LUND UNIVERSITY

Order restricted inference over countable preordered sets. Statistical aspects of neutron detection

Pastukhov, Vladimir

2018

Document Version:
Publisher's PDF, also known as Version of record

– CENTRUM SCIENTIARUM MATHEMATICARUM –

Order restricted inference over countable preordered sets. Statistical aspects of neutron detection.

VLADIMIR PASTUKHOV

Lund University
Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Order restricted inference over countable preordered sets. Statistical aspects of neutron detection.

by
Vladimir Pastukhov

LUND
UNIVERSITY

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

A doctoral thesis at a university in Sweden takes either the form of a single, cohesive research study (monograph) or a summary of research papers (compilation thesis), which the doctoral student has written alone or together with one or several other author(s).

In the latter case the thesis consists of two parts. An introductory text puts the research work into context and summarizes the main points of the papers. Then, the research publications themselves are reproduced, together with a description of the individual contributions of the authors. The research papers may either have been already published or are manuscripts at various stages (in press, submitted, or in draft).

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118
SE-211 00 Lund
Sweden

http://maths.lu.se/

Doctoral Theses in Mathematical Sciences 2018:8
ISSN: 1404-0034

ISBN: 978-91-7753-808-0 (print)
ISBN: 978-91-7753-809-7 (pdf)
ISRN: LUNFMS-1025-2018

Printed in Sweden by Media-Tryck, Lund 2018

Contents

Abstract

This thesis consists of four papers.

In the first paper, we study the isotonic regression estimator over a general countable preordered set. We obtain the limiting distribution of the estimator and study its properties. Also, it is shown that the isotonisation preserves the rate of convergence of the underlying estimator. We apply these results to the problems of estimation of a bimonotone regression function and estimation of a bimonotone probability mass function.

In the second paper, we propose a new method of estimating a discrete monotone probability mass function. We introduce a two-step procedure. First, we perform a model selection introducing the Akaike-type information criterion ($CMAIC$). Second, using the selected class of models we construct a modified Grenander estimator by grouping the parameters in the constant regions and then projecting the grouped empirical estimator onto the isotonic cone. It is shown that the post-model-selection estimator performs asymptotically better, in l_2-sense, than the regular Grenander estimator.

In the third paper, we use a stochastic process approach to determine the neutron energy in a novel detector. The data from a multi-layer detector consists of counts of the number of absorbed neutrons along the sequence of the detector's layers, in which the neutron absorption probability is unknown. These results are combined with known results on the relation between the absorption probability and the wavelength to derive an estimator of the wavelength and to show consistency and asymptotic normality.

In the forth paper, the results of the third paper are generalised to the case of a multimode Poisson beam. We study the asymptotic properties of the maximum likelihood estimator of the spectrum and thinning parameters for the spectrum's components.

Keywords: Constrained inference, Isotonic regression, Density estimaion, Grenander estimator, Limit distribution, Neutron detection.

Acknowledgements

First, I would like to express my thanks to my supervisor Dragi Anevski for his support, encouragement and patience throughout the last five years.

Second, I would like to thank all my colleagues for a great working environment and assistance. In particular, I would like to express my thanks to Richard Hall-Wilton and Kalliopi Kanaki for rewarding discussions about modeling in detector physics.

Third, I would like to especially thank Tatyana Turova for the advise in the proof of Lemma 4.4 of Paper A; Magnus Wiktorsson for the suggestions in assessing the accuracy of the asymptotic results of Paper C; Victor Ufnarovski and Andrey Ghulchak for their kind help with Lemma 3 of Paper D.

Last but not least, the biggest acknowledgments are for my family and my friends for their help and support.

Lund, 2018 *Pastukhov Vladimir*

List of papers

This thesis is based on the following papers, referred to by their Latin capitals:

A **The asymptotic distribution of the isotonic regression estimator over a countable preordered set**

Dragi Anevski, Vladimir Pastukhov
Submitted (2018).

B **Estimation of a discrete monotone distribution with model selection.**

Dragi Anevski, Vladimir Pastukhov
Submitted (2018).

C **A stochastic process approach to multilayer neutron detectors**

Dragi Anevski, Richard Hall-Wilton, Kalliopi Kanaki, Vladimir Pastukhov
Submitted (2016).

D **Estimating the distribution and thinning parameters of a homogeneous multimode Poisson process**

Dragi Anevski, Vladimir Pastukhov
Preprint (2018).

Introduction

The thesis consists of two parts.

The first part (Papers A and B) is dedicated to the order-restricted inference over general countable preordered sets.

In the second part (Papers C and D) we study the feasibility of a statistical determination of neutron wavelength and a spectrum for the new generation of neutron detectors being developed at the European Spallation Source (ESS).

1 Order-restricted inference over countable preordered sets

Let $\mathcal{X}$ be a countable set $\{x_1, x_2, \dots\}$ with $|\mathcal{X}| \leq \infty$, with a preorder $\preceq$ defined on it. We begin with the definitions of the order relations on an arbitrary set $\mathcal{X}$ and of an isotonic regression over it, cf. [4, 18, 19].

Definition 1.1. *A binary relation $\preceq$ on $\mathcal{X}$ is a simple order if*

(i) it is reflexive, i.e. $x \preceq x$ for $x \in \mathcal{X}$;

(ii) it is transitive, i.e. $x_1, x_2, x_3 \in \mathcal{X}$, $x_1 \preceq x_2$ and $x_2 \preceq x_3$ imply $x_1 \preceq x_3$;

(iii) it is antisymmetric, i.e. $x_1, x_2 \in \mathcal{X}$, $x_1 \preceq x_2$ and $x_2 \preceq x_1$ imply $x_1 = x_2$;

(iv) every two elements of $\mathcal{X}$ are comparable, i.e. $x_1, x_2 \in \mathcal{X}$ implies that either $x_1 \preceq x_2$ or $x_2 \preceq x_1$.

A binary relation $\preceq$ on $\mathcal{X}$ is a partial order if it is reflexive, transitive and antisymmetric, but there may be noncomparable elements. A preorder is reflexive and transitive but not necessary antisymmetric and the set $\mathcal{X}$ can have noncomparable elements. Note, that in some literature the preorder is called as a quasi-order.

Let us introduce the notation $x_1 \sim x_2$, if x_1 and x_2 are comparable, i.e. if $x_1 \preceq x_2$ or $x_2 \preceq x_1$.

Definition 1.2. *A function $f(x) : \mathcal{X} \to \mathbb{R}$ is isotonic if $x_i, x_j \in \mathcal{X}$ and $x_i \preceq x_j$ imply $f(x_i) \leq f(x_j)$.*

Let $\mathcal{F}^{is} = \mathcal{F}^{is}(\mathcal{X})$ denote the family of real valued bounded functions f on a set $\mathcal{X}$, which are isotonic with respect to the preorder $\preceq$ on $\mathcal{X}$. In the case when $|\mathcal{X}| = \infty$ we consider the functions from the space l_2^w, the Hilbert space of real-valued functions on $\mathcal{X}$, which are square summable with some given non-negative weights $\boldsymbol{w} = \{w_1, w_2, \dots\}$, i.e. any $g \in l_2^w$ satisfies $\sum_{i=1}^{\infty} g(x_i)^2 w_i < \infty$.

Definition 1.3. *A function $g^* : \mathcal{X} \to \mathbb{R}$ is the isotonic regression of a function $g : \mathcal{X} \to \mathbb{R}$ over the preordered set $\mathcal{X}$ with weights $\boldsymbol{w} \in \mathbb{R}^s_+$, with $s \leq \infty$, if*

$$g^* = \operatorname*{argmin}_{f \in \mathcal{F}^{is}} \sum_{x \in \mathcal{X}} (f(x) - g(x))^2 w_x,$$

where $w_{x_i} = w_i$, for $i = 1, \dots, s$.

Similarly one can define an isotonic vector in $\mathbb{R}^s$, with $s \leq \infty$, and the isotonic regression of an arbitrary vector in $\mathbb{R}^s$. Let us consider a set of indices $\mathcal{I} = \{1, \dots, s\}$, with $s \leq \infty$, with some preorder $\preceq$ defined on it.

Definition 1.4. *A vector $\boldsymbol{\theta} \in \mathbb{R}^s$, with $s \leq \infty$, is isotonic if $i_1, i_2 \in \mathcal{I}$ and $i_1 \preceq i_2$ imply $\theta_{i_1} \leq \theta_{i_2}$.*

We denote the set of isotonic vectors in $\mathbb{R}^s$, with $s \leq \infty$, by $\boldsymbol{\mathcal{F}}^{is} = \boldsymbol{\mathcal{F}}^{is}(\mathcal{I})$. In the case of an infinite index set we consider the square summable vectors (with weights $\boldsymbol{w}$) from $\boldsymbol{l}_2^w$, the Hilbert space of all square summable vectors with weights $\boldsymbol{w}$.

Definition 1.5. *A vector $\boldsymbol{\theta}^* \in \mathbb{R}^s$, with $s \leq \infty$, is the isotonic regression of an arbitrary vector $\boldsymbol{\theta} \in \mathbb{R}^s$ (or $\boldsymbol{\theta} \in \boldsymbol{l}_2^w$, if $s = \infty$) over the preordered index set $\mathcal{I}$ with weights $\boldsymbol{w} \in \mathbb{R}^s_+$ if*

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\xi} \in \boldsymbol{\mathcal{F}}^{is}} \sum_{i \in \mathcal{I}} (\xi_i - \theta_i)^2 w_i.$$

1.1 The isotonic regression estimator

Let $\mathring{g} \in \mathcal{F}^{is}$ be a fixed unknown function. Assume we are given observations z_i, $i = 1, \dots, n$, independent or not, that depend on $\mathring{g}$ in some way.

Now assume that

$$\hat{\boldsymbol{g}}_n = \hat{\boldsymbol{g}}_n(z_1, \ldots, z_n)$$

is a $\mathbb{R}^s$ (or $\boldsymbol{l}_2^{\boldsymbol{w}}$)-valued statistic. We will call the sequence $\{\hat{\boldsymbol{g}}_n\}_{n\geq 1}$ the basic estimator of $\mathring{\boldsymbol{g}}$. In order to discuss consistency and asymptotic distribution result we introduce the following basic topologies. When $s < \infty$, we study the Hilbert space with the inner product $\langle \boldsymbol{g}_1, \boldsymbol{g}_2 \rangle = \sum_{i=1}^{s} g_{1,i} g_{2,i} w_i$, for $\boldsymbol{g}_1, \boldsymbol{g}_2 \in \mathbb{R}^s$, endowed with its Borel σ-algebra $\mathcal{B} = \mathcal{B}(\mathbb{R}^s)$ and when $s = \infty$ we study the space $\boldsymbol{l}_2^{\boldsymbol{w}}$ with the inner product $\langle \boldsymbol{g}_1, \boldsymbol{g}_2 \rangle = \sum_{i=1}^{\infty} g_{1,i} g_{2,i} w_i$, for a fixed weight vector $\boldsymbol{w}$ satisfying

$$\begin{cases} \inf\limits_i \{w_i\} > 0 \\ \sup\limits_i \{w_i\} < \infty \end{cases} \tag{1}$$

and we equip $\boldsymbol{l}_2^{\boldsymbol{w}}$ with its Borel σ-algebra $\mathcal{B} = \mathcal{B}(\boldsymbol{l}_2^{\boldsymbol{w}})$.

Now define the isotonized estimator $\hat{\boldsymbol{g}}_n^*$ by

$$\hat{\boldsymbol{g}}_n^* = \underset{\boldsymbol{\xi} \in \boldsymbol{\mathcal{F}}^{is}}{\operatorname{argmin}} \sum_{i \in \mathcal{I}} (\xi_i - \hat{g}_{n,i})^2 w_i. \tag{2}$$

We make the following assumptions on the basic estimator $\hat{\boldsymbol{g}}_n$, for the finite, $s < \infty$, and the infinite, $s = \infty$, support case, respectively.

Assumption 1.1. *Suppose that $s < \infty$. Assume that $\hat{\boldsymbol{g}}_n \xrightarrow{p} \mathring{\boldsymbol{g}}$ for some $\mathring{\boldsymbol{g}} \in \boldsymbol{\mathcal{F}}^{is}$ and $B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}}) \xrightarrow{d} \boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is a random vector in $(\mathbb{R}^s, \mathcal{B})$ and B_n is a diagonal $s \times s$ matrix with elements $[B_n]_{ii} = n^{q_i}$ with q_i being real positive numbers.*

Assumption 1.2. *Suppose that $s = \infty$. Let $\hat{\boldsymbol{g}}_n$, for $n = 1, 2, 3, \ldots$, be a sequence of random vectors taking values in the Hilbert space $\boldsymbol{l}_2^{\boldsymbol{w}}$. Assume that $\hat{\boldsymbol{g}}_n \xrightarrow{p} \mathring{\boldsymbol{g}}$ for some $\mathring{\boldsymbol{g}} \in \boldsymbol{\mathcal{F}}^{is}$, and $B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}}) \xrightarrow{d} \boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is a random vector in $(\boldsymbol{l}_2^{\boldsymbol{w}}, \mathcal{B})$ and B_n is a linear operator $\boldsymbol{l}_2^{\boldsymbol{w}} \to \boldsymbol{l}_2^{\boldsymbol{w}}$, such that for any $\boldsymbol{g} \in \boldsymbol{l}_2^{\boldsymbol{w}}$ it holds that $(B_n \boldsymbol{g})_i = n^{q_i} g_i$, with q_i being the real positive numbers. Suppose also that any finite s-dimensional cylinder set in $\boldsymbol{l}_2^{\boldsymbol{w}}$ is a continuity set for the law of $\boldsymbol{\lambda}$.*

Note that the matrix B_n in Assumption 1.1 and the operator B_n in Assumption 1.2 allow for different rates of convergence for different components of $\hat{\boldsymbol{g}}_n$, i.e. the rates q_i can be all the same but they do need to.

The main goal of Paper A is to study the asymptotic behaviour of $\hat{\boldsymbol{g}}_n^*$, as $n \to \infty$. For a general introduction to the subject of constrained inference we refer to [4, 18, 19]. In the continuous case, the asymptotic behaviour of the regression estimates under monotonic restriction was investigated in, for example, [8, 21]. The problem of estimating a monotone probability density function (pdf) was considered in [2, 9, 11, 17]. In [2] the authors studied a general scheme for order constrained inference in a continuous setup.

This research is mostly motivated by the paper [12], where the authors considered the problem of estimation of a discrete monotone probability mass function (pmf). It was shown that the limiting distribution of the constrained maximum likelihood estimator of a pmf is a concatenation of the isotonic regressions of Gaussian vectors over the periods of constancy of the true pmf $\boldsymbol{p}$. Compare to [12], in our work we do not require strong consistency of a basic estimator $\hat{\boldsymbol{g}}_n$, and we consider general preorder constraints.

In the discrete case some recent results are [5, 6, 10, 12]. The asymptotic distribution of the restricted parametric extremum estimator was considered in [1]. The asymptotic results in Paper A (even in the finite case) does not directly follow from the paper [1] when the basic estimator is linearly constrained, as, for example, in the case of a pmf estimation.

The computational aspects of the least squares estimation under bimonotonicity constraints, which is an example of a partial order, were studied in [7]. In the paper [20] the authors proposed the algorithms for weighted isotonic regression under order constraints specified by a directed acyclic graph.

1.2 Review of Paper A

In Paper A we study the estimators for the problem of estimating real valued functions that are defined on a general countable set and that are isotonic with respect to a preorder defined on that set. Given the Assumption 1.1 (or Assumption 1.2 in the infinite-dimensional case) we establish a limit distribution result for the proposed estimator $\hat{\boldsymbol{g}}_n^*$ of the form

$$B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi(\boldsymbol{\lambda})$$

in Theorems 3.7 and 4.5 of Paper A, where φ is a certain isotonic regression operator defined in (A.33).

Also, we consider the case of non-constant weights $\boldsymbol{w}$, i.e. when in (2) the vector of weights $\boldsymbol{w}$ depends on n. In this case the limit distribution of the isotonized estimator is given in Theorems 3.8 and 4.7.

The asymptotic results are applied to the problems of a bimonotone pmf and a regression function estimation.

1.3 Post-model-selection estimation of a discrete monotone distribution

Assume that $x_1, x_2, \ldots, x_n$ is an i.i.d. sample of random variables with unknown pmf $\boldsymbol{p}$. Suppose that $\boldsymbol{p} = \{p_i\}_{i \in \mathbb{N}_+}$ is a monotone decreasing pmf with the support in $\mathbb{N}_+$. Let $k = \sup\{i : p_i > 0\}$, with both cases $k < \infty$ and $k = \infty$ allowed. Assume that $\boldsymbol{p}$ has constant regions of the form

$$\begin{aligned} p_{q_1} = \cdots = p_{q_1+v_1-1} > p_{q_2} &= \cdots = p_{q_2+v_2-1} > \cdots > \\ p_{q_m} &= \cdots = p_k, \end{aligned} \tag{3}$$

where q_j, for $j = 1, \ldots m$, is the index of the first element in the j-th constant region, $p_{q_1} = p_1$, m is the total number of flat regions of $\boldsymbol{p}$, $\boldsymbol{v} = (v_1, \ldots, v_m)$ is the vector of the lengths (the numbers of points) in the constant regions of $\boldsymbol{p}$, so that $\sum_{j=1}^{m} v_j = k$.

In order to estimate $\boldsymbol{p}$ we introduce a two-step procedure. First, we perform a model selection to choose a class of the form

$$\begin{aligned} \mathcal{F}^*_{k,\boldsymbol{w}} = \Big\{ \boldsymbol{f} \in \mathbb{R}^k : f_1 = \cdots = f_{w_1} &\geq \\ f_{t_2} = \cdots = f_{t_2+w_2-1} \geq \ldots &\geq f_{t_s} = \cdots = f_k \Big\}, \end{aligned} \tag{4}$$

In the case of an infinite support, when $k = \infty$, (or when k is very large) we pick a finite r and introduce $\mathcal{F}^*_{k,\boldsymbol{w}} = \mathcal{F}^*_{k,\boldsymbol{w},r}$ as the following cone in l^2

$$\begin{aligned} \mathcal{F}^*_{k,\boldsymbol{w},r} = \Big\{ \boldsymbol{f} \in l^2 : f_1 = \cdots = f_{w_1} \geq \\ f_{t_2} = \cdots = f_{t_2+w_2-1} \geq \cdots \geq f_{t_d} = \cdots = f_r \geq \\ f_{r+1} \geq f_{r+2} \geq f_{r+3} \geq \cdots \Big\}. \end{aligned} \tag{5}$$

In general, $\boldsymbol{v}$, the vector of constant regions of the true pmf $\boldsymbol{p}$, and, consequently, $\boldsymbol{q}$, the vector of indices of the first elements in the constant region,

defined in (3), are unknown. Furthermore, for any choice of a candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$, given in (4) or (5), in general, the vector of constant regions $\boldsymbol{w}$ of the candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$ may or may not be equal to the vector $\boldsymbol{v}$ of constant regions of the true pmf $\boldsymbol{p}$, thus we may or may not have $\mathcal{F}^*_{k,\boldsymbol{w}} = \mathcal{F}^*_{k,\boldsymbol{v}}$.

We aim to find the class $\mathcal{F}^*_{k,\boldsymbol{w}}$ (or, equivalently, the vector $\boldsymbol{w}$), which contains $\boldsymbol{p}$. We derive the following Akaike-type information criterion

$$CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}}, n) \quad = \quad -2l(x_1, \ldots, x_n|\hat{\boldsymbol{p}}^*_n) + B(\boldsymbol{w}), \tag{6}$$

where $l(x_1, \ldots, x_n|\hat{\boldsymbol{p}}^*_n)$ is the log-likelihood and the bias term $B(\boldsymbol{w})$ is given by a certain sum of the level probabilities, defined in Paper B.

The selected model class, based on CMAIC, is

$$\hat{\boldsymbol{w}}_n \quad = \quad \text{argmin}_{\boldsymbol{w}} CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}}, n)$$

and the post-model-selection estimator $\hat{\boldsymbol{p}}^*_n$ of a decreasing pmf $\boldsymbol{p}$ is then given by

$$\hat{\boldsymbol{p}}^*_n = \sum_{j=1}^{S} \hat{\boldsymbol{p}}^*_n(\boldsymbol{w}_j) 1\{\hat{\boldsymbol{w}}_n = \boldsymbol{w}_j\}, \tag{7}$$

where

$$\hat{\boldsymbol{p}}^*_n(\boldsymbol{w}_j) = \underset{f \in \mathcal{F}^*_{k,\boldsymbol{w}_j}}{\text{argmin}} \sum_i [\hat{p}_{n,i} - f_i]^2, \tag{8}$$

is the projection of the empirical estimator $\hat{\boldsymbol{p}}_n$ onto the cone $\mathcal{F}^*_{k,\boldsymbol{w}_j}$.

The main goal of Paper B is to study the asymptotic behaviour of $\hat{\boldsymbol{p}}^*_n$. This research is also motivated by the paper by Jankowski and Wellner [12], discussed above. Next, we mention the paper [22], where the author studied the problem of isotonic regression based on i.i.d. data of an estimand with continuous support and proposed grouping of adjacent observations, isotonisation of the corresponding means and then interpolation to the whole support.

An information criterion for the parameters under simple order restrictions was proposed in [3]. A generalization of this criterion in the one-way ANOVA (ORIC) was proposed in [13]. A further generalisation of ORIC to multivariate normal linear models (GORIC) is given in [14]. These results on model selection are not directly applicable to our problem of selecting a proper class $\mathcal{F}^*_{k,\boldsymbol{w}}$ in (4).

1.4 Review of Paper B

In Paper B we study the post-model-selection estimator of a decreasing pmf, defined in (7). First, we describe a computational algorithm for $\hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_j)$, which is the l^2 projection onto the cone $\mathcal{F}^*_{k,\boldsymbol{w}_j}$, cf. (8), and obtain its limit distribution. Second, we derive the information criterion $CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}}, n)$ and prove that it provides a conservative model selection procedure. Third, we prove that the model selection based estimator $\hat{\boldsymbol{p}}_n^*$ performs asymptotically better, in l^2-sense, that the regular Grenander estimator, in the sense of having an almost surely asymptotically smaller l^2-loss, i.e. it satisfies

$$\mathbb{P}[\liminf_{n\to\infty}\{||\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}||_2^2 \leq ||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2^2\}] = 1$$

and, consequently, there exists n_1 such that for all $n > n_1$ one has

$$\mathbb{E}[n||\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}||_2^2] \leq \mathbb{E}[n||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2^2].$$

2 Statistical aspects of neutron detection

In the second part of the thesis, in two papers, we discuss the possibilities of a statistical determination of a wavelength and a wavelength distribution, respectively, for the new generation of multilayer neutron detectors, being developed at the European Spallation Source (ESS), situated in Lund, Sweden.

2.1 Case of a unimodal process

We assume that the incident beam $X_0(t)$ is a Poisson process with intensity λ. First, we consider the case when all particles in the beam have the same wavelength μ, i.e. a unimodal case. Assume that an incident beam of neutrons hits the first layer of the detector, cf. Figure 1. At the layer a neutron can possibly be absorbed and detected. If a neutron is not absorbed it will go through the detector's layer. We assume that these are the only two possibilities for the neutron interaction with a layer. Let p be the probability of an absorption of a neutron, so that $1 - p$ is the probability of its transmission. If a neutron is absorbed, it will then be detected.

Let $X_i(t)$ be the number of neutrons absorbed at the layer i in the time interval $[0, t]$. The values of $X_i(t)$, for $i = 1, \ldots, k$, represent the data available at the

experiment. Estimators of the wavelength μ can be indirectly obtained via estimates of the thinning parameters p, using a functional relation between the wavelength and the thinning probability, as explained in Paper C. Therefore, we aim to estimate the parameters (p, λ).

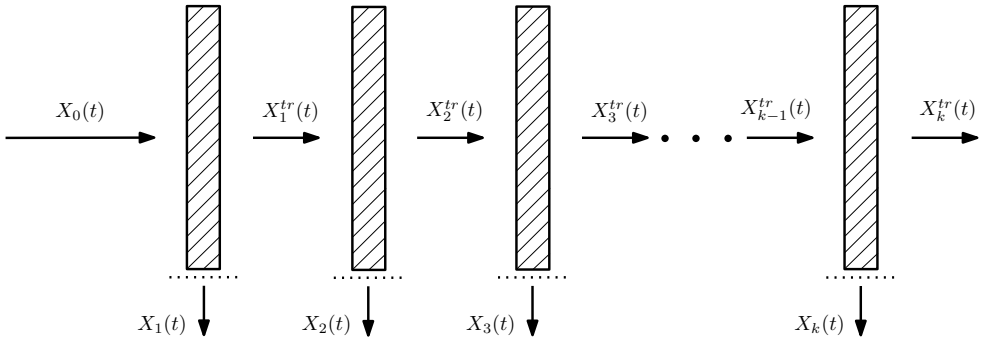

Figure 1: The scheme of the detector.

2.2 Review of Paper C

We study the maximum likelihood estimator for the parameters (p, λ) and show its consistency and asymptotic normality as the number of incoming neutrons goes to infinity. We combine these results with known results on the relation between the absorption probability p and the wavelength μ to derive an estimator of the wavelength and to show its consistency and asymptotic normality.

2.3 Case of a multimode process

Next, we consider the case of a multimode Poisson process. Assume that the neutron beam, i.e. the process $X_0(t)$, has a constant intensity λ. Assume, furthermore, that there are $s > 1$ different kinds of neutrons in the beam, with different wavelengths $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_s)$, such that

$$\mu_1 < \mu_2 < ... < \mu_s. \tag{9}$$

The values of the wavelengths are assumed to be unknown.

We model the neutron beam, or counting process $X_0(t)$, as the sum of the counting processes that count the number of neutrons that arrive at the face

of the detector in $[0,t]$, for the individual type neutrons. Thus, we let the number of neutrons with wavelength μ_r, which we may label r-neutrons, be denoted by $X_0^{(r)}(t)$, where $X_0^{(r)}(t)$ is a counting process such that $X_0^{(r)}(t) = 0$ and with intensity λ_r, for $r = 1, \ldots, s$. We write $X_0(t) = \sum_{r=1}^{s} X_0^{(r)}(t)$ for the total number of neutrons that arrive at the face of the detector; then $X_0(t)$ is a counting process with $X_0(0) = 0$.

For a given number $X_0(t) = x_0$ of the total incoming neutrons in the time interval $[0,t]$, the vector $\left(X_0^{(1)}(t), X_0^{(2)}(t), ..., X_0^{(s)}(t)\right)$ is assumed to follow a multinomial distribution with parameters $\left(q_1, q_2, ..., q_s\right)$, i.e.

$$(X_0^{(1)} = x_0^{(1)}, ..., X_0^{(s)} = x_0^{(s)} | X_0 = x_0) \quad \in \quad Mult(x_0, q_1, q_2, ..., q_s), \qquad (10)$$

with

$$\begin{aligned} x_0^{(1)} + \cdots + x_0^{(s)} &= x_0, \\ q_1 + q_2 + ... + q_s &= 1. \end{aligned}$$

The vector of proportions of numbers of different neutrons $\boldsymbol{q} = (q_1, q_2, ..., q_s)$ is the spectrum, or distribution, of an incoming neutron beam $X_0(t)$. We note that $q_r = \lambda_r / \lambda$ and assume that $\boldsymbol{q}$ does not depend on t.

2.4 Review of Paper D

In this paper we propose estimators of the distribution of events of different kinds $\boldsymbol{q}$ in a multimode Poisson process. We give an explicit solution for the maximum likelihood estimator. The inference problem gives rise the Sylvester-Ramanujan system of equations, cf. [15, 16]. We derive strong consistency and asymptotic normality of the estimator. Also, we consider the case of a decreasing spectrum of an incident beam.

3 Bibliography

[1] Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica* **67**, 1341–1383.

[2] Anevski, D. and Hössjer O. (2006). A general asymptotic scheme for inference under order restrictions. *The Annals of Statistics* **34**, 1874–1930.

[3] Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika* **86** 141–152.

[4] Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). Statistical inference under order restrictions. John Wiley & Sons, London-New York-Sydney.

[5] Balabdaoui, F., Durot, C., Koladjo, F. (2014). On asymptotics of the discrete convex LSE of a pmf. *Bernoulli* **23**, 1449–1480.

[6] Balabdaoui, F. and Jankowski, H. (2016). Maximum likelihood estimation of a unimodal probability mass function. *Statistica Sinica* **26**, 1061–1086.

[7] Beran, R. and Dümbgen, L. (2010). Least squares and shrinkage estimation under bimonotonicity constraints. *Statistics and Computing* **20**, 177–189.

[8] Brunk, H. D. (1970). Estimation of isotonic regression. Nonparametric Techniques in Statistical Inference. 177–195. Cambridge University Press

[9] Carolan, C. and Dykstra, R. (1999). Asymptotic behavior of the Grenander estimator at density flat regions. *The Canadian Journal of Statistics* **27** 557–566.

[10] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics* **43**, 1774–1800.

[11] Grenander, U. (1956). On the theory of mortality measurement. *Skand. Aktuarietidskr.* **39** 125–153.

[12] Jankowski, H. K. and Wellner, J. A. (2009). Estimation of a discrete monotone distribution. *Electronic journal of statistics* **39**, 125–153.

[13] Kuiper, R. M., Hoijtink, H. and Silvapulle, M. J. (2011). An Akaike-type information criterion for model selection under inequality constraints. *Biometrika* **98** 495–501.

[14] Kuiper, R. M., Hoijtink, H. and Silvapulle, M. J. (2011). Generalization of the order-restricted information criterion for multivariate normal linear models. *Journal of Statistical Planning and Inference* **142** 2454–2463.

[15] Lyubich Y. I., (2004). The Sylvester-Ramanujan system of equations and the complex power moment problem. *The Ramanujan Journal* **8**, 23–45

[16] Ramanujan S. (1912). Note on a set of simultaneous equations. *Journal of Indian Mathematical Society* **IV**, 94–96.

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

[18] Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). Order restricted statistical inference. John Wiley & Sons, Ltd., Chichester.

[19] Silvapulle, M. J. and Sen, P. K. (2005). Constrained Statistical Inference John Wiley & Sons, Ink., Hoboken, New Jersey.

[20] Stout, Q. F. (2013). Isotonic Regression via Partitioning. *Algorithmica* **66**, 93–112.

[21] Wright, F. T. (1981). The asymptotic behaviour of monotone regression estimates. *The Annals of Statistics* **9**, 443–448.

[22] Wright, F. T. (1982). Monotone regression estimates for grouped observations. *The Annals of Mathematical Statistics* **10** 278–286.

A

Paper A

The asymptotic distribution of the isotonic regression estimator over a general countable preordered set

Dragi Anevski and Vladimir Pastukhov

Centre for Mathematical Sciences, Lund University

Abstract

We study the isotonic regression estimator over a general countable preordered set. We obtain the limiting distribution of the estimator and study its properties. It is proved that, under some general assumptions, the limiting distribution of the isotonized estimator is given by the concatenation of the separate isotonic regressions of the restrictions of an underlying estimator's asymptotic distribution to the comparable level sets of the underlying estimator's probability limit. Also, we show that the isotonization preserves the rate of convergence of the underlying estimator. We apply these results to the problems of estimation of a bimonotone regression function and estimation of a bimonotone probability mass function.

Keywords: Constrained inference, isotonic regression, limit distribution.

1 Introduction

In this paper we study estimators for the problem of estimating real valued functions that are defined on a countable set and that are monotone with respect to a preorder defined on that set. In the situation when there exists an underlying, empirical, estimator of the function, which is not necessarily monotone, but for which one has (process) limit distribution results, we are

able to provide limit distribution results for the estimators. Our results can be applied to the special cases of probability mass function (pmf) estimation and regression function estimation. In the case of estimating a bimonotone pmf, i.e. a pmf which is monotone with respect to the usual matrix preorder on $\mathbb{Z}_+^2$, we state the limit distribution of the order restricted maximum likelihood estimator (mle), thereby generalising previously obtained results by [16], who treated the one-dimensional case, i.e. the mle of a monotone pmf on $\mathbb{Z}_+$. In fact we are able to state limit distribution results for the mle of a monotone pmf on $\mathbb{Z}_+^d$, for arbitrary $d > 1$, cf. Theorem 5.4 below. In the case of estimating a bimonotone regression function, i.e. a function defined on $\mathbb{Z}_+^2$ that is monotone with respect to the matrix preorder on $\mathbb{Z}_+^2$, we state the limit distribution of the isotonic regression estimator, again generalising previously known results for the isotonic regression on $\mathbb{Z}_+$, cf. [16]. In this setting we would also like to mention [7], that studied algorithms resulting from the minimisation of a smooth criterion function under bimonotonicity constraints. In the regression setting we are able to derive the limit distributions for the isotonic regression of functions that are monotone with respect to the matrix preorder on $\mathbb{Z}_+^d$, for arbitrary $d > 1$, cf. Theorem 5.3.

We would like to emphasize that the general approach taken in this paper allows for other preorders than the usual matrix order on $\mathbb{Z}^d$. Furthermore, our approach allows for also other starting basic empirical estimators; one could e.g. consider non-i.i.d. data settings, treating e.g. stationary (spatially homogenous) dependent data. In fact one can consider our estimator as the final step in a, at least, two-step, approach where in the first, or next-to-last, step, one provides the "empirical" estimator $\hat{\boldsymbol{g}}_n$ of the estimand $\mathring{\boldsymbol{g}}$, for which it is necessary to have established (process) limit distribution result of the form

$$n^{1/2}(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}}) \xrightarrow{d} \boldsymbol{\lambda} \tag{A.1}$$

on the appropriate space, e.g. l^2, see Assumptions 2.1 and 2.2 below. Note that we have simplified Assumptions 2.1 and 2.2 slightly in $(A.1)$ for illustrative purposes; the rate $n^{1/2}$ in $(A.1)$ is allowed to differ, even between the components in the vector $\hat{\boldsymbol{g}}_n$. Given the assumption $(A.1)$ we then establish a limit distribution result for the proposed estimator $\hat{\boldsymbol{g}}_n^*$ of the form

$$n^{1/2}(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi(\boldsymbol{\lambda}), \tag{A.2}$$

in Theorems 3.7 and 4.5, where φ is a certain isotonic regression operator defined in the sequel.

The general approach in this paper is somewhat reminiscent to the approach taken in [1], in which one considered a two-step general procedure for isotonization, allowing e.g. different types of dependence structures on the data. The difference to our paper is that we treat the, arguably, more complex notion of monotonicity with respect to preorders on d-dimensional spaces, whereas [1] only treated monotonicity in the one dimensional setting, and, furthermore, that we treat only functions with discrete or countable support, such as pmfs, whereas [1] treated functions with continuous support, such as pdfs.

This work is mainly motivated by the results obtained in [5, 16]. In [16] the problem of estimation of a discrete monotone distribution was studied in detail. It was shown that the limiting distribution of the constrained mle of a pmf is a concatenation of the isotonic regressions of Gaussian vectors over the periods of constancy of the true pmf $\boldsymbol{p}$, cf. Theorem 3.8 in [16]. In the derivation of the limiting distribution in [16] the authors used the strong consistency of the empirical estimator of $\boldsymbol{p}$ as well as the fact that the constrained mle is given by the least concave majorant (lcm) of the empirical cumulative distribution function (ecdf).

The problem of maximum likelihood estimation of a unimodal pmf was studied in [5]. That problem is different from the one being considered here, since [5] treats only pmfs on $\mathbb{Z}$, whereas we are able to treat multivariate problems with our approach.

In our work we do not require strong consistency of a basic estimator $\hat{\boldsymbol{g}}_n$, and we consider general preorder constraints, resulting in an expression for the isotonic regression that is more complicated than the lcm of the ecdf, c.f Assumptions 2.1 and 2.2. Also it turns out that the limiting distribution of the isotonized estimator $\hat{\boldsymbol{g}}_n^*$ can be split deeper than to the level sets of $\mathring{\boldsymbol{g}}$, which are the analogues of the periods of constancy of $\mathring{\boldsymbol{g}}$ in the univariate case.

For a general introduction to the subject of constrained inference we refer to the monographs: Barlow R. E. et al. [3], Robertson T. et al. [20], Silvapulle M. J. [22] and Groeneboom P. et al. [14]. In these monographs the problem of an isotonic regression has been considered in different settings, and, in particular, basic questions such as existence and uniqueness of the estimators have been addressed. In Lemmas 3.1 and 4.2 below we list those properties which will be used in the proofs of our results.

The asymptotic behaviour of the regression estimates over a continuous setup under monotonic restriction was first studied in [9, 24], where it was shown

that the difference of the regression function and its estimate multiplied by $n^{1/3}$, at a point with a positive slope, has a nondegenerate limiting distribution. The problem of estimating a monotone pdf was studied, for example, in [1, 10, 13, 19]. In [1] the authors studied a general scheme for order constrained inference in a continuous setup. In the discrete case some recent results are [4, 5, 11, 16]. In [11] the authors studied risk bounds for isotonic regression.

The remainder of this paper is organised as follows. In Section 2 we introduce some notations and define the estimators. In Section 3 we consider the finite dimensional case. Theorem 3.7 gives the asymptotic distribution of the isotonized estimator. Next, in Section 4 we consider the infinite dimensional case, which is quite different from the finite one. Theorem 4.5 describes the asymptotic behaviour of the isotonized estimator for the infinite dimensional case. In Section 5 we first discuss the application of the obtained results to the problems of estimation of a bimonotone regression function and of a bimonotone probability mass function, respectively, and then the corresponding limit distribution result for d-monotone functions, for an arbitrary $d > 1$. The limit distributions are stated in Theorems 5.1, 5.2, 5.3 and 5.4. In Section 5 we make some final comments about our results and relations to similar problems. We have gathered proofs of some intermediate results that are stated in the main body of the paper in an Appendix.

2 The inference problem and notations

In order to introduce the inference problem in detail, we start by introducing some notations. Let $\mathcal{X}$ be a countable set $\{x_1, x_2, \dots\}$ with $|\mathcal{X}| \leq \infty$, with a preorder $\preceq$ defined on it. We begin with the definitions of the order relations on an arbitrary set $\mathcal{X}$ and of an isotonic regression over it, cf. also [3, 20, 22].

Definition 2.1. *A binary relation $\preceq$ on $\mathcal{X}$ is a simple order if*

(i) it is reflexive, i.e. $x \preceq x$ for $x \in \mathcal{X}$;

(ii) it is transitive, i.e. $x_1, x_2, x_3 \in \mathcal{X}$, $x_1 \preceq x_2$ and $x_2 \preceq x_3$ imply $x_1 \preceq x_3$;

(iii) it is antisymmetric, i.e. $x_1, x_2 \in \mathcal{X}$, $x_1 \preceq x_2$ and $x_2 \preceq x_1$ imply $x_1 = x_2$;

(iv) every two elements of $\mathcal{X}$ are comparable, i.e. $x_1, x_2 \in \mathcal{X}$ implies that either $x_1 \preceq x_2$ or $x_2 \preceq x_1$.

A binary relation $\preceq$ on $\mathcal{X}$ is a partial order if it is reflexive, transitive and antisymmetric, but there may be noncomparable elements. A preorder is reflexive and transitive but not necessary antisymmetric and the set $\mathcal{X}$ can have noncomparable elements. Note, that in some literature the preorder is called as a quasi-order.

