{\frac{\varepsilon f_o(x/c)}{x/c}} = \frac{-x}{c^2}$

Therefore, we have showed that for large x ,

$$-\frac{F_\varepsilon^{(1)}(x)}{1 - F_\varepsilon(x)} \simeq \frac{F_\varepsilon^{(2)}(x)}{F_\varepsilon^{(1)}(x)} \simeq \frac{F_\varepsilon^{(3)}(x)}{F_\varepsilon^{(2)}(x)} \simeq \dots$$

■

3.3.3 The Normalizing Constants from Normal Mixture Distribution $F_\varepsilon(x)$

A classic theorem (Theorem 4.4.5, p181, Sen, 1993) says that the maximum from Exponential type distribution, properly normalized by the largest characteristic and a function of it, has the extreme value distribution of the third type.

Theorem 3.2 *Let $\{X_n\}$ be a random sample corresponding to random variable with distribution function F of the Exponential type. Let η_n denote the largest characteristic observation of F . Let $M_n = \max\{X_1, \dots, X_n\}$. Then there exist sequences of constants $\{a_n\}$ and $\{b_n\}$ such that*

$$\Pr\{a_n(M_n - b_n) \leq x\} = \Pr\{M_n \leq u_n\} \rightarrow \exp(-e^{-x})$$

where

$$u_n = x/a_n + b_n$$

Here, a_n and b_n may be taken as

$$a_n = nf(\eta_n); \quad b_n = \eta_n$$

for standard normal distribution,

$$a_n \sim \sqrt{2 \log n}; \quad b_n \sim \sqrt{2 \log n} \left\{ 1 - \frac{\frac{1}{2}[\log 4\pi + \log \log(n)]}{2 \log(n)} \right\}$$

The largest characteristic observation is defined as follows:

Definition 3.4 For a give random variable X with distribution function F . The largest characteristic observation for a given distribution is defined as the solution η_n to $F(\eta_n) = 1 - 1/n$.

Our objective in this section is to investigate on finding largest characteristic observation from the normal mixture distribution, more specifically, we'd like to find solution η_n such that

$$F_\varepsilon(\eta_n) = (1 - \varepsilon)F_o(\eta_n) + \varepsilon F_o\left(\frac{\eta_n}{c}\right) = 1 - \frac{1}{n}$$

3.3.3.1 Properties of the solution η_n

For normal mixture distribution, if we let

$$h(x) = F_\varepsilon(x) - 1 + 1/n = (1 - \varepsilon)F_o(x) + \varepsilon F_o\left(\frac{x}{c}\right) - 1 + \frac{1}{n}$$

we then seek η_n such that $h(\eta_n) = 0$. The largest characteristic observation has some nice properties:

(1) existence: η_n exists and $0 < \eta_n < c \eta_n^*$

let η_n^* be the largest characteristic observation from the standard normal distribution, that is, $F_o(\eta_n^*) = 1 - 1/n$. Since

$$\begin{aligned}
 h(0) &= (1 - \varepsilon)0.5 + \varepsilon 0.5 - 1 + 1/n = -0.5 + 1/n < 0 \text{ for } n > 2 \\
 h(c\eta_n^*) &= (1 - \varepsilon)F_o(c\eta_n^*) + \varepsilon F_o(\eta_n^*) - 1 + 1/n \\
 &= (1 - \varepsilon)F_o(c\eta_n^*) + \varepsilon(1 - 1/n) - 1 + 1/n \\
 &= (1 - \varepsilon)F_o(c\eta_n^*) - (1 - \varepsilon)(1 - 1/n) \\
 &= (1 - \varepsilon)\{F_o(c\eta_n^*) - (1 - 1/n)\} \\
 &= d > 0 \text{ since } F_o(c\eta_n^*) > (1 - 1/n) \text{ for } c > 1.
 \end{aligned}$$

Therefore, by Intermediate Value Theorem, η_n exists between 0 and $c\eta_n^*$.

(2) Uniqueness: η_n is unique.

Since

$$\begin{aligned}
 \frac{d}{dx}h(x) &= (1 - \varepsilon)f_o(x) + \frac{\varepsilon}{c}f_o\left(\frac{x}{c}\right) \\
 &= \frac{1 - \varepsilon}{\sqrt{2\pi}}e^{-x^2/2} + \frac{\varepsilon}{c\sqrt{2\pi}}e^{-x^2/2c^2} > 0, \text{ for all } x, 0 < \varepsilon < 1, c > 1.
 \end{aligned}$$

$h(x)$ is monotone increasing.

Also, $h(x)$ is bounded between $h(0) = -1/2 + 1/n$ and some positive constant $h(c\eta_n^*) = d$, so there is unique solution.

(3) Exact solution: can always be found using numerical methods such as Newton's method.

Provided an initial value sufficiently close to the true value is chosen, the solution to $h(x) = 0$ can be found iteratively by $x_n = x_{n-1} - \frac{h(x_{n-1})}{\frac{d}{dx}h(x_{n-1})}$.

(4) η_n is an increasing function of c and ε .

Since $\frac{d}{dx}h(x) > 0$, $h(x)$ is monotone increasing $\forall x, \forall n, c > 1, 0 < \varepsilon < 1$,

also, since

$$\begin{aligned}\frac{d}{dx^2}h(x) &= \frac{d}{dx}\left\{\frac{1-\varepsilon}{\sqrt{2\pi}}e^{-x^2/2} + \frac{\varepsilon}{c\sqrt{2\pi}}e^{-x^2/2c^2}\right\} \\ &= -x\left\{\frac{1-\varepsilon}{\sqrt{2\pi}}e^{-x^2/2} + \frac{\varepsilon}{c^3\sqrt{2\pi}}e^{-x^2/2c^2}\right\} < 0\end{aligned}$$

so, $h(x)$ is also a concave function for $\forall x > 0, \forall n, c > 1, 0 < \varepsilon < 1$. Now, to solve

$$h(x) = (1 - \varepsilon)F_o(x) + \varepsilon F_o\left(\frac{x}{c}\right) - 1 + \frac{1}{n} = 0$$

as c increases, $F_o(x/c)$ decreases, and $h(x)$ decreases $\forall x$. So the solution η_n increases.

Similarly, as ε increases, since $h(x) = F_o(x) - \varepsilon\{F_o(x) - F_o(x/c)\}$, $h(x)$ decreases, so again, η_n increases. This relationship can be seen more easily in Fig 3.3 in appendix.

We next discuss the approximation for η_n when c is large and when $c \approx 1$.

3.3.3.2 Case I: when c is large ($c \gg 1$)

First we find analytic formula for η_n when c is large. Table 1 and 2 in appendix show that the formula gives pretty accurate approximations when $n = 200$, $n = 400$ and $c \geq 2$.

Lemma 3.4 For normal mixture distribution

$$F_\varepsilon(y) = (1 - \varepsilon)F_o(y) + \varepsilon F_o\left(\frac{y}{c}\right), \text{ where } 0 < \varepsilon < 1, c \geq 1$$

the normalizing constants in Theorem 3.2 are

$$\begin{aligned} a_n &= n f_\varepsilon(\eta_n) \sim \frac{\sqrt{2 \log n \varepsilon}}{c} \\ b_n &= \eta_n \sim c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2} [\log 4\pi + \log \log(n\varepsilon)]}{2 \log(n\varepsilon)} \right\} \end{aligned}$$

where $f_\varepsilon(x) = (1 - \varepsilon)f_o(x) + \frac{\varepsilon}{c}f_o\left(\frac{x}{c}\right)$ and η_n is the largest characteristic observation from f_ε , that is $F_\varepsilon(\eta_n) = 1 - 1/n$.

Proof. First, we find b_n . Since η_n increases as c increase, for large c , $F_o(\eta_n) \approx 1$, so

$$\begin{aligned} h(\eta_n) &= (1 - \varepsilon)F_o(\eta_n) + \varepsilon F_o\left(\frac{\eta_n}{c}\right) - 1 + \frac{1}{n} \\ &\approx (1 - \varepsilon) + \varepsilon F_o\left(\frac{\eta_n}{c}\right) - 1 + \frac{1}{n} = 0 \end{aligned}$$

This implies that

$$F_o\left(\frac{\eta_n}{c}\right) \approx \frac{\varepsilon - 1/n}{\varepsilon} = 1 - \frac{1}{n\varepsilon}$$

since $F_o^{-1}(1 - \frac{1}{n}) \sim \sqrt{2 \log n} \left\{ 1 - \frac{\frac{1}{2} [\log 4\pi + \log \log(n)]}{2 \log(n)} \right\}$, so

$$b_n = \eta_n \approx c F_o^{-1}\left(1 - \frac{1}{n\varepsilon}\right) \sim c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2} [\log 4\pi + \log \log(n\varepsilon)]}{2 \log(n\varepsilon)} \right\}$$

Next, we find a_n . As was shown in the proof of 3.3, as $n \rightarrow \infty$, for fixed c , and ε ,

$$\begin{aligned} f_\varepsilon(b_n) &= (1 - \varepsilon)f_o(b_n) + \frac{\varepsilon}{c}f_o\left(\frac{b_n}{c}\right) \\ &\approx \frac{\varepsilon}{c}f_o\left(\frac{b_n}{c}\right) \\ &= \frac{\varepsilon}{c} \frac{1}{\sqrt{2\pi}} e^{-\frac{b_n^2}{2c^2}} \end{aligned}$$

$$\begin{aligned}
\frac{b_n^2}{c^2} &= 2 \log n\epsilon \left\{ 1 - \frac{\log 4\pi + \log \log(n\epsilon)}{2 \log(n\epsilon)} + \left(\frac{\log 4\pi + \log \log(n\epsilon)}{4 \log(n\epsilon)} \right)^2 \right\} \\
&\approx 2 \log n\epsilon - \log 4\pi - \log \log(n\epsilon) \\
e^{-\frac{b_n^2}{2c^2}} &\approx e^{-\log n\epsilon} e^{\frac{1}{2} \log 4\pi} e^{\frac{1}{2} \log \log(n\epsilon)} \\
&= \frac{1}{n\epsilon} \sqrt{4\pi} \sqrt{\log(n\epsilon)}
\end{aligned}$$

$$\begin{aligned}
\frac{\epsilon}{c} \frac{1}{\sqrt{2\pi}} e^{-\frac{b_n^2}{2c^2}} &\approx \frac{\epsilon}{c} \frac{1}{\sqrt{2\pi}} \frac{1}{n\epsilon} \sqrt{4\pi} \sqrt{\log(n\epsilon)} \\
&= \frac{\sqrt{2 \log(n\epsilon)}}{cn}
\end{aligned}$$

therefore

$$a_n = n f_\epsilon(b_n) \sim \frac{\sqrt{2 \log(n\epsilon)}}{c}$$

■

An Improved Solution

We next try to improve the approximation for b_n by finding an expression for the residue term which is of the order $o(\frac{1}{\sqrt{\log n\epsilon}})$. The idea is to first set $\eta_n = c\sqrt{2 \log n\epsilon} \{1 - \frac{\frac{1}{2} \log 4\pi + \frac{1}{2} \log \log n\epsilon}{2 \log n\epsilon}\} + h_n$ where $h_n \sim o(\frac{1}{\sqrt{\log n\epsilon}})$, and then solve for an expression for h_n .

For $c > 1$ large,

$$\begin{aligned}
F_\epsilon(x) &= (1 - \epsilon)F_o(x) + \epsilon F_o\left(\frac{x}{c}\right) \\
\text{so } 1 - F_\epsilon(x) &= (1 - \epsilon)\{1 - F_o(x)\} + \epsilon\{1 - F_o\left(\frac{x}{c}\right)\}
\end{aligned}$$

Also,

$$1 - F_o(y) = \frac{1}{y} f_o(y) \{1 - \frac{1}{y^2} + O(y^{-4})\} \text{ for large } y$$

Then

$$\begin{aligned}
1 - F_\varepsilon(x) &= \frac{1-\varepsilon}{y} f_o(y) \{1 - \frac{1}{y^2} + O(y^{-4})\} + \\
&\quad \frac{c\varepsilon}{y} f_o(\frac{y}{c}) \{1 - \frac{c^2}{y^2} + O(y^{-4})\} \\
&= \frac{1-\varepsilon}{y\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \{1 - \frac{1}{y^2} + O(y^{-4})\} \\
&\quad + \frac{c\varepsilon}{y\sqrt{2\pi}} e^{-\frac{1}{2c^2}y^2} \{1 - \frac{c^2}{y^2} + O(y^{-4})\} \\
&= \frac{c\varepsilon}{y\sqrt{2\pi}} e^{-\frac{1}{2c^2}y^2} \{1 - \frac{c^2}{y^2} + O(y^{-4})\} \\
&\quad + \frac{1-\varepsilon}{c\varepsilon} e^{-\frac{y^2}{2c^2}(c^2-1)} [1 - \frac{1}{y^2} + O(y^{-4})]
\end{aligned} \tag{3.11}$$

define η_n such that $1 - F_\varepsilon(\eta_n) = \frac{1}{n}$, then

$$\eta_n = c\sqrt{2\log n\varepsilon} \{1 - \frac{\frac{1}{2}\log 4\pi + \frac{1}{2}\log \log n\varepsilon}{2\log n\varepsilon} + o(\frac{1}{\log n\varepsilon})\} \tag{3.12}$$

$$\frac{\eta_n^2}{2c^2}(c^2 - 1) = [\log n\varepsilon](c^2 - 1) \{1 - O(\frac{\log \log n\varepsilon}{\log n\varepsilon})\} \tag{3.13}$$

Substitute equation (3.12) and (3.13) into (3.11), we get

$$\begin{aligned}
\frac{1}{n} &= \frac{c\varepsilon}{\eta_n\sqrt{2\pi}} e^{-\frac{1}{2c^2}\eta_n^2} \{1 - \frac{c^2}{\eta_n^2} + O(\frac{1}{(\log n\varepsilon)^2})\} \\
&\quad + \frac{1-\varepsilon}{c\varepsilon} (n\varepsilon)^{-(c^2-1)} \{1 - O(\frac{1}{\log n\varepsilon})\}
\end{aligned}$$

or

$$0 = \log \frac{n\varepsilon}{\sqrt{2\pi}} - \log \eta_n - \frac{\eta_n^2}{2c^2} + \log \{1 - \frac{c^2}{\eta_n^2} + O(\frac{1}{(\log n\varepsilon)^2}) + O(n^{-k})\}, k > 0$$

or

$$\begin{aligned}
\log \frac{n\varepsilon}{\sqrt{2\pi}} &= \log(\frac{\eta_n}{c}) + \frac{\eta_n^2}{2c^2} - \log \{1 - \frac{c^2}{\eta_n^2} + O(\frac{1}{(\log n\varepsilon)^2})\} \\
\log \frac{n\varepsilon}{\sqrt{2\pi}} &= \log(\frac{\eta_n}{c}) + \frac{\eta_n^2}{2c^2} + \frac{c^2}{\eta_n^2} + O(\frac{1}{(\log n\varepsilon)^2})
\end{aligned}$$

Set $v_n = \eta_n/c$, then solve for

$$\log \frac{n\varepsilon}{\sqrt{2\pi}} = \frac{v_n^2}{2} + \log v_n + \frac{1}{v_n^2} + O\left(\frac{1}{(\log n\varepsilon)^2}\right) \quad (3.14)$$

Take trial solution as

$$\begin{aligned} v_n &= \sqrt{2 \log n\varepsilon} - \frac{\frac{1}{2} \log 4\pi + \frac{1}{2} \log \log n\varepsilon}{\sqrt{2 \log n\varepsilon}} + h_n \\ &= \sqrt{2 \log n\varepsilon} \left\{ 1 - \frac{\log \sqrt{4\pi \log(n\varepsilon)}}{2 \log n\varepsilon} \right\} + h_n \end{aligned}$$

where $h_n \sim o(\frac{1}{\sqrt{\log n\varepsilon}})$. Now,

$$\begin{aligned} v_n^2 &= 2 \log n\varepsilon + \frac{[\log \sqrt{4\pi \log n\varepsilon}]^2}{2 \log n\varepsilon} + h_n^2 - 2 \log \sqrt{4\pi \log(n\varepsilon)} \\ &\quad + 2h_n \sqrt{2 \log n\varepsilon} - 2h_n \frac{\log \sqrt{4\pi \log(n\varepsilon)}}{\sqrt{2 \log n\varepsilon}} \\ \log v_n &= \log \left\{ \sqrt{2 \log n\varepsilon} \left[1 - \frac{\log \sqrt{4\pi \log n\varepsilon}}{2 \log n\varepsilon} + o\left(\frac{1}{\log n\varepsilon}\right) \right] \right\} \\ &= \log \sqrt{2 \log n\varepsilon} + \log \left[1 - \frac{\log \sqrt{4\pi \log n\varepsilon}}{2 \log n\varepsilon} + o\left(\frac{1}{\log n\varepsilon}\right) \right] \\ &= \log \sqrt{2 \log n\varepsilon} - \frac{\log \sqrt{4\pi \log n\varepsilon}}{2 \log n\varepsilon} + o\left(\frac{1}{\log n\varepsilon}\right) \end{aligned} \quad (3.15)$$

also,

$$\begin{aligned} \frac{1}{v_n^2} &= \frac{1}{2 \log n\varepsilon \left[1 - \frac{\log \sqrt{4\pi \log(n\varepsilon)}}{2 \log(n\varepsilon)} + o\left(\frac{1}{\log n\varepsilon}\right) \right]^2} \\ &= \frac{1}{2 \log n\varepsilon \left\{ \left[1 - \frac{\log \sqrt{4\pi \log(n\varepsilon)}}{2 \log(n\varepsilon)} \right]^2 + o\left(\frac{1}{\log n\varepsilon}\right)^2 + 2\left(1 - \frac{\log \sqrt{4\pi \log(n\varepsilon)}}{2 \log(n\varepsilon)}\right) o\left(\frac{1}{\log n\varepsilon}\right) \right\}} \\ &= \frac{1}{2 \log n\varepsilon \left[1 - \frac{\log \sqrt{4\pi \log(n\varepsilon)}}{2 \log(n\varepsilon)} \right]^2 + o\left(\frac{1}{\log n\varepsilon}\right) + o(1)} \end{aligned} \quad (3.16)$$

For large n , since $h_n^2 \sim o(\frac{1}{\log n\varepsilon})$ is small, so $\log \frac{n\varepsilon}{\sqrt{2\pi}} = \frac{v_n^2}{2} + \log v_n + \frac{1}{v_n^2} + O(\frac{1}{(\log n\varepsilon)^2})$ becomes

$$\begin{aligned}
0 = & \log n\varepsilon + \frac{[\log \sqrt{4\pi \log n\varepsilon}]^2}{4 \log n\varepsilon} - \log \sqrt{4\pi \log n\varepsilon} + \\
& h_n \left\{ \sqrt{2 \log n\varepsilon} - \frac{\log \sqrt{4\pi \log(n\varepsilon)}}{\sqrt{2 \log n\varepsilon}} \right\} + o\left(\frac{1}{\log n\varepsilon}\right) \\
& + \log \sqrt{2 \log n\varepsilon} - \frac{\log \sqrt{4\pi \log n\varepsilon}}{2 \log n\varepsilon} + o\left(\frac{1}{\log n\varepsilon}\right) \\
& + \frac{1}{2 \log n\varepsilon \left[1 - \frac{\log \sqrt{4\pi \log(n\varepsilon)}}{\sqrt{2 \log n\varepsilon}}\right]^2} + o\left(\frac{1}{\log n\varepsilon}\right)^2 + o(1) \\
& - \log \frac{n\varepsilon}{\sqrt{2\pi}} + O\left(\frac{1}{(\log n\varepsilon)^2}\right)
\end{aligned} \tag{3.17}$$

After some algebra, the expression for h_n is

$$\begin{aligned}
h_n = & \frac{1}{\sqrt{2 \log n\varepsilon} - \frac{\log \sqrt{4\pi \log(n\varepsilon)}}{\sqrt{2 \log n\varepsilon}}} \\
& \times \left\{ \frac{1}{\sqrt{\pi}} - \frac{\log \sqrt{4\pi \log n\varepsilon} [\log \sqrt{4\pi \log n\varepsilon} - 2]}{4 \log n\varepsilon} \right. \\
& \left. - \frac{1}{[\sqrt{2 \log n\varepsilon} - \log \sqrt{4\pi \log n\varepsilon}]^2} \right\} + o\left(\frac{1}{(\log n\varepsilon)^{3/2}}\right)
\end{aligned}$$

Discussion

We thus have introduced various ways of approximating the largest characteristic observation for the normal mixture distribution when c is large. The formula

$c\sqrt{2 \log n\varepsilon} \left\{ 1 - \frac{\frac{1}{2}[\log 4\pi + \log \log(n\varepsilon)]}{2 \log(n\varepsilon)} \right\}$ in Theorem 3.4 shows the asymptotic behavior of the estimate and relationships of the estimate with parameters c and ε . We can improve this estimate by working out the expression for the remainder term h_n . Note that in

equation (3.15) and (3.16), we have reduced the formula for $\log v$ and $1/v^2$, a more accurate formula for h_n can be found by using equation (3.14) directly, without the subsequent reductions. Also, in equation (3.17), we have ignored the term h_n^2 which is of the order $o(\frac{1}{\log n \varepsilon})$, a more accurate estimate can be found using the quadratic approximation instead of linear approximation. On the other hand, methods such as Newton's method gives exact results with any desired error bounded, although numerical approximation often does not allow one to study the asymptotic behavior of the estimate. If the inverse of normal c.d.f function can be evaluated with satisfactory accuracy using built in function in software (e.g. PROBNORM in SAS), the formula $\eta_n \approx c F_o^{-1}(1 - \frac{1}{n\varepsilon})$ also gives good approximation with errors less than 0.005 for $\varepsilon \geq 0.1$ and $200 \leq n \leq 600$.

3.3.3.3 Case II: when $c \approx 1$ (local contamination)

Analytic Formula

In this section, we consider the case of local contamination ($c \approx 1$). First, we define η_n^* and η_n such that

$$F_o(\eta_n^*) = 1 - 1/n = F_\varepsilon(\eta_n)$$

so that η_n^* and η_n are the largest characteristic observation from F_o and F_ε respectively.

When $c \approx 1$, $\eta_n = \eta_n^* + r_n \approx \eta_n^*$, where r_n is small. The idea is to solve for an expression for r_n , since η_n^* is known, we then have an expression for η_n .

