Lund University

Faculty of Engineering LTH

Master's Thesis

Maximum likelihood estimation of a monotone probability mass function with unknown labels

Supervisor: Dragi Anevski

Examiner: Nader Tajvidi

Author: Pietro Greselin

June 2019

Dedicated to my family,
without whom all of this would not have been possible.

Acknowledgments

I would like to express my deep and sincere gratitude to my research supervisor Prof. Dragi Anevski for giving me the opportunity to work with him and for providing invaluable guidance throughout this work. To be able to work in close contact with him and the discussions we had on this thesis have been for me an inestimable chance for the progress of my knowledge of this field.
I am also grateful to Vladimir Pastukhov for his wise advices. The conversations we had concerning this work played a significant role in its writing.

I would also like to thank here those people who played a significant role in my academic career, in particular in the discover and growth of my passion for mathematics, specially Giancarlo Tondi, Elena Fusini, Luca Ferrari, Prof. Marco Vignati and Prof. Giacomo Aletti.

I am grateful to Lill and Ishael Siroiney for their support, encouragement and friendship throughout this year in Lund.

Finally I would like to thank my whole family. This work is dedicated to them.

Abstract

In this paper we discuss methods to estimate a monotone frequency of species in a population when the ordering of species is unknown.

The question arises from [5], where the case of known order was taken into account, and the method is inspired by [2].

We discuss first a regression approach that however only leads to the trivial solution, and then a likelihood approach that allows some generalisations and provides inspiration for future work. We discuss existence and uniqueness of the solution, and future work might be related to discussing consistency as well as asymptotic distribution results and if possible to look for an analytic expression of it.

Contents

Introduction

In this work we compare estimators for a finite support probability mass function with unknown labels.

The motivation for this study comes from neutron detection, in particular from [5]. There the problem is to estimate the energy or, equivalently, the wavelength distribution $\mathbf{q}$ as well as the probability of transmission $\mathbf{p}$ of a neutron beam pointed at a detector. The authors give an explicit solution for the maximum likelihood estimator $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ of $(\mathbf{p}, \mathbf{q})$ and derived strong consistency and asymptotic distribution of the estimator.

In this work we address the problem of finding a new estimator, based on the one found in [5], that satisfies a certain order restriction.

This can be done e.g. by writing a suitable cost function based on a previous estimator and then looking for the estimator that minimises the cost function. We have used this in a regression approach in Chapter 2.

Another way of addressing the problem is by assuming a certain distribution for a basic estimator and then formulating a maximum likelihood problem based on the basic estimator. This has been done in Chapter 3.

Chapter 1

Statement of the problem

The motivation for the problem treated in this paper originates in a large scale physics research facility, the European Spallation Source (ESS), currently being built in Lund, Sweden. In this particular research problem one is interested in estimating the energy or, equivalently, the wavelength distribution of a neutron beam. The neutron beam consists of a finite number s of distinct wavelength neutrons. The beam is pointed at a detector which consists of a finite number k of layers. At each layer an individual neutron can be absorbed, and therefore be detected, or it can be transmitted to the subsequent layer.

The neutron beam can be characterised as consisting of individual neutrons with different wavelengths

$$\boldsymbol{\mu} = \{\mu_\alpha\}_{\alpha \in \mathcal{A}},$$

and the neutrons are distributed in the neutron beam with corresponding frequencies

$$\mathbf{q} = \{q_\alpha\}_{\alpha \in \mathcal{A}},$$

where $\mathcal{A}$ is a set such that $|\mathcal{A}| = s$. The values μ_α are positive whereas the values q_α are positive and adding up to 1. The reason for using the indices $\alpha \in \mathcal{A}$ instead of $1, 2, \ldots, s$ is that we will assume that the wavelengths are, without loss of generality, sorted in decreasing order. We will assume in this paper that the parameters $\boldsymbol{\mu}$ and $\mathbf{q}$ are unknown as well as the ordering of values in the index set $\mathcal{A}$. This is an extension of the inference problem treated in [5], where the ordering was assumed to be known.

The model for the experiment can be described as follows. The number of neutrons that arrives at the detector in the time interval $[0, t]$ can be modelled as a counting process $X_0(t)$ with constant intensity λ. The process $X_0(t)$ can be seen as a sum of the individual counting processes $X_0^\alpha(t)$, for $\alpha = 1, 2, \ldots, s$, that count the number of neutrons of type α that arrive at the face of the detector in $[0, t]$.

The detection of neutrons occurs in the following way. When the incident beam of neutrons hits a particular layer of the detector, each neutron in the beam can possibly be absorbed, and then detected, or otherwise not be absorbed. If the neutron is not absorbed it will go through the present layer and will subsequently arrive at the next layer. The assumptions in this setting are that, at each layer, absorption or transmission are the only possibilities for the neutron interactions with the layer and, moreover, that at each layer different neutron particles interact with the layer independently of each other, i.e. at each layer the absorptions of different neutrons are independent events. Each neutron type has a different probability of transmission p_α and hence a probability of absorption $1 - p_\alpha$, for $\alpha = 1, 2, \ldots, s$, that depends on the wavelength.

At every layer i the number of neutrons that are detected in the time interval $[0, t]$ can be written as

$$X_i(t) = \sum_{\alpha \in \mathcal{A}} X_i^{\alpha}(t)$$

where $X_i^{\alpha}(t)$ is the number of detected neutrons of type α. The number of transmitted neutrons at layer i is $Y_i(t) = Y_{i-1}(t) - X_i(t)$ for $i = 1, 2, \ldots, k$ and with $Y_0(t) = X_0(t)$. Both $X_i(t)$ and $Y_i(t)$ are non-decreasing counting processes, obtained by the thinning of the original Poisson process $X_0(t)$, so that $X_i(t)$ and $Y_i(t)$ are independent Poisson processes with intensities

$$\sum_{\alpha \in \mathcal{A}} p_{\alpha}^{i-1}(1 - p_{\alpha}) q_{\alpha} \lambda \quad \text{and} \quad \sum_{\alpha \in \mathcal{A}} p_{\alpha}^{i} q_{\alpha} \lambda,$$

respectively, cf. [6], [8]. Moreover the processes $X_i(t)$ for $i = 1, 2, \ldots, k$ are jointly independent.

Therefore the vector $(X_0^1(t), X_0^2(t), \ldots, X_0^s(t))$ of the numbers of neutrons of each type reaching the detector in the time interval $[0, t]$ is assumed to follow a multinomial distribution with parameters $q_1, q_2, \ldots, q_s$, i.e.

$$\left(X_0^1(t) = x_0^1, X_0^2(t) = x_0^2, \ldots, X_0^s(t) = x_0^s \mid X_0 = x_0\right) \in \text{Multi}(x_0, q_1, q_2, \ldots q_s)$$

with

$$x_0^1 + x_0^2 + \cdots + x_0^s = x_0,$$

$$q_1 + q_2 + \cdots + q_s = 1.$$

Note that in this setting $q_r = \lambda_r / \lambda$ where λ_r is the intensity of the beam of neutrons of type r and q_r is assumed to be independent of t. This model was first introduced in [5]. However, in [5], the order of the wavelengths is

assumed to be known and hence the wavelengths are written as

$$\mu_1 < \mu_2 < \cdots < \mu_s.$$

Now, it is a physical property of the neutron beam that the probability of transmission decreases with the neutron wavelength, cf. [4] for the functional form of that dependence in the case of a monochromatic neutron beam. Therefore, the model assumption that the wavelengths are increasing implies that the thinning parameters can be modelled as a decreasing sequence

$$1 > p_1 > p_2 > \cdots > p_s > 0.$$

The experiment consists of a finite number n of trials and thus it is assumed that during a fixed time interval $[0, t]$ and for fixed intensity λ of an incident beam there are n repeated measurements. Let $x_{ij}(t)$ be the total number of neutrons that have been observed at layer i at the experiment round j and in the time interval $[0, t]$. Hence, $x_{ij}(t)$ are realisations of $X_{ij}(t)$, and are therefore Poisson distributed random variables with expectations

$$\sum_{\alpha \in \mathcal{A}} (1 - p_\alpha) p_\alpha^{i-1} q \lambda t,$$

for $i = 1, 2, \ldots, k$. Moreover, if we let $X_j = (X_{1j}.X_{2j}, \ldots, X_{kj})$ denote the experiment round j, the vectors $X_1, X_2, \ldots, X_n$ are assumed to be independent.