Let us introduce the notation $x_1 \sim x_2$, if x_1 and x_2 are comparable, i.e. if $x_1 \preceq x_2$ or $x_2 \preceq x_1$.

Definition 2.2. *A function $f(x) : \mathcal{X} \to \mathbb{R}$ is isotonic if $x_i, x_j \in \mathcal{X}$ and $x_i \preceq x_j$ imply $f(x_i) \leq f(x_j)$.*

Let $\mathcal{F}^{is} = \mathcal{F}^{is}(\mathcal{X})$ denote the family of real valued bounded functions f on a set $\mathcal{X}$, which are isotonic with respect to the preorder $\preceq$ on $\mathcal{X}$. In the case when $|\mathcal{X}| = \infty$ we consider the functions from the space l_2^w, the Hilbert space of real-valued functions on $\mathcal{X}$, which are square summable with some given non-negative weights $\boldsymbol{w} = \{w_1, w_2, \dots\}$, i.e. any $g \in l_2^w$ satisfies $\sum_{i=1}^{\infty} g(x_i)^2 w_i < \infty$. We use the same notation $\mathcal{F}^{is}$ to denote the functions from l_2^w which are isotonic with respect to the preorder $\preceq$.

Definition 2.3. *A function $g^* : \mathcal{X} \to \mathbb{R}$ is the isotonic regression of a function $g : \mathcal{X} \to \mathbb{R}$ over the preordered set $\mathcal{X}$ with weights $\boldsymbol{w} \in \mathbb{R}^s_+$, with $s \leq \infty$, if*

$$g^* = \operatorname*{argmin}_{f \in \mathcal{F}^{is}} \sum_{x \in \mathcal{X}} (f(x) - g(x))^2 w_x,$$

where $w_{x_i} = w_i$, for $i = 1, \dots, s$.

Conditions for existence and uniqueness of g^* will be stated below.

Similarly one can define an isotonic vector in $\mathbb{R}^s$, with $s \leq \infty$, and the isotonic regression of an arbitrary vector in $\mathbb{R}^s$. Let us consider a set of indices $\mathcal{I} = \{1, \dots, s\}$, with $s \leq \infty$, with some preorder $\preceq$ defined on it.

Definition 2.4. *A vector $\boldsymbol{\theta} \in \mathbb{R}^s$, with $s \leq \infty$, is isotonic if $i_1, i_2 \in \mathcal{I}$ and $i_1 \preceq i_2$ imply $\theta_{i_1} \leq \theta_{i_2}$.*

We denote the set of isotonic vectors in $\mathbb{R}^s$, with $s \leq \infty$, by $\boldsymbol{\mathcal{F}}^{is} = \boldsymbol{\mathcal{F}}^{is}(\mathcal{I})$. In the case of an infinite index set we consider the square summable vectors (with weights $\boldsymbol{w}$) from l_2^w, the Hilbert space of all square summable vectors with weights $\boldsymbol{w}$.

Definition 2.5. *A vector $\boldsymbol{\theta}^* \in \mathbb{R}^s$, with $s \leq \infty$, is the isotonic regression of an arbitrary vector $\boldsymbol{\theta} \in \mathbb{R}^s$ (or $\boldsymbol{\theta} \in \boldsymbol{l}_2^w$, if $s = \infty$) over the preordered index set $\mathcal{I}$ with weights $\boldsymbol{w} \in \mathbb{R}^s_+$ if*

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\xi} \in \boldsymbol{\mathcal{F}}^{is}}{\operatorname{argmin}} \sum_{i \in \mathcal{I}} (\xi_i - \theta_i)^2 w_i.$$

Given a set $\mathcal{X}$ with a preorder $\preceq$ on it one can generate a preorder on the set $\mathcal{I} = \{1, 2 \dots\}$ of indices of the domain in $\mathcal{X}$ as follows. For $i_1, i_2 \in \mathcal{I}$, $i_1 \preceq i_2$ if and only if $x_{i_1} \preceq x_{i_2}$. This preorder on the index set $\mathcal{I}$ will be called the preorder induced by the set $\mathcal{X}$ and will be denoted by the same symbol $\preceq$. Conversely, if one starts with the set $\mathcal{I}$ consisting of the indices of the elements in $\mathcal{X}$, and $\preceq$ is a preorder on $\mathcal{I}$, the above correspondence defines a preorder on $\mathcal{X}$. Therefore, in the sequel of the paper a bold symbol, e.g. $\boldsymbol{g}$, will denote a vector in $\mathbb{R}^s$, with $s \leq \infty$, whose i-th component is given by $g_i = g(x_i)$, for $i = 1, \dots, s$, where $g(x)$ is a bounded real valued function on $\mathcal{X}$. In this case we will say that the vector $\boldsymbol{g}$ corresponds to the function $g(x)$ on $\mathcal{X}$ and vice versa.

Note 1. *A real valued function $f(x)$ on the countable set $\mathcal{X}$ with the preorder $\preceq$, defined on it, is isotonic if and only if its corresponding vector $\boldsymbol{f} \in \mathbb{R}^s$, with $s \leq \infty$, is an isotonic vector with respect to the corresponding preorder $\preceq$ on its index set $\mathcal{I} = \{1, 2, \dots\}$, induced by the preorder on $\mathcal{X}$. A real valued function $g^*(x)$ on the set $\mathcal{X}$ is the isotonic regression of a function $g(x)$ with weights $\boldsymbol{w}$ if and only if its corresponding vector $\boldsymbol{g}^* \in \mathbb{R}^s$ is the isotonic regression of the vector $\boldsymbol{g} \in \mathbb{R}^s$ with respect to the corresponding preorder $\preceq$ on its index set $\mathcal{I} = \{1, 2, \dots\}$ with weights $\boldsymbol{w}$.*

To state the inference problem treated in this paper, suppose that $\mathcal{X}$ is a finite or an infinite countable preordered set and $\mathring{g} \in \mathcal{F}^{is}$ is a fixed unknown function. Suppose we are given observations z_i, $i = 1, \dots, n$, independent or not, that depend on the (parameter) $\mathring{g}$ in some way. In the sequel we will treat in detail two important cases:
The data $z_1, \dots, z_n$ are observations of either of

(i) Z_i, $i = 1, \dots, n$ independent identically distributed random variables taking values in $\mathcal{X}$, with probability mass function $\mathring{g}$.

(ii) $Z_i = (\boldsymbol{x}_i, Y_i)$, $i = 1, \dots, n$, with $\boldsymbol{x}_i$ deterministic (design) points in $\mathcal{X}$ and Y_i real valued random variables defined in the regression model

$$Y_i = \mathring{g}(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where ε_i is a sequence of identically distributed random variables with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2 < \infty$.

Now assume that

$$\hat{\boldsymbol{g}}_n = \hat{\boldsymbol{g}}_n(z_1, \ldots, z_n)$$

is a $\mathbb{R}^s$-valued statistic. We will call the sequence $\{\hat{\boldsymbol{g}}_n\}_{n\geq 1}$ the basic estimator of $\mathring{\boldsymbol{g}}$. In order to discuss consistency and asymptotic distribution result we introduce the following basic topologies: When $s < \infty$, we study the Hilbert space with the inner product $\langle \boldsymbol{g}_1, \boldsymbol{g}_2 \rangle = \sum_{i=1}^{s} g_{1,i} g_{2,i} w_i$, for $\boldsymbol{g}_1, \boldsymbol{g}_2 \in \mathbb{R}^s$, endowed with its Borel σ-algebra $\mathcal{B} = \mathcal{B}(\mathbb{R}^s)$ and when $s = \infty$ we study the space $\boldsymbol{l}_2^{\boldsymbol{w}}$ with the inner product $\langle \boldsymbol{g}_1, \boldsymbol{g}_2 \rangle = \sum_{i=1}^{\infty} g_{1,i} g_{2,i} w_i$, for a fixed weight vector $\boldsymbol{w}$ satisfying

$$\begin{cases} \inf\limits_i \{w_i\} > 0 \\ \sup\limits_i \{w_i\} < \infty, \end{cases} \tag{A.3}$$

and we equip $\boldsymbol{l}_2^{\boldsymbol{w}}$ with its Borel σ-algebra $\mathcal{B} = \mathcal{B}(\boldsymbol{l}_2^{\boldsymbol{w}})$.

Now define the isotonized estimator $\hat{\boldsymbol{g}}_n^*$ by

$$\hat{\boldsymbol{g}}_n^* = \operatorname*{argmin}_{\boldsymbol{\xi} \in \boldsymbol{\mathcal{F}}^{is}} \sum_{i \in \mathcal{I}} (\xi_i - \hat{g}_{n,i})^2 w_i. \tag{A.4}$$

The main goal of this paper is to study the asymptotic behaviour of $\hat{\boldsymbol{g}}_n^*$, as $n \to \infty$.

We make the following assumptions on the basic estimator $\hat{\boldsymbol{g}}_n$, for the finite, $s < \infty$, and the infinite, $s = \infty$, support case, respectively.

Assumption 2.1. *Suppose that $s < \infty$. Assume that $\hat{\boldsymbol{g}}_n \xrightarrow{p} \mathring{\boldsymbol{g}}$ for some $\mathring{\boldsymbol{g}} \in \boldsymbol{\mathcal{F}}^{is}$ and $B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}}) \xrightarrow{d} \boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is a random vector in $(\mathbb{R}^s, \mathcal{B})$ and B_n is a diagonal $s \times s$ matrix with elements $[B_n]_{ii} = n^{q_i}$ with q_i being real positive numbers.*

Assumption 2.2. *Suppose that $s = \infty$. Let $\hat{\boldsymbol{g}}_n$, for $n = 1, 2, 3, \ldots$, be a tight sequence of random vectors taking values in the Hilbert space $\boldsymbol{l}_2^{\boldsymbol{w}}$. Assume that $\hat{\boldsymbol{g}}_n \xrightarrow{p} \mathring{\boldsymbol{g}}$ for some $\mathring{\boldsymbol{g}} \in \boldsymbol{\mathcal{F}}^{is}$, and $B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}}) \xrightarrow{d} \boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is a random vector in $(\boldsymbol{l}_2^{\boldsymbol{w}}, \mathcal{B})$ and B_n is a linear operator $\boldsymbol{l}_2^{\boldsymbol{w}} \to \boldsymbol{l}_2^{\boldsymbol{w}}$, such that for any $\boldsymbol{g} \in \boldsymbol{l}_2^{\boldsymbol{w}}$ it holds that $(B_n \boldsymbol{g})_i = n^{q_i} g_i$, with q_i being the real positive numbers. Suppose also that any finite s-dimensional cylinder set in $\boldsymbol{l}_2^{\boldsymbol{w}}$ is a continuity set for the law of $\boldsymbol{\lambda}$.*

Note that the matrix B_n in Assumption 2.1 and the operator B_n in Assumption 2.2 allow for different rates of convergence for different components of $\hat{\boldsymbol{g}}_n$, i.e. the rates q_i can be all the same but they do need to. The values of q_i will be specified later.

3 The case of finitely supported functions

Let us assume that $s < \infty$, i.e. that the basic estimator $\{\hat{\boldsymbol{g}}_n\}_{n\geq 1}$ is a sequence of finite-dimensional vectors. The next lemma states some well-known general properties of the isotonic regression of a finitely supported function.

Lemma 3.1. *Suppose Assumption 2.1 holds. Let $\hat{\boldsymbol{g}}_n^* \in \mathbb{R}^s$ be the isotonic regression of the vector $\hat{\boldsymbol{g}}_n$, defined in (A.4), for $n = 1, 2, 3, \ldots$. Assume also that $a \leq \hat{g}_{n,i} \leq b$ holds for some constants $-\infty < a < b < \infty$, for all $n = 1, 2, \ldots$ and $i = 1, \ldots, s$. Then the following hold:*

(i) $\hat{\boldsymbol{g}}_n^*$ *exists and it is unique.*

(ii) $\sum_{i=1}^s \hat{g}_{n,i} w_i = \sum_{i=1}^s \hat{g}_{n,i}^* w_i$, *for all* $n = 1, 2, \ldots$.

(iii) $\hat{\boldsymbol{g}}_n^*$, *viewed as a mapping from* $\mathbb{R}^s$ *into* $\mathbb{R}^s$, *is continuous. Moreover, it is also continuous if it is viewed as a function on the* $2s$*-tuples of real numbers* $(w_1, w_2, \ldots, w_s, g_1, g_2, \ldots, g_s)$, *with* $w_i > 0$.

(iv) $\hat{\boldsymbol{g}}_n^*$ *satisfies the same bounds as the basic estimator, i.e.* $a \leq \hat{g}_{n,i}^* \leq b$, *for all* $n = 1, 2, \ldots$ *and* $i = 1, \ldots, s$.

(v) $\hat{\boldsymbol{g}}_n^*$ *is a consistent estimator of* $\mathring{\boldsymbol{g}}$, *i.e.* $\hat{\boldsymbol{g}}_n^* \xrightarrow{p} \mathring{\boldsymbol{g}}$.

(vi) $(\hat{\boldsymbol{g}}_n + \boldsymbol{c})^* = \hat{\boldsymbol{g}}_n^* + \boldsymbol{c}$ *for all constant vectors* $\boldsymbol{c} \in \mathbb{R}^s$,
$(c\hat{\boldsymbol{g}}_n)^* = c\hat{\boldsymbol{g}}_n^*$ *for all* $c \in \mathbb{R}_+$.

We make a partition of the original set into comparable sets $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(k)}$

$$\mathcal{X} = \cup_{v=1}^k \mathcal{X}^{(v)}, \tag{A.5}$$

where each partition set $\mathcal{X}^{(v)}$ contains elements such that if $x \in \mathcal{X}^{(v)}$, then x is comparable with at least one different element in $\mathcal{X}^{(v)}$ (if there are any), but not with any other element in $\mathcal{X}^{(\mu)}$ for any $\mu \neq v$. In fact, the partition can be constructed even for an infinite set $\mathcal{X} = \{x_1, x_2, \ldots\}$, since a preorder can

be represented as the directed graph and any graph can be partitioned into isolated connected components and the partition is unique, cf. [12].

Now assume that $\mathcal{X}$ is finite, that we are given the partition (A.5) and let $g^{(v)}(x)$, for $v = 1, \ldots, k$, be real valued functions defined on the sets $\mathcal{X}^{(v)}$ as $g^{(v)}(x) = g(x)$, whenever $x \in \mathcal{X}^{(v)}$, i.e. $g^{(v)}(x)$ is the restriction of the function $g(x)$ to the set $\mathcal{X}^{(v)}$. The family of functions $g^{(v)}(x)$ defined on the set $\mathcal{X}^{(v)}$, which are isotonic with respect to the preorder, will be denoted by $\mathcal{F}^{is}_{(v)}$, for $v = 1, \ldots, k$.

The next lemma states a natural result of an isotonic regression on $\mathcal{X}$, that it can be obtained as a concatenation of the individual isotonic regressions of the restrictions to the comparable sets.

Lemma 3.2. *Let $g(x)$ be an arbitrary real valued function on the finite set $\mathcal{X}$ with a preorder $\preceq$ defined on it, and assume that the partition (A.5) is given. Then the isotonic regression of $g(x)$ with any positive weights $\boldsymbol{w}$ with respect to the preorder $\preceq$ is equal to*

$$g^*(x) = g^{*(v)}(x), \text{ whenever } x \in \mathcal{X}^{(v)}, \tag{A.6}$$

where $g^{*(v)}(x)$ is the isotonic regression of the function $g^{(v)}(x)$ over the set $\mathcal{X}^{(v)}$ with respect to the preorder $\preceq$.

Now let $\mathring{g}(x)$ be the fixed function defined in Assumption 2.1, assume that we are given the partition (A.5) of $\mathcal{X}$ and for an arbitrary but fixed $v \in \{1, \ldots, k\}$ let $\mathring{g}_v(x)$ be the restriction of $\mathring{g}(x)$ to $\mathcal{X}^{(v)}$. Then if $N_v = |\mathcal{X}^{(v)}|$ we can introduce the vector $\mathring{\boldsymbol{g}}_v = (\mathring{g}_{v,1}, \ldots, \mathring{g}_{v,N_v}) = (\mathring{g}_v(x_{i_1}), \ldots, \mathring{g}_v(x_{i_{N_v}}))$, where $x_{i_1}, \ldots, x_{i_{N_v}}$ are the unique points in $\mathcal{X}^{(v)}$. Given $\mathring{g}_v(x)$ we can partition the set $\mathcal{X}^{(v)}$ into m_v sets

$$\mathcal{X}^{(v)} = \cup_{l=1}^{m_v} \mathcal{X}^{(v,l)}. \tag{A.7}$$

The partition is constructed in the following way: We note first that the N_v values in the vector $\mathring{\boldsymbol{g}}_v$ are not necessarily all unique, so there are $\tilde{m}_v \leq N_v$ unique values in $\mathring{\boldsymbol{g}}_v$. Then in a first step we construct $\tilde{m}_v$ level sets

$$\tilde{\mathcal{X}}^{(v,l)} = \{x \in \mathcal{X}^{(v)} : \mathring{g}_v(x_i) = \mathring{g}_{v,l}\}$$

with $l = 1, \ldots, \tilde{m}_v$.

Next we note that for any non-singleton level set $\tilde{\mathcal{X}}^{(v,l)}$ there might be non-comparable points, i.e. $x_i, x_j \in \tilde{\mathcal{X}}^{(v,l)}$ can be such that neither $x_i \preceq x_j$ nor $x_j \preceq x_i$ hold. Therefore, in the second step for each fixed l we can partition (if necessary) the level set $\tilde{\mathcal{X}}^{(v,l)}$ into sets with comparable elements, analogously to the construction of (A.5). We can do this for every v and end up in a partition (A.7) with $\tilde{m}_v \leq m_v \leq N_v$.

In the partition (A.7) each set $\mathcal{X}^{(v,l)}$ is characterised by

(i) for every $x \in \mathcal{X}^{(v,l)}$ we have $\mathring{g}_v(x) = \mathring{g}_{v,l}$,

(ii) if $|\mathcal{X}^{(v,l)}| \geq 2$ then for every $x \in \mathcal{X}^{(v,l)}$ there is at least one $x' \in \mathcal{X}^{(v,l)}$ such that $x \sim x'$.

We have therefore proved the following lemma.

Lemma 3.3. *For any countable set $\mathcal{X}$ with the preorder $\preceq$ and any isotonic function $\mathring{g}(x)$, defined on it, there exists a unique partition $\mathcal{X} = \cup_{v=1}^{k} \cup_{l=1}^{m_v} \mathcal{X}^{(v,l)}$, satisfying the statements (i) and (ii) above. For the index set $\mathcal{I}$ with the preorder $\preceq$ generated by the set $\mathcal{X}$ and any isotonic function $\mathring{g}(x)$, defined on $\mathcal{X}$, there exists a unique partition $\mathcal{I} = \cup_{v=1}^{k} \cup_{l=1}^{m_v} \mathcal{I}^{(v,l)}$, satisfying conditions analogous to (i) and (ii) stated above.*

Definition 3.4. *The set $\mathcal{X}$ will be called decomposable if in the partition, defined in (A.5), $k > 1$. In the partition (A.7) the sets $\mathcal{X}^{(v,l)}$ will be called the comparable level sets of $\mathring{g}(x)$. In the corresponding partition of the index set $\mathcal{I}$ the sets $\mathcal{I}^{(v,l)}$ will be called the comparable level index sets of $\mathring{g}$.*

Recall that $g^{(v,l)}(x)$ is the restriction of the function $g(x)$ to the comparable level set $\mathcal{X}^{(v,l)}$, for $l = 1, \ldots, m_v$ and $v = 1, \ldots, k$.

In the case of a non-decomposable set, the full partition will be written as $\mathcal{X} = \cup_{l=1}^{m_1} \mathcal{X}^{(1,l)} \equiv \cup_{l=1}^{m} \mathcal{X}^{(l)}$, so we may then drop the index $v = 1$. Similarly, in this case $g^{(1,l)}(x) \equiv g^{(l)}(x)$ denotes the restriction of a function $g(x)$ to the comparable level set $\mathcal{X}^{(l)} \equiv \mathcal{X}^{(1,l)}$.

Next, suppose that $\mathcal{X}$ is a non-decomposable set, and let us consider an arbitrary function $\mathring{g}(x) \in \mathcal{F}^{is}$. Assume that for $\mathring{g}(x)$ there has been made a partition $\mathcal{X} = \cup_{l=1}^{m} \mathcal{X}^{(l)}$ in (A.7), satisfying (i) and (ii). Define the smallest comparable level distance of $\mathring{g}$ as

$$\begin{aligned} \tilde{\varepsilon}' \;=\; & \inf\{|\mathring{g}_{l'} - \mathring{g}_l| : l, l' = 1, \ldots, m,\ l \neq l', \\ & \exists x_1 \in \mathcal{X}^{(l)}, \exists x_2 \in \mathcal{X}^{(l')}, \text{ such that } x_1 \sim x_2\}, \end{aligned} \tag{A.8}$$

Note, that $\tilde{\varepsilon}$ is always finite and for the finite support case, $s < \infty$, also $\tilde{\varepsilon} > 0$.

Lemma 3.5. *Consider an arbitrary real valued function $g(x)$ on a non-decomposable finite set $\mathcal{X}$ with the preorder $\preceq$ and let $\tilde{\varepsilon}$ be defined in (A.8). If*

$$\sup_{x \in \mathcal{X}} \{|g(x) - \mathring{g}(x)|\} < \tilde{\varepsilon}/2, \tag{A.9}$$

then the isotonic regression of $g(x)$ is given by

$$g^*(x) = g^{*(l)}(x), \text{ whenever } x \in \mathcal{X}^{(l)}, \tag{A.10}$$

where $g^{*(l)}(x)$ is the isotonic regression of the function $g^{(l)}(x)$ over the set $\mathcal{X}^{(l)}$ with respect to the preorder $\preceq$. Therefore, the function $g^*(x)$ is a concatenation of the isotonic regressions of the restrictions of $g(x)$ to the comparable level sets of $\mathring{g}(x)$.

The next lemma is an auxiliary result which will be used later in the proof of the asymptotic distribution of $\hat{\boldsymbol{g}}_n^*$.

Lemma 3.6. *Assume $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$ are sequences of random vectors, taking values in the space $\mathbb{R}^s$, for $s \leq \infty$, with some metric on it, endowed with its Borel σ-algebra. If $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ and $\lim_{n\to\infty} \mathbb{P}[\boldsymbol{X}_n = \boldsymbol{Y}_n] = 1$, then $\boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{X}$.*

Let us consider the sequence $B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}})$, where $\hat{\boldsymbol{g}}_n^*$ is the isotonic regression of $\hat{\boldsymbol{g}}_n$, which was defined in Assumption 2.1, and with a specified matrix B_n. As mentioned in Assumption 2.1, we allow different rates of convergence n^{q_i} for different components of $\hat{\boldsymbol{g}}_n$. We however require q_i, for $i = 1, \ldots, s$, to be equal on the comparable level index sets $\mathcal{I}^{(v,l)}$ of $\mathring{\boldsymbol{g}}$, i.e. q_i, for $i = 1, \ldots, s$, are real positive numbers such that $q_{i_1} = q_{i_2}$, whenever $i_1, i_2 \in \mathcal{I}^{(v,l)}$.

We introduce an operator $\varphi : \mathbb{R}^s \to \mathbb{R}^s$ defined in the following way. First, for any vector $\boldsymbol{\theta} \in \mathbb{R}^s$ we define the coordinate evaluation map $\theta(x) : \mathcal{X} \to \mathbb{R}$, corresponding to the vector $\boldsymbol{\theta}$, by $\theta(x_i) = \theta_i$, for $i = 1, \ldots, s$. Then, let $\theta^{*(v',l')}(x)$ be the isotonic regression of the restriction of $\theta(x)$ to the comparable level set $\mathcal{X}^{(v',l')}$ of $\mathring{g}(x)$, and define

$$\varphi(\boldsymbol{\theta})_i = \theta^{*(v',l')}(x_i), \tag{A.11}$$

for $i = 1, \ldots, s$, with (v', l') the (unique) indices such that $x_i \in \mathcal{X}^{(v',l')}$.

The asymptotic distribution of $B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}})$ is given in the following theorem.

Theorem 3.7. *Suppose that Assumption 2.1 holds. Then*

$$B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi(\boldsymbol{\lambda}), \tag{A.12}$$

where φ is the operator, defined in (A.11).

Proof.

First, from Lemma 3.3 we have that any preordered set $\mathcal{X}$ can be uniquely partitioned as

$$\begin{aligned} \mathcal{X} &= \cup_{v=1}^{k} \mathcal{X}^{(v)}, \\ \mathcal{X}^{(v)} &= \cup_{l=1}^{m_v} \mathcal{X}^{(v,l)}, \end{aligned} \tag{A.13}$$

and with the partition (A.13) of $\mathcal{X}^{(v)}$ determined by the isotonic vector $\mathring{\boldsymbol{g}}$.

Second, as shown in Lemma 3.2, the isotonic regression of $g(x)$ on the original set $\mathcal{X}$ can be obtained as a concatenation of the separate isotonic regressions of the restrictions of $g(x)$ to the non-decomposable sets in the partition (A.5). Therefore, without loss of generality, we can assume that the original set $\mathcal{X}$ is non-decomposable. Thus, any $x \in \mathcal{X}$ is comparable with at least one different element of $\mathcal{X}$, $k = 1$, and

$$\begin{aligned} \mathcal{X} &= \cup_{l=1}^{m_1} \mathcal{X}^{(1,l)} \\ &\equiv \cup_{l=1}^{m} \mathcal{X}^{(l)} \end{aligned}$$

and $\mathring{g}_{1,l} \equiv \mathring{g}_l$. Note, that we have dropped the index v.

Third, since $\hat{\boldsymbol{g}}_n$ is consistent, by Assumption 2.1, for any $\varepsilon > 0$,

$$\mathbb{P}[\sup_{x \in \mathcal{X}} \{|\hat{g}_n(x) - \mathring{g}(x)|\} < \varepsilon] \to 1, \tag{A.14}$$

as $n \to \infty$. Note that the comparable level distance $\tilde{\varepsilon}$ of $\mathring{g}$, defined in (A.8), satisfies $\tilde{\varepsilon} > 0$, and take $\varepsilon = \tilde{\varepsilon}/2$. Then from Lemma 3.5 we obtain

$$\{\sup_{x \in \mathcal{X}} \{|\hat{g}_n(x) - \mathring{g}(x)|\} < \tilde{\varepsilon}/2\} \subseteq \{\hat{\boldsymbol{g}}_n^* = \varphi(\hat{\boldsymbol{g}}_n)\}. \tag{A.15}$$

Therefore, (A.14) and (A.15) imply

$$\mathbb{P}[\hat{\boldsymbol{g}}_n^* = \varphi(\hat{\boldsymbol{g}}_n)] \to 1, \tag{A.16}$$

as $n \to \infty$.

Next, since the isotonic regression is a continuous map (statement (iii) of Lemma 3.1), the operator φ is a continuous map from $\mathbb{R}^s$ to $\mathbb{R}^s$. Therefore, using the continuous mapping theorem, cf. [23], we get

$$\varphi(B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}})) \xrightarrow{d} \varphi(\boldsymbol{\lambda}). \tag{A.17}$$

Furthermore, using statement (vi) of Lemma 3.1 and taking into account the definition of the matrix B_n, we get

$$\varphi(B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}})) = B_n(\varphi(\hat{\boldsymbol{g}}_n) - \mathring{\boldsymbol{g}}). \tag{A.18}$$

Then (A.16), (A.17) and (A.18) imply that

$$\mathbb{P}[B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}}) = B_n(\varphi(\hat{\boldsymbol{g}}_n) - \mathring{\boldsymbol{g}})] \to 1, \tag{A.19}$$

as $n \to \infty$. Finally, using Lemma 3.6, from (A.17) and (A.19) we prove that

$$B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi(\boldsymbol{\lambda}),$$

as $n \to \infty$. □

For a given preorder $\preceq$ on $\mathcal{X}$ there exists a matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{g} \geq \boldsymbol{0}$ is equivalent to $\boldsymbol{g}$ is isotonic with respect to $\preceq$, cf. Proposition 2.3.1 in [22]. Therefore, if there are no linear constraints imposed on the basic estimator $\hat{\boldsymbol{g}}_n$, Theorem 3.7 can also be established by using the results on estimation when a parameter is on a boundary, in Section 6 in [2].

Assume that each vector $\hat{\boldsymbol{g}}_n$ has the following linear constraint $\sum_{i=1}^s \hat{g}_{n,i} w_i = c$ (for example, in the case of estimation of a probability mass function it would be $\sum_{i=1}^s \hat{g}_{n,i} = 1$). Then, the expression for a limiting distribution in Theorem 3.7 does not follow directly from the results in [2] in the case when $\hat{\boldsymbol{g}}_n$ is linearly constrained. However, the result of Theorem 3.7 holds, because, as established in statement (ii) of Lemma 3.1, isotonic regression with weights $\boldsymbol{w}$ preserves the corresponding linear constraint.

Next we consider the case when the vector of weights $\boldsymbol{w}$ is not a constant, i.e. we assume that some non-random sequence $\{\boldsymbol{w}_n\}_{n\geq 1}$, where each vector $\boldsymbol{w}_n$ satisfies the condition (A.3), converges to some non-random vector $\boldsymbol{w}$, which

also satisfies (A.3). We denote by $\theta^{*\boldsymbol{w}}(x)$ the isotonic regression of $\theta(x)$ with weights $\boldsymbol{w}$ and analogously to (A.11) we introduce the notation $\varphi^{\boldsymbol{w}}(\boldsymbol{\theta})$

$$\varphi^{\boldsymbol{w}}(\boldsymbol{\theta})_i = \theta^{*\boldsymbol{w}(v',l')}(x_i), \tag{A.20}$$

where $\theta^{*\boldsymbol{w}(v',l')}(x)$ is the isotonic regression, with weights $\boldsymbol{w}$, of the restriction of $\theta(x)$ to the comparable level set $\mathcal{X}^{(v',l')}$ of $\mathring{g}(x)$, where the indices v' and l' are such that $x_i \in \mathcal{X}^{(v',l')}$. Define the isotonic regression $\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n}$ of the basic estimator $\hat{\boldsymbol{g}}_n$. The next theorem gives the limiting distribution of $\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n}$.

Theorem 3.8. *Suppose that Assumption 2.1 holds. Then the asymptotic distribution of the isotonic regression $\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n}$ of the basic estimator $\hat{\boldsymbol{g}}_n$ is given by*

$$B_n(\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n} - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi^{\boldsymbol{w}}(\boldsymbol{\lambda}), \tag{A.21}$$

where $\varphi^{\boldsymbol{w}}$ is the operator, defined in (A.20).

Proof. Without loss of generality, we can assume that the original set $\mathcal{X}$ is non-decomposable. First, since the sequence $\hat{\boldsymbol{g}}_n$ is consistent, then for any

$$\mathbb{P}[\sup_{x\in\mathcal{X}}\{|\hat{g}_n(x) - \mathring{g}(x)|\} < \tilde{\varepsilon}/2] \to 1, \tag{A.22}$$

as $n \to \infty$, with $\tilde{\varepsilon}$ defined in A.8. Using the statement of Lemma 3.5, we obtain

$$\{\sup_{x\in\mathcal{X}}\{|\hat{g}_n(x) - \mathring{g}(x)|\} < \tilde{\varepsilon}/2\} \subseteq \{\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n} = \varphi^{\boldsymbol{w}_n}(\hat{\boldsymbol{g}}_n)\}. \tag{A.23}$$

Note that the result of Lemma 3.5 holds for any weights $\boldsymbol{w}_n$.

Therefore, from (A.22) and (A.23) we have

$$\mathbb{P}[\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n} = \varphi(\hat{\boldsymbol{g}}_n)] \to 1, \tag{A.24}$$

as $n \to \infty$.

Second, from statement (iii) of Lemma 3.1, the operators $\varphi^{\boldsymbol{w}_n}$, $\varphi^{\boldsymbol{w}}$ are continuous maps from $\mathbb{R}^{2s}$ to $\mathbb{R}^s$, for all weights $\boldsymbol{w}_n$, $\boldsymbol{w}$ satisfying (A.3). Using the (extended) continuous mapping theorem, cf. [23], we get

$$\varphi^{\boldsymbol{w}_n}(B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}})) \xrightarrow{d} \varphi^{\boldsymbol{w}}(\boldsymbol{\lambda}), \tag{A.25}$$

where $\boldsymbol{w}$ is the limit of the sequence $\{\boldsymbol{w}_n\}_{n\geq 1}$.

Third, using statement (vi) of Lemma 3.1 and the definition of the matrix B_n we obtain

$$\varphi^{\boldsymbol{w}_n}(B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}})) = B_n(\varphi^{\boldsymbol{w}_n}(\hat{\boldsymbol{g}}_n) - \mathring{\boldsymbol{g}}). \tag{A.26}$$

Therefore, (A.26) gives us

$$\mathbb{P}[B_n(\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n} - \mathring{\boldsymbol{g}}) = B_n(\varphi^{\boldsymbol{w}_n}(\hat{\boldsymbol{g}}_n) - \mathring{\boldsymbol{g}})] \to 1, \tag{A.27}$$

as $n \to \infty$. Finally, using Lemma 3.6, from (A.25) and (A.27) we prove that

$$B_n(\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n} - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi^{\boldsymbol{w}}(\boldsymbol{\lambda}),$$

as $n \to \infty$. □

4 The case of infinitely supported functions

In this section we assume that the original set $\mathcal{X} = \{x_1, x_2 \dots\}$ is an infinite countable enumerated set with a preorder $\preceq$ defined on it.

In the case of infinitely supported functions the isotonic regression's properties are similar to the ones in the finite case, but the proofs are slightly different. For completeness we state these properties in the following lemma.

Lemma 4.1. *Suppose Assumption 2.1 holds. Let $\hat{\boldsymbol{g}}_n^* \in \boldsymbol{l}_2^{\boldsymbol{w}}$ be the isotonic regression of the vector $\hat{\boldsymbol{g}}_n \in \boldsymbol{l}_2^{\boldsymbol{w}}$, for $n = 1, 2, 3, \dots$. Assume also that $a \leq \hat{g}_{n,i} \leq b$ holds for some constants $-\infty < a < b < \infty$, for all $n = 1, 2, \dots$ and $i = 1, \dots, \infty$. Then statements (i) - (vi) of Lemma 3.1 hold, with (iii) suitably changed to the mapping from $\boldsymbol{l}_2^{\boldsymbol{w}}$ to $\boldsymbol{l}_2^{\boldsymbol{w}}$.*

We partition the original set $\mathcal{X}$ in the same way as it was done in the finite case, i.e., first, let

$$\mathcal{X} = \cup_{v=1}^{k} \mathcal{X}^{(v)}, \tag{A.28}$$

where $k \leq \infty$ is the number of sets and each set $\mathcal{X}^{(v)}$ is such that if $x \in \mathcal{X}^{(v)}$, then x is comparable with at least one different element in $\mathcal{X}^{(v)}$ (if there are any), but not with any other elements which belong to other sets in the

partition. Note, that $\mathcal{X}^{(v)}$ can have only one element. The partition of $\mathcal{X}$ is unique and $\mathcal{X}^{(v)} \cap \mathcal{X}^{(v')} = \emptyset$, for $v \neq v'$.

Furthermore, for the fixed function $\mathring{g} \in \mathcal{F}^{is}$, defined in Assumption 2.2 (or, equivalently, its corresponding isotonic vector $\mathring{\boldsymbol{g}} \in \boldsymbol{\mathcal{F}}^{is}$), we partition each set $\mathcal{X}^{(v)}$ in (A.28) into the comparable level sets of $\mathring{g}$, i.e.

$$\mathcal{X}^{(v)} = \cup_{l=1}^{m_v} \mathcal{X}^{(v,l)}, \tag{A.29}$$

in the same way as it was done in the finite case in (A.7).

Note, that since $\mathring{\boldsymbol{g}} \in \boldsymbol{l}_2^{\boldsymbol{w}}$ and condition (A.3) is satisfied, the cardinality of any set $\mathcal{X}^{(v,l)}$ is less than infinity whenever $\mathring{g}_{v,l} \neq 0$, otherwise we would have $\sum_{i=1}^{\infty} (\mathring{g}_i)^2 w_i = \infty$, which would mean that $\mathring{g} \notin l_2^{\boldsymbol{w}}$. The set $\mathcal{X}^{(v,l)}$ can have infinitely many elements only if $\mathring{g}_{v,l} = 0$.

For the partition in (A.28) we obtain a result similar to the one obtained in Lemma 3.2 for the finite case.

Lemma 4.2. *Let $g(x)$ be an arbitrary real valued function in $l_2^{\boldsymbol{w}}$ on the set $\mathcal{X}$ with a preorder $\preceq$ defined on it. Then the isotonic regression of $g(x)$ with any positive weights $\boldsymbol{w}$ is equal to*

$$g^*(x) = g^{*(v)}(x), \text{ whenever } x \in \mathcal{X}^{(v)}, \tag{A.30}$$

where $g^{*(v)}(x)$ is the isotonic regression of the restriction of the function $g(x)$ to the set $\mathcal{X}^{(v)}$ over this set with respect to the preorder $\preceq$.

As a consequence of Lemma 4.2, without loss of generality in the sequel of the paper we can assume that the original set $\mathcal{X}$ is non-decomposable and use the same notations as in the finite case, i.e. $\mathcal{X} = \cup_{l=1}^{m} \mathcal{X}^{(l)} \equiv \cup_{l=1}^{m_1} \mathcal{X}^{(1,l)}$ and, respectively, $g^{(l)}(x) \equiv g^{(1,l)}(x)$ for the restriction of the function $g(x)$ to the set $\mathcal{X}^{(l)}$.

In the case of an infinite support the result of Lemma 3.5 is generally not applicable, because the value of $\tilde{\varepsilon}$ can in this case be zero. We therefore make the following slight modification of Lemma 3.5. Thus, assume that for a function $\mathring{g}(x) \in \mathcal{F}^{is}$ we have made a partition $\mathcal{X} = \cup_{l=1}^{m} \mathcal{X}^{(l)}$ with $m \leq \infty$. Furthermore, for any finite positive integer number $m' < m \leq \infty$ we choose m' comparable level sets $\mathcal{X}^{(l_j)}$, such that the values of the function $\mathring{g}(x)$ on them satisfy $|\mathring{g}_{l_1}| \geq |\mathring{g}_{l_2}| \geq \cdots \geq |\mathring{g}_{l_{m'}}|$. Next, we rewrite the partition as

$$\mathcal{X} = \mathcal{X}^{(l_1)} \cup \mathcal{X}^{(l_2)} \cup \cdots \cup \mathcal{X}^{(l_{m'})} \cup \mathcal{X}^{(l_{m'+1})}, \tag{A.31}$$

where $\mathcal{X}^{(l_{m'+1})} = \mathcal{X} \setminus \mathcal{X}^{(l_1)} \cup \mathcal{X}^{(l_2)} \cup \cdots \cup \mathcal{X}^{(l_{m'})}$. Define

$$\begin{aligned} \tilde{\varepsilon}' \quad = \quad & \inf\{|\mathring{g}_{l'} - \mathring{g}_l| : l' \in \{l_1, \ldots, l_{m'}\}, \\ & l \in \{1, \ldots, m\}, \exists x_1 \in \mathcal{X}^{(l)}, \\ & \exists x_2 \in \mathcal{X}^{(l')}, \text{ such that } x_1 \sim x_2\}, \end{aligned} \tag{A.32}$$

and note that $\tilde{\varepsilon}'$ is always positive.

