SAPIENZA
UNIVERSITÀ DI ROMA

Dottorato di Ricerca in Statistica Metodologica
Tesi di Dottorato XXXI Ciclo

Dipartimento di Scienze Statistiche

Population size estimation via alternative parametrizations for Poisson mixture models

Francesco Catenacci

Advisor:

Luca Tardella

Dipartimento di Scienze Statistiche, Università La Sapienza, Roma

Ad Emanuela ed Ugo,
che hanno reso tutto questo possibile.

Summary

Chapter 1

Introduction

1.1 Population size estimation with zero truncated count data

Researchers interested in questions related to finite populations often wonder how best to estimate its size. Due to sampling limitation and elusiveness of certain units, a part of the population is often not observed, leading to the so-called truncated (or zero-truncated) count data. However, in many cases, the estimation of the number of unobserved zero counts is an important issue since the entire size of the population is the sum of counts that have been detected at least once and those who have not been counted at all. One of the oldest example of estimation of the unobserved units is given by Kermack et al. (1927) who analysed the number of individuals with cholera in 223 households in a village in India. The authors argued that there could be households with no cases of cholera because their members had not been exposed to the disease or because they just had not contracted the infection. In order to estimate the number of individuals who were exposed to the disease without showing the symptoms, they ignored the 168 households with zero cases and developed an estimator of the number of zeros using the other observations based on the zero-truncated Poisson distribution. Many of the practical examples pertaining to the estimation of the number of zeros are also related to the capture-recapture sampling scheme used in biology and ecology for monitoring the conservation of real wildlife animal population. Capture-recapture methods have been widely used in the last century. They rely on catching and marking a random and representative sample of the population which is then released into the population itself. After some time, another random sample is captured and the number of marked and unmarked entities is counted. Samples must be independent and everyone in the population should have an equal chance of being captured. The above mentioned method can be utilized when it is not straightforward to count all the individuals in the population. This technique has been also used to estimate population size for hard-to-reach people such as migrant workers, since it is more accurate than methods where no individual is marked at all (e.g. census) even though it is more expensive and time-consuming. In fact, the original capture-recapture idea can be dated back to 1786 when Laplace attempted to estimate the population size of France. A century later, these procedures were more rigorously and formally introduced by Petersen

(1986) and Lincoln (1930). Models and methods for estimating unobserved zeros are useful also in other contexts where the set of units is not a real population. An interesting famous literature example was studied in Efron and Thisted (1976) who tried to answer the following question: how many words did Shakespeare know? The information they took into account was that 31,534 different words were counted in Shakespeare's works, 14,376 words were used exactly once, 4,343 words were used exactly twice, 2,292 were used exactly three times, and so forth. Detailed data were reported in their Table 1. In this context, the frequency of zeros to be inferred would represent the number of words that Shakespeare knew but did not use in any of his known works.
In the present work, we consider the problem of inferring the total number of units in a finite population in the presence of count data where during an experiment or an observation stage all the units are potentially observable multiple times but only those who are observed at least once are in fact enumerated in the sample. The problem of estimating N, the total number of units of a finite population, is a recurrent problem in statistics. Other example of elusive populations can be found in a widespread range of fields such as public health and medicine, agriculture and veterinary science, software engineering, illegal behavior research and ecological sciences (Bunge and Fitzpatrick, 1993) (Chao et al., 2001) (Hay and Smit, 2003) (Roberts Jr and Brewer, 2006).

Let us now illustrate some real datasets that we will analyze in the following sections.

1. *Cholera.* The oldest example of estimation of the unobserved units is given by Kermack et al. (1927). The authors analysed the number of individuals with cholera in 223 households in a village in India.

k	1	2	3	4	n
(f_k)	32	16	6	1	55

Table 1.1. Cholera data-frequency distribution

2. *Expressed Sequence Tag (EST).* An EST is a partial sequence identifying a gene locus; ESTs are generated by sequencing randomly selected clones in a cDNA library made from an mRNA pool. In the experiment, 2586 possibly replicated sequence tags were detected from which $n = 1825$ genes were found.

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	23	27	n
f_k	1434	253	71	33	11	6	2	3	1	2	2	1	1	1	2	1	1	1825

Table 1.2. EST data-frequency distribution

3. *Traffic data.* This popular dataset, first introduced in Simar (1976), reports the number of accident claims submitted in a single year to la Royale Belge insurance Company out of $9,461$ policies covering both "business" and "tourist" automobiles. Indeed in this example the number of policies corresponds to the population size so that it is very often used as benchmark example.

k	1	2	3	4	5	6	7	n
(f_k)	1317	239	42	14	4	4	1	1621

Table 1.3. Traffic data-frequency distribution

4. *Colorectal polyps.* From medical research experiences it is well recognized that adenomatous polyps can be under-counted due to a misclassification in colonoscopy. We will use data from Alberts et al. (2000) where subjects with previous history of colorectal adenomatous polyps are allocated to one of two treatment groups, low fiber and high fiber, in order to evaluate the recurrence of colorectal adenomatous polyps in individuals with a history of the above mentioned disease.

k	1	2	3	4	5	6	7	8	9	10	11	12	n
f_k^{low}	145	66	39	17	8	8	7	3	1	0	2	3	299
f_k^{high}	144	61	55	37	17	5	4	6	5	1	1	5	341

Table 1.4. Polyps data-frequency distribution

1.2 Model setup with Poisson mixtures for count data

Let us assume that there is a system (a trapping mechanism, a marker, a diagnostic device, etc.) that identifies n distinct units from the population by counting how many times each unit replicates in the observation process. In order to understand the available data (which we will need to frame into a rigorous statistical model) we formalize a toy example by starting with the unavailable uncensored counts of a finite population with $N = 10$ distinct units. We denote by X_i the random number of times (count) that the population unit labeled with i is detected. Let us suppose that the full list of counts is the following:

$$(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0, X_5 = 1, X_6 = 0, X_7 = 2, X_8 = 0, X_9 = 3, X_{10} = 2).$$

Indeed, the actually observed data will be made of the unlabeled list of counts in a possible arbitrary order. Here we use a conventional non-decreasing order

$$(0, 0, 0, 0, 1, 1, 1, 2, 2, 3)$$

which makes it easy to realize that the relevant information is contained in the so-called frequencies of frequencies corresponding to the list of the number of distinct units counted once, twice, three times and so forth. We will denote by the generic f_j the number of distinct units counted exactly j times:

$$\boldsymbol{f_0 = 4, f_1 = 3, f_2 = 2, f_3 = 1}.$$

To fix notation, the finite population size is denoted by $N \in \mathbb{N}$. The population size will be the most important unknown parameter to be inferred based on the zero-truncated observed counts. For illustrative purpose, in the previous toy example N was known ($N = 10$). Our statistical model setup starts from providing a distribution

for the possibly unobserved/censored X_i for each unit i of the population with i ranging in $\{1, \ldots, N\}$. We assume that these random counts are drawn from a Poisson distribution, in symbols $X_i|\lambda_i \sim Poisson(\lambda_i)$ for $i = 1, \ldots, N$. There might be circumstances where the λ_i are all approximately equal hence the inferential problem reduces to making inference of an unknown number of homogeneous sets of possibly undetected (by zero truncation) Poisson count data. Albeit this situation has been deeply treated in literature, it is not flexible enough and it barely holds in practice since it imposes a certain relationship between mean and variance which is often violated. This most common heterogeneity in the λ_i associated to an over-dispersion of the counts suggests that a hierarchical structure, where the λ_i are drawn from a second stage distribution ν, might be preferable:

$$\begin{cases} \lambda_i|\nu \overset{iid}{\sim} \nu, \quad i = 1, \ldots, N \\ X_i|\lambda_i \overset{i}{\sim} Poisson(\lambda_i). \end{cases}$$

In this case we get the corresponding count distribution derived as marginal probability (with respect to the random λ) depending only on ν

$$P(X_i = j; \nu) = P_j(\nu) = \int_0^\infty \frac{e^{-\lambda}\lambda^j}{j!} d\nu(\lambda). \tag{1.1}$$

In this work, we will refer to this count distribution as Poisson mixture distribution with ν being the so-called mixing distribution.

In general, there are several ways to model the mixing distribution ν:

- *parametric* - ν is known up to the knowledge of few parameters; an example might be when the mixing is a gamma distribution with unknown rate and shape parameter; in this case the resulting count distribution is a negative binomial and the inference will involve also their two parameters;
- *finite mixture* - the mixing distribution ν is conceived as a distribution supported on a finite number of points; this is actually a flexible strategy, a compromise between the parametric and nonparametric approach;
- *nonparametric-ν* is left completely unspecified.

In this thesis, we will consider the most general case where ν is left completely unspecified. We will then denote by $P(X = j; \nu)$ the conditional probability (conditional on λ) of observing a single count equal to j.

Now, before specifying completely our statistical model, we should focus on explaining what is really available for inferring the unknown parameters N and ν. From our toy example it is clear that the only available counts are those X_i for which $X_i > 0$; However, we have also pointed out that they are exchangeable and that we have only an arbitrary way of labelling the observed distinct units. Furthermore, the labels of unobserved zero count are unavailable. All these assertions imply that the most appropriate and rigorous way of modelling the available data is to consider as observations the frequencies of frequencies $\boldsymbol{f}_+ = (f_1, f_2, \ldots)$. Indeed, it is better to introduce some further notations and clarify the relations among them. We will denote by M the random positive integer corresponding to the maximum

observed count and with n the random number of observed distinct units that is $n = \sum_{j=1}^{\infty} f_j = \sum_{j=1}^{M} f_j = \sum_{i=1}^{N} \mathbb{1}_{X_i>0}$. However, it will be clear that the most relevant part of the observed frequencies $\boldsymbol{f}_+$ is contained in the first M components

$$\boldsymbol{f}_M = (f_1, f_2, \ldots, f_M).$$

We can extend the vector of the frequencies of frequencies including also the unobserved frequency of zeros, $f_0 = \sum_{i=1}^{N} \mathbb{1}_{X_i=0} = N - n$, so that

$$\boldsymbol{f} = (f_0, f_1, f_2, \ldots).$$

Of course we also have that $\sum_{j=0}^{\infty} f_j = f_0 + n = N$ and that $\sum_{j=1}^{\infty} f_j = N - f_0 = n$.

We can now specify our statistical Poisson mixture model for inferring the size N of a finite population based on zero-truncated count observations as follows

$$(\mathbb{F}_+, p(f_+; \boldsymbol{\theta}), \Theta)\,.$$

The unknown parameter $\boldsymbol{\theta} = (N, \nu)$ is made of two components, the first being an integer valued $N \in \mathbb{N}$ and the latter being a probability measure ν on the set $[0, \infty)$ of possible Poisson rate parameters including 0 as the lower boundary case. In general, the set of probability measures on a measurable space $(E, \sigma(E))$ will be denoted by $\mathcal{P}([0, \infty))$ so that $\nu \in \mathcal{P}([0, \infty))$ and $\Theta = \mathbb{N} \times \mathcal{P}(E)$. From now on, when denoting a model specification, we will clearly highlight the components of the parameter space dropping the generic parameter space Θ. We thus get the following statistical model:

$$\mathcal{M}_{PM}^{orig} = \{\mathbb{F}_+, p(\boldsymbol{f}_+; N, \nu), \mathbb{N} \times \mathcal{P}([0, \infty)\}\,. \tag{1.2}$$

Now, in order to write down the sampling distribution for the observable frequencies of frequencies for a fixed value of $\boldsymbol{\theta} = (N, \nu)$, we will also need the probability of the unobservable zero count

$$P(X_i = 0; \nu) = P_0(\nu) = \int_0^{\infty} e^{-\lambda} d\nu(\lambda).$$

In fact, we have that for all $\boldsymbol{f}_+ \in \mathbb{F}_+$ for which $\sum_{j=0}^{\infty} f_j \leq N$

$$p(\boldsymbol{f}_+; \boldsymbol{\theta}) = \binom{N}{f_0\ f_1\ f_2\ \ldots} \prod_{j=0}^{\infty} P_j(\nu)^{f_j} = \binom{N}{f_0\ f_1\ f_2\ \ldots f_M} \prod_{j=0}^{M} P_j(\nu)^{f_j}. \tag{1.3}$$

In this hierarchical representation, the likelihood function corresponding to the observable vector of counts or, more rigorously, to the vector $\boldsymbol{f}_+$ is a restricted multinomial distribution

$$L(\boldsymbol{\theta}; \boldsymbol{f}_+) = L(N, \nu; \boldsymbol{f}_+) = \binom{N}{f_0\ f_1\ f_2\ \ldots} \prod_{j=0}^{\infty} P_j(\nu)^{f_j} \tag{1.4}$$

with the integral constraints

$$P(X_i = j; \nu) = P_j(\nu) = P_j = \int_0^{\infty} \frac{e^{-\lambda}\lambda^j}{j!}\, d\nu(\lambda) \quad j = 0, 1, \ldots \tag{1.5}$$

Considering that there is no contribution of the marginal count probability $P_j(\nu)$ for $j = M+1, M+2, \ldots$ in the likelihood, the likelihood function provides information only on a finite (random) number M of features (functionals) of ν. We will argue that a suitable reparametrization of the original parameter $\boldsymbol{\theta} = (N, \nu)$ might make this finite dimensional information even more explicit in terms of a finite number of components of a new infinite-dimensional parameter which is one-to-one with $\boldsymbol{\theta}$. In this work, we will simplify notation when needed and shorten all the $P_j(\nu)$ as P_j. We will then denote the space of the corresponding sequence of Poisson mixture probabilities by

$$\mathcal{P}_{PoisMix} = \left\{ (P_1, ..., P_j, ...) : P_j = \int_0^\infty e^{-\lambda} \frac{\lambda^j}{j!} d\nu(\lambda), \nu \in \mathcal{P}([0,\infty)), j = 0, 1, ... \right\}. \tag{1.6}$$

From the fact that $f_j = \sum_{i=1}^N \mathbb{1}_{X_i=j}$, we can derive the expected value of the j-th count frequency f_j as follows:

$$\mathbb{E}(f_j) = NP(X_i = j; \nu) = NP_j(\nu) = N \int_0^\infty \frac{e^{-\lambda}\lambda^j}{j!} d\nu(\lambda)$$

and therefore:

$$\mathbb{E}(f_0) = NP_0(\nu) = N \int_0^\infty e^{-\lambda} d\nu(\lambda).$$

The likelihood function (1.4) can be factorized in terms of the so-called conditional and residual likelihood where the conditional likelihood is based on the probability that a generic unit is counted j times i.e. $X_i = j$, conditionally on being observed, that is conditionally on $X_i > 0$. This conditional probability is denoted by $\overline{P}_j(\nu)$ and can be written as

$$\overline{P}_j = \overline{P}_j(\nu) = \frac{P_j(\nu)}{1 - P_0(\nu)},$$

so that, by defining the event $C_+ = \{\text{only positive counts are observed}\}$, one can write

$$L_C(\nu; \boldsymbol{f}_+) = p(\boldsymbol{f}_+ | C_+) = \binom{n}{f_1, f_2, \ldots} \prod_{i=1}^\infty \left(\overline{P}_j(\nu)\right)^{f_j} \tag{1.7}$$

$$= \frac{1}{(1 - P_0(\nu))^n} \binom{n}{f_1\ f_2\ \ldots f_M} \prod_{i=1}^M P_j(\nu)^{f_j}. \tag{1.8}$$

On the other hand the residual likelihood is

$$L_R(N, \nu; \boldsymbol{f}_+) = \binom{N}{f_0} P_0(\nu)^{f_0} \left(1 - P_0(\nu)\right)^{N - f_0} \tag{1.9}$$

$$= \binom{N}{N-n} P_0(\nu)^{N-n} \left(1 - P_0(\nu)\right)^n \tag{1.10}$$

where we have used the fact that $f_0 = N - n$ and that $\binom{N}{N-n} = \binom{N}{n}$ to get the last expression. Note that both components of the parameter $\boldsymbol{\theta} = (N, \nu)$ appear in the residual likelihood with ν determining P_0 while the conditional likelihood involves

only ν by means of $\overline{P}_j(\nu)$. Hence, one can express the full likelihood as the product of the conditional likelihood and the residual likelihood so that:

$$L(N,\nu;\boldsymbol{f}) = \binom{N}{n} P_0(\nu)^{N-n}\left(1-P_0(\nu)\right)^n \binom{n}{f_1\ f_2,\ldots,f_M} \prod_{i=1}^{M} \overline{P}(\nu)_j^{f_j}. \tag{1.11}$$

From now on, for ease of notation, we will omit the ν argument in the marginal count probability so that we will have P_j instead of $P_j(\nu)$. To help the reader with the growing number of symbols and notations we have set up a reference list of symbols in Appendix B.

From the work of Sanathanan (1972), the classical inferential approach consists in dividing the estimation process in two steps:

- estimate ν or its corresponding parameters through the conditional likelihood $L_C(\nu;\boldsymbol{f}_+)$, obtaining
$$\hat{\nu}_C = \arg\max_{\nu} L_C(\nu;\boldsymbol{f}_+);$$
- plug $\hat{\nu}_C$ in $P_0(\nu)$ in the residual likelihood and hence obtain an estimate of the population size N by maximizing $L_R(N,\hat{\nu}_C;\boldsymbol{f}_+)$. Since the residual likelihood has a binomial structure with unknown size parameter, we get the following explicit expression:
$$\hat{N}_C = \hat{N}_{P_0(\hat{\nu}_C)} = \arg\max_{N} L_R(N,\hat{\nu}_C) = \left\lceil \frac{n}{1-P_0(\hat{\nu}_C)} \right\rceil. \tag{1.12}$$

The most straightforward way to infer on the unknown parameter $\boldsymbol{\theta}$ is to rely on the full likelihood. However, from the previous remark, the likelihood function provides information on $P_0,\ldots,P_M$ and this may suggest a first reparametrization of ν in terms of the sequence $\boldsymbol{P} = (P_0, P_1, \ldots, P_M, \ldots)$. On the other hand we have also pointed out the presence of integral constraints which discourage us to proceed further in this direction. Indeed, it might be still convenient to highlight that in the maximization of the full likelihood there is an explicit relation between the components N and $\boldsymbol{P}$; In fact, if the MLE estimates are denoted by

$$\left(\hat{N},\widehat{\boldsymbol{P}}\right) = \underset{(N,\boldsymbol{P})\in\mathbb{N}\times\mathcal{P}_{PoisMix}}{\arg\max} L(N,\boldsymbol{P};\boldsymbol{f}_+)$$

we have that:

$$\hat{N} = \lceil n/\left\{1-P_0(\hat{\nu})\right\}\rceil.$$

1.3 Literature review and thesis outline

In the literature, there were several authors who utilized Poisson mixtures distributions in the context of population size estimation with zero-truncated counts. Norris and Pollock (2004) proposed to estimate ν by nonparametric maximum likelihood. Wang and Lindsay (2005) suggested a penalized nonparametric maximum likelihood estimator, which is stabler than that of Norris and Pollock (2004). Nevertheless,

severe underestimation can occur due to the interplay of factors including inadequate sampling effort, heterogeneity and skewness of the abundance curve. This weakness is also observed in other non-likelihood-based nonparametric approaches. Wang and Lindsay (2008) proposed a partial prior approach for ν in order to improve the estimation of N. Böhning et al. (2005) analyzed finite (nonparametric) mixtures of Poisson and Binomial.

One of the best performing methods has been proposed in Wang (2010). The author considered a Poisson compound gamma model estimating the mixture by a nonparametric penalized maximum likelihood approach using a least squares cross-validation procedure for the choice of the common shape parameter. Several authors relied on a conditional likelihood and investigated the non-identifiability issues for which it can be argued that only a lower bound on P_0 can be estimated in the presence of a finite sample. This yields a lower bound estimate for N which can be argued to consistently achieve the true N when N grows indefinitely. One of the pioneering works in this direction is Mao and Lindsay (2007) with further insights in Mao (2006), Mao et al. (2013). One of the most famous lower bound estimator is the Chao lower bound (Chao, 1984). To derive this estimator, initially developed for species richness estimation, the author only used the singletons and doubletons (i.e. the count frequencies of the rarest units). A similar estimator was derived again by Chao for incidence data. Chiu et al. (2014) derived an improved nonparametric lower bound starting from the Good-Turing frequency formula. This proposed lower bound is nonparametric since it is universally valid for any mixing distribution and it can be derived for both species richness and incidence data. Böhning et al. (2018) proposed an innovative lower bound estimator in case of one-inflation data using only the doubletons and the tripletons. An interesting and inspiring contribution came from Daley and Smith (2016) who extended the popular moment-based lower bound introduced by Chao and derived a non-parametric moment-based estimator that is both computationally efficient and is sufficiently flexible to account for heterogeneity in the mixing distribution. They started from an innovative work of Harris (1959) who did not manage to utilize moments of higher orders due to the lack of algorithms to solve complex system of nonlinear equations.

From a Bayesian perspective, Barger and Bunge (2010) derived the form of two objective priors, using Bernardo's reference method and Jeffreys' rule, based on the mixed-Poisson likelihood used in the single-abundance sample species problem. The authors have been the first ones to examine the objective priors in depth in the context of species sampling. Guindani et al. (2014) considered a Bayesian semi-parametric approach to implement inference for sequence abundances distribution. Differently from the latter, where a nonparametric Dirichlet process prior is used for the mixing distribution ν, Alunni Fegatelli and Tardella (2018) proposed an alternative nonparametric estimate of the population size based on the unconditional likelihood reparametrized in terms of a finite number of moments of a suitable mixing distribution.

In order to better understand the outline of what we have planned to do in this thesis, it will be useful introducing the most relevant mapping which can be used to reparametrize the original statistical model. In fact, instead of using the original probability measure ν, we will find it useful to consider the following finite measure:

$\phi : \sigma([0,\infty)) \in [0,1]$

$$\phi(A) = \int_A d\phi(\lambda) = \int_A e^{-\lambda} d\nu(\lambda).$$

This transformation of ν was introduced in Teicher (1960). Although a finite measure is unusual for a nonparametric component in a statistical model, we will show that it can be a convenient way to conceive an explicit infinite dimensional moment-based parameter for our nonparametric setting. If we denote the sequence of all ordinary moments of the finite measure ϕ by

$$\boldsymbol{s} = (s_0, s_1, s_2, \ldots, s_j, \ldots)$$

where

$$s_j = \int_{[0,\infty)} \lambda^j d\phi(\lambda) \tag{1.13}$$

we have that s_0 corresponds to the total mass of ϕ

$$s_0 = \phi([0,\infty)) = P_0 \tag{1.14}$$

and

$$s_j = \int_{[0,\infty)} \lambda^j d\phi(\lambda) = \int_{[0,\infty)} \lambda^j e^{-\lambda} d\nu(\lambda) = j! P_j. \tag{1.15}$$

Indeed, the finite measure ϕ can be identified by the sequence of its moments since it admits, for all $0 \leq t \leq 1$, a finite value of its moment generating function

$$M_\phi(t) = \int_0^\infty e^{t\lambda} d\phi(\lambda) < \infty$$

with the constraint that

$$M_\phi(1) = \sum_{j=0}^{\infty} \frac{s_j}{j!} = \int_0^\infty e^{\lambda} d\phi(\lambda) = \int_0^\infty e^{\lambda} e^{-\lambda} d\nu(\lambda) = 1. \tag{1.16}$$

Moreover, if we consider the normalized probability measure

$$\widetilde{\phi}(A) = \frac{\phi(A)}{\phi([0,\infty))}$$

the corresponding sequence of moments will be

$$\widetilde{s}_j = \int_{[0,\infty)} \lambda^j d\widetilde{\phi}(\lambda) = \frac{s_j}{s_0} = \frac{j!}{P_0} P_j \tag{1.17}$$

and we could use equivalently a sequence of moments of a probability measure

$$\widetilde{\boldsymbol{s}} = (\widetilde{s}_0, \widetilde{s}_1, \widetilde{s}_2, \ldots, \widetilde{s}_j, \ldots)$$

with the constraint implied by (1.16)

$$\sum_{j=1}^{\infty} \frac{\widetilde{s}_j}{j!} = \frac{1}{s_0}$$

which uniquely identifies $\widetilde{\phi}$ through its moment generating function.

In this way, the new measure ϕ (or, equivalently, its normalized version $\widetilde{\phi}$) and the sequence of its constrained moments will enable us to single out an alternative moment-based reparametrization

$$\boldsymbol{\vartheta} = (N, \boldsymbol{s})$$

which it allows to specify a fully identifiable statistical model.
We will argue that, for the structure of the likelihood function (5.1), there is information only for the first $M+1$ moments (up to the M-th order moment including the total mass) of ϕ; this fact will allows us to restrict inference, once observed a finite sample, only to a finite dimensional component of the infinite dimensional new parameter for which we get information from the likelihood, namely

$$(N, \boldsymbol{s}_M)$$

where $\boldsymbol{s}_M = (s_0, s_1, ..., s_M)$ is a truncated moment sequence.
If we denote $\mathcal{F}(E)$ as the set of all finite measures on the measurable space $(E, \sigma(E))$ with $\sigma(E)$ as a suitable σ-field of subsets of E, we can extend this notation in order to refer to the restricted set of finite measures corresponding to the nonparametric component ϕ on $[0, \infty)$ (with total mass in (0,1]) of the reparametrized statistical model. We denote this space by

$$\mathcal{F}_{(0,1]}([0,\infty)) = \{\phi \in \mathcal{F}([0,\infty)) : \phi([0,\infty)) = s_0(\phi) \in (0,1]\}.$$

However, we should restrict this class in order to account for the constraint (1.16) with the following notation

$$\overline{\mathcal{F}}_{(0,1]}([0,\infty)) = \{\phi \in \mathcal{F}([0,\infty)) : \phi([0,\infty)) = s_0(\phi) \in (0,1], M_\phi(1) = 1\}.$$

Similarly, we should point out that the class of probability measures $\widetilde{\phi}$ must lie in the restricted space of probability measures with moment generating functions denoted by

$$\overline{\mathcal{P}}_{[1,\infty)}([0,\infty)) = \left\{\widetilde{\phi} \in \mathcal{P}([0,\infty)) : M_{\widetilde{\phi}}(1) \in [1,\infty)\right\}.$$

If we temporarily overlook the constraint just highlighted, we have motivated the interest for finite dimensional truncated moment sequences $\boldsymbol{s}_M$ as well as $\widetilde{\boldsymbol{s}}_M$ with the relation that

$$\boldsymbol{s}_M = s_0 \cdot \widetilde{\boldsymbol{s}}_M$$

where $\widetilde{\boldsymbol{s}}_M$ can be thought of as an element of the M-truncated moment space corresponding to probability distributions

$$\mathcal{S}^{[M]} = \left\{(\widetilde{s}_0, \widetilde{s}_1 \ldots, \widetilde{s}_M) : \widetilde{s}_k = \int_0^\infty \lambda^k d\widetilde{\phi}(\lambda), \widetilde{\phi} \in \mathcal{P}([0,\infty))\right\}. \tag{1.18}$$

In this thesis we will review in chapter 2 some theory behind distributions and moments, moment spaces and their truncated versions; moreover, we will introduce suitable mappings which allows one-to-one relations with alternative moment reparametrizations of a distribution. These reparametrizations will play a key role in:

- investigating inferential issues such as identifiability and the ability of inferring the target population size via sharpest lower bound (chapter 3);
- providing new computational device for lower bounds estimation of population size in a classical setting (chapter 4);
- deriving fully Bayesian inference with a suitably defined original prior distribution on the model component (chapter 5).

Chapter 2

Moment problems and useful reparametrizations for Poisson mixtures

In this work we are considering the problem of inferring the total number of units of a finite population by using and comparing different inferential methods. Despite the fact that these methods are poles apart, all of them have a common denominator: the usage of suitable reparametrizations of the nonparametric component of the statistical model. One of the key goals of these reparametrizations is avoiding to deal with annoying integral constraints such as those related with either the original Poisson mixtures probabilities (1.1) or the ordinary moments of measures such as those in (1.17) or (1.13). The key tools will be the one-to-one mappings from ordinary moments to unconstrained product spaces of real intervals. We will show how these mappings can be considered as an extension of the Skibinsky canonical moments mapping for probability measures on $[0, 1]$ and its relation with three-terms recurrence equations for orthogonal polynomials systems. This chapter will be focused on reviewing theoretical results about these mappings following Dette and Studden (1997), Dette and Nagel (2012) and Tomecki (2018).