Since $F_\varepsilon(\eta_n) = (1 - \varepsilon)F_o(\eta_n) + \varepsilon F_o(\frac{\eta_n}{c})$, using Taylor expansion evaluated at η_n^* , we

get

$$\begin{aligned}
F_o(\eta_n) &= F_o(\eta_n^*) + f_o(\eta_n^*)(\eta_n - \eta_n^*) + \frac{f_o'(\eta_n^*)}{2}(\eta_n - \eta_n^*)^2 \\
&\quad + O(\eta_n - \eta_n^*)^2 \\
&= (1 - \frac{1}{n}) + f_o(\eta_n^*)r_n - \frac{\eta_n^* f_o(\eta_n^*)}{2}r_n^2 + O(r_n^2) \\
\text{and } F_o(\frac{\eta_n}{c}) &= F_o(\frac{\eta_n^*}{c}) + f_o(\frac{\eta_n^*}{c})(\frac{\eta_n}{c} - \frac{\eta_n^*}{c}) + \frac{\eta_n^* f_o'(\frac{\eta_n^*}{c})}{2}(\frac{\eta_n}{c} - \frac{\eta_n^*}{c})^2 \\
&\quad + O(\frac{\eta_n}{c} - \frac{\eta_n^*}{c})^2 \\
&= F_o(\frac{\eta_n^*}{c}) + f_o(\frac{\eta_n^*}{c})\frac{r_n}{c} - \frac{\eta_n^*}{2c^3}f_o(\frac{\eta_n^*}{c})r_n^2 + O(r_n^2)
\end{aligned}$$

So,

$$\begin{aligned}
0 &= (1 - \varepsilon)\{(1 - \frac{1}{n}) + f_o(\eta_n^*)r_n - \frac{\eta_n^* f_o(\eta_n^*)}{2}r_n^2 + O(r_n^2)\} \\
&\quad + \varepsilon\{F_o(\frac{\eta_n^*}{c}) + f_o(\frac{\eta_n^*}{c})\frac{r_n}{c} - \frac{\eta_n^*}{2c^3}f_o(\frac{\eta_n^*}{c})r_n^2 + O(r_n^2)\} - 1 + 1/n
\end{aligned}$$

Or

$$\begin{aligned}
0 &= r_n^2\{(1 - \varepsilon)(\frac{-\eta_n^* f_o(\eta_n^*)}{2}) - \frac{\varepsilon \eta_n^* f_o(\frac{\eta_n^*}{c})}{2c^3}\} \\
&\quad + r_n\{(1 - \varepsilon)f_o(\eta_n^*) + \frac{\varepsilon f_o(\frac{\eta_n^*}{c})}{c}\} \\
&\quad + (1 - \varepsilon)(1 - \frac{1}{n}) + \varepsilon F_o(\frac{\eta_n^*}{c}) - 1 + \frac{1}{n}
\end{aligned}$$

Now, let $u = (1 - \varepsilon)f_o(\eta_n^*) + \frac{\varepsilon}{c}f_o(\frac{\eta_n^*}{c})$, when $c^3 \approx 1$ (e.g. $c - 1 < 0.01$), then

$$\begin{aligned}
r_n^2(\frac{-\eta_n^* u}{2}) + r_n u + \varepsilon[F_o(\frac{\eta_n^*}{c}) - 1 + \frac{1}{n}] &= 0 \\
\text{or } r_n^2 - r_n(\frac{2}{\eta_n^*}) - \frac{2\varepsilon[F_o(\frac{\eta_n^*}{c}) - 1 + \frac{1}{n}]}{\eta_n^* u} &= 0
\end{aligned} \tag{3.18}$$

So we get

$$\begin{aligned} r_n &= \frac{1}{2} \left\{ \frac{2}{\eta_n^*} \pm \sqrt{\frac{4}{\eta_n^{*2}} + \frac{4 \times 2\varepsilon[F_o(\frac{\eta_n^*}{c}) - 1 + \frac{1}{n}]}{\eta_n^* u}} \right\} \\ &= \frac{1}{\eta_n^*} \left\{ 1 \pm \sqrt{1 + \frac{2\eta_n^* \varepsilon[F_o(\frac{\eta_n^*}{c}) - 1 + \frac{1}{n}]}{u}} \right\} \end{aligned}$$

On the other hand, if $c^3 \approx 1$, then replace equation (3.18) by

$$r_n^2 \left\{ \frac{-\eta_n^*}{2} \left[u - \frac{\varepsilon f_o(\frac{\eta_n^*}{c})}{c} \left(1 - \frac{1}{c^2} \right) \right] \right\} + r_n u + \varepsilon \left[F_o(\frac{\eta_n^*}{c}) - 1 + \frac{1}{n} \right] = 0$$

since

$$(1 - \varepsilon) \left\{ \frac{-\eta_n^* f_o(\eta_n^*)}{2} - \frac{\varepsilon \eta_n^* f_o(\frac{\eta_n^*}{c})}{2c^3} \right\} = -\frac{\eta_n^*}{2} \left\{ u - \frac{\varepsilon f_o(\frac{\eta_n^*}{c})}{c} \left(1 - \frac{1}{c^2} \right) \right\}$$

the solution is then

$$r_n = \frac{u \pm \sqrt{u^2 + 2\eta_n^* \left[u - \frac{\varepsilon f_o(\frac{\eta_n^*}{c})}{c} \left(1 - \frac{1}{c^2} \right) \right] \varepsilon \left[F_o(\frac{\eta_n^*}{c}) - 1 + \frac{1}{n} \right]}}{\eta_n^* \left[u - \frac{\varepsilon f_o(\frac{\eta_n^*}{c})}{c} \left(1 - \frac{1}{c^2} \right) \right]}$$

Finally, if r_n is very small (e.g. $r_n^2 < 10^{-3}$), a linear approximation with

$$r_n = \frac{\varepsilon \left[F_o(\frac{\eta_n^*}{c}) - 1 + \frac{1}{n} \right]}{-u}$$

may suffice.

A Numerical Recipe: Linear Interpolation and Iteration Method

Alternatively, one can also employ a linear interpolation and iteration method to approximate η_n :

(1) for local contamination, we get $\eta_n = \eta_n^* + r_n = h(c, \eta_n^*)$.

(2) Expand $h(c, \eta_n^*)$ around $c = 1$ to get

$$\eta_n \approx h(1, \eta_n^*) + (c - 1)h^{(1)}(1, \eta_n^*) + \frac{(c - 1)^2}{2}h^{(2)}(1, \eta_n^*)$$

(3) Since $\frac{\eta_n}{c} < \eta_n^* < \eta_n$, use linear interpolation to get

$$f(\eta_n^*) = \frac{(\eta_n - \eta_n^*)f(\eta_n/c) + (\eta_n^* - \eta_n/c)f(\eta_n)}{\eta_n - \eta_n/c}$$

Repeat steps 2 and 3 until convergence.

This algorithm assumes that for $c \simeq 1$, $f_o(\eta_n/c)$, $f_o(\eta_n^*)$, and $f_o(\eta_n)$ lie on a line when η_n/c , η_n^* , and η_n are close and in the neighborhood of $c \simeq 1$. The algorithm provides a good approximation for η_n .

3.4 Order of $\bar{X}_n - \mu\sqrt{m}$ and $\frac{S_n}{\sigma} - 1$

In this section, we calculate the order of $\bar{X}_n - \mu\sqrt{m}$ and $\frac{S_n}{\sigma} - 1$, which will be needed in the later proof where we derive the asymptotic distribution of the normalized maximum score when means and variances of the scores are unknown. Note that when the means and variances for the profile scores are known, the set of X_{nj} s is a set of m -dependent random variables. When means and variances of the profile scores are not known, however, the set of X_{nj} s are more dependent than m -dependent, since the same profile is involved with all the X_{nj} s. These dependencies need to be taken correctly to calculate the orders for $\frac{S_n}{\sigma}$ and $\bar{X}_n - \mu\sqrt{m}$.

3.4.1 Order of $\bar{X}_n - \mu\sqrt{m}$

First we calculate the order of $Cov(X_{nj}, X_{n,j+\delta})$, where δ is a positive integer, and find $Var(\bar{X}_n)$, then we use Chebychev Inequality to find order for $\bar{X}_n - \mu\sqrt{m}$.

Lemma 3.5 Let X_{n1}, \dots, X_{nn} be defined as in (3.1), let m and N be number of columns and rows of the profile \mathbf{l} respectively. Let $\delta \in \mathbb{N}$. Then

$$\text{Cov}(X_{nj}, X_{n,j+\delta}) \sim \begin{cases} O(1) & \text{if } \delta < m \\ O(\frac{1}{N}) & \text{if } \delta \geq m \end{cases}$$

Proof.

$$\begin{aligned} & \text{Cov}(X_{nj}, X_{n,j+\delta}) \quad \text{for } \delta \geq m \\ &= \text{Cov}(X_{n1}, X_{n,1+\delta}) \\ &= \text{Cov} \left\{ \frac{1}{\sqrt{m}} \sum_{i=1}^m P_{ii}, \frac{1}{\sqrt{m}} \sum_{i=1}^m P_{i,i+\delta} \right\} \\ &= \frac{1}{m} \sum_{i=1}^m \text{Cov}\{P_{ii}, P_{i,i+\delta}\} + \frac{1}{m} \sum_{i \neq j} \text{Cov}\{P_{ii}, P_{j,j+\delta}\} \\ &= \frac{1}{m} \sum_{i=1}^m \text{Cov}\{P_{ii}, P_{i,i+\delta}\} \end{aligned}$$

The second term is 0 since all the columns in \mathbf{l} are independent, and all the letters in \mathbf{L} are independent. As discussed before, let $n_{i\alpha}$ be the number of letter α in the i th column, then

$$\begin{aligned} P_i(L) &= \frac{1}{N} \sum_{\alpha} I(L = \alpha) n_{i\alpha} \\ &= \frac{1}{N} \sum_{\alpha} I(L = \alpha) \sum_{k=1}^N I(l_{ki} = \alpha) \\ &= \frac{1}{N} \sum_{\alpha} \sum_{k=1}^N I(L = \alpha) I(l_{ki} = \alpha). \end{aligned}$$

$$\begin{aligned} & \text{Cov}(P_{ii}, P_{i,i+\delta}) \\ &= E(P_{ii}, P_{i,i+\delta}) - E(P_{ii})E(P_{i,i+\delta}) \end{aligned}$$

$$\begin{aligned}
E(P_{ii}) &= E \left\{ \frac{1}{N} \sum_{\alpha} \sum_{k=1}^N I(L_i = \alpha) I(l_{ki} = \alpha) \right\} \\
&= \frac{1}{N} \sum_{\alpha} \sum_{k=1}^N E[I(L_i = \alpha)] E[I(l_{ki} = \alpha)] \\
&= \frac{1}{N} \sum_{\alpha} \sum_{k=1}^N \pi_{\alpha} \pi_{i\alpha} \\
&= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} = \mu_i < \infty
\end{aligned}$$

Similarly, since L'_i 's are i.i.d., $E(P_{i,i+\delta}) = \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} = \mu_i$

$$\begin{aligned}
&P_{ii} \times P_{i,i+\delta} \\
&= \left(\frac{1}{N} \sum_{\alpha} \sum_{k=1}^N I(L_i = \alpha) I(l_{ki} = \alpha) \right) \times \left(\frac{1}{N} \sum_{\alpha} \sum_{t=1}^N I(L_{i+\delta} = \alpha) I(l_{ti} = \alpha) \right) \\
&= \frac{1}{N^2} \left[\sum_{\alpha} \left(\sum_{k=1}^N I(L_i = \alpha) I(l_{ki} = \alpha) \right) \left(\sum_{t=1}^N I(L_{i+\delta} = \alpha) I(l_{ti} = \alpha) \right) \right. \\
&\quad \left. + \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \left(\sum_{k=1}^N I(L_i = \alpha) I(l_{ki} = \alpha) \right) \left(\sum_{t=1}^N I(L_{i+\delta} = \beta) I(l_{ti} = \beta) \right) \right] \\
&= \frac{1}{N^2} \sum_{\alpha} \left(\sum_{k=1}^N I(L_i = \alpha) I(L_{i+\delta} = \alpha) I^2(l_{ki} = \alpha) \right) \\
&\quad + \frac{1}{N^2} \sum_{\alpha} \left(\sum_{k \neq t} I(L_i = \alpha) I(L_{i+\delta} = \alpha) I(l_{ki} = \alpha) I(l_{ti} = \alpha) \right) \\
&\quad + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \left(\sum_{k=1}^N I(L_i = \alpha) I(L_{i+\delta} = \beta) I(l_{ki} = \alpha) I(l_{ki} = \beta) \right) \\
&\quad + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \left(\sum_{k \neq t} I(L_i = \alpha) I(L_{i+\delta} = \beta) I(l_{ki} = \alpha) I(l_{ti} = \beta) \right)
\end{aligned}$$

Since $L_i, L_{i+\delta}, l_{ki}, l_{ti}$ are independent,

$$\begin{aligned}
& E(P_{ii} \times P_{i,i+\delta}) \\
&= \frac{1}{N^2} \sum_{\alpha} \sum_{k=1}^N E[I(L_i = \alpha)] E[I(L_{i+\delta} = \alpha)] E[I^2(l_{ki} = \alpha)] \\
&\quad + \frac{1}{N^2} \sum_{\alpha} \sum_{k \neq t} E[I(L_i = \alpha)] E[I(L_{i+\delta} = \alpha)] E[I(l_{ki} = \alpha)] E[I(l_{ti} = \alpha)] \\
&\quad + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \sum_{k=1}^N E[I(L_i = \alpha)] E[I(L_{i+\delta} = \beta)] E[I(l_{ki} = \alpha) I(l_{ki} = \beta)] \\
&\quad + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \sum_{k \neq t} E[I(L_i = \alpha)] E[I(L_{i+\delta} = \beta)] E[I(l_{ki} = \alpha)] E[I(l_{ti} = \beta)] \\
&= \frac{1}{N^2} \sum_{\alpha} \sum_{k=1}^N \pi_{\alpha} \pi_{\alpha} \pi_{i\alpha} + \frac{1}{N^2} \sum_{\alpha} \sum_{k \neq t} \pi_{\alpha} \pi_{\alpha} \pi_{i\alpha} \pi_{i\alpha} \\
&\quad + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \sum_{k=1}^N \pi_{\alpha} \pi_{\beta} \times 0 + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \sum_{k \neq t} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta} \\
&= \frac{1}{N} \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha} + \frac{N(N-1)}{N^2} \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha}^2 + \frac{N(N-1)}{N^2} \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta} \\
&= \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha}^2 + \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta} + \frac{1}{N} \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha} (1 - \pi_{i\alpha}) - \frac{1}{N} \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta}
\end{aligned}$$

$$\begin{aligned}
& E(P_{ii}) E(P_{i,i+\delta}) \\
&= \left(\sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} \right) \left(\sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} \right) \\
&= \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha}^2 + \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta}
\end{aligned}$$

$$\begin{aligned}
& Cov(P_{ii}, P_{i,i+\delta}) \\
&= E(P_{ii}, P_{i,i+\delta}) - E(P_{ii}) E(P_{i,i+\delta}) \\
&= \frac{1}{N} \left(\sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha} (1 - \pi_{i\alpha}) - \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta} \right) \\
&\sim O\left(\frac{1}{N}\right)
\end{aligned}$$

$$Cov(X_{nj}, X_{n,j+\delta}) = \frac{1}{m} \sum_{i=1}^m Cov\{P_{ii}, P_{i,i+\delta}\} \sim O\left(\frac{1}{N}\right)$$

$$\begin{aligned}
& Cov(X_{nj}, X_{n,j+\delta}) \quad \text{for } \delta < m \\
&= Cov(X_{n1}, X_{n,1+\delta}) \\
&= Cov\left\{\frac{1}{\sqrt{m}} \sum_{i=1}^m P_{ii}, \frac{1}{\sqrt{m}} \sum_{i=1}^m P_{i,i+\delta}\right\} \\
&= \frac{1}{m} \sum_{i=1}^m Cov\{P_{ii}, P_{i,i+\delta}\} + \frac{1}{m} \sum_{i=1}^{m-\delta} Cov\{P_{i+\delta,i+\delta}, P_{i,i+\delta}\} \\
&\quad + \frac{1}{m} \sum_{i \neq j} Cov\{P_{ii}, P_{j,j+\delta}\} \\
&= O\left(\frac{1}{N}\right) + \frac{1}{m} \sum_{i=1}^{m-\delta} Cov\{P_{i+\delta,i+\delta}, P_{i,i+\delta}\} + 0
\end{aligned}$$

For the second term,

$$\begin{aligned}
& Cov(P_{i+\delta,i+\delta}, P_{i,i+\delta}) \\
&= E(P_{i+\delta,i+\delta}, P_{i,i+\delta}) - E(P_{i+\delta,i+\delta})E(P_{i,i+\delta})
\end{aligned}$$

$$\begin{aligned}
& P_{i+\delta,i+\delta} \times P_{i,i+\delta} \\
&= \left(\frac{1}{N} \sum_{\alpha} \sum_{k=1}^N I(L_{i+\delta} = \alpha) I(l_{k,i+\delta} = \alpha) \right) \times \left(\frac{1}{N} \sum_{\alpha} \sum_{t=1}^N I(L_{i+\delta} = \alpha) I(l_{ti} = \alpha) \right) \\
&= \frac{1}{N^2} \left[\sum_{\alpha} \left(\sum_{k=1}^N I(L_{i+\delta} = \alpha) I(l_{k,i+\delta} = \alpha) \right) \left(\sum_{t=1}^N I(L_{i+\delta} = \alpha) I(l_{ti} = \alpha) \right) \right. \\
&\quad \left. + \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \left(\sum_{k=1}^N I(L_{i+\delta} = \alpha) I(l_{k,i+\delta} = \alpha) \right) \left(\sum_{t=1}^N I(L_{i+\delta} = \beta) I(l_{ti} = \beta) \right) \right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{N^2} \sum_{\alpha} \sum_{k=1}^N I(L_{i+\delta} = \alpha) I(L_{i+\delta} = \alpha) I(l_{k,i+\delta} = \alpha) I(l_{ki} = \alpha) \\
&\quad + \frac{1}{N^2} \sum_{\alpha} \sum_{k \neq i} I(L_{i+\delta} = \alpha) I(L_{i+\delta} = \alpha) I(l_{k,i+\delta} = \alpha) I(l_{ti} = \alpha) \\
&\quad + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \sum_{k=1}^N I(L_{i+\delta} = \alpha) I(L_{i+\delta} = \beta) I(l_{k,i+\delta} = \alpha) I(L_{k,i} = \beta) \\
&\quad + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \sum_{k \neq i} I(L_{i+\delta} = \alpha) I(L_{i+\delta} = \beta) I(l_{k,i+\delta} = \alpha) I(l_{ti} = \beta)
\end{aligned}$$

Since $L_{i+\delta}$, l_{ki} , $l_{k,i+\delta}$, l_{ti} are independent,

$$\begin{aligned}
&E(P_{i+\delta,i+\delta} \times P_{i,i+\delta}) \\
&= \frac{1}{N^2} \sum_{\alpha} \sum_{k=1}^N E[I(L_{i+\delta} = \alpha)] E[I(l_{k,i+\delta} = \alpha)] E[I(l_{ki} = \alpha)] \\
&\quad + \frac{1}{N^2} \sum_{\alpha} \sum_{k \neq i} E[I(L_{i+\delta} = \alpha)] E[I(l_{k,i+\delta} = \alpha)] E[I(l_{ti} = \alpha)] \\
&\quad + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \sum_{k=1}^N E[I(L_{i+\delta} = \alpha) I(L_{i+\delta} = \beta)] E[I(l_{k,i+\delta} = \alpha)] E[I(L_{k,i} = \beta)] \\
&\quad + \frac{1}{N^2} \sum_{\substack{\alpha, \beta \\ \alpha \neq \beta}} \sum_{k \neq i} E[I(L_{i+\delta} = \alpha) I(L_{i+\delta} = \beta)] E[I(l_{k,i+\delta} = \alpha)] E[I(l_{ti} = \beta)] \\
&= \frac{1}{N^2} \sum_{\alpha} \sum_{k=1}^N \pi_{\alpha} \pi_{i+\delta, \alpha} \pi_{i\alpha} + \frac{1}{N^2} \sum_{\alpha} \sum_{k \neq i} \pi_{\alpha} \pi_{i+\delta, \alpha} \pi_{i\alpha} + 0 \\
&= \left\{ \frac{1}{N} + \frac{N(N-1)}{N^2} \right\} \sum_{\alpha} \pi_{\alpha} \pi_{i+\delta, \alpha} \pi_{i\alpha} \\
&= \sum_{\alpha} \pi_{\alpha} \pi_{i+\delta, \alpha} \pi_{i\alpha}
\end{aligned}$$

$$\begin{aligned}
&E(P_{i+\delta,i+\delta}) E(P_{i,i+\delta}) \\
&= \left(\sum_{\alpha} \pi_{\alpha} \pi_{i+\delta, \alpha} \right) \left(\sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} \right) \\
&= \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha} \pi_{i+\delta, \alpha} + \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i+\delta, \beta}
\end{aligned}$$

$$\begin{aligned}
& Cov(P_{i+\delta, i+\delta}, P_{i, i+\delta}) \\
&= E(P_{i+\delta, i+\delta}, P_{i, i+\delta}) - E(P_{i+\delta, i+\delta})E(P_{i, i+\delta}) \\
&= \sum_{\alpha} \pi_{\alpha} \pi_{i+\delta, \alpha} \pi_{i\alpha} - \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha} \pi_{i+\delta, \alpha} - \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i+\delta, \beta} \\
&= \sum_{\alpha} \pi_{\alpha} (1 - \pi_{\alpha}) \pi_{i+\delta, \alpha} \pi_{i\alpha} - \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i+\delta, \beta} \\
&< \infty
\end{aligned}$$

$$\begin{aligned}
& Cov(X_{nj}, X_{n, j+\delta}) \quad \text{so for } \delta < m \\
&= \frac{1}{m} \sum_{i=1}^m Cov\{P_{ii}, P_{i, i+\delta}\} + \frac{1}{m} \sum_{i=1}^{m-\delta} Cov\{P_{i+\delta, i+\delta}, P_{i, i+\delta}\} \\
&\quad + \frac{1}{m} \sum_{i \neq j} Cov\{P_{ii}, P_{j, j+\delta}\} \\
&\sim O\left(\frac{1}{N}\right) + O(1) + 0 \\
&= O(1)
\end{aligned}$$

■

We are now ready to find order of $\bar{X}_n - \mu\sqrt{m}$ using Chebychev Inequality.

Lemma 3.6 (Chebychev Inequality) *Let U be a non-negative r.v. with a finite mean $\mu = EU$. Then for every $t > 0$,*

$$\Pr\{U > t\mu\} \leq \frac{1}{t}$$

Definition 3.5 *If for a sequence $\{X_n\}$ of random variables and another sequence $\{b_n\}$, for every $\varepsilon > 0$, there exists K_{ε} and a positive integer n_{ε} such that*

$$\Pr\left\{\left|\frac{X_n}{b_n}\right| > K_{\varepsilon}\right\} < \varepsilon, \quad \forall n \geq n_{\varepsilon}$$

then we say $X_n = O_p(b_n)$.

Lemma 3.7 Let X_{nj} be defined as in (3.1), let $\bar{X}_n = \frac{\sum_{j=1}^n X_{nj}}{n}$, let $N \sim O(\frac{n}{m})$, then

$$\bar{X}_n - \mu\sqrt{m} \sim O_p\left(\sqrt{\frac{m}{n}}\right)$$

Proof. Let $U_n = (\bar{X}_n - \mu\sqrt{m})^2$, let $\sigma_n^2 = E(\bar{X}_n - \mu\sqrt{m})^2 = \text{Var } \bar{X}_n$, then by Chebychev Inequality,

$$\begin{aligned} \Pr\left\{(\bar{X}_n - \mu\sqrt{m})^2 > K^2 \sigma_n^2\right\} &\leq \frac{1}{K^2} \\ \Pr\left\{\left|\frac{(\bar{X}_n - \mu\sqrt{m})}{\sigma_n}\right| > K\right\} &\leq \frac{1}{K^2} \end{aligned}$$

for a given ε , choose $\frac{1}{K^2} < \varepsilon$ or $K > \sqrt{\frac{1}{\varepsilon}}$, we then have

$$\Pr\left\{\left|\frac{(\bar{X}_n - \mu\sqrt{m})}{\sigma_n}\right| > K\right\} \leq \frac{1}{K^2} < \varepsilon$$

so $\bar{X}_n - \mu\sqrt{m} \sim O_p(\sqrt{\text{Var } \bar{X}_n})$.