Lemma 4.3. *Consider an arbitrary real valued function $g(x) \in l_2^w$ on a non-decomposable infinite countable set $\mathcal{X}$ with the preorder $\preceq$ defined on it. Suppose that $\tilde{\varepsilon}'$ is defined in (A.32). If for some $\mathring{g}(x) \in \mathcal{F}^{is}$ we have*

$$\sup_{x \in \mathcal{X}} \{|g(x) - \mathring{g}(x)|\} < \tilde{\varepsilon}'/2,$$

then the isotonic regression of $g(x)$ is given by

$$g^*(x) = g^{*(l'')}(x), \text{ whenever } x \in \mathcal{X}^{(l'')}, \text{ for } l'' \in \{l_1, \ldots, l_{m'}, l_{m'+1}\},$$

where $g^{*(l'')}(x)$ is the isotonic regression of the function $g^{(l')}(x)$ over the set $\mathcal{X}^{(l')}$ with respect to the preorder $\preceq$. Therefore, the function $g^*(x)$ is a concatenation of the isotonic regressions of the restrictions of $g(x)$ to the sets $\mathcal{X}^{(l_1)}, \mathcal{X}^{(l_2)}, \ldots, \mathcal{X}^{(l_{m'})}$ and $\mathcal{X}^{(l_{m'+1})}$.

Next we state and prove an auxiliary lemma, see also Problem III.6.3 in [21], which will be used in the final theorem.

Lemma 4.4. *Let $\boldsymbol{Z}_n$, for $n = 1, \ldots, \infty$, be a tight sequence of random vectors in $\boldsymbol{l}_2^w$, endowed with its Borel σ-algebra $\mathcal{B}$. Consider the set of indices $\mathcal{I} = \{1, 2, \ldots, \infty\}$ of the components of the vectors $\boldsymbol{Z}_n$. Assume that for some random vector $\boldsymbol{Z}$ in $(\boldsymbol{l}_2^w, \mathcal{B})$ and some rearrangement $\tilde{\mathcal{I}}$ of the original index set $\mathcal{I}$ the following holds: For any positive finite integer s we have $\tilde{\boldsymbol{Z}}_n^{(1,s)} \xrightarrow{d} \tilde{\boldsymbol{Z}}^{(1,s)}$, where $\tilde{\boldsymbol{Z}}_n^{(1,s)}$ and $\tilde{\boldsymbol{Z}}^{(1,s)}$ are vectors in $\mathbb{R}^s$ constructed from the elements of the vectors $\boldsymbol{Z}_n$ and $\boldsymbol{Z}$ in such a way that the j-th elements of $\tilde{\boldsymbol{Z}}_n^{(1,s)}$ and $\tilde{\boldsymbol{Z}}^{(1,s)}$ are equal to the $\tilde{i}_j$-th elements of the vectors $\boldsymbol{Z}_n$ and $\boldsymbol{Z}$, respectively, with $\tilde{i}_j$ being the j-th index from the rearranged index set $\tilde{\mathcal{I}}$. In addition, assume that any cylinder set in $\boldsymbol{l}_2^w$ is a continuity set for the law of $\tilde{\boldsymbol{Z}}^{(1,s)}$. Then $\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z}$.*

Finally, the next theorem gives the limiting distribution of $\hat{\boldsymbol{g}}_n^*$. Similarly to the finite case we introduce the operator $\varphi : \boldsymbol{l}_2^w \to \boldsymbol{l}_2^w$, defined in the following

way. For any vector $\boldsymbol{\theta} \in \boldsymbol{l}_2^w$ we consider the coordinate evaluation map $\theta(x) : \mathcal{X} \to \mathbb{R}$ defined as $\theta(x_i) = \theta_i$, for $i = 1, \ldots, \infty$. Then, let

$$\varphi(\boldsymbol{\theta})_i = \theta^{*(v',l')}(x_i), \tag{A.33}$$

where $\theta^{*(v',l')}(x)$ is the isotonic regression of the restriction of $\theta(x)$ to the set $\mathcal{X}^{(v',l')}$ in the partition of $\mathcal{X}$. The indices v' and l' are such that $x_i \in \mathcal{X}^{(v',l')}$. The restriction of $\varphi(\boldsymbol{\theta})$ to the comparable index level set $\mathcal{I}^{(v,l)}$ will be denoted by $[\varphi(\boldsymbol{\theta})]^{(v,l)}$

Theorem 4.5. *Suppose Assumption 2.2 holds. Then the asymptotic distribution of the isotonized estimator $\hat{\boldsymbol{g}}_n^*$ is given by*

$$B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi(\boldsymbol{\lambda}), \tag{A.34}$$

where φ is the operator defined in (A.33).

Proof. Let us consider the partition of the original set $\mathcal{X} = \cup_{l=1}^m \mathcal{X}^{(l)}$ made for the function $\mathring{g}(x)$. As it was shown above, the cardinality $|\mathcal{X}^{(l)}|$ of each comparable level set in the partition must be less than infinity, unless $\mathring{g}_l = 0$, in which case it can have infinite cardinality. Note that if the number of terms in the partition is less than infinity, i.e. $m < \infty$, then some terms (or just one) in the partition are such that the function $\mathring{g}(x)$ is equal to zero on them, i.e. $\mathring{g}_l = 0$. Therefore, in this case we can use the same approach as in the case of the finite set $\mathcal{X}$ (Lemma 3.5), because in this case the smallest comparable level distance $\tilde{\varepsilon}$, defined in (A.8), is greater than zero.

Therefore, further in the proof we assume that $m = \infty$ and write the partition as $\mathcal{X} = \cup_{l=1}^{\infty} \mathcal{X}^{(l)}$. First, for any positive integer $m' < \infty$ let us take m' terms from the partition of $\mathcal{X}$ which satisfy $|\mathring{g}_{l_1}| \geq |\mathring{g}_{l_2}| \geq \cdots \geq |\mathring{g}_{l_{m'}}|$.

Second, since the sequence $\hat{\boldsymbol{g}}_n$ is consistent, then for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}[\sup_{i\in\mathcal{I}}\{|\hat{g}_{n,i} - \mathring{g}_i|\} < \varepsilon] = 1.$$

Therefore, letting $\varepsilon = \tilde{\varepsilon}'/2$, with $\tilde{\varepsilon}'$ defined in (A.32), by Lemma 8.2 we obtain that, for the isotonic regression $\hat{\boldsymbol{g}}_n^*$ of $\hat{\boldsymbol{g}}_n$

$$\lim_{n\to\infty} \mathbb{P}[\hat{g}_i^* = \hat{g}_i^{*(l'')}] = 1, \text{ whenever } i \in \mathcal{I}^{(l'')}, \quad \text{for } l'' \in \{l_1, \ldots, l_{m'+1}\}, \tag{A.35}$$

where $\mathcal{I}^{(l')}$, for $l' \in \{l_1, \ldots, l_{m'}\}$, are the comparable level sets and $\mathcal{I}^{(m'+1)}$ is the index set of $\mathcal{X}^{(m'+1)} = \mathcal{X} \setminus \mathcal{X}^{(l_1)} \cup \mathcal{X}^{(l_2)} \cup \cdots \cup \mathcal{X}^{(l_{m'})}$.

Third, let us introduce a linear operator $A^{(m')} : \boldsymbol{l}_2^{\boldsymbol{w}} \to \mathbb{R}^s$, with $s = \sum_{l \in \{l_1, \ldots l_{m'}\}} |\mathcal{X}^{(l)}|$, such that for any $\boldsymbol{g} \in \boldsymbol{l}_2^{\boldsymbol{w}}$ the first $|\mathcal{X}^{(l_1)}|$ elements of the vector $A^{(m')}\boldsymbol{g}$ are equal to ones taken from $\boldsymbol{g}$ whose indices are in $\mathcal{I}^{(l_1)}$, the second $|\mathcal{X}^{(l_2)}|$ elements are the ones from $\boldsymbol{g}$ whose indices are from $\mathcal{I}^{(l_2)}$ and so on. Therefore, using the result in (A.35), the definition of B_n and statement (vi) of Lemma 3.1, we have that

$$\lim_{n \to \infty} \mathbb{P}[A^{(m')}\varphi(B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}})) = A^{(m')}B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}})] = 1. \tag{A.36}$$

Next, since φ is a continuous map, which follows from statement (iii) of Lemma 4.2, and $A^{(m')}$ is a linear operator, then from the continuous mapping theorem it follows that

$$A^{(m')}\varphi(B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}})) \xrightarrow{d} A^{(m')}\varphi(\boldsymbol{\lambda}). \tag{A.37}$$

and, using Lemma 3.6 and result (A.36), we prove

$$A^{(m')}B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}}) \xrightarrow{d} A^{(m')}\varphi(\boldsymbol{\lambda}).$$

Note, that the number m' is an arbitrary finite integer. Also, since φ is a continuous map, then the law of $\varphi(\boldsymbol{\lambda})$ has the same continuity sets as $\boldsymbol{\lambda}$. Moreover, the sequence $B_n(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}})$ is tight, because $B_n(\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}})$ has a limit in distribution and $||\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}})||_2 \leq ||\hat{\boldsymbol{g}}_n - \mathring{\boldsymbol{g}})||_2$. Using Lemma 4.4 we finish the proof of the theorem. □

Recall that the cardinality of any comparable level set $\mathcal{X}^{(v,l)}$ is less than infinity whenever $\mathring{g}_{v,l} \neq 0$. Then, as in the finite case, we note that the order constraints on $\mathcal{X}^{(v,l)}$ can be expressed in the form $\boldsymbol{A}\boldsymbol{g} \geq \boldsymbol{0}$, for some matrix $\boldsymbol{A}$. Therefore, one can use the results in [2] to describe the behaviour of $[\varphi(\boldsymbol{g})]^{(v,l)}$ when $|\mathcal{X}^{(v,l)}| < \infty$. It follows from Theorem 5 in [2] that the distribution of $[\varphi(\boldsymbol{g})]^{(v,l)}$ is a mixture of $2^{|\mathcal{X}^{(v,l)}|}$ distributions of the projections of $\boldsymbol{g}$ onto the cone $\boldsymbol{A}_t\boldsymbol{g} \geq \boldsymbol{0}$, where the matrixes $\boldsymbol{A}_t$, for $t = 1, \ldots, 2^{|\mathcal{X}^{(v,l)}|}$, are comprised of the rows of the matrix $\boldsymbol{A}$.

Next, let us consider the case of non-constant weights $\boldsymbol{w}$. In this section until now we assumed that the vector of weights satisfies the condition in (A.3), it

is fixed, $\boldsymbol{w}_n = \boldsymbol{w}$, so it does not depend on n, and the random elements $\hat{\boldsymbol{g}}_n$ in Assumption 2.2 all take their values in $(\boldsymbol{l}_2^{\boldsymbol{w}}, \mathcal{B})$, for some fixed $\boldsymbol{w}$, with $\mathcal{B}$ the Borel σ-algebra generated by the topology which is generated by the natural norm of $\boldsymbol{l}_2^{\boldsymbol{w}}$.

Now we consider some non-random sequence $\{\boldsymbol{w}_n\}_{n\geq 1}$, taking values in the space $\mathbb{R}^\infty$, where each $\boldsymbol{w}_n$ satisfies the condition in (A.3). The sequence $\{\boldsymbol{w}_n\}_{n\geq 1}$ converges in some norm $\|\cdot\|_R$ on $\mathbb{R}^\infty$ to some non-random vector $\boldsymbol{w}$, which also satisfies the condition in (A.3). Next, let $\mathcal{B}_n$ denotes the Borel σ-algebra generated by the topology which is generated by the natural norm in $\boldsymbol{l}_2^{\boldsymbol{w}_n}$. The next lemma shows that the normed spaces $\boldsymbol{l}_2^{\boldsymbol{w}_n}$ are all equivalent.

Lemma 4.6. *Suppose that two vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ satisfy the condition in (A.3). Then the normed spaces $\boldsymbol{l}_2^{\boldsymbol{w}_1}$ and $\boldsymbol{l}_2^{\boldsymbol{w}_2}$ are equivalent.*

Therefore, since the normed spaces $\boldsymbol{l}_2^{\boldsymbol{w}_n}$ are all equivalent, then the topologies generated by these norms are the same. Then, the Borel σ-algebras $\mathcal{B}_n$ generated by these topologies are also the same. Therefore, the measurable spaces $(\boldsymbol{l}_2^{\boldsymbol{w}_n}, \mathcal{B}_n)$ are all the same and we will suppress the index n.

Next, analogously to the finite case, let us introduce the notation $\varphi^{\boldsymbol{w}}(\boldsymbol{\theta})$

$$\varphi^{\boldsymbol{w}}(\boldsymbol{\theta})_i = \theta^{*\boldsymbol{w}(v',l')}(x_i), \tag{A.38}$$

where $\theta^{*\boldsymbol{w}(v',l')}(x)$ is the isotonic regression with weights $\boldsymbol{w}$ of the restriction of $\theta(x)$ to the comparable level set $\mathcal{X}^{(v',l')}$ of $\mathring{g}(x)$, where the indices v' and l' are such that $x_i \in \mathcal{X}^{(v',l')}$. The next theorem gives the limiting distribution of $\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n}$.

Theorem 4.7. *Suppose the Assumption 2.2 holds. Then the asymptotic distribution of the isotonic regression $\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n}$ of the basic estimator $\hat{\boldsymbol{g}}_n$ is given by*

$$B_n(\hat{\boldsymbol{g}}_n^{*\boldsymbol{w}_n} - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi^{\boldsymbol{w}}(\boldsymbol{\lambda}), \tag{A.39}$$

where φ is the operator, defined in (A.38).

Proof. First, we note that the result of Lemma 4.4 holds, it we assume that the random vectors $\boldsymbol{Z}_n$, for $n = 1, \ldots, \infty$ take their values in $\boldsymbol{l}_2^{\boldsymbol{w}_n}$, if all elements of $\boldsymbol{w}_n$ and its limit $\boldsymbol{w}$ satisfy the condition in (A.3): This follows from the fact that the measurable spaces $(\boldsymbol{l}_2^{\boldsymbol{w}_n}, \mathcal{B}_n)$ are equivalent, which was proved in Lemma 4.6.

The rest of the proof is exactly the same as for Theorem 4.5 with φ and $\hat{g}_n^*$ suitable changed to $\varphi^{\boldsymbol{w}}$ and $\hat{g}_n^{*\boldsymbol{w}_n}$. Also, recall that the result of Lemma 8.2 does not depend on the weights $\boldsymbol{w}_n$. □

5 Application to bimonotone probability mass function and regression function estimation, and extensions to d-dimensional problems

In this section we consider the problems of estimation of a bimonotone regression function, in subsection 5.1, and of a bimonotone probability mass function, in subsection 5.2. Also, we consider the generalisation to the case of d-dimensional support, in subsection 5.3.

First, let us introduce a bimonotone order relation $\preceq$ on a set $\mathcal{X} := \{\boldsymbol{x} = (i_1, i_2)^T : i_1 = 1, 2, \ldots, r_1, i_2 = 1, 2, \ldots, r_2\}$, with $r_1, r_2 \leq \infty$ in the following way. For any $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in $\mathcal{X}$ we have $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$ if and only if $x_{1,1} \leq x_{2,1}$ and $x_{1,2} \leq x_{2,2}$. The order relation $\preceq$ is a partial order, because it is reflexive, transitive, antisymmetric, but there are elements in $\mathcal{X}$ which are noncomparable.

Second, note that $\mathcal{X}$ with the order relation $\preceq$ defined above is non-decomposable, because for any $\boldsymbol{x}_1 = (x_{1,1}, x_{1,2})$ and $\boldsymbol{x}_2 = (x_{2,1}, x_{2,2})$ in $\mathcal{X}$ there exist $\boldsymbol{x}_3 = (x_{3,1}, x_{3,2}) \in \mathcal{X}$ such that $x_{3,1} \geq x_{1,1}$, $x_{3,1} \geq x_{2,1}$ and $x_{3,2} \geq x_{1,2}$, $x_{3,2} \geq x_{2,2}$, which means that $\boldsymbol{x}_1 \sim \boldsymbol{x}_3$ and $\boldsymbol{x}_2 \sim \boldsymbol{x}_3$, or $\boldsymbol{x}_4 = (x_{4,1}, x_{4,2}) \in \mathcal{X}$ such that $x_{4,1} \leq x_{1,1}$, $x_{4,1} \leq x_{2,1}$ and $x_{4,2} \geq x_{1,2}$, $x_{4,2} \geq x_{2,2}$, which means that $\boldsymbol{x}_1 \sim \boldsymbol{x}_4$ and $\boldsymbol{x}_2 \sim \boldsymbol{x}_4$. Therefore, in a partition (A.5) $k = 1$. Also, following our notations above we denote by $\mathcal{I}$ the set of indices of the domain $\mathcal{X}$ and use the same notation $\preceq$ for the order relation on $\mathcal{I}$ generated by $\mathcal{X}$.

A real valued function $g(\boldsymbol{x})$ is bimonotone increasing, i.e. isotonic with respect to bimonotone order relation $\preceq$ on a set $\mathcal{X}$, if whenever $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$ one has $g(\boldsymbol{x}_1) \leq g(\boldsymbol{x}_2)$, cf. [7]. A real valued function $h(\boldsymbol{x})$ is called a bimonotone decreasing function, if whenever $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$ one has $h(\boldsymbol{x}_1) \geq h(\boldsymbol{x}_2)$. In the last case the function $h(\boldsymbol{x})$ is called antitonic with respect to the order relation $\preceq$ on a set $\mathcal{X}$, cf. [3, 20]. Note that a fucntion $h(\boldsymbol{x})$ is antitonic if and only if $g(\boldsymbol{x}) = -h(\boldsymbol{x})$ is isotonic with respect to the order relation $\preceq$ on the set $\mathcal{X}$.

5.1 Estimation of a bimonotone increasing regression function

The problem of estimation of a bimonotone regression function via least squares was studied in detail in [7], where the authors described an algorithm for minimization of a smooth function under bimonotone order constraints.

Suppose we have observed $Z_i = (\boldsymbol{x}_i, Y_i)$, $i = 1, \ldots, n$, with $\boldsymbol{x}_i$ the design points taking values from the set $\mathcal{X} := \{\boldsymbol{x} = (i_1, i_2)^T : i_1 = 1, 2, \ldots, r_1, i_1 = 1, 2, \ldots, r_2, \}$, with $r_1, r_2 < \infty$ and Y_i real valued random variables defined in the regression model

$$Y_i = \mathring{g}(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i is a sequence of identically distributed random variables with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2 < \infty$.

The least squares estimate of $\mathring{g}(\boldsymbol{x})$ under bimonotone constraints is given by

$$g_n^* = \underset{f \in \mathcal{F}^{is}}{\operatorname{argmin}} \sum_{\boldsymbol{x} \in \mathcal{X}} (f(\boldsymbol{x}) - \hat{g}_n(\boldsymbol{x}))^2 w_{\boldsymbol{x}}^{(n)}, \tag{A.40}$$

where $\mathcal{F}^{is}$ denotes the set of all bounded bimonotone increasing functions on $\mathcal{X}$, $\hat{g}_n(\boldsymbol{x})$ is the average of Y_i, $i = 1, \ldots, n$, over the design element $\boldsymbol{x}$, i.e.

$$\hat{g}_n(\boldsymbol{x}) = \frac{\sum_{i=1}^n Y_i 1\{\boldsymbol{x}_i = \boldsymbol{x}\}}{\sum_{i=1}^n 1\{\boldsymbol{x}_i = \boldsymbol{x}\}} \tag{A.41}$$

and

$$w_{\boldsymbol{x}}^{(n)} = \frac{\sum_{i=1}^n 1\{\boldsymbol{x}_i = \boldsymbol{x}\}}{n}. \tag{A.42}$$

Note that $g_n(\boldsymbol{x})$ in (A.41) is the unconstrained least squares estimate of $\mathring{g}(\boldsymbol{x})$. The asymptotic properties of nonlinear least squares estimators were studied in [17, 25]. Assume that the design points $\boldsymbol{x}_i$, with $i = 1, \ldots, n$, satisfy the following condition

$$\boldsymbol{w}^{(n)} \to \boldsymbol{w}, \tag{A.43}$$

as $n \to \infty$, where $\boldsymbol{w}^{(n)}$ is a sequence of vectors in $\mathbb{R}_+^{r_1 \times r_2}$ whose components are from (A.42), and $\boldsymbol{w} \in \mathbb{R}_+^{r_1 \times r_2}$. Given the condition in (A.43) is satisfied, the basic estimator $\hat{g}_n(\boldsymbol{x})$ is consistent and has the following asymptotic distribution

$$n^{1/2}(\hat{\boldsymbol{g}}_n - \boldsymbol{g}) \xrightarrow{d} \boldsymbol{Y}_{\boldsymbol{0},\Sigma}, \tag{A.44}$$

where $\boldsymbol{Y}_{\mathbf{0},\Sigma}$ is a Gausian vector with mean zero and diagonal covariance matrix Σ, whose elements are given by $\Sigma_{ii} = \sigma^2 w_i$, for $i = 1, \ldots, r \times s$, cf. Theorem 5 in [25].

We next derive the asymptotic distribution of the regression function under bimonotone constraints.

Theorem 5.1. *Given that the condition (A.43) on the design points is satisfied, the asymptotic distribution of the regression function $\hat{\boldsymbol{g}}_n^*(\boldsymbol{x})$ under bimonotone constraints is given by*

$$n^{1/2}(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi^{\boldsymbol{w}}(\boldsymbol{Y}_{\mathbf{0},\Sigma}), \tag{A.45}$$

where $\varphi^{\boldsymbol{w}}$ is the operator defined in (A.20) and $\boldsymbol{Y}_{\mathbf{0},\Sigma}$ is a Gaussian vector defined in (A.44).

Proof. The requirements of Assumption 2.1 are satisfied. Therefore the result follows from Theorem 4.5. □

5.2 Estimation of a bimonotone decreasing probability mass function

In this subsection we treat the problem of estimating a bimonotone *decreasing* probability mass function on $\mathbb{Z}_2^+$. Note that this is a natural order restriction on the pmfs defined on $\mathbb{Z}_2^+$, since a positive bimonotone increasing function on $\mathbb{Z}_2^+$ does not belong to l_2.

Suppose that we have observed $Z_1, Z_2, \ldots, Z_n$ i.i.d. random variables taking values in $\mathcal{X} = \mathbb{Z}_2^+ := \{(i_1, i_2)^T : i_1 = 1, 2, \ldots, \infty, i_2 = 1, 2, \ldots, \infty, \}$ with probability mass function $\boldsymbol{p}$. The empirical estimator of $\boldsymbol{p}$ is then given by

$$\hat{p}_{n,\boldsymbol{i}} = \frac{n_{\boldsymbol{i}}}{n}, \quad n_{\boldsymbol{i}} = \sum_{j=1}^{n} 1\{Z_j = \boldsymbol{x}_{\boldsymbol{i}}\}, \quad \boldsymbol{i} \in \mathcal{I}, \tag{A.46}$$

and it is also the unrestricted mle, which generally does not satisfy the bimonotonisity constraints introduced above. However, $\hat{\boldsymbol{p}}_n$ is consistent, i.e. $\hat{\boldsymbol{p}}_n \xrightarrow{p} \boldsymbol{p}$ and asymptotically Gaussian

$$n^{1/2}(\hat{\boldsymbol{p}}_n - \boldsymbol{p}) \xrightarrow{d} \boldsymbol{Y}_{\mathbf{0},C}, \tag{A.47}$$

where $\boldsymbol{Y}_{\mathbf{0},C}$ is a Gaussian process in l_2, with mean zero and the covariance operator C such that $\langle C\boldsymbol{e_i}, \boldsymbol{e_{i'}}\rangle = p_i\delta_{i,i'} - p_i p_{i'}$, with $\boldsymbol{e_i} \in \boldsymbol{l}_2$ the orthonormal basis in $\boldsymbol{l}_2$ such that in a vector $\boldsymbol{e_i}$ all elements are equal to zero but the one with the index $\boldsymbol{i}$ is equal to 1, and $\delta_{ij} = 1$, if $i = j$ and 0 otherwise, cf. [16].

The constrained mle $\hat{\boldsymbol{p}}_n^*$ of $\boldsymbol{p}$ is then given by the isotonic regression of the empirical estimator $\hat{\boldsymbol{p}}_n$ over the set $\mathcal{X}$ with respect to the preorder $\preceq$

$$\hat{\boldsymbol{p}}_n^* = \underset{\xi \in \mathcal{F}^{an}}{\operatorname{argmin}} \sum_{\boldsymbol{x} \in \mathcal{X}} (\xi_{\boldsymbol{x}} - \hat{p}_{n,\boldsymbol{x}})^2, \tag{A.48}$$

where $\mathcal{F}^{an}$ denotes the set of all bimonotone decreasing (antitonic with respect to $\preceq$) functions on $\mathcal{X}$. This result shown on pages 45–46 in [3] and pages 38–39 in [20].

Next we make the following substitution

$$\begin{aligned} \boldsymbol{\theta} &= -\boldsymbol{p}, \\ \hat{\boldsymbol{\theta}}_n &= -\hat{\boldsymbol{p}}_n, \\ \hat{\boldsymbol{\theta}}_n^* &= -\hat{\boldsymbol{p}}_n^*. \end{aligned} \tag{A.49}$$

Therefore $\hat{\boldsymbol{\theta}}_n^*$ is the isotonic regression of $\hat{\boldsymbol{\theta}}_n$, i.e.

$$\hat{\boldsymbol{\theta}}_n^* = \underset{\xi \in \mathcal{F}^{is}}{\operatorname{argmin}} \sum_{\boldsymbol{x} \in \mathcal{X}} (\xi_{\boldsymbol{x}} - \hat{\theta}_{n,\boldsymbol{x}})^2, \tag{A.50}$$

where $\mathcal{F}^{is}$ denotes the set of all bimonotone increasing (isotonic with respect to $\preceq$) functions on $\mathcal{X}$.

We next derive the asymptotic distribution of the bimonotone mle $\hat{\boldsymbol{p}}_n^*$ as a corollary of Theorem 4.5.

Theorem 5.2. *The asymptotic distribution of the constrained mle $\hat{\boldsymbol{p}}_n^*$ of a bimonotone probability mass function $\boldsymbol{p}$ is given by*

$$n^{1/2}(\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}) \xrightarrow{d} \varphi(\boldsymbol{Y}_{\mathbf{0},C}), \tag{A.51}$$

where φ is the operator defined in (A.33) and $\boldsymbol{Y}_{\mathbf{0},C}$ is a Gaussian process in l_2 defined in (A.47).

Proof. The requirements of Assumption 2.2 for the sequence $\hat{\boldsymbol{\theta}}_n$ defined in (A.49) are satisfied. Therefore from Theorem 4.5 it follows that

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}) \xrightarrow{d} \varphi(\boldsymbol{Y}_{\mathbf{0},C})$$

and using the substitution (A.49) we finish the proof. □

5.3 Generalisation to the case of d-dimensional monotone functions

The results obtained in Theorems 5.1 and 5.2 can be directly generalised to the case of estimation of a d-dimensional monotone (d-monotone) regression function and a d-monotone pmf.

Let us consider a set

$$\mathcal{X} := \{\boldsymbol{x} = (i_1, i_2, \ldots, i_d)^T : i_1 = 1, 2, \ldots, r_1, i_2 = 1, 2, \ldots, r_2, \ldots, i_d = 1, 2, \ldots, r_d\}, \text{ with } d < \infty, r_1, r_2, \ldots, r_d \leq \infty \tag{A.52}$$

and introduce a d-monotone order relation $\preceq$ on it in the following way. For any $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in $\mathcal{X}$ we have $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$ if and only if $x_{1,1} \leq x_{2,1}, x_{1,2} \leq x_{2,2}, \ldots, x_{1,d} \leq x_{2,d}$. Similarly to the bimonotone case, it can be shown that the order relation $\preceq$ is a partial order and $\mathcal{X}$ is non-decomposable.

Suppose we have observed $Z_i = (\boldsymbol{x}_i, Y_i)$, $i = 1, \ldots, n$, with $\boldsymbol{x}_i$ the design points taking values from the set $\mathcal{X}$ defined in (A.52), with $r_1, r_2, \ldots, r_d < \infty$ and Y_i real valued random variables defined in the regression model

$$Y_i = \mathring{g}(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i is a sequence of identically distributed random variables with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2 < \infty$.

The least squares estimate of $\mathring{g}(\boldsymbol{x})$ under bimonotone constraints is given by

$$g_n^* = \underset{f \in \mathcal{F}^{is}}{\operatorname{argmin}} \sum_{\boldsymbol{x} \in \mathcal{X}} (f(\boldsymbol{x}) - \hat{g}_n(\boldsymbol{x}))^2 w_{\boldsymbol{x}}^{(n)},$$

where $\mathcal{F}^{is}$ denotes the set of all bounded d-monotone functions on $\mathcal{X}$, the expressions for $\hat{g}_n(\boldsymbol{x})$ and $w_{\boldsymbol{x}}^{(n)}$ are the same as in bimonotone case, i.e. given in (A.41) and (A.42), respectively. Therefore, under condition (A.43) on the design points $\boldsymbol{x}_i$, we obtain the following corollary.

Theorem 5.3. *The asymptotic distribution of the regression function $\hat{\boldsymbol{g}}_n^*(\boldsymbol{x})$ under d-monotone constraints is given by*

$$n^{1/2}(\hat{\boldsymbol{g}}_n^* - \mathring{\boldsymbol{g}}) \xrightarrow{d} \varphi^{\boldsymbol{w}}(\boldsymbol{Y}_{\boldsymbol{0},\Sigma}),$$

where φ^w is the operator defined in (A.20) and $\mathbf{Y}_{\mathbf{0},\Sigma}$ is a Gaussian vector defined in (A.44).

Proof. The requirements of Assumption 2.1 are satisfied. Therefore the result follows from Theorem 4.5. □

Next suppose that we have observed $Z_1, Z_2, \ldots, Z_n$ i.i.d. random variables taking values in $\mathcal{X}$ defined in (A.52), with $r_1, r_2, \ldots, r_d \leq \infty$ with probability mass function $\boldsymbol{p}$. The mle $\hat{\boldsymbol{p}}_n^*$ of $\boldsymbol{p}$ with d-monotone decreasing constraints is then given by

$$\hat{\boldsymbol{p}}_n^* = \underset{\xi \in \mathcal{F}^{an}}{\operatorname{argmin}} \sum_{\boldsymbol{x} \in \mathcal{X}} (\xi_{\boldsymbol{x}} - \hat{p}_{n,\boldsymbol{x}})^2,$$

where $\hat{\boldsymbol{p}}_n$ is the empirical estimator defined in (B.2), $\mathcal{F}^{an}$ denotes the set of all d-monotone decreasing functions on $\mathcal{X}$. The asymptotic distribution of $\hat{\boldsymbol{p}}_n^*$ is given in the following corollary.

Theorem 5.4. *The asymptotic distribution of the constrained mle $\hat{\boldsymbol{p}}_n^*$ of a d-monotone probability mass function $\boldsymbol{p}$ is given by*

$$n^{1/2}(\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}) \xrightarrow{d} \varphi(\mathbf{Y}_{\mathbf{0},C}), \tag{A.53}$$

where φ is the operator defined in (A.33) and $\mathbf{Y}_{\mathbf{0},C}$ is a Gaussian process in l_2 defined in (A.47).

Proof. Making the same substitution as in a bimonotone case, i.e. in (A.49) we note that the requirements of Assumption 2.2 are satisfied. Therefore the result follows from Theorem 4.5. □

6 Conclusions and discussion

We have derived the limit distribution of an estimator that is obtained as the l^2 projection of a basic preliminary estimator on the space of functions that are defined on a countable set, and that are monotone with respect to a preorder

on that countable set. Immediate applications that we have stated results for are to the estimation of $d-$monotone pmfs and regression functions.

We would like to emphasize a qualitative difference between the estimation of a pmf over a subset of $\mathbb{Z}_+^d$ and the estimation of a pdf over a subset of $\mathbb{R}_+^d$. We note first that limit distribution results for monotone pdf estimators, to our knowledge, exist only for the case $d = 1$, cf. however [18] for the limit distribution of the non-parametric maximum likelihood estimator (npmle) of a bimonotone pdf (so when $d = 2$), indexed by (the Lebesgue measure of) lower layers. For the case $d = 1$, the isotonic regression estimator of a pdf is, for independent data, equivalent to the npmle, i.e. the Grenander estimator, and for dependent data does not have the interpretation of an npmle, cf. [1] for the limit distribution results for the monotone restricted pdf estimator for arbitrary dependence assumptions on the data. The limit distribution in the independent data case is then the well known Chernoff distribution mentioned above, and for dependent data different, cf. Theorem 10 (ii) and Theorem 11 in [1].

Note also that, in the case $d = 1$, the order restricted estimator of a pdf is a *local* estimator, in the sense that it uses data in a shrinking neighbourhood around the point of interest, say $t_0 \in \mathbb{R}$, to calculate the value of the pdf at t_0, and the size of the neighbourhood is of the order $n^{-1/3}$ for independent data, and of a different order for dependent data, cf. Table 1 of Section 5 in [1] for an overview of the possible orders related to the dependence of the data. Any sensible estimator of a monotone pdf for $d \geq 2$ will also use data in a shrinking neighbourhood around the point of interest, cf. e.g. [15] for a discussion about rates in this connection. Furthermore, as argued e.g. in [15], the rates in higher dimensions are slower for monotone pdf estimation. This is in sharp contrast to the problems treated in this paper, on monotone pmf estimation, and is explained by the fact that the resulting estimator for those problems is a *global* estimator, i.e. it uses data points in a set of size $O(1)$ around the point of interest to obtain the estimator, irrespective of the dimension d. The fact that estimators of pdf are local and of pmf are global, also accounts for that one is able to obtain process limit distribution results for the pmf estimator, whereas it is only possible to obtain pointwise limit distribution results for the pdf estimators.

The results stated in this paper are general in terms of the demands on the basic estimator and on the underlying empirical process. In fact, Assumptions 2.1 and 2.2 only require that there is a limit process for the basic estimator, and

do not specify any requirements of e.g. dependence for the data. Note, however, that if one does not require independence of the data, then the identity between the isotonic regression of a pmf and the mle of a pmf vanishes, since the product of the marginal pmfs is then not the full likelihood.

By allowing dependent data, we are in a position to straight-forwardly obtain limit distributions in general situations. One problem that comes to mind is that of isotonic regression of an ordered pmf on a DAG. The assumption of monotonicity of the pmf with respect to the natural tree order on the DAG is sensible; one can e.g. imagine the DAG describing the, say, three categories that may influence the monthly salary of an employee at a large facility, with the DAG structure given by the (matrix) preorder on the three categories. Then, given data on employees salary and covariate readings for the three categories, one may first construct the empirical estimate of the pmf and next isotonize that. Knowing the limit distribution of the empirical estimator immediately gives us the limit distribution of the isotonized estimator, irrespective of whether data are independent or not.

7 Appendix

Proof of Lemma 3.1. The statements (i), (ii), (iii) and (iv) are from [20] (Theorems 1.3.1, 1.3.3, 1.4.4 and 1.3.4). The statements (v) and (vi) are proved in [3] (Theorems 2.2 and 1.8).

Note that statement (ii) means that if the basic estimator $\hat{\boldsymbol{g}}_n$ satisfies a linear restriction, e.g. $\sum_{i=1}^{s} w_i \hat{g}_{n,i} = c$, with some positive reals w_i, then the same holds for its isotonic regression with the weights $\boldsymbol{w}$, i.e. for $\hat{\boldsymbol{g}}_n^*$ one has $\sum_{i=1}^{s} w_i \hat{g}_{n,i}^* = c$. □

Proof of Lemma 3.2. Let $g(x)$ be an arbitrary real-valued function defined on

$\mathcal{X}$. From the definition of the isotonic regression

$$\begin{aligned} g^* &= \operatorname*{argmin}_{f\in\mathcal{F}^{is}} \sum_{x\in\mathcal{X}} (f(x)-g(x))^2 w_x \\ &= \operatorname*{argmin}_{f\in\mathcal{F}^{is}} \sum_{v=1}^{k} \sum_{x\in\mathcal{X}^{(v)}} (f(x)-g(x))^2 w_x \\ &= \sum_{v=1}^{k} \operatorname*{argmin}_{f^{(v)}\in\mathcal{F}^{is}_{(v)}} \sum_{x\in\mathcal{X}^{(v)}} (f^{(v)}(x)-g^{(v)}(x))^2 w_x, \end{aligned}$$

where $f^{(v)}$ is the restriction of the function $f : \mathcal{X} \to \mathbb{R}$ to the set $\mathcal{X}^{(v)}$. The second equality follows from $(A.5)$ and the last equality follows from the fact that since the elements from the different partition sets $\mathcal{X}^{(v)}$ are non-comparable, then any function $f \in \mathcal{F}^{is}$ can be written as a concatenation of $f^{(v)} \in \mathcal{F}^{is}_{(v)}$, with no restrictions imposed on the values of $f^{(v_1)}$ and $f^{(v_2)}$ for $v_1 \neq v_2$. □

Proof of Lemma 3.5. First, note that if the condition of the lemma is satisfied, then the function $g^*(x)$ defined in (A.10) on the set $\mathcal{X}$ is isotonic. This follows from Lemma 3.1, statement (iv). Second, assume that the function $g^*(x)$ defined in (A.10) is not an isotonic regression of $g(x)$. This means that there exists another function $\tilde{g}(x)$, such that

$$\sum_{x\in\mathcal{X}} (\tilde{g}(x)-g(x))^2 w_x < \sum_{x\in\mathcal{X}} (g^*(x)-g(x))^2 w_x, \tag{A.54}$$

Using the partition of $\mathcal{X}$, (A.54) can be rewritten as

$$\sum_{l=1}^{m} \sum_{x\in\mathcal{X}^{(l)}} (\tilde{g}(x)-g(x))^2 w_x < \sum_{l=1}^{m} \sum_{x\in\mathcal{X}^{(l)}} (g^*(x)-g(x))^2 w_x.$$

Therefore, for some l' we must have

$$\sum_{x\in\mathcal{X}^{(l')}} (\tilde{g}(x)-g(x))^2 w_x < \sum_{x\in\mathcal{X}^{(l')}} (g^*(x)-g(x))^2 w_x$$

or, equivalently,

$$\sum_{x\in\mathcal{X}^{(l')}} (\tilde{g}^{(l')}(x)-g^{(l')}(x))^2 w_x < \sum_{x\in\mathcal{X}^{(l')}} (g^{*(l')}(x)-g^{(l')}(x))^2 w_x,$$

with $g^{(l')}(x)$, $\tilde{g}^{(l')}(x)$ and $g^{*(l')}(x)$ the restrictions to the comparable level set $\mathcal{X}^{(l')}$ of $g(x)$, $\tilde{g}(x)$ and $g^*(x)$, respectively. Since the function $g^{*(l')}(x)$ is the isotonic regression of the function $g^{(l')}(x)$ on the set $\mathcal{X}^{(l')}$, the last inequality contradicts the property of the uniqueness and existence of the isotonic regression $g^{*(l')}(x)$ (statement (i) of Lemma 3.1). □

Proof of Lemma 3.6. This result follows from Theorem 3.1 in [6]. □

Proof of Lemma 4.2. Statements (i), (ii) and (iii) follow from Theorem 8.2.1, Corollary B of Theorem 8.2.7 and Theorem 8.2.5, respectively, in [20], statements (iv), (v) and (vi) follow from Corollary B of Theorem 7.9, Theorems 2.2 and Theorems 7.5 and 7.8, respectively, in [3]. □

Proof of Lemma 4.2. The proof is exactly the same as in the finite case (Lemma 3.2).