2.1 Hausdorff moment problem

For a generic measurable set $(E, \sigma(E)) \subset \mathbb{R}$ we can denote by $\mathcal{P}(E)$ the set of probability measures on $(E, \sigma(E))$. If all the moments of $\mu \in \mathcal{P}(E)$ exist, we can write the k-th moment of $\mu \in \mathcal{P}(E)$ as:

$$s_k = s_k(\mu) = \int_E x^k \, d\mu(x), \quad k = 1, 2, \ldots \tag{2.1}$$

and the moment space corresponding to the space $\mathcal{P}_{afm}(E)$ of all probability measures with all finite moments as:

$$\mathcal{S}(E) = \left\{ (s_0, s_1, s_2, \ldots) : s_k = \int_E x^k \, d\mu(x), k = 0, 1, 2, \ldots \quad , \mu \in \mathcal{P}_{afm}(E) \right\}. \tag{2.2}$$

In this section we focus on $\mathcal{S}([0, 1])$, the moment space in the particular case of $E = [0, 1]$. Indeed in this case $\mathcal{P}_{afm}(E) = \mathcal{P}([0, 1])$ since all moments exist and

are bounded in $[0,1]$. Moreover, all the moments of order $k+1$ can be bounded from below and above when the moments of lower order $1,\dots,k$ are fixed. Let us illustrate this assertion for $k=1$. Starting from the simple observation that in the space $E=[0,1]$ we have that $x \geq x^2 \ \forall\, x \in [0,1]$ and from the monotonicity property of the Lebesgue integral, we easily find out that there's a first inequality constraint between first and second order moment:

$$s_1(\mu) = \int_0^1 x\, d\mu(x) \geq \int_0^1 x^2\, d\mu(x) = s_2(\mu).$$

Moreover, from the well known Jensen's inequality, we have that

$$s_1(\mu)^2 = \left(\int_0^1 x\, d\mu(x)\right)^2 \leq \int_0^1 x^2\, d\mu(x) = s_2(\mu)$$

and we therefore derive an explicit description of the 2-truncated moment space:

$$\mathcal{S}^{[2]}\left([0,1]\right) = \left\{(s_1, s_2) \in [0,1]^2 \mid {s_1}^2 \leq s_2 \leq s_1\right\}.$$

This result gives us a first intuition: the upper bound strictly depends on the fact that the measure lies in $[0,1]$ and thus $x \geq x^2$ is always valid and an upper bound, if some conditions are met, exists. If we fix $k \in \mathbb{N}$, we can define the extremal moments of order $k+1$ as:

$$\begin{aligned} s_{k+1}^+ &= s_{k+1}^+(s_1,\dots,s_k) = sup\left\{s_{k+1}(\mu) \mid \mu \in \mathbb{P}\left([0,1]\right), s_i(\mu) = s_i \ \forall i = 1,\dots,k\right\} \\ s_{k+1}^- &= s_{k+1}^-(s_1,\dots,s_k) = inf\left\{s_{k+1}(\mu) \mid \mu \in \mathbb{P}\left([0,1]\right), s_i(\mu) = s_i \ \forall i = 1,\dots,k\right\} \end{aligned} \tag{2.3}$$

where $s_1^+ = 1$ and $s_1^- = 0$. This provides us with a lower bound and an upper bound for the $k+1$-th moment of μ when $(s_1(\mu) = s_1, \dots, s_k(\mu) = s_k)$ are fixed. In order to derive the extremal moments, we need to define the so-called Hankel matrices

$$\underline{H}_{2m} = \begin{bmatrix} s_0 & s_1 & \cdots & s_m \\ s_1 & s_2 & \cdots & s_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ s_m & s_{m+1} & \cdots & s_{2m} \end{bmatrix}, \quad \overline{H}_{2m} = \begin{bmatrix} s_1 - s_2, & \cdots & s_{m-1} - s_m \\ s_2 - s_3 & \cdots & s_m - s_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m-1} - s_m & \cdots & s_{2m-1} - s_{2m} \end{bmatrix}$$

and

$$\underline{H}_{2m+1} = \begin{bmatrix} s_1 & \cdots & s_{m+1} \\ s_2 & \cdots & s_{m+2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m+1} & \cdots & s_{2m+1} \end{bmatrix}, \quad \overline{H}_{2m+1} = \begin{bmatrix} s_0 - s_1, & \cdots & s_m - s_{m+1} \\ s_1 - s_2 & \cdots & s_{m+1} - s_{m+2} \\ \vdots & \vdots & \ddots & \vdots \\ s_m - s_{m+1} & \cdots & s_{2m} - s_{2m+1} \end{bmatrix}.$$

By theorem IV.1.1 in Rice (1967) for any finite sequence $(s_1,\dots,s_n)$ we have:

$$\begin{aligned} (s_1,\dots,s_n) \in int\ \mathcal{S}^{[n]}\left([0,1]\right) &\iff \underline{H}_n \text{ and } \overline{H}_n \text{are positive definite}, \\ (s_1,\dots,s_n) \in \mathcal{S}^{[n]}\left([0,1]\right) &\iff \underline{H}_n \text{ and } \overline{H}_n \text{are positive semidefinite}. \end{aligned} \tag{2.4}$$

We can now define the so-called *canonical moments* (Skibinsky, 1968) as:

$$c_l = \frac{s_l - s_l^-}{s_l^+ - s_l}, \quad l = 1, 2, \ldots. \tag{2.5}$$

The sequence $(c_1, \ldots, c_n)$ is the sequence of *canonical moments* corresponding to $(s_1, \ldots, s_n)$. By defining the complementary canonical moment $q_i = 1 - c_i$ we can then state the following relationship:

$$\prod_{i=1}^{n-1} q_i c_i = s_n^+ - s_n^-. \tag{2.6}$$

The canonical moments are not easy to calculate. However, the determinants of the predefined Hankel Matrices can be used to derive their expression:

$$c_n = \frac{det\underline{H}_n det\overline{H}_{n-2}}{det\underline{H}_{n-1} det\overline{H}_{n-1}}, \quad q_n = \frac{det\underline{H}_{n-2} det\overline{H}_n}{det\underline{H}_{n-1} det\overline{H}_{n-1}}. \tag{2.7}$$

From definition (2.5), we can explicitly denote the one-to-one mapping

$$\rho_n \ : (c_1, \ldots, c_n) \to (s_1, \ldots, s_n) \tag{2.8}$$

whose Jacobian determinant is given in Dette and Studden (1997)

$$\left|\frac{\partial \rho_n}{\partial s_n}\right| = \prod_{k=1}^{n} \frac{\partial c_k(s_n)}{\partial s_k} = \prod_{k=1}^{n} (s_k^+ - s_k^-)^{-1} = \prod_{k=1}^{n} (c_k(1 - c_k))^{-(n-k)}. \tag{2.9}$$

If we extend this mapping to all the sequence of moments, the corresponding extended one-to-one mapping $\rho : [0,1]^\infty \to \mathcal{S}([0,1])$ and its inverse $\rho^{-1} : \mathcal{S}([0,1]) \to [0,1]^\infty$ will allow us to reparametrize the space $\mathcal{P}([0,1])$ or, equivalently, its moment space $\mathcal{S}([0,1])$ in terms of a more convenient product space $[0,1]^\infty$. This idea has been put forward in the context of nonparametric population size estimation in Tardella (2002) and Alunni Fegatelli and Tardella (2018). In particular, we point out that Alunni Fegatelli and Tardella (2018), instead of directly addressing the statistical model (1.11), relied on a convenient approximation which bypassed the problem of dealing with moments of probability measures on the unbounded space $[0, \infty)$ for the Poisson rate parameter λ. Before moving to the moment space $\mathcal{S}([0, \infty))$ it is better to introduce, as intermediate step, the moment space $\mathcal{S}([a, b])$ for $0 \leq a < b < \infty$. Given any measure μ defined in $[0, 1]$ and given any finite real numbers a and b with $a < b$ there is a natural one-to-one map which links $\mu \in \mathcal{P}([0, 1])$ with a corresponding $\mu' \in \mathcal{P}([a, b])$. We can understand this map using the (one-to-one) linear map corresponding to underlying random variables: if $X \sim \mu$ then $Y = a + (b - a) * X \sim \mu'$. This linear map is very useful since canonical moments $c_1, \ldots, c_n$ are invariant under linear map (Dette and Studden, 1997) while the canonical moment mapping associated to this new measure

$$\chi_n : [0,1]^n \to \mathcal{S}^{[n]}([a, b])$$

has Jacobian determinant equal to:

$$\left|\frac{\partial \chi_n}{\partial s_n}\right| = \prod_{k=1}^{n} \frac{\partial c_k(s_n)}{\partial s_k} = \prod_{k=1}^{n} (s_k^+ - s_k^-)^{-1} = (b - a)^{-n(n+1)/2} \prod_{k=1}^{n} (c_k(1 - c_k))^{-(n-k)}. \tag{2.10}$$

2.2 Stieltjes moment problem

Let us now consider spaces of probability measures with non-negative support so that the corresponding space $E = [0, \infty)$ is unbounded. This is the appropriate space for dealing with probability measures and finite measures related to the Poisson parameter rate λ if we want to avoid to restrict our attention to approximations of the nonparametric component of Poisson mixtures as in Alunni Fegatelli and Tardella (2018). However, one cannot immediately extend the previous idea of canonical moments mapping since the new moment space $\mathcal{S}([0, \infty))$ does not form a compact set and there is no finite upper bound s_{k+1}^{+}. Moreover, in general, dealing with moment sequences in $\mathcal{S}([0, \infty))$ is not trivial since, differently from the compact case, there is no guarantee that a moment sequence uniquely identifies a probability measure unless appropriate moment conditions are imposed such as those in Carleman (1923). Luckily, as discussed in the previous chapter, for some probability measures that we are going to consider in our Poisson mixtures setting there exist sufficient conditions for the moment problem determinacy. Moreover, we will still use known properties of the moment space such as the equivalences in Rice (1967)

$$\begin{aligned}(s_1, \ldots, s_n) \in int\mathcal{S}^{[n]}([0, \infty)) \text{ if and only if } \underline{H}_n \textit{ and } \underline{H}_{n-1} \text{ are positive definite;}\\ (s_1, \ldots, s_n) \in \mathcal{S}^{[n]}([0, \infty)) \text{ if and only if } \underline{H}_n \textit{ and } \underline{H}_{n-1} \text{ are positive semidefinite;}\end{aligned} \tag{2.11}$$

and the fact that a non trivial lower bound s_{k+1}^{-} always exists. In order to adapt in $\mathcal{S}([0, \infty))$ the previous strategy in $\mathcal{S}([0, 1])$ relying on canonical moments, we can consider the following quantities:

$$z_j = \frac{s_j - s_j^{-}}{s_{j-1} - s_{j-1}^{-}} \tag{2.12}$$

with $s_0^{-} = 0$. The inferior extremal moments s_i^{-} always exist and in particular $s_i^{-} < s_i \ \forall i$.

Definition 2.2.1. *Let $\underline{H}_{2m}$ be the moment Hankel matrix define in* (2.4)*. We define:*

- $\underline{h}_{2m}^{T} = (s_m, \ldots, s_{2m})$
- $\underline{h}_{2m-1}^{T} = (s_{m-1}, \ldots, s_{2m-1})$

Then, whenever the inverse of the Hankel Matrix exists,

$$\begin{aligned}\underline{s}_{2m}^{-} &= \underline{h}_{2m-1}^{T} \underline{H}_{2m-2}^{-1} \underline{h}_{2m-1}\\ \underline{s}_{2m+1}^{-} &= \underline{h}_{2m}^{T} \underline{H}_{2m-1}^{-1} \underline{h}_{2m}\end{aligned} \tag{2.13}$$

with $s_1^{-} = 0$

From the above definition it follows that $s_2^{-} = s_1 s_0^{-1} s_1$; since $s_0 = 1$ we have that $s_2^{-} = s_1^2$.

We then denote by

$$\xi_n : [0,\infty)^n \to \mathcal{S}^{[n]}([0,\infty))$$

the mapping corresponding to the previous quantities so that

$$\boldsymbol{z}_n = (z_1, ..., z_n) \xrightarrow{\xi_n} (s_1, ..., s_n) = \boldsymbol{s}_n.$$

The moment space $\mathcal{S}([0,\infty))$ is strictly related to $\mathcal{S}([0,b])$. In fact, for a fixed point $s_n \in \mathcal{S}^{[n]}([0,\infty))$, there exists a $b_0 \in \mathbb{N}$ with $s_n \in \mathcal{S}^{[n]}([0,b])$ for all $b \geq b_0$. Let

$$\boldsymbol{c}_n = \rho^{-1}(\boldsymbol{s}_n)$$

denote the canonical moments corresponding to the moment sequence $\boldsymbol{s}_n$ in the moment space $\mathcal{S}^{[n]}([0,b])$. Dette and Nagel (2012) showed that:

$$c_k = \frac{z_k(\boldsymbol{s_n})}{b} \quad k = 1, 2, \ldots$$

and for $b \to \infty$ we get the result.
We will refer to $(z_1, \ldots, z_n)$ as *pseudo-canonical moments*. We put the prefix *pseudo* because they differ from the original canonical moments defined by Skibinsky (1968) but still they produce a bijection between the interior of the moment space $\mathcal{S}^{[n]}([0,\infty))$ and the product space $[0,\infty)^n$.

As it happens for the original canonical moments, the parameters z_i can be calculated directly in terms of the Hankel determinants. Before proceeding with the derivation of the pseudo-canonical moments, we need to recall a concept from the linear algebra. By Sylvester's criterion (cf. theorem 7.2.5 in Horn and Johnson (1985)) a symmetric matrix is positive definite if and only if all of its principal minors are positive (we recall that a minor of a given matrix B is simply the determinant of some smaller square matrix, obtained from B by removing one or more of its rows and columns). We can then extend the sequence $\boldsymbol{s_{n-1}} = (s_1, \ldots, s_{n-1})$ by one more moment s_n. The new sequence lies in the interior of $\mathcal{S}^{[n]}([0,\infty))$ if and only if $detH_n$ is positive. In general, the determinant is not linear with respect to the entire Hankel matrix: it is in fact a linear function of each column of the matrix only when the other columns are held fixed. In particular, when we write the determinant as a function of only the last moment of order n, $f(s_n)$, we can see that this function is affine-linear. If $f(s_n) > 0$ holds, the aforementioned moment sequence lies in the interior of the moment space by Silvester's criterion. When $f(s_n) < 0$, the moment sequence lies outside of the closure of the truncated moment space and the Hankel matrix cannot be positive semi-definite. For the linear-affinity of f, $f(s_n^-) = 0$ must hold. So we have

$$det\underline{H}_{2k} = \begin{bmatrix} s_0 & s_1 & \cdots & s_k \\ s_1 & s_2 & \cdots & s_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ s_k & s_{k+1} & \cdots & s_{2k}^- \end{bmatrix} + \begin{bmatrix} s_0 & s_1 & \cdots & 0 \\ s_1 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ s_k & s_{k+1} & \cdots & s_{2k} - s_{2k}^- \end{bmatrix} = \left(s_{2k} - s_{2k}^-\right) det\underline{H}_{2k-2} \tag{2.14}$$

holding for any sequence $(s_1, \ldots, s_n) \in int\mathcal{S}^{[n]}([0,\infty))$. Similar calculations for H_{n-1} allows us to define the z_i's in terms of ratio of Hankel determinants:

$$z_k = \frac{det(\underline{H}_k)det(\underline{H}_{k-3})}{det(\underline{H}_{k-1})det(\underline{H}_{k-2})}. \tag{2.15}$$

We define the Jacobian determinant for ξ as follows:

$$\left|\frac{\partial \boldsymbol{s_n}}{\partial \boldsymbol{z_n}}\right| = \prod_{k=1}^{n}\left|\frac{\partial s_k}{\partial z_k}\right| = \prod_{k=1}^{n}(s_{k-1} - s_{k-1}^{-}) = \prod_{k=2}^{n}(z_1, \ldots, z_{k-1}) = \prod_{k=1}^{n} z_k^{n-k}. \tag{2.16}$$

In the next two sections, in order to facilitate the introduction of the last useful mapping for our reparametrization, it is useful to connect the pseudo-canonical moments of a probability measure on $[0,\infty)$ with a suitable sequence of coefficients related to a symmetric probability measure μ_s in $(-\infty,\infty)$ which is one-to-one with μ. We will now focus on the moment problem for probability measures on $E = (-\infty,\infty)$ and its connection with the theory of orthogonal polynomials.

2.3 Hamburger moment problem

In the Hausdorff moment problem we have introduced one-to-one mappings ρ_n and ρ_n^{-1} between the ordinary moments and the canonical moments. In the Stieltjes moment problem the canonical moments do not exist and they can be replaced by the pseudo canonical moments z_i which allow us to elicit one-to-one mappings ξ_n and ξ_n^{-1} between the ordinary moments and, precisely, these parameters. In the Hamburger moment problem, when we move to the moment space $\mathcal{S}(\mathbb{R})$, neither canonical moments nor pseudo canonical moments can be used because both the inferior extremal moments and the superior extremal moments are not finite. How can we find out a suitable mapping from this moment space into an appropriate unconstrained product space? To answer this question, we need to introduce the concept of orthogonal polynomials. These polynomials have been extensively treated, among others, by Chihara (1978). First, we need to explain the concept of moment functional.

Definition 2.3.1. *For a measure $\mu \in \mathcal{P}(\mathbb{R})$, let us denote by $\mathcal{B}(\mathbb{R},\mu)$ the real-valued μ-integrable Borel functions on $\mathbb{R}$. We say that $\mathcal{L}_\mu$ is a moment functional on $\mathcal{B}(\mathbb{R},\mu)$ mapping the sequence of functions $\boldsymbol{f} = (f_0, f_1, f_2, \ldots)$ with $f_j \in \mathcal{B}(\mathbb{R},\mu)$ to the moment sequence $(s_0, s_1, s_2, \ldots)$ or, equivalently,*

$$\mathcal{L}_\mu(f) = \int_{\mathbb{R}} f(x)\, d\mu$$

if

$$s_j = \int_{\mathbb{R}} f_j(x)\, d\mu \qquad j = 1, 2, \ldots$$

Definition 2.3.2. *A sequence $\{P_n(x)\}_{n=0}^{\infty}$ is called an orthogonal polynomial sequence with respect to a moment functional $\mathcal{L}_\mu$ provided for all non-negative integers m and n,*

(i) $P_n(x)$ *is a polynomial of degree* n,

(ii) $\mathcal{L}_\mu [P_m(x)P_n(x)] = 0$ *for* $m \neq n$,

(iii) $\mathcal{L}_\mu [P_n^2(x)] \neq 0$.

If $\mathcal{L}_\mu [P_n^2(x)]$=1 then the sequence is *orthonormal*. In the general case, conditions (ii) and (iii) can be replaced by:

$$\mathcal{L}_\mu [P_m(x)P_n(x)] = K_n \delta_{m,n} \tag{2.17}$$

where $\delta_{m,n}$ is the Kronecker's delta defined as:

$$\delta_{m,n} = \begin{cases} 0, & if\ m \neq n \\ 1, & if\ m = n \end{cases}. \tag{2.18}$$

Definition 2.3.2 can then be re-formulated as follows:

Theorem 2.3.3. *Let* $\mathcal{L}_\mu$ *be a moment functional and let* $\{P_n(x)\}$ *be a sequence. Then the following assertions are equivalent:*

(a) $\{P_n(x)\}$ *are orthogonal polynomials with respect to* $\mathcal{L}_\mu$

(b) $\mathcal{L}_\mu [\pi_m(x)P_n(x)] = 0$ *for every polynomial* $\pi(x)$ *of degree* $m < n$ *while* $\mathcal{L}_\mu [\pi_m(x)P_n(x)] \neq 0$ *if* $m = n$

(c) $\mathcal{L}_\mu [x^m P_n(x)] = \mathbb{K}_n \delta_{m,n}$ *where* $K_n \neq 0, m = 0, 1, \ldots, n$.

See Chihara (1978), Theorem 2.1 for the proof. The basic idea to demonstrate (b) is that every polynomials of grade m can be written as linear combination of orthogonal polynomials:

$$\pi(x) = \sum_{k=0}^{m} g_k P_k(x), \quad c_m \neq 0 \tag{2.19}$$

and we define g_k as:

$$g_k = \frac{\mathcal{L}_\mu [\pi_m(x)P_n(x)]}{\mathcal{L}_\mu [P_n^2(x)]}. \tag{2.20}$$

We now illustrate another important property of the orthogonal polynomials.

Theorem 2.3.4. *Let* $\{P_n(x)\}$ *be orthogonal polynomials with respect to* $\mathcal{L}_\mu$. *Then, for any polynomial* $\pi_n(x)$ *of degree* n *we have:*

$$\mathcal{L}_\mu [\pi_m(x)P_n(x)] = a_n [x^n P_n(x)] = \frac{a_n k_n \underline{H}_{2n}}{\underline{H}_{2n}} \tag{2.21}$$

where a_n *denotes the leading coefficient of* $\pi_n(x)$ *and* k_n *the leading coefficient of* $P_n(x)$.

To simplify, k_n, the leading coefficient of $P_n(x)$, is set equal to 1 i.e. $P_n(x)$ is monic. For our purposes, the most important property of the orthogonal polynomials is the following.

Theorem 2.3.5. *Let $\{P_n(x)\}$ be the monic orthogonal polynomial sequence with respect to the moment functional $\mathcal{L}_\mu$. Then, there exist a sequence of constants a_n and b_n such that:*

$$P_{n+1}(x) = (x - a_n)P_n(x) - b_n P_{n-1}(x), \quad n = 0, 1, 2, \dots \tag{2.22}$$

where

$$b_i = \frac{\mathcal{L}_\mu\left[x^{i-1}P_{i-1}(x)\right]}{\mathcal{L}_\mu\left[x^{i-2}P_{i-2}(x)\right]} = \frac{det\underline{H}_{2i}det\underline{H}_{2i-4}}{det\underline{H}_{2i-2}det\underline{H}_{2i-2}} \quad for\ i = 1, \dots, \left\lceil \frac{n}{2} \right\rceil$$

$$a_i = \frac{\mathcal{L}_\mu\left[xP_{i-1}^2(x)\right]}{\mathcal{L}_\mu\left[P_{i-2}^2(x)\right]} = \frac{det\underline{H}_{2i-2}det\underline{H}_{2i+1}}{det\underline{H}_{2i}det\underline{H}_{2i-1}} + \frac{det\underline{H}_{2i}det\underline{H}_{2i-3}}{det\underline{H}_{2i-2}det\underline{H}_{2i-1}} \quad for\ i = 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor.$$

The sequence of coefficients $\{a_n, b_n\}$ is called the sequence of recurrence coefficients. If we consider the following integral

$$\int_{\mathbb{R}} x^k P_k(x) d\mu(x)$$

it is clear that the integral can be written as a linear combination of moments of μ; in fact, there are remarkable relations between the moment sequence on μ and the recurrence coefficients, namely:

$$\begin{aligned} \int_{\mathbb{R}} x^k P_k(x) d\mu(x) &= \int_{\mathbb{R}} x^{k-1}(P_{k+1}(x) + a_k P_k(x) + b_k P_{k-1}(x)) d\mu(x) \\ &= b_k \int_{\mathbb{R}} x^{k-1} P_{k-1}(x) d\mu(x) \\ &= b_1 \dots b_k \end{aligned} \tag{2.23}$$

and

$$\begin{aligned} \int_{\mathbb{R}} x^{k+1} P_k(x) d\mu(x) &= \int_{\mathbb{R}} x^k(P_{k+1}(x) + a_k P_k(x) + b_k P_{k-1}(x)) d\mu(x) \\ &= a_k \int_{\mathbb{R}} x^k P_k x\, d\mu(x) + b_k \int_{\mathbb{R}} x^k P_{k-1} x\, d\mu(x) \\ &= b_1 \dots b_k(a_1 + \dots + a_k). \end{aligned}$$

We can formalize the mapping between the recurrence coefficients and the ordinary moments with the following notation:

$$\psi_n : (a_1, b_1 \dots, a_n, b_n) \to (s_1, \dots, s_{2n}).$$

It is proved in Tomecki (2018) that the Jacobian determinant of ψ_n is equal to:

$$detD\psi_n^{\mathbb{R}} = \prod_{i=1}^{n-1} b_i^{2n-2i}. \tag{2.24}$$

Note that in the Jacobian determinant of the mapping ψ_n, only the b_i coefficients appear. On the other hand the a_i coefficients associated to a probability measure on $(-\infty, \infty)$ play a role in characterizing the symmetry of that distribution. In fact, if μ is symmetric around 0, one has $a_j = 0\ \forall j \in \mathbb{N}$. This property will be the key for connecting the pseudo canonical moments of a probability measure in $[0, \infty)$ with a suitable symmetric probability measure in $(-\infty, \infty)$.

2.4 Connections between the Stieltjes moment problem and the Hamburger moment problem

Let us consider a probability measure μ on $[0,\infty)$ characterized by its moment sequence $\boldsymbol{s} \in \mathcal{S}([0,\infty))$ or equivalently by the sequence of pseudo-canonical moments $\boldsymbol{z} \in [0,\infty)^{\infty}$. One can show that there exists a mirror probability measure μ_S on $(-\infty,\infty)$ with the following properties:

1. μ_S is symmetric around 0;
2. if $X \sim \mu$ and $S \sim \mu_S$ then $X \overset{d}{=} S^2$.

In fact, we can determine the probability measure μ_S by taking a random variable $X \sim \mu$ and another random variable $Z \sim Bernoulli(1/2)$ such that the random variable

$$S = \sqrt{X} \cdot (2 \cdot Z - 1) = \begin{cases} -\sqrt{X} & \text{with probability } \frac{1}{2} \\ \sqrt{X} & \text{with probability } \frac{1}{2} \end{cases}$$

enjoys the two aforementioned properties.

Now we can look at the particular case of the Hamburger moment problem corresponding to μ_S called symmetric Hamburger moment problem. In this case we have that all the odd moments of μ_S are equal to zero (see Schmudgen (2017) for more details). This means that the only characterizing sequence of moments for a symmetric distribution is the sequence of even moments

$$E[S^{2j}] = \int_{-\infty}^{\infty} x^{2j} d\mu_S(x).$$

Moreover, since we have that $X \overset{d}{=} S^2$, this allows us to connect every symmetric Hamburger moment problem with a corresponding Stieltjes moment problem and vice versa by noting that

$$\int_0^{\infty} x^{2j} d\mu(x) = E[X^j] = E[S^{2j}] = \int_{-\infty}^{\infty} x^{2j} d\mu_S(x).$$

Let us now analyze the orthogonal polynomials Q_n associated to μ on $[0,\infty) \subset (-\infty,\infty)$. Chihara (1978) argued that the zeros of Q_n lie in $[0,\infty)$ hence there is a corresponding sequence of symmetric orthogonal polynomials $S_n(x)$ (associated to μ_S) such that $S_{2n}(x) = Q_n(x^2)$. These symmetric orthogonal polynomials satisfy the recurrence relation:

$$S_n(x) = xS_{n-1}(x) - \lambda_{n-2}S_{n-2}(x), \tag{2.25}$$

where $a_1 = \lambda_2$ and

$$a_{n+1} = \lambda_{2n+1} + \lambda_{2n+2}, b_{n+1} = \lambda_{2n}\lambda_{2n+1}. \tag{2.26}$$

Chapter 3

Inferential issues and population size estimation with Poisson mixtures

3.1 Model identifiability

Once set up the statistical model represented by the usual triple

$$\mathcal{M} = \{\mathcal{X}, p(x;\theta), \Theta\}$$

before considering using data for making inference, it is worth investigating the identifiability of the underlying parameters. The triple basically determines the family of distributions among which we assume there is the true underlying random mechanism generating the observable data. However, we should point out that identifiability is fundamentally related to the parametrization used for specifying the family of distributions, rather than to the family itself. To quote Joseph B. Kadane, *identification is a property of the likelihood function, and is the same whether considered classically or from the Bayesian approach.* For a formal general definition of identifiability we refer to Basu (1963).

A statistical model $\mathcal{M}$ is said to be identified when for any arbitrarily fixed true parameter $\theta^* \in \Theta$ generating the observed data $X \in \mathcal{X}$ at random according to $p(x;\theta^*)$ one has that there is no other $\theta' \in \Theta$ such that:

$$p(x;\theta^*) = p(x;\theta') \quad \forall x \in \mathcal{X}. \tag{3.1}$$

When there are at least two distinct parameters θ^* and θ' for which (3.1) holds, we also say that the model parameter θ is not identified.
In our original model setup (1.3), for inferring the population size by means of zero-truncated counts, the sampling distribution is specified as $p(\boldsymbol{f}_+;\boldsymbol{\theta})$ and it is defined for any vector $\boldsymbol{f}_+$ in the set of vectors with non-negative integer components $\mathbb{F}_+$ and for any parameter $\boldsymbol{\theta} = (N,\nu) \in \Theta = \mathbb{N} \times \mathcal{P}([0,\infty))$. The first component parameter N lives in the positive integer set $\mathbb{N}$ whereas ν lives in the space of all probability distributions in $(0,\infty)$. We can therefore formally write down the model specification in the original parametrization as follows:

$$\mathcal{M}_{PM}^{orig} = \{\mathbb{F}, p(\boldsymbol{f}_+; N, \nu), \mathbb{N} \times \mathcal{P}([0,\infty))\} \tag{3.2}$$

Indeed, following the arguments illustrated in the introduction, we can consider an alternative model specification based on the moment sequence $\boldsymbol{s} = (s_0, s_1, \ldots)$ of the mixing measure ϕ such that

$$d\phi(\lambda) = e^{-\lambda} d\nu(\lambda).$$

In fact, we have argued that $\phi \in \mathcal{F}_{(0,1]}([0,\infty))$ and the latter subset of finite measures is one-to-one with $\mathcal{S}_{(0,1]}$ via the moment sequence mapping since

$$\mathcal{S}_{(0,1]} = \cup_{t\in(0,1]}\mathcal{S}_t = \cup_{t\in(0,1]} \left\{(t\cdot\widetilde{s}_0, t\cdot\widetilde{s}_1, \ldots, t\cdot\widetilde{s}_j, ...) : \widetilde{\boldsymbol{s}} = (\widetilde{s}_0, \widetilde{s}_1, ...) \in \mathcal{S}_1\right\}.$$

Hence, the moment based model specification in terms of the alternative parameter $\boldsymbol{\vartheta} = (N, \boldsymbol{s})$ is the following:

$$\mathcal{M}_{PM}^{mom} = \left\{\mathbb{F}; p(\boldsymbol{f}; N, \boldsymbol{s}), \mathbb{N}\times\mathcal{S}_{(0,1]}\right\} \tag{3.3}$$

where

$$p(\boldsymbol{f}_+; N, \boldsymbol{s}) = p(\boldsymbol{f}_+; \boldsymbol{\vartheta}) = \binom{N}{f_0\ f_1\ f_2\ \ldots} \prod_{j=0}^{\infty}\left(\frac{s_j}{j!}\right)^{f_j} = \binom{N}{f_0\ f_1\ f_2\ \ldots f_M} \prod_{j=0}^{M}\left(\frac{s_j}{j!}\right)^{f_j}$$

follows from (1.3) and the fact that P_j can be re-expressed as follows:

$$P_j = \int_0^\infty e^{-\lambda}\frac{\lambda^j}{j!}d\nu(\lambda) = \int_0^\infty \frac{\lambda^j}{j!}d\phi(\lambda) = \frac{s_j}{j!} \tag{3.4}$$

with s_j being the $j-th$ moment of the measure ϕ. It is worth noting that the measure ϕ is not a probability measure but it is a finite measure with total mass in the unit interval

$$\phi([0,\infty)) = \int_0^\infty d\phi(\lambda) = \int_0^\infty e^{-\lambda}d\nu(\lambda) = P(X_i = 0) = P_0 = s_0\text{`} \in [0,1).$$

We can therefore consider the normalized measure $\widetilde{\phi}(\lambda) = \frac{\phi(\lambda)}{\int_0^\infty d\phi(\lambda)} = \frac{\phi(\lambda)}{s_0}$ which integrates to 1. Hence, the probabilities P_j can also be re-expressed as function of the moments of new probability measure $\widetilde{\phi}(\cdot)$

$$P_j = \int_0^\infty e^{-\lambda}\frac{\lambda^j}{j!}d\nu(\lambda) = \int_0^\infty e^{-\lambda}\frac{\lambda^j}{j!}d\nu(\lambda) = \int_0^\infty \frac{\lambda^j}{j!}d\phi(\lambda) = \frac{s_j}{j!} = s_0\frac{\widetilde{s}_j}{j!} \tag{3.5}$$

where $\widetilde{s}_j = j!\frac{P_j}{P_0}$ is the j-th moment of $\widetilde{\phi}$ i.e. the generic component of a moment sequence for the Stieltjes moment problem.