To find the order of $\text{Var } \bar{X}_n$, note that as was shown in Lemma 3.2, $\text{Var } X_{nj} < \infty$ and by Lemma 3.5,

$$\text{Cov}(X_{nj}, X_{n,j+\delta}) = \begin{cases} O(1) & \text{if } \delta < m \\ O(\frac{1}{N}) & \text{if } \delta \geq m \end{cases}$$

Now,

$$\begin{aligned} \text{Var} \left\{ \sum_{j=1}^n X_{nj} \right\} &= \sum_{j=1}^n \text{Var } X_{nj} + \sum_{\delta=1}^{m-1} (n-\delta) \text{Cov}(X_{nj}, X_{n,j+\delta} | \delta < m) \\ &\quad + n(n-m) \text{Cov}(X_{nj}, X_{n,j+\delta} | \delta > m) \\ &= O(n) + O\{(n-1) + (n-2) + \dots + (n-m+1)\} + O(n^2 \frac{1}{N}) \\ &= O(n) + O\{(m-1)n - \frac{m(m-1)}{2}\} + O(n^2 \frac{1}{N}) \\ &= O(mn) + O(n^2 \frac{1}{N}) \\ \text{Var } \bar{X}_n &= O(\frac{m}{n}) + O(\frac{1}{N}) \end{aligned}$$

if $N \sim O(\frac{n}{m})$, then

$$\bar{X}_n - \mu\sqrt{m} \sim O_p(\sqrt{\text{Var}\bar{X}_n}) = O_p\left(\sqrt{\frac{m}{n}}\right)$$

■

3.4.2 order of $\frac{s_n}{\sigma} - 1$

First, we calculate $\text{Cov}\{(X_{n1} - \mu\sqrt{m})^2, (X_{n1+\delta} - \mu\sqrt{m})^2\}$

Lemma 3.8 *Let X_{nj} be defined as in (3.1), let $\bar{X}_n = \sum_{j=1}^n X_{nj}/n$, let m and N be number of columns and rows of the profile \mathbf{l} respectively. Let $\delta \in \mathbb{N}$, then for $\delta \geq m$,*

$$\text{Cov}\{(X_{n1} - \mu\sqrt{m})^2, (X_{n1+\delta} - \mu\sqrt{m})^2\} = O\left(\frac{1}{n}\right)$$

Proof. Note that as before,

$$E(P_{ii}) = \mu_i$$

$$E(X_{nj}) = E(X_{n1}) = E\left\{\frac{1}{\sqrt{m}} \sum_{i=1}^m P_{ii}\right\} = \frac{1}{\sqrt{m}} \sum_{i=1}^m \mu_i = \sqrt{m}\mu$$

$$\begin{aligned} & \text{Cov}\{(X_{n1} - \mu\sqrt{m})^2, (X_{n1+\delta} - \mu\sqrt{m})^2\} \quad \delta \geq m \\ &= \text{Cov}\left\{\left(\frac{1}{\sqrt{m}} \sum_{i=1}^m (P_{i,i} - \mu_i)\right)^2, \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m (P_{i,i+\delta} - \mu_i)\right)^2\right\} \\ &= \frac{1}{m^2} \text{Cov}\left\{\begin{aligned} & \left(\sum_{i=1}^m (P_{i,i} - \mu_i)^2 + 2 \sum_{i=1}^m \sum_{j \neq i} (P_{ii} - \mu_i)(P_{jj} - \mu_j)\right), \\ & \left(\sum_{i=1}^m (P_{i,i+\delta} - \mu_i)^2 + 2 \sum_{i=1}^m \sum_{j \neq i} (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j)\right) \end{aligned}\right\} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{m^2} Cov \left(\sum_{i=1}^m (P_{i,i} - \mu_i)^2, \sum_{i=1}^m (P_{i,i+\delta} - \mu_i)^2 \right) \\
&+ \frac{2}{m^2} Cov \left(\sum_{i=1}^m (P_{i,i} - \mu_i)^2, \sum_{i=1}^m \sum_{j \neq i} (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \right) \\
&+ \frac{2}{m^2} Cov \left(\sum_{i=1}^m (P_{i,i+\delta} - \mu_i)^2, \sum_{i=1}^m \sum_{j \neq i} (P_{ii} - \mu_i)(P_{jj} - \mu_j) \right) \\
&+ \frac{4}{m^2} Cov \left(\sum_{i=1}^m \sum_{j \neq i} (P_{ii} - \mu_i)(P_{jj} - \mu_j), \sum_{i=1}^m \sum_{j \neq i} (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \right)
\end{aligned}$$

for the first term, $Cov \left(\sum_{i=1}^m (P_{i,i} - \mu_i)^2, \sum_{i=1}^m (P_{i,i+\delta} - \mu_i)^2 \right)$, for each i and δ , (the detailed calculations are shown in the next Lemma)

	order	number of terms
(1) $Cov \{ (P_{i,i} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)^2 \}$	$O(\frac{1}{N})$	1
$Cov \{ (P_{i,i} - \mu_i)^2, (P_{j,j+\delta} - \mu_j)^2 \}$	$= 0$	$m - 1$
$Cov \{ (P_{i+\delta,i+\delta} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)^2 \}$		not possible for $\delta \geq m$
total for each i	$O(\frac{1}{N})$	m

For the second term, $Cov \left(\sum_{i=1}^m (P_{i,i} - \mu_i)^2, \sum_{i=1}^m \sum_{j \neq i} (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \right)$, for each i and δ , we have three cases:

	order	number of terms
(2) $Cov \{ (P_{i,i} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \}$	$= 0$	$2(m - 1)$
$Cov \{ (P_{i,i} - \mu_i)^2, (P_{j,j+\delta} - \mu_j)(P_{k,k+\delta} - \mu_k) \}$	$= 0$	$(m - 1)(m - 2)$
$Cov \{ (P_{i+\delta,i+\delta} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \}$		not possible for $\delta \geq m$
total for each i	0	$m(m - 1)$

For the third term, $Cov \left(\sum_{i=1}^m (P_{i,i+\delta} - \mu_i)^2, \sum_{i=1}^m \sum_{j \neq i} (P_{ii} - \mu_i)(P_{jj} - \mu_j) \right)$, for each i and δ ,

	order	number of terms
(3) $Cov \{ (P_{i,i+\delta} - \mu_i)^2, (P_{ii} - \mu_i)(P_{jj} - \mu_j) \}$	$= 0$	$2(m-1)$
$Cov \{ (P_{i,i+\delta} - \mu_i)^2, (P_{j,j+\delta} - \mu_j)(P_{k,k+\delta} - \mu_k) \}$	$= 0$	$(m-1)(m-2)$
$Cov \{ (P_{i,i+\delta} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \}$		not possible for $\delta \geq m$
total for each i	0	$m(m-1)$

for the fourth term, for each i and δ ,

$$Cov \left(\sum_{i=1}^m \sum_{j \neq i} (P_{ii} - \mu_i)(P_{jj} - \mu_j), \sum_{i=1}^m \sum_{j \neq i} (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \right)$$

there are five cases:

- (4) A $Cov \{ (P_{i,i} - \mu_i)(P_{jj} - \mu_j), (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \}$
 B $Cov \{ (P_{i+\delta,i+\delta} - \mu_i)(P_{j+\delta,j+\delta} - \mu_j), (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \}$
- (5) C $Cov \{ (P_{ii} - \mu_i)(P_{jj} - \mu_j), (P_{i,i+\delta} - \mu_i)(P_{k,k+\delta} - \mu_k) \}$
 D $Cov \{ (P_{i+\delta,i+\delta} - \mu_i)(P_{jj} - \mu_j), (P_{i,i+\delta} - \mu_i)(P_{k,k+\delta} - \mu_k) \}$
 E $Cov \{ (P_{i,i} - \mu_i)(P_{jj} - \mu_j), (P_{k,k+\delta} - \mu_k)(P_{q,q+\delta} - \mu_q) \}$
- total for each i

	order	number of terms
A	$O(\frac{1}{N^2})$	$2(m-1)$
B		not possible for $\delta \geq m$
C	$= 0$	$4(m-1)(m-2)$
D		not possible for $\delta \geq m$
E	$= 0$	$(m-1)(m-2)(m-3)$
total for each i	$O(\frac{m}{N^2})$	$m(m-1)^2$

therefore,

$$\begin{aligned}
& Cov\{(X_{n1} - \mu\sqrt{m})^2, (X_{n1+\delta} - \mu\sqrt{m})^2\} \quad \delta \geq m \\
&= \frac{1}{m^2} Cov\left(\sum_{i=1}^m (P_{i,i} - \mu_i)^2, \sum_{i=1}^m (P_{i,i+\delta} - \mu_i)^2\right) \\
&+ \frac{2}{m^2} Cov\left(\sum_{i=1}^m (P_{i,i} - \mu_i)^2, \sum_{i=1}^m \sum_{j \neq i} (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j)\right) \\
&+ \frac{2}{m^2} Cov\left(\sum_{i=1}^m (P_{i,i+\delta} - \mu_i)^2, \sum_{i=1}^m \sum_{j \neq i} (P_{ii} - \mu_i)(P_{jj} - \mu_j)\right) \\
&+ \frac{4}{m^2} Cov\left(\sum_{i=1}^m \sum_{j \neq i} (P_{ii} - \mu_i)(P_{jj} - \mu_j), \sum_{i=1}^m \sum_{j \neq i} (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j)\right) \\
&= O(\frac{1}{m^2} \frac{m}{N}) + 0 + 0 + O(\frac{1}{m^2} \frac{m^2}{N^2}) + 0 \\
&= O(\frac{1}{mN}) = O(\frac{1}{n})
\end{aligned}$$

■

Lemma 3.9 Let P_{ij} be defined as in (3.2), let $\mu_i = E(P_{ij})$, then for $\delta \geq m$,

$$Cov\left\{(P_{i,i} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)^2\right\} = O(\frac{1}{N})$$

Proof.

$$\begin{aligned}
& Cov \left\{ (P_{i,i} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)^2 \right\} \\
&= Cov \left\{ (P_{ii}^2 + \mu_i^2 - 2P_{ii}), (P_{i,i+\delta}^2 + \mu_i^2 - 2P_{i,i+\delta}) \right\} \\
&= Cov(P_{ii}^2, P_{i,i+\delta}^2) - 2Cov(P_{ii}^2, P_{i,i+\delta}) - 2Cov(P_{ii}, P_{i,i+\delta}^2) + 4Cov(P_{ii}, P_{i,i+\delta}) \\
(1.1) \quad & Cov(P_{ii}^2, P_{i,i+\delta}^2) = E(P_{ii}^2 P_{i,i+\delta}^2) - E(P_{ii}^2)E(P_{i,i+\delta}^2).
\end{aligned}$$

$$\begin{aligned}
& P_{ii}^2 P_{i,i+\delta}^2 \quad \text{for } \delta \geq m \\
&= \left\{ \frac{1}{N} \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\}^2 \left\{ \frac{1}{N} \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\}^2 \\
&= \frac{1}{N^4} \left\{ \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \left\{ \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \\
&\quad \times \left\{ \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\} \left\{ \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^4} \left\{ \sum_{\alpha} I^2(L_i = \alpha) n_{i\alpha}^2 + \sum_{\alpha, \beta} I(L_i = \alpha) I(L_i = \beta) n_{i\alpha} n_{i\beta} \right\} \\
&\quad \times \left\{ \sum_{\alpha} I^2(L_{i+\delta} = \alpha) n_{i\alpha}^2 + \sum_{\alpha, \beta} I(L_{i+\delta} = \alpha) I(L_{i+\delta} = \beta) n_{i\alpha} n_{i\beta} \right\} \\
&= \frac{1}{N^4} \left\{ \sum_{\alpha} I^2(L_i = \alpha) n_{i\alpha}^2 + 0 \right\} \left\{ \sum_{\alpha} I^2(L_{i+\delta} = \alpha) n_{i\alpha}^2 + 0 \right\} \\
&= \frac{1}{N^4} \left\{ \begin{aligned} & \sum_{\alpha} I^2(L_i = \alpha) I^2(L_{i+\delta} = \alpha) n_{i\alpha}^4 + \sum_{\alpha, \beta} I^2(L_i = \alpha) I^2(L_{i+\delta} = \beta) n_{i\alpha}^2 n_{i\beta}^2 \\ & + I^2(L_{i+\delta} = \alpha) I^2(L_i = \beta) n_{i\alpha}^2 n_{i\beta}^2 \end{aligned} \right\}
\end{aligned}$$

$$\begin{aligned}
& E(P_{i,i}^2 \times P_{i,i+\delta}^2) \\
&= \frac{1}{N^4} \sum_{\alpha} \pi_{\alpha}^2 N^4 \pi_{i\alpha}^4 + \frac{1}{N^4} \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} N^4 \pi_{i\alpha}^2 \pi_{i\beta}^2 + O\left(\frac{1}{N}\right) \\
&= \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha}^4 + 2 \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha}^2 \pi_{i\beta}^2 + O\left(\frac{1}{N}\right)
\end{aligned}$$

$$\begin{aligned}
EP_{i,i+\delta}^2 &= EP_{ii}^2 = \frac{1}{N^2} E \left\{ \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^2} E \left\{ \sum_{\alpha} I(L_i = \alpha)^2 n_{i\alpha}^2 + \sum_{\alpha, \beta} I(L_i = \alpha) I(L_i = \beta) n_{i\alpha} n_{i\beta} \right\} \\
&= \frac{1}{N^2} \left\{ \sum_{\alpha} \Pr(L_i = \alpha) E(n_{i\alpha}^2) \right\} \\
&= \frac{1}{N^2} \left\{ \sum_{\alpha} \pi_{\alpha} [N^2 \pi_{i\alpha}^2 + N \pi_{i\alpha} (1 - \pi_{i\alpha})] \right\} \\
&= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}^2 + \frac{1}{N} \sum_{\alpha} \pi_{i\alpha} (1 - \pi_{i\alpha}) \pi_{\alpha}
\end{aligned}$$

$$\begin{aligned}
Cov(P_{i,i}^2, P_{i,i+\delta}^2) &= E(P_{i,i}^2 P_{i,i+\delta}^2) - E(P_{i,i}^2) E(P_{i,i+\delta}^2) \\
&= \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha}^4 + 2 \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha}^2 \pi_{i\beta}^2 + O\left(\frac{1}{N}\right) \\
&\quad - \left(\sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}^2 + O\left(\frac{1}{N}\right) \right)^2 \\
&= O\left(\frac{1}{N}\right)
\end{aligned}$$

$$(1.2) Cov(P_{ii}^2, P_{i,i+\delta})$$

$$\begin{aligned}
&P_{ii}^2 P_{i,i+\delta} \\
&= \left\{ \frac{1}{N} \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\}^2 \left\{ \frac{1}{N} \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^3} \left\{ \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \left\{ \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \left\{ \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^3} \left\{ \sum_{\alpha} I^2(L_i = \alpha) n_{i\alpha}^2 + \sum_{\alpha, \beta} I(L_i = \alpha) I(L_i = \beta) n_{i\alpha} n_{i\beta} \right\} \left\{ \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\}
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{N^3} \left\{ \sum_{\alpha} I^2(L_i = \alpha) n_{i\alpha}^2 + 0 \right\} \left\{ \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^3} \left\{ \begin{aligned} &\sum_{\alpha} I^2(L_i = \alpha) I(L_{i+\delta} = \alpha) n_{i\alpha}^3 \\ &+ \sum_{\alpha, \beta} I^2(L_i = \alpha) I(L_{i+\delta} = \beta) n_{i\alpha}^2 n_{i\beta} + I(L_{i+\delta} = \alpha) I^2(L_i = \beta) n_{i\alpha} n_{i\beta}^2 \end{aligned} \right\}
\end{aligned}$$

$$\begin{aligned}
&E(P_{i,i}^2 \times P_{i,i+\delta}^2) \\
&= \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha}^3 + \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha}^2 \pi_{i\beta} + \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta}^2 + O\left(\frac{1}{N}\right)
\end{aligned}$$

$$\begin{aligned}
E(P_{ii}^2) &= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}^2 + \frac{1}{N} \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} (1 - \pi_{i\alpha}) \\
E(P_{i,i+\delta}) &= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}
\end{aligned}$$

$$\begin{aligned}
Cov(P_{i,i}^2, P_{i,i+\delta}) &= E(P_{i,i}^2 P_{i,i+\delta}) - E(P_{i,i}^2) E(P_{i,i+\delta}) \\
&= \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha}^3 + \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha}^2 \pi_{i\beta} + \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta}^2 + O\left(\frac{1}{N}\right) \\
&\quad - \left(\sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}^2 + O\left(\frac{1}{N}\right) \right) \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} \\
&= O\left(\frac{1}{N}\right)
\end{aligned}$$

$$(1.3) Cov(P_{ii}, P_{i,i+\delta}^2)$$

$$\begin{aligned}
&P_{i,i+\delta}^2 P_{ii} \\
&= \left\{ \frac{1}{N} \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\}^2 \left\{ \frac{1}{N} \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^3} \left\{ \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\} \left\{ \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\} \left\{ \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^3} \left\{ \sum_{\alpha} I^2(L_{i+\delta} = \alpha) n_{i\alpha}^2 + \sum_{\alpha, \beta} I(L_{i+\delta} = \alpha) I(L_{i+\delta} = \beta) n_{i\alpha} n_{i\beta} \right\}
\end{aligned}$$

$$\begin{aligned}
& \times \left\{ \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^3} \left\{ \sum_{\alpha} I^2(L_{i+\delta} = \alpha) n_{i\alpha}^2 + 0 \right\} \left\{ \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^3} \left\{ \begin{aligned} & \sum_{\alpha} I^2(L_{i+\delta} = \alpha) I(L_i = \alpha) n_{i\alpha}^3 \\ & + \sum_{\alpha, \beta} I^2(L_{i+\delta} = \alpha) I(L_i = \beta) n_{i\alpha}^2 n_{i\beta} + I(L_i = \alpha) I^2(L_{i+\delta} = \beta) n_{i\alpha} n_{i\beta}^2 \end{aligned} \right\}
\end{aligned}$$

$$\begin{aligned}
& E(P_{i,i} \times P_{i,i+\delta}^2) \\
&= \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha}^3 + \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha}^2 \pi_{i\beta} + \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta}^2 + O\left(\frac{1}{N}\right)
\end{aligned}$$

$$\begin{aligned}
E(P_{i,i+\delta}^2) &= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}^2 + O\left(\frac{1}{N}\right) \\
E(P_{i,i}) &= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}
\end{aligned}$$

$$\begin{aligned}
Cov(P_{i,i}, P_{i,i+\delta}^2) &= E(P_{i,i} P_{i,i+\delta}^2) - E(P_{i,i}) E(P_{i,i+\delta}^2) \\
&= \sum_{\alpha} \pi_{\alpha}^2 \pi_{i\alpha}^3 + \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha}^2 \pi_{i\beta} + \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta}^2 + O\left(\frac{1}{N}\right) \\
&\quad - \left(\sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}^2 + O\left(\frac{1}{N}\right) \right) \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} \\
&= O\left(\frac{1}{N}\right)
\end{aligned}$$

$$(1.4) Cov(P_{ii}, P_{i,i+\delta})$$

$$\begin{aligned}
& P_{i,i} \times P_{i,i+\delta} \\
&= \left\{ \frac{1}{N} \sum_{\alpha} I(L_i = \alpha) n_{i\alpha} \right\} \left\{ \frac{1}{N} \sum_{\alpha} I(L_{i+\delta} = \alpha) n_{i\alpha} \right\} \\
&= \frac{1}{N^2} \left\{ \begin{aligned} & \sum_{\alpha} I(L_i = \alpha) I(L_{i+\delta} = \alpha) n_{i\alpha}^2 \\ & + \sum_{\alpha, \beta} I(L_i = \alpha) I(L_{i+\delta} = \beta) n_{i\alpha} n_{i\beta} + I(L_{i+\delta} = \alpha) I(L_i = \beta) n_{i\alpha} n_{i\beta} \end{aligned} \right\}
\end{aligned}$$

$$\begin{aligned}
& E(P_{i,i} \times P_{i,i+\delta}) \\
&= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}^2 + 2 \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta} + O\left(\frac{1}{N}\right) \\
\\
& E(P_{ii}) = E(P_{i,i+\delta}) = \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} \\
Cov(P_{ii}, P_{i,i+\delta}) &= E(P_{i,i} P_{i,i+\delta}) - E(P_{ii}) E(P_{i,i+\delta}) \\
&= \sum_{\alpha} \pi_{\alpha} \pi_{i\alpha}^2 + 2 \sum_{\alpha, \beta} \pi_{\alpha} \pi_{\beta} \pi_{i\alpha} \pi_{i\beta} + O\left(\frac{1}{N}\right) - \left(\sum_{\alpha} \pi_{\alpha} \pi_{i\alpha} \right)^2 \\
&= O\left(\frac{1}{N}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
& Cov \left\{ (P_{i,i} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)^2 \right\} \\
&= Cov \left\{ (P_{ii}^2 + \mu_i^2 - 2P_{ii}), (P_{i,i+\delta}^2 + \mu_i^2 - 2P_{i,i+\delta}) \right\} \\
&= Cov(P_{ii}^2, P_{i,i+\delta}^2) - 2Cov(P_{ii}^2, P_{i,i+\delta}) - 2Cov(P_{ii}, P_{i,i+\delta}^2) + 4Cov(P_{ii}, P_{i,i+\delta}) \\
&= O\left(\frac{1}{N}\right)
\end{aligned}$$

■

Lemma 3.10 Let P_{ij} be defined as in 3.2, let $\mu_i = E(P_{ij})$, then for $\delta \geq m$, $j \neq i$,

$$(2) \quad Cov \left\{ (P_{i,i} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \right\} = 0$$

$$(3) \quad Cov \left\{ (P_{i,i+\delta} - \mu_i)^2, (P_{ii} - \mu_i)(P_{jj} - \mu_j) \right\} = 0$$

$$(5) \quad Cov \{ (P_{ii} - \mu_i)(P_{jj} - \mu_j), (P_{i,i+\delta} - \mu_i)(P_{k,k+\delta} - \mu_k) \} = 0$$

Proof.

$$\begin{aligned}
(2) \quad & Cov \left\{ (P_{i,i} - \mu_i)^2, (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \right\} \quad \delta \geq m \\
&= E \{ (P_{i,i} - \mu_i)^2 (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j) \} \\
&\quad - E(P_{i,i} - \mu_i)^2 E(P_{i,i+\delta} - \mu_i) E(P_{j,j+\delta} - \mu_j) \\
&= E \{ (P_{i,i} - \mu_i)^2 (P_{i,i+\delta} - \mu_i) \} E(P_{j,j+\delta} - \mu_j) \\
&\quad - E(P_{i,i} - \mu_i)^2 E(P_{i,i+\delta} - \mu_i) E(P_{j,j+\delta} - \mu_j) \\
&= E(P_{j,j+\delta} - \mu_j) \{ E[(P_{i,i} - \mu_i)^2 (P_{i,i+\delta} - \mu_i)] \\
&\quad - E(P_{i,i} - \mu_i)^2 E(P_{i,i+\delta} - \mu_i) \} \\
&= 0 \times Cov \{ (P_{i,i} - \mu_i)^2, (P_{i,i+\delta} - \mu_i) \} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
(3) \quad & Cov \left\{ (P_{i,i+\delta} - \mu_i)^2, (P_{ii} - \mu_i)(P_{jj} - \mu_j) \right\} \quad \delta \geq m \\
&= E \left\{ (P_{i,i+\delta} - \mu_i)^2 (P_{ii} - \mu_i)(P_{jj} - \mu_j) \right\} \\
&\quad - E(P_{i,i+\delta} - \mu_i)^2 E[(P_{ii} - \mu_i)(P_{jj} - \mu_j)] \\
&= E \{ (P_{i,i+\delta} - \mu_i)^2 (P_{ii} - \mu_i) \} E(P_{jj} - \mu_j) \\
&\quad - E(P_{i,i+\delta} - \mu_i)^2 E(P_{ii} - \mu_i) E(P_{jj} - \mu_j) \\
&= E(P_{jj} - \mu_j) \{ E[(P_{i,i+\delta} - \mu_i)^2 (P_{ii} - \mu_i)] \\
&\quad - E(P_{i,i+\delta} - \mu_i)^2 E(P_{ii} - \mu_i) \} \\
&= 0 \times Cov \left\{ (P_{i,i+\delta} - \mu_i)^2, (P_{ii} - \mu_i) \right\} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
(5) \quad & Cov\{(P_{ii} - \mu_i)(P_{jj} - \mu_j), (P_{i,i+\delta} - \mu_i)(P_{k,k+\delta} - \mu_k)\} \quad \delta \geq m \\
&= E[(P_{ii} - \mu_i)(P_{jj} - \mu_j)(P_{i,i+\delta} - \mu_i)(P_{k,k+\delta} - \mu_k)] \\
&\quad - E[(P_{ii} - \mu_i)(P_{jj} - \mu_j)] E[(P_{i,i+\delta} - \mu_i)(P_{k,k+\delta} - \mu_k)] \\
&= E[(P_{ii} - \mu_i)(P_{i,i+\delta} - \mu_i)] E[(P_{jj} - \mu_j)(P_{k,k+\delta} - \mu_k)] \\
&\quad - E(P_{ii} - \mu_i) E(P_{jj} - \mu_j) E(P_{i,i+\delta} - \mu_i) E(P_{k,k+\delta} - \mu_k) \\
&= E[(P_{ii} - \mu_i)(P_{i,i+\delta} - \mu_i)] E(P_{jj} - \mu_j) E(P_{k,k+\delta} - \mu_k) \\
&\quad - E(P_{ii} - \mu_i) E(P_{jj} - \mu_j) E(P_{i,i+\delta} - \mu_i) E(P_{k,k+\delta} - \mu_k) \\
&= E(P_{jj} - \mu_j) E(P_{k,k+\delta} - \mu_k) \{ E[(P_{ii} - \mu_i)(P_{i,i+\delta} - \mu_i)] \\
&\quad - E(P_{ii} - \mu_i) E(P_{i,i+\delta} - \mu_i) \} \\
&= 0 \times Cov(P_{ii}, P_{i,i+\delta}) \\
&= 0
\end{aligned}$$

■

Lemma 3.11 *Let P_{ij} be defined as in (3.2), let $\mu_i = E(P_{ij})$, then for $\delta \geq m$,*

$$(4) \quad Cov\{(P_{i,i} - \mu_i)(P_{jj} - \mu_j), (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j)\} = O\left(\frac{1}{N^2}\right)$$

Proof.

$$\begin{aligned}
& Cov\{(P_{i,i} - \mu_i)(P_{jj} - \mu_j), (P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j)\} \quad \delta \geq m \\
&= E\{(P_{i,i} - \mu_i)(P_{jj} - \mu_j)(P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j)\} \\
&\quad - E\{(P_{i,i} - \mu_i)(P_{jj} - \mu_j)\} E\{(P_{i,i+\delta} - \mu_i)(P_{j,j+\delta} - \mu_j)\}
\end{aligned}$$

$$\begin{aligned}
&= E[(P_{i,i} - \mu_i)(P_{i,i+\delta} - \mu_i)] E[(P_{jj} - \mu_j)(P_{j,j+\delta} - \mu_j)] \\
&\quad - E(P_{i,i} - \mu_i)E(P_{i,i+\delta} - \mu_i)E(P_{jj} - \mu_j)E(P_{j,j+\delta} - \mu_j) \\
&= E[(P_{i,i} - \mu_i)(P_{i,i+\delta} - \mu_i)] E[(P_{jj} - \mu_j)(P_{j,j+\delta} - \mu_j)] \\
&\quad - E(P_{i,i} - \mu_i)E(P_{i,i+\delta} - \mu_i)E(P_{jj} - \mu_j)E(P_{j,j+\delta} - \mu_j) \\
&= Cov(P_{i,i} - \mu_i, P_{i,i+\delta} - \mu_i) Cov(P_{jj} - \mu_j, P_{j,j+\delta} - \mu_j) \\
&= Cov(P_{ii}, P_{i,i+\delta})Cov(P_{jj}, P_{j,j+\delta}) \\
&= O\left(\frac{1}{N}\right) O\left(\frac{1}{N}\right) \\
&= O\left(\frac{1}{N^2}\right)
\end{aligned}$$

■

We next show that $Var\{(X_{nj} - \mu\sqrt{m})^2\} < \infty$. The following theorem is from Bengt von Bahr (1964).

