Parametric Modelling of Genomic Sequences Distance [*]

Aluísio de Souza Pinheiro
Departamento de Estatística
UNICAMP - SP, Brazil

Pranab Kumar Sen
Department of Biostatistics
UNC-Chapel Hill

Hildete Prisco Pinheiro
Departamento de Estatística
UNICAMP - SP, Brazil

Abstract

The paper considers the problem of homogeneity among groups by comparison of genomic sequences. Among the problems in that kind of analysis two points are specially addressed here. Genetic data perceives information as categorical variables and as a consequence the overall view of it generates strong dependence between genetic sites. The second problem is that usually models are built on the *cleaned* data (functional genetic such as genes) and the rest of the data that also carries information is dismissed as useless. We proceed here with emphasis on the available heuristic evidences of great diversity in statistical distributions for the categorical data available. A fully operational parametric statistical model is proposed. The model is built with flexibility to withstand use in several different organisms and adapt itself to that usually dismissed material and the dependence between sites. Consistency of the estimators and of derived test procedures are shown.

Keywords: Amino Acid; Asymptotic distribution; Maximum Likelihood Estimation; Categorical Data; Genome; Nucleotide; Statistical Genetics.

1 Introduction

Nowadays, a great interest in research is the understanding of the structure and evolution of genes and genomes. In order to understand these biological structures, it is important to know the general statistical characteristics of the *intron-exon* structures of eukaryotic genes. In view of this, a brief biological

[*] MSC 2000 subject classifications. Primary- 62F12, 62F03; secondary- 62M05, 92D20. Key words and phrases: DNA sequences, Categorical Data, Dependent Data, Genome, Exons, Introns, Codon, Nucleotide, Log-ratio test, Asymptotic normality, Fisher Information

This research was funded in part by Fundação de Amparo à Pesquisa do Estado de São Paulo (2002/04332-3).

background is given on section 2. Previous massive analysis of genetic data are used as motivation for a fully parametric model which is developed in section 3. Maximum likelihood estimators are found for that model are found in section 4. In section 5 asymptotic properties are investigated while in section 6 biologically hypotheses procedures are suggested on the spirit of maximum likelihood ratio tests.

2 Biological Motivation

A *gene*, in a general concept, can be defined as a sequence of genomic DNA (Deoxyribonucleic acid) or RNA (Ribonucleic acid) that is essential for a specific function. Gene, in the Mendelian setup, is the basic unit of inheritance. Genes occur at definite sites or *loci*, on *chromosomes*, which are strings of DNA, the basic genetic material in a *cell* and the carrier of genetic information for all organisms, except for some viruses. DNA is a double-helical model; it is a polymer, made up of nucleotides which are four in number, and can be distinguished by the four bases: A (*adenine*), C (*cytosine*), G (*guanine*) and T (*thymine*). Like the DNA, RNA and proteins are also macromolecules of a cell, though they differ in their forms and constitution. RNA differs from DNA by having ribose, instead of deoxyribose, also the T is replaced by U (*uracil*) in the RNA, and it has a single strand. Proteins are also polymers, and there are 20 amino acids. Most human cells contain 46 chromosomes, in 23 pairs; one pair relates to the *sex* chromosomes, while the other 22 homologous pairs are termed *autosomes.* There are about 50,000 genes embedded within the human genome. *Genetic data*, for diploid organisms, relate to traits determined by autosomal Mendelian loci, so that DNA plays a basic role in genetic data analysis (Li (1991), Waterman (1995), Lange (1997), Ewens and Grant (2001)).

Principles of molecular genetics, such as the central dogma that DNA makes RNA makes protein, govern computational sequence analysis (CSA). The transfer of genetic information from DNA to DNA (called *replication*) means that the molecule can be copied; the loop from DNA to RNA called *transcription* precedes the loop from RNA to protein, called *translation.* The RNA which is translated into protein is termed the *messenger* RNA (or $mRNA$), and the *transfer* RNA (or $tRNA$) translates the genetic code into amino acids. If we accept the basic role of DNA as the genetic information carrier, then it is natural to conclude that evolution is directly related to changes in DNA. This is the genesis of molecular evolution. Substitutions between purines only ($A \leftrightarrow G$) or pyrimidines only ($C \leftrightarrow T$) are called *transitions*, while substitutions between a purine and a pyrimidine ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$, or $G \leftrightarrow T$) are called *transversions* (recall that in a DNA, A pairs with T and G pairs with C).

Next, note that amino acids are encoded by triplets of nucleotides, called *codons.* Let us define $\mathcal{N}_R = \{A, C, G, U\}$, and let $\mathcal{C} = \{(x_1, x_2, x_3) : x_j \in \mathcal{N}_R, j = 1, 2, 3\}$ be the codon. Finally, let $\mathcal{X}$ be the set of aminoacids and termination codon, which can be seen on Table 1. The codon **AUG** specifies the aminoacid Methinonine (*Met/M*) and also serves as the starting codon for polypeptide synthesis. Any of the codons - **UAA**, **UAG**, or **UGA** - specifies the end, or termination of polypeptide synthesis. There are $2^3 = 64$ possible codons, but only 20 aminoacids. Then the *genetic code* can be defined as a map: $g : \mathcal{C} \rightarrow \mathcal{X},\ g \in \mathcal{G}$, so that $\mathcal{G}$ is the set of all genetic codes. As the human genome project is heading for a completion, there are some formidable statistical tasks which are surfacing into the research efforts to fathom out the mystery of the genomic code.