Finally, given the data, the goal is now to estimate the wavelength distribution $\mathbf{q}$, as well as the transmission probabilities $\mathbf{p}$, i.e. the wavelength values. Then, [5] introduced the maximum likelihood estimator of $(\mathbf{q}, \mathbf{p})$ as

$$(\hat{\mathbf{q}}_n, \hat{\mathbf{p}}_n) = \arg \max_{(\mathbf{q}, \mathbf{p}) \in \mathcal{F}} l_n(\mathbf{q}, \mathbf{p}),$$

where

$$l_n(\mathbf{q},\mathbf{p}) = \sum_{j=1}^{n}\sum_{i=1}^{k}\left(-\lambda t m_i + x_{i,j}\log(m_i) + x_{i,j}\log(\lambda t) - \log(x_{i,j}!)\right)$$

is the likelihood, with

$$m_i = \sum_{r=1}^{s}(1-p_r)p_r^{i-1}q_r$$

the total expected number of absorbed neutrons at layer i divided by the intensity λ and the time t and

$$\mathcal{F} = \left\{(\mathbf{q},\mathbf{p}) \in \mathbb{R}_+^{2s} : \sum_{i=1}^{s} q_i = 1,\ 1 > p_1 > p_2 > \cdots > p_s > 0\right\}$$

the parameter space.

The present work is an extension of [5] in the sense that we assume that the order of the wavelengths is unknown. In particular we assume that the estimated order in the maximum likelihood estimator $\hat{\mathbf{q}}_n$ is not necessarily correct. The dependence of the estimator $\hat{\mathbf{q}}_n$ on the amount of data n is here subsumed and will be dropped in the sequel. Another important assumption that we make in this work is that in the sample, which dimension is n, at least one occurrence of each species in the population has been measured. This in particular requires that $n > s$ and implies that the estimated frequency of every species is non-zero, i.e. $\hat{\mathbf{q}}_i \neq 0$ for all indices $i = 1, 2, \ldots, s$.

The goal in this paper is thus to estimate the frequency vector $\mathbf{q}$ under the assumption that it is a discrete probability density function but excluding the condition that we know the order.

A similar problem to the one we will treat here was stated in [2], in which the aim was to estimate the relative frequencies of the species in a population

when the total number of species is allowed to be infinite and their underlying order is assumed to be unknown.

The setting in [2] can be illustrated as follows. First, assume we are given an i.i.d. sample of dimension n drawn from a population of animals that belong to different species. Individuals from the population are sampled one at a time and counted according to their species.

Then, the sample is reduced to the count list $N = (N_1, N_2, \dots)$, where N_i is the number of observed individuals belonging to the i-th most frequent species in the sample. Clearly the number of observed species in the sample is finite. However, since in [2] the total number of species is allowed to be infinite, to the list N it is appended an infinite list of zeros.

Finally, since the data are ordered according to the observed numbers and not according to the real, population-frequency, order provided by nature, it is assumed that there exists a map from the species as ordered by the sample frequencies to the species as ordered by population probabilities, i.e.

$$\chi : \mathbb{N} \longrightarrow \mathcal{A}$$

defined as $\chi(i) = \alpha$ if and only if the i-th most frequent species in the sample is the α-th most frequent species in the population, with the tie-breaking rule

$$N_i = N_j \qquad \Longrightarrow \qquad \chi(i) < \chi(j).$$

The map χ is a permutation and an essential feature of it is that it is random and unobserved.

Denoting the unknown population frequency vector, indexed by the species labels $\mathcal{A}$ and therefore ordered, by $\boldsymbol{\theta}$, the maximum likelihood estimator was

defined as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\mathcal{F}} l(\boldsymbol{\theta}),$$

where the likelihood is

$$l(\boldsymbol{\theta}) = \sum_{\chi} \prod_{i} \theta_{\chi(i)}^{N_i}, \tag{1.1}$$

and the parameter space

$$\mathcal{F} = \left\{ \boldsymbol{\theta} \in \mathbb{R}_{+}^{\infty} : \theta_1 \geq \theta_2 \geq \ldots, \sum_i \theta_i = 1 \right\}.$$

Note that (1.1) is the sum over the space S of permutations of likelihoods for fixed orderings. The outer summing, over all permutations, is in order to allow for all possible ordering of the labels, since in [2], it is assumed that the correct order is unknown.

However, the solution to this problem in the infinite support case, i.e. $\mathcal{A} = +\infty$, does not always exist. Because of this, one instead considers the extended model maximum likelihood estimator which, in addition to the discrete probability part, also includes a continuum probability mass part. Here, the further assumption is that the discrete part of the distribution, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$, only satisfies $\sum_i \theta_i \leq 1$, allowing a remaining probability mass $\theta_0 = 1 - \sum_i \theta_i$ to be positive.

In this model, the deficit θ_0 equals the probability, when we observe just one individual, that it belongs to one of those species which individually each have zero probability. Each such species can only be observed at most once in a sample of dimension n. The converse is not true: if an individual is observed only once in our sample, we do not know whether it belongs to a zero probability species or to a positive probability species.

Then the extended MLE, or the pattern maximum likelihood estimator

(PML), is defined in [2] as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\mathcal{F}} l_{ext}(\boldsymbol{\theta}),$$

$$l_{ext}(\boldsymbol{\theta}) = \sum_{\chi} \frac{n!}{N_0! \prod_{i\geq 1} N_i!} \theta_0^{N_0} \prod_{\alpha=1}^{\infty} \theta_\alpha^{N_{\chi^{-1}(\alpha)}},$$

where

$$\mathcal{F} = \left\{ \boldsymbol{\theta} \in \mathbb{R}_+^\infty : \theta_1 \geq \theta_2 \geq \ldots, \sum_i \theta_i \leq 1 \right\}.$$

Here, the number N_0 of observations of species assigned to the continuous distribution is defined as

$$N_0 = n - \sum_{\alpha=1}^{\infty} N_{\chi^{-1}(\alpha)}.$$

Note however that the function χ used here is not the same as defined above due to the introduction of the continuum mass part. In particular, here the mappings $\chi : \mathbb{N} \to \mathcal{A}$ satisfy that for every $\alpha \geq 1$ there exists exactly one i such that $\chi(i) = \alpha$, with the tie-breaking rule

$$N_i = N_j \qquad \Longrightarrow \qquad \chi(i) < \chi(j),$$

and such that

$$\chi(i) = 0 \qquad \Longrightarrow \qquad N_i \in \{0, 1\}.$$

Note that since the data ends in a block of 1's, with N_0 of them belonging to continuum mass species, then

$$\sum_{i\geq 1} N_i = n + N_0.$$

To this problem, [2] provides an existence result and also an algorithm to estimate the value of $\hat{\boldsymbol{\theta}}$.

The aim for the present work is hence to try to use the work that has been done in [2] to give an estimator of the frequency vector for the setting provided in [5] and described earlier in the case that the underlying order of the frequencies is assumed to be unknown. It is worth stressing that, even if this works arises from [5], it can be thought as to be independent of it.

The rest of the paper is organised as follows.

In Chapter 2 we used a regression approach to improve the estimate provided by a given estimator $\hat{\mathbf{q}}$. First we considered the setting to be as in [5] but then, since it not suitable for the purposes of this work, we detached from it. For this reason we had to make assumptions on the distribution of the components of the estimator $\hat{\mathbf{q}}$ in order to formulate an appropriate regression problem. Finally we analysed the choice of the weights.

In Chapter 3 we moved from a regression approach to a maximum likelihood approach. Again one has to assume a certain distribution for the components of the estimator $\hat{\mathbf{q}}$ in order to write the likelihood function and here we focused in particular on the multivariate Gaussian case.

Chapter 2

The regression problem

The aim of this chapter is to present regression methods for improving the estimation of an ordered frequency vector for species in a population when their underlying order is unknown.

Let $\mathbf{q}$ be the unknown frequency vector of species. Here it is assumed that the vector $\mathbf{q}$ is sorted with respect to an unknown index order. Further assume we are given an estimator $\hat{\mathbf{q}}$ of the frequencies, in some order. The vector $\hat{\mathbf{q}}$ is however not necessarily in the correct order, provided by nature. The aim is to device a method to improve on the estimator, possibly improving on the estimation of the order. We will have a least squares regression approach to find a new estimator $\tilde{\mathbf{q}}$ of $\mathbf{q}$, based on the starting estimator $\hat{\mathbf{q}}$.

2.1 Regression of the maximum likelihood estimator

In this section we will use the estimator of $\mathbf{q}$ provided by [5] as our preliminary estimator $\hat{\mathbf{q}}$.

The estimator $\hat{\mathbf{q}} = \hat{\mathbf{q}}_n$ is obtained as the first component in the solution of a likelihood problem stated in [5], namely

$$(\hat{\mathbf{q}}_n, \hat{\mathbf{p}}_n) = \arg \max_{(\mathbf{q},\mathbf{p})\in\mathcal{F}} l_n(\mathbf{q}, \mathbf{p}),$$

where the log-likelihood

$$l_n(\mathbf{q}, \mathbf{p}) = \sum_{j=1}^{n}\sum_{i=1}^{k}(-\lambda t m_i + x_{i,j}\log(m_i) + x_{i,j}\log(\lambda t) - \log(x_{i,j}!))$$

only depends on the data as the total amount of neutrons $x_{i,j}$ that are detected at the i-th layer at experiment round j. Therefore the log-likelihood is independent of the neutrons' kinds, and therefore also independent of the order.