Lemma 3.12 *Let X_1, X_2, \dots, X_n be a sequence of independent r.v's with zero mean and finite variance σ_i^2 , and let*

$$Y_n = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$

If $\beta_{ri} = E|X_i|^r < \infty$, $1 \leq i \leq n$, $r > 2$, $B_{kn} = \frac{1}{n} \sum_{i=1}^n \beta_{ki}$, $\rho_{kn} = \frac{B_{kn}}{(B_{2n})^{k/2}}$. We have for every positive $v \leq r$

$$\left| E|Y_n|^v - \int_{-\infty}^{\infty} |x|^v d\Phi(x) - \sum_{j=1}^{[r/2]-1} \frac{\int_{-\infty}^{\infty} |x|^v dP_{2j,n}(-\Phi)(x)}{n^j} \right| \leq CR(v, r)$$

where

$$R(v, r) = \begin{cases} \frac{\rho_{rn}^2}{n^{(r-2)/2}} + \frac{\rho_{rn}^{(v+1)/(r-2)}}{n^{(v+1)/2}} \exp \left\{ c\sqrt{n} \sum_{i=1}^n \left(\frac{\sigma_i}{s_n} \right)^3 \left(\frac{\beta_{ri}}{\rho_{rn}\sigma_i} \right)^{1/(r-2)} \right\} & \text{if } 2 < r < 3 \\ \frac{\rho_{rn}^{3+1/r}}{n^{(r-2)/2}} + \frac{\rho_{rn}^{3(v+1)/r}}{n^{(v+1)/2}} & \text{if } 3 \leq r < 4 \\ \frac{\rho_{rn}^{3+1/r}}{n^{(r-2)/2}} + \frac{\rho_{rn}^{3(v+1)/r}}{n^{(v+1)/2}} + \frac{\rho_{rn}^{3(v+r)/r}}{n^{(v+r)/2}} & \text{if } r \geq 4 \end{cases}$$

and C and c are finite constants only depending on r .

We apply this theorem to the profile scores.

Lemma 3.13 Let X_{ni} be defined as in (3.1), let $\mu = EX_{ni}/m$, let $Var(P_{ii}) = \sigma_i^2$, and $\frac{\sum_{i=1}^m \sigma_i^2}{m} = \sigma^2$, then

$$Var\{(X_{nj} - \mu\sqrt{m})^2\} < \infty$$

Proof. for $i = 1, \dots, m$, $E(P_{ii}) = \mu_i$, and $Var(P_{ii}) = \sigma_i^2$. assuming $E|P_{ii}|^4 < \infty$.
let $\frac{\sum_{i=1}^m \sigma_i^2}{m} = \sigma^2 \rightarrow \sigma\sqrt{m} = \sqrt{\sum_{i=1}^m \sigma_i^2}$

$$\begin{aligned} & \frac{X_{n1} - \sqrt{m}\mu}{\sigma} \\ &= \frac{1}{\sigma} \left\{ \frac{1}{\sqrt{m}} \sum_{i=1}^m P_{ii} - \frac{1}{\sqrt{m}} \sum_{i=1}^m \mu_i \right\} \\ &= \frac{1}{\sigma\sqrt{m}} \sum_{i=1}^m (P_{ii} - \mu_i) \\ &= \frac{\sum_{i=1}^m (P_{ii} - \mu_i)}{\sqrt{\sum_{i=1}^m \sigma_i^2}} \end{aligned}$$

by Lemma 3.12, let $v = 4$, $r = 4$, then

$$\left| E \left| \frac{X_{n1} - \sqrt{m}\mu}{\sigma} \right|^4 - \int_{-\infty}^{\infty} |x|^4 d\Phi(x) - \frac{\int_{-\infty}^{\infty} |x|^4 dP_{2,n}(-\Phi)(x)}{m} \right| \leq CR(4, 4)$$

since

$$\begin{aligned}\rho_{4m} &= \frac{B_{4m}}{(B_{2m})^2} \\ B_{4m} &= \frac{\sum_{i=1}^m \beta_{4i}}{m} = \frac{\sum_{i=1}^m E|P_{ii}|^4}{m} < \infty \\ B_{2m} &< \infty \text{ similarly.}\end{aligned}$$

$$R(4, 4) = \frac{\rho_{4m}^{3+1/4}}{m} + \frac{\rho_{4m}^{3 \times 5/4}}{m^{5/2}} + \frac{\rho_{4m}^{3 \times 2}}{m^{8/2}} \sim O\left(\frac{1}{m}\right)$$

Now, for standard normal distribution, $\int_{-\infty}^{\infty} |x|^v d\Phi(x) = \frac{(2v)!}{2^v v!}$ so $\int_{-\infty}^{\infty} |x|^4 d\Phi(x) < \infty$,

moreover,

$$\frac{\int_{-\infty}^{\infty} |x|^4 dP_{2,n}(-\Phi)(x)}{m} = O\left(\frac{1}{m}\right)$$

therefore,

$$E \left| \frac{X_{n1} - \sqrt{m}\mu}{\sigma} \right|^4 < \infty$$

So

$$Var\{(X_{nj} - \mu\sqrt{m})^2\} = E(X_{nj} - \mu\sqrt{m})^4 - \{E(X_{nj} - \mu\sqrt{m})^2\}^2 < \infty$$

■

Lemma 3.14 Let X_{nj} be defined as in (3.1), let $\bar{X}_n = \frac{\sum_{j=1}^n X_{nj}}{n}$, $s_n^2 = \frac{1}{n[1+(c^2-1)\epsilon]} \sum_{j=1}^n (X_{nj} - \bar{X}_n)^2$. Then

$$\begin{aligned}s_n^2 - \sigma^2 &= O_p\left(\frac{m}{n}\right) \\ \frac{s_n}{\sigma} - 1 &= O_p\left(\sqrt{\frac{m}{n}}\right)\end{aligned}$$

Proof.

$$\begin{aligned}
s_n^2 - \sigma^2 &= \frac{\sum_{j=1}^n (X_{nj} - \bar{X}_n)^2}{n \{1 + (c^2 - 1)\varepsilon\}} - \sigma^2 \\
&= \frac{1}{n \{1 + (c^2 - 1)\varepsilon\}} \left\{ \sum_{j=1}^n (X_{nj} - \mu\sqrt{m})^2 - n(\bar{X}_n - \mu\sqrt{m})^2 \right\} - \sigma^2 \\
&= \frac{\sum_{j=1}^n (X_{nj} - \mu\sqrt{m})^2}{n \{1 + (c^2 - 1)\varepsilon\}} - \sigma^2 - \frac{(\bar{X}_n - \mu\sqrt{m})^2}{\{1 + (c^2 - 1)\varepsilon\}}
\end{aligned}$$

Now,

$$\begin{aligned}
E(\bar{X}_n - \mu\sqrt{m})^2 &= \text{Var}(\bar{X}_n) = O\left(\frac{m}{n}\right) \\
\frac{(\bar{X}_n - \mu\sqrt{m})^2}{\{1 + (c^2 - 1)\varepsilon\}} &= O_p\left(\frac{m}{n}\right)
\end{aligned}$$

Also, from Lemma 3.2,

$$\begin{aligned}
E(X_{nj} - \mu\sqrt{m})^2 &= \text{Var}(X_{nj}) = \sigma^2[1 + (c^2 - 1)\varepsilon] \\
E\left\{ \frac{\sum_{j=1}^n (X_{nj} - \mu\sqrt{m})^2}{n \{1 + (c^2 - 1)\varepsilon\}} \right\} &= \sigma^2
\end{aligned}$$

by Chebychev's inequality,

$$\begin{aligned}
\frac{\sum_{j=1}^n (X_{nj} - \mu\sqrt{m})^2}{n \{1 + (c^2 - 1)\varepsilon\}} - \sigma^2 &\sim O_p \left\{ \sqrt{\text{Var} \left(\frac{\sum_{j=1}^n (X_{nj} - \mu\sqrt{m})^2}{n \{1 + (c^2 - 1)\varepsilon\}} \right)} \right\} \\
&= \text{Var} \left\{ \sum_{j=1}^n (X_{nj} - \mu\sqrt{m})^2 \right\} \\
&= \sum_{j=1}^n \text{Var}(X_{nj} - \mu\sqrt{m})^2 + 2 \sum_{i < j} \text{Cov} \left\{ (X_{ni} - \mu\sqrt{m})^2, (X_{nj} - \mu\sqrt{m})^2 \right\} \\
&= n \text{Var}(X_{n1} - \mu\sqrt{m})^2 + \sum_{\delta=1}^{m-1} (n - \delta) \text{Cov} \left\{ (X_{n1} - \mu\sqrt{m})^2, (X_{n1+\delta} - \mu\sqrt{m})^2 \right\} \\
&\quad + [n(n - m) + \frac{(m - 1)^2}{2}] \text{Cov} \left[(X_{n1} - \mu\sqrt{m})^2, (X_{n,t} - \mu\sqrt{m})^2 \mid t - 1 \geq m \right] \\
&= O(n) + O(nm) + n^2 O\left(\frac{1}{n}\right) = O(nm)
\end{aligned}$$

$$\begin{aligned} Var\left(\frac{\sum_{j=1}^n (X_{nj} - \mu\sqrt{m})^2}{n\{1 + (c^2 - 1)\varepsilon\}}\right) &= \frac{1}{n^2} O(nm) \\ &= O\left(\frac{m}{n}\right) \end{aligned}$$

Therefore

$$s_n^2 - \sigma^2 = O_p\left(\sqrt{\frac{m}{n}}\right) + O_p\left(\frac{m}{n}\right) = O_p\left(\frac{m}{n}\right)$$

By definition, given $\varepsilon > 0$, $\exists K_\varepsilon$ such that

$$\Pr\left\{\left|s_n^2 - \sigma^2\right| > K_\varepsilon \frac{m}{n}\right\} < \varepsilon$$

Now, given $\varepsilon > 0$,

$$\begin{aligned} &\Pr\left\{\left|\frac{s_n}{\sigma} - 1\right| > K_\varepsilon \sqrt{\frac{m}{n}}\right\} \\ &= \Pr\left\{\left(\frac{s_n}{\sigma} - 1\right)^2 > K_\varepsilon^2 \frac{m}{n}\right\} \\ &= \Pr\left\{(s_n - \sigma)^2 > K_\varepsilon^2 \frac{m}{n} \sigma^2\right\} \\ &\leq \Pr\left\{\left|s_n^2 - \sigma^2\right| > K_\varepsilon^2 \frac{m}{n} \sigma^2\right\} \\ &< \varepsilon \end{aligned}$$

Therefore,

$$\frac{s_n}{\sigma} - 1 = O_p\left(\sqrt{\frac{m}{n}}\right)$$

■

3.5 Distribution for the Maximum of Profile Scores

In this section, we derive the distribution for M_n^o and M_n for the cases (1) when μ and σ are known (2) when μ and σ are unknown. We start with the Lemma which gives bound

for the joint probability of Y_{nj}^o and Y_{nk}^o and enable us to calculate bound b_2 in Chen Stein Theorem. Next, we apply Chen Stein Theorem to the profile scores to show the maximum of $Y_{nj}^o = \frac{X_{nj} - \sqrt{m}\mu}{\sigma}$ converges to a modified extreme value distribution of the third type. When the profile is random, we don't actually know μ , and σ , so instead we use the scores $Y_{nj} = \frac{X_{nj} - \bar{X}_n}{s_n}$, where \bar{X}_n and s_n^2 are the estimators for μ and σ^2 . We then show that the maximum of profile scores Y_{nj} also converges to a modified extreme value distribution of the third type.

The next Lemma is a modification of Lemma 2 in Goldstein (1994).

Lemma 3.15 *Let $Y_{n1}^o, Y_{n2}^o, \dots, Y_{nn}^o$ be defined as in (3.8). Let*

$$\rho_\delta = \text{Corr}(Y_{nj}^o, Y_{nk}^o) \text{ for } \delta = |k - j|$$

Let

$$\rho = \sup_{1 \leq n < \infty} \{|\rho_\delta^{(n)}| : 1 \leq n \leq \infty \text{ and } 1 \leq \delta < m\}$$

Assume that $0 \leq \rho < 1$. Then there exists a constant C such that if $v_m = o(m^{1/6})$, then for all $1 \leq |j - k| \leq m, 1 \leq j, k \leq n$,

$$p_{jk} = \Pr(Y_{nj}^o > v_m, Y_{nk}^o > v_m) \leq C \{1 - F_\varepsilon(v_m)\}^{\frac{2}{1+\rho}} v_m^{\frac{1-\rho}{1+\rho}}$$

where

$$F_\varepsilon(x) = (1 - \varepsilon)F_o(x) + \varepsilon F_o\left(\frac{x}{c}\right)$$

is d.f. for two component normal mixture distribution, and F_o is d.f. for standard normal distribution.

Proof. Note that

$$\Pr(Y_{nj}^o > v_m, Y_{nk}^o > v_m) \leq \Pr(Y_{nj}^o + Y_{nk}^o > 2v_m)$$

with $\delta = |k - j|$, since $|\rho_\delta| \leq \rho < 1$,

$$\begin{aligned} & \Pr(Y_{nj}^o + Y_{nk}^o > 2v_m) \\ &= \Pr \left\{ \frac{Y_{nj}^o + Y_{nk}^o}{\sqrt{2(1 + \rho_\delta)}} > \frac{2v_m}{\sqrt{2(1 + \rho_\delta)}} \right\} \\ &\leq \Pr \left\{ \frac{Y_{nj}^o + Y_{nk}^o}{\sqrt{2(1 + \rho_\delta)}} > \sqrt{\frac{2}{(1 + \rho)}} v_m \right\} \end{aligned}$$

Assuming $j < k$, and assume without loss of generality, P_i has mean zero, let

$$R_i^{(n)} = \begin{cases} P_i(L_{i+j-1}) & \text{for } 1 \leq i \leq \delta \\ P_i(L_{i+j-1}) + P_{i-\delta}(L_{i+j-1}) & \text{for } \delta + 1 \leq i \leq m \\ P_{(i-\delta)}(L_{i+j-1}) & \text{for } m + 1 \leq i \leq m + \delta \end{cases}$$

then $Y_{nj}^o + Y_{nk}^o = \sum_{i=1}^{m+\delta} R_i^{(n)}$. By Lemma 3.1, $\Pr(Y_{nj}^o + Y_{nk}^o > y) = \{1 - F_o(y)\} \left\{1 + O\left(\frac{y^3}{\sqrt{m}}\right)\right\}$.

So

$$\begin{aligned} & \Pr \left\{ \frac{Y_{nj}^o + Y_{nk}^o}{\sqrt{2(1 + \rho_\delta)}} > \sqrt{\frac{2}{(1 + \rho)}} v_m \right\} \\ &= \left\{ 1 - F_o \left(\sqrt{\frac{2}{(1 + \rho)}} v_m \right) \right\} \left\{ 1 + O \left(\frac{v_m^3}{\sqrt{m}} \right) \right\} \end{aligned}$$

Now, if there is a few gaps and the profile looks like fig 2, then by Lemma 3.2,

$$\begin{aligned} & \Pr \left\{ \frac{Y_{nj}^o + Y_{nk}^o}{\sqrt{2(1 + \rho_\delta)}} > \sqrt{\frac{2}{(1 + \rho)}} v_m \right\} \\ &= \left\{ 1 - F_\varepsilon \left(\sqrt{\frac{2}{(1 + \rho)}} v_m \right) \right\} \left\{ 1 + O \left(\frac{v_m^3}{\sqrt{m}} \right) \right\} \\ &= \text{constant} \left\{ (1 - \varepsilon) \left[1 - F_o \left(\sqrt{\frac{2}{(1 + \rho)}} v_m \right) \right] + \varepsilon \left[1 - F_o \left(\sqrt{\frac{2}{1 + \rho}} \frac{v_m}{c} \right) \right] \right\} \end{aligned}$$

since $1 - F_o(y) = \frac{f_o(y)}{y}(1 + O(1/y^2))$,

$$\begin{aligned}
& 1 - F_o\left(\sqrt{\frac{2}{(1+\rho)}}v_m\right) \\
& \leq C_1 \frac{f_o\left(\sqrt{\frac{2}{(1+\rho)}}v_m\right)}{v_m} \leq C_2 \frac{[f_o(v_m)]^{\frac{2}{1+\rho}}}{v_m} \\
& \leq C_3 [1 - F_o(v_m)]^{\frac{2}{1+\rho}} v_m^{\frac{1-\rho}{1+\rho}} \\
& 1 - F_o\left(\sqrt{\frac{2}{(1+\rho)}}\frac{v_m}{c}\right) \\
& \leq C_1 \frac{f_o\left(\sqrt{\frac{2}{(1+\rho)}}\frac{v_m}{c}\right)}{v_m} \leq C_2 \frac{[f_o(v_m/c)]^{\frac{2}{1+\rho}}}{v_m} \\
& \leq C_3 [1 - F_o(v_m/c)]^{\frac{2}{1+\rho}} v_m^{\frac{1-\rho}{1+\rho}}
\end{aligned}$$

so

$$\begin{aligned}
& \Pr\left\{\frac{Y_{nj}^o + Y_{nk}^o}{\sqrt{2(1+\rho_\delta)}} > \sqrt{\frac{2}{(1+\rho)}}v_m\right\} \\
& \leq C v_m^{\frac{1-\rho}{1+\rho}} \left\{(1-\varepsilon)[1 - F_o(v_m)]^{\frac{2}{1+\rho}} + \varepsilon[1 - F_o(v_m/c)]^{\frac{2}{1+\rho}}\right\} \\
& \leq C v_m^{\frac{1-\rho}{1+\rho}} \{(1-\varepsilon)[1 - F_o(v_m)] + \varepsilon[1 - F_o(v_m/c)]\}^{\frac{2}{1+\rho}} \\
& = C \{1 - F_\varepsilon(v_m)\}^{\frac{2}{1+\rho}} v_m^{\frac{1-\rho}{1+\rho}}
\end{aligned}$$

■

The next Lemma is useful for estimating the rate of convergence for the maximum of a set of normal mixture random variables.

Lemma 3.16 *For the normal mixture distribution F_ε , let*

$$\mu_n = n[1 - F_\varepsilon(u_n)] = n(1 - \varepsilon)[1 - F_o(u_n)] + n\varepsilon[1 - F_o(\frac{u_n}{c})]$$

where

$$\begin{aligned} u_n &= \frac{y}{a_n} + b_n \\ a_n &\sim \frac{\sqrt{2 \log n \varepsilon}}{c} \\ b_n &\sim c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2} [\log 4\pi + \log \log(n\varepsilon)]}{2 \log(n\varepsilon)} \right\} \end{aligned}$$

then for $c > 1$ and $0 < \varepsilon < 1$,

$$\mu_n = e^{-y} \left\{ 1 + \left(\frac{\sqrt{4\pi \log n \varepsilon}}{n \varepsilon e^y} \right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon} \right\} \left\{ 1 + O\left(\frac{\log \log n}{\log n} \right) \right\}$$

and for large n ,

$$\mu_n \approx e^{-y} \left\{ 1 + \left(\frac{\sqrt{4\pi \log n \varepsilon}}{n \varepsilon e^y} \right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon} \right\}$$

Proof. Since $1 - F_o(u_n) = \frac{f_o(u_n)}{u_n} [1 + O(1/u_n^2)] = \frac{1}{u_n \sqrt{2\pi}} e^{-u_n^2/2} \{1 + O(1/u_n^2)\}$, for $u_n = \frac{y}{a_n} + b_n$ where $a_n = \frac{\sqrt{2 \log n \varepsilon}}{c}$, $b_n = c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2} (\log 4\pi + \log \log n \varepsilon)}{2 \log n \varepsilon} \right\}$, we have

$$\begin{aligned} \frac{u_n^2}{2} &= \frac{(\frac{y}{a_n} + b_n)^2}{2} = \frac{y^2}{2a_n^2} + \frac{b_n^2}{2} + \frac{yb_n}{a_n} \\ \frac{y^2}{2a_n^2} &= \frac{y^2}{2 \frac{2 \log n \varepsilon}{c^2}} = \frac{y^2 c^2}{4 \log n \varepsilon} \\ \frac{b_n^2}{2} &= c^2 \log n \varepsilon \left\{ 1 - \frac{\log 4\pi + \log \log(n\varepsilon)}{4 \log(n\varepsilon)} \right\}^2 \\ &= c^2 \log n \varepsilon \left\{ 1 + \frac{(\log 4\pi + \log \log n \varepsilon)^2}{16 (\log n \varepsilon)^2} - \frac{\log 4\pi + \log \log n \varepsilon}{2 \log n \varepsilon} \right\} \\ &= c^2 \left\{ \log n \varepsilon - \log \sqrt{4\pi} - \log \sqrt{\log n \varepsilon} \right\} + \frac{c^2 (\log 4\pi + \log \log n \varepsilon)^2}{16 \log n \varepsilon} \\ \frac{yb_n}{a_n} &= \frac{yc \sqrt{2 \log n \varepsilon}}{\frac{\sqrt{2 \log n \varepsilon}}{c}} \left\{ 1 - \frac{\frac{1}{2} (\log 4\pi + \log \log n \varepsilon)}{2 \log n \varepsilon} \right\} \\ &= yc^2 - \frac{yc^2 \log 4\pi + yc^2 \log \log n \varepsilon}{4 \log n \varepsilon} \end{aligned}$$

Therefore,

$$\begin{aligned}
& e^{-u_n^2/2} \\
&= \exp \left\{ -\frac{y^2 c^2}{4 \log n \varepsilon} \right\} \times \exp \left\{ -c^2 \log n \varepsilon \right\} \exp \left\{ c^2 \log \sqrt{4\pi} \right\} \exp \left\{ c^2 \log \sqrt{\log n \varepsilon} \right\} \\
&\quad \times \exp \left\{ -\frac{c^2 (\log 4\pi + \log \log n \varepsilon)^2}{16 \log n \varepsilon} \right\} \times \exp \left\{ -yc^2 + \frac{yc^2 \log 4\pi + yc^2 \log \log n \varepsilon}{4 \log n \varepsilon} \right\} \\
&= \frac{1}{(n\varepsilon)^{c^2}} (\sqrt{4\pi} \sqrt{\log n \varepsilon})^{c^2} e^{-yc^2} \\
&\quad \times \exp \left\{ c^2 \frac{-4y^2 + 4y \log 4\pi + 4y \log \log n \varepsilon - (\log 4\pi + \log \log n \varepsilon)^2}{16 \log n \varepsilon} \right\}
\end{aligned}$$

So, for $u_n = \frac{y}{a_n} + b_n$ where $a_n = \frac{\sqrt{2 \log n \varepsilon}}{c}$, $b_n = c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2}(\log 4\pi + \log \log n \varepsilon)}{2 \log n \varepsilon} \right\}$,

$$\begin{aligned}
& n(1 - \varepsilon) \{1 - F_o(u_n)\} \\
&= n(1 - \varepsilon) \frac{1}{u_n \sqrt{2\pi}} e^{-u_n^2/2} \{1 + O(1/u_n^2)\} \\
&= n(1 - \varepsilon) \frac{1}{\left\{ \frac{y}{\frac{\sqrt{2 \log n \varepsilon}}{c}} + c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2}(\log 4\pi + \log \log n \varepsilon)}{2 \log n \varepsilon} \right\} \right\} \sqrt{2\pi}} \\
&\quad \times \frac{1}{(n\varepsilon)^{c^2}} (\sqrt{4\pi} \sqrt{\log n \varepsilon})^{c^2} e^{-yc^2} \\
&\quad \times \exp \left\{ c^2 \frac{-4y^2 + 4y \log 4\pi + 4y \log \log n \varepsilon - (\log 4\pi + \log \log n \varepsilon)^2}{16 \log n \varepsilon} \right\} \\
&\quad \times \{1 + O(\frac{1}{\log n \varepsilon})\} \\
&= \frac{n(1 - \varepsilon)}{(n\varepsilon)^{c^2}} \frac{(\sqrt{4\pi} \sqrt{\log n \varepsilon})^{c^2}}{c \sqrt{2\pi} \sqrt{2 \log n \varepsilon} \left\{ \frac{y}{2 \log n \varepsilon} + 1 - \frac{\frac{1}{2}(\log 4\pi + \log \log n \varepsilon)}{2 \log n \varepsilon} \right\}} \\
&\quad \times e^{-yc^2} \left\{ 1 + O\left(\frac{\log \log n \varepsilon}{\log n \varepsilon}\right) \right\} \\
&= \left(\frac{\sqrt{4\pi} \log n \varepsilon}{n} \right)^{c^2-1} \frac{1 - \varepsilon}{c \varepsilon^{c^2}} \times e^{-yc^2} \left\{ 1 + O\left(\frac{\log \log n \varepsilon}{\log n \varepsilon}\right) \right\} \\
&= \left(\frac{\sqrt{4\pi} \log n \varepsilon}{n \varepsilon} \right)^{c^2-1} \frac{1 - \varepsilon}{c \varepsilon} \times e^{-yc^2} \left\{ 1 + O\left(\frac{\log \log n \varepsilon}{\log n \varepsilon}\right) \right\}
\end{aligned}$$