Table 1: The standard Genetic Code - Codon and correspondent aminoacid

Glycine (GLY)	Serine (SER)	Arginine (ARG)	Phenylalanine (PHE)
Alanine (ALA)	Threonine (THR)	Asparagine (ASN)	Tyrosine (TYR)
Valine (VAL)	Aspartic Acid (ASP)	Glutamine (GLN)	Tryptophan (TRP)
Leucine (LEU)	Glutamic Acid (GLU)	Cysteine (CYS)	Histidine (HIS)
Isoleucine (ILE)	Lysine (LYS)	Methionine (MET)	Proline (PRO)

UUU	Phe/F	**UCU**	Ser/S	**UAU**	Tyr/Y	**UGU**	Cys/C
UUC	Phe/F	**UCC**	Ser/S	**UAC**	Tyr/Y	**UGC**	Cys/C
UUA	Leu/L	**UCA**	Ser/S	**UAA**	*Stop*	**UGA**	*Stop*
UUG	Leu/L	**UCG**	Ser/S	**UAG**	*Stop*	**UGG**	Trp/W
CUU	Leu/L	**CCU**	Pro/P	**CAU**	His/H	**CGU**	Arg/R
CUC	Leu/L	**CCC**	Pro/P	**CAC**	His/H	**CGC**	Arg/R
CUA	Leu/L	**CCA**	Pro/P	**CAA**	Gln/Q	**CGA**	Arg/R
CUG	Leu/L	**CCG**	Pro/P	**CAG**	Gln/Q	**CGG**	Arg/R
AUU	Ile/I	**ACU**	Thr/T	**AAU**	Asn/N	**AGU**	Ser/S
AUC	Ile/I	**ACC**	Thr/T	**AAC**	Asn/N	**AGC**	Ser/S
AUA	Ile/I	**ACA**	Thr/T	**AAA**	Lys/K	**AGA**	Arg/R
AUG	*Met/M*	**ACG**	Thr/T	**AAG**	Lys/K	**AGG**	Arg/R
GUU	Val/V	**GCU**	Ala/A	**GAU**	Asp/D	**GGU**	Gly/G
GUC	Val/V	**GCC**	Ala/A	**GAC**	Asp/D	**GGC**	Gly/G
GUA	Val/V	**GCA**	Ala/A	**GAA**	Glu/E	**GGA**	Gly/G
GUG	Val/V	**GCG**	Ala/A	**GAG**	Glu/E	**GGG**	Gly/G

The transcribed RNA contains *untranslated regions*, *exons* and *introns*. *Introns*, or intervening sequences, are those transcribed sequences that are excised during the processing of the *pre-messenger RNA* (pre-mRNA) molecule. All genomic sequences that remain in the mature mRNA following *splicing* are referred as *exons*. *Exons* that are translated are referred as coding regions. The distribution of intron-exon structures of eukaryotic genes was studied by Deusch and Long (1999) and some interesting statistical phenomena was pointed out by them: ”Genome size seems to be correlated with total intron length per gene. For example, invertebrate introns are smaller than those of human genes. However, this correlation is weak, suggesting that other factors besides genome size may also affect intron size.” For humans, the size of exons ranges from 1-1644, with mean 50.9 and standard deviation 58.7. The size of introns, for humans, ranges from 25-54916, with mean 34131.1 and standard deviation 6552.6. The exon and total intron length distributions are both skewed to the left. For those with smaller length, the exon length distribution has large concentration between 1 and 75. For those smaller sequences one does find again left skewness.

Some empirically supported aspects of genetic sequences must be taken into account in order to assure correct probabilistic modelling and statistical analysis. One of those aspects is that exons and intron related distribution form a very rich class which poses some hard task in modelling. Even skewness and relative dispersion (taken here as the ratio *std. dev./mean*) can vary from group to group being studied. Control regions tend to have higher rate of substitutions. For that reason models must present certain versatility to those variations.

The model formulation we will discuss here can also be applied to mitochondrial sequences. In this case, some notational changes ought to be done. There are no introns in those sequences. One has promoter regions and exons. Promoter regions, as happens in usual DNA sequences with introns, are not very selective in their mutation process. Therefore, they tend to pass substitutions in higher rates when compared to exons. For our model which deals with introns (or promoter regions) only as distance between exons, no adaptation other than notational needs to be done when dealing with introns or promoter regions.

As one can see on Table 1 there are many codons that specify the same aminoacid, i.e., synonymous codon usage. One of the measures for synonymous codon usage is the *effective number of codons* (ENC), which is a measure of departure from equal codon usage that is independent of gene length, aminoacid composition and any reference set of genes (Wright (1990)). A low ENC corresponds to high codon usage bias, a high ENC to low codon usage bias. The minimum value of 20 occurs when on codon is used exclusively for each amino acid, and the maximum is 61 when synonymous codons are used equally. The typical range of ENC is $25 - 55$ (Hartl (2000)).

3 Notation and Model Formulation

Most problems in genomic sequence analysis are essentially statistical. For instance, stochastic evolutionary forces act on genomes. In genomic sequence analysis, typically, we encounter data on a large number (K) of positions or *sites*, and in each position, we have a purely qualitative (nucleotides or amino acid labels) categorical response with 4 to 20 categories depending on the DNA or protein sequence. The spatial (functional as well as stochastic) dependence (or association) patterns of these sites may not be known, nor can they be taken to be *stochastically independent*. Also, as has been mentioned before, regular and nearly identical structures of DNA calls for statistical appraisal based on other variational properties which exhibit more statistical variation and information too.

Some notation will be introduced here. The model concerns statistical comparisons of some characteristic(s) from several sequences. The difference in those characteristics is supposed to be due to some biologically natural grouping. There will G groups, and from each group will be sampled n_g sequences, $g = 1, 2, \ldots, G$. The sequences are divided in introns and exons. The i-th sequence from the g-th group for example has $k_{gi} + 1$ codons. In between the j-th and $[j + 1]$-th codon there is an intron of length d_{gij} say. The total length of introns on the i-th sequence of the g-th group is written as $N^\star_{gi}$. Let's consider $\mathscr{X}$ to be the set of codon configurations as seen in Table 1 and take $\|\mathscr{X}\|$ as its cardinality. Notice that $\|\mathscr{X}\|$ is not 64 because there are three stop codons. Let also $\mathbf{x}_{gij} = (x_{gij1}, x_{gij2}, x_{gij3})'$ be the categorical variable representing the configuration of the j-th codon on the i-th sequence of the g-th group. With no loss of generality, take $\mathbf{x}^\star = \mathbf{GGG}$ and $\mathscr{X}^\star = \mathscr{X} \setminus \{\mathbf{x}^\star\}$.