□

Proof of Lemma 8.2. The proof is exactly the same as in the case of a finite support (Lemma 3.5). □

Proof of Lemma 4.4. This proof is reminiscent of the proofs done for the spaces $\mathbb{R}^\infty$ and $\mathbb{C}(0,1)$, the space of continuous functions on unit interval with uniform topology, cf. Chapters 2 and 3 in [6].

The space $\boldsymbol{l}_2^w$ is separable and complete. Then, from Prokhorov's theorem [21], it follows that the sequence $\boldsymbol{Z}_n$ is relatively compact, which means that every sequence from $\boldsymbol{Z}_n$ contains a subsequence, which converges weakly to some vector $\boldsymbol{Z}$. If the limits of the convergent subsequences are the same, then the result of the lemma holds.

Since the space $\boldsymbol{l}_2^w$ is separable, the Borel σ-algebra equals the σ-algebra gener-

ated by open balls in $\boldsymbol{l}_2^{\boldsymbol{w}}$ [8]. Therefore, it is enough to show that the limit laws agree on open balls, since finite intersections of open balls in $\boldsymbol{l}_2^{\boldsymbol{w}}$ constitute a π-system. To show that the limit laws agree on finite intersections of open balls, we note that the open ball $B(z,\varepsilon)$ in $\boldsymbol{l}_2^{\boldsymbol{w}}$ can be written as

$$B(z,\varepsilon) \quad = \quad \cap_{M\geq 1} B_M,$$

where

$$\begin{aligned} B_M &= \cup_{n\geq 1} A_n^M, \\ A_n^M &= \{y \in \boldsymbol{l}_2^{\boldsymbol{w}} : \sum_{j\in \tilde{i}_1,\ldots,\tilde{i}_M} |z_j - y_j|^2 w_j < \varepsilon^2 - \frac{1}{n}\}, \end{aligned}$$

where the indices $\tilde{i}_1,\ldots,\tilde{i}_M$ are the M first indices from $\tilde{\mathcal{I}}$.

The sequence of vectors $\tilde{\boldsymbol{Z}}_n^{(1,M)}$ converges weakly to $\tilde{\boldsymbol{Z}}^{(1,M)}$ for all finite M, therefore any subsequence of $\tilde{\boldsymbol{Z}}_n^{(1,M)}$ converges weakly to $\tilde{\boldsymbol{Z}}^{(1,M)}$. That means that, with $\mathbb{P}_n^M$ the laws of an arbitrary but fixed subsequence of $\tilde{\boldsymbol{Z}}_n^{(1,M)}$, and $\mathbb{P}^M$ the law of $\tilde{\boldsymbol{Z}}^{(1,M)}$, $\mathbb{P}_n^{(M)}(A) \to \mathbb{P}^{(M)}(A)$ for any $\mathbb{P}^{(M)}$-continuity set A. Therefore, since the cylinder set A_n^M is a continuity set for the limit law $P^{(M)}$, and by the continuity properties of a probability measure, we obtain

$$\begin{aligned} \mathbb{P}(B(z,\varepsilon)) &= \lim_{M\to\infty} \mathbb{P}(B_M) \\ &= \lim_{M\to\infty} \lim_{n\to\infty} \mathbb{P}(A_n^M) \\ &= \lim_{M\to\infty} \lim_{n\to\infty} \mathbb{P}^{(M)}(A_n^M), \end{aligned}$$

where $\mathbb{P}$ is the law of $\boldsymbol{Z}$.

Thus, we have shown that the limit laws, $\mathbb{P}$, of the convergent subsequences of $\{\boldsymbol{Z}_n\}$ agree on the open balls $B(z,\varepsilon)$, and, therefore, also on the finite intersections of these open balls. Since the laws agree on the π-system (they are all equal to $\mathbb{P}$), they agree on the Borel σ-algebra. □

Proof of Lemma 4.6. First, we prove that if $\boldsymbol{w}$ satisfies the condition in (A.3), then $\boldsymbol{x} \in \boldsymbol{l}_2^{\boldsymbol{w}}$ if and only if $\boldsymbol{x} \in \boldsymbol{l}_2$ ($\boldsymbol{l}_2$ is the space of all square summable sequences, i.e. $\boldsymbol{w} = \{1,1,\ldots\}$). Let $\boldsymbol{x} \in \boldsymbol{l}_2^{\boldsymbol{w}}$, then $\sum_{i=1}^{\infty} x_i^2 w_i < \infty$ and we have

$$(\inf_i \{w_i\}) \sum_{i=1}^{\infty} x_i^2 \leq \sum_{i=1}^{\infty} x_i^2 w_i < \infty.$$

Therefore, since $\inf_i\{w_i\} > 0$, we have that $\sum_{i=1}^{\infty} x_i^2 < \infty$, which means that $\boldsymbol{x} \in \boldsymbol{l}_2$.

Next, let $\boldsymbol{x} \in \boldsymbol{l}_2$, then $\sum_{i=1}^{\infty} x_i^2 < \infty$ and we have

$$\sum_{i=1}^{\infty} x_i^2 w_i \leq (\sup_i\{w_i\}) \sum_{i=1}^{\infty} x_i^2 < \infty,$$

since $\sup_i\{w_i\} < \infty$. Therefore, $\boldsymbol{x} \in \boldsymbol{l}_2^{\boldsymbol{w}}$.

Second, let $\|\cdot\|_{\boldsymbol{w}}$ and $\|\cdot\|$ denote the natural norms in $\boldsymbol{l}_2^{\boldsymbol{w}}$ and $\boldsymbol{l}_2$. We can prove that if $\boldsymbol{w}$ satisfies the condition in (A.3), then $\boldsymbol{l}_2^{\boldsymbol{w}}$ and $\boldsymbol{l}_2$ are equivalent, i.e. there exist two positive constants c_1 and c_2 such that

$$c_1\|\boldsymbol{x}\| \leq \|\boldsymbol{x}\|_{\boldsymbol{w}} \leq c_2\|\boldsymbol{x}\|, \tag{A.55}$$

if, for example, $c_1 = \inf_i\{w_i\}$ and $c_2 = \sup_i\{w_i\}$. Therefore, since the equivalence of norms is transitive, then $\boldsymbol{l}_2^{\boldsymbol{w_1}}$ and $\boldsymbol{l}_2^{\boldsymbol{w_2}}$ are equivalent, provided $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ satisfy the condition in (A.3). □

Acknowledgements

VP's research is fully supported and DA's research is partially supported by the Swedish Research Council, whose support is gratefully acknowledged.

8 Bibliography

[1] Anevski, D. and Hössjer O. (2006). A general asymptotic scheme for inference under order restrictions. *The Annals of Statistics* **34**, 1874–1930.

[2] Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica* **67**, 1341–1383.

[3] Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical inference under order restrictions* John Wiley & Sons, London-New York-Sydney.

[4] Balabdaoui, F., Durot, C., Koladjo, F. (2014). On asymptotics of the discrete convex LSE of a pmf. *Bernoulli* **23**, 1449–1480.

[5] Balabdaoui, F. and Jankowski, H. (2016). Maximum likelihood estimation of a unimodal probability mass function. *Statistica Sinica* **26**, 1061–1086.

[6] Billingsley, P. (2013). *Convergence of probability measures.* John Wiley&Sons.

[7] Beran, R. and Dümbgen, L. (2010). Least squares and shrinkage estimation under bimonotonicity constraints. *Statistics and Computing* **20**, 177–189.

[8] Bogachev, V. I. (2007). *Measure theory*. Vol. I. Springer-Verlag, Berlin.

[9] Brunk, H. D. (1970). *Estimation of isotonic regression. Nonparametric Techniques in Statistical Inference*. 177–195. Cambridge University Press

[10] Carolan, C. and Dykstra, R. (1999). Asymptotic behavior of the Grenander estimator at density flat regions. *The Canadian Journal of Statistics*, **27** 557–566.

[11] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics* **43**, 1774–1800.

[12] Du, D. Z. and Pardalos, P. M. (1999). Handbook of Combinatorial Optimization: Supplement Volume A. Springer Science+Business Media Dordrecht

[13] Grenander, U. (1956). On the theory of mortality measurement. *Skand. Aktuarietidskr.*, **39** 125–153.

[14] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric estimation under shape constraints*. Cambridge University Press.

[15] Han, Q., Wang, T., Chatterjee, S. and Samworth, R.J. (2017) Isotonic regression in general dimensions, arxiv:1708:09468

[16] Jankowski, H. K. and Wellner, J. A. (2009). Estimation of a discrete monotone distribution. *Electronic journal of statistics* **39**, 125–153.

[17] Jennrich, R. I. (1969). Asymptotic properties of nonlinear least squares estimators. *The Annals of Statistics* **40**, 663–643.

[18] Polonik, W. (1998) The silhouette, concentration functions and ML-density estimation under order restrictions. *The Annals of Statistics* **26**, 1857–1877.

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

[20] Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order restricted statistical inference*. John Wiley & Sons, Ltd., Chichester.

[21] Shiryaev, A. (2007). *Probability*. Springer, New York.

[22] Silvapulle, M. J. and Sen, P. K. (2005). *Constrained Statistical Inference*. John Wiley & Sons, Ink., Hoboken, New Jersey.

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

[24] Wright, F. T. (1981). The asymptotic behaviour of monotone regression estimates. *The Annals of Statistics* **9**, 443–448.

[25] Wu, C. (1981). Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics* **9**, 501–513.

B

Paper B

Estimation of a discrete monotone distribution with model selection

Dragi Anevski and Vladimir Pastukhov

Centre for Mathematical Sciences, Lund University

Abstract

We introduce a new method of estimating a discrete monotone probability mass function. We propose a two-step procedure. First, we perform a model selection introducing the Akaike-type information criterion ($CMAIC$). Second, using the selected class of models we construct a modified Grenander estimator by grouping the parameters in the constant regions and then projecting the grouped empirical estimator onto the isotonic cone. We show that the post-model-selection estimator performs asymptotically better, in l_2-sense, than the regular Grenander estimator.

Keywords: Constrained inference, isotonic regression, density estimation, Grenander estimator, limit distribution.

1 Introduction

In this paper we study a two step procedure for estimating a monotone probability mass function (pmf). The procedure consists of an order restricted estimation step, which takes into account detailed information about the shape of the estimand, preceded by a model selection procedure step, for selecting the appropriate class of shape restricted pmfs.

Our procedure is slightly reminiscent of the pioneering paper by Jankowski and Wellner [12], which was the first to introduce and study two estimators

that satisfy order restrictions on the unknown pmf, and in fact the work in [12] is a main motivation for our paper. The two estimators that were introduced in [12] are the order restricted maximum likelihood estimator (mle) $\hat{\boldsymbol{p}}_n^G$ and the monotone rearrangement of the empirical estimator $\hat{\boldsymbol{p}}_n^R$, respectively. The limit distributions of the two estimators were established, and it was shown that the order restricted mle $\hat{\boldsymbol{p}}_n^G$ had a smaller l^2-risk than the monotone rearrangement estimator $\hat{\boldsymbol{p}}_n^R$.

Our work is also motivated by [24], which studied the problem of isotonic regression based on i.i.d. data of an estimand with continuous support and proposed grouping of adjacent observations, isotonization of the corresponding means and then interpolation to the whole support. The procedure studied in [24] gives a better rate of convergence and a normal limiting distribution of the restricted estimator, as opposed to the standard $n^{-1/3}$ rate and the Chernoff limit distribution that is common in the asymptotic theory for order restricted inference. The author also discussed the interpolation scheme and a proper way to make a partition of the support.

The limit distributions of both the rearrangement $\hat{\boldsymbol{p}}_n^R$ and the order restricted mle $\hat{\boldsymbol{p}}_n^G$, depend on the regions of constancy of the estimand, the true pmf $\boldsymbol{p}$, which are unknown in general. It is, therefore, not straightforward to use those limiting distribution results to, for example, construct confidence intervals. Furthermore, in our paper we show that it is not optimal to ignore the existence of the constant regions in the process of constrained estimation. In particular, we show that our proposed post-model-selection estimator $\hat{\boldsymbol{p}}_n^*$, performs better than the order restricted mle $\hat{\boldsymbol{p}}_n^G$ in the sense of having an almost surely asymptotically smaller l^2-risk, i.e. it satisfies

$$\mathbb{P}[\liminf_{n\to\infty}\{||\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}||_2 \leq ||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2\}] = 1, \tag{B.1}$$

and it is satisfied with the inequality in $(B.28)$ changed to an equality if $\boldsymbol{p}$ is strictly decreasing. Since it was shown in [12] that $\hat{\boldsymbol{p}}_n^G$ performs better than $\hat{\boldsymbol{p}}_n^R$ in this sense, our estimator also performs better than the monotone rearrangement estimator $\hat{\boldsymbol{p}}_n^R$.

The estimator $\hat{\boldsymbol{p}}_n^*$ that we propose is a version of the order restricted mle first introduced in [12], that we however apply an approach which is slightly reminiscent to the approach for the continuous support case, used in [24], to and that we in addition introduce a novel model selection criterion for. The model assumption for the unknown pmf is that it is monotone, but that it may be not everywhere strictly monotone, and thus we assume that it may have sets of

(adjacent) points in its discrete support on which it is constant. Our algorithm consists of, first, changing the inference problem slightly, by modifying the estimand $\boldsymbol{p}$ into a new estimand $\boldsymbol{p}'$, that is also monotone, and that is possibly and ideally *strictly* monotone, by grouping the values of $\boldsymbol{p}$ at the levels of constancy of $\boldsymbol{p}$. Next, we calculate the order restricted mle of this modified monotone pmf, and, finally, we expand or interpolate the order restricted mle into an estimator of the original estimand. We show in Theorem 4.2 below that, when the grouping of the values is done in an appropriate way, this resulting estimator has a smaller l^2 risk than the unmodified order restricted mle introduced in [12]. However, the procedure for calculating the "estimator" involves a grouping of the data according to the levels of constancy of the estimand, and thus it can not be calculated on merely the data; in fact the algorithm is *not* an estimator, since its calculation depends on $\boldsymbol{p}$, the unknown parameter. If we know the levels of constancy of the estimand it would however be an estimator, which would outperform the unmodified order restricted mle $\hat{\boldsymbol{p}}_n^G$ that was introduced in [12]. A natural idea is then to try to estimate the levels of constancy, or rather to make a *model selection* of the appropriate pfm, where the outcome from the model selection procedure is the levels of constancy of the unknown pmf, and then for this selected pmf construct the above modified order restricted mle. This is in fact the approach we use in our paper. Thus, we first obtain the regions of constancy of an underlying pmf $\boldsymbol{p}$ (or at least some of them), using model selection. Previous results in model selection under order restrictions, cf. [4] and the discussion below, are not directly applicable to our problem, and in Section 5 we introduce a novel constrained monotone Akaike-type information criterion ($CMAIC$). Next, we use this information in the construction of an order restricted estimator.

An information criterion for the parameters under (simple) order restrictions was proposed in [4], which studied sampling from normal distributions with either known variances or with known variance ratios, as well as sampling from exponential families. A generalization of this model selection criterion in the one-way analysis of variance model when the population means may be restricted by a mixture of linear equality and inequality constraints (ORIC) was proposed in [13]. A further generalisation of ORIC to multivariate normal linear models (GORIC) is given in [14].

An approach which is somewhat related to ours, in testing for monotone parameters, was proposed in [23], where the authors studied tests in which both the null and the alternative hypotheses describe order restrictions for a finite

set of parameters, namely

$$H_0 \quad : \theta_1 = \cdots = \theta_{q_2-1} \geq \theta_{q_2} = \cdots = \theta_{q_3-1} \geq \cdots \geq \theta_{q_m} = \cdots = \theta_k$$
$$H_1 \quad : \theta_1 \geq \theta_2 \geq \cdots \geq \theta_{k-1} \geq \theta_k,$$

respectively, with $(\theta_1, \ldots, \theta_k)$ the vector of parameters of interest. This problem apparently arises in psychiatric research of unipolar affective disorder, cf. [23]. The authors considered multinomial sampling as well as independent samples from k populations, with each population following an exponential family distribution.

In relation to our obtained results on l^2-risk bounds, in [7] the authors considered the problem of estimating a vector $\boldsymbol{\theta} \in \mathbb{R}^n$ under isotonic constraints and studied the risk bound in isotonic regression. They proved that the rate of convergence of the risk depends on the shape of the vector $\boldsymbol{\theta}$, i.e. on the constant regions in $\boldsymbol{\theta}$.

The paper is organised as follows. In Section 2 we give a short review of some previous estimators of a decreasing pmf. In Section 3 we make a formal statement of the problem and introduce some notation. In Section 4 we consider the case when the model class, denoted $\mathcal{F}^*_{k,\boldsymbol{w}}$ in the sequel, is fixed, i.e. when it has been chosen in advance (i.e. not based on the data) and, therefore, does not change with n. We distinguish between two possibilities for the candidate class; namely that (i) it contains properly the true model class, when the grouping of the values is done in an appropriate way, and (ii) that it does not contain the true model class, cf. $(B.17)$ in Section 3 for the proper definition. First, in Subsection 4.1, we study the case (i), when the selected class contains the true class, i.e. when $\mathcal{F}^*_{k,\boldsymbol{w}} \supseteq \mathcal{F}^*_{k,\boldsymbol{v}}$, or, equivalently, when $\mathcal{F}^*_{k,\boldsymbol{w}}$ contains $\boldsymbol{p}$. In Theorem 4.2 it is shown that when $\mathcal{F}^*_{k,\boldsymbol{w}}$ contains $\boldsymbol{p}$, the estimator $\hat{\boldsymbol{p}}^*_n$ has, properly scaled, asymptotically smaller l^2-risk, compared to the order restricted mle $\hat{\boldsymbol{p}}^G_n$. Second, in Subsection 4.2 we study the case when the selected class $\mathcal{F}^*_{k,\boldsymbol{w}}$ does not contain the true class. Next, in Section 5, using the results of Section 4, we describe the model selection procedure, derive the Akaike-type information criterion (CMAIC) and study its performance. In Section 6 we study the post-model-selection estimator $\hat{\boldsymbol{p}}^*_n$, taking into account that the selected class depends on data. We show that the post-model-selection estimator $\hat{\boldsymbol{p}}^*_n$ has asymptotically smaller l^2-risk than the regular order restricted mle. In Section 7 a simulation study illustrates the behaviour of $\hat{\boldsymbol{p}}^*_n$ in comparison with $\hat{\boldsymbol{p}}^G_n$. The proofs of all results are given in an Appendix and in the Supplementary material. The R code for $CMAIC$ and for Algorithms 3.1 and 5.1 is available upon request.

2 Review of previous estimators of a decreasing probability mass function

Suppose that we have observed $X_1, X_2, \ldots, X_n$ i.i.d. random variables with pmf $\boldsymbol{p}$ defined on $\mathbb{N}_+$, let $k = \sup\{i : p_i > 0\}$, and note that we allow both $k < \infty$ and $k = \infty$. The model assumption for $\boldsymbol{p}$ is that it is decreasing. The empirical estimator $\hat{\boldsymbol{p}}_n$ of $\boldsymbol{p}$ is then given by

$$\hat{p}_{n,i} = \frac{n_i}{n}, \tag{B.2}$$

where

$$n_i = \sum_{j=1}^{n} 1\{X_j = i\}, \tag{B.3}$$

for $i \in \mathbb{N}_+$, and it is equivalent to the unrestricted mle

$$\hat{\boldsymbol{p}}_n = \operatorname*{argmax}_{\boldsymbol{f} \in \mathcal{G}_k} \prod_i f_i^{n_i}, \tag{B.4}$$

where

$$\mathcal{G}_k = \Big\{ \boldsymbol{f} \in \mathbb{R}_+^k : \sum_{i=1}^{k} f_i = 1 \Big\}.$$

The empirical estimator $\hat{\boldsymbol{p}}_n$ is unbiased, consistent and asymptotically normal, cf. [12, 20]. It, however, does not necessarily satisfy the order restriction

$$\hat{p}_{n,1} \geq \hat{p}_{n,2} \geq \cdots \geq \hat{p}_{n,k}. \tag{B.5}$$

An estimator which does satisfy the order restricton (B.5) is the monotone rearrangement of the empirical estimator, $\hat{\boldsymbol{p}}_n^R$, defined by

$$\hat{\boldsymbol{p}}_n^R = rear(\hat{\boldsymbol{p}}_n), \tag{B.6}$$

where $\hat{\boldsymbol{p}}_n$ is the unrestricted mle in (B.2) and $rear(v)$ for a vector $v = (v_1, \ldots, v_k)$ is the reverse-ordered vector. The monotone-rearrangement estimator, as an estimator of a pmf, was introduced by [12], and was first used in a statistical framework, and then as an estimator of a probability density function (pdf), in [10], cf. also [2] for results on the use of this estimator for pdfs and for regression functions.

The order restricted mle, $\hat{\boldsymbol{p}}_n^G$, is defined by

$$\hat{\boldsymbol{p}}_n^G \;=\; \underset{f\in\mathcal{H}}{\operatorname{argmax}} \prod_i f_i^{n_i}, \tag{B.7}$$

where $\mathcal{H} = \left\{ f \in \mathbb{R}_+^k : \sum_{i=1}^k f_i = 1,\, f_1 \;\geq f_2 \geq \cdots \geq f_k \right\}$, and it is known to be equivalent to the isotonic regression of the unrestricted mle, see [5, 12, 17], i.e. $\hat{\boldsymbol{p}}_n^G = \hat{\boldsymbol{p}}_n^{IS}$, where

$$\hat{\boldsymbol{p}}_n^{IS} \;=\; \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^k (\hat{p}_{n,i} - f_i)^2,$$

with $\mathcal{F} = \left\{ f \in \mathbb{R}^k : \; f_1 \;\geq f_2 \geq \cdots \geq f_k \right\}$ and where $\hat{p}_{n,i}$ is the empirical estimator defined in (B.2).

The estimator $\hat{\boldsymbol{p}}_n^G$ is called the Grenander estimator in [12] and is derived using an algorithm which can be described as the vector of left derivatives of the least concave majorant of the cumulative sum diagram $(j, \mathbb{F}_n(j))$, for $j = 1, \ldots, k$, where $\mathbb{F}_n(x)$ is the empirical distribution function $\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^n 1\{X_i \leq x\}$. Incidently, the algorithm is the same as the one used to calculate the nonparametric mle of a decreasing pdf on $\mathbb{R}_+$, i.e. the original Grenander estimator.

In [12] the authors showed the consistency of the estimators $\hat{\boldsymbol{p}}_n^R$ and $\hat{\boldsymbol{p}}_n^G$. Furthermore, they first established that the empirical estimator is asymptotically Gaussian, namely that $\sqrt{n}(\hat{\boldsymbol{p}}_n - \boldsymbol{p})$ converges weakly in l^2 to an infinite dimensional Gaussian vector $Y_{\mathbf{0},\boldsymbol{B}}$ with mean zero and the covariance operator $\boldsymbol{B}$

$$< \boldsymbol{B} e_i, e_{i'} >= p_i \delta_{i,i'} - p_i p_{i'}, \tag{B.8}$$

with e_i the orthonormal basis in l^2 space, cf. e.g. [9] for Gaussian measures in infinite dimensional Hilbert spaces. Next, [12] derived the limit distribution result for the Grenander estimator

$$\sqrt{n}(\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}) \xrightarrow{d} \varphi(Y_{\mathbf{0},\boldsymbol{B}}), \tag{B.9}$$

where $\varphi : \mathbb{R}^k \to \mathbb{R}^k$ (or $\varphi : l^2 \to l^2$ if $k = \infty$) is an operator defined as follows: for any $Y \in \mathbb{R}^k$ (or $Y \in l^2$), for all constant regions of $\boldsymbol{p}$

$$[\varphi(Y)]^{(r,s)} = isot\{Y^{(r,s)}\},$$

where $[\boldsymbol{Z}]^{(r,s)}$ denotes the restriction of $\boldsymbol{Z} \in \mathbb{R}^k$ to the index set (r,s) and $isot\{\cdot\} : \mathbb{R}^{s-r+1} \to \mathbb{R}^{s-r+1}$ denotes the isotonic operator, cf. Theorems 3.1 and 3.8 in [12]. In [12] it was also shown that the Grenander estimator $\hat{\boldsymbol{p}}_n^G$ has a smaller l^2-risk than both the rearrangement estimator $\hat{\boldsymbol{p}}_n^R$ and the empirical estimator $\hat{\boldsymbol{p}}_n$.

3 Statement of the problem and notation

In this section we state the inference problem and introduce some notations.

Assumption 3.1. *Assume that $X_1, X_2, \ldots, X_n$ is an i.i.d. sample of random variables with unknown pmf $\boldsymbol{p}$. Suppose that $\boldsymbol{p} = \{p_i\}_{i\in\mathbb{N}_+}$ is a monotone decreasing pmf with support in $\mathbb{N}_+$. Let $k = \sup\{i : p_i > 0\}$, with both cases $k < \infty$ and $k = \infty$ allowed. Assume that $\boldsymbol{p}$ has flat regions, of the form*

$$\begin{aligned} p_{q_1} = \cdots = p_{q_1+v_1-1} > p_{q_2} &= \cdots = p_{q_2+v_2-1} > \cdots > \\ p_{q_m} &= \cdots = p_k, \end{aligned} \tag{B.10}$$

where q_j for $j = 1, \ldots m$ is the index of the first element in the j-th flat region, $p_{q_1} = p_1$, m is the total number of flat regions of $\boldsymbol{p}$, $\boldsymbol{v} = (v_1, \ldots, v_m)$ is the vector of the lengths (the numbers of points) of the flat regions of $\boldsymbol{p}$, so that $\sum_{j=1}^m v_j = k$.

Note, that we allow the flat regions to be one-point sets, namely at the places where $\boldsymbol{p}$ is strictly decreasing. Thus, if $\boldsymbol{p}$ is strictly decreasing at some index i and there are $j-1$ flat regions to the left of i, some of which may be one-point sets, then we put $v_j = 1$, so the size of the j-th flat region is 1. Furthermore, if $\boldsymbol{p}$ is strictly decreasing on the whole support, then $m = k$ and $\boldsymbol{v}$ is a vector of ones with length k. Also, since $\sum_{i\in\mathbb{N}_+} p_i = 1$, we must have $v_j < \infty$ for all j, i.e. each flat region of $\boldsymbol{p}$ has a finite number of points.

We consider estimation of both the finitely, $k < \infty$, and infinitely, $k = \infty$, supported pmfs $\boldsymbol{p}$. In the case of finite support, i.e. when $k < \infty$, the candidate class $\mathcal{H}^*_{k,\boldsymbol{w}}$ of pmfs is of the form

$$\begin{aligned} \mathcal{H}^*_{k,\boldsymbol{w}} = \Big\{ \boldsymbol{f} \in \mathbb{R}^k_+ : \sum_{i=1}^k f_i = 1,\, f_1 = \cdots = f_{w_1} \geq \\ f_{t_2} = \cdots = f_{t_2+w_2-1} \geq \cdots \geq f_{t_s} = \cdots = f_k \Big\}, \end{aligned} \tag{B.11}$$

where $\boldsymbol{t} = (t_1, \dots t_s)$, with $t_1 = 1$, is the vector of indices of the first elements of the constant regions, $\boldsymbol{w} = (w_1, w_2, \dots, w_s)$ is the vector of their lengths and $s = |\boldsymbol{w}| = |\boldsymbol{t}|$ is the number of constant regions in $\mathcal{H}^*_{k,\boldsymbol{w}}$. We can then define the order restricted mle, $\hat{\boldsymbol{p}}^*_n$, of $\boldsymbol{p}$, as

$$\hat{\boldsymbol{p}}^*_n = \underset{\boldsymbol{f} \in \mathcal{H}^*_{k,\boldsymbol{w}}}{\operatorname{argmax}} \prod_i f_i^{n_i}, \tag{B.12}$$

where the full data is collapsed, by sufficiency, to the count data n_i, $i = 1, \dots, k$, defined in $(B.3)$.

In the case of an infinite support, i.e. when $k = \infty$ (and note that we actually use the following approach also when k is finite but very large), we take some fixed finite r and consider the candidate class $\mathcal{H}^*_{k,\boldsymbol{w}} = \mathcal{H}^*_{k,\boldsymbol{w},r}$ defined by

$$\begin{aligned} \mathcal{H}^*_{k,\boldsymbol{w},r} \quad = \quad & \Big\{ \boldsymbol{f} \in l^2 : f_i \geq 0, \sum_{i=1}^{k} f_i = 1, f_1 = \dots = f_{w_1} \\ & \geq f_{t_2} = \dots = f_{t_2+w_2-1} \geq \dots \geq f_{t_d} = \dots = f_r \\ & \geq f_{r+1} \geq f_{r+2} \geq f_{r+3} \geq \dots \Big\}, \end{aligned} \tag{B.13}$$

where $\boldsymbol{t} = (t_1, \dots t_d, r+1, r+2, \dots)$, with $t_1 = 1$ and $(t_1, \dots t_d)$ are the indices of the first elements of the constant regions for the elements with the indices less or equal than r. Consequently, the vector of the lengths of the constant regions is $\boldsymbol{w} = (w_1, w_2, \dots, w_d, 1, 1, \dots)$. Note, that the true pmf $\boldsymbol{p}$ may have constant regions for $i > r$, but in the case of an infinite (or large) support we only search for the constant regions of $\boldsymbol{p}$ up to some fixed finite index r, and the constraints for $i > r$ are not active (i.e. they are treated as $\geq$). Thus, for this candidate class $\mathcal{H}^*_{k,\boldsymbol{w},r}$ we take into account flat regions of the pmf only on the finite part $\{1, \dots, r\}$ of the whole support set. The order restricted mle $\hat{\boldsymbol{p}}^*_n$, in the case of $k = \infty$, is defined as in $(B.12)$.

It is convenient to introduce also the cones of vectors that satisfy the appropriate order restrictions, but that do not necessarily satisfy the hyperplane condition that the vectors sum to one. Thus, in the case of finite support, we introduce $\mathcal{F}^*_{k,\boldsymbol{w}}$ as the following cone in $\mathbb{R}^k$

$$\begin{aligned} \mathcal{F}^*_{k,\boldsymbol{w}} = \Big\{ \boldsymbol{f} \in \mathbb{R}^k : f_1 = \dots = f_{w_1} \quad \geq \\ f_{t_2} = \dots = f_{t_2+w_2-1} \geq \dots \quad \geq \quad f_{t_s} = \dots = f_k \Big\}, \end{aligned} \tag{B.14}$$

with $\boldsymbol{t}, \boldsymbol{w}$ and s defined identically to as in the definition of $\mathcal{H}^*_{k,\boldsymbol{w}}$. In the case of an infinite support, when $k = \infty$, (or when k is very large) we again pick a

finite r and introduce $\mathcal{F}^*_{k,\boldsymbol{w}} = \mathcal{F}^*_{k,\boldsymbol{w},r}$ as the following cone in l^2

$$\begin{aligned}\mathcal{F}^*_{k,\boldsymbol{w},r} \;=\; & \Big\{\boldsymbol{f} \in l^2 :\, f_1 = \cdots = f_{w_1} \geq \\ & f_{t_2} = \cdots = f_{t_2+w_2-1} \geq \cdots \geq f_{t_d} = \cdots = f_r \geq \\ & f_{r+1} \geq f_{r+2} \geq f_{r+3} \geq \cdots \Big\},\end{aligned} \tag{B.15}$$

where $\boldsymbol{t}, \boldsymbol{w}$ are defined identially to as in the definition of $\mathcal{H}^*_{k,\boldsymbol{w},r}$.

In Section 4, we show that the solution to (B.12), under the constraints (B.11), is given by

$$\hat{\boldsymbol{p}}^*_n = \underset{\boldsymbol{f} \in \mathcal{F}^*_{k,\boldsymbol{w}}}{\operatorname{argmin}} \sum_i [\hat{p}_{n,i} - f_i]^2, \tag{B.16}$$

where the empirical estimator $\hat{p}_{n,i}$ is defined in (B.4), and, under the constraints (B.13), it is given by $(B.16)$ with $\mathcal{F}^*_{k,\boldsymbol{w}}$ replaced by $\mathcal{F}^*_{k,\boldsymbol{w},r}$.

The choice of an appropriate r will be discussed in Section 5. Most of the limit results in the paper hold for both finite and infinite k, and in order to keep the notations simpler we use the same notation $\mathcal{F}^*_{k,\boldsymbol{w}}$ for both cases and when the infinite case is different we will emphasise it.

We will say that the class $\mathcal{F}^*_{k,\boldsymbol{w}}$ is generated by the vector $\boldsymbol{w}$. Furthermore, we will say that for given k, $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, the class $\mathcal{F}^*_{k,\boldsymbol{w_1}}$, generated by $\boldsymbol{w_1}$, is bigger than the class $\mathcal{F}^*_{k,\boldsymbol{w_2}}$, generated by $\boldsymbol{w_2}$, if $\mathcal{F}^*_{k,\boldsymbol{w_2}} \subseteq \mathcal{F}^*_{k,\boldsymbol{w_1}}$, with $\subseteq$ given the ordinary set theoretic meaning. Therefore, for a given class $\mathcal{F}^*_{k,\boldsymbol{w}}$, defined in (B.14), there are $2^{\sum_{j=1}^{s} w_j - s} = 2^{k-s}$ classes bigger or equal to $\mathcal{F}^*_{k,\boldsymbol{w}}$, and for a given class $\mathcal{F}^*_{k,\boldsymbol{w}}$, defined in (B.15), there are $2^{\sum_{j=1}^{d} w_j - d} = 2^{r-d}$ classes bigger or equal to $\mathcal{F}^*_{k,\boldsymbol{w}}$. For a fixed k, there is a one-to-one correspondence between $\boldsymbol{w}$ and $\mathcal{F}^*_{k,\boldsymbol{w}}$, therefore, in the sequel of the paper we will sometimes write $\boldsymbol{w}$ instead of $\mathcal{F}^*_{k,\boldsymbol{w}}$ to denote the corresponding class $\mathcal{F}^*_{k,\boldsymbol{w}}$.

Note that in general, $\boldsymbol{v}$, the vector of constant regions of the true pmf $\boldsymbol{p}$, and, consequently, $\boldsymbol{q}$, the vector of indices of the first elements in the constant region, defined in (D.27), are unknown. Furthermore, for any choice of candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$, given in (B.14) or (B.15), in general, the vector of constant regions $\boldsymbol{w}$ of the candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$ may or may not be equal to the vector $\boldsymbol{v}$ of constant regions of the true pmf $\boldsymbol{p}$, thus we may or may not have $\mathcal{F}^*_{k,\boldsymbol{w}} = \mathcal{F}^*_{k,\boldsymbol{v}}$. In fact, we can have two possibilities

$$\begin{aligned} &(i) \qquad \mathcal{F}^*_{k,\boldsymbol{v}} = \mathcal{F}^*_{k,\boldsymbol{w}} \text{ or } \mathcal{F}^*_{k,\boldsymbol{v}} \subset \mathcal{F}^*_{k,\boldsymbol{w}}, \\ &(ii) \qquad \mathcal{F}^*_{k,\boldsymbol{v}} \not\subset \mathcal{F}^*_{k,\boldsymbol{w}}. \end{aligned} \tag{B.17}$$

We note that for any candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$, as defined in either (B.14) or (B.15), in order to have $\boldsymbol{p} \in \mathcal{F}^*_{k,\boldsymbol{w}}$, i.e. in order to have either of the two subcases in (i) of $(B.17)$, one needs to have at least as many constant regions in the candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$ as there are constant regions in $\boldsymbol{p}$, i.e. one needs $s \geq m$, and that for any $j_1 \in \{1, \ldots, m\}$ there exist $j_2 \in \{1, \ldots, s\}$ such that $\sum_{i=1}^{j_1} v_i = \sum_{i=1}^{j_2} w_i$. Therefore, $\boldsymbol{p} \in \mathcal{F}^*_{k,\boldsymbol{w}}$ if and only if there are no active constraints in the class $\mathcal{F}^*_{k,\boldsymbol{w}}$ in between the following pairs of elements $(f_{q_2-1}, f_{q_2}), (f_{q_3-1}, f_{q_3}) \ldots$.

We also introduce the following notation for restricting a subset $A \subset \mathbb{R}^\infty$ to a coordinate set I. Let $I \subset \{1, 2, 3, \ldots\}$ be a set of indices. Then we define

$$[A]^I \quad = \quad \{x(I) \in \mathbb{R}^{|I|} : x \in A\}, \tag{B.18}$$

where $x(I)$ denotes the length-$|I|$ vector consisting of the I coordinates of x.

Finally, let us introduce the notation $\Pi(\boldsymbol{y}|\mathcal{F})$ for the l^2 projection of a vector $\boldsymbol{y} \in l^2$ onto a fixed but arbitrary closed convex cone $\mathcal{F}$ in l^2, i.e.

$$\Pi(\boldsymbol{y}|\mathcal{F}) = \underset{z \in \mathcal{F}}{\operatorname{argmin}} \sum_i (z_i - y_i)^2. \tag{B.19}$$

3.1 An algorithm for a monotone pmf estimator with prior model selection

In this subsection we describe an algorithm for our proposed estimator. Assume that we are given a data set $(x_1, \ldots, x_n)$ of observations from n i.i.d. random variables $X_1, X_2, \ldots, X_n$ from the pmf $\boldsymbol{p}$. To estimate the underlying decreasing pmf $\boldsymbol{p}$ we propose the following model selection based algorithm. Note that the algorithm is valid in the two cases $k < \infty$ and $k = \infty$.

Algorithm 3.1.

1. ***The model selection.***

   First, we perform a model selection to obtain the class $\mathcal{F}^*_{k,\boldsymbol{w}}$. This is described in detail in Section 5.

2. ***Grouping of the parameters.***

   Then, we make a reparametrization by grouping the probabilities which are required to be equal, according to the constant regions $\boldsymbol{w}$ of the selected class $\mathcal{F}^*_{k,\boldsymbol{w}}$. This is done by, at the j-th flat region of $\boldsymbol{p}$, setting the value of the new

parameter p'_j to be equal to the common probability values of $\boldsymbol{p}$ at the j-th flat region. The resulting reparametrized true pmf $\boldsymbol{p}' = (p'_1, \dots p'_s)$ then satisfies

$$\sum_{j=1}^{s} w_j p'_j = 1$$

and the values p'_j are strictly decreasing, if the chosen class $\mathcal{F}^*_{k,\boldsymbol{w}}$ is exactly equal to the true one.