Now, since $1 - F_o(\frac{u_n}{c}) = \frac{f_o(u_n/c)}{u_n/c} \{1 + O(c^2/u_n^2)\} = \frac{c}{u_n\sqrt{2\pi}} e^{-u_n^2/2c^2} \{1 + O(c^2/u_n^2)\}$, and

$$\frac{u_n^2}{2c^2} = \frac{1}{2c^2} \left(\frac{y}{a_n} + b_n \right)^2 = \frac{y^2}{2c^2 a_n^2} + \frac{b_n^2}{2c^2} + \frac{y b_n}{c^2 a_n}$$

and

$$\frac{y^2}{2c^2 a_n^2} = \frac{y^2}{2c^2 \frac{2 \log n \varepsilon}{c^2}} = \frac{y^2}{4 \log n \varepsilon}$$

$$\frac{b_n^2}{2c^2} = \log n \varepsilon - \frac{1}{2} \log 4\pi - \frac{1}{2} \log \log n \varepsilon + O\left(\frac{\log \log n}{\log n}\right)$$

the last term

$$\frac{y b_n}{c^2 a_n} = \frac{y c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2} [\log 4\pi + \log \log(n \varepsilon)]}{2 \log(n \varepsilon)} \right\}}{c^2 \frac{\sqrt{2 \log n \varepsilon}}{c}} = y \left\{ 1 - \frac{\frac{1}{2} [\log 4\pi + \log \log(n \varepsilon)]}{2 \log(n \varepsilon)} \right\}$$

so

$$\begin{aligned} & n \varepsilon [1 - F_o(\frac{u_n}{c})] \\ &= n \varepsilon \frac{c}{u_n \sqrt{2\pi}} e^{-u_n^2/2c^2} \{1 + O(1/u_n^2)\} \\ &= n \varepsilon \frac{c}{\sqrt{2\pi} \left\{ \frac{y}{\frac{\sqrt{2 \log n \varepsilon}}{c}} + c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2} (\log 4\pi + \log \log n \varepsilon)}{2 \log n \varepsilon} \right\} \right\}} \\ & \quad \times \exp \left\{ \frac{-y^2}{4 \log n \varepsilon} - \log n \varepsilon + \log \sqrt{4\pi} + \log \sqrt{\log n \varepsilon} - O\left(\frac{\log \log n}{\log n}\right) - y \right\} \\ & \quad \times \exp \left\{ \frac{\frac{1}{2} [\log 4\pi + \log \log(n \varepsilon)]}{2 \log(n \varepsilon)} \right\} \\ &= e^{-y} \frac{n \varepsilon}{\sqrt{2 \log n \varepsilon} \sqrt{2\pi} \left\{ \frac{y}{2 \log n \varepsilon} + 1 - \frac{\frac{1}{2} (\log 4\pi + \log \log n \varepsilon)}{2 \log n \varepsilon} \right\}} \\ & \quad \times \exp \left\{ -\log n \varepsilon + \log \sqrt{4\pi} + \log \sqrt{\log n \varepsilon} \right\} \\ & \quad \times \exp \left\{ O\left(\frac{\log \log n}{\log n}\right) \right\} \\ &= e^{-y} \left\{ 1 + O\left(\frac{\log \log n}{\log n}\right) \right\} \end{aligned}$$

Therefore,

$$\begin{aligned}
\mu_n &= n(1 - \varepsilon)[1 - F_o(u_n)] + n\varepsilon[1 - F_o(\frac{u_n}{c})] \\
&= \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon} \right)^{c^2-1} \frac{1 - \varepsilon}{c\varepsilon} \times e^{-yc^2} \left\{ 1 + O\left(\frac{\log \log n\varepsilon}{\log n\varepsilon} \right) \right\} \\
&\quad + e^{-y} \left\{ 1 + O\left(\frac{\log \log n}{\log n} \right) \right\} \\
&= e^{-y} \left\{ 1 + e^{-y(c^2-1)} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon} \right)^{c^2-1} \frac{1 - \varepsilon}{c\varepsilon} \right\} \left\{ 1 + O\left(\frac{\log \log n}{\log n} \right) \right\} \\
&= e^{-y} \left\{ 1 + \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon e^y} \right)^{c^2-1} \frac{1 - \varepsilon}{c\varepsilon} \right\} \left\{ 1 + O\left(\frac{\log \log n}{\log n} \right) \right\}
\end{aligned}$$

■

3.5.1 When μ and σ are known

We next state the Chen Stein Theorem again and apply it to the profile scores $Y_{nj}^o = \frac{X_{nj} - \sqrt{m}\mu}{\sigma}$ to derive the asymptotic distribution of maximum M_n^o .

Lemma 3.17 (Chen, Stein) *Let $I = \{1, 2, \dots, n\}$ and for each $j \in I$, let B_j be a Bernoulli random variable with $p_j \equiv \Pr(B_j = 1) = 1 - \Pr(B_j = 0) \in (0, 1)$. Let*

$$W_n \equiv \sum_{j \in I} B_j, \text{ and } \lambda_n \equiv EW_n \equiv \sum_{j \in I} p_j$$

For each $j \in I$, suppose there is a set of dependence for B_j , $N_j \subset I$, with $j \in N_j$, such that

$$B_j \text{ is independent of } \{B_k : k \notin N_j\}$$

Define

$$b_1 \equiv \sum_{j \in I} \sum_{k \in N_j} p_j p_k \text{ and}$$

$$b_2 \equiv \sum_{j \in I} \sum_{j \neq k \in N_j} p_{jk}, \text{ where } p_{jk} \equiv E(B_j B_k)$$

Then

$$|\Pr(W_n = 0) - e^{-\lambda_n}| \leq b_1 + b_2$$

Theorem 3.3 Let $L = \{L_1, L_2, \dots, L_{n+m-1}\}$ be a sequence of i.i.d. letters. Let the standardized profile scores be

$$Y_{nj}^o = \frac{X_{nj} - \sqrt{m}\mu}{\sigma}$$

where X_{nj}, μ, σ are defined as in section 3.1. Let M_n^o be the maximum profile score,

$$M_n^o = \max_{1 \leq j \leq n} Y_{nj}^o$$

Suppose that the maximum correlation of the profile scores is bounded strictly by 1:

$$\rho_\delta = \text{Corr}(Y_{nj}^o, Y_{nk}^o) \text{ for } \delta = |k - j|$$

$$\rho = \sup_{1 \leq n \leq \infty} \{|\rho_\delta| : 1 \leq n < \infty \text{ and } 1 \leq \delta < m\} < 1$$

and for given y , let $u_n = \frac{y}{a_n} + b_n$, where

$$a_n \sim \frac{\sqrt{2 \log n \varepsilon}}{c}$$

$$b_n \sim c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2} [\log 4\pi + \log \log(n\varepsilon)]}{2 \log(n\varepsilon)} \right\}$$

for $0 < \varepsilon < 1$, and $c > 1$, F_o is the d.f. for standard normal distribution, let

$$\lambda_n = n \Pr(Y_{n1}^o > u_n)$$

$$\mu_n = n \{1 - F_\varepsilon(u_n)\} = n(1 - \varepsilon)[1 - F_o(u_n)] + n\varepsilon[1 - F_o(\frac{u_n}{c})]$$

Suppose that $m \sim O(n^k)$ where $k \in (0, \frac{1-\rho}{1+\rho})$, then

$$|\Pr \{a_n(M_n^o - c_n) \leq y\} - e^{-\lambda_n}| = o(n^{-\gamma}) \text{ for every } \gamma \in (0, \frac{1-\rho}{1+\rho} - k)$$

$$\Pr \{a_n(M_n^o - c_n) \leq y\} \rightarrow \exp \left(-e^{-y} \left\{ 1 + \left(\frac{\sqrt{4\pi \log n \varepsilon}}{n \varepsilon e^y} \right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon} \right\} \right)$$

Proof. For $j \in \{1, 2, \dots, n\}$, let

$$B_j = I(Y_{nj}^o > u_n) \text{ and } W_n = \sum_{j=1}^n B_j,$$

then

$$\lambda_n = E W_n = n \Pr(Y_{n1}^o > u_n)$$

and since by Lemma 3.2, $\frac{\lambda_n}{\mu_n} = 1 + O(\frac{u_n^3}{\sqrt{m}})$, and by Lemma 3.16, $\frac{\mu_n}{e^{-y}} = 1 + O\left(\frac{\log \log n}{\log n}\right)$,

$$\begin{aligned} b_1 &= \sum_{j \in I} \sum_{k \in N_j} p_j p_k = |I| |N_j| p_j^2 \leq 2mn [\Pr(Y_{n1} > u_n)]^2 = 2m \frac{\lambda_n^2}{n} \\ &= \frac{2m}{n} \frac{\lambda_n^2}{\mu_n^2} \frac{\mu_n^2}{e^{-2y}} \\ &= \frac{m}{n} \left\{ 1 + O\left(\frac{(\log n)^{3/2}}{\sqrt{m}}\right) \right\} \left\{ 1 + O\left(\frac{\log \log n}{\log n}\right) \right\} \\ &= \frac{m}{n} = n^{k-1} \end{aligned}$$

by Lemma 3.15, $p_{jk} = \Pr(Y_{nj}^o > u_n, Y_{nk}^o > u_n) \leq C \{1 - F_\varepsilon(u_n)\}^{\frac{2}{1+\rho}} u_n^{\frac{1-\rho}{1+\rho}}$, also, since

$$\mu_n = e^{-y} \left\{ 1 + O\left(\frac{\log \log n}{\log n}\right) \right\} < \infty$$

$$\begin{aligned} b_2 &= \sum_{j \in I} \sum_{j \neq k \in N_j} p_{jk} \leq A |I| |N_j| \{1 - F_\varepsilon(v_m)\}^{\frac{2}{(1+\rho)}} u_n^{\frac{1-\rho}{1+\rho}} \\ &\leq B n m \left(\frac{\mu_n}{n}\right)^{\frac{2}{1+\rho}} u_n^{\frac{1-\rho}{1+\rho}} \\ &= B n^{1+k-\frac{2}{1+\rho}} u_n^{\frac{1-\rho}{1+\rho}} \mu_n^{\frac{2}{1+\rho}} \sim O\left(\frac{1}{n^{\frac{1-\rho}{1+\rho}-k}}\right) \end{aligned}$$

now,

$$\Pr\{W_n = 0\} = \Pr\{M_n^o \leq y\} = \Pr\{a_n(M_n^o - b_n)\}$$

so

$$|\Pr\{a_n(M_n^o - b_n)\} - e^{-\lambda_n}| \leq O\left(\frac{1}{n^{1-k}}\right) + O\left(\frac{1}{n^{\frac{1-\rho}{1+\rho}-k}}\right) = O\left(\frac{1}{n^{\frac{1-\rho}{1+\rho}-k}}\right)$$

Now, let $\mu_n = n[1 - F_\varepsilon(u_n)]$, since, $\frac{\lambda_n}{\mu_n} = 1 + O\left(\frac{1}{n^k}\right)$, and by Lemma 3.16, $\frac{\mu_n}{e^{-y}} = 1 + O\left(\frac{\log \log n}{\log n}\right)$, so

$$\begin{aligned} \frac{\lambda_n}{e^{-y} \left\{ 1 + \left(\frac{\sqrt{4\pi \log n \varepsilon}}{n \varepsilon e^y} \right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon} \right\}} &= \frac{\lambda_n}{\mu_n} \frac{\mu_n}{e^{-y} \left\{ 1 + \left(\frac{\sqrt{4\pi \log n \varepsilon}}{n \varepsilon e^y} \right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon} \right\}} \\ &= \left\{ 1 + O\left(\frac{1}{n^k}\right) \right\} \left\{ 1 + O\left(\frac{\log \log n}{\log n}\right) \right\} \\ &= \left\{ 1 + O\left(\frac{\log \log n}{\log n}\right) \right\} \end{aligned}$$

and

$$\begin{aligned} &\left| \Pr\{a_n(M_n^o - b_n) \leq y\} - \exp\left(-e^{-y} \left\{ 1 + \left(\frac{\sqrt{4\pi \log n \varepsilon}}{n \varepsilon e^y} \right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon} \right\}\right) \right| \\ &\leq \left| \Pr\{a_n(M_n^o - b_n) \leq y\} - e^{-\lambda_n} \right| \\ &\quad + \left| e^{-\lambda_n} - \exp\left(-e^{-y} \left\{ 1 + \left(\frac{\sqrt{4\pi \log n \varepsilon}}{n \varepsilon e^y} \right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon} \right\}\right) \right| \\ &= O\left(\frac{1}{n^{\frac{1-\rho}{1+\rho}-k}}\right) + O\left(\frac{\log \log n}{\log n}\right) \\ &= O\left(\frac{\log \log n}{\log n}\right) \end{aligned}$$

■

Here, as we can see from the proof above, the slow rate of convergence have mainly resulted from the $\frac{\log \log n}{\log n}$ order for convergence of maximum of normal mixture random

variables to extreme value distribution. For random profiles, since μ , and σ are unknown, naturally we would use scores $Y_{nj} = \frac{X_{nj} - \bar{X}_n}{s_n}$, where \bar{X}_n and s_n^2 are the estimators of μ and σ^2 .

3.5.2 When μ and σ are unknown

The next theorem shows that the maximum of scores Y_{nj} s also converge to a modified extreme value distribution too.

Theorem 3.4 *Let $\mathbf{L} = \{L_1, L_2, \dots, L_{n+m-1}\}$ be a sequence of i.i.d. letters. Let $Y_{nj}^o = \frac{X_{nj} - \sqrt{m}\mu}{\sigma}$ where X_{nj} , μ , σ are defined as in 3.1 and let $M_n^o = \max_{1 \leq j \leq n} Y_{nj}^o$. When the profile is random, μ and σ are unknown. Let the profile scores and their maximum be*

$$Y_{nj} = \frac{X_{nj} - \bar{X}_n}{s_n}$$

$$M_n = \max_{1 \leq j \leq n} Y_{nj}$$

where $\bar{X}_n = \frac{\sum_{j=1}^n X_{nj}}{n}$, $s_n^2 = \frac{1}{n[1+\varepsilon(c^2-1)]} \sum_{j=1}^n (X_{nj} - \bar{X}_n)^2$. Suppose that the maximum correlation of the profile scores is bounded strictly by 1:

$$\rho_\delta = \text{Corr}(Y_{nj}^o, Y_{nk}^o) \text{ for } \delta = |k - j|$$

$$\rho = \sup_{1 \leq n \leq \infty} \{|\rho_\delta| : 1 \leq n < \infty \text{ and } 1 \leq \delta < m\} < 1$$

and for given y , let $u_n = \frac{y}{a_n} + b_n$, where

$$a_n \sim \frac{\sqrt{2 \log n \varepsilon}}{c}$$

$$b_n \sim c \sqrt{2 \log n \varepsilon} \left\{ 1 - \frac{\frac{1}{2} [\log 4\pi + \log \log(n\varepsilon)]}{2 \log(n\varepsilon)} \right\}$$

for $0 < \varepsilon < 1$, and $c > 1$, F_o is the d.f. for standard normal distribution, let

$$\lambda_n = n \Pr(Y_{n1}^o > u_n)$$

$$\mu_n = n \{1 - F_\varepsilon(u_n)\} = n(1 - \varepsilon)[1 - F_o(u_n)] + n\varepsilon[1 - F_o(\frac{u_n}{c})]$$

Suppose that $m \sim O(n^k)$ where $k \in (0, \frac{1-\rho}{1+\rho})$, then

$$\begin{aligned} & \left| \Pr\{a_n(M_n - c_n) \leq y\} - \exp\left(-e^{-y} \left\{1 + \frac{1-\varepsilon}{c\varepsilon} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon}\right)^{c^2-1} e^{-y(c^2-1)}\right\}\right) \right| \\ &= O\left(\frac{\log \log n}{\log n}\right) \end{aligned}$$

Proof. Let

$$Y_{nj}^o = \frac{X - \mu\sqrt{m}}{\sigma}$$

$$M_n^o = \max_{1 \leq i \leq n} Y_{nj}^o$$

then

$$\begin{aligned} Y_{nj} &= \frac{X_i - \bar{X}_n}{s_n} \\ &= \frac{\sigma}{s_n} \frac{X - \mu\sqrt{m}}{\sigma} - \frac{\bar{X}_n - \mu\sqrt{m}}{s_n} \\ &= \frac{\sigma}{s_n} Y_{nj}^o - \frac{\bar{X}_n - \mu\sqrt{m}}{s_n} \end{aligned}$$

and

$$M_n = \max_{1 \leq i \leq n} Y_{nj} = \max_{1 \leq i \leq n} \frac{X_i - \bar{X}_n}{s_n} = \frac{\sigma}{s_n} M_n^o - \frac{\bar{X}_n - \mu\sqrt{m}}{s_n}$$

So,

$$\begin{aligned}
& \Pr\{a_n(M_n - b_n) \leq y\} \\
&= \Pr\{a_n(\frac{\sigma}{s_n}M_n^o - \frac{\bar{X}_n - \mu\sqrt{m}}{s_n} - b_n) \leq y\} \\
&= \Pr\{a_n\frac{\sigma}{s_n}M_n^o - a_nb_n \leq y + a_n\frac{\bar{X}_n - \mu\sqrt{m}}{s_n}\} \\
&= \Pr\{a_nM_n^o - a_nb_n(\frac{s_n}{\sigma}) \leq \frac{s_n}{\sigma}y + a_n\frac{\bar{X}_n - \mu\sqrt{m}}{\sigma}\} \\
&= \Pr\{a_nM_n^o - a_nb_n(\frac{s_n}{\sigma}) - a_nb_n + a_nb_n \leq \frac{s_n}{\sigma}y + a_n\frac{\bar{X}_n - \mu\sqrt{m}}{\sigma}\} \\
&= \Pr\{a_n(M_n^o - b_n) \leq y\frac{s_n}{\sigma} - y + y - a_nb_n(1 - \frac{s_n}{\sigma}) + a_n\frac{\bar{X}_n - \mu\sqrt{m}}{\sigma}\} \\
&= \Pr\{a_n(M_n^o - b_n) \leq y(\frac{s_n}{\sigma} - 1) + y + a_nb_n(\frac{s_n}{\sigma} - 1) + a_n\frac{\bar{X}_n - \mu\sqrt{m}}{\sigma}\} \\
&= \Pr\{a_n(M_n^o - b_n) \leq y + yv + a_nb_nv + a_nw\}
\end{aligned}$$

where

$$\begin{aligned}
v &= \frac{s_n}{\sigma} - 1 \\
w &= \frac{\bar{X}_n - \mu\sqrt{m}}{\sigma}
\end{aligned}$$

By Theorem 3.3, for $u_n = \frac{\log \log n}{\log n}$,

$$\Pr\{a_n(M_n^o - b_n) \leq y\} = (1 + u_n) \exp\left(-e^{-y} \left\{1 + \frac{1-\varepsilon}{c\varepsilon} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon}\right)^{c^2-1} e^{-y(c^2-1)}\right\}\right)$$

let $B = B(n, c, \varepsilon) = \frac{1-\varepsilon}{c\varepsilon} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon}\right)^{c^2-1}$, then

$$\begin{aligned}
\Pr\{a_n(M_n^o - b_n) \leq y\} &= (1 + u_n) \exp\left(-e^{-y} \left\{1 + Be^{-y(c^2-1)}\right\}\right) \\
&= (1 + u_n) \exp\left(-e^{-y} - Be^{-yc^2}\right)
\end{aligned}$$

now, let $h = h_n(y, v, w) = y + yv + a_nb_nv + a_nw$,

$$\begin{aligned}
& \Pr\{a_n(M_n - b_n) \leq y\} \\
&= \Pr\{a_n(M_n^o - b_n) \leq y + yv + a_nb_nv + a_nw\} \\
&= \exp\left(-e^{-h} - Be^{-hc^2}\right)(1 + u_n) \\
&= G_n(y, v, w)(1 + u_n)
\end{aligned}$$

First, we derive the expression for $G_n(y, v, w)$.

Let $G = G_n(y, v, w) = \exp\left(-e^{-h} - Be^{-hc^2}\right)$, then

$$\begin{aligned}
\frac{\partial h}{\partial y} &= 1 + v & \frac{\partial h}{\partial v} &= y + a_nb_n & \frac{\partial h}{\partial w} &= a_n \\
\frac{\partial h}{\partial y \partial v} &= 1 & \frac{\partial h}{\partial y \partial w} &= 0 & \frac{\partial h}{\partial v \partial w} &= 0 \\
\frac{\partial h}{\partial v^2} &= \frac{\partial h}{\partial y^2} = \frac{\partial h}{\partial w^2} = 0 \\
h(y, 0, 0) &= y & G(y, 0, 0) &= \exp(-e^{-y} - Be^{-yc^2})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial G}{\partial v} &= G \frac{\partial}{\partial h} \left(-e^{-h} - Be^{-hc^2}\right) \frac{\partial h}{\partial v} \\
&= G(e^{-h} + Bc^2e^{-hc^2})(y + a_nb_n) \\
\frac{\partial G}{\partial v}_{(y,0,0)} &= [\exp(-e^{-y} - Be^{-yc^2})](e^{-y} + Bc^2e^{-yc^2})(y + a_nb_n)
\end{aligned}$$

similarly,

$$\begin{aligned}
\frac{\partial G}{\partial w} &= G \frac{\partial}{\partial h} \left(-e^{-h} - Be^{-hc^2}\right) \frac{\partial h}{\partial w} \\
&= G(e^{-h} + Bc^2e^{-hc^2})a_n \\
\frac{\partial G}{\partial w}_{(y,0,0)} &= [\exp(-e^{-y} - Be^{-yc^2})](e^{-y} + Bc^2e^{-yc^2})a_n
\end{aligned}$$

Furthermore, since $\frac{\partial}{\partial v}(e^{-h} + Bc^2e^{-hc^2}) = (-e^{-h} - Bc^4e^{-hc^2})(y + a_nb_n)$

$$\begin{aligned}
\frac{\partial^2 G}{\partial v \partial v} &= [G(e^{-h} + Bc^2e^{-hc^2})(y + a_nb_n)](e^{-h} + Bc^2e^{-hc^2})(y + a_nb_n) \\
&\quad - G[(e^{-h} + Bc^4e^{-hc^2})(y + a_nb_n)](y + a_nb_n) \\
&= G(e^{-h} + Bc^2e^{-hc^2})^2(y + a_nb_n)^2 - G(y + a_nb_n)^2(e^{-h} + Bc^4e^{-hc^2}) \\
&= G(y + a_nb_n)^2[(e^{-h} + Bc^2e^{-hc^2})^2 - (e^{-h} + Bc^4e^{-hc^2})] \\
\frac{\partial^2 G}{\partial v \partial v}_{(y,0,0)} &= [\exp(-e^{-y} - Be^{-yc^2})](y + a_nb_n)^2[(e^{-y} + Bc^2e^{-yc^2})^2 - (e^{-y} + Bc^4e^{-yc^2})]
\end{aligned}$$