We can then argue that $\widetilde{\phi}$ admits the existence of a moment generating function

$$M(t) = \int_0^\infty e^{t\lambda}d\widetilde{\phi}(\lambda)$$

since the following integral

$$\int_0^\infty e^{\lambda}d\widetilde{\phi}(\lambda) = \int_0^\infty e^{\lambda}\frac{d\phi(\lambda)}{P_0} = \frac{1}{P_0}\int_0^\infty e^{\lambda}e^{-\lambda}d\mu(\lambda) = \frac{1}{P_0}\int_0^\infty d\mu(\lambda) = \frac{1}{P_0} < \infty$$

corresponds to the moment generating function $M_{\widetilde{\phi}}(\cdot)$ evaluated in $t = 1$ and $M_{\widetilde{\phi}}(t) \leq M_{\widetilde{\phi}}(1) < \infty$ for all $t \leq 1$. Therefore $\widetilde{\phi}$ is determined by the moment generating function of $\widetilde{\phi}$ or, equivalently, by the sequence of all the moments of non negative integer order of $\widetilde{\phi}$ which exist and are finite.

Moreover, the following holds '

$$\int_0^\infty e^\lambda d\widetilde{\phi}(\lambda) = \int_0^\infty \sum_{j=0}^\infty \frac{\lambda^j}{j!} d\widetilde{\phi}(\lambda) = \sum_{j=0}^\infty \frac{\widetilde{s}_j}{j!} = \frac{1}{P_0}. \tag{3.6}$$

.

Theorem 3.1.1. *The parametric model*

$$\mathcal{M}_{PM}^{mom} = \left\{ \mathbb{F}_+; p(\boldsymbol{f}_+; N, \boldsymbol{s}), \mathbb{N} \times \mathcal{S}_{(0,1]} \right\}$$

is identified.

Proof. We need to show that for $\vartheta = (N, \boldsymbol{s}) \neq \vartheta' = (N', \boldsymbol{s}')$ in the parameter space $\mathbb{N} \times \mathcal{S}_{(0,1]}$, ice of $\boldsymbol{f}_+ \in \mathbb{F}_+$, one cannot have that:

$$p(\boldsymbol{f}_+; \boldsymbol{\vartheta}) = p(\boldsymbol{f}_+; \boldsymbol{\vartheta}') \quad \forall\ \boldsymbol{f}_+ \in \mathbb{F}_+.$$

We consider two different cases: $N \neq N'$ (w.l.o.g. $N > N'$) or $N = N'$ and $\boldsymbol{s} \neq \boldsymbol{s}'$. If $N > N'$, the sampling distributions corresponding to the two different parameters have different support; therefore, by taking $\boldsymbol{f}_+ = (f_1, \ldots, f_M)$ such that $\sum_{j=1}^M f_j = n = N$ and hence $f_0 = 0$, we have that

$$p(\boldsymbol{f}_+; \boldsymbol{\vartheta}) > 0$$

while

$$p(\boldsymbol{f}_+; \boldsymbol{\vartheta}') = 0.$$

Let us consider the other case when $N = N'$ but $\boldsymbol{s} \neq \boldsymbol{s}'$.There must exist a $j^* \geq 1$ such that $s_{j^*} \neq s_j$. We can consider the frequencies of frequencies corresponding to all observed units being counted j^* times so that $\boldsymbol{f}_+$ is such that $f_0 = N - n$, $0 < f_{j*} = n \leq N$ and $f_j = 0$ for all $j \in \mathbb{N}$ and $j \neq j^*$. In this case

$$\frac{p(\boldsymbol{f}_+; \boldsymbol{\vartheta})}{p(\boldsymbol{f}_+; \boldsymbol{\vartheta}')} = \left(\frac{s_{j^*}}{s_j} \right)^n \neq 1.$$

□

Note that it is easy to argue that, in the model specification $\mathcal{M}_{PM}^{mom}$ in (3.3), the infinite-dimensional moment component equates to a linear (affine) reparametrization of the most natural count distribution probabilities corresponding to the sequence of Poisson mixture probabilities P_j so that (3.3) reparametrizes the more standard model specification

$$\mathcal{M}_{PM}^{pmp} = \left\{ \mathbb{F}_+; p(\boldsymbol{f}_+; N, \boldsymbol{P}), \mathbb{N} \times \mathcal{P}_{PoisMix} \right\}. \tag{3.7}$$

3.2 Some non identifiability issues in the conditional likelihood

Starting from the pioneering work of Sanathanan (1972), there has been always interest in simplifying the inferential task of estimating the population size by means of the conditional likelihood (1.7) due to the asymptotically minor contribution of the residual likelihood. However, similarly to what has been raised by Link (2003) and Farcomeni and Tardella (2012) in the binomial mixture settings, an annoying non identifiability issue affects the conditional likelihood if one aims at estimating P_0 by maximizing (1.7).

To understand this issue, one can argue that for any mixing distribution ν with corresponding $P_0 < 1$, there is an infinite collection of other mixing distributions $\nu_\epsilon(A) = (1-\epsilon)\nu(A) + \epsilon\delta_0(A)$ for any $\epsilon \in (0,1)$ for which $L_c(\nu; \boldsymbol{f}_+) = L_c(\nu_\epsilon; \boldsymbol{f}_+)$, where δ_0 is the degenerate distribution at 0. We can write $P(X_i = j; \nu_\epsilon)$ as follows:

$$P'(X_i = j; \nu) = P'_j = (1-\epsilon)\int_0^\infty e^{-\lambda}\frac{\lambda^j}{j!}d\nu(\lambda).$$

P_0 under the same mixing distribution ν_ϵ is:

$$P'(X_i = 0; \nu) = P'_0 = \epsilon + (1-\epsilon)\int_0^\infty e^{-\lambda}d\nu(\lambda)$$

and therefore the conditional probability is equal to:

$$\frac{P'_j}{1-P'_0} = \frac{(1-\epsilon)P_j}{(1-\epsilon)(1-P_0)} = \frac{P_j}{1-P_0}.$$

This means that there is structural non-uniqueness and potential degeneracy of the conditional maximum likelihood estimation of P_0 and N as highlighted in Mao and Lindsay (2007) and in Wang (2010). In fact, if we denote by $\widehat{\nu}^C$ the conditional maximum likelihood estimate we can argue that any other $\widehat{\nu}^C_\epsilon$ maximizes the conditional likelihood and provides two different conditional MLE for N

$$\hat{N}_C = \frac{n}{1-P_0(\hat{\nu})} \quad \hat{N}_{C,\epsilon} = \frac{n}{(1-\epsilon)(1-P_0(\hat{\nu}))}$$

so that with an arbitrarily large value of ϵ one can get an unbounded conditional MLE estimate of N. The same identifiability issue also holds when we restrict the attention to a reduced statistical model $\mathcal{M}^{mom}_{PM}$ involving only the first M moments of the mixing measure ϕ associated to ν for which there is information in the original likelihood. In fact, if we rewrite the factorization

$$\begin{aligned} L(N, \boldsymbol{s}; \boldsymbol{f}_+) &= p(\boldsymbol{f}_+; N, \boldsymbol{s}) = \binom{N}{f_0\ f_1\ f_2\ \dots f_M}\prod_{i=0}^{M}(j!s_j)^{f_j} \\ &= \binom{n}{f_1, f_2, \dots f_M}\prod_{j=1}^{M}\left(\frac{j!s_j}{1-s_0}\right)^{f_j} \times \binom{N}{f_0}{s_0}^{N-n}(1-s_0)^n \\ &= L_C(\boldsymbol{s}; \boldsymbol{f}_+) \times L_R(N, \boldsymbol{s}; \boldsymbol{f}_+) = L_C(\boldsymbol{s}; \boldsymbol{f}_+) \times L_R(N, s_0; \boldsymbol{f}_+) \end{aligned}$$

the observational equivalence of ν and ν_ϵ with respect to the conditional likelihood $L_C(\boldsymbol{s}; \boldsymbol{f}_+)$ is turned into the observational equivalence of the corresponding first M moments of the corresponding measures ϕ and ϕ_ϵ, due to the proportionality

$$\boldsymbol{s}(\phi_\epsilon) = (1-\epsilon)\boldsymbol{s}(\phi)$$

which yields the equality

$$\frac{\boldsymbol{s}(\phi_\epsilon)}{1-s_0(\phi_\epsilon)} = \frac{(1-\epsilon)\boldsymbol{s}(\phi)}{(1-\epsilon)(1-s_0)}.$$

Moreover, since the total masses of ϕ and ϕ_ϵ are ,respectively, s_0 and ϵs_0 one can also argue the equality of the moments of the corresponding normalized probability measures $\widetilde{\phi}$ and $\widetilde{\phi_\epsilon}$.

Although we have argued that the unconditional likelihood associated to model specification (3.3) overcomes the identifiability problem for a finite population size N there is still some interest in relying on the unidentifiable conditional likelihood model to define a quantity associated with $\mathcal{C}(\boldsymbol{P}^*_M)$ the equivalence class of parameters corresponding to the same conditional likelihood probabilities $\boldsymbol{P}^*_M = \overline{P}^*_1, ..., \overline{P}^*_M$. More formally

$$\begin{aligned} \mathcal{C}(\boldsymbol{P}^*_M) &= \left\{ \boldsymbol{s}_M \in \mathcal{S}^{[M]}_{(0,1]} : \overline{P}(\boldsymbol{s}_M) = \boldsymbol{P}^*_M, \quad j = 1, \ldots, M \right\} \qquad (3.8) \\ &= \left\{ \boldsymbol{s}_M = (s_0, \ldots, s_M) \in \mathcal{S}^{[M]}_{(0,1]} : \frac{s_j}{j!1-s_0} = \overline{P}_j, \quad j = 1, \ldots, M \right\}. \end{aligned}$$

3.3 Model approximation

Our main goal consists in making inference on ν and, consequently, on P_0 and N. As seen in the previous paragraphs, by reparametrizing the mixing distribution ν using $d\phi(\lambda) = e^{-\lambda} d\nu(\lambda)$, one gets a one-to-one correspondence between $(P_0, P_1, P_2, \ldots)$ and the moment sequence $(s_0, s_1, s_2, \ldots)$ of a finite measure ϕ with total mass $s_0 = \int_0^\infty d\phi(\lambda) \in (0, 1]$ uniquely identified by its moment sequence with $s_j = \int_0^\infty \lambda^j d\phi(\lambda)$ for $j = 0, 1, 2, ...$ and by its moment generating function

$$M(t) = \int_0^\infty e^t \lambda d\phi(\lambda)$$

that is, for $t = 1$, equal to:

$$M(1) = \int_0^\infty e^\lambda d\phi(\lambda) = 1.$$

In turn, the sequence $(s_0, s_1, s_2 \ldots)$ is one to one with the sequence

$$(s_0, \widetilde{s}_1 = \frac{s_1}{s_0}, \widetilde{s}_2 = \frac{s_2}{s_0}, \ldots).$$

Therefore, the unconditional likelihood corresponding to this parametrization can be reformulated as follows:

$$L(N, (s_0, \widetilde{s}_1, \widetilde{s}_2, \ldots); \boldsymbol{f}_+) \propto \binom{N}{n} \prod_{j=0}^{\infty} \left(\frac{\widetilde{s}_j}{j!} s_0 \right)^{f_j} \qquad (3.9)$$

where $\widetilde{s}_j = \int_0^\infty \lambda^j d\widetilde{\phi}(\lambda)$ for $j = 0, 1, 2, \ldots$ is the generic jth moment of the probability measure $\widetilde{\phi}(\lambda) = \frac{\phi(\lambda)}{\int_0^\infty d\phi(\lambda)}$.
Note however that the components of this reparametrization are constrained to the following series equality:

$$\sum_{j=0}^{\infty} \frac{\widetilde{s}_j}{j!} = \sum_{j=0}^{\infty} \frac{s_j}{j! s_0} = \frac{1}{s_0} = \frac{1}{P_0}. \tag{3.10}$$

Equivalently, one can write $s_0 = \frac{1}{\sum_{j=0}^{\infty} \frac{\widetilde{s}_j}{j!}}$; this constraint does not allow us to directly exploit the alternative reparametrizations described before.
However, we will rely on the finite sum approximation of the sampling distribution

$$\widetilde{p}(\boldsymbol{f}_+; N, \widetilde{\boldsymbol{s}}_T) = \binom{N}{f_0\ f_1\ f_2 \ldots,\ f_T} \prod_{j=0}^{T} \left[\frac{\widetilde{s}_j}{j! \sum_{k=0}^{T} \frac{\widetilde{s}_k}{k!}} \right]^{f_j}.$$

This approximation is well grounded by the fact that $\widetilde{p}(\boldsymbol{f}_+; N, \widetilde{\boldsymbol{s}}_T)$ defines a rigorous and flexible sampling distribution. If we consider the corresponding likelihood function

$$\widetilde{L}_T(N, \widetilde{\boldsymbol{s}}_T; \boldsymbol{f}_+) = \widetilde{p}(\boldsymbol{f}_+; N, \widetilde{\boldsymbol{s}}_T)$$

we have that, for $T \to \infty$, the approximating likelihood $\widetilde{L}_T(N, \widetilde{\boldsymbol{s}}_T; \boldsymbol{f}_+)$ achieves the original $L(N, \widetilde{\boldsymbol{s}}; \boldsymbol{f}_+)$. In this thesis we propose to use the following statistical model

$$\widetilde{\mathcal{M}}_{PM}^{appr(T)} = \left\{ \mathbb{F}_T +; \widetilde{p}(\boldsymbol{f}_+; N, \widetilde{\boldsymbol{s}}), \mathbb{N} \times \mathcal{S}_1^{[T]} \right\} \tag{3.11}$$

as a surrogate model for $\mathcal{M}_{PM}^{mom}$ in (3.3); the likelihood function can then be re-expressed as follows:

$$\widetilde{L}_T(N, \widetilde{\boldsymbol{s}}_T; \boldsymbol{f}_+) = \widetilde{p}(\boldsymbol{f}_+; N, \widetilde{\boldsymbol{s}}_T) = \binom{N}{f_0\ f_1 \ldots f_T} \prod_{j=0}^{T} \left[\frac{\widetilde{s}_j}{j! \sum_{k=0}^{T} \frac{\widetilde{s}_k}{k!}} \right]^{f_j}. \tag{3.12}$$

We point out that the approximation depends on the choice of T. Suitable choice for T can be $T = M$ but it can also be $T < M$. Moreover, if $T < M$, this model overlooks the information coming from the observed frequencies $f_j > 0$ with $j > T$. We note that the choice $T > M$ should be avoided since the observed $\boldsymbol{f}_+$ contains no information on moments s_j with $j > M$. In fact, changing the values of moments s_j with $j > M$ and leaving s_r with $r <= M$ fixed, the likelihood function does not change. A similar approximation has been put forward by Alunni Fegatelli and Tardella (2018) in a Bayesian framework setting. However, in their proposal there is a further approximation building block which limits the original mixing measure ν to have a bounded support. Now we need to discuss in the following section how one can attempt to estimate model parameters by using suitable reparametrizations of the underlying truncated moment space $\mathcal{S}_1^{[T]}$.

3.4 Useful reparametrizations for model estimation

The above reformulation of the approximating likelihood (3.12) suggests us that the most general inferential problem of Poisson mixtures in a so-called non-parametric setting reduces in fact to a finite-dimensional inference. So far we have explicitly seen that the problem can be parametrized either in terms of the marginal Poisson mixture probabilities $P \in \mathcal{P}_{PoisMix}$ where the range of the Poisson mixture probabilities is a convex subset of the T-dimensional simplex Δ_T (1.6) or in terms of the first T moments $(\widetilde{s}_1, \ldots, \widetilde{s}_T)$ of the mixing distribution $\widetilde{\phi}$ whose range is in the truncated moment space (1.18). One notes that they are constrained T-dimensional subsets namely convex bodies which are one-to-one linearly related through $P_j = \frac{\widetilde{s}_j}{j!\left(\sum_{k=0}^{T} \frac{\widetilde{s}_k}{k!}\right)}$, hence the space of the Poisson mixture probabilities (1.6) can be reinterpreted as an image through an affine map of the convex body of the T-truncated moment space $\mathcal{S}_1^{[T]}$. The inferential problem seems to be well-posed so one could just think to use standard maximization routines to get an estimate for P_0 and therefore for N. Computational aspects would be a rather easy obstacle to overcome if the truncated moment space involved in the parametrization of the likelihood function were an easy-to-handle unrestricted space. Unfortunately this is not the case. Luckily, there is another less known equivalent saturated parametrization of the first T moments which turns out to map either $P_{PoisMix}$ or $\mathcal{S}_1^{[T]}$ onto a convenient product space. This parametrization is expressed in terms of the so-called recurrence coefficients (or, equivalently, pseudo canonical moments) which are quantities in $[0, \infty)$, otherwise unconstrained, which can be mapped back and forth from the ordinary moments in terms of either the QD algorithm or the Chebyschev algorithm, both involving only summations and multiplications.

.The two equivalent reparametrized models can be denoted as follows:

$$\widetilde{\mathcal{M}}_{PM}^{pcm} = \{\mathbb{F}_+; \widetilde{p}(\boldsymbol{f}_+; N, \boldsymbol{z}_M), \mathbb{N} \times [0, \infty)^\infty\} \tag{3.13}$$

and

$$\widetilde{\mathcal{M}}_{PM}^{rec} = \left\{\mathbb{F}_+; \widetilde{p}(\boldsymbol{f}_+; N, \boldsymbol{a}_{M/2}, \boldsymbol{b}_{M/2-1}), \mathbb{N} \times [0, \infty)^\infty\right\}. \tag{3.14}$$

To our knowledge, there is no other attempt in the literature to use these recurrence coefficients for inferential purposes for Poisson mixtures models. Few authors have instead implemented inferential procedures using canonical moments. Tardella (2002) used them in a particular case of binomial mixture problems related to the estimation of a capture-recapture model. Alunni Fegatelli and Tardella (2018) utilized canonical moments in a context similar to ours (Poisson mixtures) but in a Bayesian settings. Canonical moments, being quantities in the unit interval, are easier to deal with and their initialization for numerical search of the MLE is straightforward and does not constitute an issue. The recurrence coefficients are directly related with the pseudo-canonical moments z_i since $b_i = z_{2i-1} z_{2i}$ and $a_i = z_{2i-1} + z_{2i-2}$. One notes that the extremal inferior moment of second order is $\widetilde{s}_2^- = \widetilde{s}_1 \widetilde{s}_0^{-1} \widetilde{s}_1$ and since $\widetilde{s}_0 = 1$ we have that $\widetilde{s}_2^- = \widetilde{s}_1^2$. So the inferior extremal moment of second order is equal to the square of the expected value of $\widetilde{\phi}$ (with

$\widetilde{s}_1^- = 0$). We remind that z_j is defined as

$$z_j = \frac{s_j - s_j^-}{s_{j-1} - s_{j-1}^-}$$

Hence, $z_1 = \widetilde{s}_1$. Then, if we consider a random variable $X \sim \widetilde{\phi}$, we can derive an interesting relation:

$$z_2 = \frac{s_2 - s_2^-}{s_1 - s_1^-} \rightarrow z_1 z_2 = s_2 - s_2^- = \mathbb{E}_{\widetilde{\phi}}[X^2] - \mathbb{E}^2_{\widetilde{\phi}}[X] = \mathbb{V}_{\widetilde{\phi}}[X]\,. \tag{3.15}$$

Therefore, b_1 can be regarded as the variance of $\widetilde{\phi}$ and a_1 as its expected value. Interpreting recurrent coefficients of higher order is complicated therefore, to overcome initialization issues, we will adopt three strategies:

- compactification of the mixing distribution $\widetilde{\phi}$ to get recurrence coefficients restricted to a compact support $[0, u]$,
- automatic initialization routine based on the probability ratios,
- usage of pseudo-canonical moments (z parameters) instead of the recurrence coefficients.

We will exploit the first two strategies in the next two paragraphs.

3.4.1 Mixing distribution compactification

Let's begin by showing that our model can be approximated arbitrarily well by a model in which the mixing distribution ν has a compact support in $[0, u]$ for a suitable choice of u.

Theorem 3.4.1. *Let ν be a generic probability distribution with support on $[0, \infty)$;* $\forall\ \eta > 0\ \exists\ u_{\eta,\nu}$ *such that*

$$d_{TV}(P_j(\nu), P_j(\nu_{u_{\eta,\nu}})) \leq \eta, \quad \forall j$$

Proof can be found in Alunni Fegatelli and Tardella (2018). We can then define $d\phi_u(\cdot) = e^{\lambda} d\nu_u(\cdot)$ so that:

$$P_j(\nu_u) = \int_0^u \frac{e^{-\lambda}\lambda^k}{k!}\, d\nu_u(\lambda) = \frac{1}{k!}\int_0^u \lambda^k\, d\phi_u(\lambda) = \frac{s_k(\phi_u)}{k!}.$$

One can then reparametrize the approximated likelihood (3.12) as follows:

$$\widetilde{L}_T(N, \widetilde{\boldsymbol{s}}_{\boldsymbol{u}}; \boldsymbol{f}_+) = \binom{N}{f_0\ f_1 \ldots f_T} \prod_{j=0}^{T} \left[\frac{\widetilde{s}_j(\widetilde{\phi}_u)}{j! \left(\sum_{k=0}^{T} \frac{\widetilde{s}_k(\widetilde{\phi}_u)}{k!} \right)} \right]^{f_j}.$$

We can make a further simplification by separating the dependence of $\widetilde{s}_j(\widetilde{\phi}_u)$ from u and the moments of a single probability distribution $\widetilde{\phi}_1$ supported on $[0,1]$, namely:

$$\widetilde{s}_j(\widetilde{\phi}_u) = u^j \widetilde{s}_j(\widetilde{\phi}_1)$$

which corresponds to the change of measure for $\widetilde{\phi}_u$ due to a scale factor u for the rate parameter λ. For ease of notation, let us denote $\widetilde{s}_j(\widetilde{\phi}_1)$ with $\widetilde{\widetilde{s}}_j$: $(\widetilde{\widetilde{s}}_1, \dots, \widetilde{\widetilde{s}}_T)$ will then be the vector of first T moments of the new mixing distribution $\widetilde{\phi}_1$. The approximated likelihood can then be re-expressed as follows:

$$\widetilde{L}(N, \widetilde{\widetilde{\boldsymbol{s}}}_T; \boldsymbol{f}_+) = \binom{N}{f_0,\ f_1, \dots f_T} \prod_{j=0}^{T} \left[\frac{u^j \widetilde{\widetilde{s}}_j}{j! \left(\sum_{k=0}^{T} \frac{u^k \widetilde{\widetilde{s}}_k}{k!} \right)} \right]^{f_j} \tag{3.16}$$

where the T-truncated moment space $\mathcal{S}_1^{[T]}$ is such that:

$$\mathcal{S}_1^{[T]}([0,1]) = \left\{ (s_0, \widetilde{\widetilde{s}}_1, \dots, \widetilde{\widetilde{s}}_T) : \widetilde{\widetilde{s}}_k = \int_0^1 \lambda^k d\widetilde{\phi}_1(\lambda), \widetilde{\phi}_1 \in \mathcal{P}([0,1]) \right\} \tag{3.17}$$

with $\mathcal{P}([0,1])$ being the class of probability distributions with support in $[0,1]$. At this point, one can exploit the results reviewed in chapter 2 and reparametrize the T-truncated sequence in terms of the recurrence coefficients $(a_{T/2}) \in [0,1]^{T/2}$ and $(b_{T/2-1}) \in [0,1]^{T/2-1}$, initialize the maximization routine by choosing any value between $[0,1]$ for these coefficients (and for the upper bound u) and hence reparametrize back into the space of ordinary moments.

3.4.2 Recurrence coefficients initialization based on ratio regression

In the context of estimation of unobserved units of a population, the maximum likelihood estimation often appears to be prone to numerical problems, even when reparametrized moments are used. Moreover, numerical issues can be favored by unsuitable starting values of numerical maximization routines. The situation might be improved if convenient starting values of the parameters are provided for the numerical search of the ML estimates. To do so, we will take advantage of suitable sequences of probability ratios

$$r_j = (j+1)E(f_{j+1})/E(f_j) = (j+1)P_{j+1}/P_j$$

. We recall that the marginal distribution of power series mixtures satisfies the following monotonicity property $r_j \leq r_{j+1}$ with $r_j = (j+1)E(f_{j+1})/E(f_j)$. Both Rocchetti et al. (2011) and Böhning et al. (2016) proposed a ratio regression approach to identify an appropriate distributional form without the need to parametrically specify the mixing density ν. In particular Rocchetti et al. (2011) re-expressed the probability ratio as follows:

$$(x+1)E(f_{x+1})/E(f_x) = \gamma + \delta x \tag{3.18}$$

where γ and δ are real constants. This structure characterizes the Katz family of distributions first introduced in Johnson et al. (2005). Rocchetti et al. (2011) highlight that the non-decreasing linear relation characterizes the following special cases (with $\gamma > 0$ and $delta < 1$ otherwise no probability distribution can be defined):

- $0 < \delta < 1$ corresponds to negative binomial distributions,
- $\delta = 0$ corresponds to Poisson distributions .

It is then clear that δ regulates the heterogeneity of the distribution. In the paper no reference is made to the Poisson mixtures even though, due to the monotonicity of the ratio, the linear relation above can be regarded as a first-order linear approximation for any Poisson mixture (not just gamma), thus justifying a degree of robustness of the method across a wide range of heterogeneity models. Recalling that $P_j = s_j/j!$ (where $s_j/j!$ is the j-th moment of ϕ), we can then re-express the ratio as function of the moments and get the following expression:

$$r_j = \frac{(j+1)s_{j+1}/(j+1)!}{s_j/j!} = \frac{s_{j+1}}{s_j} = \frac{\widetilde{s}_{j+1}}{\widetilde{s}_j} \tag{3.19}$$

where $\widetilde{s}_j = s_j/s_0$. At this point, we can just substitute the moment ratio in the left-hand side of the linear relation and obtain:

$$\widetilde{s}_{x+1}/\widetilde{s}_x = \gamma + \delta x. \tag{3.20}$$

It is then straightforward to show that $\widetilde{s}_j = \prod_{i=1}^{j-1} \gamma + \delta i$. Following the approach of Rocchetti et al. (2011), γ and δ can be estimated by considering the linear regression model

$$r_x = \gamma + \delta x + \epsilon$$

so that

$$\hat{s}_j = \prod_{i=1}^{j-1} (\hat{\gamma} + \hat{\delta} i).$$

Indeed, this approach can be generalized and possibly improved ,by replacing the standard linear regression by means of a log linear regression or any other isotonic regression function. Finally, we recall that there exists a one-to-one mapping ψ_n between recurrence coefficients and ordinary moments, namely

$$\psi_n : (a_1, b_1 \ldots, a_n, b_n) \to (s_1, \ldots, s_{2n}) .$$

One can then use the inverse mapping ψ_n^{-1} to obtain a suitable initialization of the recurrence coefficients from the estimated ordinary moments:

$$\psi_n^{-1}(\hat{s}_1, \ldots, \hat{s}_n) = (\hat{a}_1, \hat{b}_1, \ldots, \hat{a}_{n/2}, \hat{b}_{n/2-1}).$$

Chapter 4

Sharpest lower bounds for estimating the population size

Despite the non-identifiability of the conditional likelihood (1.7), Mao (2006) and Mao and Lindsay (2007) showed that it can still be of interest to draw inference on a non trivial, uniquely determined, sharpest lower bound of $P_0(\nu)$, or equivalently, N. To emphasize the importance of the last sentence, we must quote Irving John Good, one of the first to tackle this problem (Good, 1953) *"I don't believe it is usually possible to estimate the number of species, but only an appropriate lower bound to that number. This is because there is nearly always a good chance that there are a very large number of extremely rare species"*(Bunge and Fitzpatrick, 1993). As pointed out in the previous chapters, the count distribution conditionally on all the $X_i > 0$ and n, follows a conditional zero truncated Poisson mixture distribution with the following probability masses

$$\overline{P}_j = \frac{P_j}{1-P_0} = \int_0^\infty e^{-\lambda}\frac{\lambda^j}{j!}\frac{d\nu(\lambda)}{1-P_0} = \frac{s_j}{j!(1-s_0)} \qquad j = 1, 2, \ldots$$

Alternatively, we could instead follow Mao and Lindsay (2007) and define a probability measure γ as $d\gamma(\lambda) = \frac{(1-e^{-\lambda})d\nu(\lambda)}{\int_0^\infty (1-e^{-u})d\nu(u)}$. The conditional probabilities can then be re-expressed as follows:

$$\overline{P}_j = \int_0^\infty \frac{\lambda^j}{j!(e^\lambda - 1)} d\gamma(\lambda) \qquad j = 1, 2, \ldots.$$

They will find it useful to consider the odds that a unit is unseen in the sample, that is

$$\varphi = \varphi(\gamma) = \int_0^\infty (e^\lambda - 1)^{-1} d\gamma(\lambda).$$

Mao and Lindsay (2007) derived some interesting properties for the odds φ :

- $\varphi(\cdot)$ is Hellinger discontinuous at any γ;
- $\varphi(\cdot)$ is Kolmogorov lower-semicontinuous at any γ;
- φ has no locally unbiased and locally informative estimator;

- φ has no locally asymptotically unbiased and locally asymptotically informative estimator.

However, from its lower-semicontinuity, the odds φ could be partially inferred in terms of a suitable lower bound for φ once a consistent sequence of estimators of the conditional probabilities is used. Note that φ is a strictly increasing one-to-one function of P_0 and hence we prefer to focus on the corresponding lower bound for P_0. This is usually referred to as the sharpest lower bound corresponding to a fixed value of the conditional probabilities $\overline{\boldsymbol{P}}^*_M = (\overline{P}^*_1, \ldots, \overline{P}^*_M)$ with $\overline{P}^*_j = \frac{P_j(\boldsymbol{s}_M)^*}{1-P_0(\boldsymbol{s}_M)^*}$ for $j = 1, \ldots, M$.

Definition 4.0.1. *For any fixed value of the conditional probabilities $\overline{\boldsymbol{P}}^*_M$ corresponding to the conditional likelihood* (1.7) *we define the sharpest lower bound for P_0 arising from the equivalence class $\mathcal{C}(\overline{\boldsymbol{P}}^*_M)$ in* (3.8) *as follows*

$$\tau^*_M = \tau^*_M(\overline{\boldsymbol{P}}^*_M) = \inf_{\boldsymbol{s}_M \in \mathcal{C}(\overline{\boldsymbol{P}}^*_M)} P_0(\boldsymbol{s}_M). \tag{4.1}$$

From the above mentioned relation, the corresponding sharpest lower bound for φ is $\varphi^*_M = \frac{\tau^*_M}{1+\tau^*_M}$.