Hence an attempt here may be to fix the order according to the estimated values of $\hat{\mathbf{q}}$, and then to do isotonic regression of $\hat{\mathbf{q}}$ with respect to that order. This means that one estimates the order first and then estimates the value of $\hat{\mathbf{q}}$.

However, since this new sorted estimator $\hat{\mathbf{q}}$ is already ordered, there is no need to further do regression. In fact a similar approach in the context of estimation of a probability mass function was used in [7], and then called the monotone rearrangement, cf. also [1] for the corresponding probability density function estimation problem.

2.2 The multinomial case

To introduce an alternative to the setting given in [5], one can instead formulate the problem along the lines of [2].

Hence, the setting to the regression problem can be described as follows. Assume that a certain population, of neutrons or of animals, is composed of a finite number s of different species and denote by $\mathbf{q}$ the vector of relative frequencies of the different species. Further, assume that the relative frequencies can be sorted in decreasing order as

$$q_1 \geq q_2 \geq \cdots \geq q_s$$

where q_i denotes the relative frequency of the i-th most frequent species in nature.

Assume now that an i.i.d. sample of n individuals has been drawn from the population. Individuals in the sample can be divided according to their species and therefore the sample can be written as a vector $(X_1, X_2, \ldots, X_s)$, where X_i denotes the number of times that an individual belonging to the i-th most frequent species in nature has been observed. This implies that the data $(X_1, X_2, \ldots, X_s)$ can be written as following a multinomial distribution, i.e.

$$(x_1, x_2, \ldots, x_s) \in \text{Multi}(n, q_1, q_2, \ldots, q_s).$$

Now, since the underlying order of the species is assumed to be unknown, the actual data that one collects is unordered with respect to the order provided by nature, and can without loss of generality be sorted in any order e.g. in decreasing order according to the observation, and hence they can be written as

$$(n_1, n_2, \ldots, n_s) = \text{sort}(x_1, x_2, \ldots, x_s),$$

where the function sort($\mathbf{v}$) sorts a vector $\mathbf{v}$ in decreasing order. Thus n_i denotes the number of times that an individual belonging to the i-th most frequently observed species has been detected.

A naive estimator of $\mathbf{q}$ is provided by

$$\hat{\mathbf{q}} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_s}{n} \right).$$

To improve on the naive estimator $\hat{\mathbf{q}}$, we introduce a regression approach. We define the estimator

$$\tilde{\mathbf{q}} = \arg\min_{\mathbf{q} \in \mathcal{F}} r(\mathbf{q}), \tag{2.1}$$

where the criterion function is

$$r(\mathbf{q}) = \sum_{\chi \in S_s} \sum_{i=1}^{s} \frac{1}{n q_{\chi(i)} \left(1 - q_{\chi(i)}\right)} \left(n_i - n q_{\chi(i)}\right)^2, \tag{2.2}$$

and the parameter space

$$\mathcal{F} = \left\{ \mathbf{q} \in \mathbb{R}_+^s : q_1 \geq q_2 \geq \cdots \geq q_s, \ \sum_{i=1}^{s} q_i = 1 \right\}.$$

In (2.2), S_s denotes the symmetric group of order s and hence the functions χ are permutations of the indices. The criterion function $r(\mathbf{q})$ in (2.2) is a sum over all possible reshuffling of the indices of wighted sums of squares of differences between observed and expected values, with weights provided by the variances.

In this setting one can prove an existence and uniqueness result for the solution of (2.1). The proofs follow by considering the gradient of the criterion function (2.2). In order to simplify the derivations we rewrite the criterion function. Note that the permutations χ are bijective functions of $\{1, 2, \ldots, s\}$

onto $\mathcal{A}$ so that we can write

$$r(\mathbf{q}) = \sum_{i=1}^{s} \frac{1}{nq_i\,(1-q_i)} \left(\sum_{\chi \in S_s} \left(n_{\chi^{-1}(i)} - nq_i \right)^2 \right). \tag{2.3}$$

Now since for each i, χ runs through all permutations in S_s in the sum

$$\sum_{\chi \in S_s} \left(n_{\chi^{-1}(i)} - nq_i \right)^2,$$

then each $j \in \{1, 2, \ldots, s\}$ is exhibited in the index $n_{\chi^{-1}(i)}$, i.e. $n_{\chi^{-1}(i)} = n_j$ a number $s!/s = (s-1)!$ of times, because of symmetry. Thus

$$r(\mathbf{q}) = (s-1)! \sum_{i=1}^{s} \frac{1}{nq_i\,(1-q_i)} \left(\sum_{j=1}^{s} (n_j - nq_i)^2 \right). \tag{2.4}$$

The inner sum in (2.4) can be simplified as

$$\begin{aligned}
\sum_{j=1}^{s} (n_j - nq_i)^2 &= \sum_{j=1}^{s} n_j^2 - 2nq_i \sum_{j=1}^{s} n_j + sn^2 q_i^2 \\
&= \sum_{j=1}^{s} n_j^2 - 2n^2 q_i + sn^2 q_i^2 \\
&= n^2 \left(\frac{1}{n^2} \sum_{j=1}^{s} n_j^2 - 2q_i + sq_i^2 \right) \\
&=: n^2 \left(c - 2q_i + sq_i^2 \right).
\end{aligned}$$

Note that the constant c depends on the data $(n_1, n_2, \ldots, n_s)$ and therefore is random. However for fixed setting of the experiment, i.e. for fixed sample size n and number of species s, c assumes values only in the interval $[1/s, 1]$. This holds since the two extremal possibilities for the data are either $(n/s, n/s, \ldots, n/s)$, leading to $c = 1/s$, or $(n, 0, \ldots, 0)$, leading to $c = 1$.

Therefore (2.4) can be written as

$$r(\mathbf{q}) = n(s-1)! \sum_{i=1}^{s} \frac{sq_i^2 - 2q_i + c}{q_i\,(1-q_i)}, \tag{2.5}$$

and since constant factors do not affect the location of extreme value one can instead consider

$$\tilde{r}(\mathbf{q}) = \sum_{i=1}^{s} \frac{sq_i^2 - 2q_i + c}{q_i\,(1-q_i)}. \tag{2.6}$$

Proposition 1 (Existence and uniqueness of the solution). *The solution to (2.1) exists and is unique for every choice of the number of species s and sample size n.*

Proof. In order to prove existence and uniqueness of the solution it is sufficient to show that the criterion function (2.6) admits a unique minimum in the interval $(0,1)^s$.

Consider first the case of $s = 2$ when

$$\tilde{r}(\mathbf{q}) = \sum_{i=1}^{2} \frac{2q_i^2 - 2q_i + c}{q_i\,(1-q_i)}.$$

It can be easily checked that the value $q_i = 1/2$ is the only minimum of the function $2q_i^2 - 2q_i + c$ and it is the only maximum of $q_i\,(1-q_i)$. This in particular implies that the point

$$\left(\frac{1}{2}, \frac{1}{2}\right)$$

is the only minimum of the criterion function in the interval $(0,1)^2$.

In the general case $s > 2$, note first that every function

$$\frac{sq_i^2 - 2q_i + c}{q_i\,(1-q_i)}$$

is such that

$$\lim_{q_i \to 0^+} \frac{sq_i^2 - 2q_i + c}{q_i\,(1 - q_i)} = \lim_{q_i \to 1^-} \frac{sq_i^2 - 2q_i + c}{q_i\,(1 - q_i)} = +\infty.$$

Then it is sufficient to show that every such function admits one and only one extreme value in the interval $(0, 1)$.

Thus consider the directional derivative of the regression function in direction q_i

$$\begin{aligned} \frac{\partial r}{\partial q_i} &= \frac{\partial}{\partial q_i} \frac{c - 2q_i + sq_i^2}{q_i\,(1 - q_i)} \\ &= \frac{(s-2)q_i^2 + 2cq_i - c}{q_i^2\,(1 - q_i)^2}, \end{aligned}$$

which vanishes at

$$q_i = \frac{-c \pm \sqrt{c(s-1)}}{s-2}. \tag{2.7}$$

The solution corresponding to the choice of $-$ in (2.7) is not permissible since it is negative. However the solution corresponding to the choice of $+$ is a value in $(0, 1)$. In fact we will establish this by showing that the solution is positive and that is strictly less than 1 separately.

To show that the solution is in fact positive we note that

$$-c + \sqrt{c(s-1)} > 0 \quad \Longleftrightarrow \quad \sqrt{c(s-1)} > c.$$

Since the expressions on both sides of the last inequality are positive, one can square the expression and then obtains

$$c^2 - c(s-1) < 0. \tag{2.8}$$

The function $c^2 - c(s-1)$ is a convex parabola that vanishes at $c = 0$ and $c = s - 1$ implying that the inequality (2.8) holds for every c and hence that

(2.7) is always positive.