Since $\frac{\partial}{\partial w}(e^{-h} + Bc^2e^{-hc^2}) = (-e^{-h} - Bc^4e^{-hc^2})a_n$

$$\begin{aligned}
\frac{\partial^2 G}{\partial w \partial w} &= [G(e^{-h} + Bc^2e^{-hc^2})a_n](e^{-h} + Bc^2e^{-hc^2})a_n \\
&\quad - G[(e^{-h} + Bc^4e^{-hc^2})a_n]a_n \\
&= G(e^{-h} + Bc^2e^{-hc^2})^2a_n^2 - Ga_n^2(e^{-h} + Bc^4e^{-hc^2}) \\
&= Ga_n^2[(e^{-h} + Bc^2e^{-hc^2})^2 - (e^{-h} + Bc^4e^{-hc^2})] \\
\frac{\partial^2 G}{\partial w \partial w}_{(y,0,0)} &= [\exp(-e^{-y} - Be^{-yc^2})]a_n^2[(e^{-y} + Bc^2e^{-yc^2})^2 - (e^{-y} + Bc^4e^{-yc^2})] \\
\frac{\partial^2 G}{\partial w \partial v} &= [G(e^{-h} + Bc^2e^{-hc^2})a_n](e^{-h} + Bc^2e^{-hc^2})(y + a_nb_n) \\
&\quad - G[(e^{-h} + Bc^4e^{-hc^2})a_n](y + a_nb_n) \\
&= G(e^{-h} + Bc^2e^{-hc^2})^2a_n(y + a_nb_n) - Ga_n(y + a_nb_n)(e^{-h} + Bc^4e^{-hc^2}) \\
&= Ga_n(y + a_nb_n)[(e^{-h} + Bc^2e^{-hc^2})^2 - (e^{-h} + Bc^4e^{-hc^2})] \\
\frac{\partial^2 G}{\partial w \partial v}_{(y,0,0)} &= [\exp(-e^{-y} - Be^{-yc^2})]a_n(y + a_nb_n)[(e^{-y} + Bc^2e^{-yc^2})^2 \\
&\quad - (e^{-y} + Bc^4e^{-yc^2})]
\end{aligned}$$

$$\begin{aligned}
G_n(y, v, w) &= G_n(y, 0, 0) + \begin{bmatrix} 0 & v & w \end{bmatrix} \begin{bmatrix} \partial G / \partial y \\ \partial G / \partial v \\ \partial G / \partial w \end{bmatrix}_{(y, 0, 0)} \\
&\quad + \begin{bmatrix} 0 \\ v \\ w \end{bmatrix}^t \begin{bmatrix} \partial^2 G / \partial y^2 & \partial^2 G / \partial y \partial v & \partial^2 G / \partial y \partial w \\ \partial^2 G / \partial v \partial y & \partial^2 G / \partial v^2 & \partial^2 G / \partial v \partial w \\ \partial^2 G / \partial w \partial y & \partial^2 G / \partial w \partial v & \partial^2 G / \partial w^2 \end{bmatrix} \begin{bmatrix} 0 \\ v \\ w \end{bmatrix}_{(y, 0, 0)} \\
&\quad + o(v^2, w^2, vw) \\
&= \exp(-e^{-y} - Be^{-yc^2}) + v \frac{\partial G}{\partial v}_{(y, 0, 0)} + w \frac{\partial G}{\partial w}_{(y, 0, 0)} \\
&\quad + \begin{bmatrix} v \partial^2 G / \partial v \partial y + w \partial^2 G / \partial w \partial y \\ v \partial^2 G / \partial v^2 + w \partial^2 G / \partial w \partial v \\ v \partial^2 G / \partial v \partial w + w \partial^2 G / \partial w^2 \end{bmatrix}^t \begin{bmatrix} 0 \\ v \\ w \end{bmatrix}_{(y, 0, 0)} \\
&\quad + o(v^2, w^2, vw) \\
&= \exp(-e^{-y} - Be^{-yc^2}) + v \frac{\partial G}{\partial v}_{(y, 0, 0)} + w \frac{\partial G}{\partial w}_{(y, 0, 0)} \\
&\quad + (v^2 \frac{\partial^2 G}{\partial v^2}_{(y, 0, 0)} + vw \frac{\partial^2 G}{\partial v \partial w}_{(y, 0, 0)}) \\
&\quad + (vw \frac{\partial^2 G}{\partial v \partial w}_{(y, 0, 0)} + w^2 \frac{\partial^2 G}{\partial w^2}_{(y, 0, 0)}) + o(v^2, w^2, vw)
\end{aligned}$$

$$\begin{aligned}
&G_n(y, v, w) \\
&= \exp(-e^{-y} - Be^{-yc^2}) \left\{ 1 + (e^{-y} + Bc^2 e^{-yc^2})[v(y + a_n b_n) + w a_n] \right\} \\
&\quad + \exp(-e^{-y} - Be^{-yc^2}) \left\{ \begin{aligned} &[(e^{-y} + Bc^2 e^{-yc^2})^2 - (e^{-y} + Bc^4 e^{-yc^2})] \\ &\times [v^2(y + a_n b_n)^2 + w^2 a_n^2 + 2vwa_n(y + a_n b_n)] \end{aligned} \right\} \\
&\quad + o(v^2, w^2, vw)
\end{aligned}$$

Let $c > 0, d > 0$ be two constants, we next find upper and lower bound for the expression of $G_n(y, v, w)$, and show their difference goes to 0 asymptotically. From Lemma 3.14 and 3.7, it was shown that

$$\begin{aligned} v &= \frac{s_n}{\sigma} - 1 \sim O_p\left(\sqrt{\frac{m}{n}}\right) = O_p\left(\frac{1}{(\sqrt{n})^{1-k}}\right) \\ w &= \frac{\bar{X}_n - \mu\sqrt{m}}{\sigma} \sim O_p\left(\sqrt{\frac{m}{n}}\right) = O_p\left(\frac{1}{(\sqrt{n})^{1-k}}\right) \end{aligned}$$

so

$$\begin{aligned} & \Pr\{a_n(M_n^o - b_n) \leq y(1+v) + a_nb_nv + a_nw\} \\ = & \Pr\{a_n(M_n^o - b_n) \leq y(1+v) + a_nb_nv + a_nw; \\ & |v| \leq \frac{c}{(\sqrt{n})^{1-k-\rho}}, |w| \leq \frac{d}{(\sqrt{n})^{1-k-\rho}}\} \\ & + \Pr\{a_n(M_n^o - b_n) \leq y(1+v) + a_nb_nv + a_nw; \\ & |v| \leq \frac{c}{(\sqrt{n})^{1-k-\rho}}, |w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}\} \\ & + \Pr\{a_n(M_n^o - b_n) \leq y(1+v) + a_nb_nv + a_nw; \\ & |v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}, |w| \leq \frac{d}{(\sqrt{n})^{1-k-\rho}}\} \\ & + \Pr\{a_n(M_n^o - b_n) \leq y(1+v) + a_nb_nv + a_nw; \\ & |v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}, |w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}\} \end{aligned}$$

The and the upper bound is

$$\begin{aligned} & \Pr\{a_n(M_n^o - b_n) \leq y(1+v) + a_nb_nv + a_nw\} \tag{3.19} \\ \leq & \Pr\{a_n(M_n^o - b_n) \leq y(1+v) + a_nb_nv + a_nw; \\ & |v| \leq \frac{c}{(\sqrt{n})^{1-k-\rho}}, |w| \leq \frac{d}{(\sqrt{n})^{1-k-\rho}}\} \end{aligned}$$

$$\begin{aligned}
& + \Pr\{|w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}\} + 2 \Pr\{|v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}\} \\
& \rightarrow \exp(-e^{-y} - B e^{-y c^2}) \\
& \quad \times \left\{ \begin{aligned} & 1 + (e^{-y} + B c^2 e^{-y c^2}) \left[\frac{c}{(\sqrt{n})^{1-k-\rho}} (y + a_n b_n) + \frac{d}{(\sqrt{n})^{1-k-\rho}} a_n \right] \\ & + [(e^{-y} + B c^2 e^{-y c^2})^2 - (e^{-y} + B c^4 e^{-y c^2})] \\ & \times \left[\frac{c^2}{n^{1-k-\rho}} (y + a_n b_n)^2 + \frac{d^2}{n^{1-k-\rho}} a_n^2 + \frac{2cd}{n^{1-k-\rho}} a_n (y + a_n b_n) \right] \\ & + o\left(\frac{1}{n^{1-k-\rho}}\right) \end{aligned} \right\} \\
& \times (1 + u_n) \\
& + \Pr\{|w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}\} + 2 \Pr\{|v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}\} \\
& = H_{n1}(y, c, d) + \Pr\{|w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}\} + 2 \Pr\{|v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}\}
\end{aligned}$$

Similarly, the lower bound can be obtained by

$$\begin{aligned}
& \Pr\{a_n(M_n^o - b_n) \leq y(1 + v) + a_n b_n v + a_n w\} \\
& \geq \Pr\{a_n(M_n^o - b_n) \leq y(1 + v) + a_n b_n v + a_n w ; \\
& \quad |v| \leq \frac{c}{(\sqrt{n})^{1-k-\rho}}, |w| \leq \frac{d}{(\sqrt{n})^{1-k-\rho}}\} \\
& \geq \Pr\{a_n(M_n^o - b_n) \leq y[1 - \frac{c}{(\sqrt{n})^{1-k-\rho}}] - a_n b_n \frac{c}{(\sqrt{n})^{1-k-\rho}} \\
& \quad - a_n \frac{d}{(\sqrt{n})^{1-k-\rho}}; |v| \leq \frac{c}{(\sqrt{n})^{1-k-\rho}}, |w| \leq \frac{d}{(\sqrt{n})^{1-k-\rho}}\}
\end{aligned}$$

Let A = the event that $a_n(M_n^o - b_n) \leq y(1 + v) + a_n b_n v + a_n w$, B = the event $|v| \leq \frac{c}{(\sqrt{n})^{1-k-\rho}}$, and C = the event $|w| \leq \frac{d}{(\sqrt{n})^{1-k-\rho}}$,

$$\begin{aligned}
\Pr(A) &= \Pr(ABC) + \Pr[A(BC)^c] \\
&= \Pr(ABC) + \Pr[A(B^c \cup C^c)] \\
&= \Pr(ABC) + \Pr[AB^c \cup AC^c] \\
\Pr(ABC) &= \Pr(A) - \Pr[AB^c \cup AC^c] \\
&= \Pr(A) - \Pr(AB^c) - \Pr(AC^c) + \Pr(AB^c \cap AC^c) \\
&\geq \Pr(A) - \Pr(AB^c) - \Pr(AC^c) \\
&\geq \Pr(A) - \Pr(B^c) - \Pr(C^c)
\end{aligned}$$

$$\begin{aligned}
&\Pr\{a_n(M_n^o - b_n) \leq y(1+v) + a_nb_nv + a_nw\} \\
\geq &\Pr\{a_n(M_n^o - b_n) \leq y[1 - \frac{c}{(\sqrt{n})^{1-k-\rho}}] - a_nb_n\frac{c}{(\sqrt{n})^{1-k-\rho}} - a_n\frac{d}{(\sqrt{n})^{1-k-\rho}}\} \\
&- \Pr\{|v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}\} - \Pr\{|w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}\}
\end{aligned}$$

$$\begin{aligned}
&\rightarrow \exp(-e^{-y} - Be^{-yc^2}) \\
&\times \left\{ \begin{aligned} &1 - (e^{-y} + Bc^2e^{-yc^2})[\frac{c}{(\sqrt{n})^{1-k-\rho}}(y + a_nb_n) + \frac{d}{(\sqrt{n})^{1-k-\rho}}a_n] \\ &+ [(e^{-y} + Bc^2e^{-yc^2})^2 - (e^{-y} + Bc^4e^{-yc^2})] \\ &\times [\frac{c^2}{n^{1-k-\rho}}(y + a_nb_n)^2 + \frac{d^2}{n^{1-k-\rho}}a_n^2 + \frac{2cd}{n^{1-k-\rho}}a_n(y + a_nb_n)] \\ &+ o(\frac{1}{n^{1-k-\rho}}) \end{aligned} \right\} \\
&\times (1 + u_n) \\
&- \Pr\{|v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}\} - \Pr\{|w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}\}
\end{aligned}$$

$$= H_{n2}(y, c, d) - \Pr(|v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}) - \Pr(|w| > \frac{d}{(\sqrt{n})^{1-k-\rho}})$$

But for $w = \frac{\bar{X}_n - \mu\sqrt{m}}{\sigma}$, $Var(w) \sim O(\frac{m}{n}) = O(\frac{1}{n^{1-k}})$,

$$\begin{aligned} & \Pr \left\{ |w| > \frac{d}{(\sqrt{n})^{1-k-\rho}} \right\} \\ &= \Pr \left\{ \left(\frac{\bar{X}_n - \mu\sqrt{m}}{\sigma} \right)^2 > \frac{d^2}{n^{1-k-\rho}} Var(\bar{X}_n) \right\} \\ &= \Pr \left\{ \left(\frac{\bar{X}_n - \mu\sqrt{m}}{\sigma} \right)^2 > \frac{d^2}{n^{1-k-\rho}} n^{1-k} \right\} \\ &\leq \frac{1}{dn^\rho} = O(n^{-\rho}) \end{aligned}$$

and for $v = \frac{s_n}{\sigma} - 1 \sim O_p(\frac{1}{\sqrt{n}^{1-k}})$, $Var(v) \sim O(\frac{m}{n}) = O(\frac{1}{n^{1-k}})$,

$$\begin{aligned} & \Pr \left\{ |v| > \frac{c}{(\sqrt{n})^{1-k-\rho}} \right\} \\ &= \Pr \left\{ \left(\frac{s_n}{\sigma} - 1 \right)^2 > \frac{c^2}{n^{1-k-\rho}} Var(\frac{s_n}{\sigma}) \right\} \\ &= \Pr \left\{ \left(\frac{s_n}{\sigma} - 1 \right)^2 > \frac{c^2}{n^{1-k-\rho}} n^{1-k} \right\} \\ &\leq \frac{1}{c^2 n^\rho} = O(n^{-\rho}) \end{aligned}$$

Now since

$$\begin{aligned} & H_2(y, c, d) - \Pr(|v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}) - \Pr(|w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}) \\ &\leq G_n(y, v, w) \\ &\leq H_1(y, c, d) + \Pr(|w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}) + 2 \Pr(|v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}) \end{aligned}$$

For $a_n \sim \sqrt{2 \log n}$ and $b_n \sim \sqrt{2 \log n}$, The difference between upper and lower bound is

right side - left side

$$\begin{aligned}
&= 3 \Pr(|v| > \frac{c}{(\sqrt{n})^{1-k-\rho}}) + 2 \Pr(|w| > \frac{d}{(\sqrt{n})^{1-k-\rho}}) \\
&\quad + \left\{ \begin{aligned} &2 \exp(-e^{-y} - B e^{-y c^2}) \times (e^{-y} + B c^2 e^{-y c^2}) \\ &\times [\frac{c}{(\sqrt{n})^{1-k-\rho}} (y + a_n b_n) + \frac{d}{(\sqrt{n})^{1-k-\rho}} a_n] + o(\frac{1}{n^{1-k-\rho}}) \end{aligned} \right\} \\
&\quad \times (1 + u_n) \\
&\rightarrow 0
\end{aligned}$$

Therefore, the probability $\Pr\{a_n(M_n^o - b_n) \leq y(1 + v) + a_n b_n v + a_n w\}$ converges to the upper bound as $n \rightarrow \infty$, and

$$\begin{aligned}
&\Pr\{a_n(M_n^o - b_n) \leq y(1 + v) + a_n b_n v + a_n w\} \\
&\rightarrow \exp(-e^{-y} - B e^{-y c^2}) \left\{ 1 + O\left(\frac{(\log n)^2}{\sqrt{n}^{1-k-\rho}}\right) \right\} \left\{ 1 + O\left(\frac{\log \log n}{\log n}\right) \right\} + O\left(\frac{1}{n^\rho}\right) \\
&= \exp(-e^{-y} - B e^{-y c^2}) \left\{ 1 + O\left(\frac{\log \log n}{\log n}\right) \right\}
\end{aligned}$$

■

Table3.1 Comparison of Estimates and True x (from Neton's method) for n=200

c	estimate	true x	error	F(x)	h(x)
n=200 epsilon=0.05					
2	2.5631	2.9518	0.13168	0.99842	-9.77E-06
3	3.8447	3.8627	0.00468	0.99994	-7.21E-05
5	6.4078	6.4075	0.00004	1	-8.05E-05
10	12.8155	12.8155	0	1	-1.32E-06
n=200 epsilon=0.1					
2	3.2897	3.3598	0.020854	0.99961	-6.72E-06
3	4.9346	4.9346	0.000003	1	-6.24E-05
5	8.2243	8.2243	0	1	-3.3E-06
10	16.4485	16.4485	0.000001	1	-1.11E-05
n=200 epsilon=0.2					
2	3.9199	3.9259	0.00151024	0.99996	-1.39E-05
3	5.8799	5.8799	0.000000044	1	-2.75E-06
5	9.7998	9.7998	0.000001905	1	-2.29E-05
10	19.5996	19.5994	0.000011608	1	-5.67E-05
n=200 epsilon=0.5					
2	4.6527	4.6528	0.000026403	1	-9.14E-07
3	6.979	6.979	0.000010099	1	-6.03E-05
5	11.6317	11.6317	0.000000114	1	-6.38E-06
10	23.2635	23.2635	0.000001234	1	-2.1E-05
epsilon	estimate	true x	error	F(x)	h(x)
n=200 c=2					
0.05	2.5631	2.95179	0.13168	0.99842	-9.77E-06
0.1	3.28971	3.35977	0.02085	0.99961	-6.72E-06
0.2	3.91993	3.92586	0.00151	0.99996	-1.39E-05
0.5	4.6527	4.65282	0.00003	1	-9.14E-07
n=200 c=3					
0.05	3.84465	3.86274	0.004682749	0.99994	-7.21E-05
0.1	4.93456	4.93458	0.000003208	1	-6.24E-05
0.2	5.87989	5.87989	0.000000044	1	-2.75E-06
0.5	6.97904	6.97897	0.000010099	1	-6.03E-05
n=200 c=5					
0.05	6.4078	6.4075	0.000041196	1	-8.05E-05
0.1	8.2243	8.2243	0.000000051	1	-3.3E-06
0.2	9.7998	9.7998	0.000001905	1	-2.29E-05
0.5	11.6317	11.6317	0.000000114	1	-6.38E-06
n=200 c=1					
0.05	12.8155	12.8155	0.000000011	1	-1.32E-06
0.1	16.4485	16.4485	0.000000574	1	-1.11E-05
0.2	19.5996	19.5994	0.000011608	1	-5.67E-05
0.5	23.2635	23.2635	0.000001234	1	-2.1E-05

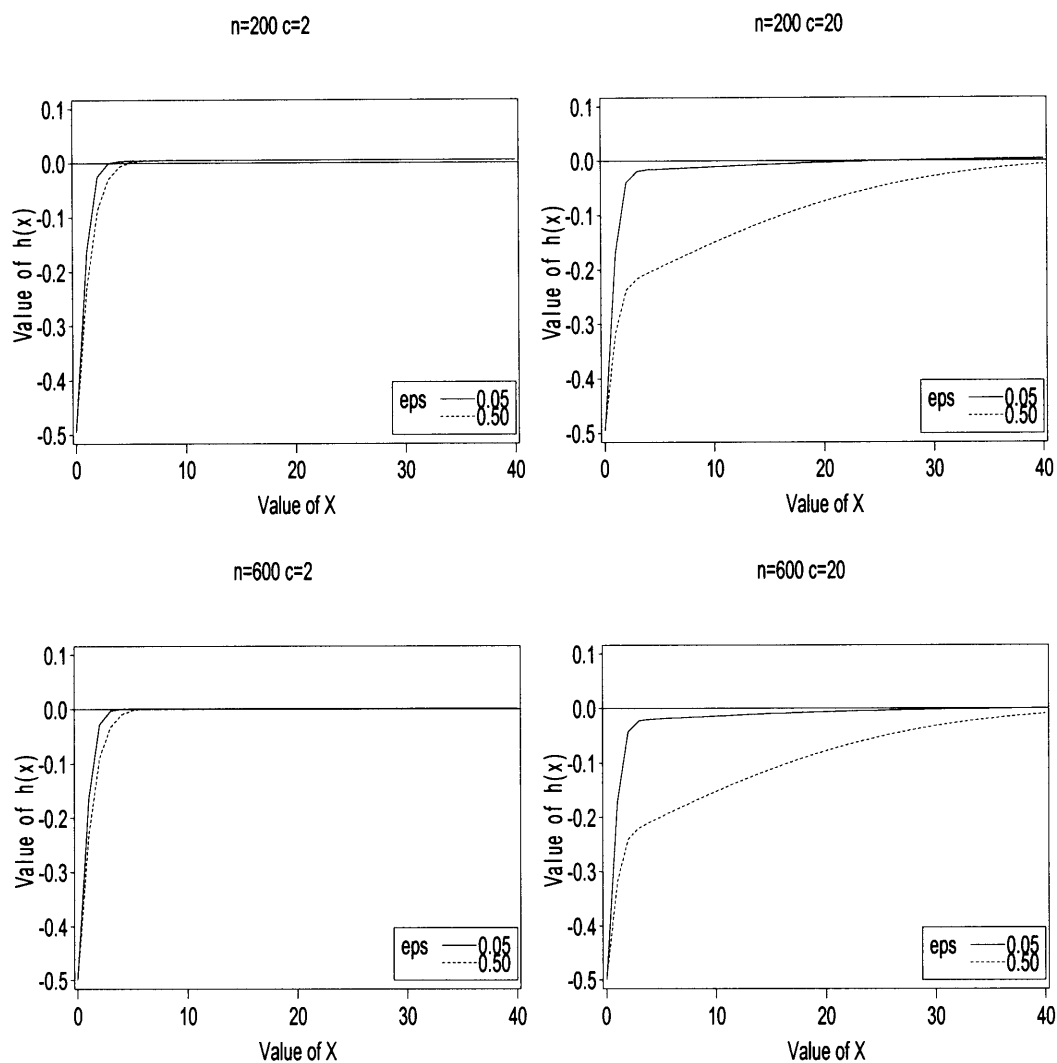
* error = |estimate - true x| / true x

Table3.2 Comparison of Estimates and True x (from Neton's method) for n=400

c	estimate	true x	error	F(x)	h(x)
n=400 epsilon=0.05					
2	3.2897	3.4139	0.036369	0.99968	-1.94E-05
3	4.9346	4.9342	0.000067	1	-7.84E-05
5	8.2243	8.2243	0	1	-4.26E-06
10	16.4485	16.4485	0.000003	1	-1.31E-05
n=400 epsilon=0.1					
2	3.9199	3.9329	0.003301484	0.99996	-1.83E-05
3	5.8799	5.8799	0.00000004	1	-2.88E-06
5	9.7998	9.7998	0.000005375	1	-1.93E-05
10	19.5996	19.5991	0.000029104	1	-4.51E-05
n=400 epsilon=0.2					
2	4.4828	4.4836	0.000180685	1	-4.29E-05
3	6.7242	6.7242	0.000004184	1	-1.88E-05
5	11.207	11.2059	0.000097229	1	-9.23E-05
10	22.414	22.414	0.00000031	1	-5.11E-06
n=400 epsilon=0.5					
2	5.1517	5.1517	0.000002964	1	-7.25E-06
3	7.7275	7.7275	0.00000025	1	-5.12E-06
5	12.8791	12.8789	0.000019241	1	-4.54E-05
10	25.7583	25.7583	0.000000034	1	-1.89E-06
epsilon	estimate	true x	error	F(x)	h(x)
n=400 c=2					
0.05	3.28971	3.41387	0.036369	0.99968	-1.94E-05
0.1	3.91993	3.93291	0.003301	0.99996	-1.83E-05
0.2	4.48281	4.48362	0.000181	1	-4.29E-05
0.5	5.15166	5.15167	0.000003	1	-7.25E-06
n=400 c=3					
0.05	4.93456	4.93423	0.000067007	1	-7.84E-05
0.1	5.87989	5.87989	0.00000004	1	-2.88E-06
0.2	6.72421	6.72418	0.000004184	1	-1.88E-05
0.5	7.72749	7.72749	0.00000025	1	-5.12E-06
n=400 c=5					
0.05	8.2243	8.2243	0.00000034	1	-4.26E-06
0.1	9.7998	9.7998	0.000005375	1	-1.93E-05
0.2	11.207	11.2059	0.000097229	1	-9.23E-05
0.5	12.8791	12.8789	0.000019241	1	-4.54E-05
n=400 c=10					
0.05	16.4485	16.4485	0.000003183	1	-1.31E-05
0.1	19.5996	19.5991	0.000029104	1	-4.51E-05
0.2	22.414	22.414	0.00000031	1	-5.11E-06
0.5	25.7583	25.7583	0.000000034	1	-1.89E-06

* error = |estimate - true x| / true x

Figure 3.3 Largest Characteristic Observation: $h(x)$ vs. x



$$*h(x) = (1-\epsilon) F_o(x) + \epsilon F_o(x/c) - 1 + 1/n$$

CHAPTER 4

NUMERICAL EXAMPLE

4.1 Description of the Data

In this section, we demonstrate the utility of our theory by applying our main theorem to a real data set. We need two types of data: a protein profile and a protein databank. The Ig profile for Immunoglobulin domain (accession number PF00047) consisting 113 sequences was downloaded from pfam website at <http://pfam.wustl.edu/cgi-bin/getdesc?name=ig>. Pfam is a database of high-quality, manually checked, well characterized multiple alignments provided by Washington University at St. Louis and the Sanger Centre in United Kingdom. A consecutive portion of 25 columns from the profile was selected for this example (Figure 4.1). The immunoglobulin domain is found in many diverse proteins. The primary functions of this domain include protein-protein and protein-ligand interactions. We computed the column statistics from the profile, that is, the proportion of each amino acid that appears in each column. This is shown in Figure 4.2.

Next, the UniProt protein database in FASTA format was downloaded from the European Bioinformatics Institute website at <http://www.ebi.ac.uk/FTP/>. This is a comprehen-

sive database consisting of fully classified, accurately annotated protein sequences provided freely to the scientific community by the UniProt Consortium which consists the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). Of all the protein sequences, 4654 sequences with length more than 600 amino acids were selected. Figure 4.3 shows some examples of entries used for computation. The first line contains protein ID, name and descriptions. The next lines are the actual amino acids composition of the sequences.

4.2 Analysis Strategies

PERL program was used to compute the scores X_{nj} . The Ig protein profile was moved along each of the $N = 4654$ sequences, and each comparison yielded a score X_{nj} . To standardize X_{nj} , we need to compute \bar{X}_n, s_n^2 where $\bar{X}_n = \frac{\sum_{j=1}^n X_{nj}}{n}$, and $s_n^2 = \frac{1}{n[1+(c^2-1)\epsilon]} \sum_{j=1}^n (X_{nj} - \bar{X}_n)^2$ so that $Y_{nj} = \frac{X_{nj} - \bar{X}_n}{s_n}$. This requires knowing parameters c and ϵ . We used the Maximum Likelihood method (ML) introduced by Mott (1992) to estimate c , and ϵ . This approach assesses the significance of each comparison using a distribution fitted to the set of scores obtained from a data-bank search. As noted by the author, this procedure is valid since the query will be unrelated to the vast majority of the data-bank sequences, and consequently nearly all the scores obtained from a search may be treated as a random sample. Mott (1992) has shown that for pairwise alignment with gaps, the distribution of Smith-Waterman alignment scores estimated from ML is very similar to those obtained by simulation. On the other hand, compared to the time-consuming sim-

ulation of random sequences, the maximum likelihood method can be implemented more rapidly and requires less computing resource.