As stated by Mao and Lindsay (2007), the conditional likelihood approach consistently estimates τ^*_M with the conditional MLE estimator

$$\hat{\tau}_C = inf \left\{ P_0(\boldsymbol{s}') : \boldsymbol{s}' = \arg\max_{\boldsymbol{s} \in \mathcal{S}_I^{[M]}} L_c(\boldsymbol{s}_M; \boldsymbol{f}_+) \right\}. \tag{4.2}$$

There is not a unique best and safe way to numerically compute the sharpest lower bound (4.2) corresponding to any truncated sequence $\overline{\boldsymbol{P}}_M$ or, equivalently, $\boldsymbol{s}_M$. In the next subsections we are going to present four methods to approach its evaluation:

A algebraic lower bound evaluation;

B quadrature lower bound evaluation;

C lower bound evaluation through sequential moment condition check;

D likelihood-based lower bound evaluation.

The first method is introduced in Mao (2006) and the second has been recently proposed in Daley and Smith (2016); the third method, to the best of our knowledge, is new and represents an original contribution. The fourth method is basically a way of exploiting the observed data $f_1, ..., f_M$ and some inequality relations between the conditional and unconditional MLE of P_0 for estimating from below the sharpest lower bound of P_0 corresponding to the true unknown mixing ν.

4.1 Algebraic lower bound approach

In Mao and Lindsay (2007) it has been proposed a moment based approach to derive lower bounds for φ: this method yields what we refer to as algebraic lower bound for the odds φ. Of course, a lower bound for the odds φ can be easily transformed to become a lower bound for $P_0 = \frac{\varphi}{1+\varphi}$.

Let $o_j = \int_0^\infty \lambda^j d\omega(\lambda)$ be the j-th moment of finite a measure ω over $(0, \infty)$ with $d\omega(\lambda) = (e^\lambda - 1)^{-1} d\gamma(\lambda)$. Note that, for $j = 0$, o_0 is the total mass of ω.

Proposition 4.1.1. *The moment sequence* $\boldsymbol{o} = (o_0, o_1, o_2, \ldots)$ *satisfies:*

- $o_0 = \varphi$
- $o_j = j!\overline{P}_j$

From the above proposition, we consider the problem of determining the sharpest lower bound to the total mass $o_0 = \varphi$ corresponding to a finite measure with fixed truncated higher order moment sequence, an issue in the scope of the truncated Stieltjes moment problem.

Hence, one can also argue that

$$\tau_M^*(\overline{\boldsymbol{P}}_M^*) = \inf_{\boldsymbol{s}_M \in \mathcal{C}(\overline{\boldsymbol{P}}_M)} P_0(\boldsymbol{s}_M) = \inf_{\boldsymbol{o}_M \in \mathcal{S}_I^{[M]} | o_1 = \overline{P}_1, o_2 = 2\overline{P}_2, \ldots, o_M = M!\overline{P}_M} \frac{o_0}{1 + o_0}. \quad (4.3)$$

If we now focus on the odds and the moments of ω, we can find quite easily one of the most trivial and least stringent lower bound for φ from the first two moments and the Cauchy-Schwartz inequality. In fact, the following holds

$$o_0 o_2 \geq o_1^2 \Longrightarrow \varphi = o_0 \geq o_1^2/o_2$$

and yields the famous Chao lower bound (Chao, 1984).

The most remarkable result in Mao and Lindsay (2007) states that when the Hankel matrix of the truncated moment sequence $\boldsymbol{o}_k$ is positive definite one gets that:

$$\varphi_k = \boldsymbol{o}_k^t A_k^{-1} \boldsymbol{o}_k \quad (4.4)$$

where A_k is a $k \times k$ matrix with generic entry o_{i+j} corresponding to row i and column j for $i, j = 1, \ldots, k$. Note that in order to compute φ_k one needs to provide $M = 2k$ moments to fill in all the entries of A_k. Moreover, $\varphi_k = \tau_{2k}^*(\overline{\boldsymbol{P}}_{2k})$ where $\overline{P}_1 = o_1,\ \overline{P}_2 = \frac{o_2}{2!}, \ldots$
Other two important results are here reported.

Theorem 4.1.2. *Let* $\chi(\gamma)$ *be the number of support points of* γ*. If the number of support points is finite then* $\varphi_1 < \ldots < \varphi_{\chi(\gamma)} = \varphi(\gamma) = \frac{P_0(\gamma)}{1-P_0(\gamma)}$*; if* $\chi(\gamma) = \infty$*, then* $\varphi_1 < \ldots < \lim_{k \leftarrow \infty} \varphi_k = \varphi(\gamma) = \frac{P_0(\gamma)}{1-P_0(\gamma)}$.

Theorem 4.1.3. *Let us consider partitioning the space of conditional Poisson mixture probabilities* $\mathcal{P}_{PM}^C$ *into "sieves" of discrete distributions with* k *positive masses* $\mathcal{P}_{PM,k}^C = \left\{\overline{P} = \overline{P}(\gamma) : \chi(\gamma) = k\right\}$*. Then, if* $\overline{P}(\gamma) \in \mathcal{P}_{PM,k}^C$*,* $\varphi(\gamma) = \varphi_k$ *with* $o_1 = \overline{P}_1(\gamma),\ o_2 = \overline{P}_2(\gamma)/2, \ldots$*; if instead* $\overline{P} \in \cup_{k=j}^\infty \mathcal{P}_{PM,j}^C$*, then* $\varphi_k \leq \varphi(\gamma)$.

Theorem 4.1.3 guarantees that φ_k is Fisher consistent. Proves of both theorems can be found in Mao and Lindsay (2007) with a slightly different notation.

To estimate φ_k, Mao and Lindsay (2007) considered the empirical moments $\hat{o}_k = k!\frac{f_k}{n}$ and the corresponding empirical shifted Hankel matrix so that:

$$\hat{\varphi}_k = \hat{\boldsymbol{o}}_k^t \hat{A}_k^{-1} \hat{\boldsymbol{o}}_k$$

for $k \leq \hat{\chi}_n$ where $\hat{\chi}_n$ is the maximum k for which $\hat{A}_k$ is positive definite. For finite k the quantity $\hat{\varphi}_k$ is well defined, it exists almost surely and enjoys good asymptotic properties. However, it is not easy to compute, especially when the Hankel moment matrix is large. The main issue of this methodology is that, as we will see in a dedicated simulation study, the matrix inversion $\hat{A}_k^{-1}$ becomes unstable when k grows. In order to escape the Hankel matrix inversion, in the next paragraph we will review an alternative method first introduced in Harris (1959) and then refined in Daley and Smith (2016) and for which we provide an original improvement.

4.2 Harris transform and moment equation approach

For a reason that will be clarified within this paragraph, we might find convenient to reparametrize ν with another finite measure μ as follows:

$$d\mu(\lambda) = \lambda e^{-\lambda} d\nu(\lambda).$$

This parametrization was first introduced by Harris (1959) and also used by Chao (1984) to derive the moment-based Chao estimator. It should be noted that the total mass of μ is related to the Poisson mixture probability mass P_1. In fact

$$\int_0^\infty d\mu(\lambda) = \int_0^\infty \lambda e^{-\lambda} d\nu(\lambda) = P(X_i = 1|\lambda) = P_1 \leq 1.$$

Correspondingly we could consider the normalized reparametrized mixing distribution $\widetilde{\mu}(\lambda) = \frac{\mu(\lambda)}{\int_0^\infty d\mu(\lambda)}$ which is well defined as long as $P_1 > 0$. The moments of $\widetilde{\mu}$ are then equal to

$$\begin{aligned} \widetilde{m}_j &= \int_0^\infty \lambda^j \, d\widetilde{\mu}(\lambda) = \frac{N \int_0^\infty \lambda^{j+1} e^{-\lambda} d\nu(\lambda)}{N \int_0^\infty \lambda e^{-\lambda} d\nu(\lambda)} = \\ &= \frac{(j+1)! \mathbb{E}(f_{j+1})}{\mathbb{E}(f_1)} \quad for \quad j = 0, 1, 2, \ldots . \end{aligned} \tag{4.5}$$

Therefore, the moments of $\tilde{\mu}$ can be related to the expected values of the frequencies of frequencies. Since the expected frequencies of frequencies can be easily estimated by the observed frequencies of frequencies, we can plug them in (4.5) and get

$$\hat{m}_j = \frac{(j+1)! f_{j+1}}{f_1} \quad for \quad j = 0, 1, 2, \ldots . \tag{4.6}$$

This reparametrization based on the moments $\widetilde{m}_j$ yields a new representation of the unobserved expected frequency $\mathbb{E}(f_0) = N * P_0$ which links it to the expected observable frequency of frequencies $\mathbb{E}(f_1)$. In fact:

$$\begin{aligned}\mathbb{E}(f_0) &= N\int_0^\infty e^{-\lambda}d\nu(\lambda) = N\int_0^\infty \lambda^{-1}d\mu(\lambda) = \\ &= N\int_0^\infty d\mu(\lambda)\int_0^\infty \lambda^{-1}d\widetilde{\mu}(\lambda) = \mathbb{E}(f_1)\int_0^\infty \lambda^{-1}d\widetilde{\mu}(\lambda).\end{aligned} \tag{4.7}$$

The above expression is interesting: the expected value of f_0 is equal to the product between the expected value of the singletons and the functional $\int_0^\infty \lambda^{-1}d\widetilde{\mu}(\lambda)$. The information provided by the observed data on $\widetilde{\mu}$ can be conveyed through the moments that we can estimate using the observed count frequencies. We can now show how one can try to estimate $f_0 = N - n$ by means of a moment-constrained problem as follows:

$$\begin{aligned}&\text{estimate } \; E[f_0] = \int_0^\infty \lambda^{-1}d\widetilde{\mu}(\lambda) \\ &\text{subject to } \; \widetilde{m}_j = \hat{m}_j = (j+1)!f_{j+1}/f_1 \quad j = 1,\ldots,M.\end{aligned} \tag{4.8}$$

Estimates for the above integral functional of $\widetilde{\mu}$, subject to the moment constraints, can be substituted in equation (4.7) to obtain moment-based estimates of $E(f_0)$ and thus estimates of N. We will see in the next paragraphs that all the measures $\widetilde{\mu}$ on the positive real line that satisfy the expected moment constraints in equation (4.8) form a truncated moment class and the functional $\int_0^\infty \lambda^{-1}d\mu(\lambda)$ will attain its minimum and maximum on the boundary of this truncated moment class. In particular, the lower bound will be obtained by solving the system of equations involving the support points of a boundary measure $\widetilde{\mu}$ and its corresponding weights. We will argue that the solutions of the complex system of equations can be obtained easily if we consider the system of monic polynomials associated to the measure $\widetilde{\mu}$. These orthogonal polynomials satisfy a three-term recurrence equation whose coefficients are in a one-to-one correspondence with the moments of the measure $\widetilde{\mu}$.

4.2.1 Systems of moment equations

As anticipated in the previous paragraph, our focus is to obtain a moment-based lower bound of $E(f_0)$ and thus of N. We note that the constraints in (4.8) are based on the finite number of positive frequencies of frequencies $f_1,\ldots,f_M$.

We remind that the range of the first M moments forms a convex body and is usually referred to as the M-truncated moment space of $\widetilde{\mu}$, in symbols:

$$\mathcal{S}_1^{[M]} = \left\{(\widetilde{m}_1,\widetilde{m}_2,..,\widetilde{m}_M) : \widetilde{m}_k = \int_0^\infty \lambda^k d\widetilde{\mu}(\lambda), k = 1,...,M \;\; for \;\widetilde{\mu} \in \mathcal{P}([0,\infty))\right\}. \tag{4.9}$$

The space of the Poisson mixture probabilities can be reinterpreted as an image through an affine map of the convex body of the M truncated moment space and it is clear that $\mathcal{S}_1^{[M]}$ is closed and convex (Theorem 1 of Harris (1959)). Let us consider a fixed probability measure $\widetilde{\mu}^* \in \mathcal{P}([0,\infty))$. Let S be the number of support points of $\widetilde{\mu}^*$. When the support is not a finite collection of points on $[0,\infty)$, we set $S = \infty$. Then, for $M > 2S-1$, the class of probability measures $\widetilde{\mathcal{M}}^{[M]}$

$$\widetilde{\mathcal{M}}^{[M]}(\widetilde{m}_1^*,\ldots,\widetilde{m}_M^*) = \left\{\widetilde{\mu} \in \mathcal{P}([0,\infty)) : \int_0^\infty \lambda^k d\widetilde{\mu}(\lambda) = \widetilde{m}_k^*, \;\; for \; k = 1,...,M\right\}. \tag{4.10}$$

contains only a single probability measure, namely $\widetilde{\mu}^*$. On the other hand, if $M \leq 2S-1$ then $\widetilde{\mathcal{M}}^{[M]}$ contains not only $\widetilde{\mu}^*$, but also the other probability measures whose first M moments correspond to the first M moments of $\widetilde{\mu}^*$. As argued in the previous chapter, if all of the expected count frequencies $\{\mathbb{E}[f_1], \mathbb{E}[f_2], \ldots, \}$ were known, then the measure ν corresponding to ν^* could be perfectly recovered.

Let $g(\lambda)$ be a strictly convex or concave (and $g(\lambda) = \lambda^{-1}$ is the former); then, the linear functional $\mathcal{L}_{\widetilde{\mu}}$ defined by :

$$\mathcal{L}_{\widetilde{\mu}}(g) = \int_0^\infty g(\lambda) d\widetilde{\mu}(\lambda) \tag{4.11}$$

will attain its minima and maxima on the boundary of $\widetilde{\mathcal{M}}^{[M]}$. The boundary consists of discrete measures of minimal degree that satisfy the moment constraints (see Harris (1959) for more details). We can compute the extrema by finding the unique set of points $0 \leq e_1 < \ldots < e_S \leq \infty$ and positive weights $v_1, \ldots, v_S$ that satisfy the moment constraints given by:

$$\begin{aligned} v_1 + \ldots + v_S &= \hat{m}_0 \\ v_1 e_1 + \ldots + v_S e_S &= \hat{m}_1 \\ &\vdots \\ v_1 e_1^{2S-1} + \ldots + v_S e_S^{2S-1} &= \hat{m}_{2S-1}. \end{aligned} \tag{4.12}$$

Inferior extremal estimates (i.e. lower bounds) can be found by simply taking M odd and $S = (M+1)/2$ (we will argue this statement in the next paragraph). If one can solve the above system of equations, then the lower bound for the number of unobserved units is given by:

$$\hat{f}_0 = f_1 * \left(\frac{v_1}{e_1} + \ldots + \frac{v_S}{e_S} \right). \tag{4.13}$$

If $S = 1$ and $M = 1$, then the system of equations is simply given by the two equations $v_1 = 1$ and $e_1 = \hat{m}_1 = \frac{2f_2}{f_1}$ since

$$\widetilde{m}_1 = \frac{2!\mathbb{E}(f_2)}{\mathbb{E}(f_1)}.$$

It follows that $\hat{f}_0$ is equal to:

$$\hat{f}_0 = \frac{f_1}{e_1} = \frac{f_1^2}{2f_2}$$

which is exactly equal to Chao's lower bound. The basic idea is that, by considering moments of higher orders, one could get a lower bound which is larger than Chao's lower bound and thus find an estimate which is closer to the true value of $f_0 = N - n$. In general, it is not easy to derive the solutions of the system of equations for large S and M. Harris (1959) managed to find an explicit solution for $S = 2$ (corresponding to 2 points and 3 moments). Higher order analytic solutions cannot be found since they involve finding the roots of polynomials of degree 5 or higher. To our knowledge, in the population size estimation context, the first attempt to rely on the

quadrature formula (4.13) to find the sharpest lower bound by numerically solving the previous of system of equations for large S and M has been put forward in Daley and Smith (2016). In the following chapter, we will propose a slightly different approach which relies on checking if the empirical moment sequence of $\widetilde{\mu}$ is in fact a proper moment sequence corresponding to a measure on the interval $[0, \infty)$. The presented methodology is based on the pseudo canonical moments z_i, introduced in chapter 2 and for which we provide an explicit derivation in the Appendix.

4.2.2 Admissibility of moment sequences

We recall the Hankel matrices defined in (2.4). We showed that the equivalences

$$\begin{aligned}(\widetilde{m}_1, \ldots, \widetilde{m}_M) \in int\mathcal{S}_!^{[M]}([0,\infty)) &\text{ if and only if } \underline{\widetilde{H}}_M \textit{ and } \underline{\widetilde{H}}_{M-1} \text{ are positive definite;}\\ (\widetilde{m}_1, \ldots, \widetilde{m}_M) \in \mathcal{S}_1^{[M]}([0,\infty)) &\text{ if and only if } \underline{\widetilde{H}}_M \textit{ and } \underline{\widetilde{H}}_{M-1} \text{ are positive semidefinite;}\end{aligned} \tag{4.14}$$

hold.
In this sense, the positive semi-definiteness of the Hankel Matrices is both necessary and sufficient for the existence of a probability measure $\hat{\mu}$ with moments $\widetilde{m}_0, \widetilde{m}_1, \ldots, \widetilde{m}_M$. However, there is an equivalent condition based on the pseudo-canonical mapping ξ_M^{-1}. In Dette and Studden (2002) it is remarked that

$$\widetilde{m}_0, \widetilde{m}_1, \ldots, \widetilde{m}_M$$

is a moment sequence if and only if

$$z_1 \ldots z_M > 0 \tag{4.15}$$

where $(z_1, \ldots, z_M) = \xi_M^{-1}(\widetilde{m}_1, \ldots, \widetilde{m}_M)$. Consequently, (4.15) can be used to check if a given sequence is in fact a moment sequence corresponding to a probability measure on the interval $[0, \infty)$. Since we do not observe true moments but just empirical moments obtained via the count frequencies, we are not guaranteed that they form a legitimate M-moment sequence. For this reason, in practice, we deal with k-truncated moment sequence where k is the maximum $k \leq M$ for which $(z_1, \ldots, z_k) = \xi_k^{-1}(\widetilde{m}_1, \ldots, \widetilde{m}_k)$ and

$$z_1 \ldots z_k > 0.$$

4.2.3 Quadrature lower bounds via Jacobi matrix eigendecomposition

In the previous paragraph, we have analyzed the necessary conditions for the admissibility of a moment sequence and thus for the existence of solutions of the system of equations (4.12). The main issue we face is that, despite the solutions are unique, we might still find some of them that satisfy the system of equations to numerical precision but are quite far to the true values. This is not a surprise since Gautschi (1985) proved the ill-conditioning of the mapping between moments and quadrature rules. This is particularly true when the system of equations is large (i.e.

there are many moments).
The Gaussian quadrature is a technique strictly related to the moment problem. Moments and quadrature have been always connected since the classical works of Chebyshev and Stieltjes on continued fractions and the moment problem. A quadrature formula with degree of precision $S \in \mathbb{N}$ for the probability measure $\widetilde{\mu}$ consists of real numbers $(e_1, \ldots, e_S)$, called nodes and corresponding quadrature coefficients or weights $(v_1, \ldots, v_S)$, such that

$$\int_0^\infty \lambda^m d\widetilde{\mu}(\lambda) = v_1 e_1^m + \ldots + v_S e_S^m + R_S(\lambda^m)$$

which has maximum algebraic degree of exactness $2S$–1 (i.e. $R_S(\lambda^{2S-1}) = 0$). Let us now consider the case where $m = -1$; since the even order derivatives of $g(\lambda) = \frac{1}{\lambda}$ are positive, we have that $R_S(\lambda^{-1}) < 0$ when M is odd: in this case, the Gaussian quadrature approximation is a lower bound for $\int_0^\infty \lambda^{-1} d\tilde{\mu}(\lambda)$ (Golub and Meurant, 1994). In turns, since the odd order derivatives of $g(\lambda) = \frac{1}{\lambda}$ are negative, $R_S(\lambda^{-1}) > 0$ when M is even which implies that the Gaussian quadrature approximation is an upper bound for $\int_0^\infty \lambda^{-1} d\tilde{\mu}(\lambda)$. Algorithms for computing Gaussian quadrature rules involve estimating the orthogonal polynomials associated with the underlying measure. Daley and Smith (2016) found that this methodology gives stabler estimates than the available non-linear equation solvers. We know that, for $k \leq S$, the nodes $(e_1, \ldots, e_k)$ are the zeros of monic orthogonal polynomials $P_k(\cdot; d\widetilde{\mu})$ (depending on the measure $\widetilde{\mu}$) with recurrence coefficients $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$. It follows that the nodes $(e_1, \ldots, e_k)$ are the eigenvalues of the Jacobi matrix J_k

$$J_k = \begin{bmatrix} a_0 & \sqrt{b}_1 & 0 & \cdots & 0 \\ \sqrt{b}_1 & a_1 & \sqrt{b}_1 & \cdots & 0 \\ 0 & \sqrt{b}_2 & a_2 & \sqrt{b}_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sqrt{b}_{k-1} & a_{k-1} \end{bmatrix}. \tag{4.16}$$

The weights $(v_1, \ldots, v_k)$ are then given by the squared first component of the eigenvectors of J_k. The recurrence coefficients can be derived trough either the Chebyshev algorithm or the QD algorithm. Both algorithms are detailed in two separate appendices. Given the estimated recurrence coefficients, the quadrature points and weights can be obtained using any algorithm that calculates the eigenvalues and eigenvectors. In our scenario, since the matrix is already tridiagonal, there is no need to calculate all the components of the eigenvectors since only the first component is required. In Daley and Smith (2016) a modified QR algorithm is instead used. Such modified algorithm has complexity that is linear in both space and time and this represents a great advantage. Moreover, it is claimed that it is well-conditioned and small changes in the estimated recurrence coefficients won't cause large changes in the estimated nodes and weights. In our proposal, to calculate the eigenvalues and eigenvectors of J_k, we have utilized an Eigenvalue Decomposition using the R package RSpectra (Qiu and Mei, 2019). In R, the function `eigen()` is already available in CRAN but it still raises some issues: when the matrix becomes large,

it can be very time-consuming and the complexity to calculate all the eigenvalues of a $k \times k$ matrix is $O(k^3)$. Furthermore, we usually need to compute only few eigenvalues or eigenvectors (as it happens in our case where we only need the first component of eigenvectors). In `eigen()` , there is no option to limit the number of eigenvalues to be computed therefore we need to go for a full computation. For the basic singular value decomposition (SVD), it happens exactly the same: we might need a truncated or partial SVD but in standard R this option is not available. The package RSpectra allows us to overcome these problems providing an R interface to the Spectra library, which is used to solve large scale eigenvalue problems. The core part of Spectra and RSpectra has been developed in efficient C++ code and it can handle large scale matrices in either dense or sparse formats. Moreover, this package can limit the algorithm to provide only the first components.

4.2.4 Stabilizing numerical methods for moment based quadrature

The estimation of the truncated moment sequence $\boldsymbol{m}_M$ of the reparametrized (and normalized) mixing distribution $\widetilde{\mu}$ is a fundamental part of our work. We have seen in the previous chapters that these moments are directly related to the expected values of the count frequencies. In practice, we dispose only of observed count frequencies and the risk is that they might be just a spurious observation and thus lead to bad estimates of the truncated moment sequences (and consequently bad estimates of the quadrature nodes and weights).

Let us suppose we have two different vectors of observed count frequencies:

$$\mathbf{f}_1 = (2899, 1706, 803, 391, 353, 336, 288, 251, 223, 128, 97, 52, 26, 2, 6, 2, 1)$$
$$\mathbf{f}_2 = (2853, 1664, 878, 409, 341, 317, 296, 275, 219, 124, 100, 48, 22, 3, 9, 4, 1)$$

The above observed count frequencies produce the following truncated moment sequences (for ease of writing we should truncate the sequence to moments of order 5)

$$\boldsymbol{m}_1 = (1, 1.18, 1.66, 3.24, 14.62)$$
$$\boldsymbol{m}_2 = (1, 1.17, 1.85, 3.44, 14.35)$$

on the other hand, the corresponding pseudo canonical moments will be:

$$\boldsymbol{z}_1 = (1.17, 0.24, 3.18, 6.73)$$
$$\boldsymbol{z}_2 = (1.67, 0.42, 1.07, 14.508)$$

It is clear that small variations in the truncated moment sequence can cause a big variation in the corresponding pseudo canonical moments (and therefore in the recurrence coefficients). How can we mitigate this instability ? We are in a situation where the bootstrapped samples can actually help. Bootstrapping is a re-sampling method It is described as random sampling with replacement.

In essence, it gives us duplicates/perturbed observations. It might be particularly useful when the dataset is not big and the algorithm requires many data (as it happens in our setting). In Daley and Smith (2016), the Bagging (bootstrap aggregating), an alternative bootstrap algorithm (originally presented in Kuhnert et al. (2008)), is

utilized. When repeated samples and hence repeated evaluations of the corresponding lower bounds are available, one could try to take advantage of them by aggregating the collection of lower bounds by means of a robust summary like the median. In our setting, in order to avoid the use of a single sample $f_+ = (f_1, f_2, \ldots, f_M)$ which might turn in an unstable estimate, we follow Daley and Smith (2016) and use the median of all lower bounds derived from the bootstrapped samples. Why should we utilize the median instead of the mean? An important difference between these two statistics is that the mean is much more sensitive to outliers than the median. That is, one or two outliers can drastically affect the mean but do not change the median that much (in other words, the median is much more robust than the mean).

We have implemented this idea for deriving alternative estimates of population size relying on different ways of computing lower bounds for $\mathbb{E}(f_0)$ by exploiting a suitable Dirichlet distribution for resampling. In fact, if we multiply by n a random draw $(\pi_1, ..., \pi_M)$ from the following Dirichlet distribution

$$\boldsymbol{\pi}_M = (\pi_1, ..., \pi_M) \sim Dir\left(\frac{f_1 + 1}{n + M}, \ldots, \frac{f_M + 1}{n + M}\right)$$

we can mimic in a smoother way the behaviour of the standard resampling from the observed frequencies:

- B vectors of probabilities $\boldsymbol{\pi}_M^{(b)}$ are sampled from a Dirichlet distribution with parameter components $a_j = (f_j + 1)/(n + M)$ and the corresponding generic bootstrap frequencies are
$$\boldsymbol{f}_M^{(b)} = n * \boldsymbol{\pi}_M^{(b)}$$
so that $m_k^{(b)}$ is
$$\hat{m}_k^{(b)} = \frac{(k+1)! f_{k+1}^{(b)}}{f_1^{(b)}}$$

- for each $b = 1, ...B$, starting from the $\hat{m}_k^{(b)}$, the recurrence coefficients are calculated via the QD algorithm and the Jacobi matrix J is built. Finally, the nodes (eigenvalues of J) and weights (squared first component of the eigenvectors) are calculated and $\hat{n}_0^{(b)}$ and the corresponding $\hat{N}^{(b)}$ are obtained.