To show that it is strictly less than 1, note that

$$\frac{-c+\sqrt{c(s-1)}}{s-2}<1 \quad \iff \quad \sqrt{c(s-1)}<c+s-2.$$

Again the expressions on both sides of the last inequality are positive, thus one can square the expression obtaining

$$c^2+(s-3)c+s^2+4-4s>0. \tag{2.9}$$

The function $c^2+(s-3)c+s^2+4-4s$ is a convex parabola and it can be easily seen that it is strictly positive implying that the inequality (2.9) holds for every c and hence that (2.7) is always strictly less than 1.

Therefore the function (2.6) has unique minimum in the interval $(0,1)^s$ and the solution to (2.1) is the projection of it to the space $\mathcal{F}$. □

However, since the expression in (2.7) does not depend on i, the estimator $\tilde{\mathbf{q}}$, defined in (2.1), is the trivial on, i.e.

$$\tilde{\mathbf{q}}=\left(\frac{1}{s},\frac{1}{s},\ldots,\frac{1}{s}\right)$$

independently of the sample size n.

2.3 Other weighted regressions

Suppose we are given an estimator for the frequency of the species of a certain population, based on a sample from that population. Assume that the number of species s is at most finite and, further, assume that it is known. Assume furthermore that the sample dimension n is large enough for all the

species to have been detected at least once. Denote the given estimator as

$$\hat{\mathbf{q}} = (\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_s),$$

where the dependence on the sample size n is here suppressed in the notation. Let

$$\mathbf{q} = \{q_\alpha\}_{\alpha \in \mathcal{A}}$$

denote the unknown probaility mass function for the different species. The estimator $\hat{\mathbf{q}}$ is sorted in decreasing order with respect to the observations, meaning that $\hat{q}_i$ represents the estimated frequency of the i-th most frequent specie in the sample, while $\mathbf{q}$ is sorted in decreasing order with respect to the unknown ordering of the species. The two ordering might not be the same and hence there exists an unknown map

$$\chi : \quad \{1, 2, \ldots, s\} \quad \longrightarrow \quad \mathcal{A}$$

defined as $\chi(i) = \alpha$ if and only if the i-th most frequent species in the sample is the α-th most frequent species in the population, with the tie-breaking rule that if the estimated frequency of 2 species $i,\ j$ is the same, i.e. $\hat{q}_i = \hat{q}_j$, then $\chi(i) < \chi(j)$. The map χ is assumed to be random and not observed.

In this section our aim is to find suitable weights for a regression problem similar to (2.1). The weighted regression problem can be stated as

$$\tilde{\mathbf{q}} = \arg\min_{\mathbf{q} \in \mathcal{F}} r(\hat{\mathbf{q}}), \tag{2.10}$$

with

$$r(\mathbf{q}) = \sum_{\chi \in S_s} \sum_{i=1}^{s} \omega(i, \chi) \left(\hat{q}_i - q_{\chi(i)}\right)^2, \tag{2.11}$$

and where

$$\mathcal{F} = \left\{ \mathbf{q} \in \mathbb{R}_+^s : q_1 \geq q_2 \geq \cdots \geq q_s, \ \sum_{i=1}^{s} q_i = 1 \right\}.$$

Note that in this setting we allow weights to be depend on both the empirical ordering and the permutations of such order.

We suggest the following weigths

$$\omega(i, \chi) = \frac{\sigma_i(\chi(i))}{v_{\chi(i)}}$$

where $v_{\chi(i)}$ is the empirical variance of the $\chi(i)$-th species and $\sigma_i(\chi(i))$ is a distance-based weight, to be specified below. The empirical variance depends on the underlying probability model that could be e.g. a Poisson distribution or a multinomial distribution. On the other hand, we introduce the weight $\sigma_i(\chi(i))$ since we would like to give more importance, while estimating the real value of the frequency of the $\chi(i)$-th specie, to the species that appear close to it in the empirical ordering. For this reason we have been looking for weights σ_i having a bell shape with the peak over the $\chi(i)$-th specie and decreasing with the distance from it. Moreover, even though it is not stressed here, the weights depend on the sample size n and in particular we picked σ_i such that

$$\sigma_i \longrightarrow \delta_i,$$

as $n \to +\infty$, where δ_i is the Kronecker delta .

We considered 2 different scenarios depending on the weights σ_i.

The first one comes from setting σ_i to be a binomial weight, i.e. to be the probability mass functions of a binomial distribution $B(s, i)$ evaluated at

the point $\chi(i)$.

The second is to set σ_i to be

$$\sigma_i(\chi(i)) = \begin{cases} \frac{\log_s(n)}{c(\log_s(n)+s)} & \text{for } i = \chi(i) \\ \frac{1}{c(|i-\chi(i)|\log_s(n)+s)} & \text{for } i \neq \chi(i) \end{cases}$$

where c is a normalising constant.

In both the setting we have introduced in this section, one can prove an existence and uniqueness result for the solution to (2.10) with a method analogous to the one used to prove the result in previous section.

However, since neither of the two choices lead to a reshuffling of the order provided by $\hat{\mathbf{q}}$, we moved to a different problem and did not consider consistency for this estimator.

Chapter 3

The likelihood problem

The inference problem defined above will next be stated as a maximum likelihood problem. We would here like to emphasise that our suggestion can be seen as a novel approach to estimating a monotone probability mass function, when one has no knowledge about the labels.

Let $\mathbf{q}$ be the unknown frequency vector of species. It is assumed that the vector $\mathbf{q}$ is sorted with respect to an unknown index order. Again assume we are given an estimator $\hat{\mathbf{q}}$ of the frequencies which is not necessarily ordered in the correct species order. The aim is to improve on this estimator, and possibly to correct the order. For this reason we will make assumptions on the distributions of the estimator's components, in order to state an appropriate likelihood function for the estimator. Finally the new estimator $\tilde{\mathbf{q}}$ is derived as the maximiser of the likelihood function over the parameter space $\mathcal{F}$.

Improvements to this method can be done if an asymptotic distribution result is provided for the estimator $\hat{\mathbf{q}}$.

3.1 The multinomial case

The setting we consider here is the same as the one described in Section 2.2. In particular we assume that the partially unobserved data is modelled by the multinomial distribution, i.e.

$$(x_1, x_2, \ldots, x_s) \in \text{Multi}(n, q_1, q_2, \ldots, q_s).$$

Since the underlying order is unknown, the actual observed data is

$$(n_1, n_2, \ldots, n_s) = \text{sort}(x_1, x_2, \ldots, x_s),$$

thus n_i denotes the number of times that an individual belonging to the i-th most frequently observed species has been detected.

One, naive, estimator of $\mathbf{q}$ consists of the relative frequencies of the observed species, sorted by their observed occurencies, i.e.

$$\hat{\mathbf{q}} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_s}{n}\right).$$

This is also the maximum likelihood estimator of $\mathbf{q}$, under the assumption that the observed species order is the correct one.

We introduce a maximum likelihood estimator as

$$\tilde{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathcal{F}} L(\mathbf{q}), \tag{3.1}$$

where

$$\begin{aligned} L(\mathbf{q}) &= \sum_{\chi\in S_s} l(\mathbf{q}, \chi) \\ &= \sum_{\chi\in S_s} \prod_{i=1}^{s} q_{\chi(i)}^{n_i}, \end{aligned} \tag{3.2}$$

is a sum over all permutations of the likelihoods, and where the parameter space is

$$\mathcal{F} = \left\{ \mathbf{q} \in \mathbb{R}_+^s : q_1 \geq q_2 \geq \cdots \geq q_s,\ \sum_{i=1}^{s} q_i = 1 \right\}.$$

Note that in (3.2) we could actually write the full likelihood by introducing the multinomial factors

$$\binom{n}{n_1 \dots n_s},$$

which, since they are the same for all terms in the sum, can be factored out.

One can immediately note that the function $L(\mathbf{q})$, defined in (3.2), is continuous over the closed parameter space $\mathcal{F}$ implying that the estimator $\tilde{\mathbf{q}}$, defined in (3.1), exists. This proves the following result.

Proposition 2 (Existence of the solution). *In the setting described in this section, the maximum likelihood estimator, defined in (3.1), exists.*

Now in order to discuss the uniqueness of the estimator $\tilde{\mathbf{q}}$, we make two claims that we will prove afterwards. First we claim that the likelihood function $L(\mathbf{q})$ is invariant under permutations of the axis and then we also claim that every likelihood function $l(\mathbf{q}, \chi)$ admits a unique maximum.