We next illustrate the application of ML method to our setup: our main theorem says under suitable conditions,

$$\Pr \{a_n(M_n - c_n) \leq y\} \rightarrow \exp \left(-e^{-y} \left\{ 1 + \frac{1-\varepsilon}{c\varepsilon} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon} \right)^{c^2-1} e^{-y(c^2-1)} \right\} \right)$$

where $M_n = \max_{1 \leq j \leq n} Y_{nj}$, $Y_{nj} = \frac{X_{nj} - \bar{X}_n}{s_n}$ and $s_n^2 = \frac{1}{n[1+(c^2-1)\varepsilon]} \sum_{j=1}^n (X_{nj} - \bar{X}_n)^2$. We write

$$\begin{aligned} Y_{nj} &= \frac{X_{nj} - \bar{X}_n}{s_n} \\ &= \frac{X_{nj} - \bar{X}_n}{\sqrt{\frac{\sum_{j=1}^n (X_{nj} - \bar{X}_n)^2}{n[1+\varepsilon(c^2-1)]}}} \\ &= \sqrt{[1+\varepsilon(c^2-1)]} \frac{X_{nj} - \bar{X}_n}{\sqrt{\sum_{j=1}^n (X_{nj} - \bar{X}_n)^2/n}} \\ &= \sqrt{[1+\varepsilon(c^2-1)]} U_n \end{aligned}$$

Without knowledge for c and ε , the statistics we get initially is U_n . So,

$$\begin{aligned} &\Pr \{a_n(M_n - c_n) \leq y\} \\ &= \Pr \left\{ a_n \left(\sqrt{[1+\varepsilon(c^2-1)]} U_n - c_n \right) \leq y \right\} \\ &= \Pr \left\{ \sqrt{[1+\varepsilon(c^2-1)]} U_n \leq \frac{y}{a_n} + c_n \right\} \\ &= \Pr \left\{ U_n \leq \frac{y}{a_n \sqrt{[1+\varepsilon(c^2-1)]}} + \frac{c_n}{\sqrt{[1+\varepsilon(c^2-1)]}} \right\} \end{aligned}$$

Let $z = \frac{y}{a_n \sqrt{[1+\varepsilon(c^2-1)]}} + \frac{c_n}{\sqrt{[1+\varepsilon(c^2-1)]}}$, then $y = g(z) = z a_n \sqrt{[1+\varepsilon(c^2-1)]} - a_n c_n$, so

$$\begin{aligned}
& \Pr\{U_n \leq z\} \\
& \rightarrow \exp\left(-e^{-g(z)} \left\{1 + \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon e^{g(z)}}\right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon}\right\}\right) \\
& = \exp\left\{-e^{-g(z)} - e^{-g(z)c^2} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon}\right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon}\right\} \\
& = F(z, n, c, \varepsilon)
\end{aligned}$$

Taking derivative with respect to z , we get

$$\begin{aligned}
f(z, n; c, \varepsilon) &= \frac{\partial F(z, n, c, \varepsilon)}{\partial z} = \frac{\partial F(z, n, c, \varepsilon)}{\partial g(z)} \frac{\partial g(z)}{\partial z} \\
&= \exp\left\{-e^{-g(z)} - e^{-g(z)c^2} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon}\right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon}\right\} \\
&\quad \times \left\{e^{-g(z)} + c^2 e^{-g(z)c^2} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon}\right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon}\right\} \times a_n \sqrt{[1+\varepsilon(c^2-1)]}
\end{aligned}$$

Now taking log, we get

$$\begin{aligned}
\log f(z, n; c, \varepsilon) &= \left\{-e^{-g(z)} - e^{-g(z)c^2} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon}\right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon}\right\} \\
&\quad + \log\left\{e^{-g(z)} \left[1 + c^2 \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon e^{g(z)}}\right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon}\right]\right\} \\
&\quad + \log\left(a_n \sqrt{[1+\varepsilon(c^2-1)]}\right)
\end{aligned}$$

$$\begin{aligned}
&= \left\{ -e^{-g(z)} - e^{-g(z)c^2} \left(\frac{\sqrt{4\pi \log n\epsilon}}{n\epsilon} \right)^{c^2-1} \frac{1-\epsilon}{c\epsilon} \right\} - g(z) \\
&\quad + \log \left\{ 1 + c^2 \left(\frac{\sqrt{4\pi \log n\epsilon}}{n\epsilon e^{g(z)}} \right)^{c^2-1} \frac{1-\epsilon}{c\epsilon} \right\} \\
&\quad + \log \left(a_n \sqrt{[1 + \epsilon(c^2 - 1)]} \right) \\
&= \left\{ -e^{-g(z)} \left[1 + \left(\frac{\sqrt{4\pi \log n\epsilon}}{n\epsilon e^{g(z)}} \right)^{c^2-1} \frac{1-\epsilon}{c\epsilon} \right] \right\} - g(z) \\
&\quad + \log \left\{ 1 + c^2 \left(\frac{\sqrt{4\pi \log n\epsilon}}{n\epsilon e^{g(z)}} \right)^{c^2-1} \frac{1-\epsilon}{c\epsilon} \right\} \\
&\quad + \log \left(a_n \sqrt{[1 + \epsilon(c^2 - 1)]} \right)
\end{aligned}$$

Maximizing $L(c, \epsilon) = \sum_{i=1}^N \log f(z, n; c, \epsilon)$ with respect to c and ϵ then gives the maximum likelihood estimates \hat{c} and $\hat{\epsilon}$.

We implemented the maximization step via Newton-Raphson Ridge Optimization in PROC NLP (Nonlinear Programming Procedure) by SAS. This procedure computes analytic formula for gradient functions automatically and requires only the objective function and the constraint for the parameters to be specified. Therefore, the results are more accurate in that errors in typing and derivatives calculations are eliminated. For our problem, the parameter constraints are: $c > 1$, and $0 < \epsilon < 1$. We provided the initial estimate $c_0 = 1.5$, $\epsilon_0 = 0.05$, the program returned the final estimates $\hat{c} = 2.233004$ and $\hat{\epsilon} = 0.248135$ with gradient 0.001159 and -0.001621 respectively. We then normalized the scores X_{nj} by computing $Y_{nj} = \frac{X_{nj} - \bar{X}_n}{s_n}$ using these estimates. Finally, the significance of each score is tested by computing its predicted p-value $1 - F(Y_{nj}, n, c, \epsilon)$. Therefore, with this method, although the parameters and distribution of the score Y_{nj} is estimated using information from all the data, the significance of each comparison is tested against a different null hypothesis

that allows for difference in length.

4.3 Results

We next assessed the goodness of fit for the approximation using Quantile-Quantile plot (Figure 4.4). According to the theorem, the j th largest normalized maximum score should be approximately equal to the $j/(N + 1)$ th quantile from the extreme value distribution $G(y) = \exp\left(-e^{-y} \left\{1 + \left(\frac{\sqrt{4\pi \log n \varepsilon}}{n \varepsilon e^y}\right)^{c^2-1} \frac{1-\varepsilon}{c\varepsilon}\right\}\right)$. Since the sequence in the database have different length n , we compared to the quantiles from $e^{-e^{-y}}$ instead. The normalized maximum scores were ordered and plotted against the quantiles from the extreme value distribution. In Figure 4.4, the solid line at 45° represents perfect fit, although there is some departure at the tail, our results show very good fit for scores less than 6. Examining P values for the scores show the cut off point for $\alpha = 0.01$ occurs at 4.58206, and the cut off point for $\alpha = 0.05$ occurs at 2.96917. Therefore, the theorem provides accurate approximation in regions where estimates of critical values at 1% and 5% are identified. Moreover, comparing our graph with that was shown in Goldstein and Waterman (1994), where the scores were treated as fixed and modeled with normal distributions, our results show much better approximation.

Table 4.1 shows detailed information for the top 20 sequences with smallest P values that were found to be significantly associated with the Ig domain. Comparison of the 12th sequence with the 15th sequence shows that although AKH_BUCBP with 816 residues has smaller raw score 4.064 compared to the raw score of 4.516 from CAML_FUGRU

with 1277 residues, AKH_BUCBP was found to be more significantly related to Ig domain with smaller P value. Therefore, our approximation provides a way to adjust for the fact that a shorter sequence is less likely to match the profile well than a longer one by chance alone.

Figure 4.1 The Immunoglobulin Profile

GVSKGNFSIGRMTQDLA.GTYRCYG
VNRTFQADSPLDPATHG.GAYRCFG
RTFQADFPLGPATHGG...TYRCFG
RIFQESFNMSPVTTAHA.GNYTCRG
KYNMKVLYLSAFTSKDE.GTYTCAL
GNGTFSVVLNQLTAEDE.GFYWCVS
ENGTFVVNIAQLSQDDS.GRYKCGL
RKNLFLVEVTQLTESDS.GVYACGA
STSSFNFTITASQVVDS.AVYFCAL
SNSSFHLRKASVHWSDS.AVYFCAV
ASLHFSLHIRDSQPSDS.ALYLCAV
SNSPCSLEIQSSEAGDS.ALYLCAS
NQSHSTLKIQSTQPQDS.AVYLCAS
PDSHSELNMTSLELTDS.ALYLCAS
KKSSFSLTVTSAQKNEM.TVFLCAS
KREHFSLILDSAKTNQT.SVYFCAQ
SKNQVVL SMNTVGP GDT.ATYYCAR
SRNQFSLNLRSM SAADT.AMYYCAR
KTSTTVDLKTS LPTEDT.ATYFCAR
AKDSL YLQMNSLRAEDT.AVYYCAP
SFNQA HMELSSLFSED T.AVYYCAR
SGTSASLAISGLQSENE.ADYFCAT
SGSDYSLIIGSLESEDF.ADY YCLQ
SGTZFTLTLSDVZCDDA.ATYYCGG
PGQPAQLQLNATESDDG.RSFFCSA
DDGRFQLQLQNVQPPSS.GIYSATY
GSAIYTFRIQNIEVS DM.GPYECQV
ERYDASILLWKLQFDDN.GTYTCQV
EKGSFPLIINKLKMEDS.QTYICEL
RTTHSSVLIITPRPQDHGTNLTCQV
DDSSSTLTIYNANIDDA.GIYKCVV
KTNDNSLTI AKTMELDS.GEYTCVA
GHYGKSLVIRQTNFDDA.GTYTCDV
GLGVLSLLIRRP GPFDG.GTYGCRA
LLNGSQLVLHGLELGHS.GLYACFH
FIEDGRLVIHSLDYSDQ.GNYSCVA
HARVSSLTLKSIQYTDA.GEYVCTA
GTFSSVLTLTNLTGLDT.GEYFCTH
TTTGNTLVLRDVQLSDT.GDYLC SL
QNHNKTLQLLKVGEEDD.GEYRCLA
LSGNATLTLIAMRMEDS.GIYVCEG
IPSDATLEIQNLR SNDS.GIYRCEV
SQDNLTITMHLQLSDT.GTYTCQA
SQGNTTLSINPVKREDA.GTYWCEV

Figure 4.1 The Immunoglobulin Profile

RNTSSEYHIARAEREDA.GFYWCEV
RVITSRLKINPVKEEDA.TTFTCMA
VVEQDGLTILNVTEMDD.GTYTCRA
KGRISTLTFFNVSEKDY.GNYTCVA
YQKVLTLNLDHVSFQDA.GNYSCTA
NVHAINLTLVNVTSEDNGFTLTCIA
. .DSATLQLYDLVKD. . .TSATCVA
NHAQFSLQVANITED. . .TTFNCVA
GATGTTLTVRATARTDG.TRYRAVF
.VSQTQDFTKIASKSDS.GTYICTA
QVGQTMLHISKVSKEAE.GSYM CVV
.TSTPSYRITSASVNDS.GEYRCQR
YERQATLTISSARVNDS.GVFM CYA
SEDELSIRLSNITVHDE.GVYKCY
VAQGAQLLIRPVDKPIN.TTFICNV
TPEKSQLTISNLDVNVDPGTYVCNA
SYHSYRLHINRLHGSDN.GTYYCCT
VKMDLGLLFLRVRKSDA.GTYFCQT
DRTTAEFEVPSLTLGDS.GFWE CRV
RRFIASFNVVNTTKRDA.GKYRCMI
IRPFYTLRVSPVTPEDS.GTYRCRL
NSTLPFLKIMDVKPEDS.GFYFCAM
RTSGKRLLFKTTLPEDE.GVYTCEV
RITGEEVEVRDAVPEDS.GLYACMT
MIRILLAFVSSVGRNDT.GYYTCSS
NESRVNLTIQGLRAVDT.GLYLCKV
AEGSVAVRIQEVKASDD.GEYRCFF
GLLNTSITFWNTTLDDE.GCYMCLF
DLQDLSIFITNVTYNHS.GDYECHV
TLRASQILIIKEIYRNE. .EFTCVS
NYIEVPLIFDPVTREDLHMDFKCVV
GLSFQAFSYLNFTPEPS.DIFSCIV
DYSFHKFHYLTFVPSAED.FYDCRV
DHSFFKISYLTFLPSADE.IYDCKV
GTYNIFSTLRFTPVEGD. .IYSCSV
SFYLLSHAEFTPN SKD. . .QYSCRV
WSFYLLYYTEFTPNEKD. .EYACRV
TFQILVMLEMPQRGD. . .VYTCHV
WTYQLLVLLETTPPRGL. .TYSCQV
WTYQLLVLLEIPPQLGV. .SYTCQV
WFYQI HSHLEYTPRSGE. .KISCVV
WYYQTHSHLEYTPRSGE. .KISCVV
WYYQI HSHLEYTPKSGE. .KISCMV
FYQI HSELE.YTPKSGE. .KISCMV
DGTYQMEKQTD FEPSPD.AKFSCEV

Figure 4.1 The Immunoglobulin Profile

DGTFQKWAAVVPSGQE.QRYTCHV
QIWSVLRLPVALSPSLD..TYTCVV
MSSTLTMSSEEFKYS...TMTCEV
YSMSSTLTLTKEDEYERH.NSYTCEA
YSMSSTLSLTKADYESH.NLYTCEV
FQKSTHLTVTPEEWKNN..KYQCVV
TEEKVHIPKILPWHAG...TYS CVA
KESQLNFDSISPEDAG...SYSCWV
TYQTLSHLALTPSYGD...TYTCVV
TLCSLTIEKVMPEDGG...EYKCIA
GISILRIEPVRAGRDD..APYECVA
VDGVATVCASEWDGGD...GYVCKV
IYSKLNMKTSKWEKTD...SFSCNV
AHSILTVSEEEWNTGE...TYTCVV
VTSILRVAAEDWKKGD...TFSCMV
DSVISTVDISTQAWLSE.AVFYCVV
ITSILPVVAKDWIEGY...GYQCIV
QHSRLTLPRSLWNAGT...SVTCTL
VTSILRVKDPKTQVGK...EVICQV
ATSQVLVPSKDV LQGTE.EYLVCKV
TNNKYVLTLNKFSKENE.GYYFCSV
KSPGYVLDLIVTPQNKS.TFYTCQV
YTLSSSVTPSSPRPSE..TVTCNV
TTTSTLTVLAYGPNS...TATCLV

Table 4.1 Profile Statistics for Ig Profile

Col	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	4.5	2.7	1.8	1.8	11.5	1.8	1.8	3.5	4.4	5.4	5.3	9.7	1.8	7.1	3.5
C	0.0	0.0	0.9	0.0	0.9	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.9	0.0
D	8.2	3.6	3.5	6.2	0.9	0.9	1.8	3.5	0.9	2.7	8.8	1.8	4.4	1.8	9.7
E	2.7	5.4	4.4	1.8	1.8	3.5	1.8	7.1	3.5	8.0	3.5	1.8	11.5	7.1	22.1
F	2.7	3.6	4.4	4.4	13.3	1.8	8.8	2.7	4.4	1.8	1.8	6.2	0.9	2.7	0.9
G	10.0	6.3	8.8	7.1	1.8	1.8	0.0	0.0	0.0	2.7	2.7	0.9	4.4	2.7	14.2
H	0.9	4.5	2.7	4.4	1.8	6.2	1.8	7.1	0.0	2.7	0.9	0.0	1.8	1.8	0.9
I	3.6	5.4	1.8	10.6	3.5	0.0	8.8	3.5	29.2	5.4	1.8	3.5	1.8	0.9	0.0
K	7.3	5.4	1.8	4.4	3.5	2.7	0.0	5.3	3.5	4.5	9.7	0.0	7.1	10.6	3.5
L	1.8	6.3	2.7	2.7	21.2	7.1	48.7	12.4	27.4	6.3	2.7	19.5	2.7	5.3	3.5
M	1.8	0.0	2.7	0.9	0.0	1.8	3.5	0.0	4.4	1.8	0.9	2.7	0.9	1.8	0.9
N	5.5	7.1	7.1	5.3	1.8	5.3	0.0	5.3	0.0	8.0	13.3	0.0	4.4	1.8	8.0
P	1.8	1.8	1.8	1.8	1.8	2.7	0.0	3.5	1.8	3.6	7.1	8.8	12.4	13.3	2.7
Q	3.6	4.5	5.3	15.0	2.7	7.1	0.0	5.3	0.9	6.3	3.5	0.9	10.6	4.4	2.7
R	8.2	5.4	5.3	2.7	0.0	6.2	0.9	5.3	0.9	8.0	6.2	0.0	6.2	9.7	2.7
S	12.7	8.9	24.8	16.8	14.2	22.1	3.5	7.1	3.5	13.4	16.8	3.5	8.8	8.0	18.6
T	8.2	16.1	10.6	8.0	8.0	21.2	0.9	19.5	2.7	10.7	8.8	14.2	15.9	3.5	4.4
V	6.4	5.4	0.9	3.5	6.2	7.1	14.2	6.2	9.7	5.4	1.8	21.2	2.7	7.1	1.8
W	5.5	0.0	0.9	0.0	0.0	0.0	0.9	0.0	0.0	1.8	0.0	5.3	0.9	2.7	0.0
Y	4.5	8.0	8.0	1.8	5.3	0.9	2.7	1.8	2.7	1.8	4.4	0.0	0.0	7.1	0.0

Col	16	17	18	19	20	21	22	23	24	25
A	2.8	12.8	0.0	20.5	0.9	1.8	3.5	1.8	14.2	19.5
C	0.0	0.0	0.0	0.0	0.9	0.0	0.0	98.2	0.9	0.0
D	61.5	8.5	16.7	1.2	4.4	0.0	1.8	0.0	0.9	0.0
E	1.8	19.1	16.7	1.2	8.8	0.0	3.5	0.0	8.8	0.0
F	0.0	1.1	0.0	1.2	5.3	11.5	8.8	0.0	3.5	2.7
G	10.1	4.3	33.3	57.8	1.8	0.0	0.9	0.0	2.7	5.3
H	3.7	3.2	16.7	0.0	0.0	0.0	0.0	0.0	2.7	1.8
I	0.9	0.0	0.0	0.0	6.2	3.5	3.5	0.0	3.5	0.9
K	2.8	0.0	0.0	0.0	6.2	0.0	4.4	0.0	3.5	0.0
L	1.8	2.1	0.0	0.0	6.2	2.7	6.2	0.0	3.5	6.2
M	0.0	2.1	0.0	1.2	0.9	0.9	2.7	0.0	5.3	0.9
N	3.7	5.3	0.0	2.4	4.4	0.0	0.9	0.0	3.5	0.0
P	0.9	1.1	16.7	0.0	1.8	0.0	0.0	0.0	0.0	0.9
Q	1.8	1.1	0.0	2.4	0.9	0.0	1.8	0.0	8.8	1.8
R	0.9	0.0	0.0	1.2	2.7	0.0	8.8	0.0	7.1	4.4
S	3.7	25.5	0.0	1.2	7.1	0.0	14.2	0.0	4.4	6.2
T	1.8	11.7	0.0	9.6	27.4	0.0	24.8	0.0	5.3	3.5
V	0.9	1.1	0.0	0.0	11.5	2.7	4.4	0.0	17.7	44.2
W	0.0	0.0	0.0	0.0	0.0	0.9	2.7	0.0	0.9	0.0
Y	0.9	1.1	0.0	0.0	2.7	76.1	7.1	0.0	2.7	1.8

Figure 4.2 Protein Sequences in FASTA format

```
>104K_THEPA (P15711) 104 kDa microneme-rhoptry antigen
MKFLILLFNILCLFPVLAADNHGVGPQGASGVDPITFDINSNQTGPAFLTAVEMAGVKYL
QVQHGSNVNIHRLVEGNVVIWENASTPLYTGAIVTNNDGPY MAYVEVLGDPNLQFFIKSG
DAWVTLSEHEYLA KLQEIRQAVHIESVFSLNMAFQLENNKYEVETHAKNGANMVTFIPRN
GHICKMVYHKNVRIYKATGNDTVTSVVGFFRGLRLLLINVFSIDDNGMMSNRYFQHVDDK
YVPISQKNYETGIVKLKDYKHAYHPVDLDIKDIDYTMFHLADATYHEPCFKIIPNTGFICI
TKLFDGDQVLYESFNPLIHCINEVHIYDRNNGSIICLHLNYSPPSYKAYLVLKDTGW EAT
THPLLEEKIEELQDQ RACELDVNFISDKDLVVAALTNADLN YTMVTPRPHRDVIRVSDGS
EVLWYYEGLDNFLVCAWIYVSDGVASLVHLRIKDRI PANNDIYVLKGDLYWTRITKIQFT
QEIKRLVKKSKKKLAPITEEDSDKHDEPPEGPGASGLPPKAPGDKEGSEGHKGPSKGS DS
SKEGKKPGSGKKPGPAREHKPSKIPTLSKKPSGPKDPKHPRDPKEPRKSKSPRTASPTRR
PSPKLPQLSKLPKSTSPRSPPPTRPSSPERPEGTKIIKTSKPPSPKPPFDP SFKEKFYD
DYSKAASRSKETKTTVVLD ESFESILKETLPETPGTPTFTT PRVPVPPKRPRTPE SPFEPPK
DPDSPSTSPSEFFTPPESKRTRFHETPADTLPDVT AELFKEPDVTAETKSPDEAMKRPR
SPSEYEDTSPGDYPSLPMKRHLRLRLTTTETMETDPGRMAKDASGKPVKLKRSKSFDDL
TTVELAPEPKASRIVDDEGTEADDEETHPPEERQKTEVRRRRRPPKKPSKSPRPSKPKKP
KKPDSAYIP SILAILVVS LIVGIL

>11S3_HELAN (P19084) 11S globulin seed storage protein G3
precursor (Helianthinin G3)
MASKATLLLAFTLLFATCIARHQQRQQQONQCQLQNI EALEPIEVIQAEAGVTEIWDAYD
QQFQCAWSILFDTGFNLVAFSCLPTSTPLFWPSSREGVILPGCRRTYEYSQEQQFSGEGG
RRGGGEGTFRTVIRKLENLKEGDVVAIPTGTAHWLHNDGNTEL VVVFLDTQNHENQLDEN
QRRFFLAGNPQAQAQSQQQQQRQPRQQSPQRQRQRQ RQGQGNAGNIFNGFTPELIAQSF
NVDQETAQKLGQNDQRGHIVNVGQDLQIVRPPQDRRS PRQQQE QATS PRQQQEQQQGRR
GGWSNGVEETICSMKFKNIDNPSQADFNVPQAGSIANLNSFKFP ILEHLRLSVERGELR
PNAIQSPHWTINAHNLLYVTEGALRVQIVDNQGNSVFDNELREGQVVVIPQNFAVIKRAN
EQGSRWVSFKTNDNAMIANLAGRVSASAASPLTLWANRYQLSREEAQQ LKFSQRETVLFA
PSFSRGQGIRASR

>120K_RICRI (P14914) 120 kDa surface-exposed protein
MVIQSANATGQVNF RHIVDVGADGTTAFKTAASKVTITQDSNFGNTDFGNLAAQIKVPNA
ITLTGNFTGDASNPGNTAGVITFDANGTLESASADANVAVTNNITAEASGAGVVQLSGT
HAAELRLGNAGSIFKLADGTVINGKVNQTALVGGALAA GTITLDGSATITGDIGNAGGAA
ALQRITLANDAKKTLTLGGANIIGAGGGTIDLQANGGTIKLTSTQNNIVVDFDLAIATDQ
TGVVDASSLTNAQTLTINGKIGTIGANNKTLGQFNIGSSKTVLSNGNVAINELVIGNDGA
VQFAHDTYLI TRTTNAAGQGKIIIFNPVVNNGTTLAAGTNLGSATNPLAEINFGSKGVNVD
TVLNVGEGVNLYATNITTTDANVGSFVFNAGGTNIVSGTVGGQQGNKFNTVALENGTTVK
FLGNATFNNGNTTIAANSTLQIGGNYTADCVASADGTGIVEFVNTGPITVTLNKQAAPVNA
LKQITVSGPGNVVINEIGNAGNHGAVTDTIAFENSSLGAVVFLPRGIPFNDAGNTMPLT
IKSTVGNKTAKGFDVPSVVVLGVDSVIADGQVIGDQNNIVGLGLGSDNGIIVNATTLYAG
ISTLNNNQGTVTLSGGVPNTPGTVYGLGTGIGASKFKQVTF TTDYNNLGNIIATNATIND
GVTVTTGGIAGIGFDGKITLGSVNGNGNVRFADGILSNSTSMIGTTKANNGTVTYLGN AF
VGNIGSDSTPVASVRFTGSDSGAGLQGNIIYSQVIDFGTYNLGIVNSNIILGGGTTAINGK
IDLVTNTLTFASGTSTWGNNTSIETTLTLANGNIGHIVILEGAQVNTTTTGT TTIKVQDN
ANANFSGTQTYTLIQGARFNGTLGSPNFAVTGSNRFVNYSLIRAANQDYVITRTNNAEN
VVTNDIANSPFGGAPGVDQNVTTFVNATNTAA YNNLLLLAKNSANSANFVGAIVTDTSAAI
TNVQLDLAKDIIQAQLGNRLGALRYLGTPETAEMAGPEAGAI SAAVAAGDEAIDNVAYGIW
```