- The median of $\hat{N}$ across all the bootstrapped samples is then computed
$$\hat{N} = Median\left\{\hat{N}^{(1)}, \ldots, \hat{N}^{(B)}\right\}.$$

4.3 Sequential moment condition check approach

Analogously to what has been argued in the previous paragraphs, one could verify the admissibility of the truncated moment sequence $\widetilde{\boldsymbol{s}}_M = (\widetilde{s}_0 = 1, \widetilde{s}_1, ..., \widetilde{s}_j, ..., \widetilde{s}_M)$ of the probability measure $\widetilde{\phi}$ defined from normalizing the finite measure $d\phi(\lambda) = e^{-\lambda} d\nu(\lambda)$ by checking the positiveness of the corresponding pseudo-canonical moments $\boldsymbol{z}_M$. This truncated moment sequence, together with any fixed value of the total mass

$s_0 \in (0,1]$, allows us to uniquely determine the truncated moment sequence of a finite measure $\phi \in \overline{\mathcal{F}}_{(0,1]}([0,\infty))$ using the relation $s_j = s_0 \widetilde{s}_j$. This suggests us a different approach for determining the lower bound of $P_0 = s_0$ when s_0 is the total mass of a finite measure with moment constraints underlying the class

$$\begin{aligned}\mathcal{C}(\overline{P}_1^*, \ldots, \overline{P}_M^*) &= \left\{ \boldsymbol{s}_M \in \mathcal{S}_{(0,1]}^{[M]} : \overline{P}(\boldsymbol{s}_M) = \overline{\boldsymbol{P}}_M^*, \quad j = 1, \ldots, M \right\} \\ &= \left\{ \boldsymbol{s}_M = (s_0, \ldots, s_M) \in \mathcal{S}_{(0,1]}^{[M]} : \frac{j! s_j}{1 - s_0} = \overline{P}_j, \quad j = 1, \ldots, M \right\}.\end{aligned}$$

From the last expression one can use a candidate $s_0^* \in (0,1]$ and all the $\overline{P}_j$s to verify if the corresponding putative truncated sequence

$$(s_0^*, s_1^* = \overline{P}_1 \cdot (1 - s_0^*), \ldots, s_j^* = \frac{\overline{P}_j}{j!} \cdot (1 - s_0^*), \ldots, s_M^* = \frac{\overline{P}_M}{M!} \cdot (1 - s_0^*))$$

does indeed correspond to a valid truncated moment sequence of the underlying normalized probability measure with moment sequence

$$\widetilde{\boldsymbol{s}}_{\boldsymbol{M}}^* = \left(\widetilde{s}_0^* = 1, \widetilde{s}_1^* = \overline{P}_1 \cdot \frac{(1 - s_0^*)}{s_0^*}, \ldots, \widetilde{s}_j^* = \frac{\overline{P}_j}{j!} \cdot \frac{(1 - s_0^*)}{s_0^*}, \ldots, \widetilde{s}_M^* = \frac{\overline{P}_M}{M!} \cdot \frac{(1 - s_0^*)}{s_0^*} \right).$$

Hence, one could attempt to sequentially approximate up to arbitrary precision the sharpest lower bound with a bisection search of the smaller $s_0^* \in (0,1]$ such that the normalized moment sequence $\widetilde{\boldsymbol{s}}_M^*$ has a pseudo canonical mapping

$$(z_1^*, \ldots, z_j^*, \ldots, z_M^*) = \xi^{-1}(\widetilde{s}_1^*, \ldots, \widetilde{s}_j^*, \ldots, \widetilde{s}_M^*)$$

and

$$z_1^* \ldots z_M^* > 0.$$

4.4 Numerical accuracy of alternative methods for sharpest lower bound computations

In this section, for comparison purposes, we report the numerical findings obtained from fixing some mixing ν and deriving for different values of M the corresponding conditional probabilities $(\overline{P}_1(\nu), \ldots, \overline{P}_M(\nu))$ and computing the sharpest lower bound

$$\tau_M^*(\overline{\boldsymbol{P}}_M) = \inf_{\boldsymbol{s}_M \in \mathcal{C}(\overline{\boldsymbol{P}}_M)} P_0(\boldsymbol{s}_M) = \inf_{\boldsymbol{o}_M \in \mathcal{S}_I^{[M]} | o_1 = \overline{P}_1, o_2 = 2\overline{P}_2, \ldots, o_M = M! \overline{P}_M} \frac{o_0}{1 + o_0}$$

using the algebraic lower bound approach (A-Mao), the quadrature approach (Q-DS) and the sequential moment condition check approach (S-Z). We consider 5 different mixing distributions ν corresponding to the first five settings of Wang (2010). In particular, as detailed in Table 4.1), the mixing distributions ν are gamma distribution (setting 1,2,3) and two component mixtures of gamma (4,5). For each setting, we first need to derive ϕ as $\phi(\lambda) = e^{\lambda}\nu(\lambda)$ and then its corresponding moments. For the first three settings, we have that:

$$\phi(\lambda) = e^{\lambda} Ga(\alpha, \mu)$$

and it is easy to check that:

$$s_k = \lambda^k d\phi(\lambda) = \left(\frac{\alpha}{\mu}\right)^{\alpha} \left(\frac{1}{\frac{\alpha}{\mu}+1}\right)^{(k+\alpha)} \frac{(\alpha+k-1)!}{(\alpha-1)!}.$$

The derivation of P_0 is then straightforward:

$$P_0 = s_0 = \left(\frac{\alpha}{\mu}+1\right)^{-\alpha} \left(\frac{\alpha}{\mu}\right)^{\alpha}.$$

For the last two settings we have instead the following reparametrized finite measure:

$$\phi(\lambda) = e^{\lambda}(w * Ga(\alpha,\mu) + (1-w)Ga(\alpha_1,\mu_1).$$

Since the integration is a linear operator, the generic k-th moment can be easily derived:

$$\begin{aligned} s_k =& \lambda^k d\phi(\lambda) = w * \left[\left(\frac{\alpha}{\mu}\right)^{\alpha} \left(\frac{1}{\frac{\alpha}{\mu}+1}\right)^{(k+\alpha)} \frac{(\alpha+k-1)!}{(\alpha-1)!}\right] + \\ & (1-w)\left[\left(\frac{\alpha_1}{\mu_1}\right)^{\alpha_1} \left(\frac{1}{\frac{\alpha_1}{\mu_1}+1}\right)^{(k+\alpha_1)} \frac{(\alpha_1+k-1)!}{(\alpha_1-1)!}\right]. \end{aligned}$$

The calculation of P_0 is again immediate:

$$P_0 = s_0 = w * \left[\left(\frac{\alpha}{\mu}+1\right)^{-\alpha} \left(\frac{\alpha}{\mu}\right)^{\alpha}\right] + (1-w) * \left[\left(\frac{\alpha_1}{\mu_1}+1\right)^{-\alpha_1} \left(\frac{\alpha_1}{\mu_1}\right)^{\alpha_1}\right].$$

Setting	Distribution (ν)	P_0
	Gamma	
1	Ga(4,3.25)	0.09266
2	Ga(4,1)	0.4096
3	Ga(1,0.25)	0.8000
	Gamma Mixture	
4	$0.5 \cdot$ Ga(2,1)+$0.5 \cdot$ Ga(2,2)	0.3472
5	$0.5 \cdot$ Ga(2,1)+$0.5 \cdot$ Ga(4,1)	0.4270

Table 4.1. Wang Simulation Settings (2010

sim	α	μ	M	A-Mao	S-Z	Q-DS
1	4	3.25	4	0.08702	0.08702	0.08702
1	4	3.25	6	0.09025	0.09026	0.09025
1	4	3.25	8	0.09146	0.09146	0.09146
1	4	3.25	10	0.09199	0.09199	0.09199
1	4	3.25	12	0.09226	0.09226	0.09226
1	4	3.25	14	0.09240	0.09241	0.09240
1	4	3.25	16	0.09249	0.09249	0.09249
1	4	3.25	18	0.09254	0.09255	0.09254
1	4	3.25	20	0.09257	0.09258	0.09258
1	4	3.25	22	0.09260	0.09260	0.09260
1	4	3.25	24	0.09261	0.09262	0.09261
1	4	3.25	26	0.09282	0.09263	0.09262
1	4	3.25	28	0.09149	0.09264	0.09263
1	4	3.25	30	0.11434	0.09264	0.09264
1	4	3.25	32	0.36151	0.09265	0.09264
1	4	3.25	34	0.84919	0.09265	0.09264
2	4	1	4	0.39303	0.39305	0.39303
2	4	1	6	0.40261	0.40264	0.40261
2	4	1	8	0.40612	0.40616	0.40612
2	4	1	10	0.40767	0.40767	0.40767
2	4	1	12	0.40845	0.40845	0.40845
2	4	1	14	0.40887	0.40890	0.40887
2	4	1	16	0.40911	0.40915	0.40911
2	4	1	18	0.40926	0.40927	0.40926
2	4	1	20	0.40936	0.40940	0.40936
2	4	1	22	0.40942	0.40944	0.40942
2	4	1	24	0.40949	0.40948	0.40947
2	4	1	26	0.41001	0.40952	0.40950
2	4	1	28	0.40933	0.40956	0.40952
2	4	1	30	0.41601	0.40956	0.40954
2	4	1	32	0.93617	0.40956	0.40955
2	4	1	34			0.40958
2	4	1	36			0.40955
2	4	1	38			0.40955

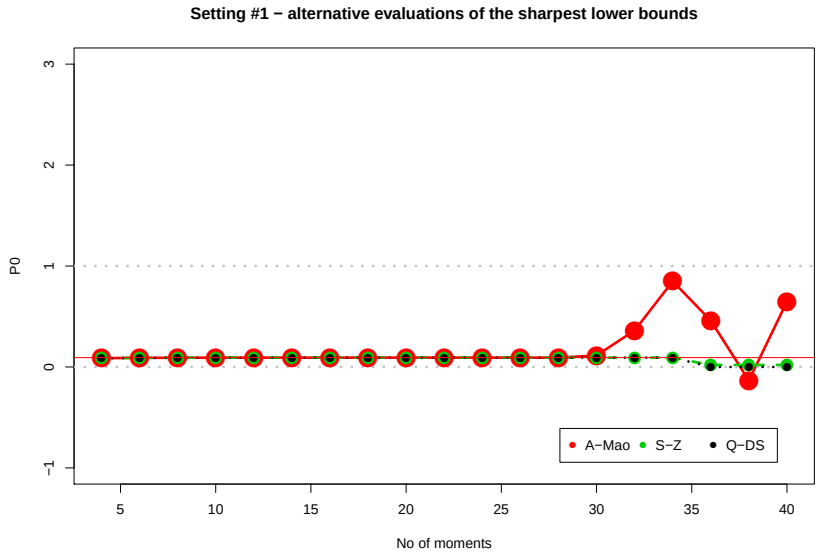
Table 4.2. *Calculating the sharpest lower bounds in the first 3 settings of the Wang simulation study*

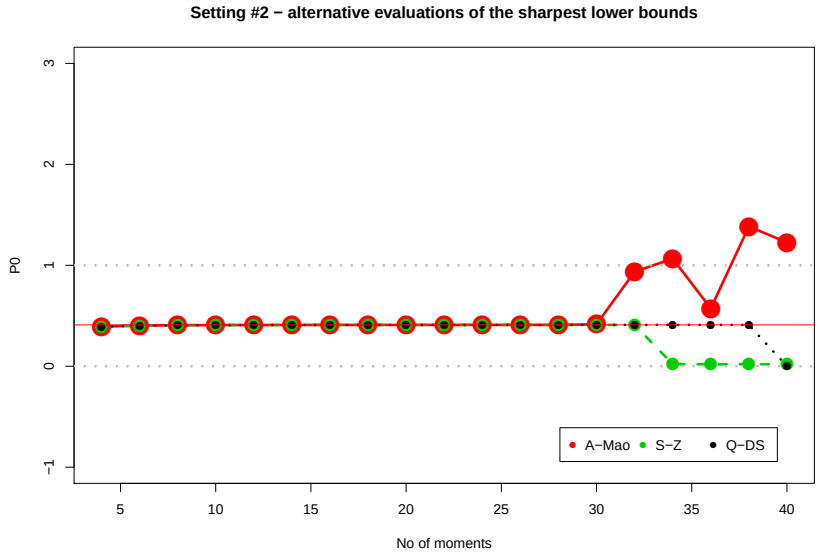
Sim	α	μ	M	A-Mao	S-Z	Q-DS
3	1	0.25	4	0.72727	0.72727	0.72727
3	1	0.25	6	0.75000	0.75008	0.75000
3	1	0.25	8	0.76190	0.76192	0.76190
3	1	0.25	10	0.76923	0.76928	0.76923
3	1	0.25	12	0.77419	0.77424	0.77419
3	1	0.25	14	0.77778	0.77784	0.77778
3	1	0.25	16	0.78049	0.78056	0.78049
3	1	0.25	18	0.78261	0.78264	0.78261
3	1	0.25	20	0.78431	0.78432	0.78431
3	1	0.25	22	0.78571	0.78576	0.78571
3	1	0.25	24	0.78688	0.78696	0.78689
3	1	0.25	26	0.78789	0.78792	0.78788
3	1	0.25	28	0.78870	0.78880	0.78873
3	1	0.25	30	0.78875	0.78952	0.78947
3	1	0.25	32	0.77128	0.79016	0.79012
3	1	0.25	34	1.34661	0.79064	0.79067
3	1	0.25	36	1.30425	0.79104	0.79092
3	1	0.25	38	0.62282	0.79208	

Table 4.3. *Calculating the sharpest lower bounds in the first 3 settings of the Wang simulation study*

sim	w	α_1	μ_1	α_2	μ_2	M	A-Mao	S-Z	Q-DS
4	0.50	2	1	2	2	4	0.30077	0.30078	0.30077
4	0.50	2	1	2	2	6	0.31884	0.31885	0.31884
4	0.50	2	1	2	2	8	0.32792	0.32795	0.32792
4	0.50	2	1	2	2	10	0.33318	0.33318	0.33318
4	0.50	2	1	2	2	12	0.33651	0.33652	0.33651
4	0.50	2	1	2	2	14	0.33877	0.33878	0.33877
4	0.50	2	1	2	2	16	0.34038	0.34038	0.34038
4	0.50	2	1	2	2	18	0.34156	0.34156	0.34156
4	0.50	2	1	2	2	20	0.34245	0.34248	0.34245
4	0.50	2	1	2	2	22	0.34315	0.34316	0.34315
4	0.50	2	1	2	2	24	0.34370	0.34372	0.34370
4	0.50	2	1	2	2	26	0.34411	0.34418	0.34415
4	0.50	2	1	2	2	28	0.34508	0.34454	0.34452
4	0.50	2	1	2	2	30	0.35976	0.34483	0.34482
4	0.50	2	1	2	2	32	0.41044	0.34510	0.34508
4	0.50	2	1	2	2	34	0.60132	0.34529	0.34530
4	0.50	2	1	2	2	36	0.38147	0.34529	0.34538
4	0.50	2	1	2	2	38	0.54271	0.34555	
5	0.50	2	1	4	1	4	0.39579	0.39580	0.39579
5	0.50	2	1	4	1	6	0.40973	0.40976	0.40973
5	0.50	2	1	4	1	8	0.41603	0.41603	0.41603
5	0.50	2	1	4	1	10	0.41942	0.41945	0.41942
5	0.50	2	1	4	1	12	0.42145	0.42149	0.42145
5	0.50	2	1	4	1	14	0.42276	0.42279	0.42276
5	0.50	2	1	4	1	16	0.42366	0.42368	0.42366
5	0.50	2	1	4	1	18	0.42430	0.42434	0.42430
5	0.50	2	1	4	1	20	0.42478	0.42478	0.42478
5	0.50	2	1	4	1	22	0.42513	0.42515	0.42514
5	0.50	2	1	4	1	24	0.42541	0.42543	0.42542
5	0.50	2	1	4	1	26	0.42570	0.42568	0.42564
5	0.50	2	1	4	1	28	0.42515	0.42584	0.42582
5	0.50	2	1	4	1	30	0.38285	0.42596	0.42596
5	0.50	2	1	4	1	32	0.18048	0.42609	0.42608
5	0.50	2	1	4	1	34	0.10521	0.42621	0.42614
5	0.50	2	1	4	1	36	0.65169	0.42621	0.42615

Table 4.4. *Calculating the sharpest lower bounds in settings 4 & 5 of the Wang simulation study*





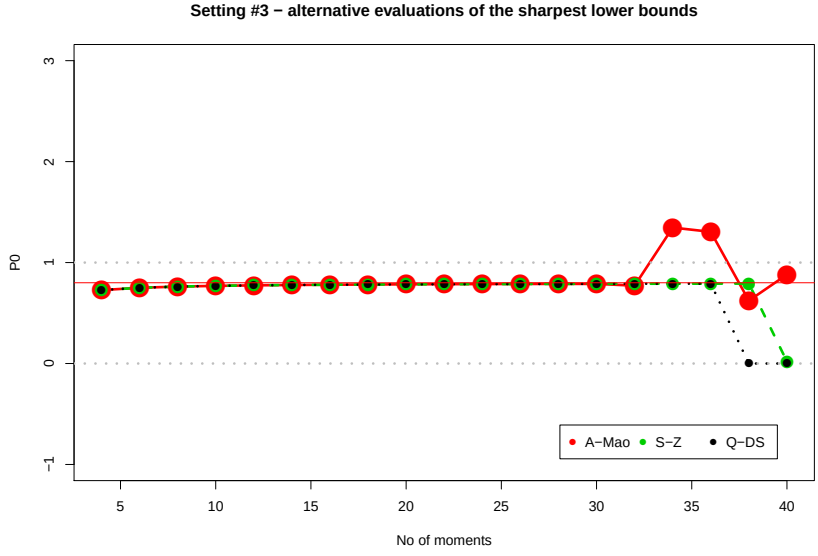


Figure 4.1. settings 1 &2 & 3: convergence of the numerical methods to P_{0_*}

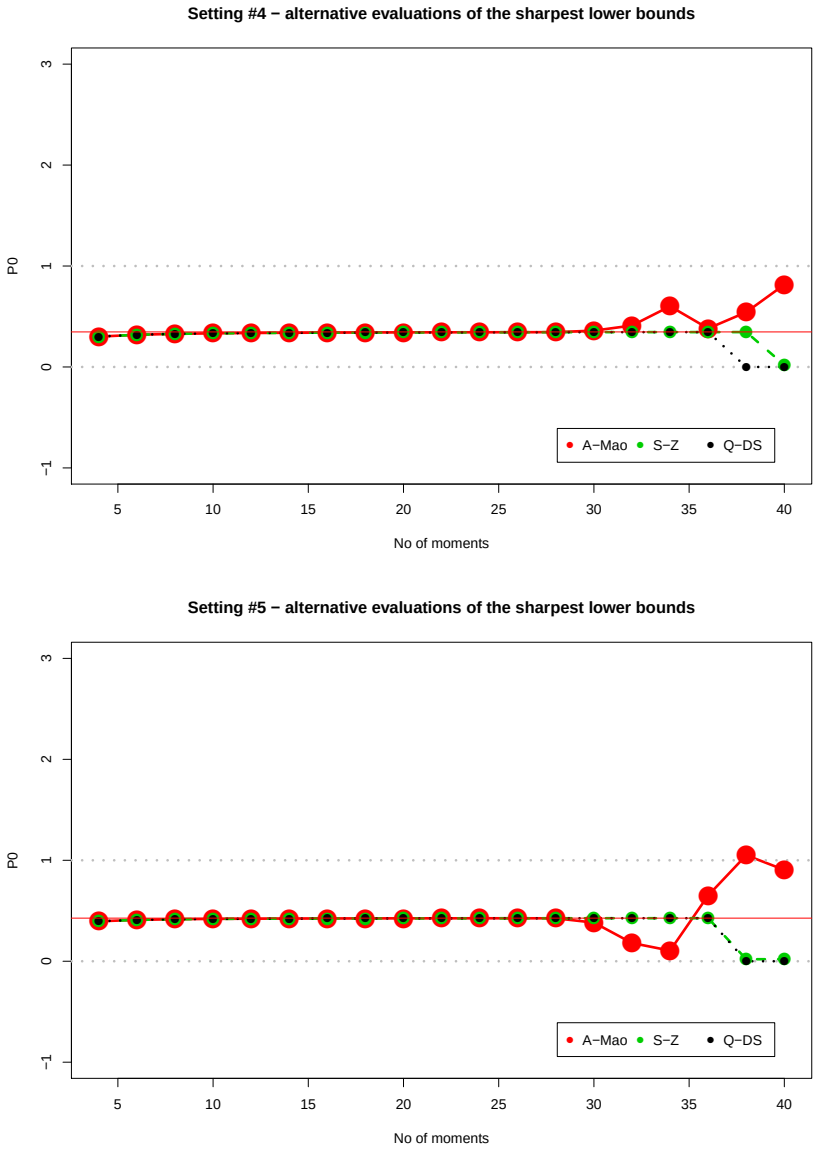


Figure 4.2. settings 4 & 5: convergence of the numerical methods to P_{0_*}

Comparative results for each setting are shown in tables 4.2, 4.3 and 4.4 and graphically displayed in Figure 4.1 and 4.2. The Algebraic lower bound (A-Mao) gets unstable when M is greater than 25 in all the three settings. This is likely due to fact that inverting the moment matrix becomes problematic for moderately large M. The sequential methods (Z-S) and the DS method (Q-DS) are similar and start suffering instability for larger value of M without a manifest superiority of one method over the other. Unwittingly, we are comparing the performance of the QD algorithm and of the Chebyshev algorithm. There has been a large debate in the literature on which algorithm one should use. Although it has been noted by Wheeler (1974) that QD algorithm is more likely to break down due to degeneracies, in this specific simulation study we do not see one algorithm winning over the other.

Indeed, one could try to improve the numerical accuracy of the sequential method by using the stabler but more computationally demanding QD algorithm illustrated in Stokes (1980).

4.5 Unconditional MLE for estimation of the sharpest lower bound

In the previous paragraphs we have illustrated a methodology to compute numerically the lower bound of the probability $P_0(\nu)$ corresponding to all mixing probability measures ν which share the same fixed conditional probabilities $(\overline{P}_1, \ldots, \overline{P}_M)$ using alternative numerical algorithms for different values of M. Of course, these methodologies can be employed for estimating the sharpest lower bound of the unknown P_0 and hence of the unknown N once an estimate of the conditional probabilities $(\overline{P}_1, \ldots, \overline{P}_M)$ is available from maximizing the conditional likelihood. If we start from the theoretical results shown in the third chapter, we could infer on the unknown parameters by maximizing the unconditional likelihood reparametrized using the moments of the transformed measure ϕ corresponding to the underlying mixing distribution ν. The unconditional maximum likelihood estimator for $(N, \boldsymbol{s}_M)$ is denoted by:

$$(\hat{N}, \hat{\boldsymbol{s}}_M) = \underset{(N, \boldsymbol{s}_M) \in \mathbb{N} \times \mathcal{S}_I^{[M]}}{\arg\max} L(N, \boldsymbol{s}_M; \boldsymbol{f}_+). \tag{4.17}$$

One can argue that the first component of $(\hat{N}, \hat{\boldsymbol{s}}_M)$ can be expressed as a function of the second component according to the following:

$$(\hat{N}, \hat{\boldsymbol{s}}_M) = \left(\left\lceil \frac{n}{1 - P_0(\hat{\boldsymbol{s}}_M)} \right\rceil, \hat{\boldsymbol{s}}_M \right) = (\hat{N}, \hat{\boldsymbol{s}}_M) = \left(\left\lceil \frac{n}{1 - \hat{s}_0} \right\rceil, \hat{\boldsymbol{s}}_M \right)$$

The moment-based approach for the conditional maximum likelihood estimation relies instead on two steps:

- estimate $\boldsymbol{s}_M$ through the conditional likelihood $L_C(\boldsymbol{s}_M; \boldsymbol{f}_+)$ obtaining:

$$\hat{\boldsymbol{s}}_M^C = (\hat{s}_0^C, \hat{s}_1^C, \ldots, \hat{s}_M^C) = \underset{\boldsymbol{s}_M \in \mathcal{S}_I^{[M]}}{\arg\max} L_C(\boldsymbol{s}_M; \boldsymbol{f}_+);$$

- plug $\hat{\boldsymbol{s}}_M^C$ in $P_0(\boldsymbol{s}_M)$ in the residual likelihood and hence obtain an estimate of the population size N by maximizing $L_R(N, \hat{\boldsymbol{s}}_M^C; \boldsymbol{f}_+)$, namely:

$$\hat{N}_C = \hat{N}_{P_0(\hat{\boldsymbol{s}}_M^C)} = \underset{N \in \mathbb{N}}{\arg\max} L_R(N, \hat{\boldsymbol{s}}_M^C; \boldsymbol{f}_+) = \left\lceil \frac{n}{1 - P_0(\hat{\boldsymbol{s}}_M^C)} \right\rceil = \left\lceil \frac{n}{1 - \hat{s}_0^C} \right\rceil. \tag{4.18}$$

We can then argue that $P_0(\hat{\boldsymbol{s}}_M^C) \geq P_0(\hat{\boldsymbol{s}}_M)$ and, consequently, $\hat{N}_C \geq \hat{N}$. The argument comes from the fact that the unconditional MLE, whether parametric or nonparametric, can be obtained as penalized version of the conditional MLE with a penalization term which depends only on $s_0 = P_0(\boldsymbol{s}_M)$ and is increasing with respect to s_0. Thus, the corresponding unconditional $P_0(\hat{\boldsymbol{s}}_M) = \hat{s}_0$ is always no greater

than $P_0(\hat{\boldsymbol{s}}_M^C) = \hat{s}_0^C$. To demonstrate this result, let us first write the unconditional log-likelihood

$$\ell(N, \boldsymbol{s}_M; \boldsymbol{f}_+) = \ell_R(N, \boldsymbol{s}_M; \boldsymbol{f}_+) + \ell_C(\boldsymbol{s}_M; \boldsymbol{f}_+).$$

We could then use the fact that, taking

$$N(s_0) = \left\lceil \frac{n}{1 - s_0} \right\rceil$$

one gets

$$\begin{aligned} \ell_R(N, \boldsymbol{s}_M; \boldsymbol{f}_+) &= log \left[\binom{N}{f_0} P_0(\boldsymbol{s}_M)^{f_0} \left(1 - P_0(\boldsymbol{s}_M)\right)^{N - f_0} \right] \leq \\ &\leq log \left[\binom{N(s_0)}{f_0} s_0^{f_0} \left(1 - s_0\right)^{N(s_0) - f_0} \right] \equiv \ell_R^*(s_0; \boldsymbol{f}_+) \end{aligned}$$

and hence

$$\ell(N, \boldsymbol{s}_M; \boldsymbol{f}_+) = \ell_R(N, \boldsymbol{s}_M; \boldsymbol{f}_+) + \ell_C(\boldsymbol{s}_M; \boldsymbol{f}_+) \leq \ell_R^*(s_0; \boldsymbol{f}_+) + \ell_C(\boldsymbol{s}_M; \boldsymbol{f}_+).$$

We could then look at the rhs of the above inequality as unconditional profile likelihood which turns out to be a penalized version of the conditional likelihood where the penalty coefficient is $\gamma = 1$ and the penalty function is $h(s_0; \boldsymbol{f}_+) = -\ell_R^*(s_0; \boldsymbol{f}_+)$. Following the argument in Appendix C of Wang and Lindsay (2005) with a slightly different notation one can prove that $\ell_R^*(s_0; \boldsymbol{f}_+)$ decreases monotonously in s_0 so that the penalization function increases with s_0 which in turn implies that

$$P_0(\hat{\boldsymbol{s}}_M) = \hat{s}_0 \leq \hat{s}_0^C = P_0(\hat{\boldsymbol{s}}_M^C).$$

This in fact holds for any $\hat{\boldsymbol{s}}_M^C$ which maximizes the conditional likelihood and hence

$$P_0(\hat{\boldsymbol{s}}_M) \leq \hat{\tau}_C = inf \left\{ P_0(\boldsymbol{s}') : \boldsymbol{s}' \in \underset{\boldsymbol{s}_M \in \mathcal{S}_I^{[M]}}{\arg\max} L_c(\boldsymbol{s}_M; \boldsymbol{f}_+) \right\} \tag{4.19}$$

with $\hat{\tau}_C$ being the sharpest lower bound corresponding to the conditional MLE. This fact reveals that the unconditional likelihood always provides an estimate which represents an approximation from below of the estimate provided by the sharpest lower bound estimation.

4.6 Simulation study

In order to evaluate the performance of our proposals, we have implemented a simulation study according to the same settings considered in Wang (2010) as described in the table below. Different mixing distributions are considered (one per each setting) and 100 simulated set of data are drawn and used to repeat the estimation procedure. Averages of Median, MSE, lower confidence interval, upper confidence interval and coverage are considered for each estimator. We compare our proposed estimator (likelihood-based lower bound) with the Daley-Smith lower bounds (DS lower bound) and its variant obtained through the QD algorithm (Z lower bound, presented in the previous chapter for which the derivation is detailed in paragraph 4.2.4). The simulation study is structured as follows:

- we compare the likelihood-based lower bounds with the famous nonparametric MLE proposed by Norris and Pollock (NP) over $N = 10000$;
- we compare the performances of the "quadrature" bound Z against the proposal in Daley and Smith (2016) (that we will call DS) over $N = 1000,\ 10000,$ 100000 and 1000000;
- we finally compare DS lower bound with our likelihood-based lower bounds for $N = 10000$.

For Z and DS we have considered 20000 bootstrapped samples whereas for the likelihood based lower bound we have considered 20 bootstrapped samples due to their larger computational burden.
In order to derive the unconditional MLE, we have considered three different maximization routines: one (rec01) considering bounded mixing distributions between $[0, 1]$, the other (rec) considering standard unbounded mixing distributions over $[0, \infty)$ with suitable initialization for the recurrence coefficients and the last one (zm) considering the pseudo canonical moments instead of the recurrence coefficients. For all these methods, we have avoided to maximize over the integer N by using a reparametrized version of the profile likelihood. In order to choose T for each simulation setting, we have looked at the ratio plot (Böhning et al., 2013) across the 12 simulation settings; we have then truncated the frequencies of frequencies when the monotonicity conditions are not met. All the 12 settings' ratio plot, averaged over the 100 simulated dataset, confirmed that choosing $T = min\{10, M\}$ as the number of moments of the probability distribution $\widetilde{\phi}$ (i.e. number of frequencies of frequencies to be considered) was the right choice for all the settings besides the discrete ones (10,11,12) for which we have fixed $T = min\{4, M\}$.