To prove the symmetry of the function $L(\mathbf{q})$ consider first the function $l(\mathbf{q}, \chi)$ for a fixed but arbitrary permutation χ. It is easy to see that

$$l(\mathbf{q}, \chi) = l(\mathbf{q}_{\chi^{-1}}, id), \tag{3.3}$$

where

$$\mathbf{q}_{\chi^{-1}} = \left(q_{\chi^{-1}(1)}, q_{\chi^{-1}(2)}, \dots, q_{\chi^{-1}(s)} \right),$$

and id denotes the identity element in S_s. This can equivalently be stated as, if χ_1 and χ_2 are two different permutations, then

$$l(\mathbf{q}, \chi_1) = l(\mathbf{q}_{\chi_2 \circ \chi_1^{-1}}, \chi_2).$$

Now using (3.3) we see that the we can rewrite $L(\mathbf{q})$ as

$$\begin{aligned} L(\mathbf{q}) &= \sum_{\chi \in S_s} l(\mathbf{q}_{\chi^{-1}}, id) \\ &= \sum_{\chi \in S_s} l(\mathbf{q}_{\chi}, id), \end{aligned} \tag{3.4}$$

where the last equality holds since χ^{-1} is a permutation if and only if χ is. The expression (3.4) implies that $L(\mathbf{q})$ is invariant under permutation of the axis, i.e. that

$$L(\mathbf{q}) = L(\mathbf{q}_{\chi})$$

for any permutation $\chi \in S_s$.

To instead prove the second claim, without loss of generality we may consider the function $l(\mathbf{q}, id)$, since for every other χ, the likelihood $l\mathbf{q}, \chi)$ is obtained from $l(\mathbf{q}, id)$ by an appropriate permutation of the axis. Then one can make use of the condition

$$\sum_{i=1}^{s} q_i = 1$$

to replace one of the q_i, say q_s, in the formulation of $l(\mathbf{q}, id)$ thus obtaining

$$l(\mathbf{q}, id) = \underbrace{\prod_{i=1}^{s-1} q_i^{n_i}}_{f(\mathbf{q})} \underbrace{\left(1 - \sum_{i=1}^{s-1} q_i\right)^{n_s}}_{g(\mathbf{q})}. \tag{3.5}$$

where the functions

$$f, g : \mathcal{G} \longrightarrow \mathbb{R}_+$$

are defined on the set

$$\mathcal{G} = \left\{ \mathbf{q} \in \mathbb{R}_+^{s-1} : q_1 \geq q_2 \geq \cdots \geq q_{s-1},\ \sum_{i=1}^{s-1} q_i \leq 1 \right\}.$$

The functions f and g are non-negative on the set $\mathcal{G}$. Furthermore it is easy to prove that, with respect to the following partial order defined on $\mathcal{G}$, sometimes called the matrix order,

$$\mathbf{x},\ \mathbf{y} \in \mathcal{G}, \quad \mathbf{x} \leq \mathbf{y} \quad \text{iff} \quad x_i \leq y_i\ \ \forall i,$$

the function f is increasing, whereas the function g is decreasing. Note also that the function f vanishes on the set $\mathcal{Z}_f$ defined as

$$\mathcal{Z}_f = \{ \mathbf{q} \in \mathcal{G} :\ \exists j \text{ s.t. } q_j = 0 \},$$

whereas the function g vanishes on the set $\mathcal{Z}_g$ defined as

$$\mathcal{Z}_g = \left\{ \mathbf{q} \in \mathcal{G} :\ \sum_{i=1}^{s-1} q_i = 1 \right\}.$$

Since

$$\partial\mathcal{G} = \mathcal{Z}_f \cup \mathcal{Z}_g$$

then the function $l(\mathbf{q}, id)$ defined in (3.5) vanishes at $\mathcal{G}$'s boundary and since it is non-negative then it has at least one maximum in $\mathcal{G}$.

To complete the proof of the second claim then one has to prove that there is exactly one maximum. In order to do so observe that the directional

derivative of the function $l(\mathbf{q}, id)$ in the direction j is

$$\begin{aligned}\frac{\partial}{\partial q_j} l(\mathbf{q}, id) &= \frac{\partial}{\partial q_j} \prod_{i=1}^{s-1} q_i^{n_i} \underbrace{\left(1 - \sum_{i=1}^{s-1} q_i\right)}_{h(\mathbf{q})}^{n_s} \\ &= \prod_{\substack{i=1 \\ i \neq j}}^{s-1} q_i^{n_i} \cdot n_j q_j^{n_j - 1} \cdot h(\mathbf{q})^{n_s} - f(\mathbf{q}) \cdot n_s \cdot h(\mathbf{q})^{n_s - 1} \\ &= \underbrace{\prod_{\substack{i=1 \\ i \neq j}}^{s-1} q_i^{n_i} \cdot q_j^{n_j - 1} \cdot h(\mathbf{q})^{n_s - 1}}_{>0} \cdot \left(n_j h(\mathbf{q}) - n_s q_j\right),\end{aligned}$$

where $f(\mathbf{q})$ is defined in (3.5).

Therefore the directional derivative vanishes if and only if

$$n_j h(\mathbf{q}) - n_s q_j = 0$$

allowing for the condition $\nabla l(\mathbf{q}, id) = 0$ to be written as

$$\begin{cases} n_1 h(\mathbf{q}) - n_s q_1 = 0 \\ n_2 h(\mathbf{q}) - n_s q_2 = 0 \\ \quad \vdots \\ n_{s-1} h(\mathbf{q}) - n_s q_{s-1} = 0 \end{cases} \tag{3.6}$$

Now since $h(\mathbf{q}) = q_s$, one can rewrite the system of equation (3.6) as

$$\begin{cases} q_1 = \frac{n_1}{n_s} q_s \\ q_2 = \frac{n_2}{n_s} q_s \\ \quad \vdots \\ q_{s-1} = \frac{n_{s-1}}{n_s} q_s \end{cases}$$

and, recalling that $\sum_{i=1}^{s} \_i = 1$ has to hold, it is easy to see that the maximum is attained at the point

$$\mathbf{q} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_s}{n} \right).$$

Motivated by a simulation study, defined below, we would like to claim that, independently on $\hat{\mathbf{q}}$, the function $L(\mathbf{q})$ has exactly one peak that satisfies the order restriction provided by $\mathcal{F}$. Therefore we propose the following conjecture.

Conjecture 1 (Uniqueness of the solution). *In the setting described in this section, the maximum likelihood estimator, defined in (3.1), exists and is unique.*

The motivation for the last claim about the uniqueness, that we have not been able to prove, comes from some simulations we did in dimension $s = 3$, one of which is illustrated in Figure 3.1 and Figure 3.2. We have chosen this dimension since it is the biggest in which one can still plot the likelihood function. However the solution to (3.1) can be estimated also in bigger dimension and analogous results can be shown as well.

In the following s is assumed to be 3 and $\tilde{\mathbf{q}}$ is estimated 4 times on samples of dimension 15, 30, 100, 150 respectively. The choices for the sample sizes is to show the different behaviours of the likelihood function depending on the relation between the peaks location and amplitude.

Figure 3.1 shows the level sets of the likelihood function $L(\mathbf{q})$ together with the location of the permutations of $\mathbf{q}$ whereas Figure 3.2 shows the likelihood function. Note that in Figure 3.1 a circle has been used to mark

the real solution while its permutations have been marked with a dot. The same notation will be used in next figures. We have chosen to plot the results in barycentric coordinates in order to improve the visualisation of the domain of the likelihood function that is a subspace of the simplex bounded by the points $(1,0,0),\ (0,1,0),\ (0,0,1)$ in $\mathbb{R}^3$.