3. ***Isotonisation of the grouped empirical estimator.*** *First, we define $\hat{\boldsymbol{p}}'_n = (\hat{p}'_{n,1}, \dots \hat{p}'_{n,s})$, the unrestricted (i.e. without order restrictions) mle, as*

$$\hat{\boldsymbol{p}}'_n = \underset{\boldsymbol{f}' \in \mathcal{G}'_{s,\boldsymbol{w}}}{\operatorname{argmax}} \prod_j {f'}_j^{n'_j}, \tag{B.20}$$

where

$$\mathcal{G}'_{s,\boldsymbol{w}} = \Big\{ \boldsymbol{f}' \in \mathbb{R}^s_+ : \sum_{j=1}^{s} w_j f'_j = 1 \Big\}, \tag{B.21}$$

and

$$n'_j = \sum_{l=1}^{n} 1\{X_l \in \{t_j, t_j + w_j - 1\}\}. \tag{B.22}$$

Next, we find the order restricted mle $\hat{\boldsymbol{p}}_n^{'G}$ of $\boldsymbol{p}' = (p'_1, \dots p'_s)$ by, equivalently, finding the isotonic regression of $\hat{\boldsymbol{p}}'_n$ with the weights $\boldsymbol{w} = (w_1, \dots, w_s)$, i.e. as

$$\hat{\boldsymbol{p}}_n^{'G} = \underset{\boldsymbol{f}' \in \mathcal{F}'_{s,\boldsymbol{w}}}{\operatorname{argmin}} \sum_j [\hat{p}'_{n,j} - f'_j]^2 w_j,$$

where

$$\mathcal{F}'_{s,\boldsymbol{w}} = \Big\{ \boldsymbol{f}' \in \mathbb{R}^s_+ : f'_1 \geq f'_2 \geq \cdots \geq f'_s \Big\}. \tag{B.23}$$

See Section 4 below for a proof of the equivalence of the order restricted mle and the weighted isotonic regression.

4. ***Interpolation to the whole support.*** *Finally, we interpolate the estimator $\hat{\boldsymbol{p}}_n^{'G}$, which has its support on the indices $(1, \dots, s)$, to an estimator on the whole support of $\boldsymbol{p}$. This is done by writing*

$$\hat{\boldsymbol{p}}^*_n = \boldsymbol{A} \hat{\boldsymbol{p}}_n^{'G}, \tag{B.24}$$

where $\boldsymbol{A}$ is a linear operator from l^2 to l^2, the application of which $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ on a vector $\boldsymbol{x}$ of dimension $\dim(\boldsymbol{x}) = s$ gives a vector $\boldsymbol{y}$ of dimension $\dim(\boldsymbol{y}) = k$ and where the components of $\boldsymbol{y}$ are given by

$$y_{(t_j:t_j+w_j-1)} = x_j$$

Note that in the case of a finite k, $\boldsymbol{A}$ is a $k \times m$ matrix, with non-zero elements:

$$[A]_{t_j:t_j+w_j-1,j} = 1, \tag{B.25}$$

and with all other elements not defined in (B.25) equal to zero.

The goal is to investigate the resulting post-model-selection estimator $\hat{\boldsymbol{p}}_n^*$ and compare its performance with the Grenander estimator $\hat{\boldsymbol{p}}_n^G$ defined in (B.7).

4 Characterization of the estimator for a fixed model class and asymptotic results for the estimator

In this section we assume that the candidate model class is fixed, i.e. we assume that the candidate class does not depend on the data and does not change with n. To clarify, we thus assume that we *have* made a choice of the model class, but that this is a deteministic process, in the sense that it has not been influenced by the data. In Section 6 below we present a corresponding treatment of the more realistic scenario in which the selection based estimator uses a candidate class which is data dependent.

First, we note that the problem in (B.12) is equivalent to (B.16), i.e.

$$\hat{\boldsymbol{p}}_n^* = \Pi(\hat{\boldsymbol{p}}_n | \mathcal{F}_{k,w}^*) \tag{B.26}$$

with $\hat{\boldsymbol{p}}_n$ the empirical estimator, defined in (B.4), $\mathcal{F}_{k,w}^*$ defined in (B.14) or (B.15), and with $\Pi(\cdot|\mathcal{F}_{k,w}^*)$ the l^2 projection on the cone $\mathcal{F}_{k,w}^*$, defined in (B.19), cf. pages 45–46 in [5] and pages 38–39 in [17]. Thus the order restricted mle coincides with the l^2 projection of the empirical estimator onto a corresponding cone. We, therefore, in the sequel study the estimator as being defined in $(B.26)$, and note that the discussions about and choices of the candidate classes $\mathcal{H}_{k,\boldsymbol{w}}^*$ and $\mathcal{H}_{k,\boldsymbol{w},r}^*$ are transformed to the corresponding discussions about and choices of for the candidate classes $\mathcal{F}_{k,\boldsymbol{w}}^*$ and $\mathcal{F}_{k,\boldsymbol{w},r}^*$.

4.1 The class $\mathcal{F}^*_{k,w}$ contains the true model class

In this subsection we assume that $\boldsymbol{p} \in \mathcal{F}^*_{k,w}$. First we study the asymptotic properties of the estimator $\hat{\boldsymbol{p}}^*_n$. In order to introduce the limit random variable, we introduce below a cone that is adapted from $\mathcal{F}^*_{k,\boldsymbol{w}}$, by relaxing some of the restrictions.

Note that since $\boldsymbol{p} \in \mathcal{F}^*_{k,w}$, there must be $k_1, k_2, k_3, \ldots$ integers, with $k_i \geq 1$ for all $i \geq 1$, such that the union of first k_1 regions of constancy in $\mathcal{F}^*_{k,w}$ is equal to the first region of constancy of $\boldsymbol{p}$, the union of the k_2 next regions of constancy in $\mathcal{F}^*_{k,w}$ is equal to the second region of constancy of $\boldsymbol{p}$, and so on. We define the cone $\mathcal{G}^*_{k,\boldsymbol{w},p}$ as the Cartesian product

$$\mathcal{G}^*_{k,\boldsymbol{w},p} \quad = \quad \bigtimes_{j \geq 1} [\mathcal{F}^*_{k,w}]^{(q_j, q_j + v_j - 1)}, \tag{B.27}$$

where $[\mathcal{F}^*_{k,w}]^{(q_j, q_j+v_j-1)}$ is the cone consisting of the restriction of $\mathcal{F}^*_{k,w}$ to the coordinates in the j'th region of constancy of $\boldsymbol{p}$, cf. Assumption 3.1 and (B.18). Thus we have relaxed some of the previous conditions in $\mathcal{F}^*_{k,\boldsymbol{w}}$ in the definition of $\mathcal{G}^*_{k,\boldsymbol{w},p}$, so that there are no active constraints in (B.27) in between the regions of constancy of $\boldsymbol{p}$ i.e. between (x_{v_1}, x_{v_1+1}), $(x_{v_1+v_2}, x_{v_1+v_2+1}), \ldots$ $(x_{\sum_{j=1}^{k-1} v_j}, x_{\sum_{j=1}^{s-1} v_j + 1})$, where $\boldsymbol{v} = (v_1, \ldots, v_s)$ is the vector of the lengths of the regions of constancy of true pmf $\boldsymbol{p}$.

Theorem 4.1. *Given that $\boldsymbol{p} \in \mathcal{F}^*_{k,w}$, the estimator $\hat{\boldsymbol{p}}^*_n$ is strongly consistent*

$$\hat{\boldsymbol{p}}^*_n \overset{a.s.}{\to} \boldsymbol{p}$$

and its asymptotic distribution is given by

$$\sqrt{n}(\hat{\boldsymbol{p}}^*_n - \boldsymbol{p}) \overset{d}{\to} \Pi(Y_{\boldsymbol{0},\boldsymbol{B}} | \mathcal{G}^*_{k,\boldsymbol{w},p}),$$

where $Y_{\boldsymbol{0},\boldsymbol{B}}$ is a Gaussian vector in l^2 with mean zero and covariance operator $\boldsymbol{B}$ given by $< \boldsymbol{B}e_i, e_j >= \delta_{ij} p_i - p_i p_j$.

The limit random variable in Theorem 4.1 can be seen as a concatenation of separate isotonic regressions over each region of constancy $(q_j, q_j + v_j - 1)$ of the Gaussian vector $Y_{\boldsymbol{0},\boldsymbol{B}}$, but where the isotonic regression is for functions that are monotone with respect to a certain preorder that is induced by the candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$ and follows from recent results in [3]. The proof of the

theorem as well a discussion of the relevant preorder, which we denote $\preceq_{\boldsymbol{w}}$, can be found in the appendix. Note however that for the statement of the theorem it is not necessary to know about the preorder.

The next theorem shows that the estimator $\hat{\boldsymbol{p}}_n^*$ performs asymptotically better, in the l^2-sense, than the regular Grenander estimator $\hat{\boldsymbol{p}}_n^G$.

Theorem 4.2. *Assume that $\boldsymbol{p} \in \mathcal{F}_{k,w}^*$. Then*

$$\mathbb{P}[\liminf_{n\to\infty}\{||\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}||_2 \leq ||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2\}] = 1,$$

Therefore, we have proved that for any pmf $\boldsymbol{p}$ there is n_1 such that $||\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}||_2 \leq ||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2$ a.s. for all $n > n_1$, provided that $\boldsymbol{p} \in \mathcal{F}_{k,w}^*$. Next, in [12] it was shown that

$$\lim_{n\to\infty} \mathbb{E}[n||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2^2] = \sum_{j=1}^{m}\sum_{i=1}^{v_j} p_{q_j}(\frac{1}{i} - p_{q_j}),$$

where q_j for $j = 1, \dots m$ is the index of the first element in the j-th constant region, $\boldsymbol{v} = (v_1, \dots, v_m)$ is the vector of the lengthsof the constant regions of true pmf $\boldsymbol{p}$.

Therefore, we have the following corollary

Corollary 4.1. *Assume that $\boldsymbol{p} \in \mathcal{F}_{k,w}^*$. Then, there exists n_1 such that for all $n > n_1$ one has*

$$\mathbb{E}[n||\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}||_2^2] \leq \mathbb{E}[n||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2^2].$$

4.2 The class $\mathcal{F}_{k,w}^*$ does not contain the true model

Assume that $\mathcal{F}_{k,v}^* \not\subset \mathcal{F}_{k,w}^*$ or, equivalently, that $\boldsymbol{p} \notin \mathcal{F}_{k,w}^*$. Then in this case the estimator $\hat{\boldsymbol{p}}_n^*$ is not consistent. In fact, using the continuous mapping theorem we have that

$$\hat{\boldsymbol{p}}_n^* = \Pi(\hat{\boldsymbol{p}}_n|\mathcal{F}_{k,w}^*) \xrightarrow{a.s.} \Pi(\boldsymbol{p}|\mathcal{F}_{k,w}^*) \neq \boldsymbol{p},$$

since the projection operator, cf. (B.19), is a continuous map, and where the inequality on the right hand side follows since $\boldsymbol{p} \notin \mathcal{F}_{k,w}^*$. Therefore, $\limsup_{n\to\infty} nl_2^2(\hat{\boldsymbol{p}}_n^*, \boldsymbol{p})$ becomes infinite. Below, in Section 5, we will prove that the classes $\mathcal{F}_{k,\boldsymbol{w}}^*$ which do not include the configuration of the true pmf $\boldsymbol{p}$ will be asymptotically excluded in the model selection process.

5 Akaike-type information criterion for model selection

In this section we construct an Akaike-type information criterion to obtain the vector of regions of constancy $\boldsymbol{w}$. We assume that the true pmf $\boldsymbol{p}$ belongs to $\mathcal{F}_k^* = \cup_{\boldsymbol{w}} \mathcal{F}_{k,\boldsymbol{w}}^*$, i.e. that the problem is correctly specified.

5.1 The case of a finite support of p

First, we consider the case when $k < \infty$ and assume we are given a data set $(x_1, \ldots, x_n)$ of observations from n i.i.d. random variables $X_1, X_2, \ldots, X_n$ generated by a pmf $\boldsymbol{p}$. For some pmf $\boldsymbol{f} \in \mathcal{F}_{k,\boldsymbol{w}}^*$ the log-likelihood is given by

$$l(x_1, \ldots, x_n | \boldsymbol{f}) = \sum_{i=1}^{k} n_i \log f_i, \tag{B.28}$$

with $n_i = \sum_{j=1}^{n} 1\{X_j = i\}$. We aim to find $\boldsymbol{w}$, the shape of the underlying pmf $\boldsymbol{p}$, i.e. to select the class $\mathcal{F}_{k,\boldsymbol{w}}^*$ which contains $\boldsymbol{p}$. Note that each vector $\boldsymbol{w}$ is a certain composition, in the number theoretic sense, of an integer k, which is the cardinality of the support of the underlying pmf $\boldsymbol{p}$. For a given k there are 2^{k-1} different compositions, and therefore 2^{k-1} different candidate classes of the form (B.14).

We use the approach originally developed by H. Akaike in [**?**], i.e. we choose the model which gives the lowest Kullback-Leibler discrepancy, cf. [15],

$$d(\hat{\boldsymbol{p}}_n^*) = -2\mathbb{E}_{\boldsymbol{p}}[l(X_1, \ldots, X_n | \boldsymbol{f})]|_{\boldsymbol{f}=\hat{\boldsymbol{p}}_n^*}, \tag{B.29}$$

where $\mathbb{E}_{\boldsymbol{p}}[\cdot]$ stands for the expectation with respect to the true pmf $\boldsymbol{p}$, and $\hat{\boldsymbol{p}}_n^* \in \mathcal{F}_{k,\boldsymbol{w}}^*$ is the mle, defined in (B.16). Though $d(\hat{\boldsymbol{p}}_n^*)$ cannot be evaluated, because $\boldsymbol{p}$ is unknown, it can be estimated. In [**?**] the author suggested to estimate $d(\hat{\boldsymbol{p}}_n^*)$, for general parametric models, by

$$\hat{d}(\hat{\boldsymbol{p}}_n^*) = -2l(x_1, \ldots, x_n | \hat{\boldsymbol{p}}_n^*), \tag{B.30}$$

which, however, is a biased estimator. We will use the estimate $\hat{d}(\hat{\boldsymbol{p}}_n^*)$ as a basis for our proposed information criterion and bias correct it. We define the information criterion as

$$CMAIC = \hat{d}(\hat{\boldsymbol{p}}_n^*) + B(\boldsymbol{w}),$$

cf. Definition 5.2 below, where the bias correction $B(\boldsymbol{w})$ is a certain sum of level probabilities of a Gaussian vector, defined below in Theorem 5.1. The remainder of this subsection is dedicated to the derivation of an expression for $B(\boldsymbol{w})$, cf. (B.35) below.

Let us assume that $\boldsymbol{p} \in \mathcal{F}^*_{k,\boldsymbol{w}}$ for a candidate family $\mathcal{F}^*_{k,\boldsymbol{w}}$ and let us study $d(\hat{\boldsymbol{p}}^*_n) - \hat{d}(\hat{\boldsymbol{p}}^*_n)$, where $\hat{\boldsymbol{p}}^*_n$ is the constrained mle in (B.16).

First, we note that

$$d(\hat{\boldsymbol{p}}^*_n) - \hat{d}(\hat{\boldsymbol{p}}^*_n) \quad = \quad -2\mathbb{E}_{\boldsymbol{p}}[l(X_1,\ldots,X_n|\boldsymbol{f})]|_{\boldsymbol{f}=\hat{\boldsymbol{p}}^*_n} + 2l(x_1,\ldots,x_n|\hat{\boldsymbol{p}}^*_n).$$

Second, using the result (2.16) in [8], for $d(\hat{\boldsymbol{p}}^*_n) - \hat{d}(\hat{\boldsymbol{p}}^*_n)$ we have

$$d(\hat{\boldsymbol{p}}^*_n) - \hat{d}(\hat{\boldsymbol{p}}^*_n) = \boldsymbol{Z}_n + 2\boldsymbol{V}_n^T\boldsymbol{P}^{-1}\boldsymbol{V}_n + o_p(1), \tag{B.31}$$

where $\boldsymbol{Z}_n$ is a random variable such that $\mathbb{E}_{\boldsymbol{p}}[\boldsymbol{Z}_n] = 0$ for all n, $\boldsymbol{V}_n = \sqrt{n}(\hat{\boldsymbol{p}}^*_n - \boldsymbol{p})$ and $\boldsymbol{P}$ is a diagonal $k \times k$ matrix with $\boldsymbol{P}_{jj} = p_j$.

Next, let $\mathcal{D} \in \mathbb{R}^k$ be some cone and let $P(\mathcal{D}, j, k)$ denote the probability that the projection of a standard normal k-dimensional vector on the cone $\mathcal{D}$ has j distinct values, for $j = 1,\ldots,k$, cf. [5, 17, 19]. The asymptotic distribution of $\boldsymbol{V}_n^T\boldsymbol{P}^{-1}\boldsymbol{V}_n$ is given in the next theorem.

Theorem 5.1. *The limit distribution of $\boldsymbol{V}_n^T\boldsymbol{P}^{-1}\boldsymbol{V}_n$ is given by*

$$\boldsymbol{V}_n^T\boldsymbol{P}^{-1}\boldsymbol{V}_n \quad \xrightarrow{d} \quad \mathcal{V}, \tag{B.32}$$

where $\mathcal{V}$ has the following distribution

$$\mathbb{P}[\mathcal{V} \leq v] \quad = \quad \sum_{j=1}^{s} P(\mathcal{G}^*_{k,\boldsymbol{w},p}, j, s)\mathbb{P}[\chi^2_{j-1} \leq v],$$

for any real number v, χ^2_j is a chi-square random variable with j degrees of freedom and $s = |\boldsymbol{w}|$.

Moreover,

$$\mathbb{E}[\boldsymbol{V}_n^T\boldsymbol{P}^{-1}\boldsymbol{V}_n] \quad \to \quad \mathbb{E}[\mathcal{V}] \tag{B.33}$$

and

$$\sum_{j=1}^{s} P(\mathcal{F}^*_{k,\boldsymbol{w}}, j, k)(j-1) \leq \mathbb{E}[\mathcal{V}] \leq (s-1). \tag{B.34}$$

As a consequence of the obtained bounds in (B.35), we may choose the bias correction term, $B(\boldsymbol{w})$, equal to the lower bound of $2\mathbb{E}[\mathcal{V}]$, i.e. we let

$$B(\boldsymbol{w}) \quad = \quad 2\sum_{j=1}^{s} P(\mathcal{F}^*_{k,\boldsymbol{w}}, j, k)(j-1). \tag{B.35}$$

With this choice of bias correction term we finally introduce an information criterion in the following definition.

Definition 5.2. *The monotone-constrained Akaike type information criterion (CMAIC) for a model class $\mathcal{F}^*_{k,\boldsymbol{w}}$, is defined by*

$$CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}}, n) \quad = \quad -2l(x_1, \ldots, x_n|\hat{\boldsymbol{p}}^*) + B(\boldsymbol{w}). \tag{B.36}$$

Note that our choice of $B(\boldsymbol{w})$ is analogous to [4, 13, 14]. Using the CMAIC criterion we can next define the selected model class, by equivalently define the selected vector $\boldsymbol{w}$.

Definition 5.3. *The selected model class, based on CMAIC, is*

$$\hat{\boldsymbol{w}}_n \quad = \quad \text{argmin}_{\boldsymbol{w}} CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}}, n).$$

We note that the selected model class is $\mathcal{F}^*_{k,\hat{\boldsymbol{w}}_n}$, that it is random and depends on n.

5.2 The case of an infinite support of p

Assume that the underlying pmf $\boldsymbol{p}$ has an infinite support, i.e. $p_i > 0$ for all $i \in \mathbb{Z}_+$, or that the support is very large. Let us choose some finite integer r and estimate the constant regions of $\boldsymbol{p}$ only among the index set $\{1, \ldots, r\}$, i.e. we aim to select the class $\mathcal{F}^*_{k,\boldsymbol{w}}$ defined in (B.15).

We construct a distribution $\mathfrak{p}$ on $\{1, \ldots, r+1\}$, obtained from $\boldsymbol{p}$, by

$$\begin{aligned} \mathfrak{p}_i \quad &= \quad p_i \text{ for } i \leq r, \\ \mathfrak{p}_{r+1} \quad &= \quad \sum_{j=r+1}^{\infty} p_j. \end{aligned}$$

Given a data set $(x_1, \ldots, x_n)$ of observations from n i.i.d. random variables $X_1, X_2, \ldots, X_n$ generated by a pmf $\boldsymbol{p}$, the data can be grouped to give us observations $z_1, \ldots, z_n$ from $\boldsymbol{p}$, by

$$z_j = \sum_{i=1}^{r} x_i 1\{x_j = i\} + (r+1)1\{x_j \geq r+1\}$$

The empirical estimator of $\mathfrak{p}$ is given by

$$\begin{aligned} \hat{\mathfrak{p}}_i &= \hat{p}_i \text{ for } i \leq r, \\ \hat{\mathfrak{p}}_{r+1} &= \sum_{j=r+1}^{\infty} \hat{p}_j. \end{aligned}$$

Observe that a decreasing pmf $\boldsymbol{p}$ belongs to $\mathcal{F}^*_{k,\boldsymbol{w}}$, defined in (B.15), if and only if $\mathfrak{p} \in \mathfrak{F}^*_{r,\boldsymbol{w}}$, where $\mathfrak{F}^*_{r,\boldsymbol{w}}$ is the following cone in $\mathbb{R}^{r+1}$

$$\begin{aligned} \mathfrak{F}^*_{r,\boldsymbol{w}} &= \Big\{ \mathfrak{f} \in \mathbb{R}^{r+1} : \mathfrak{f}_1 = \cdots = \mathfrak{f}_{w_1} \geq \\ &\qquad \mathfrak{f}_{w_1+1} = \cdots = \mathfrak{f}_{w_1+w_2} \geq \cdots \geq \mathfrak{f}_{\sum_{j=1}^{d-1} w_j + 1} = \cdots = \mathfrak{f}_r \Big\}. \end{aligned} \tag{B.37}$$

Therefore, with $\mathfrak{f} \in \mathfrak{F}^*_{k,\boldsymbol{w}}$ the pmf corresponding to $\boldsymbol{f} \in \mathcal{F}^*_{k,\boldsymbol{w}}$, the likelihood based on the data $z_1, \ldots, z_n$ from $\boldsymbol{p}$ is given by

$$l(z_1, \ldots, z_n | \mathfrak{f}) = \sum_{i=1}^{r+1} n_i \log \mathfrak{f}_i. \tag{B.38}$$

with $n_i = \sum_{j=1}^{n} 1\{Z_j = i\}$ for $i = 1, \ldots, r+1$.

We aim to select the class $\mathfrak{F}^*_{k,\boldsymbol{w}}$, which contains $\mathfrak{p}$. First define the isotonic regression under monotonicity assumptions only on the first r points in the support of pmf, by

$$\hat{\mathfrak{p}}^*_n = \underset{\mathfrak{f} \in \mathfrak{F}^*_{k,\boldsymbol{w}}}{\operatorname{argmin}} \sum_{i=1}^{r+1} (\mathfrak{f}_i - \hat{\mathfrak{p}}_{n,i})^2.$$

Similarly to the derivation for the finite case above, we obtain the following bias correction

$$\mathfrak{B}(\boldsymbol{w}) = 2 \sum_{j=1}^{d+1} P(\mathfrak{F}^*_{r,\boldsymbol{w}}, j, r)(j-1). \tag{B.39}$$

Definition 5.4. *The monotone-constrained Akaike-type information criterion (CMAIC) for the model class $\mathfrak{F}^*_{k,\boldsymbol{w}}$, is defined as*

$$CMAIC(\mathfrak{F}^*_{k,\boldsymbol{w}}, n) = -2l(x_1, \ldots, x_n|\hat{\mathfrak{p}}^*_n) + \mathfrak{B}(\boldsymbol{w}). \tag{B.40}$$

We use the results of the model selection procedure with the use of the information criterion $CMAIC$ defined in (B.40), i.e. the obtained vector of lengths of constant regions $(w_1, \ldots, w_d)$ to construct the candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$ in (B.15). Since a candidate class is equivalently specified by the vector $\boldsymbol{w}$, we may define the selected model class, based on the data $x_1, \ldots, x_n$, with the use of the CMAIC criterion, as in the next definition.

Definition 5.5. *The selected model class, based on CMAIC, is*

$$\hat{\boldsymbol{w}}_n \quad = \quad \text{argmin}_{\boldsymbol{w}} CMAIC(\mathfrak{F}^*_{k,\boldsymbol{w}}, n).$$

Note that the selected class, $\mathfrak{F}^*_{k,\hat{\boldsymbol{w}}_n}$, is random, and depends on n.

5.3 Asymptotic properties of $CMAIC$

$CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}}, n)$ provides a conservative model selection procedure in the sense that the parametric classes $\mathcal{F}^*_{k,\boldsymbol{w}}$ which do not include the configuration of the true pmf $\boldsymbol{p}$ will be asymptotically excluded in the model selection process. We are able to state a slightly stronger result in the next theorem.

Theorem 5.6. *Let $\mathcal{F}^*_1$ and $\mathcal{F}^*_2$ be two model classes, defined in (B.14) for the finite case or in (B.15) for the infinite case, such that $\boldsymbol{p} \in \mathcal{F}^*_1$ and $\boldsymbol{p} \notin \mathcal{F}^*_2$. Then*

$$\mathbb{P}[\liminf_{n\to\infty}\{CMAIC(\mathcal{F}^*_1, n) < CMAIC(\mathcal{F}^*_2, n)\}] = 1. \tag{B.41}$$

5.4 The model selection procedure and its performance

First, we emphasise that $CMAIC$ provides a conservative model selection procedure in the sense that the chosen class will contain the underlying pmf $\boldsymbol{p}$, but that this class is not necessarily the true one. In the case of a strictly decreasing true pmf $\boldsymbol{p}$, there is only one model class $\mathcal{F}^*_{k,\boldsymbol{w}}$, generated by $\boldsymbol{w} = (1, 1, \ldots, 1)$, which contains it, and then the chosen class, using $CMAIC$, will be the true one.

In order to analyse the model selection procedure we make a simulation study. First, let us consider the following strictly decreasing pmfs:

$$\begin{aligned}
\boldsymbol{M1}: p(x) &= (4/10, 3/10, 2/10, 1/10),\\
\boldsymbol{M2}: p(x) &= (6/21, 5/21, 4/21, 3/21, 2/21, 1/21),\\
\boldsymbol{M3}: p(x) &= (8/36, 7/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36),\\
\boldsymbol{M4}: p(x) &= (10/55, 9/55, 8/55, 7/55, 6/55, 5/55, 4/55, 3/55, 2/55, 1/55),\\
\boldsymbol{M5}: p(x) &= (12/78, 11/78, 10/78, 9/78, 8/78, 7/78, 6/78, 5/78,\\
&\quad 4/78, 3/78, 2/78, 1/78),\\
\boldsymbol{M6}: p(x) &= (16/136, 15/136, 14/136, 13/136, 12/136, 11/136, 10/136,\\
&\quad 9/136, 8/136, 7/136, 6/136, 5/136, 4/136, 3/136, 2/136, 1/136).
\end{aligned}$$

Note that in all the above models, the pmfs are decreasing and equidistant between subsequent values, i.e. $p_i - p_{i+1}$ is positive and does not depend on i, for all $i = 1, \ldots, k-1$.

We use a recently developed R package "restrictor" to compute the level probabilities $P(\mathcal{F}^*_{k,\boldsymbol{w}}, j, s)$, cf. [21], needed for the bias correction $B(\boldsymbol{w})$ in the calculation of the $CMAIC$ criterion.

Figure B.1 illustrates the performance of $CMAIC$ for the models **M1**-**M6** for 1000 Monte Carlo samples. Evaluating the plots in Figure B.1 we make an empirical conclusion on the number n of data points needed to detect a strictly decreasing model model, with high accuracy, for support size k, and present the conclusions in Table B.1.

Table B.1: Size of the data set n needed for the support's size k

k	n
$k \leq 5$	$n > 100k$
$5 < k \leq 10$	$n > 500k$
$10 < k \leq 15$	$n > 1000k$
$k > 15$	$n > 2000k$

Next, we consider the following models with several constant regions:

$$\begin{aligned}
\boldsymbol{M7}: p(x) &= 0.2U(4) + 0.8U(8),\\
\boldsymbol{M8}: p(x) &= 0.25U(2) + 0.2U(4) + 0.15U(6) + 0.4U(8),\\
\boldsymbol{M9}: p(x) &= 0.15U(4) + 0.1U(8) + 0.75U(12),
\end{aligned}$$

where $U(k)$ denotes the uniform pmf on $\{1, \ldots, k\}$. These are the same probability mass functions as the ones studied in [12]. Figure B.2 illustrates the

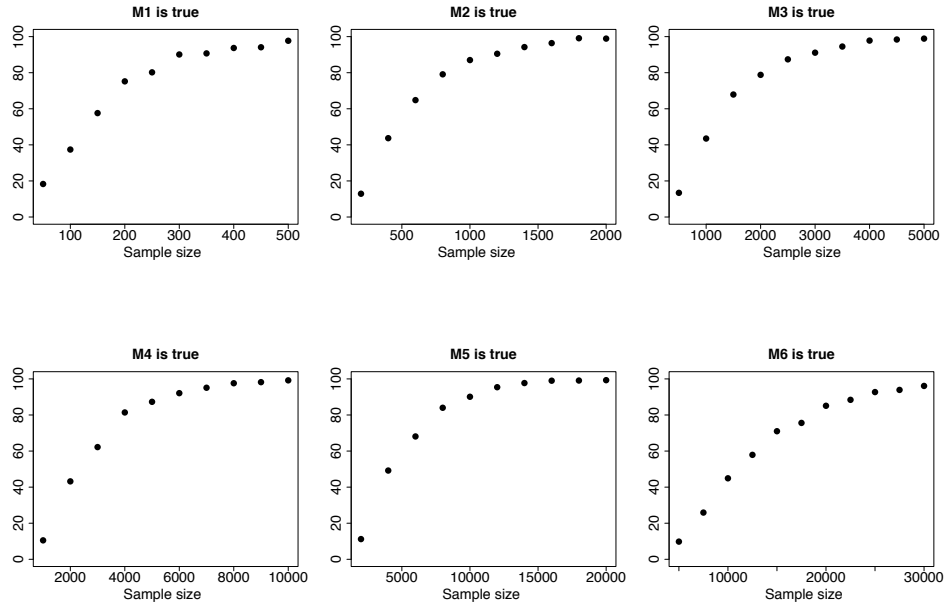

Figure B.1: Performance of $CMAIC$ for strictly decreasing models **M1**-**M6**. Percentage of times that the true model was chosen versus the sample size n.

performance of $CMAIC$ for the models **M7**, **M8** and **M9** for 1000 Monte Carlo simulations. One can see that as the sample size increases the probability that the selected class $\mathcal{F}^*_{k,\boldsymbol{w}}$ contains $\boldsymbol{p}$ goes to 1. However, the probability that the selected class is exactly the true one does not go to 1 as the sample size becomes larger.

In order to improve the model selection procedure, we propose the following approach. First, assume we are given a data set $(x_1, \ldots, x_n)$ of observations from n i.i.d. random variables $X_1, X_2, \ldots, X_n$, generated by $\boldsymbol{p}$ with the support $\{1, \ldots, k\}$ with $k < \infty$. Recall, that there are $S = 2^{k-1}$ candidate classes $\mathcal{F}^*_{k,\boldsymbol{w}}$ of the form (B.14) and among them there are $T = 2^{k-m}$ classes containing the true pmf $\boldsymbol{p}$.

Second, for a given data set we sort the candidate classes and obtain the sequence

$$\{\mathcal{F}^*_{k,\boldsymbol{w}_1}, \ldots, \mathcal{F}^*_{k,\boldsymbol{w}_S}\} \tag{B.42}$$

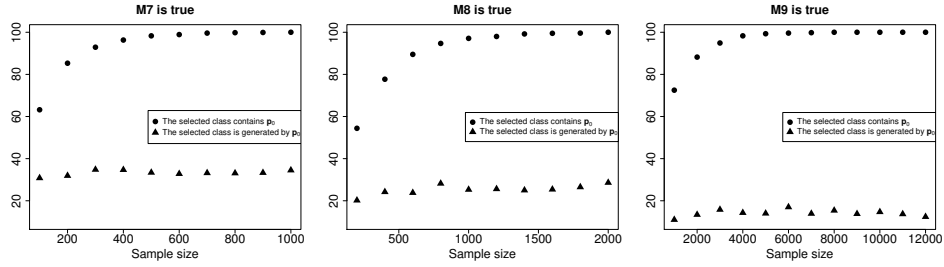

Figure B.2: Performance of $CMAIC$ for models **M7**, **M8** and **M9**. Percentage of times the correct model class was chosen versus the sample size n.

in ascending order of $CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}}, n)$, i.e. for $\mathcal{F}^*_{k,\boldsymbol{w}_i}$ in (B.42) one has

$$CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}_1}, n) < \cdots < CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}_S}, n).$$

Third, using Theorem 5.6 and (B.41), we have

$$\mathbb{P}[\liminf_{n\to\infty}\{\boldsymbol{p} \in \mathcal{F}^*_{k,\boldsymbol{w}_i}, \text{ for all } i \in \{1,\ldots,T\}\}] = 1.$$

Therefore, if the class $\mathcal{F}^*_{k,\boldsymbol{v}}$ is generated by the true pmf $\boldsymbol{p}$ then the following holds

$$\begin{aligned} \mathcal{F}^*_{k,\boldsymbol{v}} &\in \{\mathcal{F}^*_{k,\boldsymbol{w}_1}, \ldots, \mathcal{F}^*_{k,\boldsymbol{w}_T}\}, \\ \mathcal{F}^*_{k,\boldsymbol{v}} &\subset \mathcal{F}^*_{k,\boldsymbol{w}_i}, \text{ for all } i \in \{1,\ldots,T\}, \\ \mathcal{F}^*_{k,\boldsymbol{v}} &= \mathcal{F}^*_{k,\boldsymbol{w}_i}, \text{ for one of } i \in \{1,\ldots,T\} \end{aligned}$$

almost surely, as $n \to \infty$.

Then, in the model selection procedure instead of selecting the class $\mathcal{F}^*_{k,\boldsymbol{w}_1}$ which gives the smallest value of $CMAIC$, we choose the first class $\mathcal{F}^*_{k,\boldsymbol{w}'}$ from (B.42) (with the smallest index i) which satisfies

$$\begin{aligned} \mathcal{F}^*_{k,\boldsymbol{w}'} &\in \{\mathcal{F}^*_{k,\boldsymbol{w}_1}, \ldots, \mathcal{F}^*_{k,\boldsymbol{w}_{T'}}\}, \\ \mathcal{F}^*_{k,\boldsymbol{w}'} &\subset \mathcal{F}^*_{k,\boldsymbol{w}_i}, \text{for all } i \in \{1,\ldots,T'\} \end{aligned} \tag{B.43}$$

where $T' = 2^{k-m'}$ with $m = |\boldsymbol{w}'|$.

Recall, that if the true pmf $\boldsymbol{p}$ is strictly decreasing, then there is only one model class $\mathcal{F}^*_{k,\boldsymbol{w}}$ containing it. This class is the one generated by $\boldsymbol{w} = (1,1,\ldots,1)$.

Also, the model class $\mathcal{F}^*_{k,\boldsymbol{w}'}$ with $\boldsymbol{w}' = (k)$, i.e. the class with one constant region with the size of the support k, is contained in all classes $\mathcal{F}^*_{k,\boldsymbol{w}}$ of the form (B.14).

Therefore, if the first class in (B.42) is such that $\boldsymbol{w}_1 = (1, \ldots, 1)$ then we select it. Also, if in the model selection procedure we obtain that $\boldsymbol{w}' = (k)$, then, in order to avoid a possible misspecification, in this case we also select the model $\mathcal{F}^*_{k,\boldsymbol{w}_1}$, i.e. the one with the lowest $CMAIC$.

The next algorithm summarises the approach described above

Algorithm 5.1.

1. *Sort the candidate models* $\{\mathcal{F}^*_{k,\boldsymbol{w}_1}, \ldots, \mathcal{F}^*_{k,\boldsymbol{w}_S}\}$ *in an ascending order of* $CMAIC(\mathcal{F}^*_{k,\boldsymbol{w}_i}, n)$.
2. *If for* $\mathcal{F}^*_{k,\boldsymbol{w}_1}$ *the conditions in (B.43) are satisfied, then we select the class* $\mathcal{F}^*_{k,\boldsymbol{w}_1}$.
3. *If not, then we check for* $i = 2$ *and if for* $\mathcal{F}^*_{k,\boldsymbol{w}_2}$ *the conditions in (B.43) are satisfied, then we select the class* $\mathcal{F}^*_{k,\boldsymbol{w}_2}$.
4. *We repeat this procedure until the class* $\mathcal{F}^*_{k,\boldsymbol{w}_i}$ *satisfies (B.43).*
5. *If for* $i > 2$ *the selected class* $\mathcal{F}^*_{k,\boldsymbol{w}_i}$ *is such that* $\boldsymbol{w}_i = (k)$, *then we select the class* $\mathcal{F}^*_{k,\boldsymbol{w}_1}$, *i.e. the one with the lowest* $CMAIC$.

Note that if the first class $\mathcal{F}^*_{k,\boldsymbol{w}_1}$ is such that $\boldsymbol{w}_1 = (1, \ldots, 1)$, then $T' = 1$ and the conditions in (B.43) are, obviously, satisfied and, therefore, we select it. Also, in order to avoid the possible misspecification, we select the class $\mathcal{F}^*_{k,\boldsymbol{w}}$, generated by $\boldsymbol{w} = (k)$ only in the case when $\boldsymbol{w}_1 = (k)$, i.e. when it provides the lowest $CMAIC$.

Next, we note that though the procedure described in Algorithm 5.1 is conservative, there is no guarantee that the selected class $\mathcal{F}^*_{k,\boldsymbol{w}'}$ is exactly equal to the true one even asymptotically, i.e. in general

$$\mathbb{P}[\{\mathcal{F}^*_{k,\boldsymbol{w}'} = \mathcal{F}^*_{k,\boldsymbol{v}}\}] \quad \not\to \quad 1, \tag{B.44}$$

as $n \to \infty$. Therefore, the Algorithm 5.1 is also merely conservative, but not consistent in general. In the case of a strictly decreasing pmf $\boldsymbol{p}$ Algorithm 5.1 is however consistent.