Since one is interested in modelling DNA sequences taking into account intron structure a natural approach (used here) is to consider the number of substitutions from the j-th to the $[j + 1]$-th codon on the i-th sequence of the g-th group (observable and known) which will be called $m_{\mathbf{x}_{gi[j+1]}}$ and assumes one out of the possible values $\{0, 1, 2, 3\}$. The average number of feasible substitutions from a codon of type $\mathbf{x}$ is $\mu_{\mathbf{x}}$ and the probability of the next codon be $\mathbf{y}$ given that its previous is $\mathbf{x}$ and the gap noncodant material in between them has length d on the g-th group is said to be $\pi^{(g)}_{\mathbf{xy}}(d)$. Finally, there is a mixing parameter for the $\pi_{\mathbf{xy}}$'s on the g-th group, say $\beta^{(g)}$.

A typical configuration of a sampled sequence would be as follows

$$\mathbf{x}_1 \ \underline{\ d_1\ } \ \mathbf{x}_2 \ \underline{\ d_2\ } \ \cdots \ \underline{\ d_k\ } \ \mathbf{x}_{k+1}, \tag{3.1}$$

where $\mathbf{x}_j = (x_{j1}, x_{j2}, x_{j3})' \in \mathscr{X}$, $\mathscr{X} = \{x_{[1]}, \ldots, x_{[64]}\}$.

$$\pi_{\mathbf{xy}}(d) = P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}, d) = \pi_{\mathbf{xy}}(0)e^{-\beta d} + \pi_{\mathbf{y}}\left(1 - e^{-\beta d}\right)$$

i.e., $\pi_{\mathbf{xx}}(d) \uparrow \pi_{\mathbf{x}}$ as $d \uparrow \infty$, since $\pi_{\mathbf{xx}}(d) = \pi_{\mathbf{x}} + (\pi_{\mathbf{xx}}(0) - \pi_{\mathbf{x}})\, e^{-\beta d}$ and $\pi_{\mathbf{xx}}(0)$ is very small, by empirical evidence, which means that the RHS is strict smaller than $\pi_{\mathbf{x}}$ and hence $\pi_{\mathbf{xx}}(d)$ converges to $\pi_{\mathbf{x}}$ as stated, when d gets large. On the other hand,

$$\pi_{\mathbf{xy}}(d) = \pi_{\mathbf{y}} + (\pi_{\mathbf{xy}}(0) - \pi_{\mathbf{y}})\, e^{-\beta d},$$

for which there are no empirical evidences favoring monotonicity in d, but it still easy to see that $\pi_{\mathbf{xy}}(d)$ converges to $\pi_{\mathbf{y}}$ when d gets large.

When codons are adjacent, i.e., the *intron* inbetween them has length zero the modelling of the transition probabilities will be given by

$$\pi_{\mathbf{xy}}(0) = \pi_{\mathbf{y}}\left(1 + \alpha\left(m_{\mathbf{y}} - \mu_{\mathbf{x}}\right)\right),$$

where $\mu_{\mathbf{x}}$ will vary on the $\|\mathscr{X}\|$ $\mathbf{x}$'s configurations but it will still be known.

The number of introns and total number of codons need a somewhat flexible modelling. Empirical evidences for the number of introns itself vary on the species and on the part of genome considered. For that reason we will be considering a generalization of the negative binomial as follows. For the sake of simplifying notation, we will take K and $N^\star$ to be respectively the number of introns and the total length of introns on a given sequence, and we will suppose K obeys the following law:

$$P(K=k)=\binom{r+k-2}{k-1}(1-\theta)^r\,\theta^{k-1},\ \text{for } k=1,2,\ldots,$$

where $0<\theta<1$ and $r>0$ are both unknown parameters and are allowed to vary among groups.

One can show that

$$E(K)=1+r\frac{\theta}{1-\theta}\ \text{and}\ Var(K)=r\frac{\theta}{(1-\theta)^2}.$$

That means that $SD(K)=(E(K)-1)/\sqrt{(r\theta)}$ and that convenient choices of r and θ result in either $SD(K)>E(K)$ or otherwise. That enables us to model a larger class of genomic sequences. Moreover, $N^\star=D_1+D_2+\cdots+D_K$ and we assume that $N^\star|K\ \sim G(\phi_K)$. So,

$$E(N^\star|K)=\frac{1}{1-\phi_K}\ ,\ Var(N^\star|K)=\frac{\phi_K}{(1-\phi_K)^2}\ \text{and}\ SD(N_\star|K)=\frac{\sqrt{\phi_K}}{1-\phi_K}.$$

Biological evidences suggest that the average lengths of introns and the average number of codons ($3K$) are both linear on the genome total size; therefore, $\phi_K=1-(\gamma K)^{-1}$. Assuming the lengths in between codons, D_i, are identically distributed (given $N^\star$ and K), $(D_1,D_2,\ldots,D_K|N^\star,K)'\sim Multinomial(N^\star,K^{-1})$ and

$$E(D_i|N^\star,K)=N^\star K^{-1}\qquad Var(D_i|N^\star,K)=(K-1)\,K^{-2}N^\star\qquad E(D_i|K)=\gamma$$

$$Var(D_i|K)=\gamma K^{-1}\left((\gamma+1)\,K-2\right)\qquad E(N^\star)=\gamma\left(\frac{r\theta}{1-\theta}+1\right)$$

$$Var(N^\star)=(E(N^\star))^2\left(1-\frac{1-\theta}{\gamma\left((r-1)\theta+1\right)}+\frac{2r\theta}{\left((r-1)\theta+1\right)^2}\right)$$

for which it is also true that convenient choices of r, γ and θ result in wider or narrower relative dispersions (for instance, for fixed θ and γ, choose r either bigger or smaller than $(-2\theta+1+\theta^2)(2\gamma-1+\theta)/\theta$ to get respectively standard deviation smaller or larger than the expected value of $N^\star$).