Setting	Distribution (ν)	$\mathbb{E}(n/N)$
	Gamma	
1	Ga(4,3.125)	0.90
2	Ga(4,1)	0.59
3	Ga(1,0.25)	0.20
	Gamma Mixture	
4	$0.5 \cdot$ Ga(2,1)+$0.5 \cdot$ Ga(2,2)	0.65
5	$0.5 \cdot$ Ga(2,1)+$0.5 \cdot$ Ga(4,1)	0.57
	Log-Normal	
6	LN(0.75,0.75)	0.82
7	LN(-0.5,2)	0.50
8	LN(-1,1)	0.36
	Log-Normal Mixture	
9	$0.5 \cdot$ LN(-0.5,1)+ $0.5 \cdot$ LN(0.5,1)	0.61
	Finite Mixture	
10	$0.8 \cdot \delta(1.2)+ 0.2 \cdot \delta(6.7)$	0.61
11	$0.89 \cdot \delta(0.5)+ 0.11 \cdot \delta(6.7)$	0.49
12	$0.8 \cdot \delta(0.2)+ 0.2 \cdot \delta(1.3)$	0.27

Table 4.5. Wang Simulation Settings (2010)

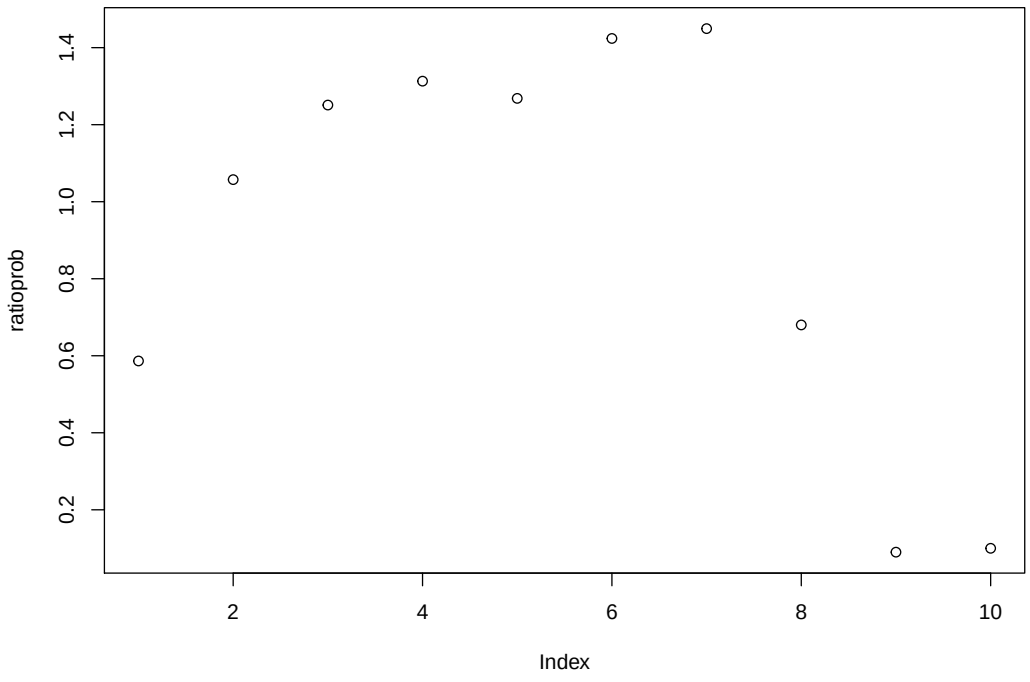


Figure 4.3. *Example*: Ratio plot for simulation setting 12 ($N = 10000$). It is clear that the monotonicity is broken at M=4

	sim	Low.CI	Median	Upp.CI	MSE	Coverage
rec	1	9726.93	9853.75	9987.46	172.66	0.41
rec01	1	9773.11	9933.25	10539.94	182.45	0.93
zm	1	9680.50	9900.00	11168.61	157.13	1.00
NP	1	9770.55	9852.75	10527.47	288.69	0.90
rec	2	9418.04	10085.25	16939.02	1762.84	0.95
rec01	2	9347.39	10271.00	16658.64	1382.46	0.91
zm	2	9282.11	9954.00	14261.97	1461.94	0.93
NP	2	9294.62	10010.25	13533.40	1769.64	0.92
rec	3	8153.15	10325.50	13467.77	893.83	0.90
rec01	3	6672.70	9600.50	11121.12	846.26	0.93
zm	3	6177.48	9785.00	42204.05	15461.56	0.97
NP	3	6098.36	8507.50	38951.56	13930.37	0.85
rec	4	8929.85	9361.25	10032.24	699.29	0.53
rec01	4	9059.62	9590.00	13497.19	914.12	0.92
zm	4	8983.55	9780.50	14844.32	640.59	0.96
NP	4	9059.30	9972.25	16222.75	6012.68	0.86
rec	5	8664.00	9316.25	9981.60	786.85	0.49
rec01	5	9217.98	9811.50	13663.82	1027.83	0.95
zm	5	9159.48	9837.50	12917.02	1876.25	0.99
NP	5	9141.46	9782.25	13718.87	3609.45	0.92
rec	6	85179.27	93949.25	182859.02	21376.90	0.97
rec01	6	83670.59	87481.25	94934.57	12642.18	0.27
zm	6	82953.49	108771.25	232861.05	49125.74	0.98
NP	6	8191.06	9615.25	24654.87	6609.45	0.91
rec	7	7462.70	8076.50	8703.76	1955.88	0.00
rec01	7	7102.16	7944.50	10401.17	2065.16	0.59
zm	7	6899.60	7524.75	14704.36	2425.49	0.89
NP	7	7832.42	9941.00	38180.50	13572.00	0.89
rec	8	7507.62	8745.75	14332.06	1784.10	0.90
rec01	8	7434.65	8368.75	13612.36	1811.29	0.82
zm	8	7544.91	10013.50	61021.50	18113.80	0.92
NP	8	7526.05	8933.04	30833.15	7711.69	0.86
rec	9	8660.64	9257.75	10091.49	895.10	0.54
rec01	9	8479.89	9162.20	12202.27	1119.21	0.86
zm	9	8282.15	8930.04	19122.51	1269.34	0.98
NP	9	8796.17	10184.50	22929.14	4160.41	0.91
rec	10	9831.66	9965.50	10117.20	107.72	0.83
rec01	10	9878.70	10026.00	10183.91	122.21	0.80
zm	10	9763.51	9920.25	12459.32	118.44	0.98
NP	10	9873.38	10001.25	10159.54	106.15	1
rec	11	10013.61	10457.75	11015.36	556.66	0.48
rec01	11	11146.46	11871.03	12373.65	1901.14	0
zm	11	9631.95	10123.75	11322.14	300.13	0.92
NP	11	9796.05	10195.50	10791.60	434.49	0.98
rec	12	7027.60	10041.50	28820.39	6448.24	0.86
rec01	12	7095.31	10295.25	67804.17	9610.93	0.85
zm	12	7033.75	10056.00	47068.42	13484.39	0.86
NP	12	9580.65	10092.10	11204.26	380.85	1

Table 4.6. N=10000. moment based maximization routine vs ratio regression vs Norris-Pollock maximization routine: Point and **Interval Estimates** with Coverage %

As stated in Wang (2010), NP suffers of significant instability. Furthermore, the computing is intensive, particularly when the extrapolation is large. In fact, it might take hours to compute the bootstrap confidence interval for just one dataset. That is why we did not replicate this simulation study for $N = 10000$ and $N = 1000000$. Among all the proposed estimators, rec01 is the one which performs better beating all the competitors in terms of MSE and coverage in all the settings. Let us now see the behaviour of the numerical lower bounds across different values of N.

	sim	lower.CI	median	upper.CI	MSE	coverage
Z	1	964.75	986.88	1063.28	21.03	0.90
DS	1	969.03	988.59	1229.90	20.72	0.95
Chao	1	964.59	983.46	1008.08	22.17	0.67
Z	2	853.55	914.97	1021.67	94.99	0.60
DS	2	861.87	919.11	1548.45	92.64	0.87
Chao	2	853.53	913.20	988.79	95.70	0.45
Z	3	487.67	609.75	820.04	385.56	0.20
DS	3	488.05	619.11	1191.86	377.53	0.59
Chao	3	487.67	609.75	805.30	386.88	0.15
Z	4	825.89	877.33	1221.80	126.96	0.75
DS	4	833.53	889.58	1661.89	119.01	0.87
Chao	4	824.72	868.01	922.15	133.92	0.05
Z	5	827.79	896.69	1048.72	116.29	0.61
DS	5	838.47	905.74	1697.06	109.93	0.91
Chao	5	826.61	890.88	968.73	119.59	0.27
Z	6	693.56	778.22	1041.19	231.38	0.52
DS	6	703.17	791.73	1991.77	219.07	0.84
Chao	6	692.71	772.24	872.75	240.41	0.05
Z	7	679.49	752.05	1303.25	254	0.73
DS	7	688.48	774.93	2081.16	235.72	0.86
Chao	7	673.56	729.10	803.77	272.31	0
Z	8	618.11	705.53	966.68	295.17	0.47
DS	8	625.79	718.37	2026.75	275.43	0.87
Chao	8	616.69	693.89	800.73	303.86	0.01
Z	9	799.55	874.03	1284.93	163.58	0.82
DS	9	813.24	885.56	2000.75	150.14	0.91
Chao	9	797.99	864.88	920.19	174.50	0.06
Z	10	957.98	1009.92	1245.57	42.03	0.89
DS	10	963.62	1008.98	1486.22	45.97	0.85
Chao	10	956.56	1004.59	1064.84	36.08	0.87
Z	11	864.39	1009.98	1543.94	184.38	0.86
DS	11	865.30	1015.74	2010.64	196.64	0.86
Chao	11	854.08	980.05	1138.96	114.42	0.80
Z	12	553.04	646.06	868.64	357.24	0.27
DS	12	560.14	668	1945.15	340.75	0.84
Chao	12	552.21	639.83	768.53	362.81	0.03

Table 4.7. N=1000. comparison between DS lower bound, Z lower bound and Chao lower bound: Point and **Interval Estimates** with Coverage %

	sim	Low.CI	median	Upp.CI	MSE	coverage
Z	1	9771.80	9913.69	10821.80	163.48	0.90
DS	1	9794.29	9958.91	11521.78	150.13	0.98
Chao	1	9757.17	9819.83	9886.53	186.53	0
Z	2	9008.50	9280.93	12751.98	734.90	0.91
DS	2	9055.14	9447.93	17015.18	583.26	0.98
Chao	2	8985.31	9197.36	9422.74	829.72	0
Z	3	5616.45	6112.25	7976.05	3888.37	0.36
DS	3	5673.83	6325.76	17522.11	3687.01	0.79
Chao	3	5610.17	6055.78	6567.53	3962.46	0
Z	4	8639.09	9201.97	11859.95	903.11	0.77
DS	4	8801.22	9366.43	13627.30	749.50	0.88
Chao	4	8593.75	8745.01	8910.06	1251.52	0
Z	5	8819.31	9343.77	12531.65	758.95	0.90
DS	5	8910.78	9542.32	14576.16	613.65	0.94
Chao	5	8783.51	8928.06	9082.35	1080.96	0
Z	6	7561.71	8534.99	14080.08	1535.40	0.85
DS	6	7745.62	8865.16	18056.94	1246.48	0.94
Chao	6	7481.60	7751.85	8052.06	2251.76	0
Z	7	7188.50	8019.21	11577.33	1997.65	0.71
DS	7	7388.43	8161.76	14231.03	1841.81	0.85
Chao	7	7042.01	7236.20	7437.73	2752.30	0
Z	8	6868.86	7933.44	13011.25	2151.86	0.76
DS	8	7053.92	8224.44	16966.50	1843.30	0.83
Chao	8	6786.28	7068.17	7377.08	2935.73	0
Z	9	8025.49	9151.89	11556.48	1209.85	0.81
DS	9	8347.25	9345.76	14098.78	1060.25	0.89
Chao	9	7811.75	8044.99	8297.16	1771.14	0
Z	10	9814.04	10017.71	10441.80	194.73	0.85
DS	10	9829.82	10018.26	10637.45	203.61	0.85
Chao	10	9797	9956.46	10127.30	118.16	0.84
Z	11	9380.79	9881.80	10655.50	380.40	0.90
DS	11	9441.89	9951.30	10886.55	375.87	0.93
Chao	11	9339.21	9757.90	10214.12	345.18	0.78
Z	12	6136.40	6513.11	19933.28	3458.06	0.77
DS	12	6249.84	7881.38	52212.59	2377.04	0.96
Chao	12	6101.68	6424.56	6780.48	3621.33	0

Table 4.8. N=10000. comparison between DS lower bound, Z lower bound and Chao lower bound: Point and **Interval Estimates** with Coverage %

	sim	lower.CI	median	upper.CI	MSE	coverage
Z	1	98034.98	99410.77	101760.47	1060.13	0.82
DS	1	98714.02	99464.45	103758.58	747.65	0.92
Chao	1	97949.24	98148.08	98355.14	1866.26	0
Z	2	91323.05	96368.29	104718.45	5862.86	0.78
DS	2	93851.77	97383.86	108632.43	4031.69	0.91
Chao	2	91122.12	91812.95	92513.10	8217.71	0
Z	3	58840.48	62182.77	109631.74	33089.98	0.54
DS	3	61127.84	72094.89	154698.79	26758.95	0.70
Chao	3	58504.81	59979.25	61504.39	40059.96	0
Z	4	87870.55	93903.06	113369.06	6480.09	0.80
DS	4	91147.13	94103.01	134678.43	6137.80	0.93
Chao	4	86924.35	87422.66	87930.53	12561.57	0
Z	5	89159.45	94866.24	107559.93	6143.72	0.74
DS	5	92261.98	94864.11	121425.64	5175.67	0.86
Chao	5	88734.33	89207.58	89688.01	10818	0
Z	6	77513.77	87854.93	117844.88	12931.75	0.74
DS	6	82526.15	88042.13	140829.72	12060.48	0.89
Chao	6	76267.51	77148.71	78051.68	22957.24	0
Z	7	78564.02	82524.41	119841.69	17280.74	0.72
DS	7	78191.18	82735.54	156926.28	17079.61	0.84
Chao	7	71815.73	72416.59	73029.50	27563.70	0
Z	8	73518.68	83529.35	125318.82	17351.78	0.74
DS	8	77294.24	83464.15	161519.43	16678.55	0.88
Chao	8	69580.56	70500.80	71447.09	29412.86	0
Z	9	90615.47	95231.62	120508.26	8145.89	0.86
DS	9	93108.21	95204.07	146633.04	7929.87	0.89
Chao	9	90044.56	90446.39	90855.47	15409.83	0
Z	10	99032.39	99928.74	101120.52	562.78	0.90
DS	10	99330.28	100102.55	101478.70	506.02	0.89
Chao	10	98962.72	99478.41	100003.26	599.05	0.52
Z	11	96767.84	99379.40	102100.84	1633.88	0.86
DS	11	98067.67	99936.51	102981.50	1179.69	0.87
Chao	11	96562.83	97938.58	99348.74	2177.29	0.24
Z	12	62727.54	84588.23	177132.48	27627.55	0.95
DS	12	78768.23	101059.09	230884.67	18170.55	0.99
Chao	12	62458.16	63478.40	64536.10	36536.08	0

Table 4.9. N=100000. comparison between DS lower bound, Z lower bound and Chao lower bound: Point and **Interval Estimates** with Coverage %

$\hat{N}$	sim	lower.CI	median	upper.CI	MSE	coverage
Z	1	991373.35	995388.55	1024956.21	4873.81	0.90
DS	1	991320.58	995284.51	1051943.92	4905.68	0.97
Chao	1	980726.76	981373.94	982031.70	18562.51	0.00
Z	2	917995.67	974144.80	1051964.30	40038.77	0.78
DS	2	959482.88	974008.25	1116276.41	26295.82	0.84
Chao	2	915822.12	918000.94	920203.27	82177.40	0.00
Z	3	599573.11	738607.95	898049.97	278377.10	0.36
DS	3	693359.46	743510.75	1066309.36	251971.17	0.58
Chao	3	595753.69	600445.58	605186.85	399811.94	0.00
Z	4	933589.80	960860.28	1144605.96	42450.05	0.80
DS	4	931032.51	961214.61	1285509.30	42582.03	0.90
Chao	4	873201.58	874807.97	876399.43	125123.49	0.00
Z	5	943795.01	962697.84	1195927.54	37463.99	0.87
DS	5	941957.82	964084.56	1428264.49	37584.22	0.92
Chao	5	890704.96	892197.25	893704.92	107891.77	0.00
Z	6	869615.40	919311.50	1204454.81	85439.13	0.78
DS	6	865186.67	919972.38	1420022.71	85759.18	0.87
Chao	6	768355.20	771188.25	774072.18	228769.28	0.00
Z	7	826361.26	867859.07	1075853.54	129127.29	0.63
DS	7	821779.29	868176.81	1320122.47	129119.18	0.84
Chao	7	722925.54	724829.18	726779.10	275179.40	0.00
Z	8	816619.97	868329.63	1174023.10	128533.99	0.71
DS	8	811930.82	868626.86	1392123.58	127962.84	0.80
Chao	8	702335.69	705244.81	708195.23	294426.56	0.00
Z	9	944437.52	956077.49	1045588.74	66027.62	0.65
DS	9	943616.27	955977.68	1093807.17	65834.03	0.78
Chao	9	903403.60	904676.39	905958.11	167589.29	0.00
Z	10	993730.59	998207.39	1003157.05	3674.98	0.85
DS	10	997798.53	1000152.00	1004130.71	1770.53	0.87
Chao	10	993521.49	995151.25	996824.01	4987.05	0.00
Z	11	976676.89	993618.30	1006709.07	13879.76	0.85
DS	11	994079.92	1000578.14	1009448.55	3876.04	0.90
Chao	11	975785.44	980198.14	984692.57	20120.61	0.00
Z	12	632735.28	640101.24	1138081.55	266759.99	0.86
DS	12	914439.22	1003893.46	1183627.86	57834.68	0.96
Chao	12	631440.94	634701.17	637963.02	364912.49	0.00

Table 4.10. N=1000000. Numerical Estimators: Point and **Interval Estimates** with Coverage

We finally run the comparison between DS lower bound and likelihood-based lower bounds (rec and rec01)

	sim	Low.CI	Median	Upp.CI	MSE	Coverage
rec	1	97899.23	98511.11	99994.49	1507.02	0.52
rec01	1	98362.56	98986.39	100058.25	1040.87	0.54
DS	1	98714.19	99464.36	103758.22	748.56	0.92
rec	2	95247.11	98785	129184.38	5231.70	0.86
rec01	2	97178.16	101328.25	124033.27	6777.18	0.88
DS	2	93851.77	97383.86	108632.43	4031.69	0.91
rec	3	65142.73	96252	126579.51	15728.54	0.97
rec01	3	70397.55	86960.25	218363.20	40464.59	0.88
DS	3	61127.84	72094.89	154698.79	26758.95	0.70
rec	4	93594.74	97929.75	113790.67	6036.814	0.87
rec01	4	92566.25	95091.75	105712.77	4922.737	0.76
DS	4	91147.13	94103.01	134678.43	6137.80	0.93
rec	5	93774.39	97288.25	108481.91	4494.47	0.81
rec01	5	94391.05	97669.25	111537.35	3894.45	0.86
DS	5	92261.98	94864.11	121425.64	5175.67	0.86
rec	6	85179.27	93949.25	182859.02	24376.90	0.97
rec01	6	83670.59	87481.25	94934.57	12642.18	0.27
DS	6	82526.15	88042.13	140829.72	12060.48	0.89
rec	7	80440.39	85795.75	132677.22	14604.84	0.92
rec01	7	75426.62	77354.25	83072.72	22643.05	0.05
DS	8	77294.24	83464.15	161519.43	16678.55	0.88
rec	8	85891.78	103961.25	117918.90	10171.42	0.96
rec01	8	75716.93	82227.25	98266.92	17706.40	0.48
DS	8	77294.24	83464.15	161519.43	16678.55	0.88
rec	9	94395.40	96699.00	100155.38	23342.80	0.40
rec01	9	88517.55	93392.50	96984.50	8554.45	0.20
DS	9	93108.21	95204.07	146633.04	7929.87	0.89
rec1	10	99310.94	99756.50	100198.09	428.93	0.67
rec011	10	99799.05	100292.50	100925.65	433.81	0.77
DS	10	99330.28	100102.55	101478.70	506.02	0.89
rec	11	95174.71	97029.75	98625.89	3109.53	0.13
rec01	11	112718.98	115825.50	118559.96	15984.44	0.00
DS	11	98067.67	99936.51	102981.50	1179.69	0.87
rec	12	81904.67	99229.25	148893.69	16387.57	0.96
rec01	12	82378.60	101314.25	237266.25	19185.43	0.96
DS	12	78768.23	101059.09	230884.67	18170.55	0.99

Table 4.11. N=100000. Numerical lower bound (DS) vs Likelihood-based lower bounds

We start from commenting the results coming from the quadrature lower bounds. It is quite evident that the DS estimator is closer to the true values of N in all the different scenarios ($N = 10000, N = 10000, N = 1000000$). Our new proposal (Z estimator) performs better than Chao lower bound but it is under-performing in simulation setting 3 (where the unobserved units are 80% of the total units) and in simulation setting 12 with respect to DS lower bound.

As concerns the comparison between likelihood-based lower bounds (rec and rec01) and the numerical lower bound DS, we can be satisfied. Our proposals (rec and rec01) compete well with their competitor (DS), although in some settings they tend to be bias upward. It is worth recalling that we were not able to draw 20000 bootstrapped samples as done in Daley and Smith (2016) and this might partially explain the instability experienced in some settings. One possible solution might be to enhance the maximization routine using $C++/Fortran$. We will see in the next chapter how a Bayesian approach, leveraging on the same statistical model 3.14, can produce better estimates in terms of efficiency and coverage.

Chapter 5

Moment-based Bayesian inference of population size

In this final chapter we present an alternative inferential approach for estimating the population size within the framework of Poisson mixtures for zero truncated count data. Indeed, we are going to rely on the same model approximation introduced in chapter 3 which will allow us to exploit the useful reparametrization related to the truncated moment space. Let us start by recalling the original moment based likelihood structure

$$L(N, \boldsymbol{s}; \boldsymbol{f}_+) = \binom{N}{f_0\ f_1\ f_2 \ldots} (s_0)^N \prod_{j=1}^{\infty} \left(\frac{\widetilde{s}_j}{j!}\right)^{f_j} = \binom{N}{f_0\ f_1\ f_2 \ldots} \prod_{j=1}^{\infty} \left(\frac{\widetilde{s}_j}{j! \sum_{k=0}^{\infty} \frac{\widetilde{s}_k}{k!}}\right)^{f_j}. \tag{5.1}$$

We have seen that one can avoid the infinite summation by defining, from the expression above, an approximation which represents a flexible parametric distribution

$$\widetilde{p}(\boldsymbol{f}_+; N, \widetilde{\boldsymbol{s}}_T) = \widetilde{L}(N, \widetilde{\boldsymbol{s}}_T; \boldsymbol{f}_+) = \binom{N}{f_0\ f_1\ \ldots f_T} \prod_{j=0}^{T} \left[\frac{\widetilde{s}_j}{j! \sum_{k=0}^{T} \frac{\widetilde{s}_k}{k!}}\right]^{f_j}. \tag{5.2}$$

We remark that the following quantities

$$g(j; \widetilde{s}_j) = \frac{\widetilde{s}_j}{j! \sum_{k=0}^{T} \frac{\widetilde{s}_k}{k!}} \qquad j = 0, 1, 2, \ldots, M$$

can be regarded as approximations for the original $P_j = \frac{s_j}{j!}$. Although we can exploit all the information provided by the observed data by setting $T = M$, it could still make sense considering $T \neq M$, especially when there is a large value of M and possibly little values of $f_j > 0$ in the right tail of the observed frequencies of frequencies. However, in most of the real data application, we have used $T = M$ and therefore we prefer leaving M for the illustration of this chapter.

From now on, our Bayesian model setup can be formalized by the the following ingredients: the likelihood function

$$\widetilde{L}(N, \widetilde{\boldsymbol{s}}_M; \boldsymbol{f}_+) = \binom{N}{f_0\ f_1\ \ldots f_M} \prod_{j=0}^{M} \left[\frac{\widetilde{s}_j}{j! (\sum_{k=0}^{M} \frac{\widetilde{s}_k}{k!}}\right]^{f_j}$$

and a suitable prior distribution

$$\pi(N, \widetilde{\boldsymbol{s}}_M)$$

on the parameter space $\mathbb{N} \times \mathcal{S}_1^{[M]}$ where $\mathcal{S}_1^{[M]}$ is the truncated moment space $\mathcal{S}_1^{[M]}$

$$\mathcal{S}_1^{[M]} = \left\{ (\widetilde{s}_0, \widetilde{s}_1, \ldots, \widetilde{s}_M) : \widetilde{s}_k = \int_0^\infty \lambda^k d\widetilde{\phi}(\lambda), \widetilde{\phi} \in \mathcal{P}([0,\infty)) \right\}$$

and $\mathcal{P}([0,\infty))$ is the class of probability measures with support in $[0,\infty)$. The ordinary truncated moment space $\mathcal{S}_1^{[M]}$ is a constrained M-dimensional convex body and might not be easy to treat especially if we need to find out a suitable approximation of the posterior distribution by simulation. Alunni Fegatelli and Tardella (2018) considered the same idea but they then decided to rely on a reparametrization of the ordinary moments in terms of the so-called canonical moments (2.5) by restricting the attention to probability measures $\widetilde{\phi}_u$ on a compact interval $[0,u]$ for a finite $u > 0$. The canonical moments, as seen in the previous paragraphs, are defined when the measure $\widetilde{\phi}$ is in $\mathcal{P}([0,1])$ which is the case of the Hausdorff Moment problem. In order to avoid to further approximate the underlying model by compactification we decided to find an alternative strategy to directly reparametrize the ordinary moments for $\widetilde{\phi} \in \mathcal{P}([0,\infty))$. We recall that, for probability measures in $\mathcal{P}([0,\infty))$, it exists a one-to-one mapping between the ordinary moments and the so called recurrence coefficients defined in (2.24):

$$\psi_M \ : (\widetilde{s}_1, \ldots, \widetilde{s}_M) \rightarrow \left(a_1, b_1 \ldots, a_{M/2}, b_{M/2}\right)$$

where $\left(a_1, b_1 \ldots, a_{M/2}, b_{M/2}\right)$ can be derived using either the QD algorithm or the Chebyshev algorithm. However, as it has been noted in Wheeler (1974), the QD algorithm is more likely to break down due to degeneracies therefore we prefer to use the Chebyshev algorithm. We then propose to do all simulations in this transformed parameter space, reparametrize back to the constrained parameter space $\mathcal{S}_1^{[M]}$ (using the reverse Chebyshev algorithm) and then implement standard MCMC algorithms which only requires the numerical evaluation of the likelihood function and the prior up to proportionality constants. As explained before, in order to set up a fully Bayesian approach, we need to elicit a suitable prior distribution for the parameter vector $\boldsymbol{\vartheta} = (N, \boldsymbol{s}_M)$. In the next paragraph we will illustrate how one can construct a suitable default prior distribution on the parameter space $\mathbb{N} \times \mathcal{S}_1^{[M]}$.

5.1 Non informative Bayesian inference

The first question we might want to ask ourselves is why we should adopt non-informative priors distribution. The primary reason is that they allow us to make automatic Bayesian statistical analysis in the absence of genuinely available subjective information and in any case favouring a choice which can be perceived as more easily shareable among different users. The term "non-informative", which is best known in the literature , is a bit misleading since it suggests the possibility that some of the probability laws can contain no information: in reality they are probability

distributions that produce Bayesian inferences through the final distribution of the parameter of interest which often enjoys also good frequentist properties. As far as prior specification for a Bayesian analysis is concerned, in this section we review a range of possible choices for a prior distribution for the parameters N and $\widetilde{\boldsymbol{s}}_M$.
First of all, it is worth noting that, for fixed N, the likelihood (5.2) is multinomial in terms of the probabilities $\mathbf{g}_M = (g(0;\widetilde{\boldsymbol{s}}_M),\ldots,g(M;\widetilde{\boldsymbol{s}}_M))$ which are in turn one-to-one related to $\widetilde{\boldsymbol{s}}_M$. Some naively intuitive choices could either be the flat prior on the truncated ordinary moment space $\mathcal{S}_1^{[M]}$ or the flat prior on the recurrence coefficients space. However, if we are looking for prior specification which is theoretically well grounded and possibly invariant under reparametrization, one could start from a Jeffreys prior on the constrained multinomial probabilities $\mathbf{g}_M$ which has the additional appealing feature of usually leading to good frequentist properties of the corresponding posterior analysis. It is known that the Jeffreys'prior for an unconstrained multinomial parameter vector is a Dirichlet distribution with parameters 1/2; one can then argue that for the count frequency probabilities which are constrained on a proper convex body contained in the M-dimensional simplex the same functional form of the Jeffreys' prior is preserved up to a different normalizing constant. So we have:

$$\pi_j(g(0;\widetilde{\boldsymbol{s}}_M),\ldots,g(M;\widetilde{\boldsymbol{s}}_M)) \propto \prod_{j=0}^{M} [g(j;\widetilde{\boldsymbol{s}}_M)]^{-\frac{1}{2}}\,. \tag{5.3}$$

Of course, we will need to transform it back in terms of a default prior on $\mathcal{S}_1^{[M]}$ using the appropriate Jacobian for the mapping $\mathbf{g}_M \to \widetilde{\boldsymbol{s}}_M$. To simplify notation and arguments related to the constrained simplified structure of the constrained multinomial probabilities, we follow the same approach presented in Alunni Fegatelli and Tardella (2018) and restrict the attention to only M components of the truncated moment sequence, namely $x_j = g(j,\widetilde{s}_j)$, $y_j = \frac{\widetilde{s}_j}{j!}$, $\boldsymbol{x} = (x_1,\ldots,x_M)$, $\boldsymbol{y} = (y_1,\ldots,y_M)$. We can then express the M count frequencies as function of $\boldsymbol{y}$

$$\boldsymbol{x} = f(\boldsymbol{y})$$

and get:

$$x_k = \frac{y_k}{\sum_{i=0}^{M} y_i} = \frac{y_k}{D_y} \qquad k = 1,...,M.$$

One notes that both vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ can be completed by x_0 and y_0 using the constraints:

$$\sum_{k=0}^{M} x_k = 1 \;\&\; y_0 = 1.$$

We can then re-write the default prior in terms of x_k

$$\pi_J(\boldsymbol{x}) = \prod_{k=0}^{M} x_k^{-\frac{1}{2}}.$$

The corresponding prior for $\boldsymbol{y}$ is then:

$$\pi_J(\boldsymbol{y}) = \pi_J(f(\boldsymbol{y})) * |J_f(\boldsymbol{y})|$$

Where J is the Jacobian of the one-to-one mapping whose diagonal elements are $\frac{D_y - y_i}{D_y^2}$ $(i = 1, \ldots, M)$ and the extra-diagonal elements are $-\frac{y_i}{D_y^2}$ $(\forall i \neq j)$. The simulation within the moment space can be simplified by considering the recurrence coefficients instead of the ordinary moments which of course require an additional reparametrization mapping ψ_M and hence the evaluation of another Jacobian. In chapter 2, we have illustrated that the Jacobian determinant for the mapping $\psi_M : \left(a_1, b_1 \ldots, a_{M/2}, b_{M/2-1}\right) \rightarrow (s_1, \ldots, s_M)$ is equal to :

$$detD\psi_M^{\mathbb{R}} = \prod_{i=1}^{M-1} b_i^{M-i}. \tag{5.4}$$

Finally, we get:

$$\pi_R(\widetilde{\boldsymbol{s}}_M) = \pi_J(f(g(\widetilde{\boldsymbol{s}}_M)) * J_f(g(\widetilde{\boldsymbol{s}}_M)) * J_{\psi_M}(\widetilde{\boldsymbol{s}}_M). \tag{5.5}$$

To complete the prior construction on the whole parameter space, we need to provide a default prior for N. In Alunni Fegatelli and Tardella (2018), three different default prior distributions are considered: uniform, $1/N$ and Rissanen prior (one of the default options for eliciting a proper non informative prior distribution on the unknown population size with tails of the order between $1/N$ and $1/N^2$). In this work, we will adopt the same priors.