Figure B.3 illustrates the performance of Algorithm 5.1 for the models **M7**, **M8** and **M9** for 1000 Monte Carlo simulations. One can see that using Algorithm

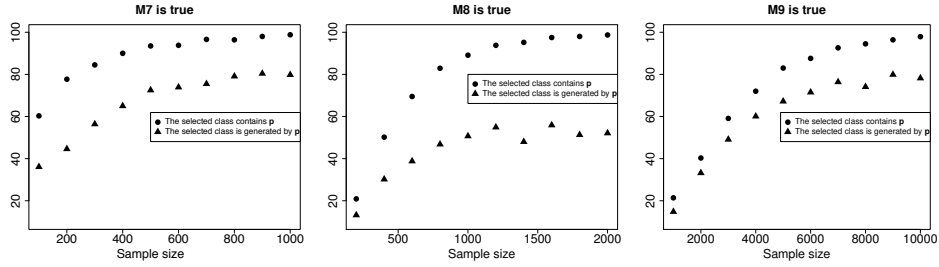

Figure B.3: Performance of the model selection procedure of Algorithm 5.1 for the models **M7**, **M8** and **M9**. Percentage of times versus the sample size.

5.1 seems to increase the asymptotic probability of selecting the exactly true class.

In the case of an infinite support of $\boldsymbol{p}$ we can use Algorithm 5.1, but with $\mathcal{F}^*_{k,\boldsymbol{w}}$ suitably changed to $\mathfrak{F}^*_{k,\boldsymbol{w}}$, defined in (B.37).

6 The asymptotic properties of the post-model-selection estimator $\hat{p}^*_n$

In this section, as apposed to the treatement in Section 4, we take into consideration that the model selection procedure is random, i.e. we let the candidate class depend on the data set and thus change with n.

Recall that the selected model $\hat{\boldsymbol{w}}_n$ is given in Definitions 5.3 and 5.5, for the finite and infinite cases, respectively. The post-model-selection estimator $\hat{\boldsymbol{p}}^*_n$ is then given by

$$\hat{\boldsymbol{p}}^*_n = \sum_{j=1}^{S} \hat{\boldsymbol{p}}^*_n(\boldsymbol{w}_j) 1\{\hat{\boldsymbol{w}}_n = \boldsymbol{w}_j\}, \tag{B.45}$$

where S is the total number of the candidate classes, i.e. $S = 2^{k-1}$ in the case of a finite support, $S = 2^{r-1}$ if k is large or infinite, and $\hat{\boldsymbol{p}}^*_n(\boldsymbol{w}_j) = \Pi(\hat{\boldsymbol{p}}_n | \mathcal{F}^*_{k,\boldsymbol{w}_j})$ denotes the estimator associated with the class $\mathcal{F}^*_{k,\boldsymbol{w}_j}$.

Theorem 6.1. *The post-model-selection estimator $\hat{\boldsymbol{p}}^*_n$ satisfies*

$$\mathbb{P}[\liminf_{n\to\infty} \{||\hat{\boldsymbol{p}}^*_n - \boldsymbol{p}||_2^2 \leq ||\hat{\boldsymbol{p}}^G_n - \boldsymbol{p}||_2^2\}] \quad = \quad 1.$$

Finally, we have the following corollary result for the risk of the estimator $\hat{\boldsymbol{p}}_n^*$.

Corollary 6.1. *For any decreasing pmf $\boldsymbol{p}$ there exists n_1 such that for all $n > n_1$ one has*

$$\mathbb{E}[n||\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}||_2^2] \leq \mathbb{E}[n||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2^2].$$

7 Comparison of the estimators and discussion

We have established limit properties for a new, model selection based, estimator of a monotone pmf, as well as performed a simulation study to assess its finite sample properties. There are two main reasons for our proposal of an estimator. Firstly, as we have established in this paper, knowledge about the shape of the distribition in terms of the levels of constancy increases the accuracy, i.e. one obtains a smaller risk compared to the unmodified order restricted mle, which is the Grenander estimator. Secondly, the limit distribution of the Grenander estimator depends on the shape of the pmf, in a very information dependent way, namely if the point of interest lies in a flat region for the pmf then the limit distribution at that point is obtained as a pooled adjacent violators (PAVA) algorithm of a Gaussian vector, whereas if the point of interest instead lies in a region where the estimand is strictly monotone then the limit distribution is Gaussian and is identical to the limit distribution of the unrestricted mle.

The estimator we have presented consists of a two step procedure. In the first step we introduce a novel information criterion, CMAIC, for model selection in order restricted inference. The second step, which is the estimation step for the pmf in the selected model class, uses a novel approach consisting of utilizing information about the shape of the pmf in a way that increases the precision of the estimate. The limit distribution for the final estimator is derived using recent results on pmf estimation for functions that are monotone with respect to a quasi order, cf. [3].

As noted in [3], the limit distributions in [3] are general and potentially applicable to arbitrary dependence structures. We believe that the approach in this paper is also potentially applicable to the estimation of a monotone pmf which is the marginal distribution of a stationary process, based on observations of that process. The resulting estimator will then not be an mle, but instead a marginal or partional mle. Limit results for such a model selection

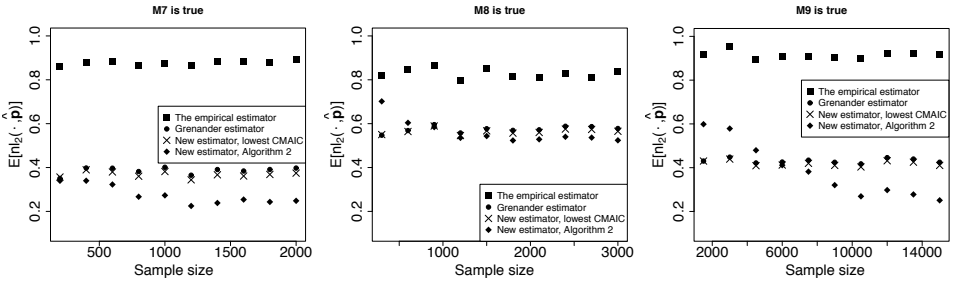

Figure B.4: The estimates of the normalised risk $\mathbb{E}[nl_2^2(\cdot, \boldsymbol{p})]$ for the estimators: the empirical estimator $\hat{\boldsymbol{p}}_n$, Grenander estimator $\hat{\boldsymbol{p}}_n^G$ and the new estimator $\hat{\boldsymbol{p}}_n^*$ for a model selection by the lowest value of $CMAIC$ and with Algorithm 5.1 for the models **M7**, **M8** and **M9**.

based estimator in the dependent data setting will then depend on extending the results in [3] to dependent data, as well as on extending the results on the model selection step to such data. This should be an interesting topic for future research.

For a visualisation of the performance of the proposed estimator $\hat{\boldsymbol{p}}_n^*$, we make a simulation study. Figure B.4 illustrates the performance of the estimator $\hat{\boldsymbol{p}}_n^*$ for the cases when the selected model is the one with the lowest value of $CMAIC(\mathcal{F}_{k,\boldsymbol{w}_1}^*, n)$ and when the selected model is obtained by Algorithm 5.1 for 1000 Monte Carlo simulations.

Figure B.5 illustrates the performance of the estimator $\hat{\boldsymbol{p}}_n^*$ for the cases when the selected model is the one with the lowest value of $CMAIC(\mathcal{F}_{k,\boldsymbol{w}_1}^*, n)$ and when the selected model is obtained by Algorithm 5.1 for 1000 Monte Carlo simulations.

The simulation study clearly illustrates that the model selection based approach has a better asymptotic performance, in l^2-sense, than both the empirical estimator and the Grenander estimator. Also, one can see that using Algorithm 5.1 for a models selection (the point 1 of Algorithm 3.1) seems to give a smaller asymptotic risk than in the case when the model selection is performed by choosing the first class $\mathcal{F}_{k,\boldsymbol{w}_1}^*$ from (B.42), i.e. the one with the lowest $CMAIC$.

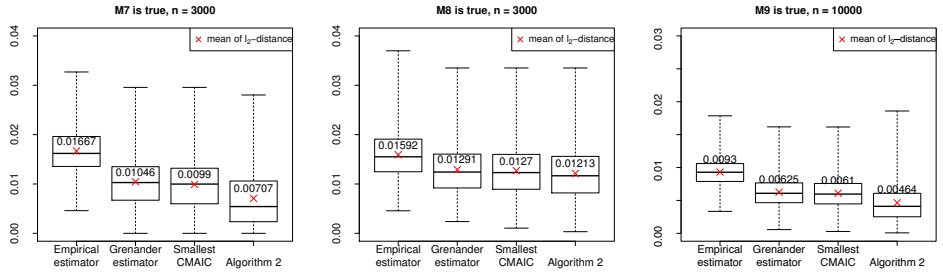

Figure B.5: The boxplots for l^2-distances for the estimators: the empirical estimator $\hat{p}_n$, Grenander estimator $\hat{p}_n^G$ and the new estimator $\hat{p}_n^*$ for a model selection by the lowest value of $CMAIC$ and with Algorithm 5.1. The numbers are the estimates of the means of l^2-distances between the estimates and p for the models **M7**, **M8** and **M9**.

8 Appendix

We first introduce some notations and state some results for the isotonic regression over a general preordered set, cf. [3]. It is possible to derive the limit properties for the proposed estimator by tracing the steps in the algorithm for its calculation, and proving that in each step one gets both consistency and limit distributions results. It is however more straightforward to derive limit properties for the proposed estimator with the use of some recent results on limit properties for isotonic regression of functions that are ordered with respect to a preorder, that were derived in [3].

Consider a set of indices $\mathcal{I} = \{1, \ldots, \infty\}$ with some preorder $\preceq$ defined on it. The preorder is assumed to be arbitrary, and we will below introduce particular preorders that are relevant for the problem at hand. Then a vector f^* is called the isotonic regression of an arbitrary vector $f \in l^2$ over the preordered index set $\mathcal{I}$ if

$$f^* = \underset{\xi \in \mathbf{\Theta}^{is}}{\operatorname{argmin}} \sum_{i \in \mathcal{I}} (\xi_i - f_i)^2,$$

where $\mathbf{\Theta}^{is}$ denotes the set of all isotonic vectors in l^2 with respect to the preorder $\preceq$.

Let $f \in \mathbf{\Theta}^{is}$ be an arbitrary but fixed vector, and assume that it satisfies the assumptions on p in Assumption 3.1. Then, for an arbitrary but fixed integer $a < m$ we may partition the original index set $\mathcal{I}$ in the following way

$$\mathcal{I} = \cup_{j=1}^{a+1} \mathcal{I}_{(j)}, \tag{B.46}$$

where the set $\mathcal{I}_{(j)}$ contains the indices of the j-th constant region of $\boldsymbol{f}$, i.e. $\mathcal{I}_{(j)} = \{q_j, \dots, q_j + v_j - 1\}$, for each $j \leq a$, and $\mathcal{I}_{(a+1)} = \{q_{a+1}, q_{a+1} + 1, \dots\}$.

Next, let $[\boldsymbol{f}]^{\mathcal{I}_{(j)}}$ denote the restriction of the vector $\boldsymbol{f} \in l^2$ to the j-th index set in the partition (B.46). We introduce an operator $\psi_a : l^2 \to l^2$, defined in the following way. For any vector $\boldsymbol{f} \in l^2$, the operator values of $\psi(\boldsymbol{f})$ on each index $\mathcal{I}_{(j)}$ set in (B.46) are given by

$$[\psi_a(\boldsymbol{f})]^{\mathcal{I}_{(j)}} = isot\{[\boldsymbol{f}]^{\mathcal{I}_{(j)}}\}, \tag{B.47}$$

where $isot\{[\boldsymbol{f}]^{\mathcal{I}_{(j)}}\}$ denotes the isotonic regression of the restriction of the vector $\boldsymbol{f} \in l^2$ to the index set $\mathcal{I}^{(j)}$ in the partition (B.46) with respect to the preorder $\preceq$. Therefore, $\psi_a(\boldsymbol{f})$ is a concatenation of the separate isotonic regressions, with respect to the preorder $\preceq$, of the restrictions of $\boldsymbol{f}$ to the index sets $\mathcal{I}_{(j)}$, for $j = 1, \dots, a+1$.

We next introduce an appropriate preorder. Consider a candidate class $\mathcal{F}^*_{k,\boldsymbol{w}} = \mathcal{F}^*_{k,\boldsymbol{w},r}$, defined in (B.15), and assumed to contain $\boldsymbol{p}$. The class $\mathcal{F}^*_{k,\boldsymbol{w}}$ generates a preorder $\preceq$ on the index set $\mathcal{I} = \{1, 2, \dots\}$, as follows. We define the order relation $\preceq$ on $\mathcal{I}$ by specifying that if $i_1, i_2 \in \mathcal{I}$ and the indices belong to different constant regions of the class $\mathcal{F}^*_{k,\boldsymbol{w}}$, then if $i_1 > i_2$ we let $i_1 \preceq i_2$. Furthermore if $i_3, i_4 \in \mathcal{I}$ are in the same constant region we specify, unconditionally, that $i_3 \preceq i_4$, and with that specification of course also follows that $i_4 \preceq i_3$. Therefore, when $i_3, i_4 \in \mathcal{I}$ are different indices belonging to the same constant region of $\mathcal{F}^*_{k,\boldsymbol{w}}$, i.e. $i_3, i_4 \in \{t_j, t_j + w_j - 1\}$, for some $j = 1, \dots, d$, we have both $i_3 \preceq i_4$ and $i_4 \preceq i_3$. Thus, the order relation $\preceq$ on $\mathcal{I}$ is not antisymmetric; it is however transitive, and thus it is a preorder, cf. also [5, 17, 19]. To emphasize that the preorder is generated by $\mathcal{F}^*_{k,\boldsymbol{w}}$, or equivalently by w, we denote it by $\preceq_{\boldsymbol{w}}$.

With this definition we see that for any vector $\boldsymbol{f} \in l^2$, $\boldsymbol{f} \in \mathcal{F}^*_{k,\boldsymbol{w}}$ if and only if it is isotonic with respect to the preorder $\preceq_{\boldsymbol{w}}$ on $\mathcal{I}$. Therefore, since From this equivalence the next result immediately follows.

Lemma 8.1. *The estimator $\hat{\boldsymbol{p}}^*_n = \Pi(\hat{\boldsymbol{p}}_n | \mathcal{F}^*_{k,\boldsymbol{w}})$ is identical to the isotonic regression of the empirical estimator $\hat{\boldsymbol{p}}_n$ with respect to the preorder $\preceq_{\boldsymbol{w}}$ on the index set $\mathcal{I}$, i.e.*

$$\hat{\boldsymbol{p}}^*_n \quad = \quad \underset{\boldsymbol{\xi} \in \boldsymbol{\Theta}^{is}}{\operatorname{argmin}} \sum_{i \in \mathcal{I}} (\xi_i - \hat{p}_{n,i})^2,$$

where $\boldsymbol{\Theta}^{is}$ denotes the set of all isotonic vectors in l^2 with respect to the preorder $\preceq_{\boldsymbol{w}}$.

We note also that the equivalence mentioned gives a second characterization of the operator value $\psi_a(\boldsymbol{f})$ of an $\boldsymbol{f} \in l^2$, namely if the preorder is $\preceq_{\boldsymbol{w}}$ as defined above and ψ_a is the corresponding operator, then

$$\psi_a(\boldsymbol{f}) \quad = \quad \Pi(\boldsymbol{f}|\mathcal{G}^*_{k,\boldsymbol{w},p}). \tag{B.48}$$

Indeed, the cone $\mathcal{G}^*_{k,\boldsymbol{w},p}$, which is defined in (B.27), consists of the Cartesian product of the cones in $\mathcal{F}^*_{k,w}$ over the regions of constancy of $\boldsymbol{p}$ but without any constraints in between the pairs of the elements (f_{v_1}, f_{v_1+1}), $(f_{v_1+v_2}, f_{v_1+v_2+1}), \ldots$ $(f_{\sum_{j=1}^{k-1} v_j}, f_{\sum_{j=1}^{s-1} v_j+1})$. Therefore, the result in (B.48) follows by a simple partition of the total sum of squares into sum of squares over each factor in the Cartesian product, cf. Lemma 7 in [3] for the strict proof.

We first recall a simple result from [3]. Define

$$\varepsilon = \inf\{|\tilde{f}_{l'} - \tilde{f}_l| : l' \in \{1, \ldots, a\}, l \in \{1, \ldots, m\}\}, \tag{B.49}$$

where $\tilde{f}_l$ is the constant value of $\boldsymbol{f}$ on the l'th constant region.

Lemma 8.2. *Consider an arbitrary vector $\hat{\boldsymbol{f}} \in l^2$ and an index set $\mathcal{I}$ with a preorder $\preceq$ defined on it. Suppose that ε is defined in (B.49). If for $\hat{\boldsymbol{f}}$ one has*

$$\sup_{i \in \mathcal{I}} \{|\hat{f}_i - f_i|\} < \varepsilon/2,$$

then the isotonic regression of $\hat{\boldsymbol{f}}$ is given by $\psi_a(\hat{\boldsymbol{f}})$, i.e.

$$\hat{\boldsymbol{f}}^* = \psi_a(\hat{\boldsymbol{f}}).$$

Therefore, the isotonic regression $\hat{\boldsymbol{f}}^*$ of $\hat{\boldsymbol{f}}$ is a concatenation of the separate isotonic regressions, with respect to the preorder $\preceq$, of the restrictions of $\hat{\boldsymbol{f}}$ to the index sets $\mathcal{I}_{(1)}, \mathcal{I}_{(2)}, \ldots, \mathcal{I}_{(a)}$ and $\mathcal{I}_{(a+1)}$, defined in (B.46).

Proof. The statement of the lemma follows from Lemma 8 in [3]. □

Lemma 8.3. *Let $\mathcal{A}$ and $\mathcal{B}$ be closed convex cones in $\mathbb{R}^k$, for $k \leq \infty$, and $\mathcal{A} \subset \mathcal{B}$. Then for any $\boldsymbol{y} \in \mathbb{R}^k$,*

$$\begin{aligned} ||\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{A})||_2 &\geq ||\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{B})||_2, \\ ||\Pi(\boldsymbol{y}|\mathcal{A})||_2 &\leq ||\Pi(\boldsymbol{y}|\mathcal{B})||_2, \end{aligned}$$

where $||\cdot||_2$ is the l^2-norm.

Proof. Note, that for any $\boldsymbol{y} \in \mathbb{R}^k$ the following equalities hold

$$\begin{aligned} ||\boldsymbol{y}||_2^2 &= ||\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{A})||_2^2 + ||\Pi(\boldsymbol{y}|\mathcal{A})||_2^2, \\ ||\boldsymbol{y}||_2^2 &= ||\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{B})||_2^2 + ||\Pi(\boldsymbol{y}|\mathcal{B})||_2^2, \end{aligned}$$

cf. Proposition 3.4.1 in [19]. Next, since $\mathcal{A} \subset \mathcal{B}$, one has $||\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{B})||_2^2 \leq ||\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{A})||_2^2$ and this proves $||\Pi(\boldsymbol{y}|\mathcal{A})||_2 \leq ||\Pi(\boldsymbol{y}|\mathcal{B})||_2$. □

Proof of Theorem 4.1. The strong consistency of the estimator $\hat{\boldsymbol{p}}_n^*$ follows from the continuous mapping theorem.

To prove the limit distribution result, let us consider a candidate class $\mathcal{F}_{k,\boldsymbol{w}}^*$ defined in (B.15), and the preorder $\preceq_{\boldsymbol{w}}$ generated by it. We established in Lemma 8.1 that the estimator $\hat{\boldsymbol{p}}_n^* = \Pi(\hat{\boldsymbol{p}}_n|\mathcal{F}_{k,\boldsymbol{w}}^*)$, is equal to the isotonic regression of the empirical estimator $\hat{\boldsymbol{p}}_n$ with respect to the preorder $\preceq_{\boldsymbol{w}}$ on the index set $\mathcal{I}$.

Therefore, the limit distribution result follows from Theorem 3 in [3], in which we established the asymptotic distribution of the isotonized estimator, over a general countable preordered set, and proved the limit distribution result

$$\sqrt{n}(\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}) \quad \xrightarrow{d} \quad \psi_a(Y_{\mathbf{0},B}),$$

where $Y_{\mathbf{0},B}$ is the weak limit of $\sqrt{n}(\hat{\boldsymbol{p}}_n - \boldsymbol{p})$, which, noting the characterization (B.48), can be written as

$$\sqrt{n}(\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}) \quad \xrightarrow{d} \quad \Pi(Y_{\mathbf{0},B}|\mathcal{G}_{k,\boldsymbol{w},p}^*).$$

□

We note that the limit distribution result in Theorem 4.1 for the finite support case was derived in Theorem 5.2.1 in [17].

Acknowledgements

VP's research is fully supported and DA's research is partially supported by the Swedish Research Council, whose support is gratefully acknowledged.

Supplementary material.

Proof of Theorem 4.2.

Analogously to the proof of Theorem 4.1, let us consider a candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$ as defined in (B.15) and containing $\boldsymbol{p}$. Furthermore, if we let $\tilde{\boldsymbol{w}} = (1, 1, \dots)$ then we obtain the candidate class $\mathcal{F}^*_{k,\tilde{\boldsymbol{w}}}$, defined as,

$$\mathcal{F}^*_{k,\tilde{\boldsymbol{w}}} = \Big\{ \boldsymbol{f} \in l^2 : f_1 \geq f_2 \geq \dots \Big\}. \tag{B.50}$$

Then the regular Grenander estimator $\hat{\boldsymbol{p}}^G_n$, which is

$$\begin{aligned} \hat{\boldsymbol{p}}^G_n &= \Pi(\hat{\boldsymbol{p}}_n | \mathcal{F}^*_{k,\tilde{\boldsymbol{w}}}) \\ &= \underset{f \in \mathcal{F}^*_{k,\tilde{\boldsymbol{w}}}}{\operatorname{argmin}} \sum_i [\hat{p}_{n,i} - f_i]^2, \end{aligned} \tag{B.51}$$

can, with the use of Lemma 8.1, equivalently be viewed as the isotonic regression of the empirical estimator $\hat{\boldsymbol{p}}_n$ with respect to the preorder $\preceq_{\tilde{\boldsymbol{w}}}$ generated by the class $\mathcal{F}^*_{k,\tilde{\boldsymbol{w}}}$, defined in (B.50). The preorder $\preceq_{\tilde{\boldsymbol{w}}}$ on $\mathcal{I}$ is simply a reverse order on the integers: for any $i_1, i_2 \in \mathcal{I}$, let $i_1 \preceq_{\tilde{\boldsymbol{w}}} i_2$ if $i_1 > i_2$. Thus we have a preorder, associated with $\mathcal{F}^*_{k,\boldsymbol{w}}$, which we denote by $\preceq_{\boldsymbol{w}}$ and a preorder, associated with $\mathcal{F}^*_{k,\tilde{\boldsymbol{w}}}$, which we denote by $\preceq_{\tilde{\boldsymbol{w}}}$.

Next, for some integer $a > r$, where r is the number of flat regions in candidate class $\mathcal{F}^*_{k,\boldsymbol{w}}$, we make a partition of the index set $\mathcal{I}$ as in (B.46) Let ε be defined as in (B.49), with $\boldsymbol{f}$ replaced by $\boldsymbol{p}$. Since the empirical estimator $\hat{\boldsymbol{p}}_n$ is strongly consistent, there exists an integer n_1 such that for all $n > n_1$ one has

$$\sup_{i \in \mathcal{I}} \{ |\hat{p}_{n,i} - p_i| \} < \varepsilon/2,$$

almost surely. Therefore, using Lemma 8.2, if $n > n_1$, we obtain

$$\begin{aligned} \hat{\boldsymbol{p}}^*_n &\overset{a.s.}{=} \psi^{\boldsymbol{w}}_a(\hat{\boldsymbol{p}}_n), \\ \hat{\boldsymbol{p}}^G_n &\overset{a.s.}{=} \psi^{\tilde{\boldsymbol{w}}}_a(\hat{\boldsymbol{p}}_n), \end{aligned} \tag{B.52}$$

where the operator ψ_a is defined in (B.47). Here we use the upper-scripts $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ to emphasise which preorder that the isotonic regression is with respect to, or equivalently, which preorder the operator ψ_a is associated to.

Next, we compare $||\hat{\boldsymbol{p}}^*_n(\boldsymbol{w}) - \boldsymbol{p}||_2$ with $||\hat{\boldsymbol{p}}^G_n - \boldsymbol{p}||_2$. We do this by a comparison separately on each partition set in (B.46). From (B.52), it follows that if $n > n_1$

then

$$\begin{aligned} ||\hat{\boldsymbol{p}}_n^*(\boldsymbol{w}) - \boldsymbol{p}||_2 &\overset{a.s.}{=} ||\psi_a^{\boldsymbol{w}}(\hat{\boldsymbol{p}}_n) - \boldsymbol{p}||_2, \\ ||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2 &\overset{a.s.}{=} ||\psi_a^{\tilde{\boldsymbol{w}}}(\hat{\boldsymbol{p}}_n) - \boldsymbol{p}||_2. \end{aligned}$$

First, we note that the separate isotonic regression on the last partition set in (B.46) is identical for the two preorders, i.e.

$$\begin{aligned} [\psi_a^{\boldsymbol{w}}(\hat{\boldsymbol{p}}_n)]^{\mathcal{I}_{(a+1)}} &= [\psi_a^{\tilde{\boldsymbol{w}}}(\hat{\boldsymbol{p}}_n)]^{\mathcal{I}_{(a+1)}} \\ &= \underset{Y_1 \geq Y_2 \geq}{\operatorname{argmin}} \sum_{i=0}^{\infty} (Y_i - \hat{p}_{n,i+q_{a+1}})^2. \end{aligned}$$

Second, without loss of generality, assume the true pmf $\boldsymbol{p}$ starts with a constant region, so that the first constant region of $\boldsymbol{p}$ has indices $(1, \ldots, v_1)$ with $v_1 > 1$. Now, since $\boldsymbol{p}$ belongs to the class $\mathcal{F}_{k,\boldsymbol{w}}^*$, there must be $k_1 \geq 1$ regions of constancy in $\mathcal{F}_{k,\boldsymbol{w}}^*$ whose union is $(1, \ldots, v_1)$, i.e. such that $v_1 = \sum_{i=1}^{k_1} w_i$, where $(w_1, \ldots, w_{k_1})$ are the lengths of the first k_1 constant regions in the class $\mathcal{F}_{k,\boldsymbol{w}}^*$, cf. also the discussion after (B.17).

Since $\boldsymbol{p}$ is constant on $\{1, \ldots, v_1\}$, by the use of Theorem 1.8 in [5] and Lemma 8.1, we have

$$\begin{aligned} [\psi_a^{\boldsymbol{w}}(\hat{\boldsymbol{p}}_n)]^{(1,v_1)} - [\boldsymbol{p}]^{(1,v_1)} &= [\psi_a^{\boldsymbol{w}}(\hat{\boldsymbol{p}}_n - \boldsymbol{p})]^{(1,v_1)} \\ &= \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \sum_{i=1}^{v_1} (Z_i - (\hat{p}_{n,i} - p_i))^2 \end{aligned}$$

where $\mathcal{C} = [\mathcal{F}_{k,\boldsymbol{w}}^*]^{(1,v_1)} \subset \mathbb{R}^{v_1}$ is the cone consisting of the first v_1 coordinates of the candidate class $\mathcal{F}_{k,\boldsymbol{w}}^*$, defined in (B.15). Similarly

$$[\psi_a^{\tilde{\boldsymbol{w}}}(\hat{\boldsymbol{p}}_n)]^{(1,v_1)} - [\boldsymbol{p}]^{(1,v_1)} = \underset{Z \in \tilde{\mathcal{C}}}{\operatorname{argmin}} \sum_{i=1}^{v_1} (Z_i - (\hat{p}_{n,i} - p_i))^2$$

where $\tilde{\mathcal{C}} = [\mathcal{F}_{k,\tilde{\boldsymbol{w}}}^*]^{(1,v_1)} \subset \mathbb{R}^{v_1}$ is the cone consisting of the first v_1 coordinates of the candidate class $\mathcal{F}_{k,\tilde{\boldsymbol{w}}}^*$, defined in (B.50). Since $\mathcal{C} \subset \tilde{\mathcal{C}}$, from Lemma 8.3 it follows that

$$||[\psi_a^{\boldsymbol{w}}(\hat{\boldsymbol{p}}_n)]^{(1,v_1)} - [\boldsymbol{p}]^{(1,v_1)}||_2 \leq ||[\psi_a^{\tilde{\boldsymbol{w}}}(\hat{\boldsymbol{p}}_n)]^{(1,v_1)} - [\boldsymbol{p}]^{(1,v_1)}||_2.$$

Similarly, for every constant region of the true pmf $\boldsymbol{p}$, up to the $(a+1)$-th, so for every $(q_j, q_j + v_j - 1)$ with $j < a+1$, one can prove that

$$||[\psi_a^{\boldsymbol{w}}(\hat{\boldsymbol{p}}_n)]^{(q_j,q_j+v_j-1)} - [\boldsymbol{p}]^{(q_j,q_j+v_j-1)}||_2 \leq \\ ||[\psi_a^{\tilde{\boldsymbol{w}}}(\hat{\boldsymbol{p}}_n)]^{(q_j,q_j+v_j-1)} - [\boldsymbol{p}]^{(q_j,q_j+v_j-1)}||_2.$$

Therefore, we have proved that for all $n > n_1$

$$||\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}||_2 \overset{a.s.}{\leq} ||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2 \tag{B.53}$$

□

Proof of Theorem 5.1. The proof is done in three steps.

Step 1. We obtain the asymptotic distribution of the second term in (B.31). The proof is reminiscent to the one of Theorem 5.2.1 in [17]. In fact, using Theorem 4.1 for $\boldsymbol{V}_n$, we have

$$\boldsymbol{V}_n \quad \overset{d}{\to} \quad \Pi(Y_{\boldsymbol{0},\boldsymbol{B}}|\mathcal{G}^*_{k,\boldsymbol{w},p}), \tag{B.54}$$

where $Y_{\boldsymbol{0},\boldsymbol{B}} \overset{d}{=} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B})$, with the covariance matrix $\boldsymbol{B}_{ij} = \delta_{ij}p_i - p_i p_j$, $\mathcal{G}^*_{k,\boldsymbol{w},p}$ is the cone defined in (B.27), and where $\boldsymbol{v} = (v_1, \ldots, v_s)$ is the vector of the lengths of the regions of constancy of the pmf $\boldsymbol{p}$.

Then, using the continuous mapping theorem, together with Lemma A of Theorem 5.2.1 in [17], we obtain

$$\boldsymbol{V}_n^T \boldsymbol{P}^{-1} \boldsymbol{V}_n \quad \overset{d}{\to} \quad \mathcal{V},$$

where

$$\mathcal{V} \quad = \quad \Pi(\boldsymbol{Z}|\mathcal{G}^*_{k,\boldsymbol{w},p})^T \boldsymbol{P} \Pi(\boldsymbol{Z}|\mathcal{G}^*_{k,\boldsymbol{w},p}).$$

where $\boldsymbol{Z} = (\boldsymbol{U} - \bar{\boldsymbol{U}})$, with $\boldsymbol{U}$ a normal vector with covariance matrix $\boldsymbol{P}^{-1}$ and $\bar{\boldsymbol{U}} = \sum_{j=1}^k p_j U_j$. The distribution of $\mathcal{V}$ is given in Theorem 5.2.1 in [17], namely

$$\mathbb{P}[\mathcal{V} \leq v] = \sum_{j=1}^{k} P(\mathcal{G}^*_{k,\boldsymbol{w},p}, j, k) \mathbb{P}[\chi^2_{j-1} \leq v], \tag{B.55}$$

for any real number v, where χ^2_j is a chi-square random variable with j degrees of freedom, $\chi^2_0 \equiv 0$, and $P(\mathcal{G}^*_{k,\boldsymbol{w},p}, j, k)$ is the probability that the projection of a

standard normal k-dimensional vector on the cone $\mathcal{G}^*_{k,\boldsymbol{w},p}$ has j distinct values, for $j = 1, \ldots, k$, cf. [5, 17, 19]. Furthermore, from Proposition 3.6.1.9 in [19] it follows that $P(\mathcal{G}^*_{k,\boldsymbol{w},p}, j, k) = 0$ for all $j > s$, where $s = |\boldsymbol{w}|$.

Step 2. We prove the statement (B.33). Recall, that $\boldsymbol{V}_n = \sqrt{n}(\hat{\boldsymbol{p}}^*_n - \boldsymbol{p})$. Then from the strong consistency of $\hat{\boldsymbol{p}}_n$, cf. Theorem 5.2.1 in [17], it follows that there exists n_1 such that for all $n > n_1$,

$$\boldsymbol{V}_n \;=\; \sqrt{n}\Pi(\hat{\boldsymbol{p}}_n|\mathcal{G}^*_{k,\boldsymbol{w},p}) - \boldsymbol{p}),$$

almost surely. Therefore, for $n > n_1$, using the reduction of error property of isotonic regression (Theorem 7.6 in [5]) one has

$$\boldsymbol{V}_n^T \boldsymbol{P}^{-1} \boldsymbol{V}_n \;\leq\; \sqrt{n}(\hat{\boldsymbol{p}}_n - \boldsymbol{p})^T \boldsymbol{P}'^{-1} \sqrt{n}(\hat{\boldsymbol{p}}_n - \boldsymbol{p}), \tag{B.56}$$

almost surely. Next, since the right hand side of (B.56) is asymptotically uniformly integrable, then $\boldsymbol{V}_n^T \boldsymbol{P}^{-1} \boldsymbol{V}_n$ is also asymptotically uniformly integrable. The statement in (B.33) now follows from Theorem 2.20 in [20].

Step 3.

The final statement of the theorem, (B.34), is proved in the discussion after Theorem 5.2.1 in [17]. We detail some parts of the proof in [17] below, in our notation. The limit random variable was shown in [17] to be equal in distribution to a more conducive expression, namely

$$\mathcal{V} \;\stackrel{d}{=}\; ||\Pi(\boldsymbol{Y}|\mathcal{G}^*_{k,\boldsymbol{w},p}) - \bar{\boldsymbol{Y}}||_2^2,$$

where $\boldsymbol{Y}$ is a standard normal k-dimensional vector and $\bar{\boldsymbol{Y}} = k^{-1} \sum_{j=1}^k Y_j$. Note that

$$\mathcal{F}^*_{k,w} \subseteq \mathcal{G}^*_{k,\boldsymbol{w},p} \subseteq \mathcal{A}^*_{k,\boldsymbol{w}} \tag{B.57}$$

where $\mathcal{A}^*_{k,\boldsymbol{w}}$ is the cone

$$\begin{aligned} \mathcal{A}^*_{k,\boldsymbol{w}} \;=\; \Big\{ \boldsymbol{f} \in \mathbb{R}^k :\; & f_1 = \cdots = f_{w_1}, \\ & f_{t_2} = \cdots = f_{t_2+w_2-1}, \ldots, f_{t_s} = \cdots = f_k \Big\}. \end{aligned}$$

Then, since for any of the choices $\mathcal{C} = \mathcal{F}^*_{k,w}, \mathcal{G}^*_{k,\boldsymbol{w},p}$ or $\mathcal{A}^*_{k,\boldsymbol{w}}$, we have

$$||\boldsymbol{Y} - \bar{\boldsymbol{Y}}||_2^2 \;=\; ||\boldsymbol{Y} - \Pi(\boldsymbol{Y}|\mathcal{C})||_2^2 + ||\Pi(\boldsymbol{Y}|\mathcal{C}) - \bar{\boldsymbol{Y}}||_2^2$$

and since (B.57) implies an ordering of $||\boldsymbol{Y} - \Pi(\boldsymbol{Y}|\mathcal{C})||_2^2$, for the three choices for $\mathcal{C}$, we obtain, almost surely,

$$\begin{aligned} ||\Pi(\boldsymbol{Y}|\mathcal{F}^*_{k,w}) - \bar{\boldsymbol{Y}}||_2^2 &\leq ||\Pi(\boldsymbol{Y}|\mathcal{G}^*_{k,\boldsymbol{w},p}) - \bar{\boldsymbol{Y}}||_2^2 \leq \\ &\leq ||\Pi(\boldsymbol{Y}|\mathcal{A}^*_{k,\boldsymbol{w}}) - \bar{\boldsymbol{Y}}||_2^2. \end{aligned} \tag{B.58}$$

Next, for any choice $\mathcal{C} = \mathcal{F}^*_{k,w}, \mathcal{G}^*_{k,\boldsymbol{w},p}, \mathcal{A}^*_{k,\boldsymbol{w}}$, we have

$$\begin{aligned} ||\Pi(\boldsymbol{Y}|\mathcal{C}) - \bar{\boldsymbol{Y}}||_2^2 &\stackrel{d}{=} \sum_{j=1}^{k} P(\mathcal{C}, j, k)\mathbb{P}[\chi^2_{j-1} \leq v], \\ P(\mathcal{C}, j, k) &= 0, \text{ for all } j > s. \end{aligned}$$

Note also the $P(\mathcal{A}^*_{k,\boldsymbol{w}}, j, k) = 0$ for $j \neq s$ and $P(\mathcal{A}^*_{k,\boldsymbol{w}}, s, k) = 1$. Therefore, $||\Pi(\boldsymbol{Y}|\mathcal{A}^*_{k,\boldsymbol{w}}) - \bar{\boldsymbol{Y}}||_2^2$ is distributed as a χ^2_{s-1} random variable. This finally shows, by takings expectations of the expression (B.58), that

$$\sum_{j=1}^{s} P(\mathcal{F}^*_{k,\boldsymbol{w}}, j, k)(j-1) \leq \mathbb{E}[\mathcal{V}] \leq s - 1,$$

which ends the proof.

□

Proof of Theorem 5.6. Let $k < \infty$ and $\mathcal{F}^*_{k,\boldsymbol{w_1}}$ and $\mathcal{F}^*_{k,\boldsymbol{w_2}}$ be two classes, such that $\boldsymbol{p} \in \mathcal{F}^*_{k,\boldsymbol{w_1}}$ and $\boldsymbol{p} \notin \mathcal{F}^*_{k,\boldsymbol{w_2}}$. Next, let $\hat{\boldsymbol{p}}^*_n(\boldsymbol{w}_1) = \Pi(\hat{\boldsymbol{p}}_n|\mathcal{F}^*_{k,\boldsymbol{w_1}})$ and $\hat{\boldsymbol{p}}^*_n(\boldsymbol{w}_2) = \Pi(\hat{\boldsymbol{p}}_n|\mathcal{F}^*_{k,\boldsymbol{w_2}})$.

First, using the continuous mapping theorem, we have

$$\begin{aligned} (\hat{\boldsymbol{p}}_n, \hat{\boldsymbol{p}}^*_n(\boldsymbol{w}_1), \hat{\boldsymbol{p}}^*_n(\boldsymbol{w}_2))^T &\stackrel{a.s.}{\to} (\boldsymbol{p}, \Pi(\boldsymbol{p}|\mathcal{F}^*_{k,\boldsymbol{w_1}}), \Pi(\boldsymbol{p}|\mathcal{F}^*_{k,\boldsymbol{w_2}}))^T \\ &= (\boldsymbol{p}, \boldsymbol{p}, \tilde{\boldsymbol{p}})^T \end{aligned} \tag{B.59}$$

and $\tilde{\boldsymbol{p}} \neq \boldsymbol{p}$, by the almost sure consistency of $\hat{\boldsymbol{p}}_n$ and since projection on any $\mathcal{F}^*_{k,\boldsymbol{w}}$ is a continuous map.