4 The Likelihood

We will consider G different groups and n_g (not necessarily balanced) sequences sampled on the g-th group, on a total of $n=\sum_{g=1}^G n_g$. Let $\mathbf{X}_{gi}$ be the i-th sequence of the g-th group: it is formed by the codons $\mathbf{x}_{gij}$, $j=1,2,\ldots,K_{gi}+1$, with intermediate noncodant materials $D_1,D_2,\ldots,D_{K_{gi}}$ on a configuration as in (3.1). We will write the likelihood function as follows:

$$L(\pi^{(g)},\alpha^{(g)},\beta^{(g)},\theta^{(g)},\gamma^{(g)}|\mathbf{X},\mathbf{K},\mathbf{N}^\star,\mathbf{D})=\prod_{g=1}^{G}\prod_{i=1}^{n_g}L(\mathbf{X}_{gi},K_{gi},N^\star_{gi},\mathbf{D}_{gi}).$$

Let's initially look at one such sequence $\mathbf{X}_{gi}$. There, using the hypotheses on the distributions of total length of introns, number of exons, substitution pattern, one can write L as:

$$
\begin{aligned}
&L(\mathbf{X}_{gi}, K_{gi}, N^\star_{gi}, \mathbf{D}_{gi}) \\
&\quad = L(\mathbf{X}_{gi}|N^\star_{gi}, \mathbf{D}_{gi}, K_{gi}) \times L(\mathbf{D}_{gi}|N^\star_{gi}, K_{gi}) \times L(N^\star_{gi}|K_{gi}) \times L(K_{gi}) \\
&\quad = \left(\prod_{j=1}^{K_{gi}+1} \pi^{(g)}_{\mathbf{x}_{gij}} \right) \prod_{j=1}^{K_{gi}} \left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}} \right) \frac{N^\star_{gi}! K_{gi}^{-N^\star_{gi}}}{\prod_{j=1}^{K_{gi}} D_{gij}!} \\
&\qquad \times \left(1 - \phi^{(g)}_{K_{gi}}\right) \phi^{(g)}_{K_{gi}}{}^{N^\star_{gi}-1} \binom{r^{(g)} + K_{gi} - 2}{K_{gi} - 1} \left(1 - \theta^{(g)}_{gi}\right)^{r^{(g)}} \theta^{(g)}{}^{K_{gi}-1}.
\end{aligned}
$$

Looking at the last RHS, one can factor $L(\mathbf{X}_{gi}, N^\star_{gi}, \mathbf{D}_{gi}, K_{gi})$ as

$$
\begin{aligned}
&L(\mathbf{X}_{gi}, N^\star_{gi}, \mathbf{D}_{gi}, K_{gi}) \propto \\
&\quad \mathbf{L}_1\left(\pi^{(g)}_{\mathbf{x}_{[1]}}, \ldots, \pi^{(g)}_{\mathbf{x}_{[\|\mathcal{X}^\star\|]}}\right) \times \mathbf{L}_2\left(\alpha^{(g)}, \beta^{(g)}\right) \times \mathbf{L}_3\left(\gamma^{(g)}\right) \times \mathbf{L}_4\left(\theta^{(g)}\right),
\end{aligned}
$$

where

$$
\mathbf{L}_1\left(\pi^{(g)}_{\mathbf{x}_{[1]}}, \ldots, \pi^{(g)}_{\mathbf{x}_{[\|\mathcal{X}^\star\|]}}\right) = \prod_{j=1}^{K_{gi}+1} \pi^{(g)}_{\mathbf{x}_{gij}}, \tag{4.2}
$$

$$
\mathbf{L}_2\left(\alpha^{(g)}, \beta^{(g)}\right) = \prod_{j=1}^{K_{gi}} \left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}} \right), \tag{4.3}
$$

$$
\mathbf{L}_3\left(\gamma_{(g)}\right) = \frac{\left(\gamma^{(g)} K_{gi} - 1\right)^{N^\star_{gi}-1}}{\left(\gamma^{(g)} K_{gi}\right)^{N^\star_{gi}}} \text{ and} \tag{4.4}
$$

$$
\mathbf{L}_4\left(\theta^{(g)}\right) = \binom{r^{(g)} + K_{gi} - 2}{K_{gi} - 1} \left(1 - \theta^{(g)}\right)^{r^{(g)}} \theta^{(g)}{}^{K_{gi}-1}. \tag{4.5}
$$

The decomposition of the log-likelihood ℓ can be written as:

$$
\begin{aligned}
&\ell(\mathbf{X}, \mathbf{K}, \mathbf{N}^\star, \mathbf{D}) \\
&\quad = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \ell(\mathbf{X}_{gi}, K_{gi}, N^\star_{gi}, \mathbf{D}_{gi}) = \ell_1(\mathbf{X}, \mathbf{K}, \mathbf{N}^\star, \mathbf{D}) \\
&\qquad + \ell_2(\mathbf{X}, \mathbf{K}, \mathbf{N}^\star, \mathbf{D}) + \ell_3(\mathbf{X}, \mathbf{K}, \mathbf{N}^\star, \mathbf{D}) + \ell_4(\mathbf{X}, \mathbf{K}, \mathbf{N}^\star, \mathbf{D}),
\end{aligned} \tag{4.6}
$$

where (being $f_{gi}(\mathbf{x})$ the frequency of observations on the category $\mathbf{x}$)

$$\ell_1(\mathbf{X}, \mathbf{K}, \mathbf{N}^\star, \mathbf{D}) = \sum_{\mathbf{x}\in\mathcal{X}} \sum_{g=1}^{G} \log\left(\pi_{\mathbf{x}}^{(g)}\right) \sum_{i=1}^{n_g} f_{gi}(\mathbf{x}) \tag{4.7}$$

$$\ell_2(\mathbf{X}, \mathbf{K}, \mathbf{N}^\star, \mathbf{D}) = \\ = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \log\left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}\right), \tag{4.8}$$

$$\ell_3(\mathbf{X}, \mathbf{K}, \mathbf{N}^\star, \mathbf{D}) = \\ = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(\left(N_{gi}^\star - 1\right) \log\left(\gamma^{(g)} K_{gi} - 1\right) - N_{gi}^\star \log\left(\gamma^{(g)} K_{gi}\right)\right), \tag{4.9}$$

$$\ell_4(\mathbf{X}, \mathbf{K}, \mathbf{N}^\star, \mathbf{D}) = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(r^{(g)} \log\left(1 - \theta^{(g)}\right) + \right. \\ \left. + \left(K_{gi} - 1\right) \log\left(\theta^{(g)}\right) + \log\left(\binom{r^{(g)} + K_{gi} - 2}{K_{gi} - 1}\right)\right). \tag{4.10}$$