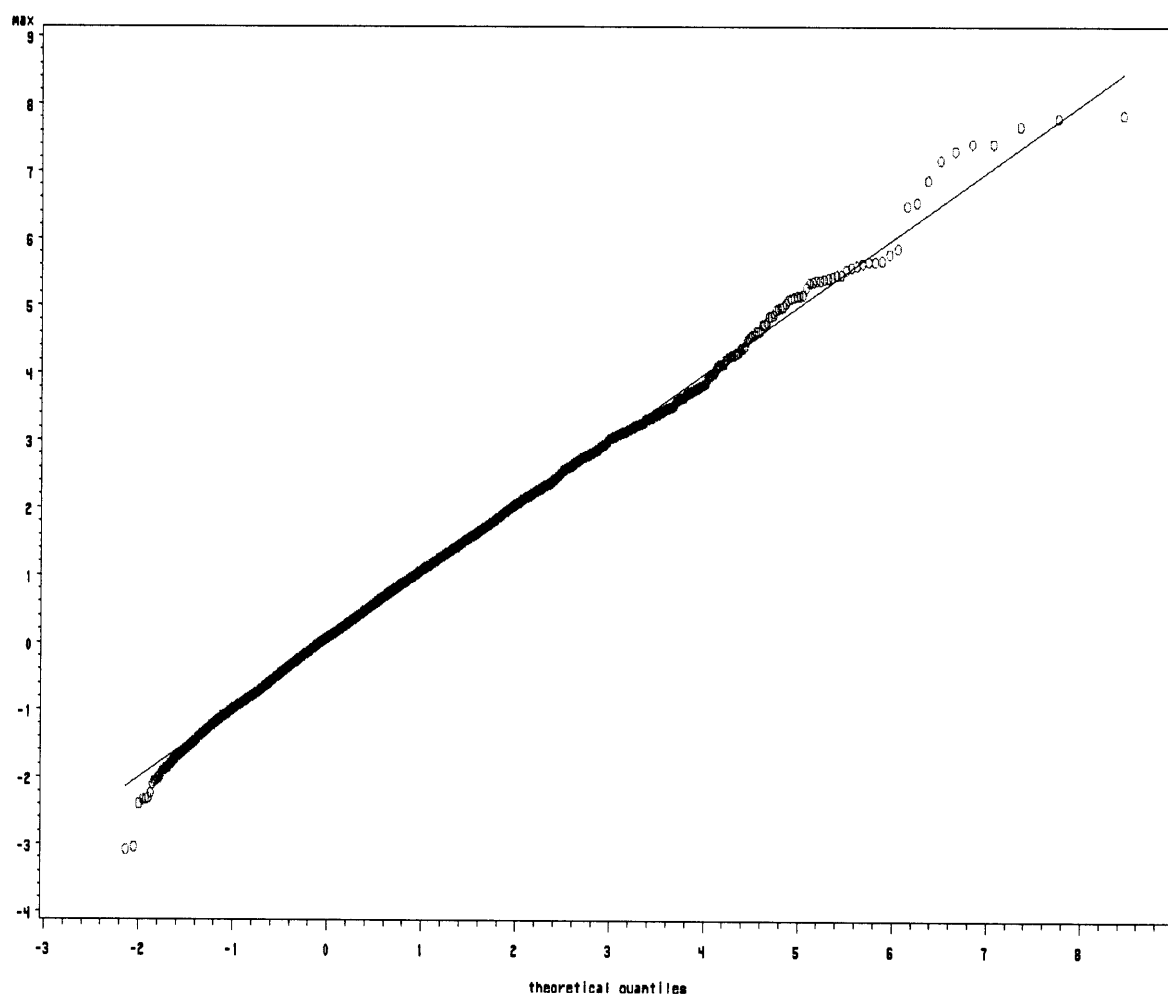
Table 4.2 Sequences found statistically most similar to Ig (Immunoglobulin) domain

	name	length	raw_max	mean	var	std_max*	an	cn	norm_max*	pvalue
1	BIP2_MAIZE	663	4.501	1.379	0.156	11.148	1.425	5.651	7.836	0.00040
2	BIP3_MAIZE	663	4.501	1.379	0.157	11.113	1.425	5.651	7.785	0.00042
3	BIP1_ARATH	669	4.459	1.380	0.155	11.030	1.427	5.658	7.664	0.00047
4	CADJ_HUMAN	772	4.802	1.427	0.191	10.892	1.447	5.772	7.410	0.00060
5	BIP2_ARATH	668	4.413	1.384	0.155	10.851	1.427	5.657	7.409	0.00061
6	BIP_SPIOL	668	4.406	1.388	0.156	10.777	1.427	5.657	7.304	0.00067
7	BIP4_TOBAC	667	4.377	1.386	0.156	10.680	1.426	5.656	7.167	0.00077
8	BIP5_TOBAC	668	4.288	1.384	0.153	10.471	1.427	5.657	6.867	0.00104
9	CAO2_CANTR	708	4.358	1.403	0.165	10.260	1.435	5.704	6.539	0.00145
10	BIP_LYCES	666	4.307	1.386	0.163	10.204	1.426	5.655	6.488	0.00152
11	4CL_VANPL	553	4.122	1.383	0.159	9.688	1.399	5.503	5.853	0.00287
12	AKH_BUCBP	816	4.064	1.369	0.151	9.781	1.455	5.816	5.772	0.00311
13	BTB9_MOUSE	612	4.311	1.391	0.184	9.601	1.414	5.586	5.676	0.00342
14	ALAB_ARATH	1203	4.224	1.384	0.165	9.861	1.509	6.111	5.658	0.00348
15	CAML_FUGRL	1277	4.516	1.429	0.194	9.885	1.517	6.156	5.658	0.00348
16	ACE1_TRIRE	733	4.231	1.387	0.173	9.644	1.440	5.731	5.634	0.00357
17	A2M2_MOUSE	1451	4.238	1.413	0.162	9.899	1.534	6.249	5.600	0.00369
18	ALA4_ARATH	1216	4.189	1.389	0.162	9.812	1.511	6.119	5.577	0.00378
19	CHS5_SCHPO	620	3.890	1.430	0.133	9.514	1.416	5.597	5.545	0.00390
20	ALAA_ARATH	1202	4.224	1.386	0.169	9.737	1.509	6.111	5.471	0.00420

std_max = (raw_max - mean)/sqrt(var/(1+eps(c*c-1)))

* norm_max = an*(std_max-bn)

Figure 4.3 Quantile-Quantile Plot of Empirical Scores
vs. Theoretical Quantiles from Extreme Value Distribution
(from searching UniProt Database with Ig Profile)



CHAPTER 5

SUMMARY AND FUTURE WORK

Although almost all of current theories on similarity scores have focused on studying statistical distributions of i.i.d. sequences, biological sequences often exhibit heterogeneities and dependencies. In Chapter two, we have studied the approximation of the statistical distribution for the fixed exact local alignment score R_n in the Headruns problem. Our study represents an important starting point for modeling the variations and dependencies inherent in biological sequences. In the first scenario, we studied independent sequences and modeled variations of the matching probabilities $\tilde{p} = (p_1, \dots, p_n)$ by random variables with a common expectation. In the second scenario, we studied Markov dependent sequences, and in addition modeled the transition probabilities $\tilde{\alpha} = (\alpha_1, \dots, \alpha_n)$ by random variables with the same expectation. To further relax the "identical" part of the assumption, this problem can be further explored by modeling the different properties at separate regions along the sequences with random variables of different expectations and possibly with different forms. One can also relax the "dependency" part of the assumption by modeling the dependencies with higher orders of Markovian structure and evaluate numerically the improvement on the approximation. We have imposed an independence condition for the

relation between the matching probabilities and the transition probabilities, an extension would be to study the situation when the matching probabilities and transition probabilities are more dependent.

Another important property of biological sequences is that during the course of evolution, sequences often undergo changes through mutations such as substitutions, insertions and deletions. This means that two related sequences may exhibit approximate rather than exact matching patterns. For i.i.d. sequences, local alignment scores for approximate matching, matching with gaps allowed, matching with shifts have been studied. See Arratia et al. (1990), Dembo et al. (1994) and Siegmund et al. (2000). While allowing for heterogeneities and dependencies, further studies are needed to study the corresponding properties of local approximate alignment scores under these more general schemes.

In each of these setups, we conjecture the Chen-Stein method of Poisson approximation will be a useful tool. For the i.i.d. case, as pointed out by Arratia et al. (1990), there are two distinct issues: the expected number of events λ must be approximated and the dependence among the events being counted must be controlled via b_1 and b_2 . The second issue is simple for fixed local alignment scores, but is very complicated for scores from alignments with shifts. We anticipate a similar picture under the more general setups.

Accompanying the approximation theory is the question of how to estimate the parameters \tilde{p} and $\tilde{\alpha}$. Because direct estimation via computing the empirical distribution function of alignment scores is time consuming and demands large computing resources, the declumping method introduced by Waterman et al. (1994) and the Maximum Likelihood

method pioneered by Mott (1992) have been the popular approaches. These methods are general and can be easily adapted to estimate parameters under more general setups. Their performance in these situations can then be evaluated numerically.

In Chapter three, we derived statistical distributions for the maximum of profile scores obtained by comparing multiple alignment profiles with sequences in databanks. We have assumed the profile is random with independent sequences and accommodated the possibilities for gaps in the profile. The simple sequence was assumed to be an i.i.d. sequence with the residues following i.i.d. distributions. As in the pairwise alignment problem noted above, further studies in more general settings that allow for heterogeneity, dependency, and gapping in the simple sequence and dependency among sequences in the profile are needed to model the properties of biological sequences more precisely. Moreover, as we have used simple indicator functions to construct matching scores, one limitation is the inability to allow for mismatches that vary in degree. A natural extension in this area would be to study profile scores with popular substitution matrices such as PAM or BLOSUM matrices. In each of these setups, the tail behaviors of Y_{nj} s needs to be approximated and the distribution for their maximum needs to be derived.

Again, as in the pairwise alignment problem, we are also faced with estimation of parameters for profile score distributions. In Chapter 4, we have used the Maximum Likelihood (ML) method (Mott, 1992) to obtain estimates for c and ε . We then compared scores $M_n = \max_{1 \leq j \leq n} Y_{nj}$ where $Y_{nj} = \frac{X_{nj} - \bar{X}_n}{s_n}$ and $s_n^2 = \frac{1}{n[1+(c^2-1)\varepsilon]} \sum_{j=1}^n (X_{nj} - \bar{X}_n)^2$ to quantiles from the extreme value distribution $e^{-e^{-y}}$. The performance of our approximation in the main

theorem, $\Pr \{a_n(M_n - c_n) \leq y\} \rightarrow \exp \left(-e^{-y} \left\{ 1 + \frac{1-\varepsilon}{c\varepsilon} \left(\frac{\sqrt{4\pi \log n\varepsilon}}{n\varepsilon} \right)^{c^2-1} e^{-y(c^2-1)} \right\} \right)$, however, can only be evaluated using simulated sequences with equal lengths. This is because real sequences with different lengths will converge to different distributions on the right side of the equation. This can be implemented by simulating a set of random sequences, say 10,000, with uniform compositions, and compare each sequence to a given profile. Then, estimate c and ε from the scores obtained. For given \hat{c} , $\hat{\varepsilon}$, and n , quantiles from distribution on the right hand side can then be computed via numerical methods and compared to empirical scores normalized using \hat{c} and $\hat{\varepsilon}$.

The next question that arises is that: will the distribution for a particular comparison with parameters (c, ε) estimated by simulation, be close to the results of a databank search and substituting in the values of c and ε ? This can be evaluated by comparing critical values derived from comparing a set of profiles, say 20, to a set of simulated random sequences with critical values identified by comparing the same set of profiles with a databank.

Finally, further numerical study can also be carried out to study the performance of the approximations with finite sequence length n . This will be important for the evaluation of asymptotic approximations to real biological sequences. On the other hand, extensions of the theories in the direction of sub-asymptotic behaviors of alignment scores will also be of interest.

REFERENCES

- Aach, J. *et al.* (2001) Computational comparison of two draft sequences of the human genome. *Nature* 409, 856-859
- Altschul, S.F. (1991) Amino acid substitution matrices from an information theoretical perspective. *Journal of Molecular Biology* 219, 555-565
- Altschul, S., Bundschuh, R., Olsen, R., Hwa, T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Research* 29, 351-361
- Aparicio, S. *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301-1310
- Arratia, R., Goldstein, L., Gordon, L. (1989) Two moments suffice for Poisson approximation: The Chen-Stein method. *Annals of Probability*, 17, 9-25
- Arratia, R., Gordon, L., Waterman, M.S. (1990) The Erdos-Renyi law in distribution, for coin tossing and sequence matching. *Annals of Statistics*, 18, 539-570
- Arratia, R., Morris, P., Waterman, M.S. (1988) Stochastic scrabble: Large deviations for sequences with scores. *Journal of Applied Probability*, 25, 106-119
- Arratia, R., Waterman, M.S. (1985) Critical phenomena in sequence matching. *Annals of Probability*, 13, 1236-1249
- Arratia, R., Waterman, M.S. (1989) The Erdos-Renyi strong law for pattern matching with a given proportion of mismatches. *Annals of Probability*, 17, 1152-1169
- Arratia, R., Waterman, M. S. (1994) A phase transition for the score in matching random sequences allowing deletion, ns. *The Annals of Applied Probability* 4, 200-225
- Baltimore, D. (2001) Our genome unveiled. *Nature* 409, 814-816
- Bentley, D.R., *et al.* (2001) The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* 409, 942-943
- Berg, J. M., Tymoczko, J. L., Stryer, L. (2002) *Biochemistry*. W.H. Freeman and Co. New York.
- Birney, E., *et al.* (2001) Mining the draft human genome. *Nature* 409, 827-828
- Bock, J. B., *et al.* (2001) A genomic perspective on membrane compartment organization. *Nature* 409, 839-841

- BOOK, S. A. (1970) Large Deviations Probabilities for Weighted Sums. *The Annals of Mathematical Statistics*, 43,1221-1234.
- Brüls, T., *et al.* (2001) A physical map of human chromosome 14. *Nature* 409, 947-948
- The Celera Genomics Sequencing Team (2001) The sequence of the human genome. *Science* 291, 1304-1351
- Casella, G., Berger, R. (1990) *Statistical Inference*. Wadsworth, Inc., Belmont, CA.
- Chen, L.H.Y. (1975) Poisson approximation for dependent trials. *Annals of Probability*, 3, 534-545
- Chvatal, V., Sankoff, D. (1975) Longest common subsequences of two random sequences. *Journal of Applied Probability*. 12, 306-315
- Clayton, J. D., *et al.* (2001) Keeping time with the human genome. *Nature* 409, 829-831
- Collins, J.F., Coulson, A.F.W. (1990) Significance of protein sequence similarities. *Methods in Enzymology* 183, 474
- Collins, J.F., Coulson, A.F.W., Lyall, A. (1988) The significance of protein sequence similarities. *Computer Applications in Biosciences*. 4, 67-71
- Coulson, A.F.W., Collins, J.F., Lyall, A. (1987) Protein and nucleic acid sequence database searching: a suitable case for parallel processing. *Computing Journal*, 30, 420
- Crollius, R.H., *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genetics*, 25, 235-238
- Cramer, H. (1893) *Random variable and probability distributions*. The University Press. Cambridge, England.
- Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. (1978) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation. Washington, DC, 5, 345-352
- Dembo, A., Karlin, S., Zeitouni, O. (1994a) Critical phenomena for sequence matching with scoring. *The Annals of Probability*. 22,1993-2021
- Dembo, A., Karlin, S., Zeitouni, O. (1994b) Limit distribution of maximal non-aligned two-sequence segmental score. *The Annals of Probability* 22, 2022-2039
- Durrett, R. (1991) *Probability: Theory and Examples*. Wadsworth, Inc., Belmont, CA.

- Enserink, M. (2001) Finding the talismans that protect against infection. *Science* 291, 1183.
- Ewens, W.J., Grant, G.R. (2001) *Statistical Methods in Bioinformatics*. Springer-Verlag, New York City
- Ewing, B., Green, P. (2000) Analysis of expressed sequence tag indicates 35,000 human genes. *Nature Genetics*, 25, 232-234
- Erdos, P., Renyi, A. (1970) On a new law of large numbers. *Journal d'Analyse Mathematique*, 22, 103-111
- Fahrer, A., *et al.* (2001) A genomic view of immunology. *Nature* 409, 836-838
- Futreal, P. A., *et al.* (2001) Cancer and genomics. *Nature* 409, 850-852.
- Goldstein, L., Waterman, M.S. (1994) Approximations to profile score distributions. *J. Comput. Biol.*, 1, 93-104.
- Gribskov, M., McLachlan, A.D., Eisenberg, D. (1987) *Proceedings of National Academy of Sciences, USA* 84, 4355-4358
- Henikoff, S., Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of National Academy of Sciences*, 89, 10915-10919
- Karlin, S., Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of National Academy of Sciences* 87, 2264-2268
- Karlin, S., Dembo, A., Kawabata, T. (1990) Statistical composition of high-scoring segments from molecular sequences. *Annals of Statistics*, 18, 571-581
- Karlin, S., Ghandour, G., Ost, F., Tavaré, S., Korn, L.J. (1983) New approaches for computer analysis of nucleic acid sequences. *Proceedings of National Academy of Sciences*, 80, 5660-5664
- Karlin, S., Ost, F. (1987) Counts of long aligned word matches among random letter sequences. *Advances in Applied Probability*. 19, 293-351
- Kotz, S. Johnson, N. L. (1981) *Encyclopedia of Statistical Sciences*, Volume 1, John Wiley and Sons, New York
- Helmuth, L. (2001) Brain calls dibs on many genes *Science* 291, 1188

- Henikoff, S., Henikoff, J.G. (1992) *Proceedings of National Academy of Sciences, USA*, 89, 10915-10919
- International Human Genome Sequencing Consortium. (2001a) Initial sequencing and analysis of the human genome. *Nature* 409, 860-921
- The International Human Genome Mapping Consortium (2001b) A physical map of the human genome. *Nature* 409, 934-941
- Klug, W.S., Cummings, M.R. (2000) *Concept of Genetics*, Prentice Hall, Upper Saddle River, NJ.
- Kotz, S., Johnson, N. L. (1981) *Encyclopedia of Statistical Sciences*, Volume 1, John Wiley and Sons, New York.
- Kulkarni, V.G., (1995) *Modeling and Analysis of Stochastic Systems*, Chapman and Hall, London, UK
- Leadbetter, M.R., Lindgren, G., Rootzen, H., (1980) *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York City
- Li, W. H., *et al.* (2001) Evolutionary analyses of the human genome. *Nature* 409, 847-849
- Marshall, E. (2001) Sharing the glory, not the credit. *Science* 291, 1189-1193.
- Marshall, E. (2001) Celera and *Science* spell out data access provisions. *Science* 291, 1191
- Marshall, E. (2001) Bermuda rules: Community spirit, with teeth *Science* 291, 1192
- Marx, J. (2001) Nailing down cancer culprits. *Science* 291, 1185.
- Montgomery, K.T., *et al.* (2001) A high-resolution map of human chromosome 12. *Nature* 409, 945-946
- Mott, R. (1992) Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bulletin of Mathematical Biology*, 54, 59-75
- Mott, R. (2000) Accurate formula for *p*-values of gapped local sequence and profile alignments. *Journal of Molecular Biology* 300, 649-659
- The Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-562

- Murray, A. W., *et al.* (2001) Can sequencing shed light on cell cycling? *Nature* 409, 844-846
- Needleman, S.B, Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, 48, 443-453.
- Nestler, E. J., *et al.* (2001) Learning about addiction from the genome. *Nature* 409, 834-835
- Neuhauser, C. (1994) A poisson approximation for sequence comparisons with insertions and deletions. *The Annals of Statistics* 22, 1603-1629
- Olsen, R., Bundschuh, R., and Hwa, T. (1999) Rapid assessment of extremal statistics for gapped local alignment. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology* AAAI Press, Menlo Park, CA, pp211-222
- Poindexter, K., Nelson, N., DuBose, R. F., Black, R. A., Cerretti, D.P. (1999) The identification of seven metalloproteinase-disintegrin (ADAM) genes from genomic libraries. *Gene* 237, 61-70
- Pollard, T. D. (2001) Genomics, the cytoskeleton and motility. *Nature* 409, 842-843
- Riethman, H. C., *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* 409, 948-951
- Roberts, L. (2001) Controversial from the start. *Science* 291, 1182-1188.
- Roberts, L., *et al.* (2001) A history of the human genome project *Science* 291, 1195
- Sagane, K., Ohya, Y., Hasegawa, Y., Tanaka, I. (1998) Metalloproteinase-like, cysteine-rich proteins MDC2 and MDC3: novel human cellular disintegrins highly expressed in the brain. *Biochemistry Journal* 334, 93-98
- Sankoff, D., Kruskal, J. B. (1983) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence comparison*. Addison-Wesley. Reading, MA.
- Schwartz, R.M., Dayhoff, M.O. (1978) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation. Washington, DC, 5, 353-358
- Sellers, P.H. (1984) Pattern recognition in genetic sequences by mismatch density. *Bulletin of Mathematical Biology*, 46, 501-514
- Sen, P.K., Singer, J.M., (1993) *Large Sample Methods in Statistics*, Chapman and Hall, London, UK

- Service, R. F. (2001) Objection #1: Big biology is bad biology. *Science* 291, 1182.
- Service, R. F. (2001) Objection #3: Impossible to do *Science* Feb 16 2001: 1186
- Sidow, A. (2002) Sequence first. Ask questions later. *Cell* 111, 13-16 (2002)
- Siegmund, D., Yakir, B. (2000) Approximate p -values for local sequence alignments. *The Annals of Statistics* 28, 657-680
- Smigielski, E. M., Sirotkin, K., Ward, M., Sherry, S. T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research* 28, 352-355
- Smith, T.F., Burks, C., Waterman, M.S. (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Research* 13, 645-656
- Stein, C.M. (1986) *Approximate Computation of Expectations*. IMS, Hayward, CA
- Tatusov, R.L., Altschul, S.F., Koonin, E.V. (1994) Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proceedings of National Academy of Sciences, USA*, 91, 12091-12095
- Tilford, C. A., *et al.* (2001) A physical map of the human Y chromosome. *Nature* 409, 943-945.
- Titterton, D. M. (1985) *Statistical analysis of finite mixture distributions*. Wiley, New York.
- Tupler, R., Perini, G., Green, M. R. (2001) Expressing the human genome. *Nature* 409, 823-833
- US Department of Energy Genome Program: www.ornl.gov/hgmis
- Vogel, G. (2001) Objection #2: Why sequence the junk? *Science* 291, 1184.
- Vogel, G. A (2001) parakeet genome project? *Science* 291, 1187
- von Bahr, Bengt (1965) On the convergence of moments in the central limit theorem. *The Annals of Mathematical Statistics*, 36, 808-818
- Waterman, M.S. (1995) *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman and Hall, London, UK
- Waterman, M.S., Gordon, L., Arratia, R. (1987) Phase transitions in sequence matches and nucleic acid structure. *Proceedings of National Academy of Sciences*, 87, 1239-1243

- Waterman, M.S., Vingron, M. (1994a) Sequence comparison significance and poisson approximation. *Statistical Science* 9, 367-381
- Waterman, M.S., Vingron, M. (1994b) Rapid and accurate estimates of statistical significance for sequence database searches. *Proceedings of National Academy of Sciences*, 91, 4625-4628
- Watson, G.S. (1954) Extreme values in samples from m -dependent stationary stochastic sequences. *Annals of Mathematical Statistics*, 25, 798-800
- Williams, D. (1991) *Probability with Martingales*. Cambridge University Press.
- Wolfsberg, T. G., *et al.* (2001) Guide to the draft human genome. *Nature* 409, 824-826