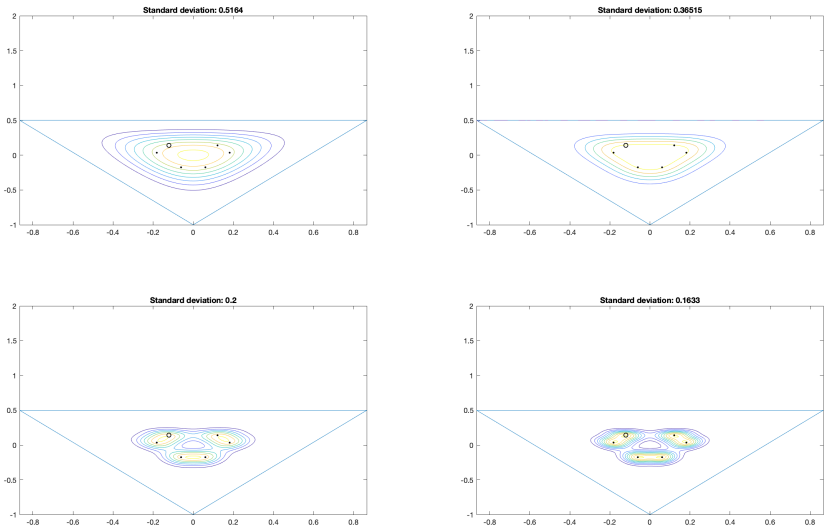

Figure 3.1: Level sets for the likelihood function

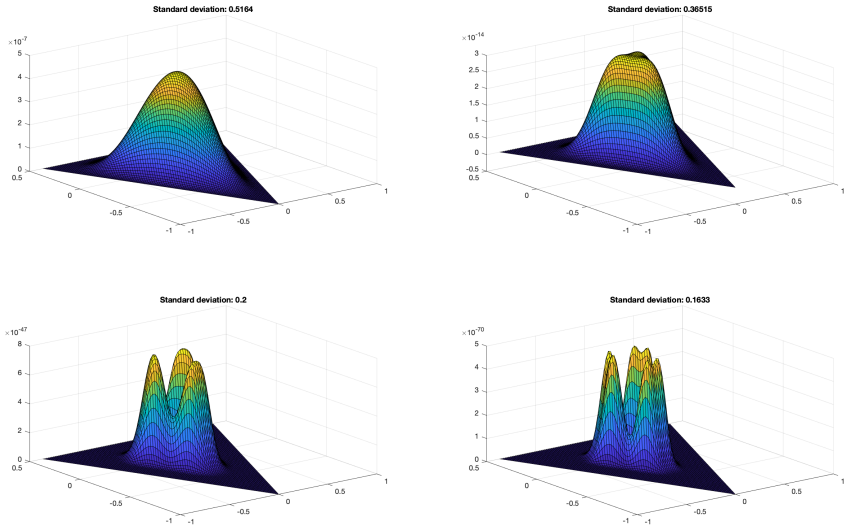

Figure 3.2: Likelihood function

From the plots in Figure 3.1 and in Figure 3.2 one can see the behaviour of the likelihood function depending on the different sample sizes n.

Observing the figures, it is worth stressing that the likelihood function is symmetric and that it has one peak corresponding to every permutation of $\mathbf{q}$. However when the sample size is small, then every peak's amplitude is big enough for all the peaks to sum up to one peak over the trivial solution. This is what happens for example in the case $n = 15$, i.e. in the plot in the top left corner in Figure 3.1 and in Figure 3.2. Then, for bigger sample sizes, each peak's amplitude gets smaller and the number of peaks increases to s, then to $s(s-1)$ and so on up to $s!$. Also one can see that for every sample dimension there is only one peak satisfying the order restriction.

3.2 The multivariate Gaussian case

Assume in this setting that the components of the estimator $\hat{\mathbf{q}}$ are distributed as independent Gaussian variables around the actual value of the frequency and with variance σ^2, i.e.

$$\hat{q}_i \sim \mathcal{N}(q_i, \sigma^2)$$

for $i = 1, 2, \ldots, s$. This is however a restrictive assumption and the setting here can be generalised in different ways. First, one would want to extend the analysis to the case in which the components of the estimator $\hat{\mathbf{q}}$ are independent Normal distributed with different variances. Second, it is of particular interest the case in which the components are not assumed to be independent. These cases will be discussed more in detail later on.

In the case of independent Normal distributed components with fixed variance σ^2, one can write the likelihood as

$$\begin{aligned} L(\mathbf{q}) &= \sum_{\chi \in S_s} l(\mathbf{q}, \chi) \\ &= \sum_{\chi \in S_s} \prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{n(\hat{q}_i - q_{\chi(i)})^2}{2\sigma^2} \right) \\ &= \sum_{\chi \in S_s} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\sum_{i=1}^{s} \frac{n(\hat{q}_i - q_{\chi(i)})^2}{2\sigma^2} \right). \end{aligned} \tag{3.7}$$

The maximum-likelihood estimator is then defined as

$$\tilde{\mathbf{q}} = \arg\max_{\mathbf{q} \in \mathcal{F}} L(\mathbf{q}), \tag{3.8}$$

where $\mathcal{F}$ is the parameter space

$$\mathcal{F} = \left\{ \mathbf{q} \in \mathbb{R}_+^s : q_1 \geq q_2 \geq \cdots \geq q_s, \ \sum_{i=1}^{s} q_i = 1 \right\}.$$

Since the function $L(\mathbf{q})$ is continuous and $\mathcal{F}$ is a closed set, the estimator $\tilde{\mathbf{q}}$, defined in (3.8), exists. Thus the following result holds.

Proposition 3 (Existence of the solution). *In the setting described in this section, the maximum likelihood estimator, defined in (3.8), exists.*

Now with a reasoning analogous to the one shown in Section 3.1, one can show that

$$\begin{aligned} L(\mathbf{q}) &= \sum_{\chi \in S_s} l(\mathbf{q}_{\chi^{-1}}, id) \\ &= \sum_{\chi \in S_s} l(\mathbf{q}_{\chi}, id), \end{aligned} \tag{3.9}$$

implying that $L(\mathbf{q})$ is invariant under permutation of the axis, i.e. that

$$L(\mathbf{q}) = L(\mathbf{q}_{\chi})$$

for any permutation $\chi \in S_s$.

Further, in this setting it is cleat that each $l(\mathbf{q}_{\chi}, id)$, being Gaussian, admits exactly one maximum that is attained at $\mathbf{q}_{\chi}$. We would like to claim that the function $L(\mathbf{q})$ has exactly one maximum that satisfies the order restriction provided by $\mathcal{F}$ allowing to state the following conjecture.

Conjecture 2 (Uniqueness of the solution). *In the setting described in this section, the maximum likelihood estimator, defined in (3.8), exists and is unique.*

We illustrate the behaviour of the estimator in Figure 3.3 and in Figure 3.4. Again, to show the behaviour of the likelihood function $L(\mathbf{q})$ with plots, s needs to be no bigger than 3.

In the following s is assumed to be 3, σ^2 to be 5 and $\tilde{\mathbf{q}}$ is estimated 4 times on samples of dimension 100, 350, 1500, 2000 respectively. The sample sizes are chosen in order to show the different behaviours of the likelihood function depending on the relation between the standard deviation and the distances among the peaks. Figure 3.3 shows the level sets of the likelihood function $L(\mathbf{q})$ together with the location of the permutations of $\mathbf{q}$ whereas Figure 3.4 shows the likelihood function. Both figures are plotted in barycentric coordinates for the 4 different choices of the sample size. For every plot the correspondent standard deviation is indicated.

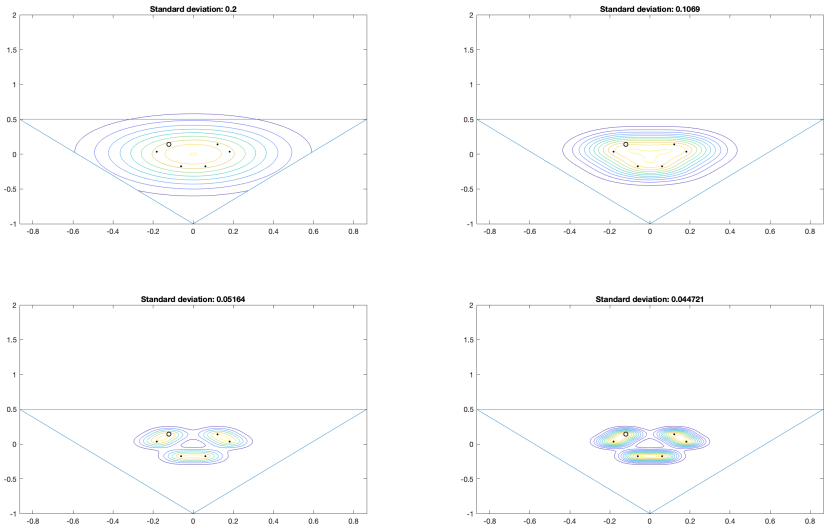

Figure 3.3: Level sets for the likelihood function

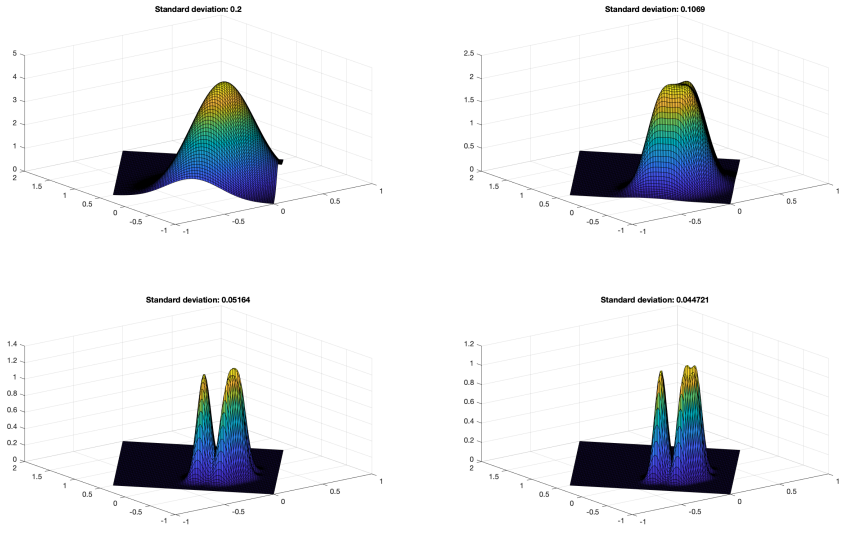

Figure 3.4: Likelihood function

From the plots in Figure 3.3 and in Figure 3.4 one can see that the behaviour of the likelihood function reflects what has been stated above.