Hence,

$$\begin{aligned}
\frac{\partial \ell}{\partial \alpha^{(g)}} &= \sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \frac{\left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_j}\right) e^{-\beta^{(g)} D_{gij}}}{1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}} \\
\frac{\partial \ell}{\partial \beta^{(g)}} &= -\alpha^{(g)} \sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \frac{D_{gij} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}}{1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}} \\
\frac{\partial^2 \ell}{\partial {\alpha^{(g)}}^2} &= -\sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \frac{\left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right)^2 e^{-2\beta^{(g)} D_{gij}}}{\left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}\right)^2} \\
\frac{\partial^2 \ell}{\partial \alpha^{(g)} \partial \beta^{(g)}} &= -\sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \frac{D_{gij} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}}{\left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}\right)^2} \\
\frac{\partial^2 \ell}{\partial {\beta^{(g)}}^2} &= \alpha^{(g)} \sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \frac{D_{gij}^2 \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}}{\left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}\right)^2}
\end{aligned}$$

$\hat{\alpha}^{(g)}$ and $\hat{\beta}^{(g)}$ are the numerical solution of the following equations:

$$\sum_{i=1}^{n_g} K_{gi} = \sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \left(1 + \hat{\alpha}^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\hat{\beta}^{(g)} D_{gij}}\right)^{-1} \tag{4.11}$$

$$\sum_{i=1}^{n_g} N_{gi}^\star = \sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} D_{gij} \left(1 + \hat{\alpha}^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\hat{\beta}^{(g)} D_{gij}}\right)^{-1} \tag{4.12}$$

for $g = 1, 2, \ldots, G$.

$$\begin{aligned}
\frac{\partial \ell}{\partial \gamma^{(g)}} &= \sum_{i=1}^{n_g} K_{gi} \left(\frac{N^\star_{gi} - 1}{\gamma^{(g)} K_{gi} - 1} - \frac{N^\star_{gi}}{\gamma^{(g)} K_{gi}} \right) \\
\frac{\partial^2 \ell}{\partial {\gamma^{(g)}}^2} &= -\sum_{i=1}^{n_g} K^2_{gi} \left(\frac{N^\star_{gi} - 1}{\left(\gamma^{(g)} K_{gi} - 1\right)^2} - \frac{N^\star_{gi}}{\left(\gamma^{(g)} K_{gi}\right)^2} \right)
\end{aligned}$$

$\hat{\gamma}^{(g)}$ is the numerical solution of the following equation:

$$n_g = \sum_{i=1}^{n_g} \frac{N^\star_{gi} - 1}{\hat{\gamma}^{(g)} K_{gi} - 1}, \quad g = 1, 2, \ldots, G. \tag{4.13}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta^{(g)}} &= -\frac{r^{(g)} n_g}{1 - \theta^{(g)}} + \frac{\sum_{i=1}^{n_g} K_{gi} - n_g}{\theta^{(g)}} \\
\frac{\partial^2 \ell}{\partial {\theta^{(g)}}^2} &= -\frac{r^{(g)} n_g}{\left(1 - \theta^{(g)}\right)^2} - \frac{\sum_{i=1}^{n_g} K_{gi} - n_g}{\left(\theta^{(g)}\right)^2} \\
\frac{\partial \ell}{\partial r^{(g)}} &= n_g \log\left(1 - \theta^{(g)}\right) + \sum_{i=1}^{n_g} \Psi\left(r^{(g)} + k_{gi} - 2\right) - n_g \Psi\left(r^{(g)}\right) \\
\frac{\partial^2 \ell}{\partial \theta^{(g)} \partial r^{(g)}} &= -\frac{n_g}{1 - \theta^{(g)}} \\
\frac{\partial^2 \ell}{\partial {r^{(g)}}^2} &= \sum_{i=1}^{n_g} \Psi^{'}\left(r^{(g)} + K_{gi} - 2\right) - n_g \Psi^{'}\left(r^{(g)}\right)
\end{aligned}$$

since

$$\frac{\partial}{\partial r^{(g)}} \log\left(\binom{r^{(g)} + K_{gi} - 2}{K_{gi} - 1}\right) = \sum_{j=0}^{K_{gi}-2} \frac{1}{r^{(g)} + j} = \Psi\left(r^{(g)} + K_{gi} - 2\right) - \Psi\left(r^{(g)}\right),$$

where $\Psi(\cdot)$ is the polygamma function.

$\hat{\theta}^{(g)}$ and $\hat{r}^{(g)}$ are given respectively by:

$$\hat{\theta}^{(g)} = 1 - \left(\frac{1}{\hat{r}^{(g)} n_g} \sum_{i=1}^{n_g} K_{gi} + \frac{\hat{r}^{(g)} - 1}{\hat{r}^{(g)}} \right)^{-1} \tag{4.14}$$

and by the numerical solution of

$$-n_g \log\left(\frac{1}{\hat{r}^{(g)} n_g} \sum_{i=1}^{n_g} K_{gi} + \frac{\hat{r}^{(g)} - 1}{\hat{r}^{(g)}} \right) + \sum_{i=1}^{n_g} \Psi\left(\hat{r}^{(g)} + k_{gi} - 2\right) - n_g \Psi\left(\hat{r}^{(g)}\right) = 0, \tag{4.15}$$

for $g = 1, 2, \ldots, G$.

$$\begin{aligned}
\frac{\partial \ell}{\partial \pi_{\mathbf{x}}^{(g)}} &= \sum_{i=1}^{n_g} \left(\frac{f_{gi}(\mathbf{x})}{\pi_{\mathbf{x}}^{(g)}} - \frac{K_{gi} + 1 - \sum_{y \in \mathcal{X}^\star} f_{gi}(\mathbf{y})}{1 - \sum_{\mathbf{y} \in \mathcal{X}^\star} \pi_{\mathbf{y}}^{(g)}} \right) \\
\frac{\partial^2 \ell}{\partial {\pi_{\mathbf{x}}^{(g)}}^2} &= -\sum_{i=1}^{n_g} \left(\frac{f_{gi}(\mathbf{x})}{{\pi_{\mathbf{x}}^{(g)}}^2} + \frac{K_{gi} + 1 - \sum_{\mathbf{y} \in \mathcal{X}^\star} f_{gi}(\mathbf{y})}{\left(1 - \sum_{\mathbf{y} \in \mathcal{X}^\star} \pi_{\mathbf{y}}^{(g)}\right)^2} \right) \\
\frac{\partial^2 \ell}{\partial \pi_{\mathbf{x}}^{(g)} \partial \pi_{\mathbf{z}}^{(g)}} &= -\sum_{i=1}^{n_g} \frac{K_{gi} + 1 - \sum_{\mathbf{y} \in \mathcal{X}^\star} f_{gi}(\mathbf{y})}{\left(1 - \sum_{\mathbf{y} \in \mathcal{X}^\star} \pi_{\mathbf{y}}^{(g)}\right)^2}
\end{aligned}$$