5.2 Simulation study

In order to test the validity of our Bayesian approach, we consider the same 12 simulation settings proposed in Wang (2010). We recall that for each setting a different mixing distribution is fixed and 100 simulated datasets are drawn and used to repeat the estimation procedure. We have fixed the population size to be $N = 1000$, N=10000 and N=100000. Bias and mean square error of the estimates are re-evaluated averaging the results over the simulated datasets. Coverage of interval estimates is obtained in the same way. We have fixed the number of moments of the probability distribution $\widetilde{\phi}$ to be $T = min\{10, M\}$ (as it was done in Alunni Fegatelli and Tardella (2018)). Although we have evaluated various prior choices for N we have reported here only the results obtained from the Rissanen prior. The building block of the MCMC algorithm for this implementation is the well known Adaptive Rejection Metropolis Sampling (ARMS). ARMS is widely used within Gibbs sampling, where automatic and fast samplers are often needed to draw from full-conditional densities.

sim	M.hat	s	COV.hpd	MSE.3	MSE.2	MSE.1	HPD.l
1	1010	19.21	98	21.60	20.18	22.57	94.04
2	1018	55.48	100	56.13	58.17	65.58	367.37
3	779	126.65	85	318.03	292.27	296.89	1039.11
4	925	38.35	86	88.65	91.88	81.30	253.75
5	948	38.31	95	68.87	72.10	62.74	247.54
6	866	66.32	90	175.09	172.08	148.61	487.96
7	860	62.29	74	176.31	175.31	155.30	421.89
8	844	71.45	86	211.29	204.19	176.38	546.49
9	910	74.91	85	130.58	131.92	117.18	292.01
10	1067	43.13	84	80.14	78.36	86.79	221.86
11	1017	135.80	90	137.12	143.09	163.37	481.41
12	688	61.33	24	343.15	335.03	313.99	413.56
sim	M.hat	s	COV.hpd	MSE.3	MSE.2	MSE.1	HPD.l
1	10094	94.15	100	137.25	115.43	143.27	640.92
2	10645	366.81	100	754.89	693.06	955.77	4078.33
3	7943	737.77	86	2278.05	2268.54	1829.24	6747.11
4	9978	346.42	99	347.29	367.65	372.87	2433.15
5	10068	322.43	100	342.61	335.61	405.46	2546.68
6	10194	617.36	96	633.31	717.14	760.93	4177.05
7	8638	465.85	83	1404.65	1450.92	1304.65	2802.70
8	9388	524.32	99	825.10	917.58	702.86	4315.62
9	9864	437.61	91	416.06	440.99	422.47	2484.95
10	10100	102.54	82	150.96	151.19	152.66	408.07
11	10480	327.34	78	600.88	602.25	642.24	1519.10
12	9177	622.36	99	1046.79	936.30	825	4586.19
sim	M.hat	s	COV.hpd	MSE.3	MSE.2	MSE.1	HPD.l
1	101512	718.03	95	1753.37	1448.90	1813.31	6065.98
2	108300	3994.82	87	9890.39	9383.35	11294.77	30976.58
3	94863	6497.69	94	8639.41	9310.03	7686.83	42712.30
4	101214	3575.99	100	4119.27	3784.09	5052.57	24772.54
5	100858	2915.88	100	3338.58	2885.07	4116.39	20978.01
6	101018	5852.22	99	6366.20	7030.52	8118.81	43999.67
7	92584	3632.17	95	7647.61	8199.30	6741.79	29629.47
8	97847	5009.59	97	5384.30	6619.94	5405.93	32302
9	101688	4590.57	100	3932.90	3996.77	4668.39	22961.70
10	100463	415.64	67	650.09	650.27	651.68	1245.46
11	102678	1489.82	25	3277.79	3282.64	3300.01	4115.08
12	119888	27347.44	39	37019.90	37342.08	37220.35	22182.67

Table 5.1. *Evaluating the median bias, mean squared error, interval length and 95% confidence interval coverage in 12 simulation settings,* $N = 1000, N = 10000, N = 100000$

Overall, these results are satisfactory. In particular, we want to comment on the inferential performances with growing N. The first consistent evidence is that the point estimation bias reduces with N in almost settings whereas the mean square error has a more alternating pattern although, averaging out all the

errors, the Bayesian estimates get closer to the underlying population size. HPD coverage often increases reaching the frequentist nominal level corresponding the posterior probability content. However, this evidence fails in the last three settings corresponding to the discrete mixing distributions. One possible explanation is that, in the discrete mixing probabilities case, with few support points the number of moments characterizing uniquely the discrete distribution are often smaller than the number of moments used in the Bayesian model. An improved strategies for setting T could enhance the performances. However, one should remark that Bayesian inference might be less performing in these last simulation settings where the true moments are at the boundary of the truncated moment space.

5.3 Comparison between different methods

We start by comparing our Bayesian approach (BPM) with its closer Bayesian competitor (TAF) proposed in Alunni Fegatelli and Tardella (2018) and the Poisson compound gamma Estimator (PCG) proposed in Wang (2010), which is considered to date one of the best performing classical estimator. All the estimators compared in this round aim at directly estimating $\hat{N}$ although with different methodologies. Let us first compare the three estimates with N equal to 10000.

Sim	median PCG	MSE PCG	Median TAF	MSE TAF	Median BPM	MSE BPM‘
1	9966	247	10177	202	10094	137
2	10018	438	11974	1970	10645	755
3	10005	586	9645	709	7943	2278
4	9932	772	10139	348	9978	347
5	10012	637	10288	389	10068	342
6	9963	615	10180	955	10194	633
7	8726	1912	9080	1033	8638	1400
8	10715	1630	9425	743	9388	825
9	10014	1137	10034	550	9864	416
10	10163	198	10886	901	10100	152
11	10338	1552	12431	2396	10480	601
12	10996	1839	12150	2565	9177	1046

Table 5.2. Comparison between Wang PCG, TAF and new version of BPM for $N = 10000$

Surprisingly, our new BPM behaves well in all the settings (however occasionally it might be beaten by the others in few settings).

We then run a comparison, in terms of MSE, between these three estimators and the three lower bounds obtained in the dedicated chapter.

Sim	MSE PCG	MSE TAF	MSE BPM	MSE CHAO	MSE Z	MSE DS
1	247	202	137	186	163	150
2	438	1970	755	830	735	583
3	586	709	2278	3962	3888	3687
4	772	348	347	1252	903	749
5	637	389	342	1080	759	614
6	615	955	633	2252	1535	1535
7	1912	1033	1400	2752	1998	1842
8	1630	743	825	2936	2152	1843
9	1137	550	416	1771	1210	1060
10	198	901	152	118	194	203
11	1552	2396	601	345	380	376
12	1839	2565	1046	3621	3458	2377

Table 5.3. Comparison Table in terms of MSE: N=10000

The same comparative analysis is done for $N = 100000$, this time including only BPM, TAF, DS and PCG.

sim	MSE BPM	MSE DS	MSE PCG	MSE TAF
1	1753	748	2791	2246
2	9890	4032	3881	11011
3	8639	26759	9232	33312
4	4119	6138	6749	4509
5	3339	5176	6946	6436
6	6366	12060	8082	9037
7	7648	17080	11431	9996
8	5384	16679	11214	7240
9	3933	7930	19624	3180
10	650	506	1684	7062
11	3278	1180	2955	24535
12	37020	18171	78960	48640

Table 5.4. Comparison Table in terms of MSE: N=100000

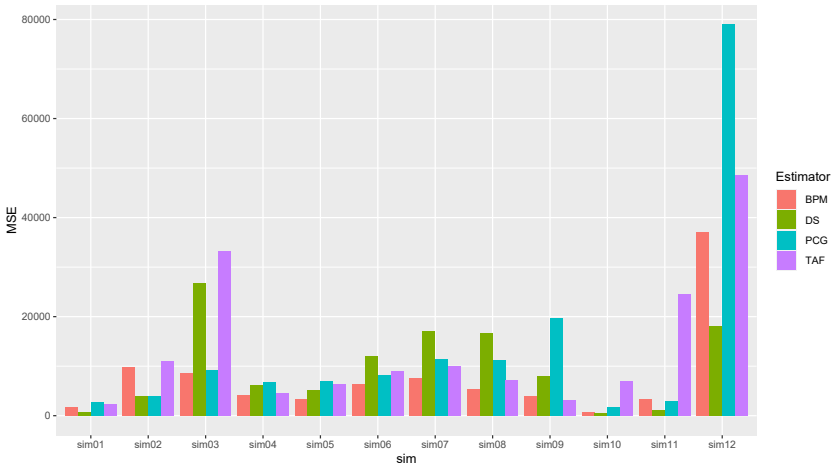


Figure 5.1. $N = 100000$ comparison in terms of MSE between BPM, PCG, TAF and DS

The BPM competes extremely well with all its competitor, beating them in 8 settings over 12, although occasionally it can be beaten in terms of efficiency.

5.4 Real data analysis

We investigate the effectiveness of our proposed estimator with several benchmark datasets introduced in the initial part of this work.

Traffic Data

The first dataset we analyze is probably one of the most famous and it is known as Traffic Data. It was originally studied in Simar (1976) and recently re-analyzed in Böhning et al. (2005) and Wang (2010). Data represent the accident counts submitted to "La Royale Belge Insurance Company" during a particular year. The real value for $N(9461)$ is known for this dataset and it is the total number of insurance policies covering both “business” and “tourist” automobiles. The complete frequency counts in table 5.5 show that the proportion of the unobserved units is very high.

k	1	2	3	4	5	6	7	n
(f_k)	1317	239	42	14	4	4	1	1621

Table 5.5. Traffic Data

Method	$\hat{N}$	$low_{0.025}$	$upp_{0.975}$
BPM	7,933	5,002	21,720
TAF	9835	4969	60666
DS	5,590	4,828	13,212
Z	5,455	4,902	11,461

Table 5.6. Traffic data: Results

Since we know the true sample size, we can assess the goodness of the estimators by looking at their "closeness" to N and/or if the confidence interval contains it. Besides Chao, for all the other proposal N is contained in the confidence interval. In terms of MSE, the closest estimator is TAF, followed by BPM. Let us now look at some convergence diagnostics for BPM, as the one for TAF was already presented in Alunni Fegatelli and Tardella (2018). We draw autocorrelation and traceplot for both N and the first recurrent coefficient a_1 and only autocorrelation for a_1.

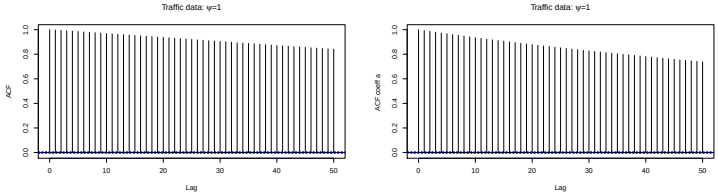


Figure 5.2. Autocorrelation N and a_1

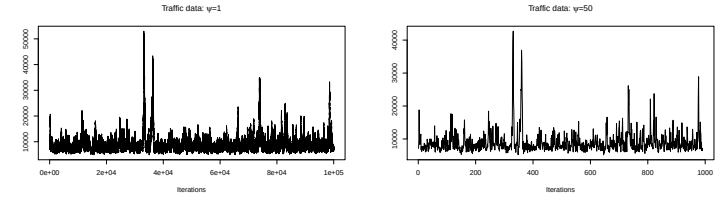


Figure 5.3. Traceplot N

As in Alunni Fegatelli and Tardella (2018) the ACF plot is not satisfying; we adopt the same strategy mentioned in their paper and we redraw the ACF and traceplot considering a thin factor $\psi = 50$ leading to 2000 iterations.

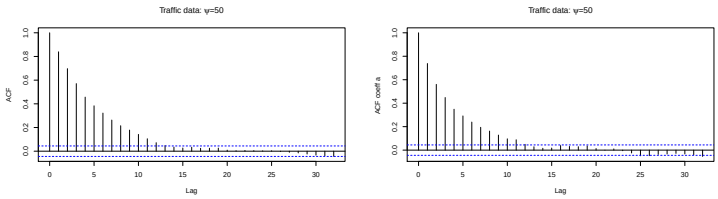


Figure 5.4. Autocorrelation N and a_1

The resulting ACF looks more reasonable.

Polyps Data

From medical research experiences it is well recognized that diagnosing adenomatous polyps can be subjected to undercount due to misclassification at colonoscopy. In this work, we have used data from Alberts et al. (2000) where, in order to evaluate the recurrence of colorectal adenomatous polyps, subjects with previous history of colorectal adenomatous polyps are allocated to one of two treatment groups, low fiber and high fiber. Polyps data-frequency distribution of recurrent adenomatous polyps per patient, by treatment group is reported in table 5.7 . 584 subjects are allocated to the low fiber group whereas 722 subjects are allocated to the high fiber group.

k	1	2	3	4	5	6	7	8	9	10	11	12	n
f_k^{low}	145	66	39	17	8	8	7	3	1	0	2	3	299
f_k^{high}	144	61	55	37	17	5	4	6	5	1	1	5	341

Table 5.7. Polyps data-frequency distribution

Method	$\hat{N}$	$low_{0.025}$	$upp_{0.975}$
BPM	575	406	846
TAF	520	434	660
DS	546	478	6078
Z	546	488	884

Table 5.8. Polyps high data: Results

Method	$\hat{N}$	$low_{0.025}$	$upp_{0.975}$
BPM	510	389	636
TAF	522	409	743
DS	487	423	2156
Z	496	437	1262

Table 5.9. Polyps low data: Results

All the estimates are quite similar and reasonable. Convergence diagnostics for N and a_1 confirms a good implementation of the MCMC procedures.

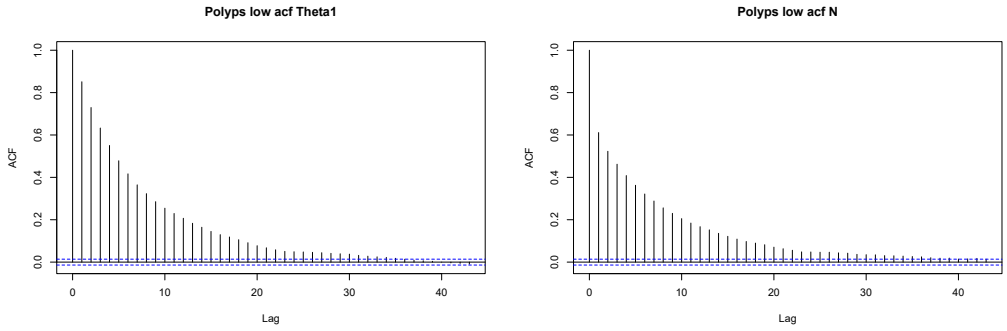


Figure 5.5. Autocorrelation N and a_1: Polyps Low Data

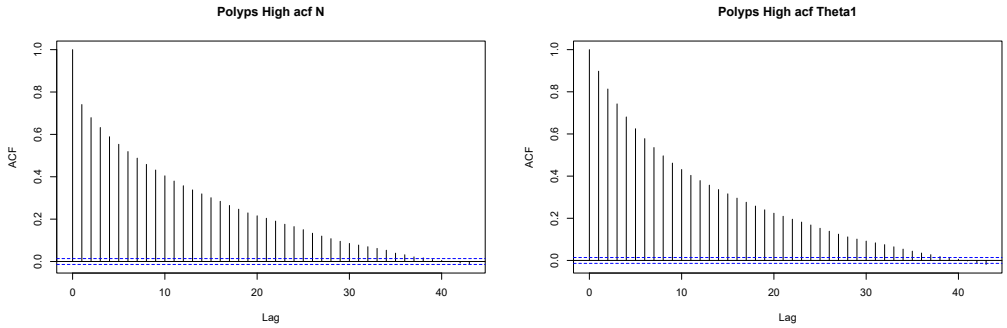


Figure 5.6. Autocorrelation N and a_1: Polyps High Data

The strong autocorrelation might be due to the strong dependency among the N and a_1. In any case, the estimated value of N is quite good and the interval estimates are wide enough to possibly include the true value of N.

Cholera

One of the oldest example of estimation of the unobserved units is given by Kermack et al. (1927). The authors analysed the number of individuals with cholera in 223 households in a village in India.

k	1	2	3	4	n
Cholera (f_k)	32	16	6	1	55

Table 5.10. Cholera data-frequency distribution

Method	$\hat{N}$	$low_{0.025}$	$upp_{0.975}$
BPM	105	75	170
TAF	125	67	1459
DS	88	73	195
Z	89	80	190

Table 5.11. Cholera data: point and interval estimates

All the estimates are reasonable even if the interval for TAF is very wide. We then check the diagnostics for N and a_1.

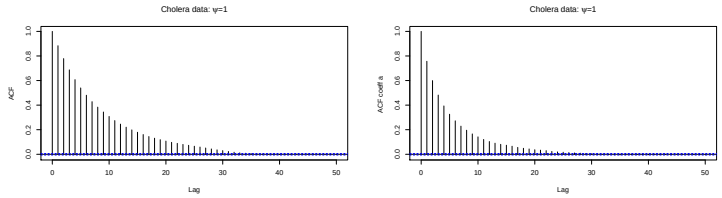


Figure 5.7. Autocorrelation N and a_1: EST Data

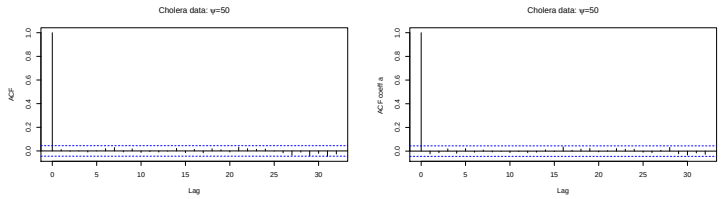


Figure 5.8. Autocorrelation N and a_1: EST Data, $\Psi = 50$

The resulting ACF for both N and a_1 looks reasonable.

Expressed Sequence Tag (EST)

This popular dataset was first introduced in Mao and Lindsay (2003). An EST is a partial sequence identifying a gene locus; ESTs are generated by sequencing randomly selected clones in a cDNA library made from an mRNA pool. In the experiment, 2586 possibly replicated sequence tags were detected from which $n = 1825$ genes were found.

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	23	27	n
f_k	1434	253	71	33	11	6	2	3	1	2	2	1	1	1	2	1	1	1825

Table 5.12. EST data-frequency distribution

Method	$\hat{N}$	$low_{0.025}$	$upp_{0.975}$
BPM	9872	6079	19153
DS	7169	5752	18918
Z	7160	5830	14234
TAF	9115	6099	16264

Table 5.13. Est data: point and interval estimates

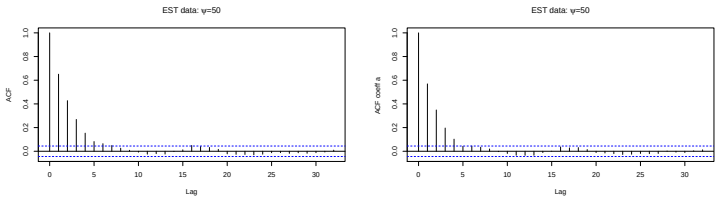


Figure 5.9. Autocorrelation N and a_1: EST Data, $\Psi = 50$

DS and Z are lower bounds and are therefore more conservative. the two Bayesian proposals are closer with BPM providing a larger interval. We check again the diagnostics for N and a_1.

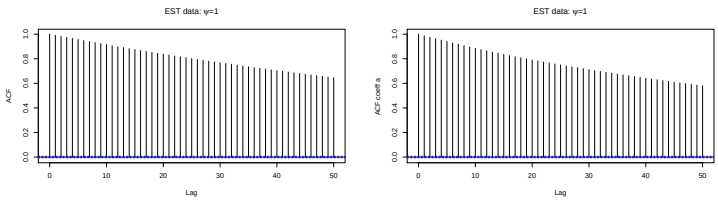


Figure 5.10. Autocorrelation N and a_1: EST Data

There is an evident autocorrelation for both N and a_1 as it happened for *traffic*. We thus considered a thin factor $\psi = 50$ so that we are left with only 2000 iterations and the resulting ACF looks much better.

Chapter 6

Final remarks

In this thesis, we have dealt with modeling individual heterogeneity within Poisson count distribution in the absence of zero counts. In the first part, we have investigated identifiability issues for both conditional likelihood and unconditional likelihood. For the unconditional likelihood, we have reparametrized the statistical model in terms of the moment sequence of a suitable finite measure and proved model identifiability. On the other hand, for the conditional likelihood, identifiability issues are not solved even when we consider a suitable moment reparametrization. Leveraging on the lower semi-continuity of P_0, we have shown that a sharpest lower bound for P_0 can always be derived. We have then verified that there is not a unique best and safe way to numerically compute the sharpest lower bound corresponding to any truncated sequence $\overline{\boldsymbol{P}}_M$ or, equivalently, $\boldsymbol{s}_M$ and we have presented four methods to approach its evaluation:

- algebraic lower bound evaluation;
- quadrature lower bound evaluation;
- lower bound evaluation through sequential moment condition check;
- likelihood-based lower bound evaluation.

For the quadrature lower bounds, extensively treated in Daley and Smith (2016), we proposed a new method to check the moment admissibility. The performances of our proposal Z were not satisfactory and one possible improvement, as mentioned in the first simulation study, might be considering stabler version of the QD algorithm. On the same topic, we have given an innovative contribution on a possible sequential evaluation of the sharpest lower bound. In the same chapter, we have tested the numerical accuracy of the algebraic lower bound approach, the quadrature lower bound approach and our new proposal obtaining encouraging results.

In the final part of the thesis we have adopted a Bayesian approach. Starting from a reparametrization of the likelihood function (5.2) in terms of the first M ordinary moments corresponding to a probability measure $\widetilde{\phi}$, we have first truncated its infinite sequence of moments to the first M moments using an explicit renormalization which formally resembles the original likelihood. Moreover, we have reparametrized the ordinary moments of $\widetilde{\phi}$ in terms of the so-called recurrence

coefficients allowing for an easier MCMC implementation. Finally, in order to set-up an appropriate prior distribution on the moment space we have noted that, conditionally on N, the likelihood function has a multinomial structure: this has allowed us to consider a standard Jeffreys'prior opportunely expressed in terms of moments with the appropriate Jacobian. In the simulation study, the BPM estimator has been tested in the 12 Simulation settings proposed by Wang (2010). The overall performance were found quite satisfactory both in terms of mean squared error and coverage for increasing values of the true underlying population size N. Our BPM inference competes well with the most recent and performing competitors and returns interval estimates with the same computational efforts whereas some of the competing methods (Wang, 2010) requires a substantial increase in computing time and makes it prohibitive for very large values of N. We have also tested our Bayesian approach with real data and the performance was really satisfying, especially in real datasets where the true underlying N is indeed known. Possible improvements on our methodology could be achieved by implementing alternative MCMC algorithms possibly relying on most recent hamiltonian MCMC variants as those implemented in STAN.

Appendix A

Algorithms

A.1 QD algorithm

Introduced in Rutishauser (1957), the QD algorithm was thought to compute the eigenvalues of a symmetric tridiagonal matrix. The algorithm, useful in numerical analysis and approximation theory, has several applications such as turning a power series into a continued fraction

$$c_0 + c_1 z + c_2 z^2 + \ldots = \frac{c_0|}{|1} - \frac{q_1^{(0)} z|}{|1} - \frac{e_1^{(0)} z|}{|1} - \frac{q_2^{(0)} z|}{|1} - \frac{e_2^{(0)} z|}{|1} - \ldots .$$

The Above algorithm relates the quantities c_i in the power series with the coefficients $q_i^{(0)}, e_i^{(0)}$ of the continued fraction. Given $c_0, c_1, \ldots$ we have that, for $n \geq 0$

$$q_1^{(n)} = \frac{c_{n+1}}{c_n}.$$

Then, for $m \geq 1$ and $n \geq 0$, we denote $q_{m+1}^{(n)}, e_m^{(n)}$ by:

$$e_m^{(n)} = q_m^{(n+1)} - q_m^{(n)} + e_{m-1}^{(n+1)}$$
$$q_{m+1}^{(n)} = (e_m^{(n)})^{-1} q_m^{(n+1)} e_m^{(n+1)}.$$

For our analysis, we will concentrate on the case

$$c_k = s_k = \int_0^\infty x^k \, d\mu(x) \quad . \tag{A.1}$$

We know that, by the Stieltjes Transform:

$$\int_0^\infty \frac{d\mu(x)}{1 - zx} = c_0 + c_1 z + c_2 z^2 + \ldots \tag{A.2}$$

and

$$z_{2m-1} = q_m^{(0)}, \quad z_{2m} = e_m^{(0)}, \quad m \geq 1.$$

We finally recall from chapter 2 that:

$$z_k = \frac{s_k - s_k^-}{s_{k-1} - s_{k-1}^-}.$$

A.1.1 QD algorithm to derive the recurrence coefficients

$\boldsymbol{z}$ can be also utilized to calculate the recurrence coefficients $\boldsymbol{a}$ and $\boldsymbol{b}$

$$b_k = z_{2k-1} z_{2k} \quad \textit{and} \quad a_k = z_{2k} + z_{2k+1} \quad k = 1, 2, \ldots.$$

These coefficients can be explicitly obtained by solving the QD algorithm with a specific initialization.

Theorem A.1.1. *If the QD algorithm starts with:*

$$q_1^{(n)} = \frac{s_{n+1}}{s_n} \tag{A.3}$$

and $e_m^{(n)}, q_{m+1}^{(n)}, m \geq 1, n \geq 0$ *are defined recursively by* (A.1) *then*

$$e_m^{(0)} = z_{2m}, q_m^{(0)} = z_{2m-1} \tag{A.4}$$

for $m \geq 1$.

Thus, if we start from the truncated moment sequence $\boldsymbol{s_M}$, we can then derive the pseudo-canonical moments $(z_1, \ldots, z_M)$ and sub-sequentially the recurrence coefficients $(a_1, b_1 \ldots, a_{M/2}, b_{M/2-1})$.

A.2 Chebyshev algorithm

Let $\mathcal{L}_\mu$ be a functional associated to the measure μ. We have seen that:

$$\mathcal{L}_\mu \left[x^m P_n(x)\right] = \mathbb{K}_n \delta m, n \textit{ where } \mathbb{K}_n \neq 0, m = 0, 1, \ldots, n \tag{A.5}$$

and $\delta_{m,n}$ is the Kronecker delta. Under this condition, the polynomials satisfy the three-term recurrence relation. Let us now consider a second sequence τ_n of monic polynomials with τ having exact degree n. These polynomials satisfy another recurrence relation of the form:

$$\tau_{n+1}(z) = z\tau_n(z) - \sum_{k=0}^{n} \tau_{k,n} \tau_n(z), \quad n \in \mathbb{N}. \tag{A.6}$$

For the sake of our analysis, we are interested in the case where they satisfy as well a three-term recurrence relation:

$$\tau_{n+1}(z) = (z - a'_n)\tau_n(z) + b'_n \tau_{n-1}(z), \quad n \in \mathbb{N}.$$

We have now all the ingredients to define the so-called modified moments

$$s_n^* = \int_0^\infty \tau_n(z) d\mu(z)$$

and, sub-sequentially, the quantities

$$\sigma_{m,n} = \int_0^\infty \tau_m(z) P_n(z) \mu(z).$$

It can be noted that $\sigma_{m,0} = m_m^*$ and $\sigma_{m,n}$=0 for $m < n$. But why should we use the modified moments ? As mentioned before, the map from the ordinary moments to the recurrence coefficients of the orthogonal polynomials is not well-conditioned and one might hope to use the modified Chebyshev algorithm to improve the conditioning of the algorithm and therefore improve the estimation of the nodes and weights (Gautschi, 1985). However, in species sampling problems, it has been demonstrated that considering the modified moments does not bring any advantages in estimating the recurrence coefficients (Daley and Smith, 2016). Therefore, we will come back to the ordinary moments and fix $\tau_m(z) = z^m$. Then, the quantity $\sigma_{m,n}$ becomes:

$$\sigma_{m,n} = \int_0^\infty z^m P_n(z)\mu(z).$$

Let us now define the infinite row vectors:

$$P = [P_0, P_1, P_2, \ldots], \quad \tau = [\tau_0, \tau_1, \tau_2, \ldots].$$

The matrices

$$H = \begin{bmatrix} a_0 & b_1 & \cdots & \cdots & \\ 1 & a_1 & b_2 & \cdots & \\ & & 1 & a_2 & \ddots \\ & & & \ddots & \ddots \end{bmatrix}, \quad T = \begin{bmatrix} \tau_{0,0} & \tau_{0,1} & \tau_{0,2} & \cdots & \\ 1 & \tau_{1,1} & \tau_{1,2} & \cdots & \\ & & 1 & \tau_{2,2} & \ddots \\ & & & \ddots & \ddots \end{bmatrix}$$

$$S = \begin{bmatrix} \sigma_{0,0} & & & & \\ \sigma_{1,0} & \sigma_{1,1} & & & \\ \sigma_{2,0} & \sigma_{2,1} & \sigma 2,2 & & \\ \vdots & \vdots & & \vdots & \ddots \end{bmatrix}, \quad D = \begin{bmatrix} \sigma_{0,0} & & & & \\ & \sigma_{1,1} & & & \\ & & \sigma 2,2 & & \\ & & & & \ddots \end{bmatrix}$$

are also infinite. From this relation, we get

$$zP(z) = P(z)H \quad z\tau(z) = \tau(z)T$$

from which we obtain:

$$SH = T^T S. \tag{A.7}$$

One easily sees that $H = S^{-1}T^T S$ is determined by S and T. From the same relation, we realize that it suffices to know the diagonal and the first codiagonal of S. We then obtain:

$$\sigma_{m,n+1} + a_n\sigma_{m,n} + b_n\sigma_{m,n-1} = \sigma_{m+1,n} + a'_m\sigma_{m,n} + b'_m\sigma_{m-1,n} \tag{A.8}$$

with $\sigma_{m,-1} = \sigma_{-l,n} = 0$ For $m < n-1$ both sides of the above expression are zero since $\sigma_{m,n}$=0 for $m < n$. For $m = n-1$ we have:

$$b_n\sigma_{n-1,n-1} = \sigma_{n,n} \tag{A.9}$$

while for $m = n$

$$a_n\sigma_{n,n} + b_n\sigma_{n,n-1} = \sigma_{n+1,n} + \tau_{n,n}\sigma_{n,n}. \tag{A.10}$$

Since we are considering the unmodified moments (where $\tau_{n.n} = 1$), the above expression becomes:

$$a_n\sigma_{n,n} + b_n\sigma_{n,n-1} = \sigma_{n+1,n} + \sigma_{n,n}. \tag{A.11}$$

Putting together the last two expressions, one gets:

$$\begin{aligned} b_n =& \frac{\sigma_{n,n}}{\sigma_{n-1,n-1}} \\ a_n =& \frac{\sigma_{n+1,n}}{\sigma_{n,n}} - \frac{\sigma_{n,n-1}}{\sigma_{n-1,n-1}}. \end{aligned} \tag{A.12}$$

The above result is not surprising. We have in fact seen that:

$$\begin{aligned} \sigma_{n,n} &= \int_0^\infty x^n P_n(x) d\nu(x) = b_1 \dots b_n \\ \sigma_{n+1,n} &= \int_0^\infty x^{n+1} P_n(x) d\nu(x) = b_1 \dots b_n (a_1 + \dots + a_n). \end{aligned} \tag{A.13}$$

It is then clear how we can get the recurrence coefficients $\boldsymbol{a}$ and $\boldsymbol{b}$ from $\boldsymbol{\sigma}$.

But how can we initialize the algorithm ? There is no a unique answer since it really depends on the context. Daley and Smith (2016) started from the empirical moments derived from the frequencies of frequencies. In our statistical model setups, we have tried out several approaches (see chapter 3 and chapter 5 for additional details).

In general, one could start with:

$$\sigma_{m,0} = \widehat{s}_m, \quad m = 0, 1, \dots, M. \tag{A.14}$$

Then, S is built up automatically by applying (A.8)

$$\sigma_{m,n+1} = \sigma_{m+1,n} - a_n\sigma_{m,n} - b_n\sigma_{m,n-1} \tag{A.15}$$

since $\tau_{m,m} = 1$ and $\tau_{m,k} = 0 \ \ (\forall k \neq n)$. The recurrence coefficients $\boldsymbol{a}$ and $\boldsymbol{b}$ can then be calculated recursively through (A.12).

Appendix B

List of symbols

for ease for of reading, we have collected the majority of symbols used in the thesis.