It would be of great interest to study the sample dimensions correspondent to each peak's split but this has not been done here.

We will now generalise the assumptions on the starting estimator.

First, one can assume that the components of the estimator $\hat{\mathbf{q}}$ are distributed as independent Gaussian variables around the actual value of the frequency and with variance σ_i^2, i.e.

$$\hat{q}_i \sim \mathcal{N}(q_i, \sigma_i^2),$$

for$i = 1, 2, \ldots, s$. Under this assumption we can define a maximum-likelihood

estimator by

$$\tilde{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathcal{F}} L(\mathbf{q}), \tag{3.10}$$

where

$$\begin{aligned} L(\mathbf{q}) &= \sum_{\chi\in S_s} l(\mathbf{q},\chi) \\ &= \sum_{\chi\in S_s}\prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma_i^2/n}} \exp\left(-\frac{n(\hat{q}_i - q_{\chi(i)})^2}{2\sigma_i^2}\right). \end{aligned} \tag{3.11}$$

Existence of the maximum-likelihood estimator follows by the continuity of $L(\mathbf{q})$, and since the parameter space is closed.

Proposition 4 (Existence of the solution). *In the setting described in this section, the maximum likelihood estimator, defined in (3.10), exists.*

Here we conjecture that the likelihood function $L(\mathbf{q})$, defined in (3.11), admits unique maximum satisfying the order restriction provided by $\mathcal{F}$.

Conjecture 3 (Uniqueness of the solution). *In the setting described in this section, the maximum likelihood estimator, defined in (3.10), exists and is unique.*

The motivation for the claim of uniqueness of the maximum likelihood estimator comes from simulations in dimension $s = 3$, one of which is illustrated in Figure 3.5 and Figure 3.6.

In the following s is assumed to be 3, $\boldsymbol{\sigma}^2$ to be (5, 0.1, 1) and $\tilde{\mathbf{q}}$ is estimated 4 times on samples of dimension 30, 100, 200, 500 respectively.

The choices for the sample sizes is to show the different behaviours of the likelihood function depending on the relation between the peaks location and amplitude.

Figure 3.5 shows the level sets of the likelihood function $L(\mathbf{q})$ together with the location of the permutations of $\mathbf{q}$ whereas Figure 3.6 shows the likelihood function. Both figures are plotted in barycentric coordinates for the 4 different choices of the sample size.

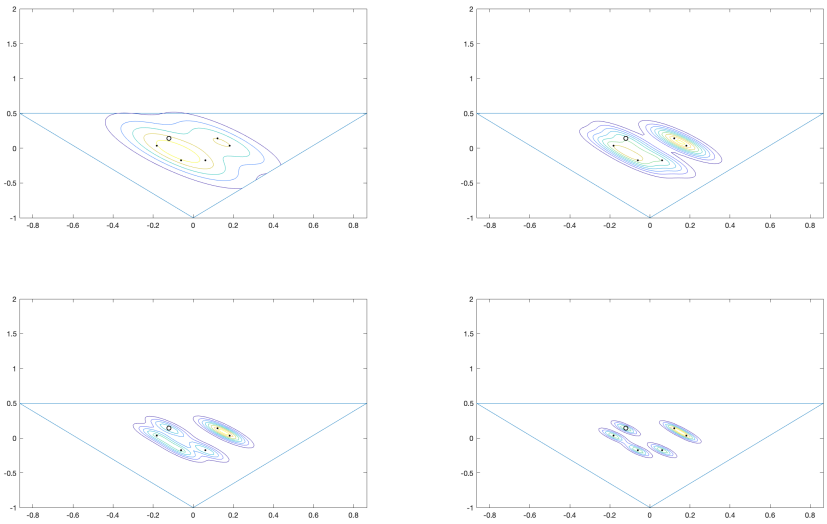

Figure 3.5: Level sets for the likelihood function

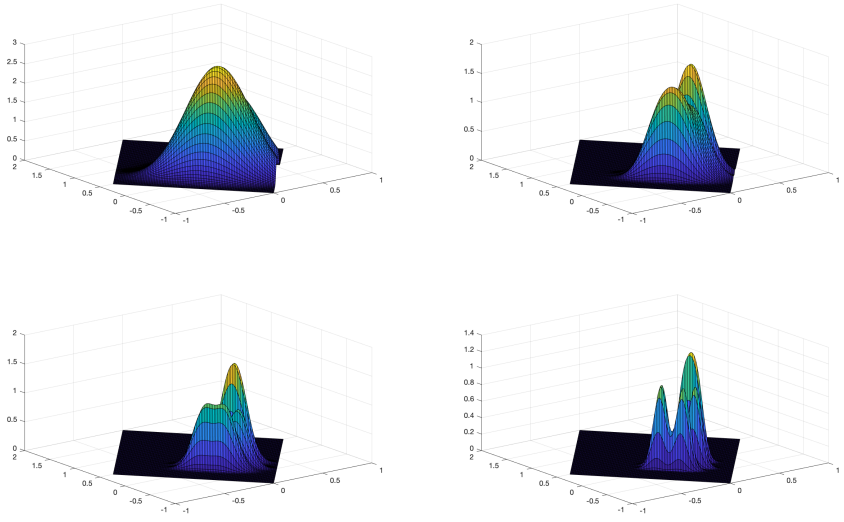

Figure 3.6: Likelihood function

From the plots in Figure 3.5 and Figure 3.6 it can be seen that the behaviour of the likelihood function is different from the previous setting in the sense that the peaks do not split together. It is also worth stressing that the orientation of the ellipses in Figure 3.2 depends on the ratio between the variances σ_i^2. In this case, being σ_1^2 bigger than the major axis of all ellipses is oriented along the q_1 direction. We have picked instead the variance vector to be (0.1, 5, 1) and all the other parameters to be same and illustrated the results in Figure 3.7.

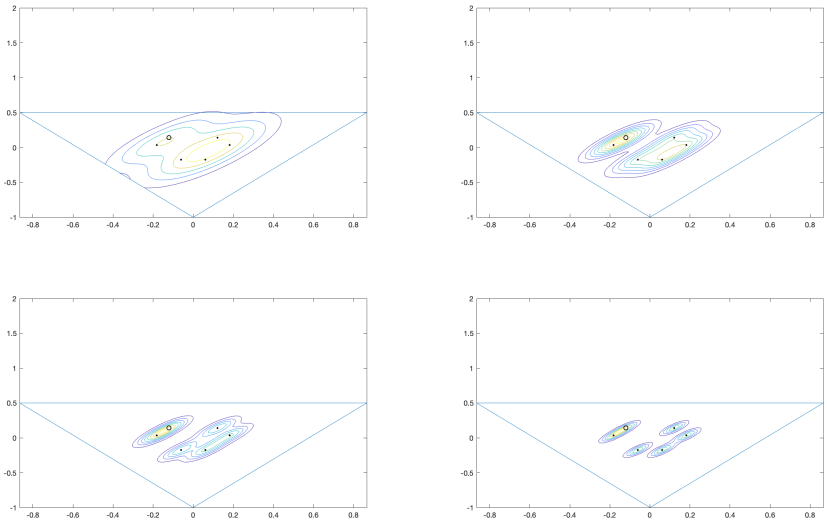

Figure 3.7: Level sets for the likelihood function

In Figure 3.7 one does not the dependence on the variance σ_3^2 since the third variable is implicitly defined by the other 2 due to the conditions provided by the parameter space $\mathcal{F}$. For this reason, in both the preceding cases σ_3^2 has been taken to be 1.

The final generalisation is to the case in which an asymptotic distribution result is given for the starting estimator. In our setting this can be for example the result provided in Theorem in [5], namely

Theorem 1. (Anevski, Pastukhov [3.4]) *The mle* $(\hat{q}_n, \hat{p}_n)$ *is strongly consistent*

$$(\hat{q}_n, \hat{p}_n) \xrightarrow[n\to+\infty]{a.s.} (q, p)$$

and asymptotically normal

$$\sqrt{n}\left((\hat{q}_n, \hat{p}_n)(q,p)\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, \mathbf{\Sigma}^2),$$

as $n \to +\infty$ *where*

$$\mathbf{\Sigma}^2 = [\boldsymbol{\partial\psi}(u)]_{1:2s,2:2s} \times \mathbf{\Sigma}_A^2 \times [\boldsymbol{\partial\psi}(u)]^T_{1:2s,2:2s}$$

and the notation $[\cdot]_{1:2s,2:2s}$ *is used to denote a matrix without the first column.*

Using the asymptotic distribution for $\hat{\mathbf{q}}$, we can state an (asymptotically valid) likelihood for the problem that we treat, namely

$$l(\mathbf{q}) = \sum_{\chi \in S_s} \frac{1}{\sqrt{(2\pi)^s |\mathbf{\Sigma}|^2}} \exp\left(-\frac{1}{2}(\hat{\mathbf{q}} - \mathbf{q}_\chi)\mathbf{\Sigma}^{-2}(\hat{\mathbf{q}} - \mathbf{q}_\chi)^T\right),$$

and introduce the maximum-likelihood estimator

$$\tilde{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathcal{F}} l(\mathbf{q}). \tag{3.12}$$

As in the previous setting, existence of the solution follows with a similar proof. We are not able to prove uniqueness and therefore state a conjecture.