Next, note that for the bias correction term $B(\boldsymbol{w})$, defined in (B.35), the following holds

$$\frac{B(\boldsymbol{w})}{n} \to 0$$

as $n \to \infty$.

Therefore, the statement of the theorem holds if

$$\frac{l(x_1, \ldots, x_n | \hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_1))}{n} - \frac{l(x_1, \ldots, x_n | \hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_2))}{n} \xrightarrow{a.s.} c > 0,$$

where $l(x_1, \ldots, x_n | \boldsymbol{f})$ is the log-likelihood defined in (B.28).

By the almost sure consistency result (B.59) and since the log-likelihood $l(x_1, \ldots, x_n | \cdot)$ is a continuous map, we get

$$\begin{aligned} &\frac{l(x_1, \ldots, x_n | \hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_1))}{n} - \frac{l(x_1, \ldots, x_n | \hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_2))}{n} \\ \xrightarrow{a.s.} &\sum_{i=1}^{k} p_i \log p_i - \sum_{i=1}^{k} p_i \log \tilde{p}_i, \end{aligned} \tag{B.60}$$

from the continuous mapping theorem. Finally since

$$\underset{\sum_i f_i = 1}{\operatorname{argmax}} \sum_{i=1}^{k} p_i \log f_i = \boldsymbol{p},$$

and by the strict concavity of the logarithm, the right hand side of (B.60) is strictly greater than zero, which proves the theorem for the finite case.

The proof for the case $k = \infty$, when the candidate class $\mathcal{F}_{k,\boldsymbol{w}}^*$ is defined in (B.15), is similar to the finite case, with $\mathcal{F}_{k,\boldsymbol{w_1}}^*$ and $\mathcal{F}_{k,\boldsymbol{w_2}}^*$ properly changed to $\mathfrak{F}_{r,\boldsymbol{w}_1}^*$ and $\mathfrak{F}_{r,\boldsymbol{w}_2}^*$, $B(\boldsymbol{w})$ to $\mathfrak{B}(\boldsymbol{w})$, $\boldsymbol{p}$ to $\mathfrak{p}$, $\hat{\boldsymbol{p}}_n$ to $\hat{\mathfrak{p}}_n$ and $\hat{\boldsymbol{p}}_n^*$ to $\hat{\mathfrak{p}}_n^*$. □

Proof of Theorem 6.1. We consider the case of an infinite support ($k = \infty$), choose a fixed finite r and a candidate class $\mathcal{F}_{k,\boldsymbol{w}}^*$, defined in (B.15), and assume that the true pmf $\boldsymbol{p}$ has the following structure

$$p_{q_1} = \cdots = p_{q_1+v_1-1} > p_{q_2} \quad = \quad \cdots = p_{q_2+v_2-1} > \cdots > .$$

Let a be the number of the constant regions of $\boldsymbol{p}$ up to the element with the index r, i.e. a is such that $r \in \{q_a, q_a + v_a - 1\}$ and $p_{q_a} = \cdots = p_r = \cdots = p_{v_a-1}$. Therefore, there are $T = 2^{r-a}$ classes of the form (B.15) containing the true pmf $\boldsymbol{p}$, which we may label $j = 1, \ldots, S$

Suppose $\hat{\boldsymbol{p}}_n^*$ is the post-model-selection estimator, defined in (B.45), with $\mathcal{F}^*_{k,\hat{\boldsymbol{w}}_n}$, or equivalently $\hat{\boldsymbol{w}}_n$, the selected class. Then, from Theorem 5.6 there exists an n_1 such that for all $n \geq n_1$

$$\mathbb{P}[\mathcal{F}^*_{k,\hat{\boldsymbol{w}}_n} \ni \boldsymbol{p}] \quad = \quad 1.$$

Therefore, for $n \geq n_1$ the post-model-selection estimator $\hat{\boldsymbol{p}}_n^*$ can be written as

$$\begin{aligned}
\hat{\boldsymbol{p}}_n^* &= \sum_{j=1}^{S} \hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_j) 1\{\boldsymbol{w}_j = \hat{\boldsymbol{w}}_n\} \\
&= \sum_{j=1,\ldots,S:\mathcal{F}^*_{k,\boldsymbol{w}} \ni \boldsymbol{p}} \hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_j) 1\{\boldsymbol{w}_j = \hat{\boldsymbol{w}}_n\} \\
&\quad + \sum_{j=1,\ldots,S:\mathcal{F}^*_{k,\boldsymbol{w}} \not\ni \boldsymbol{p}} \hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_j) 1\{\boldsymbol{w}_j = \hat{\boldsymbol{w}}_n\} \\
&\overset{a.s.}{=} \sum_{j=1}^{T} \hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_j) 1\{\boldsymbol{w}_j = \hat{\boldsymbol{w}}_n\}, \qquad \text{(B.61)}
\end{aligned}$$

where we have (re)-labeled the classes so that $\mathcal{F}^*_{k,\boldsymbol{w}_j} \ni \boldsymbol{p}$, for $j = 1, \ldots, T$.

Now, let us consider any of the T candidate classes $\mathcal{F}^*_{k,\boldsymbol{w}}$ that contain $\boldsymbol{p}$, and let $\hat{\boldsymbol{p}}_n^*(\boldsymbol{w}) = \Pi(\hat{\boldsymbol{p}}_n | \mathcal{F}^*_{k,\boldsymbol{w}})$. Then, from Theorem 4.2 it follows that there exists a finite $\tilde{n}$, such that for all $n > \tilde{n}$,

$$||\hat{\boldsymbol{p}}_n^*(\boldsymbol{w}) - \boldsymbol{p}||_2 \overset{a.s.}{\leq} ||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2.$$

If $\tilde{n}_1, \ldots, \tilde{n}_T$ are the required values of $\tilde{n}$ for the T candidate classes, and $n_2 = \max(\tilde{n}_1, \ldots, \tilde{n}_T)$, then for all $n > n_2$,

$$\max_{j \in 1,\ldots,T} ||\hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_j) - \boldsymbol{p}||_2 \overset{a.s.}{\leq} ||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2, \qquad \text{(B.62)}$$

with $\hat{\boldsymbol{p}}_n^*(\boldsymbol{w}_j) = \Pi(\hat{\boldsymbol{p}}_n | \mathcal{F}^*_{k,\boldsymbol{w}_j})$ for $j \in 1, \ldots, T$, the projections on the classes $\mathcal{F}^*_{k,\boldsymbol{w}_j}$ which contain $\boldsymbol{p}$.

Finally, let $n_3 = \max\{n_1, n_2\}$. Then, from (B.61) and (B.62), the post-model-selection estimator $\hat{\boldsymbol{p}}_n^*$, defined in (B.45), satisfies that for all $n > n_3$,

$$||\hat{\boldsymbol{p}}_n^* - \boldsymbol{p}||_2 \overset{a.s.}{\leq} ||\hat{\boldsymbol{p}}_n^G - \boldsymbol{p}||_2. \qquad \text{(B.63)}$$

□

9 Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In Proc. 2nd Information Theory, Ed. B. N. Pterov and Csako* pp. 267-81. Budapest.

[2] Anevski, D., Fougeres, A-L. (2018). Limit properties of the monotone rearrangement for density and regression function estimation *Bernoulli*, to appear.

[3] Anevski, D., Pastukhov, V. (2017). The asymptotic distribution of the isotonic regression estimator over a countable preordered set. Tech. rep., arXiv.org:1709.03807.

[4] Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika* **86** 141–152.

[5] Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical inference under order restrictions*. John Wiley & Sons, London-New York-Sydney.

[6] Bogachev V. I. (2007). *Measure theory. Vol. I*. Springer-Verlag, Berlin.

[7] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. The Annals of Statistics **43**, 1774–1800.

[8] Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

[9] Da Prato G. (2006). *An Introduction to Infinite-Dimensional Analysis*. Springer-Verlag, Berlin.

[10] Fougères, A.-L. (1997), Estimation de densités unimodales, The Canadian Journal of Statistics. La Revue Canadienne de Statistique, 25(3), 375–387.

[11] Grenander, U. (1956). On the theory of mortality measurement. *Skand. Aktuarietidskr.* **39** 125–153.

[12] Jankowski, H. K. and Wellner, J. A. (2009). Estimation of a discrete monotone distribution. *Electronic journal of statistics* **3** 1567–1605.

[13] Kuiper, R. M., Hoijtink, H. and Silvapulle, M. J. (2011). An Akaike-type information criterion for model selection under inequality constraints. *Biometrika* **98** 495–501.

[14] Kuiper, R. M., Hoijtink, H. and Silvapulle, M. J. (2011). Generalization of the order-restricted information criterion for multivariate normal linear models. *Journal of Statistical Planning and Inference* **142** 2454–2463.

[15] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22** 79–86.

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

[17] Robertson, T., Wright, F. T. and Dykstra R. L. (1988). *Order restricted statistical inference*. John Wiley & Sons, Ltd., Chichester.

[18] Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review* **56** 49–62.

[19] Silvapulle, M. J. and Sen, P. K. (2005). Constrained Statistical Inference. John Wiley & Sons, Ink., Hoboken, New Jersey.

[20] van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge University Press, Cambridge.

[21] Vanbrabant L (2017). restriktor: Restricted Statistical Estimation and Inference for Linear Models. R package version 0.1-70. http://CRAN.R-project.org/package=restriktor

[22] Warrack G. and Robertson, T. (1984) A likelihood ratio test regarding two nested but oblique order-restricted hypothesis. *Journal of the American Statistical Association* **79** 881–886.

[23] Warrack G. and Robertson, T. (1984) An application of order restricted inference methodology to a problem in psychiatry. *Psychometrica* **50** 421–427.

[24] Wright, F. T. (1982). Monotone regression estimates for grouped observations. *The Annals of Mathematical Statistics* **10** 278–286.

C

Paper C

A stochastic process approach to multilayer neutron detectors

Dragi Anevski[a], Richard Hall-Wilton[b], Kalliopi Kanaki[b] and Vladimir Pastukhov[a]

[a]*Centre for Mathematical Sciences, Lund University*
[b] *European Spallation Source*

Abstract

The sparsity of the isotope Helium-3, ongoing since 2009, has initiated a new generation of neutron detectors. One particularly promising development line for detectors is the multilayer gaseous detector. In this paper, a stochastic process approach is used to determine the neutron's energy from the additional data afforded by the multilayer nature of these novel detectors.

The data from a multi-layer detector consists of counts of the number of absorbed neutrons along the sequence of the detector's layers, in which the neutron absorption probability is unknown. We study the maximum likelihood estimator for the intensity and absorption probability, show its consistency and asymptotic normality, as the number of incoming neutrons goes to infinity. We combine these results with known results on the relation between the absorption probability and the wavelength to derive an estimator of the wavelength and to show consistency and asymptotic normality.

Keywords: Maximum Likelihood, Multinomial Thinning of Point Processes, Neutron Detection, Poisson Process, Thinned Poisson Process.

1 Introduction

The European Spallation Source[1] (ESS), sited in Lund, Sweden, is planned to be operational in 2019 and the world's leading source for the study of materials using neutrons by 2025.

In order to address the challenge of developing a new generation of neutron detectors an international collaboration of 10 neutron scattering institutes in Europe, Asia and America (the International Collaboration on the Development of Neutron Detectors[2]) was formed in 2010. The members have chosen as the three most promising technologies for investigation: Scintillator detectors, boron-10 thin film detectors and $^{10}BF_3$ gas detectors. At present boron-10 thin film detectors seem to be the only realistic solution for large area detectors (> 10 m^2 active detector area). For the ESS, novel neutron detectors represent a critical technology that need to be developed, with corresponding R&D done as contributions to the ESS design work.

In this paper we study the feasibility and possibilities of the statistical determination of neutron wavelength for the new generation of neutron detectors being developed at the ESS.

Assume that a beam of neutrons arrives at the face of the detector. The detector consists of a sequence of boron-10 coated layers, between which there are gas-filled cavities. The principle of the detector can be described in a simplified manner as follows: a neutron that goes through a boron-10 layer can sometimes interact with a boron-10 atom in the layer, temporarily exciting the atom into an unstable state from which it will fall back to a stable state and thereby emit an electrically charged particle, that will ionise the gas. This electrical potential in the gas filled chamber is detected and the instrument notes that a neutron has been absorbed, see [4]. The outcome of this is that we have a count of +1 in the number of neutrons that have passed and been detected. The probability with which a neutron is absorbed and detected is a function of the energy content of the neutron, i.e. a function of the neutron wavelength.

If we view the neutron beam as a set of particles that hit the face of the detector, then each neutron will either be absorbed or not at the first layer. If the neutron is not absorbed at the first layer, it may possibly be absorbed at the second layer, and so on. From the simplified description above it is clear

[1] `https://europeanspallationsource.se`
[2] `http://icnd.org`

that the data from a multilayer detector will consist of counts of the number of absorbed neutrons along the sequence of the detector's layers.

By a beam we mean a stream of particles with a certain fixed wavelength μ. Let the number of neutrons that arrive in the time interval $[0, t]$ be denoted by $X_0(t)$. Then $X_0(t)$ is a counting process, such that $X_0(0) = 0$.

A simple model for the process of incoming neutrons $X_0(t)$ is that of a Poisson process with intensity λ. The Poisson model assumption is reasonable since neutrons are electrically neutral particles and since there are therefore no long-distance interactions between the particles in the beam, see Chapter 2 in [10] for a discussion of the model. The intensity λ is assumed to be an unknown nuisance parameter, and will be estimated.

At a layer each neutron is absorbed with a certain probability p (the absorption efficiency). The probability of absorption p is also assumed to be an unknown parameter, its dependence on the wavelength μ of the incident neutron is, however, of a known functional form, see [4]. This property will be used to make inference about the parameter μ. For a more thorough introduction to the subject of neutron interactions we refer to Chapter 2 in [10].

As will be shown later, our data set is generated by a sequentially thinned Poisson process, which is a special case of multinomial thinning. Inference for thinned point processes was studied in detail in [5] and [2], where the authors, in particular, studied the problem of estimation of the thinning parameter p from the observation of the thinned processes. The thinning parameter p is defined as a function from an underlying compact metric space to $[0, 1]$, in [5] and [2]. In[5] the author uses a nonparametric histogram estimator of p and in [2] the author studies a kernel estimator.

Though the approaches developed in [5] and [2] are quite general, they cannot be applied to the problem considered in this paper because, first, in our case the absorption probability (thinning parameter) is homogeneous (does not depend on the time of experiment) and, therefore, we can use a parametric approach to estimate it and, second, our data come from the multinomial thinning of the original Poisson process, not a binomial one as in [5] and [2].

The problem of multinomial thinning of point processes was studied in [7], where the author, in particular, proved that a point process is Poisson if and only if the thinned processes are independent and Poisson. However, to our knowledge, the problem of inference for a sequentially thinned Poisson pro-

cess has not been studied yet. Given the data, we suggest in this paper a likelihood approach and study the maximum likelihood estimator (mle) of the two-dimensional parameter (λ, p), where λ is the intensity and p the thinning parameter (absorption probability). In this paper, we derive conditions for the existence of the mle and prove its consistency and asymptotic normality, as the experiment time (or number of incoming neutrons) goes to infinity. We combine these results with known results for the relation between the absorption probability and the wavelength to derive a final estimator of the wavelength and to show consistency and asymptotic normality for the estimator. We also state results on the precision of the estimator, by deriving a relation between the width of the confidence interval, for the unknown wavelength, and the detector construction, in terms of the number of layers used in the detector. The performance of the estimator is illustrated on simulated data.

There are two main results of this paper. The first establishes the feasibility of estimating the wavelength of a neutron beam, based only on count data of the number of detected neutrons. The second determines necessary features of the detector, which for the specific detector is the number of layers, in order to be able to estimate the wavelength with a given precision. Following the construction of the ESS research facility, we intend to apply our estimation procedures to experimental data.

The paper is organized as follows. Section 2 provides the general scheme of the neutron detector and the modeling of neutron interactions with the detector layers. Section 3 is devoted to the inference of the parameters: We derive the mle for the intensity λ of an incident beam and absorption efficiency p, in Lemma 3.1 and 3.2 we discuss the uniqueness of the solutions to the score equations, and in Theorem 3.3, which is one of the main results of this paper, we derive the strong consistency and asymptotic normality of the mle. In Corollaries 3.1 and 3.2 we derive the consistency and asymptotic normality of the mle of the wavelength. Using these final results we are able to construct confidence intervals for the wavelength. Section 4 gives a simulation study to explore the estimator's performance. Section 5 contains a discussion of the results presented in the paper and plans for future work. Proofs of all results are given in the Appendix.

2 Scheme of a discrete spacing detector

Assume that an incident beam of neutrons hits the first layer of the detector, cf. Figure C.1. At the layer a neutron can possibly be absorbed and detected. If a neutron is not absorbed it will go through the detector's layer. We assume that these are the only two possibilities for the neutron interaction with a layer, i.e. it is assumed that the probability of an inelastic scattering of a neutron in the boron layers or in the material of the layers is negligibly small. Let p be the probability of an absorption of a neutron, so that $1 - p$ is the probability of its transmission. If a neutron is absorbed, it will then be detected. Let $X_1(t)$ be the number of neutrons that are absorbed at the first layer, so that $X_1^{tr}(t) = X_0(t) - X_1(t)$ is the number of transmitted neutrons.

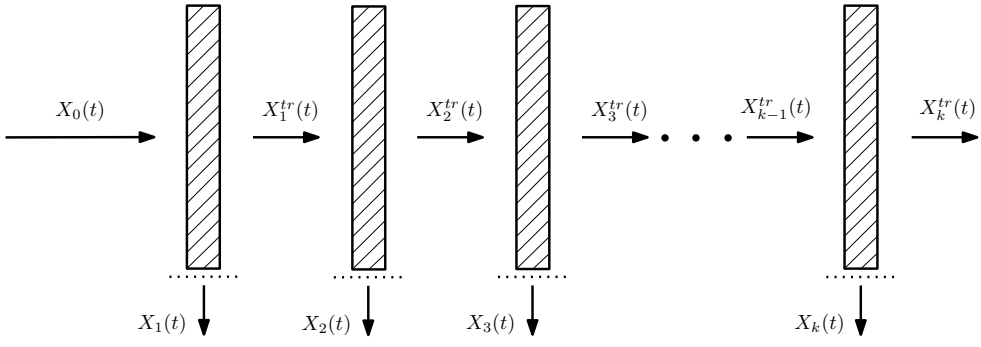

Figure C.1: The scheme of the detector.

Now assume that the beam of transmitted neutrons $X_1^{tr}(t)$ hits the next layer, at which, again, each neutron can either be absorbed (with the same probability p as at the previous layer) and then detected, or transmitted again. Let $X_2(t)$ be the number of neutrons that are absorbed at the second layer and let $X_2^{tr}(t) = X_1^{tr}(t) - X_1(t)$ be the number of transmitted neutrons. We assume that the registrations (absorptions) of different particles are independent and the times of absorption and travelling from layer to layer are negligibly small. This behaviour is repeated at each layer and gives the general scheme for the neutron beam's absorption and transmission in the detector.

Let $X_i(t)$ be the number of neutrons absorbed at the layer i in the time interval $[0, t]$ and let $X_i^{tr}(t)$ be the number of transmitted neutrons in the same time interval through the layer i, for $i = 1, \ldots, k$. Then $X_i(t)$ and $X_i^{tr}(t)$ are counting processes and $X_i(0) = 0$ and $X_i^{tr}(0) = 0$, for $i = 1, \ldots, k$. The next lemma shows that $\{X_i(t)\}_{i \geq 1}$ are jointly independent Poisson processes with

parameters $\lambda p(1-p)^{i-1}$, respectively.

Lemma 2.1. *The processes $\{X_i(t)\}_{i\geq 1}$ are jointly independent Poisson processes with intensities $p(1-p)^{i-1}\lambda$.*

The statement of Lemma 2.1 follows from the property of a multinomial thinning of a Poisson process cf. Theorem 5.17 in [6], [7], [1].

3 Inference for the parameters

Now suppose that we have run an experiment at the neutron detector, the result of which is a sequence of counts of the numbers of detected neutrons along the detector. Let us denote the data as a vector $x = (x_1, \ldots, x_k)$ of integers, with x_i the number of observed neutrons at layer i, for $i = 1, \ldots, k$. From Lemma 2.1 we know that the data are observations of independent Poisson distributed random variables, with unknown expectations $p(1-p)^{i-1}\lambda$, for $i = 1, \ldots, k$.

3.1 The mle of the thinning parameter p and the intensity of an incident process λ

We are interested in deriving consistency and asymptotic normality of the estimators. For this we need to explain what we mean by letting "the amount of data" go to infinity. There are several ways to model this. We can either let the experiment time t increase, or we can view the problem as a repeated measurement problem and thus make several, n of them, independent measurements during a fixed time interval $[0, t]$ and instead let n go to infinity. Since we use the Poisson process as a model for the neutron beam, the two approaches will give quantitatively the same limit results. We choose to view the problem as a repeated sample problem.

The inference problem can be described as follows. We perform n experiments. For each experiment $j = 1, \ldots, n$, we measure the number of neutrons X_{ij} detected at layer $i = 1 \ldots, k$ during the time interval $[0, t]$. Thus $\{X_{ij}\}_{i,j=1}^{n}$ are the random variables and $\{x_{ij}\}_{i,j=1}^{n}$ are the values which X_{ij} take. Let (p, λ) denote the parameters, that are assumed to lie in $[0, 1] \times [0, \infty)$. Introduce the vectors $\mathbf{X_j} = (X_{1j}, \ldots, X_{kj})^T$ and $\mathbf{x_j} = (x_{1j}, \ldots, x_{kj})^T$, respectively. Note that

the vectors $\mathbf{X_j}$ are independent random vectors with jointly independent components X_{ij}, by Lemma 2.1, from n independent experiment rounds. Finally denote $\mathbf{X} = [\mathbf{X_1}, \ldots, \mathbf{X_n}]$ and $\mathbf{x} = [\mathbf{x_1}, \ldots, \mathbf{x_n}]$, and note that these are $k \times n$ matrices of discrete random variables and of integers values, respectively.

Thus we let X_{ij} be the number of neutrons observed at the layer i at the experiment round j with probability mass function

$$f(x_{ij}|p,\lambda) = e^{-m_i}\frac{m_i^{x_{ij}}}{x_{ij}!},$$

where $m_i = p(1-p)^{i-1}\lambda t$. Then each vector $\mathbf{X_j} = (X_{1j}, \ldots, X_{kj})^T$ has the joint distribution

$$f(\mathbf{x_j}|p,\lambda) = \prod_{i=1}^{k} f(x_{ij}|p,\lambda) = \prod_{i=1}^{k} e^{-m_i}\frac{m_i^{x_{ij}}}{x_{ij}!}.$$

Note, that if $k = 1$, then $m = p\lambda t$ and, therefore, in this case one can only estimate the product $p\lambda$, and not p and λ separately.

Assume that $k > 1$. The log-likelihood is then given by

$$l_n(p,\lambda|\mathbf{x}) \quad = \quad \sum_{j=1}^{n}\sum_{i=1}^{k}(-m_i + x_{ij}\log m_i - \log x_{ij}!).$$

The mle $(\hat{p}_n, \hat{\lambda}_n)$ is the solution of the score equations

$$\begin{cases} \frac{1}{n}\frac{\partial l_n}{\partial \lambda} = \frac{s_n - \lambda t(1-(1-p)^k)}{\lambda} = 0, \\ \frac{1}{n}\frac{\partial l_n}{\partial p} = \frac{(1-p)(s_n+z_n) - z_n - \lambda t(k(1-p)^k - k(1-p)^{k+1})}{p(1-p)} = 0, \end{cases} \tag{C.1}$$

where $s_n = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{k} x_{ij}$ and $z_n = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{k}(i-1)x_{ij}$. If we assume that $\hat{p}_n(1-\hat{p}_n) \neq 0$, $\hat{\lambda}_n \neq 0$ we get the system of equations

$$\begin{cases} s_n - \hat{\lambda}_n t(1-\hat{y}_n^k) = 0, \\ a_n\hat{y}_n^{k+1} - b_n\hat{y}_n^k + c_n\hat{y}_n - d_n = 0, \end{cases} \tag{C.2}$$

where

$$\begin{aligned} a_n &= -s_n - z_n + ks_n, \\ b_n &= -z_n + ks_n, \\ c_n &= z_n + s_n, \\ d_n &= z_n, \\ \hat{y}_n &= 1 - \hat{p}_n. \end{aligned} \tag{C.3}$$

Obviously $(C.2)$ has exactly one solution $(\hat{p}_n, \hat{\lambda}_n)$ if and only if the second equation in $(C.2)$ has exactly one root.

Lemma 3.1. *The function*

$$f(y) = a_n y^{k+1} - b_n y^k + c_n y - d_n,$$

for $k > 1$ with coefficients given in (C.3), has one zero in the open interval $(0,1)$ when the inflection point $y_{i.p.}$ satisfies the inequality

$$y_{i.p.} := \frac{b_n(k-1)}{a_n(k+1)} < 1,$$

and no zeros in $(0,1)$ when $y_{i.p.} \geq 1$.

Lemma 3.1 gives the condition of existence and uniqueness of $(\hat{p}_n, \hat{\lambda}_n)$, but there is no guarantee that it holds for a finite n. However, the following result holds.

Lemma 3.2. *Let $A_n = \{$Equation (C.2) has exactly one root in $(0,1)\}$. Then A_n happens for all sufficiently large n almost surely.*

Asymptotic properties of the mle

Theorem 3.3. *The mle $(\hat{p}_n, \hat{\lambda}_n)$, given in (C.1), is strongly consistent*

$$(\hat{p}_n, \hat{\lambda}_n) \xrightarrow{a.s.} (p, \lambda),$$

and asymptotically normal

$$\sqrt{n}((\hat{p}_n, \hat{\lambda}_n) - (p, \lambda)) \xrightarrow{d} \mathcal{N}(\mathbf{0}, [\mathbf{I}(p, \lambda)]^{-1}),$$

as $n \to \infty$, where $\mathbf{I}(p, \lambda)$ is the information matrix

$$\mathbf{I}(p, \lambda) \quad = \quad \frac{1}{k} \sum_{i=1}^{k} \mathbf{I}_{(i)}(p, \lambda),$$

where $\mathbf{I}_{(i)}(p, \lambda)$ denotes the information matrix corresponding to $f(x_{ij}|p, \lambda)$ with fixed i.

From the theorem above, after simplification, we obtain the following asymptotic covariances

$$\begin{aligned}
\sigma_p^2(p,\lambda) = [\mathbf{I}(p,\lambda)]^{-1}_{pp} &= \frac{(1-(1-p)^k)(1-p)p^2}{\lambda t q(p,k)} \to \frac{(1-p)p^2}{\lambda t}, \\
\sigma_\lambda^2(p,\lambda) = [\mathbf{I}(p,\lambda)]^{-1}_{\lambda\lambda} &= \frac{\lambda h(p,k)}{t q(p,k)} \to \frac{\lambda}{t}, \\
\sigma_{p,\lambda}^2(p,\lambda) = [\mathbf{I}(p,\lambda)]^{-1}_{\lambda p} &= \frac{kp((1-p)^k-(1-p)^{k-1})}{t q(p,k)} \to 0,
\end{aligned}$$

as $k \to \infty$, where

$$h(p,k) = 1 - k^2(1-p)^{k+1} + (2k^2-1)(1-p)^k - k^2(1-p)^{k-1},$$

and

$$\begin{aligned}
q(p,k) &= (1-p)^{2k} - k^2(1-p)^{k+1} + 2(k^2-1)(1-p)^k \\
&\quad -k^2(1-p)^{k-1} + 1. \qquad \text{(C.4)}
\end{aligned}$$

We are mainly interested in the estimation of p, since there is a functional relation between the absorption efficiency p and the wavelength μ of the incident neutrons, cf. (C.5) and (C.6) below. Analysing the behaviour of $\sigma_p^2(p,\lambda)$, it can be shown that $\sigma_p^2(p,\lambda)$ is a strictly decreasing function of k for every $p \in (0,1)$.

3.2 Estimation of the wavelength μ of an incident beam.

We are interested in estimating the wavelength of a monochromatic neutron beam. The probability of absorption p depends on the neutron wavelength μ as (cf. Section 2.3 in [10])

$$p = 1 - e^{-\Sigma(\mu)\rho_{at}d_l}, \qquad \text{(C.5)}$$

where the parameter $\Sigma(\mu)$ is called the cross-section of absorption, ρ_{at} is the atomic density of ^{10}B in the B_4C coating and d_l is the thickness of the boron layer. Example values of parameters in a detector are $\rho_{at} = 10^{29}\ m^{-3}$, $d_l = 10^{-6}$ m, cf. [4].

The neutron cross-section $\Sigma(\mu)$ can be modelled as

$$\Sigma(\mu) = \varsigma\mu,$$

where the coefficient ς is different for different materials, see [10]. Furthermore, the coefficient ς does not depend on the neutron wavelength and has been measured experimentally, cf. [8]. From the results in [8] we conclude that the estimator $\hat{\varsigma}$ of ς is unbiased and asymptotically normal

$$\sqrt{n'}(\hat{\varsigma}_{n'} - \varsigma) \xrightarrow{d} \mathcal{N}(0, \sigma_\varsigma^2),$$

as $n' \to \infty$. Here n' is the number of runs performed in the experiment to estimate ς and σ_ς^2 is its asymptotic variance.

Let us rewrite (C.5) as

$$p = 1 - e^{-\chi\mu}, \tag{C.6}$$

where

$$\chi = \rho_{at} d_l \varsigma,$$

The plug-in estimator $\hat{\chi} = \rho_{at} d_l \hat{\varsigma}$ of χ is then asymptotically normal

$$\sqrt{n'}(\hat{\chi}_{n'} - \chi) \xrightarrow{d} \mathcal{N}(0, \sigma_\chi^2), \tag{C.7}$$

with $\chi = \rho_{at} d_l \varsigma$ and $\sigma_\chi^2 = \rho_{at}^2 d_l^2 \sigma_\varsigma^2$.

From (C.6), we obtain

$$\mu(p, \chi) = -\frac{\log(1-p)}{\chi}. \tag{C.8}$$

Next, we combine two limit distribution results, for $\hat{p}_n$ and for $\hat{\chi}_{n'}$, to get a limit distribution for the plug-in estimator of μ. In order to formalize this in a proper way, we introduce a factor γ, which is merely the (asymptotic) ratio between n' and n. The result in a practical finite-sample situation will be used in exactly that way: by letting $\gamma = n'/n$ and use the limit distribution to provide asymptotic confidence intervals or tests.

Corollary 3.1. *The plug-in estimator $\hat{\mu} = \mu(\hat{p}_n, \hat{\chi}_{n'})$ of μ is asymptotically normal*

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma_\mu^2),$$

where

$$\sigma_\mu^2 = \left[\frac{\partial \mu}{\partial p}(p, \chi)\right]^2 \sigma_p^2(p, \lambda) + \frac{1}{\gamma}\left[\frac{\partial \mu}{\partial \chi}(p, \chi)\right]^2 \sigma_\chi^2$$

as $n \to \infty$, where n is the number of measurements for $\hat{p}_n$ and $n' = \lceil \gamma n \rceil$, $\gamma > 0$, , is the number of measurement for $\hat{\chi}_{n'}$ ($\lceil \gamma n \rceil$ is smallest integer not less than γn).

Introduce the notation

$$S_n^2(\hat{p}, \hat{\lambda}, \hat{\chi}) = \left[\frac{\partial \mu}{\partial p}(\hat{p}_n, \hat{\chi}_{n'})\right]^2 \sigma_p^2(\hat{p}_n, \hat{\lambda}_n) + \frac{1}{\gamma}\left[\frac{\partial \mu}{\partial \chi}(\hat{p}_n, \hat{\chi}_{n'})\right]^2 \hat{\sigma}_\chi^2, \tag{C.9}$$

where both the estimate $\hat{\chi}_{n'}$ and the estimate of the variance $\hat{\sigma}_\chi^2$ are based on n' measurements, and $(\hat{p}_n, \hat{\lambda}_n)$ are the mle of (p, λ) based on n measurements.

The next result follows from Slutsky's theorem and the continuous mapping theorem, cf. Chapter 2 in [9].

Corollary 3.2.

$$\frac{\sqrt{n}(\hat{\mu} - \mu)}{S_n} \xrightarrow{d} \mathcal{N}(0, 1),$$

as $n \to \infty$, where $n' = \lceil \gamma n \rceil$, $\gamma > 0$ and S_n is given in (C.9).

Using the above limit distribution result for the mle $\hat{\mu}$ we can construct the approximate confidence interval for μ. The approximate $100(1-\alpha)$ percent confidence interval for μ is

$$[\mu(\hat{p}_n, \hat{\chi}_{n'}) - z_{\alpha/2}\frac{S_n}{\sqrt{n}},\ \mu(\hat{p}_n, \hat{\chi}_{n'}) + z_{\alpha/2}\frac{S_n}{\sqrt{n}}], \tag{C.10}$$

where $z_{\alpha/2}$ is the $\alpha/2$-th quantile of the standard normal distribution.

Next, let us rewrite the expression for $\frac{S_n}{\sqrt{n}}$ as

$$\frac{S_n}{\sqrt{n}} = S_{\hat{\mu}}^{(p)} + S_{\hat{\mu}}^{(\chi)}, \tag{C.11}$$

where

$$S_{\hat{\mu}}^{(p)}(\hat{p}_n, \hat{\lambda}_n, \hat{\chi}_{n'}) = \frac{1}{\sqrt{n}}\frac{\partial \mu}{\partial p}(\hat{p}_n, \hat{\chi}_{n'})\sigma_p(\hat{p}_n, \hat{\lambda}_n) = \frac{\sigma_p(\hat{p}_n, \hat{\lambda}_n)}{\sqrt{n}(1-\hat{p}_n)\hat{\chi}_{n'}}, \tag{C.12}$$

$$S_{\hat{\mu}}^{(\chi)}(\hat{p}_n, \hat{\lambda}_n, \hat{\chi}_{n'}) = \frac{\gamma}{\sqrt{n}}\frac{\partial \mu}{\partial \chi}(\hat{p}_n, \hat{\chi}_{n'})\hat{\sigma}_\chi = \frac{\log(1-\hat{p}_n)}{\sqrt{n'}\hat{\chi}_{n'}^2}\hat{\sigma}_\chi. \tag{C.13}$$

Note that $S_{\hat{\mu}}^{(\chi)}$ does note go to zero in probability as $n \to \infty$. Therefore, we can view this term as a kind of systematic error, outside of our control.

4 A simulation experiment

In this section we perform a simulation experiment to evaluate the estimator's performance. In particular, we illustrate the dependence of individual terms in (C.11) on the number of layers (Figure C.2) and on the intensity of a beam (Figure C.3), and the confidence interval width's dependence on the number of layers for several wavelengths (Figure C.5).

We simulate a Poisson process $X_0(t)$ a number of times n, for $n = 10,\ 100$, for the parameters values $p = 0.05,\ 0.07,\ 0.1$, $\lambda = 10^5\ s^{-1}$, which correspond to the wavelengths μ = 2.4, 3.4 and 4.9 Å. These are typical neutron wavelengths for the possible applications of the detector, see [4].

The mle $(\hat{p}_n, \hat{\lambda}_n)$ is calculated for the simulated data. We recall the relation between χ and ς in $(C.7)$, and note that ρ_{at} and d_l are known. The estimator of ς is assumed to be asymptotically normal, with mean value the sample mean and variance equal to a pooled variance estimate using three series of 15 measurements, which gives in total 45 experimental data points, see [8]. Using the results of [8] we have the following estimates for χ: $\hat{\chi}_{n'} = 2.142 \times 10^8\ m^{-1}$ and $\hat{\sigma}^2_{\chi} = 0.021 \times 10^8\ m^{-2}$.

First, we analyse the dependence of the approximal confidence interval on the number of detector's layers. Figure C.2 shows the dependence of $S^{(p)}_{\hat{\mu}}$ and $S^{(\chi)}_{\hat{\mu}}$, defined in (C.12) and (C.13), on the number of the layers in the detector for 10 and 100 runs of the experiment. We note, in particular, that $S^{(p)}_{\hat{\mu}}$ and $S^{(\chi)}_{\hat{\mu}}$ are of the same size at $k \approx 25$ for $n = 10$ experimental runs and at $k \approx 15$ for $n = 100$.

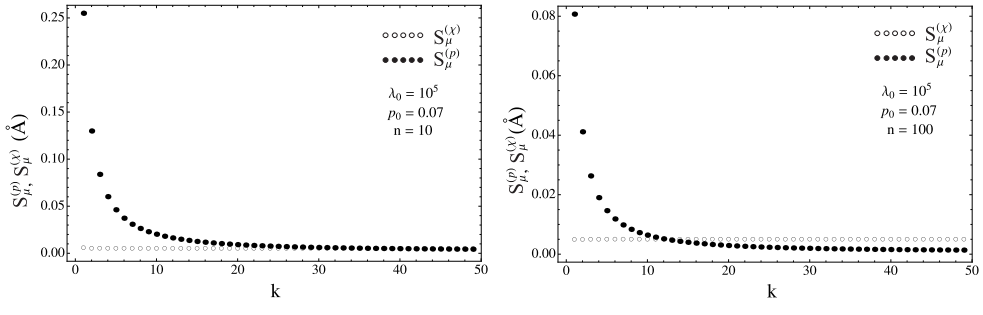

Figure C.2: The dependence of $S^{(p)}_{\hat{\mu}}$ and $S^{(\chi)}_{\hat{\mu}}$ on the number of layers k.

Second, we study the dependence of the approximate confidence interval on the intensity of an incident beam λ.

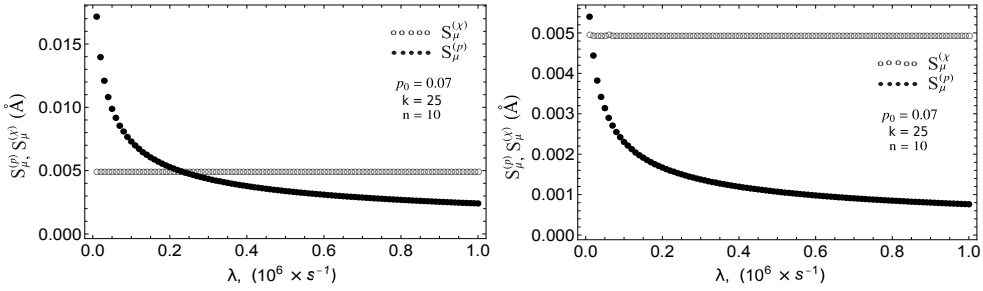

Figure C.3: The dependence of $S_{\hat{\mu}}^{(p)}$ and $S_{\hat{\mu}}^{(\chi)}$ on the the intensity of an incident beam λ.

Next, in order to assess the accuracy of the asymptotic approximation we estimate the coverage probability of the approximate confidence interval based on 5000 Monte-Carlo simulations. From Figure C.4 one can see that the deviation of the confidence band's width is less that 0.5 % even for the quite small number of repetitions $n = 10$.

In Figure C.5 we have plotted the confidence interval bars as a function of the number of layers, for μ = 2.4, 3.4 and 4.9 Å and $n = 10,\ 100$.