which is solved by

$$\hat{\pi}_{\mathbf{x}}^{(g)} = \frac{\sum_{i=1}^{n_g} f_{gi}(\mathbf{x})}{\sum_{i=1}^{n_g} K_{gi} + n_g}, \quad g = 1, 2, \ldots, G \text{ and } \mathbf{x} \in \mathcal{X}^\star. \tag{4.16}$$

All others unmentioned first and second order partial derivatives are zero.

5 Asymptotic Results

In this section we present the asymptotic properties for the estimators defined by (4.11), (4.12), (4.13), (4.14), (4.15) and (4.16). Joint asymptotic normality is shown to the whole vector of parameter estimators. Asymptotic variances and covariances (when applicable) are given for $\{\theta^{(g)}, r^{(g)}, \gamma^{(g)}, g = 1, 2, \ldots, G\}$. Due to its dependence structure only rough quotas would be available for the variance-covariance asymptotic structure of $\{\hat{\alpha}^{(g)}, \hat{\beta}^{(g)}, g = 1, 2, \ldots, G\}$. As our main concern resides in the $\{\pi^{(g)}, g = 1, 2, \ldots, G\}$ structure we only prove that the estimators defined by (4.11) and (4.12) are regular, i,e. that a joint asymptotic normality holds.

The information matrix for $\{\hat{\alpha}^{(g)}, \hat{\beta}^{(g)}, \hat{\gamma}^{(g)}, \hat{\theta}^{(g)}, \hat{r}^{(g)}, \hat{\pi}^{(g)}, g = 1, 2, \ldots, G\}$, given respectively by (4.11), (4.12), (4.13), (4.14), (4.15) and (4.16), can be written as:

$$\begin{bmatrix}
\mathbf{I}\left(\alpha, \beta\right)_{2G \times 2G} & 0_{2G \times G} & 0_{2G \times 2G} & 0_{2G \times 64G} \\
0_{G \times 2G} & \mathbf{I}\left(\gamma\right)_{G \times G} & 0_{G \times 2G} & 0_{G \times 64G} \\
0_{2G \times 2G} & 0_{2G \times G} & \mathbf{I}\left(\theta, r\right)_{2G \times 2G} & 0_{2G \times 64G} \\
0_{\|\mathcal{X}^\star\| G \times 2G} & 0_{\|\mathcal{X}^\star\| G \times G} & 0_{\|\mathcal{X}^\star\| G \times 2G} & \mathbf{I}\left(\pi\right)_{\|\mathcal{X}^\star\| G \times \|\mathcal{X}^\star\| G}
\end{bmatrix}. \tag{5.17}$$

$$\mathbf{I}\left(\theta, r\right) = diag\left(I(\theta^{(1)}, r^{(1)})_{2\times 2}, I(\theta^{(2)}, r^{(2)})_{2\times 2}, \ldots, I(\theta^{(G)}, r^{(G)})_{2\times 2} \right), \tag{5.18}$$

where

$$I(\theta^{(g)}, r^{(g)})_{1,1} = \frac{r^{(g)} n_g}{\theta^{(g)} \left(1 - \theta^{(g)}\right)^2} \qquad I(\theta^{(g)}, r^{(g)})_{1,2} = \frac{n_g}{1 - \theta^{(g)}}$$

$$I(\theta^{(g)}, r^{(g)})_{2,2} \geq \frac{n_g}{r^{(g)}} \frac{\theta^{(g)}}{1 - \theta^{(g)}}.$$

$$\mathbf{I}\left(\gamma\right) = diag\left(I(\gamma^{(1)}), I(\gamma^{(2)}, \ldots, I(\gamma^{(G)})\right), \tag{5.19}$$

where

$$I(\gamma^{(g)}) \geq \begin{cases} \frac{n_g}{(r-1)\theta{\gamma^{(g)}}^3} \left(\left(r-1\right)\theta\left(\gamma + \left(1-\theta\right)^r\right) + \left(1-\theta\right)\right) & r \neq 1 \\ \frac{n_g}{{\gamma^{(g)}}^2} \left(1 - \frac{1-\theta^{(g)}}{\gamma^{(g)}\theta^{(g)}} \log\left(1 - \theta^{(g)}\right)\right) & r = 1. \end{cases}$$

$$\mathbf{I}\left(\pi\right) = diag\left(\mathbf{I}\left(\pi^{(1)}\right), \mathbf{I}\left(\pi^{(2)}\right), \ldots, \mathbf{I}\left(\pi^{(G)}\right)\right). \tag{5.20}$$

Note that

$$\mathbf{I}\left(\pi^{(g)}\right) = \left[\mathbf{I}\left(\pi^{(g)}_{\mathbf{x}}, \pi^{(g)}_{\mathbf{z}})\right)\right]_{\mathbf{x},\mathbf{z}\in\mathcal{X}^\star},$$

where

$$\mathbf{I}\left(\pi^{(g)}_{\mathbf{x}}, \pi^{(g)}_{\mathbf{x}}\right) = n_g\left(2 + \frac{r^{(g)}\theta^{(g)}}{1-\theta^{(g))}}\right)\left(\frac{1}{\pi^{(g)}_{\mathbf{x}}} + \frac{1}{\pi^{(g)}_{\mathbf{x}^\star}}\right)$$

and

$$\mathbf{I}\left(\pi^{(g)}_{\mathbf{x}}, \pi^{(g)}_{\mathbf{z}}\right) = n_g\left(2 + \frac{r^{(g)}\theta^{(g)}}{1-\theta^{(g))}}\right)\frac{1}{\pi^{(g)}_{\mathbf{x}^\star}}.$$