- N: finite population size in the set of natural numbers $\mathbb{N}$.
- X_i: the i-th (possibly unobserved/censored) count i.e. the random number of times unit i of the population has been repeatedly counted in the detection/sampling/enumeration process. The index i ranges in $\{1, 2, \ldots, N\}$.
- $\boldsymbol{X}_{obs} = \{X_i : X_i > 0\}$: zero truncated random observable counts.
- $\boldsymbol{x}_{obs} = \{x_i : X_i > 0\}$: zero truncated observed counts.
- n: random observable number of distinct units of the population counted at least once in the sample i.e. number of indexes i for which $X_i > 0$.
- $N - n$: unknown random number of units in the population unobserved in the sample i.e. number of indexes i for which $X_i = 0$.
- f_j: count j frequency. When $j = 0$, $f_0 = N - n$. When $j > 0$, f_j is the observed number of units i of the population for which the sampling count is $X_i = j$.
- M: finite maximum observed count i.e. $M = \max\{X_1, \ldots, X_i, \ldots, X_N\} = \max\{j \in \mathbb{N} : f_j > 0\}$ so that $n = \sum_{j=1}^{\infty} f_j = N - f_0$.
- $\boldsymbol{f}_+ = (f_1, \ldots, f_j, \ldots, f_M, \ldots)$: sequence of observable frequencies of frequencies in a finite population. By definition, $\boldsymbol{f}_+$ is such that there exists a finite $M \in \mathbb{N}$ for which $\sum_{j=1}^{M} f_j = \sum_{j=1}^{\infty} f_j = n$.
- $\boldsymbol{f} = (f_0, \boldsymbol{f}_+)$: sequence of all frequencies of frequencies including the zero frequency. One has that $\sum_{j=0}^{\infty} f_j = N$. Hence, the knowledge of $\boldsymbol{f}$ is equivalent to the knowledge of $\boldsymbol{f}_+$ and N.
- $\mathbb{F}_+$: sample space of the observable frequencies of frequencies.
- $\nu(\lambda)$: the mixing probability measure or distribution on $[0, \infty)$ for the Poisson rate parameter λ in the model specification of the Poisson mixture distributions of the observable counts.

- $\boldsymbol{\theta} = (N, \nu)$ the original parametrization of the Poisson mixture distribution model.
- P_j: probability of observing a count equal to j; in the original parametrization used for the nonparametric hierarchical model specification of the Poisson mixture distributions this probability depends on the mixing distribution $\nu(\cdot)$ so that $P_j = P(X_i = j; \nu) = P_j(\nu)$. However, by using alternative parametrizations, it will be shown that P_j can be reformulated equivalently as depending on a finite number of moments of $\phi(\cdot)$, an equivalent finite measure transformation of $\nu(\cdot)$. To make notation lighter, when clear from the context, the dependence on $\nu(\cdot)$ or equivalent reparametrizations will be dropped.
- $P_0 = P(X_i = 0; \nu) = P_0(\nu)$ probability of observing a count equal to 0.
- $\overline{P}_j$: conditional probability that a generic unit of the population is counted j times ($X_i = j$) provided that it is observed ($X_i > 0$); the following holds: $\overline{P}_j = \frac{P_j}{1-P_0}$.
- $\overline{\boldsymbol{P}}_+ = (\overline{P}_1, \ldots, \overline{P}_j, \ldots, \overline{P}_M, \ldots)$: sequence of conditional Poisson mixtures probabilities. One has that $\sum_{j=1}^{\infty} \overline{P}_j = \sum_{j=1}^{M} \overline{P}_j = 1$.
- $\boldsymbol{P} = (P_1, \ldots, P_j, \ldots, P_M, \ldots)$: sequence of Poisson mixtures probabilities including P_0. One has that $\sum_{j=1}^{\infty} P_j = \sum_{j=0}^{M} P_j = 1$.
- $L(\boldsymbol{\theta}; \boldsymbol{x}_{obs})$: likelihood function for the Poisson mixture model in the original parametrization.
- $L(\boldsymbol{\theta}; \boldsymbol{f}_+)$: likelihood function for the Poisson mixture model corresponding to the observed frequencies of frequencies. The sequence $\boldsymbol{f}_+$ represents the sufficient statistics for $\boldsymbol{\theta}$.
- $L(\boldsymbol{\theta}; \boldsymbol{f}_+) = L(N, \boldsymbol{P}; \boldsymbol{f}_+) = \binom{N}{f_0 f_1 \ldots} \prod_{j=0}^{\infty} P_j^{f_j}$: likelihood function reformulated as function of the Poisson mixtures count probabilities. it will be useful to factorize the likelihood function as $L(\boldsymbol{\theta}; \boldsymbol{f}_+) = L_C(\boldsymbol{\theta}; \boldsymbol{f}_+) * L_R(\boldsymbol{\theta}; \boldsymbol{f}_+)$.
- $L_C(\boldsymbol{\theta}; \boldsymbol{f}_+) = L_C(P_+; \boldsymbol{f}_+) = \binom{n}{f_1, f_2, \ldots} \prod_{i=1}^{\infty} \left(\frac{P_j}{1-P_0}\right)^{f_j}$: conditional likelihood.
- $L_R(\boldsymbol{\theta}; \boldsymbol{f}_+) = L_R(N, P_0; \boldsymbol{f}_+) = \binom{N}{f_0} P_0^{f_0} (1 - P_0)^{N - f_0}$: residual likelihood.
- $\hat{\nu}_C$: MLE for ν with respect to the conditional likelihood.
- $\hat{\nu}$: MLE for ν with respect to the unconditional likelihood.
- $\mathcal{P}(E)$: set of all probability measures on the measurable space $(E, \sigma(E))$ with $\sigma(E)$ a suitable σ-field of subsets of E. Most relevant special cases dealt with are $E = [0, \infty)$, $E = [0, 1]$ and $E = \mathbb{R}$ with their Borel σ-field. The set $[0, \infty)$ is the parameter space for the Poisson rate λ.
- $\mathcal{F}(E)$: set of all finite measures on the measurable space $(E, \sigma(E))$ with $\sigma(E)$ as a suitable σ-field of subsets of E.

- $\mathcal{F}_{(0,1]}([0,\infty))$: subset of finite measures $\phi \in \mathcal{F}([0,\infty))$ which are determined by their moment sequence and with total mass restriction constraining $\phi([0,\infty)) \in (0,1]$.

- ϕ: finite measure on $[0,\infty)$ with $d\phi(\lambda) = e^{-\lambda} d\nu(\lambda)$ and total mass $\phi([0,\infty)) = \int_{[0,\infty)} d\phi(\lambda) = P_0$. The set of probability distributions ν on $[0,\infty)$ denoted $\mathcal{P}([0,\infty))$ is one-to-one with a proper subset of finite measures $\mathcal{F}([0,\infty))$ namely the restricted space of finite measure ϕ on $[0,\infty)$ with total mass in (0,1] denoted as $\mathcal{F}_{(0,1]}([0,\infty))$.

- $\widetilde{\phi}$: probability measure obtained by normalizing $\phi(\cdot)$ as follows: $\widetilde{\phi}(\cdot) = \frac{\phi(\cdot)}{\phi([0,\infty))}$.

- μ is a σ-finite measure on $[0,\infty)$ with $d\mu(\lambda) = \lambda e^{-\lambda} d\nu(\lambda)$. Exclusively in chapter 2 and in the Appendix, it will be used as reference probability measure.

- $d\widetilde{\mu}(\lambda)$ probability measure obtained by normalizing $\mu(\cdot)$ provided μ is finite i.e. only whenever $\int_{[0,\infty)} \lambda e^{-\lambda} d\nu(\lambda) < \infty$.

- γ: a probability measure defined as $d\gamma(\lambda) = \frac{(1-e^{-\lambda})d\nu(\lambda)}{\int (1-e^{-\lambda})d\nu(\lambda)}$.

- ω : finite measure defined as $d\omega(\lambda) = (e^{\lambda}-1)^{-1} d\gamma(\lambda)$.

- μ_S: mirror probability measure of μ n $(-\infty,\infty)$ such that if $X \sim \mu$, then μ_S is the probability measure corresponding to the distribution of the symmetric random variable $S = \sqrt{X}(2Z-1)$ where Z is an independent Bernoulli random variable with success probability 1/2. Note that μ_S is symmetric and X is such that $X \stackrel{d}{=} S^2$.

- s_j: ordinary moment of non negative integer order j of the finite measure ϕ i.e. $s_j = \int_{[0,\infty)} \lambda^j d\phi(\lambda)$. Note that $\sum_{j=1}^{\infty} s_j/j! = 1$.

- $\boldsymbol{s} = (s_0, s_1, \ldots, s_j, \ldots)$ vector of all the ordinary moments of the finite measure ϕ.

- $\boldsymbol{s}_M = (s_0, s_1, \ldots, s_M)$ finite-dimensional (projection) vector of the first $M+1$ coordinate of $\boldsymbol{s}$.

- $\widetilde{\boldsymbol{s}} = (\widetilde{s}_0, \widetilde{s}_1, \ldots, \widetilde{s}_j, \ldots)$: moment sequence of the normalized probability measure $\widetilde{\phi}$ with $\widetilde{s}_0 = 1$.

- $\widetilde{\boldsymbol{s}}_M$ vector of moments of the normalized probability measure $\widetilde{\phi}$ with $\widetilde{s}_0 = 1$.

- m_j: ordinary moment of order j (with $j \in \mathbb{Z}$) of the finite measure μ i.e. $m_j = \int_{[0,\infty)} \lambda^{j+1} d\mu(\lambda)$. Note that $\sum_{j=0}^{\infty} m_j/j! = 1$.

- $\boldsymbol{m} = (m_0, m_1, \ldots, m_j, \ldots)$ vector of all the ordinary moments of the finite measure μ.

- $\boldsymbol{m}_M = (m_0, m_1, \ldots, m_M)$ finite-dimensional (projection) vector of the first $M+1$ coordinate of $\boldsymbol{m}$.

- $\widetilde{\boldsymbol{m}} = (\widetilde{m}_0, \widetilde{m}_1, \ldots, \widetilde{m}_j, \ldots)$: moment sequence of the normalized probability measure $\widetilde{\mu}$ with $\widetilde{m}_0 = 1$.
- $\widetilde{\boldsymbol{m}}_M$ vector of moments of the normalized probability measure $\widetilde{\mu}$ with $\widetilde{m}_0 = 1$.
- o_j: ordinary moment of non negative integer order j of the finite measure ω i.e. $m_j = \int_{[0,\infty)} \lambda^j \, d\omega(\lambda)$. Note that $\sum_{j=0}^{\infty} o_j/j! = 1$.
- $\boldsymbol{o} = (o_0, o_1, \ldots, o_j, \ldots)$ vector of all the ordinary moments of the finite measure ω.
- $\boldsymbol{o}_M = (o_0, o_1, \ldots, o_M)$ finite-dimensional (projection) vector of the first $M+1$ coordinate of $\boldsymbol{o}$.
- $\widetilde{\boldsymbol{o}} = (\widetilde{o}_0, \widetilde{o}_1, \ldots, \widetilde{o}_j, \ldots)$: moment sequence of the normalized probability measure $\widetilde{\omega}$ with $\widetilde{o}_0 = 1$.
- $\widetilde{\boldsymbol{o}}_M$ vector of moments of the normalized probability measure $\widetilde{\omega}$ with $\widetilde{o}_0 = 1$.
- $\mathcal{S}_t(E)$: moment space for a finite measure ϕ supported on $\mathbb{E}$ which is determined by its moment sequence with total mass equal to t with $0 < t < \infty$ and such that moments of all orders exist. When $t = 1$ we have the standard moment space for a probability measure on $E \subseteq \mathbb{R}$ determined by its moments. More explicitly, we can write

$$\mathcal{S}_t(E) = \left\{ (s_0, s_1, s_2, \ldots) : s_j = \int_E \lambda^j \, d\phi(\lambda) \ s.t. \ s_0 = t, j = 0, 1, 2, \ldots \right\}.$$

- $\mathcal{S}_I$: moment space for a finite measure ϕ supported on $\mathbb{E}$, determined by its moment sequence and with total mass in I where I is a bounded subset of $\mathbb{R}^+$. When $t = 1$ we have the standard moment space for a probability measure on $E \subseteq \mathbb{R}$. More explicitly, we can write

$$\mathcal{S}_I(E) = \left\{ (s_0, s_1, s_2, \ldots) : s_j = \int_E \lambda^j \, d\phi(\lambda) \ s.t. \ s_0 \in I, j = 0, 1, 2, \ldots \right\}.$$

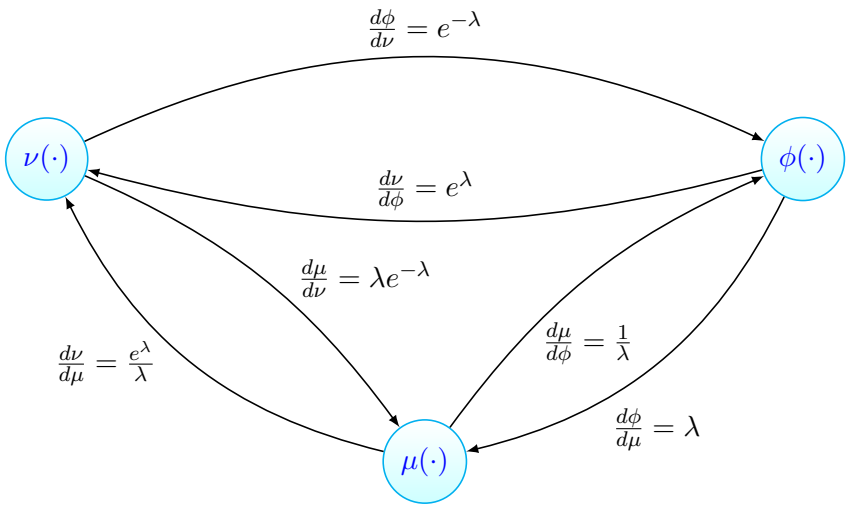
- $\mathcal{S}^{[M]}(E)$: truncated moment space for a finite measure ϕ supported on $\mathbb{E}$ which is determined by its moment sequence. When $s_0 = 1$ we have the standard truncated moment space for a probability measure on $E \subseteq \mathbb{R}$. More explicitly, we can write

$$\mathcal{S}_1^{[M]}(E) = \left\{ (s_0, s_1, s_2, \ldots, s_M) : s_j = \int_E \lambda^j \, d\phi(\lambda), j = 0, 1, 2, \ldots, M \right\}.$$

- $\mathcal{S}_I^{[M]}(E)$: truncated moment space for a finite measure ϕ supported on $\mathbb{E}$ which is determined by its moment sequence and with total mass in a bounded interval I of positive reals

$$\mathcal{S}_I^{[M]}(E) = \left\{ (s_0, s_1, s_2, \ldots, s_M) : s_0 \in I, s_j = \int_E \lambda^j \, d\phi(\lambda), j = 1, 2, \ldots, M \right\}.$$

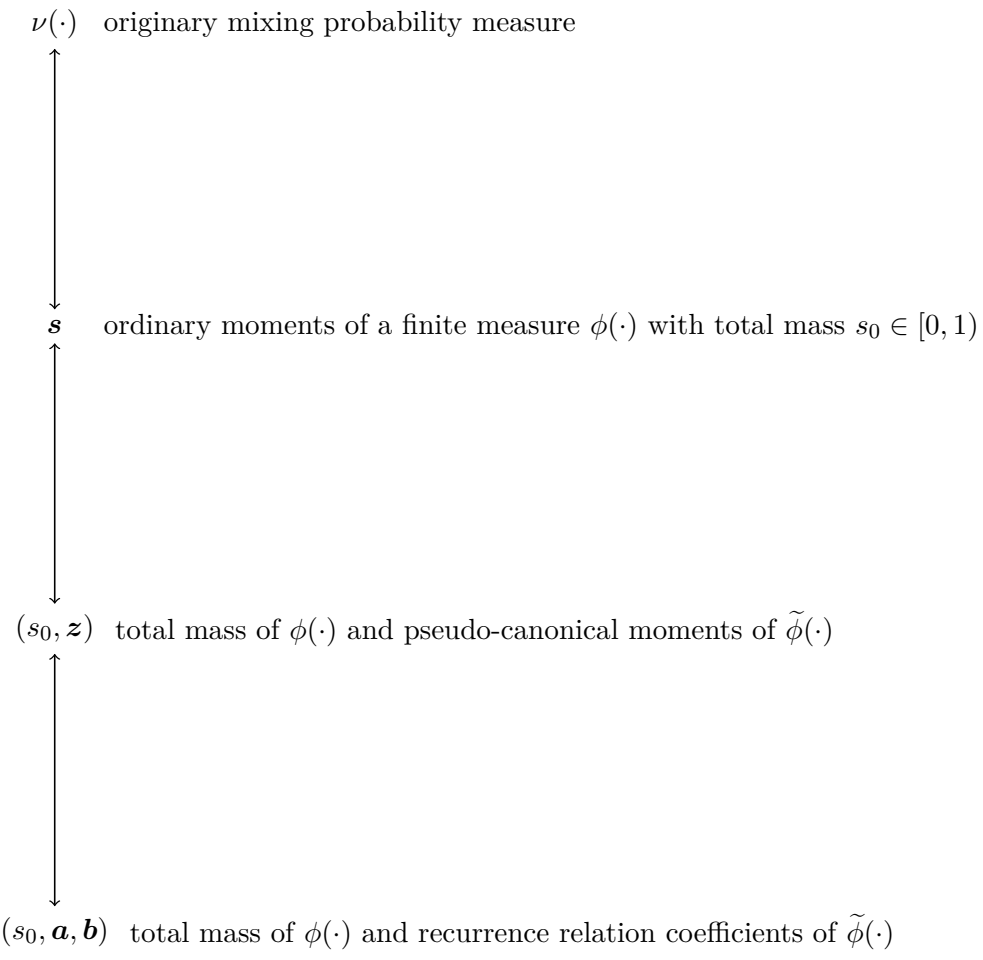
- $s^{+}_{k+1} = s^{+}_{k+1}(s_0, s_1, \ldots, s_k)$ superior extremal moment of order k+1 corresponding to fixed values $(s_0, s_1, \ldots, s_k)$ of the moments of lower non negative order order.
- $s^{-}_{k+1} = s^{-}_{k+1}(s_0, s_1, \ldots, s_k)$ inferior extremal moment of order k+1 corresponding to fixed values $(s_0, s_1, \ldots, s_k)$ of the moments of lower non negative order.



$$\boldsymbol{s}_M = (s_0, s_1, \ldots, s_M) \longleftrightarrow (s_0, c_1, \ldots, c_M)$$

$$\updownarrow$$

$$\boldsymbol{z}_M = (z_0, z_1, \ldots, z_M)$$


$\nu(\cdot)$ originary mixing probability measure
$\boldsymbol{s}$ ordinary moments of a finite measure $\phi(\cdot)$ with total mass $s_0 \in [0, 1)$
$(s_0, \boldsymbol{z})$ total mass of $\phi(\cdot)$ and pseudo-canonical moments of $\widetilde{\phi}(\cdot)$
$(s_0, \boldsymbol{a}, \boldsymbol{b})$ total mass of $\phi(\cdot)$ and recurrence relation coefficients of $\widetilde{\phi}(\cdot)$

- c_j j-th canonical moment of a probability measure in $\mathcal{P}([0,1])$.
- z_j j-th pseudo-canonical moment of a probability measure in $\mathcal{P}([0,\infty))$.
- $\{P_d(x)\}_{n=0}^{\infty}$: orthogonal polynomials on E associated to a finite measure ϕ on the measurable space $(E,\sigma(E))$; when clear from the context the dependence on ϕ is dropped from the notation. An orthogonal polynomial sequence is a countable family of polynomials defined on the space E such that any two different polynomials in the sequence are orthogonal to each other with respect to the inner product in a functional space associated to a finite measure ϕ on $(E,\mathcal{E})$: $\langle f,g\rangle = \int_E f(x)g(x)d\phi(x)$.
- a_j & b_j j-th recurrence coefficients.
- J_d: Jacobi matrix corresponding to a tridiagonal symmetric matrix determined by the first recurrence coefficients a_j & b_j.
- $e_1,\dots e_n$: eigenvalues of J_n (quadrature nodes).
- $v_1,\dots,v_n$: square of the first component of the eigenvectors of J_n (quadrature weights).
- H_M: Hankel matrix of order M associated to a point in the M-truncated moment space.
- ρ_d : $[0,1]^d \to \mathcal{S}_d([0,1))$ mapping between canonical moments and ordinary moments i.e. $(c_1,\dots,c_d) \overset{\rho_d}{\to} (s_1,\dots,s_d)$.
- $\xi_d : (z_1,\dots,z_d) \to (s_d,\dots,s_d)$ mapping between pseudo-canonical moments and ordinary moments.
- $\psi_d : (a_1,b_1\dots,a_d,b_d) \to (s_1,\dots,s_{2d})$ mapping between recurrence coefficients and ordinary moments.

Bibliography

David S. Alberts, María Elena Martínez, Denise J. Roe, José M. Guillén-Rodríguez, James R. Marshall, J. Barbara van Leeuwen, Mary E. Reid, Cheryl Ritenbaugh, Perla A. Vargas, A.B. Bhattacharyya, David L. Earnest, Dianne Parish, Kris Koonce, Lianne Fales, and Richard E. Sampliner. Lack of effect of a high-fiber cereal supplement on the recurrence of colorectal adenomas. *New England Journal of Medicine*, 342(16):1156–1162, 2000. doi: 10.1056/NEJM200004203421602. PMID: 10770980.

Danilo Alunni Fegatelli and Luca Tardella. Moment-based Bayesian Poisson mixtures for inferring unobserved units. *arXiv:1806.06489*, 2018.

Kathryn Barger and John Bunge. Objective Bayesian estimation for the number of species. *Bayesian Anal.*, 5(4):765–785, 12 2010. doi: 10.1214/10-BA527.

Asip P. Basu. Icing the tails to limit theorems, lecture notes in economics and mathematical systems, 192. *Encyclopedia of statistical sciences*, 4, 1963.

Dankmar Böhning, Panicha Kaskasamkul, and Peter G.M. van der Heijden. A modification of Chao's lower bound estimator in the case of one-inflation. *Metrika*, pages 1–24, October 2018.

John Bunge and Michael Fitzpatrick. Estimating the number of species: A review. *Journal of the American Statistical Association*, 88(421):364–373, 1993. doi: 10.1080/01621459.1993.10594330.

Dankmar Böhning, Ekkehart Dietz, Ronny Kuhnert, and Dieter Schön. Mixture models for capture-recapture count data. *Statistical Methods and Applications*, 14: 29–43, 02 2005. doi: 10.1007/BF02511573.

Dankmar Böhning, M. Fazil Baksh, Rattana Lerdsuwansri, and James Gallagher. Use of the ratio plot in capture–recapture estimation. *Journal of Computational and Graphical Statistics*, 22(1):135–155, 2013. doi: 10.1080/10618600.2011.647174.

Dankmar Böhning, Irene Rocchetti, Marco Alfó, and Heinz Holling. A flexible ratio regression approach for zero-truncated capture–recapture counts. *Biometrics*, 72 (3):697–706, 2016. doi: 10.1111/biom.12485.

Torsten Carleman. Sur les équations intégrales singulières a noyau réel et symétrique. 1923.

Anne Chao. Non-parametric estimation of the classes in a population. *Scandinavian Journal of Statistics*, 11:265–270, 01 1984. doi: 10.2307/4615964.

Anne Chao, P. K. Tsay, Sheng-Hsiang Lin, Wen-Yi Shau, and Day-Yu Chao. The applications of capture-recapture models to epidemiological data. *Statistics in Medicine*, 20(20):3123–3157, 2001. doi: 10.1002/sim.996.

Theodore S. Chihara. An introduction to orthogonal polynomials. *Mathematics and its Applications*, 1978.

Chun-Huo Chiu, Yi-Ting Wang, Bruno A. Walther, and Anne Chao. An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics*, 70(3):671–682, 2014. doi: 10.1111/biom.12200.

Timothy Daley and Andrew D. Smith. Better lower bounds for missing species: improved non-parametric moment-based estimation for large experiments, 2016.

H. Dette and W.J. Studden. . *Canonical Moments with Applications in Statistics*. Wiley and Sons, 1997.

Holger Dette and Jan Nagel. Distributions on unbounded moment spaces and random moment sequences. *Ann. Probab.*, 40(6):2690–2704, 11 2012. doi: 10.1214/11-AOP693.

Holger Dette and William Studden. A note on the matrix valued q-d algorithm and matrix orthogonal polynomials on $[0, 1]$ and $[0, \infty)$. *Journal of Computational and Applied Mathematics*, 148:349–361, 11 2002. doi: 10.1016/S0377-0427(02)00555-1.

Bradley Efron and Ronald Thisted. Estimating the number of unsen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976. ISSN 00063444.

Alessio Farcomeni and Luca Tardella. Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. *Electron. J. Statist.*, 6:2602–2626, 2012. doi: 10.1214/12-EJS758.

Walter Gautschi. Orthogonal polynomials—constructive theory and applications. *Journal of Computational and Applied Mathematics*, 12-13:61 – 76, 1985. ISSN 0377-0427. doi: 10.1016/0377-0427(85)90007-X.

Gene Golub and Gérard Meurant. Matrices, moments and quadrature. *Numerical Analysis 1993*, 303, 10 1994.

Irving J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40, 12 1953. doi: 10.2307/2333344.

Michele Guindani, Nuno Sepulveda, Carlos Paulino, and Peter Müller. A Bayesian semi-parametric approach for the differential analysis of sequence counts data. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 63:385–404, 04 2014. doi: 10.1111/rssc.12041.

Bernard Harris. Determining bounds on integrals with applications to cataloging problems. *Ann. Math. Statist.*, 30(2):521–548, 06 1959. doi: 10.1214/aoms/1177706266.

Gordon Hay and Filip Smit. Estimating the number of drug injectors from needle exchange data. *Addiction Research & Theory*, 11(4):235–243, 2003. doi: 10.1080/1606635031000135622.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985. doi: 10.1017/CBO9780511810817.

Norman L. Johnson, Adrienne W. Kemp, and Samuel Kotz. Univariate discrete distributions, third editions. 2005. doi: 10.1002/0471715816.

William Ogilvy Kermack, Anderson G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927. doi: 10.1098/rspa.1927.0118.

Ronny Kuhnert, Victor J. Del Rio Vilas, James Gallagher, and Dankmar Böhning. A bagging-based correction for the mixture model estimator of population size. *Biometrical Journal*, 50(6):993–1005, 2008. doi: 10.1002/bimj.200810485.

Frederick C. Lincoln. Calculating waterfowl abundance on the basis of banding returns. *United States Department of Agriculture Circular.*, 118:1–4, 1930.

William A. Link. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4):1123–1130, 2003. ISSN 0006-341X. doi: 10.1111/j.0006-341X.2003.00129.x.

Chang Xuan Mao. Inference on the number of species through geometric lower bounds. *Journal of the American Statistical Association*, 101(476):1663–1670, 2006. doi: 10.1198/016214506000000528.

Chang Xuan Mao and Bruce G Lindsay. Tests and diagnostics for heterogeneity in the species problem. *Computational Statistics Data Analysis*, 41(3):389 – 398, 2003. ISSN 0167-9473. doi: 10.1016/S0167-9473(02)00164-0. Recent Developments in Mixture Model.

Chang Xuan Mao and Bruce G. Lindsay. Estimating the number of classes. *Ann. Statist.*, 35(2):917–930, 04 2007. doi: 10.1214/009053606000001280.

Chang Xuan Mao, Nan Yang, and Jinhua Zhong. On population size estimators in the poisson mixture model. *Biometrics*, 69, 07 2013. doi: 10.1111/biom.12044.

James L. Norris and Kenneth Pollock. Non-parametric mle for poisson species abundance models allowing for heterogeneity between species. *Environmental and Ecological Statistics*, 5:391–402, 2004.

Johannes C. G. Petersen. The yearly immigration of young plaice into the limfjord from the german sea, ect. *Report of the Danish Biological Station for 1985*, 6: 1–48, 1986. URL `https://ci.nii.ac.jp/naid/10013009910/en/`.

Yixuan Qiu and Jiali Mei. *RSpectra: Solvers for Large-Scale Eigenvalue and SVD Problems*, 2019. R package version 0.16-0.

John R. Rice. Tchebycheff systems: With applications in analysis and statistics (samuel karlin and william j. studden). *SIAM Review*, 9(2):257–258, 1967. doi: 10.1137/1009050.

John M. Roberts Jr and Devon D. Brewer. Estimating the prevalence of male clients of prostitute women in vancouver with a simple capture–recapture method. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4): 745–756, 2006. doi: 10.1111/j.1467-985X.2006.00416.x.

Irene Rocchetti, John Bunge, and Dankmar Böhning. Population size estimation based upon ratios of recapture probabilities. *Ann. Appl. Stat.*, 5(2B):1512–1533, 06 2011. doi: 10.1214/10-AOAS436.

Heinz Rutishauser. Der quotienten - differenzen - algorithmus. *Mitteilungen aus dem Institut für angewandte Mathematik*, 1957. doi: 10.1002/zamm.19570371122.

Lalitha Sanathanan. Estimating the size of a multinomial population. *Ann. Math. Statist.*, 43(1):142–152, 02 1972. doi: 10.1214/aoms/1177692709.

Konrad Schmudgen. The moment problem. *Graduated Text in Mathematics*, 2017.

Leopold Simar. Maximum likelihood estimation of a compound poisson process. *Ann. Statist.*, 4(6):1200–1209, 11 1976. doi: 10.1214/aos/1176343651.

Morris Skibinsky. Extreme nth moments for distributions on [0, 1] and the inverse of a moment space map. *Journal of Applied Probability*, 5(3):693–701, 1968. doi: 10.2307/3211931.

Nick Stokes. A stable quotient-difference algorithm. *Mathematics of Computation*, 34(150):515–519, 1980. ISSN 00255718, 10886842.

Luca Tardella. A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity. *Biometrika*, 89(4):807–817, 2002. ISSN 00063444.

Henry Teicher. On the mixture of distributions. *Ann. Math. Statist.*, 31(1):55–73, 03 1960. doi: 10.1214/aoms/1177705987.

Dominik Tomecki. Asymptotics for random moment sequences. *Ruhr-Universität Bochum, 2018*, 2018.

Ji-Ping Wang. Estimating species richness by a poisson-compound gamma model. *Biometrika*, 97:727–740, 09 2010. doi: 10.1093/biomet/asq026.

Ji-Ping Wang and Bruce G. Lindsay. An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology*, 5(1):30 – 45, 2008. ISSN 1572-3127. doi: 10.1016/j.stamet.2007.03.004.

Ji-Ping Z. Wang and Bruce G. Lindsay. A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*, 100(471):942–959, 2005. doi: 10.1198/016214504000002005.

John C. Wheeler. Modified moments and Gaussian quadratures. *Rocky Mountain J. Math.*, 4(2):287–296, 06 1974. doi: 10.1216/RMJ-1974-4-2-287.