Conjecture 4 (Existence and uniqueness of the solution). *In the setting described in this section, the maximum likelihood estimator, defined in (3.12), exists and is unique.*

The motivation for the claim of uniqueness of the maximum likelihood estimator comes from simulations in dimension $s = 3$, one of which is illustrated in Figure 3.8 and Figure 3.9.

In the following s is assumed to be 3, the covariance matrix to be

$$\mathbf{\Sigma}^2 = \begin{bmatrix} 5 & -1 & 0.1 \\ -1 & 1 & 0.4 \\ 0.1 & 0.4 & 1 \end{bmatrix}$$

and $\tilde{\mathbf{q}}$ is estimated 4 times on samples of dimension 30, 100, 200, 500 respectively. The choices for the sample sizes is to show the different behaviours of the likelihood function depending on the relation between the peaks location and amplitude.

Figure 3.8 shows the level sets of the likelihood function $L(\mathbf{q})$ together with the location of the permutations of $\mathbf{q}$ whereas Figure 3.9 shows the likelihood function. Both figures are plotted in barycentric coordinates for the 4 different choices of the sample size.

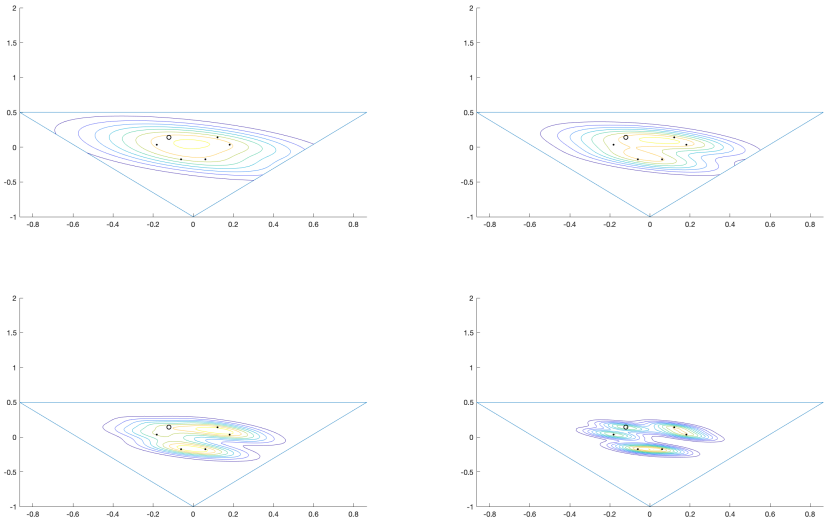

Figure 3.8: Level sets for the likelihood function

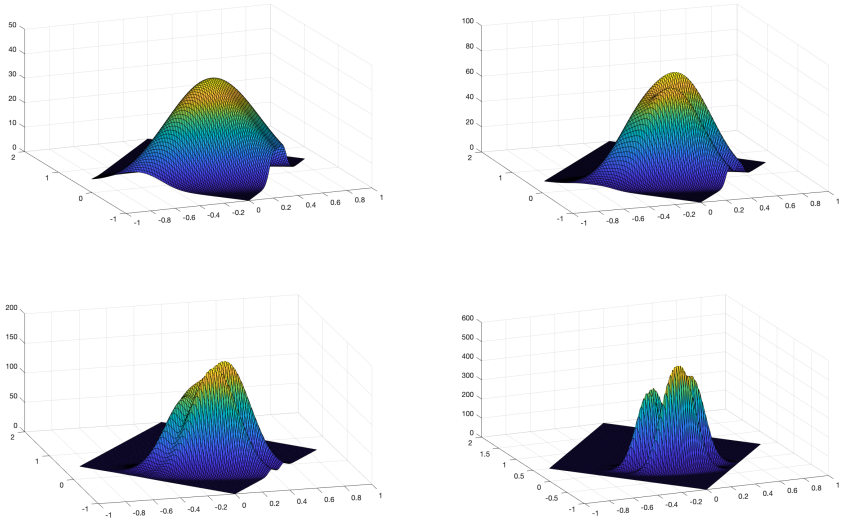

Figure 3.9: Likelihood function

From the plots in Figure 3.8 and Figure 3.9 it can be seen that now the function is completely not symmetric because of the correlation structure.

Conclusions

The goal of this thesis has been to improve on some previous results on the estimation of the wavelength distribution, and in particular if possible to correct for errors in the order. A further goal, and we would like to argue more important, has been to study estimators of a monotone probability mass function when the labels are unknown, from a general perspective.

Concerning the first goal, we have not been able to solve the rearrangement problem as we wanted to. This is because the statement of the problem as either a regression or a likelihood one on the parameter space

$$\mathcal{F} = \left\{ \mathbf{q} \in \mathbb{R}_+^s : q_1 \geq q_2 \geq \cdots \geq q_s, \ \sum_{i=1}^{s} q_i = 1 \right\}$$

does not give raise to any reshuffling of the indices. In fact the original paper in [1] on the monotone probability mass function estimation for unknown labels was not concerned with rearranging the order but rather with providing an estimate of the distribution for unseen species.

Further work related to this might be done either relaxing the order restriction

$$q_1 \geq q_2 \geq \cdots \geq q_s$$

and introducing appropriate weights or considering an analogous problem on the permutations' space S.

It is worth stressing that the present work allows for several extensions and future research.

It would also be interesting to revisit the missing species problem that was treated in [1]. It can be first considered the case in which only one species is missing and an estimator of $\mathbf{q}$ can be found. This can be later extended to the case in which a known finite number of species are missing in the sample and finally to the case in which an unknown finite number of species is missing.

Concerning the likelihood problem that has been dealt in Section 3.2, one can improve the results that have been presented here by studying consistency and asymptotic distribution of the estimator provided by (3.12). It would also be of great interest to look for an analytic expression or a characterisation of (3.12). To do so one can start by considering the i.i.d. setting as in (3.8) and then move to a more general setting.

Further, as it has been stated earlier one can also consider the issue of finding the sample dimensions corresponding to each time a peak of the likelihood function splits in the different settings that have been presented in Chapter 3. Most likely, however, a good result for this can be achieved only with numerical methods.

Related to this, one can also consider the problem of finding the least sample dimension n_ϵ required to have a fixed accuracy ϵ in the frequency estimation.

Lastly, simulations of the problem in (3.12) can be improved by writing an algorithm that considers the dependence of both the estimator $\hat{\mathbf{q}}$ and of the covariance matrix $\mathbf{\Sigma}^2$ on the sample dimension n.

Bibliography

[1] Anevski D., Fougères A. L. (2019), *Limit properties of the monotone rearrangement for density and regression function estimation*, Bernoulli, **25**(1), 549-583.

[2] Anevski D., Gill R. D., Zohren S. (2017), *Estimating a probability mass function with unknown labels*, The Annals of Statistic, Volume 45, Number 6, 2708-2735.

[3] Anevski D., Hossjer O. (2006), *A general asymptotic scheme for inference under order restrictions*, The Annals of Statistic, Volume 34, Number 4, 1874-1930.

[4] Anevski, D., Pastukhov, V. (2018), *Estimation of a discrete monotone distribution with model selection*, Tech report, arXiv.org

[5] Anevski D., Pastukhov V. (2018), *Estimating the distribution and thinning parameters of a homogeneous multimode Poisson process*, Tech report, arXiv.org.

[6] Assuncao R. M., Ferrari P. A. (2007), textitIndependence of thinned processes characterizes the Poisson process: An elementary proof and a statistical application, TEST **16**, 333-345.

[7] Jankowski H. K., Wellner J. A. (2009), *Estimation of a discrete monotone distribution*, Electronic Journal of Statistics, Volume 3, 1567-1605.

[8] Long Y. H. (1995), *Thinning and multinomial thinning of point*, Computers & Mathematics with Applications **30**: 1–4.

[9] Prakasa Rao, B. L. S. (1969). *Estimation of a unimodal density.* Sankhyā, Series A **31** 23–36.

[10] Van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press, New York.