So, up to now, because of its construction, the likelihood was factored in three parts, respectively ℓ_1, ℓ_3 and ℓ_4, which were solely related to respectively π's, γ's and θ's, whilst a fourth part, ℓ_2 is solely related to the α's and β's. We were able to directly apply the usual asymptotic techniques to ℓ_1, ℓ_3 and ℓ_4. As ℓ_2 has a correlated inner structure one must resort to a different argument. One can consider each of the partial derivatives of ℓ with respect to $\{\alpha^{(g)}, \beta^{(g)}, g = 1, 2, \ldots, G\}$ as a sum of n_g independent terms and work with their properties to enable a feasible Taylor expansion around the true parameter value. Moreover, since the $\alpha^{(g)}, \beta^{(g)}$ vectors are functionally orthogonal, we will be dealing with G two-dimensional such expansions. Each score function will have as its expected value, $E\left(\partial\ell/\partial\alpha^{(g)}\right)$

$$\sum_{i=1}^{n_g} E\left(\sum_{j=1}^{K_{gi}} E\left(\frac{\left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_j}\right)e^{-\beta^{(g)}D_{gij}}}{1 + \alpha^{(g)}\left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right)e^{-\beta^{(g)}D_{gij}}} \left| K_{g1}, \ldots, K_{gn_g}\right.\right)\right). \tag{5.21}$$

We then compute the inner expected value of (5.21) conditionally on the values of D_{gij} and, for each j, we compute it conditionally on the value of $\mathbf{x}_{gij}$, call it $E_\alpha\left(\cdot|\mathbf{K}, D_{gij}, \mathbf{x}_{gij}\right)$.

$$E_\alpha\left(\cdot|\mathbf{K}, D_{gij}, \mathbf{x}_{gij}\right) = \sum_{\mathbf{y}} \frac{\left(m_{\mathbf{y}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}}{1 + \alpha^{(g)} \left(m_{\mathbf{y}} \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}} \times \pi^{(g)}_{\mathbf{x}_{gij}\mathbf{y}}(D_{gij}) = 0.$$

On the same spirit for $\beta^{(g)}$, one has

$$E_\beta\left(\cdot|\mathbf{K}, D_{gij}, \mathbf{x}_{gij}\right) = -\alpha^{(g)} \sum_{\mathbf{y}} \frac{D_{gij}\left(m_{\mathbf{y}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}}{1 + \alpha^{(g)} \left(m_{\mathbf{y}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}} \times \pi^{(g)}_{\mathbf{x}_{gij}\mathbf{y}}(D_{gij}) = 0.$$

Moreover,

$$\begin{aligned}
\frac{\partial^3 \ell}{\partial {\alpha^{(g)}}^3} &= \sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \frac{\left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right)^3 e^{-3\beta^{(g)} D_{gij}}}{\left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}\right)^3} \\
\frac{\partial^3 \ell}{\partial {\alpha^{(g)}}^2 \partial \beta^{(g)}} &= 2\sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \frac{D_{gij}\left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right)^2 e^{-2\beta^{(g)} D_{gij}}}{\left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}\right)^3} \\
\frac{\partial^3 \ell}{\partial \alpha^{(g)} \partial {\beta^{(g)}}^2} &= \sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \frac{D^2_{gij}\left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}}{\left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}\right)^3} \\
&\quad \times \left(1 - \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}\right) \\
\frac{\partial^3 \ell}{\partial {\beta^{(g)}}^3} &= -\alpha^{(g)} \sum_{i=1}^{n_g} \sum_{j=1}^{K_{gi}} \frac{D^3_{gij}\left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}}{\left(1 + \alpha^{(g)} \left(m_{\mathbf{x}_{gi[j+1]}} - \mu_{\mathbf{x}_{gij}}\right) e^{-\beta^{(g)} D_{gij}}\right)^3}.
\end{aligned}$$

Basically, all such partial derivatives depend on terms concerning: $m_{\mathbf{y}} - \mu_{\mathbf{x}}$, which is stochastically bounded by 3, $D_{gij} e^{-a\beta D_{gij}}$, which is stochastically bounded since $D_{gij} \geq 0$, and $e^{-a\beta D_{gij}}$, which is stochastically bounded since $D_{gij} \geq 0$. Using that, each of the partial derivatives above will be bounded by variables as $M \sum_{i=1}^{n_g} K_{gi}$ which has a finite expectation, because n_g and M are real numbers and K_{gi}'s have finite expectations. One then argues on the Taylor expansions for $\sqrt{n_g}\left(\hat{\alpha}^{(g)} - \alpha^{(g)}\right)$ and $\sqrt{n_g}\left(\hat{\beta}^{(g)} - \beta^{(g)}\right)$ and the normal asymptotics will come out.

Let's take $n_0 = min\{n_1, n_2, \ldots, n_g\} \to \infty$. Let also $R_g = \lim_{n_0 \to \infty} n_0/n_g$ and suppose $R_g > 0$, $g = 1, 2, \ldots, G$. Then

$$n_0 \mathbf{Var}\left(\alpha, \beta, \gamma, \theta\, r, \pi\right) \to \left[\begin{array}{cccc}
\mathbf{V}\left(\alpha, \beta\right) & 0_{2G\times G} & 0_{2G\times 2G} & 0_{2G\times G\|\mathcal{X}\|} \\
0_{G\times 2G} & \mathbf{V}\left(\gamma\right) & 0_{G\times 2G} & 0_{G\times G\|\mathcal{X}\|} \\
0_{G\times 2G} & 0_{G\times G} & \mathbf{V}\left(\theta, r\right) & 0_{G\times G\|\mathcal{X}\|} \\
0_{G\|\mathcal{X}^\star\|\times 2G} & 0_{G\|\mathcal{X}^\star\|\times G} & 0_{G\|\mathcal{X}^\star\|\times 2G} & \mathbf{V}\left(\pi\right)
\end{array}\right],$$

where each of those variance matrices limits are found from the Fisher information matrices derived above and are all finite and bounded away from zero.

Therefore one has, for each group $\|\mathcal{X}^\star\|$ π's, one θ, one r, one α, one β and one γ, with a total of $G\left(\|\mathcal{X}^\star\|+5\right)$ parameters, for which:

- $\alpha^{(g)}$'s and $\beta^{(g)}$'s are estimated numerically, with a block diagonal $2G \times 2G$ Hessian with blocks of size 2×2
- $\gamma^{(g)}$'s are estimated numerically, with a diagonal $G \times G$ Hessian
- $\theta^{(g)}$'s and $r^{(g)}$'s are estimated numerically, with a block diagonal $2G \times 2G$ Hessian with blocks of size 2×2
- $\pi^{(g)}$'s are estimated with an explicit formulae.

So, the Fisher information matrix will be a block diagonal $G\left(\|\mathcal{X}^\star\|+5\right) \times G\left(\|\mathcal{X}^\star\|+5\right)$, for which there will be four major blocks ([α,β]; [γ]; [θ,r] and [π]). The number of non-zero elements will be roughly $4G+G+4G+\left(G\|\mathcal{X}^\star\|\right)^2$. Although large, those last $\left(G\|\mathcal{X}^\star\|\right)^2$ elements have a product multinomial structure.

6 Hypothesis Testing

One hypothesis in which there is a lot of interested is that of coding material equivalence among groups. In our model that can be expressed as

$$H_0 : \pi_{\mathbf{x}}^{(g)} = \pi_{\mathbf{x}}^{(1)} \quad \forall \mathbf{x} \in \mathcal{X} \quad g = 1, 2, \ldots, G.$$

We know, from (4.6)-(4.10) that the log-likelihood can be factored in four parts and, from those, only ℓ_1 is directly related to H_0. Therefore, the likelihood ratio statistic can be written as:

$$\Lambda_{n_0} = \ell_1\left(\hat{\pi}_{\mathbf{x}}^{(g)},\ \mathbf{x} \in \mathcal{X}\ ,\ g = 1, 2, \ldots, G\right) - \ell_1\left(\hat{\pi}_{\mathbf{x}}^{(0)},\ \mathbf{x} \in \mathcal{X}\right).$$

where

$$\ell_1\left(\hat{\pi}_{\mathbf{x}}^{(g)},\ \mathbf{x} \in \mathcal{X}\ ,\ g = 1, 2, \ldots, G\right) = \sum_{\mathbf{x}\in\mathcal{X}} \sum_{g=1}^{G} \log\left(\frac{\sum_{i=1}^{n_g} f_{gi}(\mathbf{x})}{\sum_{i=1}^{n_g} K_{gi} + n_g}\right) \sum_{i=1}^{n_g} f_{gi}(\mathbf{x})$$

Under H_0, the maximum likelihood estimators are

$$\hat{\pi}_{\mathbf{x}}^{(0)} = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n_g} f_{gi}(\mathbf{x})}{\sum_{g=1}^{G} \sum_{i=1}^{n_g} K_{gi} + n}, \quad \mathbf{x} \in \mathcal{X},$$

and the corresponding log-likelihood is given by

$$\ell_1\left(\hat{\pi}_{\mathbf{x}}^{(0)},\ \mathbf{x} \in \mathcal{X}\right) \quad = \quad \sum_{\mathbf{x}\in\mathcal{X}} \log\left(\frac{\sum_{g=1}^{G} \sum_{i=1}^{n_g} f_{gi}(\mathbf{x})}{\sum_{g=1}^{G} \sum_{i=1}^{n_g} K_{gi} + n}\right) \sum_{g=1}^{G} \sum_{i=1}^{n_g} f_{gi}(\mathbf{x}).$$

Finally,

$$\Lambda_{n_0} = \sum_{\mathbf{x}\in\mathcal{X}} \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{gi}(\mathbf{x})\right)\left[\log\left(\frac{\sum_{i=1}^{n_g} f_{gi}(\mathbf{x})/\sum_{i=1}^{n_g} K_{gi}+n_g}{\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{gi}(\mathbf{x})/\sum_{g=1}^{G}\sum_{i=1}^{n_g} K_{gi}+n}\right)\right] \tag{6.22}$$

will be such that $-2ln\left(\Lambda_{n_0}\right) \rightarrow Z^2$, where Z^2 has a chi-squared distribution with $(G-1)\|\mathcal{X}\|$ degrees of freedom.

7 References

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction* (ed. H. Solomon), Stanford Univ. Press, CA pp. 158-176.

Besag, J. (1974). Spatial interaction and the statistical analysis of life systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **48**, 192-236.

Chatterjee, S. K. and Sen, P. K. (1964). Nonparametric tests for the bivariate two-sample location problem. *Calcutta Statist. Assoc. Bull.* **13**, 18-58.

Deutsch, Michael and Long, Manyuan (1999). Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Research* **27**, 3219-3228.

Ewens, W. J. and Grant, G. R. (2001). *Statistical Methods in Bioinformatics: An Introduction*, Springer, New York, NY.

Hartl, D. L. (2000). *A Primer of Population Genetics*, Sinauer, Massachusetts.

Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*, Springer, New York, NY.

Li, Wen-Hsiung and Graur, Dan (1991). *Fundamentals of Molecular Evolution*, Sinauer Associates, Massachusetts.

Pinheiro, H., Seillier-Moiseiwitsch, Sen, P. K. and Eron, J. (2000). Genomic sequence analysis and quasi-multivariate CATANOVA. In *Handbook of Statistics, Volume 18 : Bioenvironmental and Public Health Statistics* (eds. P. K. Sen and C. R. Rao), Elsevier, Amsterdam, pp. 713-746.

Pinheiro, H., Seillier-Moiseiwitsch, and Sen, P. K. (2001). Analysis of variance for Hamming distances applied to unbalanced designs. *Research Report 30/01.* Instituto de Matemática, Estatística e Computação Científica. Universidade Estadual de Campinas.

Sen, P. K. (2001). *Excursions in Biostochastics: Biometry to Biostatistics to Bioinformatics*, Lect. Notes, Academia Sinica Inst. Statist. Sci., Taipei, ROC.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*, Chapman-Hall, UK.

Waterman, M. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman-Hall, UK.

Wright, F. (1990). The effective number of codons used in a gene. *Gene* **87**, 23-29.