The results of the simulation experiments show that the errors are rapidly decreasing as a function of the number of layers k in the detector, cf. Figure 2, where the term $S_{\hat{\mu}}^{(p)}$ we may control by increasing the number of measurements, whereas the term $S_{\hat{\mu}}^{(\chi)}$ we are not able to influence and therefore we can see as a form of systematic error contribution to the total variance (C.11). As indicated in Figure 2, for the choice of model parameters, at approximately 10-25 layers the term $S_{\hat{\mu}}^{(p)}$ that we can affect becomes smaller than the systematic error term $S_{\hat{\mu}}^{(\chi)}$. Figure 3 shows that, again, the term $S_{\hat{\mu}}^{(p)}$ decreases with increasing intensity, whereas the term $S_{\hat{\mu}}^{(\chi)}$ is almost not affected by a change in intensity.

Note that in our simulations for Figure 4, and only here, in our assessment of the coverage probability for the confidence intervals, we treat the random variable ς as a constant, since we do not have the original data from which it was estimated and since we do not know the data generating mechanism. This implies that in Figure 4 the term $S_{\hat{\mu}}^{(\chi)}$ in $(C.11)$ is not taken into account

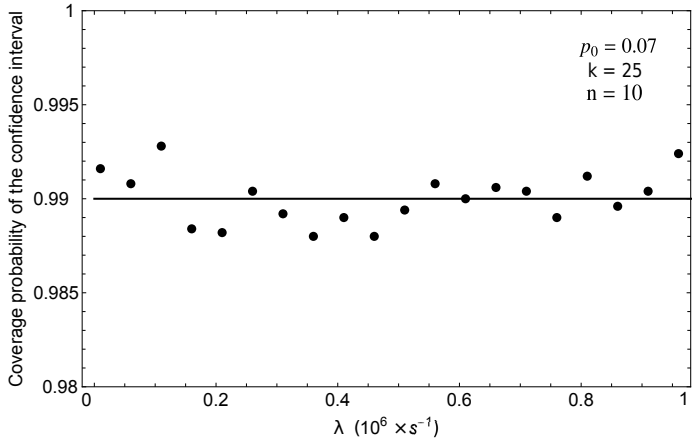

Figure C.4: Dependence of the coverage probability of the approximate confidence interval on the intensity of an incident beam λ.

in the constriction of the confidence interval.

Finally in Figure C.5 we illustrate that even for a small number of repetitions (i.e. small effective sample sizes), we obtain good efficiency in the estimation of the wavelengths.

5 Conclusions

The results here show that it is statistically possible to determine the neutron energy for a monochromatic beam with a good precision using multilayer neutron detectors. With relatively few layers ($\leq$ 15), already maximal information can be extracted and many layers do not significantly improve the precision of the results.

For neutron beams with high intensity ($\lambda \geq 10^6$ particles), a statistical precision (width of 99 % confidence interval) of less than 0.1 Å on the determination of the wavelength of the beam in the range 2.5-5 Å is possible (Fig.C.5). Uncertainty in the neutron's cross section of the boron-10 isotope becomes dominant in the regime of high intensity beams and more than 10-20 layers. This means again that more than 10-20 layers are not needed (Fig.C.2).

An interesting further outcome of our work is that it shows that it might be possible, in high intensity experiments, with a precisely determined wavelength of a monochromatic neutron beam, to improve the statistical measurement of

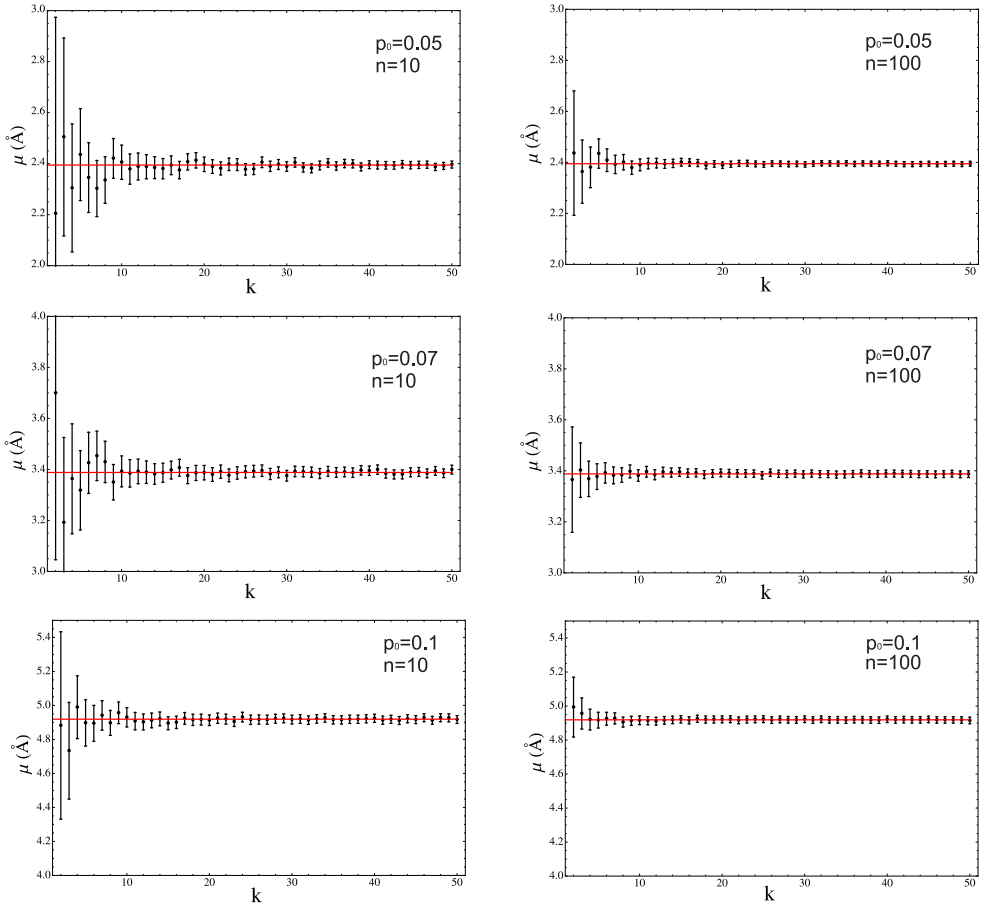

Figure C.5: 99% confidence interval for μ based on simulations for $n = 10$, 100 and $p = 0.05$, 0.07, 0.1, $\lambda = 10^5$, $t = 1$s. The red line is the true value of μ.

the boron-10 cross section by using an inverse of the method described in this manuscript. The systematic effects of such a measurement might be significant. In the limit of low intensity, a precision of 1 Å in determining the wavelength of the monochromatic neutron beam is still possible.

The asymptotic expansion used in the derivation of the asymptotic normality of the mle of the wavelength depends on two limit distribution results. The first is the asymptotic normality of the mle of the absorption probability p. Since we choose the effective number of neutrons that hits the detector ourselves, we are able to obtain an approximation which is as fine as wanted. Furthermore, the term $(C.12)$ in the total efficiency $(C.11)$, resulting from the

mle of p, can be obtained as small as desired. A possible limitation here is that a large number of effective neutrons means running the experiment for a long time. In that case the assumption of a constant intensity Poisson process as a model may become questionable. A possible remedy for this is to instead do many repeated runs, while tightly controlling the experimental apparatus, in order to obtain a homogeneous Poisson process in each run. The second asymptotic result is the asymptotic normality of the estimator of ς, which we conclude from [8]. The number of data points used for the estimation of ς in that paper is 45, and therefore arguably on the boundary of what one can accept as an asymptotic normality result. A more serious practical limitation for us is that we are not able to affect the term $(C.13)$ in $(C.11)$ resulting from the estimator of ς. This puts a limit on the total efficiency that we can obtain for the wavelength estimation in our experimental setup. It also tells us, as noted above, that building a detector with many layers is not necessary, since for such a detector the term that we can affect in $(C.11)$ becomes negligible compared to term arising from the estimation of ς, and therefore increasing the number of layers will have negligible effect on $(C.11)$.

In a real detector there may be a degradation in the result achieved coming from systematic effects resulting from defects in the detector.

In this paper we have considered the Poisson process as a model for the incoming beam. Having real data it will in the future be possible to perform goodness of fit tests, e.g. for assessing the validity of the Poisson process model. A possible alternative model for the incident beam is the negative binomial process. In fact, thinning of a negative binomial process also results in a negative binomial process, cf. [3]. However, unlike in the Poisson process case, the count processes $\{X_i(t)\}_{i\geq 1}$ will in that case not be independent, which makes the maximum likelihood approach more complicated. A possible solution could be to simplify the likelihood using some sort of quasi likelihood approach, e.g. by treating the count processes as independent and obtain similar expressions for the likelihood as in this paper. The model fit testing and negative binomial process modelling may be a direction for possible future research.

This manuscript concentrated on a monochromatic neutron beam. In the future our results will be generalised to discrete and continuous wavelength distributions for the incoming neutron beam.

6 Acknowledgements

VP's research is fully supported by the Swedish Research Council (SRC). The research of DA, RHW and KK is partially supported by the SRC. The authors gratefully acknowledge the SRC's support. The authors would furthermore like to thank the associate editor and referees for their comments that have significantly improved the exposition and readability of the paper.

Author's present address: Ällingavägen 12 lgh 1006 227 34 Lund

E-mail: pastuhov@maths.lth.se

7 Appendix

Proof Lemma 3.1 . For simplicity we skip the lower subscript n but we assume that a, b, c, d are as defined in (C.3).

We study the monotonicity and convexity/concavity of $\tilde{f}$ on $[0,\infty)$ by studying the signs of $\tilde{f}'$ and $\tilde{f}''$ on $[0,\infty)$. For $k \geq 2$ we have

$$\begin{aligned} \tilde{f}' &= a(k+1)y^k - bky^{k-1} + c, \\ \tilde{f}'' &= y^{k-2}k(a(k+1)y - b(k-1)). \end{aligned}$$

(i) : *The second derivative.*
Clearly $\tilde{f}''(0) = 0$. Factoring out $ky^{k-2} \geq 0$, we see that to study the zeros and signs of $\tilde{f}''$ is equivalent to studying the zeros and signs of

$$g(y) = a(k+1)y - b(k-1),$$

Clearly $g(0) = -b(k-1) < 0$, $g(\infty) > 0$ and $g(y)$ has a unique root

$$y_{i.p.} = \frac{b(k-1)}{a(k+1)}.$$

From the expressions in (C.3) we can see that both a and b are positive and $b > a$, which means that $y_{i.p.} \in (0,\infty)$.

Thus the function $\tilde{f}''$ is negative to the left of $y_{i.p.}$ and positive to the right of $y_{i.p.}$ which implies

a) $\tilde{f}$ is concave on $(0, y_{i.p.})$, convex on $(y_{i.p.}, \infty)$, and thus $y_{i.p.}$ is an inflection point for $\tilde{f}$.

(ii) : *The first derivative*. We see that $\tilde{f}'(0) = c > 0$. Furthermore using the expressions for a, b, c we see that $\tilde{f}'(1) = a(k+1) - kb + c = 0$. From the sign change of $\tilde{f}''$ at $y_{i.p.}$ we have that $\tilde{f}'$ is decreasing on $(0, y_{i.p.})$ and increasing on $(y_{i.p.}, \infty)$. Now there are two possible cases:

Case A : $y_{i.p.} < 1$. In this case, the sign change of $\tilde{f}''$ together with $\tilde{f}'(0) = c > 0, \tilde{f}'(1) = 0$ and the continuity of $\tilde{f}$, implies that for some $y_1 < y_{i.p.}$,

b') $\tilde{f}'$ is positive on $(0, y_1)$, negative on $(y_1, 1)$, positive on $(1, \infty)$,

which of course implies

c') $\tilde{f}$ is increasing on $(0, y_1)$, decreasing on $(y_1, 1)$, increasing on $(1, \infty)$.

Case B : $y_{i.p.} \geq 1$. In this case we know that $\tilde{f}'$ is decreasing and positive on $(0, 1)$, decreasing and negative on $(1, y_{i.p.})$ and increasing on $(y_{i.p.}, \infty)$. This implies that there is an y_2 such that $\tilde{f}'$ is negative on $(y_{i.p.}, y_2)$ and positive on (y_2, ∞). Thus the full statement becomes

b'') $\tilde{f}'$ is decreasing and positive on $(0, 1)$, decreasing and negative on $(1, y_{i.p.})$, increasing and negative on $(y_{i.p.}, y_2)$, increasing and positive on (y_2, ∞).

which implies that

c'') $\tilde{f}$ is concave and increasing on $(0, 1)$, concave and decreasing on $(1, y_{i.p.})$, convex and decreasing on $(y_{i.p.}, y_2)$, convex and increasing on (y_2, ∞).

(iii) : *The function*. We first note that $\tilde{f}(0) = -d < 0$, and that the expression for the coefficients a, b, c, d imply $\tilde{f}(1) = a - b + c - d = 0$. Now we treat the two cases separately:

Case A: From the sign changes of $\tilde{f}''$ and $\tilde{f}'$, it follows that $\tilde{f}$ is concave and increasing on $(0, y_1)$, concave and decreasing on $(y_1, y_{i.p.})$, convex and decreasing on $(y_{i.p.}, 1)$. This together with $\tilde{f}(0) = -d < 0$, $\tilde{f}(1) = 0$ implies (and in fact only the information that $\tilde{f}$ is first increasing, then decreasing is enough) that there is a zero $\tilde{y} \in (0, 1)$ for $\tilde{f}$.

Case B: In this case we have that $\tilde{f}$ is increasing and concave on $(0,1)$, which together with $\tilde{f}(0) = -d < 0$, $\tilde{f}(1) = 0$ implies that there is no zero for $\tilde{f}$ in the open $(0,1)$.

Finally noting that a zero $\tilde{y}$ of $\tilde{f}$ in $(0,\infty)$, corresponds, via $\tilde{y} = 1 - \tilde{p}$, to a zero $\tilde{p}$ of f in $(-\infty, 1)$, the Lemma follows. □

Proof of Lemma 3.2 . From Lemma 3.1, we see that

$$A_n = \left\{ \frac{b_n(k-1)}{a_n(k+1)} < 1 \right\}.$$

We will prove that

$$\frac{b_n(k-1)}{a_n(k+1)} \stackrel{a.s.}{\to} c, \tag{C.14}$$

as $n \to \infty$, for some constant $c < 1$. This immediately proves the condition of the lemma, since if $c < 1$

$$\left\{ \frac{b_n(k-1)}{a_n(k+1)} \to c \right\} \subseteq \bigcup_{n\geq 1} \bigcap_{m \geq n} A_m.$$

Now to prove (C.14), note that $\{s_j\}_{j=1}^n$ and $\{z_j\}_{j=1}^n$ in (C.3) are two sequences of i.i.d. random variables. Thus from the strong law of large numbers

$$\frac{b_n(k-1)}{a_n(k+1)} \stackrel{a.s}{\to} \frac{k-1}{k+1} \frac{k-(k+1)(1-p)+(1-p)^{k-1}}{(k-1)-k(1-p)+(1-p)^k} =: c,$$

as $n \to \infty$. One can easily prove that $c < 1$ by considering the polynomial

$$(k-1)(1-p)^{k+1} - (k+1)(1-p)^k + (k+1)(1-p) - (k-1),$$

which is negative for all $k > 1$ and $0 < p < 1$. This proves the lemma. □

Proof of Theorem 3.3 .

From Lemma 3.1 it follows that there exists n_1 such that for all $n > n_1$ the mle $(\hat{p}_n, \hat{\lambda}_n)$ is a differentiable function of (s_n, z_n), defined in (C.1). Therefore, the

strong consistency of $(\hat{p}_n, \hat{\lambda}_n)$ follows from the strong law of large numbers and the continuous mapping theorem.

Next, (s_n, z_n) is asymptotically normal, which follows from the central limit theorem. Using the delta method we prove the asymptotic normality of $(\hat{p}_n, \hat{\lambda}_n)$. □

Proof of Corollary 3.1.

Assume that there has been made n measurements for $(\hat{p}_n, \hat{\lambda}_n)$ and n' measurements for $\hat{\chi}_{n'}$, and that $(\hat{p}_n, \hat{\lambda}_n)$ and $\hat{\chi}_{n'}$ are independent. Let $n' = \lceil \gamma n \rceil$, with γ a proportionality factor that we introduce for convenience.

From the asymptotic normality of the estimators $\hat{p}_n$ and $\hat{\chi}_{n'}$ we have

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} \mathcal{N}(0, \sigma_p^2), \tag{C.15}$$

and

$$\begin{aligned} \sqrt{n}(\hat{\chi}_{n'} - \chi) &= \sqrt{\frac{n}{n'}}\sqrt{n'}(\hat{\chi}_{n'} - \chi) \\ &= \sqrt{\frac{n}{\lceil \gamma n \rceil}}\sqrt{n'}(\hat{\chi}_{n'} - \chi) \xrightarrow{d} \mathcal{N}(0, \frac{\sigma_\chi^2}{\gamma}), \end{aligned} \tag{C.16}$$

as $n \to \infty$, since $\lim_{n\to\infty} \frac{n}{\lceil \gamma n \rceil} = \frac{1}{\gamma}$. Combining $(C.15)$ and $C.16$, the result follows from the delta method, see, for example, Chapter 3 in [9]. □

8 Bibliography

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

[2] Bensaïd N. (1997) Nonparametric inference for thinned point process. Statistics & probability letters **33**, 253–258.

[3] Harremos P., Johnson O. T. & Kontoyiannis I. (2007). Thinning and the law of small numbers, Proc. ISIT 2007, 1491–1495.

[4] Kanaki K. at al. (2013). Statistical energy determination in neutron detector systems for neutron scattering science. 2013 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC).

[5] Karr A. F. (1985). Inference for thinned Poisson process, with application to Cox process. Journal of multivariate analysis **16**, 368–392.

[6] Kulkarni V. G. (2009). Modeling and analysis of stochastic systems, CRC Press.

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

[8] Schmitt H. W., Block R. C. & Bailey R, L. (1959). Total neutron cross section of B^{10} in the thermal neutron energy range. Nuclear Physics **17**, 109–115.

[9] van der Vaart A.W. (1998). Asymptotic statistics, Cambridge University Press, New York.

[10] Willis B. T. M. and Carlile C. J. (1999). Experimental neutron scattering, Oxford University Press.

Paper D

Estimating the distribution and thinning parameters of a homogeneous multimode Poisson process

Dragi Anevski and Vladimir Pastukhov

Centre for Mathematical Sciences, Lund University

Abstract

In this paper we propose estimators of the distribution of events of different kinds in a multimode Poisson process. We give the explicit solution for the maximum likelihood estimator and derive strong consistency and asymptotic normality of the estimator. We also provide an order restricted estimator and derive its consistency and asymptotic distribution. We discuss the application of the estimator to the detection of neutrons in a novel detector being developed at the European Spallation Source in Lund, Sweden. The inference problem gives rise a system of equations first studied by Ramanujan.

Keywords: Maximum Likelihood, Multinomial Thinning of Point Processes, Neutron Detection, Poisson Process, Thinned Poisson Process

1 Introduction

The motivation for the research in this paper comes from neutron detection, of importance for the European Spallation Source (ESS), which is a large scale research facility currently being built in Lund, Sweden. The main research problem from the physicists perspective in this connection is the estimation of the energy or, equivalently, the wavelength distribution of a neutron beam. The data in the neutron scattering experiment for the neutron detector type

that we are considering consists of counts of the numbers of neutrons that have been absorbed along the *layers* in the detector. Given the data, the goal is to estimate the unknown wavelength distribution in the neutron beam that one has observed. We have previously studied this problem in the simpler setting of there being exactly one wavelength in the neutron beam, which was then considered to be unknown, cf. [3]. The goal in [3] was to derive an estimator of the unknown wavelength, which was a maximum likelihood estimator (mle), and to derive properties of the estimator, in particular [3] showed the consistency and asymptotic normality of the mle. Of particular importance for the physicists are relations between the properties of the mle and the detector construction, and then, in particular, the number of layers used.

This paper can be seen as a generalisation of the study in [3], in the sense that we investigate the same detector type, but are now interested in a set of wavelengths with finite cardinality, say s, and that both the wavelengths sizes/values as well as the distribution of the wavelengths in the neutron beam are unknwon. The goal in this paper is to construct an estimate of these $2s$ parameters, and if possible to derive properties of the constructed estimator.

In [3], the neutron beam was assumed to be well described by a time homogenous Poisson process, and we take a similar approach here. We assume that the neutron beam is a sum of individual wavelength neutron beams, each being described by a Poisson process; the proportions $\boldsymbol{q}$ of the individual wavelength neutrons in the total sum is however unknown, and is in fact a parameter that we want to estimate; the total sum is, of course, still a Poisson process. The data obtained from the detector then consists of counts of neutrons that are absorbed and detected in the neutron detector, and we may use the key observation that the probability of absorption of a specific neutron is, in principle, a known function of the wavelength. Thus each neutron in the beam will be absorbed with a probability which depends on the wavelength of that neutron and one may assume that the absorptions of different neutrons, even of the same wavelength, are independent events. This points to the direction of modeling with the use of thinned Poisson processes.

In fact, we treat in this paper an inference problem that can be stated as the estimation of the wavelength distribution, as well as the thinning probabilities $\boldsymbol{p}$, i.e. the wavelength sizes, of a multimode homogeneous Poisson process.

Having stated the problem and formulated a maximum likelihood estimator,

we see that the problem becomes difficult to treat directly, if one goes trough the standard machinery of finding zeros to the score equations. In fact, the problem may be simiplified by rephrasing it into estimating algebraic functions of some of the $2s$ parameters, and then having obtained estimates of the algebraic functions, to try to solve the upcoming algebraic equations for the variables in those equations. This later problem can be seen as a problem that was studied by Ramanujan [9], namely solving a system of algebraic equations, which in our setting can be written as, solving for $(\boldsymbol{q}, \boldsymbol{p}) \in \mathbb{R}^{2s}$ the system of k equations

$$\sum_{r=1}^{s}(1-p_r)p_r^{i-1}q_r \quad = \quad \hat{b}_n^{(i)},$$

for $i = 1, \ldots, k$, where $\hat{b}_n^{(i)}$ are given numbers, cf. (D.8) below. The solution was given by Ramanujan [9] and with later refinements given e.g. in [8]. In our setting $\boldsymbol{q}$ denotes the distribution of the wavelengths, while $\boldsymbol{p}$ denotes the thinning probabilities for respective wavelength. As shown by Ramanujan, the necessary number of equations to obtain a solution is $2s-1$, and thus k above should be $2s-1$.

Using standard results on almost sure consistency and asymptotic normality for the mle, coupled with continuity and differentiability of the function that defines the solution of the Ramanujan equations, via the continuous mapping theorem and the delta method, we obtain almost sure consistency and asymptotic normality of the desired mle of $(\boldsymbol{q}, \boldsymbol{p})$. Taking into account the fact that the set of frequencies in a beam often is a basic frequency and its overtones, or equivalently that the set of wavelengths consist of a dominant wavelength and its fractions, it makes sense to model the wavelength distribution $\boldsymbol{q}$ as a decreasing sequence. This is a motivation for finding an order restricted estimator of $\boldsymbol{q}$, and we therefore propose the l^2 projection of the unrestricted mle of $\boldsymbol{q}$ on the set of decreasing probability mass functions. We are then able to use results on consistency and limit distributions for such isotonic regression estimators, see [7], [10] for results for i.i.d data and [4] for general results.

The remainder of the paper is organised as follows. In Section 2 we give a detailed description of the detector model that is being used, of the Poisson process model for the beam and of the Poisson data generated from the detector, cf. Lemma 2.1. In Section 3 we study the likelihood approach to estimating the parameters $(\boldsymbol{q}, \boldsymbol{p})$, and the system of algebraic equations that facilitates the estimation. In Theorem 3.2 we show that if the number of equations is

$k = 2s - 1$ then there is a unique mle of $(\boldsymbol{q}, \boldsymbol{p})$, obtained by the solution of the algebraic equations. In Subsection 3.2 and Theorem 3.4 we derive the consistency and asymptotic normality of the mle of $(\boldsymbol{q}, \boldsymbol{p})$. In Subsection 4.1 we define an order restricted estimator of $\boldsymbol{q}$ and state its consistency and asymptotic distribution in Theorem 4.1. Finally in Section 5 we discuss the obtained results and some remaining and interesting future problems.

2 Motivation and description of the data generating mechanism

The inference problem is motivated by the following problem that arises in neutron detection. Assume that a neutron beam is pointed at a detector. We model the number of neutrons that arrive at the face of the detector in the time interval $[0, t]$ by a counting process $X_0(t)$. Assume that the neutron beam, i.e. the process $X_0(t)$, has constant intensity λ. Assume furtherthermore that there are $s > 1$ different kinds of neutrons in the beam, with different wavelengths $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_s)$, such that

$$\mu_1 < \mu_2 < ... < \mu_s. \tag{D.1}$$

The values of the wavelengths are assumed to be unknown. We assume that we do however know the order in $(D.1)$, and can thus distinguish which label i to put on a neutron and its wavelengths placement in $(D.1)$, cf. Section 5 for a discussion on possible extensions.

We model the neutron beam, or counting process $X_0(t)$, as the sum of the counting processes that count the number of neutrons that arrive at the face of the detector in $[0, t]$, for the individual type neutrons. Thus we let the number of neutrons with wavelength μ_r, which we may label r-neutrons, be denoted by $X_0^{(r)}(t)$, where $X_0^{(r)}(t)$ is a counting process such that $X_0^{(r)}(t) = 0$ and with intensity λ_r, for $r = 1, \ldots, s$. We write $X_0(t) = \sum_{r=1}^{s} X_0^{(r)}(t)$ for the total number of neutrons that arrive at the face of the detector; then $X_0(t)$ is a counting process with $X_0(0) = 0$.

For a given number $X_0(t) = x_0$ of the total incoming neutrons in the time interval $[0, t]$, the vector $\left(X_0^{(1)}(t), X_0^{(2)}(t), ..., X_0^{(s)}(t)\right)$ is assumed to follow a multinomial distribution with parameters $\left(q_1, q_2, ..., q_s\right)$, i.e.

$$(X_0^{(1)} = x_0^{(1)}, ..., X_0^{(s)} = x_0^{(s)} | X_0 = x_0) \quad \in \quad Mult(x_0, q_1, q_2, ..., q_s), \qquad \text{(D.2)}$$

with

$$\begin{aligned} x_0^{(1)} + \cdots + x_0^{(s)} &= x_0, \\ q_1 + q_2 + ... + q_s &= 1. \end{aligned}$$

The vector of proportions of numbers of different neutrons $\boldsymbol{q} = (q_1, q_2, ..., q_s)$ is the spectrum, or distribution, of an incoming neutron beam $X_0(t)$. We note that $q_r = \lambda_r / \lambda$ and assume that $\boldsymbol{q}$ does not depend on t.

Now assume that the incident beam $X_0(t)$ is a Poisson process with intensity λ. In this case the components $X_0^{(r)}(t), r = 1, \ldots, s$ of the beam are independent Poisson processes with intensities $q_r\lambda$, for $r = 1, \ldots, s$, because the vector $\left(X_0^{(1)}(t), X_0^{(2)}(t), ..., X_0^{(s)}(t)\right)$ is the thinning of the original Poisson process, cf. e.g. [6]).

We next introduce the the so called multilayer detector that may be used in this setting. We assume that detector consists of fixed number of layers, say $k > 1$ layers, cf. Fig. D.1. The value of k will be elaborated on below, and will be shown to be in principle determined by the number of components in the spectrum, i.e. by the number of different types of neutrons that are present in the neutron beam.

The detection of neutrons in the multilayer detector can be described as follows. When an incident beam of neutrons hits a layer of the detector each neutron in the beam can possibly be absorbed, and then detected, or otherwise not be absorbed. If the neutron is not absorbed it will go through the present layer and will subsequently arrive at the next layer. We assume that at each layer, absorption or passing through are the only possibilities for the neutron interactions with the layer. We also assume that at each layer, different neutron particles interact with the layer independently of each other, i.e. at each layer the absorptions of different neutrons are independent events.

Let $\boldsymbol{p} = (p_1, \ldots, p_s)$ be the vector of probabilities of transmition (the thinning parameters), so that $1 - p_r$ is the probability of absorption for r-neutron, for $r = 1, \ldots, s$. It is a physical property of the neutron beam that the probability of transmission decreases with the neutron wavelength, cf. [2] and references therein, and therefore the thinning parameters can be modelled as a decreas-

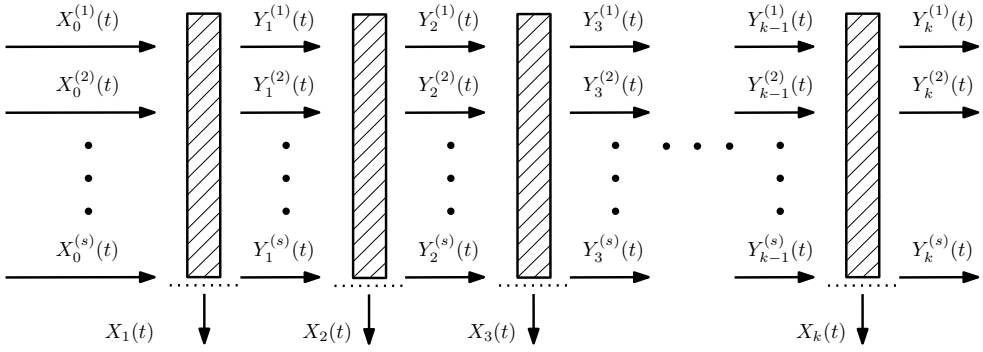

Figure D.1: The scheme of the detector. Here $Y_i^{(r)}(t)$ is the number of transmitted r-neutrons and $X_i(t) = \sum_{r=1}^{s} X_i^{(r)}(t)$ is the total number of the neutrons absorbed at the layer i.

ing sequence

$$1 > p_1 > p_2 > ... > p_s > 0. \tag{D.3}$$

Let us consider a beam of r-neutrons and denote the number of r-neutrons that are absorbed at the first layer by $X_1^{(r)}(t)$, so that $Y_1^{(r)}(t) = X_0^{(r)}(t) - X_1^{(r)}(t)$ is the number of r-neutrons transmitted. Then $X_1^{(r)}(t)$ and $Y_1^{(r)}(t) = X_0^{(r)}(t) - X_1^{(r)}(t)$ are non-decreasing counting processes, obtained by the thinning of the original Poisson process $X_0^{(r)}(t)$, so that $X_1^{(r)}(t)$ and $Y_1^{(r)}(t)$ are independent Poisson processes with intensities $(1 - p_r)q_r\lambda$ and $p_r q_r \lambda$, respectively, cf. [6].

Now assume that the transmitted beam $Y_1^{(r)}(t)$ hits the second layer, at which, again, each r-neutron can be either absorbed or transmitted. Let $X_2^{(r)}(t)$ be the number of absorbed neutrons and $Y_2^{(r)}(t) = Y_1^{(r)}(t) - X_2^{(r)}(t)$ the number of transmitted neutrons, at the second layer. Then, again, $X_2^{(r)}(t)$ and $Y_2^{(r)}(t)$ are obtained by thinning of the Poisson process $Y_1^{(r)}(t)$ and therefore they are independent Poisson processes, with intensities $p_r(1 - p_r)q_r\lambda$ and $p_r p_r q_r \lambda$, respectively [6]. By iterating the argument, cf. also [3] for a similar and more detailed reasoning, we obtain the following result.

Lemma 2.1. *$\{X_i(t)\}$, for $i = 1, \ldots, k$, are jointly independent Poisson processes with the rates $\sum_{r=1}^{s}(1 - p_r)p_r^{i-1}q_r\lambda$, respectively.*

One can state the goal in this paper as the estimation of the wavelength distribution $\boldsymbol{q}$ of the incident beam as well as of the actual values of the wavelengths $\boldsymbol{\mu}$, based on observations of the (total) Poisson process, and with the use of the

multilayer neutron detector, described above. Estimators of the wavelength values $\boldsymbol{\mu}$ can be indirectly obtained via estimates of the thinning parameters $\boldsymbol{p}$, using a functional relation between wavelength and thinning probability, as explained in [3]. The main goal of the paper will however be the estimation of wavelength distribution $\boldsymbol{q}$ but we will also state estimator for the thinning probabilities $\boldsymbol{p}$.

3 Inference for the parameters

In this section we state the inference problem, define the mle of the parameters $(\boldsymbol{p}, \boldsymbol{q})$, state conditions for its existence, and derive consistency and asymptotic normality for the mle of $(\boldsymbol{p}, \boldsymbol{q})$. Subsequently we introduce an order restricted estimator of $\boldsymbol{q}$ and derive its consistency and limit distribution.

We start by the following note on the experimental setup, and the data: In order to derive the limit properties for the estimator, we need to define what we mean by "letting n go to infinity". This may be done in, at least, two ways. We can either let the time t go to infinity, and view the data as stemming from on Poisson process run for a (very) long time, or we can keep the time t fixed and gather data from several independent Poisson process runs, cf. [3] a more detailed discussion about advantages and disadvantages with respective approach.

We will view the estimation problem as a repeated sample problem. Thus we assume that during a fixed time interval $[0, t]$ and for fixed intensity λ, i.e. fixed intensities $(\lambda_1, \ldots, \lambda_r)$, of an incident beam there are n repeated measurements. Let $x_{i,j}$ be the observed number of neutrons at layer i, for $i = 1, \ldots, k$, at the experiment round j, $j = 1, \ldots, n$. Then at each experiment round j the vector $\mathbf{X_j} = (X_{1j}, \ldots, X_{kj})$ is distributed as according to Lemma 2.1, and furthermore the vectors $\mathbf{X_1}, \ldots, \mathbf{X_n}$ are assumed to be independent.

Thus the inference problem is to, given data as above, estimate the pair $(\boldsymbol{q}, \boldsymbol{p})$ subject to them lying in the parameter space $\mathcal{F} \subset \mathbb{R}_+^{2s}$ which is given by

$$\begin{aligned} \mathcal{F} &= \{(\boldsymbol{q}, \boldsymbol{p}) \in \mathbb{R}_+^{2s} : q_1 + q_2 + \cdots + q_s = 1 \\ &\quad 1 > p_1 > p_2 > \cdots > p_s > 0\}. \end{aligned} \tag{D.4}$$

Note that $\boldsymbol{q}$ is a probability mass function while $\boldsymbol{p}$ is merely a vector of probabilities. We would like to emphasize here that the main object of study is the

wavelength distribution $\boldsymbol{q}$. The thinning probabilities $\boldsymbol{p}$ however are also of interest, since they determine the values of the wavelengths, which we assume are unknown; if we know the actual wavelength values there is no need to estimate the thinning probabilities. Note that we do however know the order of the wavelength values, cf. (D.1). See Section 5 for further comments on this.

We will use the likelihood approach for making inference about the unknown parameters $(\boldsymbol{q}, \boldsymbol{p})$. We define the maximum likelihood estimator (mle) of $(\boldsymbol{q}, \boldsymbol{p})$ by

$$(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n) \quad = \quad \underset{(\boldsymbol{q},\boldsymbol{p})\in\mathcal{F}}{\operatorname{argmax}}\, l_n(\boldsymbol{q}, \boldsymbol{p}), \tag{D.5}$$

where

$$l_n(\boldsymbol{q}, \boldsymbol{p}) = \sum_{j=1}^{n}\sum_{i=1}^{k}(-\lambda t m_i + x_{i,j}\log m_i + x_{i,j}\log(\lambda t) - \log x_{i,j}!) \tag{D.6}$$

is the log likelihood, and

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

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

3.1 Existence and uniqueness of the mle

In this subsection we prove the existence of the mle $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n)$, introduced in (D.5), and obtain an explicit expression for it.

First, we note that working directly with the parameters $(\boldsymbol{q}, \boldsymbol{p})$, we obtain the first derivatives of $l_n(\boldsymbol{q}, \boldsymbol{p}|\mathbf{x})$ and trying to solve the score equations proves to be quite cumbersome. Moreover, one can show that the log-likelihood $l_n(\boldsymbol{q}, \boldsymbol{p}|\mathbf{x})$ seen as a function on the parameter space $\mathcal{F} \subset \mathbb{R}^{2s}$ is not a concave function, which makes it difficult to find a solution $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n)$ even numerically.

We will therefore reparametrise the problem as an inference problem for the vector $(m_1, \ldots, m_k)$ of expected total numbers of observed neutrons, divided by λt, and having found a solution to this simpler inference problem, solve an upcoming system of equations for obtaining the solution to (D.5). Introduce

the notation $\hat{\boldsymbol{b}}_n = (\hat{b}_n^{(1)}, \ldots, \hat{b}_n^{(k)})$, where

$$\hat{b}_n^{(i)} = \frac{\sum_{j=1}^{n} x_{i,j}}{n\lambda t},$$

We then rewrite (D.6) as

$$\begin{aligned} g(m_1, \ldots, m_k) &:= \sum_{i=1}^{k} (-m_i + \hat{b}_n^{(i)} \log m_i) \\ &= \frac{l_n(\boldsymbol{q}, \boldsymbol{p}|\mathbf{x})}{n\lambda t} \end{aligned} \tag{D.7}$$

and note that we have dropped the last two terms in (D.6), in the last equality. The function $g(m_1, \ldots, m_k)$ reaches its unique global maximum at $\hat{m}_i \hat{b}_n^{(i)}$, for $i = 1, \ldots, k$. Therefore, if $(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$ is a solution to the following system of equations

$$\begin{cases} m_1(\boldsymbol{q}, \boldsymbol{p}) = \hat{b}_n^{(1)}, \\ m_2(\boldsymbol{q}, \boldsymbol{p}) = \hat{b}_n^{(2)}, \\ \qquad \cdots \\ m_k(\boldsymbol{q}, \boldsymbol{p}) = \hat{b}_n^{(k)}, \end{cases} \tag{D.8}$$

and if it satisfies the constraints in (D.4), then $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n) = (\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$, i.e. the solution is the mle.

In order to reformulate the system of equations (D.8) on matrix form, we introduce the vectors $\hat{\boldsymbol{a}}_n = (\hat{a}_n^{(1)}, \ldots, \hat{a}_n^{(k+1)})$, where

$$\begin{aligned} \hat{a}_n^{(1)} &= 1, \\ \hat{a}_n^{(i)} &= 1 - \sum_{l=1}^{i-1} \hat{b}_n^{(l)}, \end{aligned} \tag{D.9}$$

for $i = 2, \ldots, k+1$, and for $\boldsymbol{u} \in \mathbb{R}^{2s}$ we define the matrices $\boldsymbol{C}(\boldsymbol{u})$ and $\boldsymbol{D}(\boldsymbol{u})$ as

$$\boldsymbol{C}(\boldsymbol{u}) = \begin{pmatrix} u_s & u_{s-1} & u_{s-2} & \cdots & u_1 \\ u_{s+1} & u_s & u_{s-1} & \cdots & u_2 \\ u_{s+2} & u_{s+1} & u_s & \cdots & u_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{2s-1} & u_{2s-2} & u_{2s-3} & \cdots & u_s \end{pmatrix} \tag{D.10}$$

and

$$\boldsymbol{D}(\boldsymbol{u}) = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ u_1 & 0 & 0 & \cdots & 0 \\ u_2 & u_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{s-1} & u_{s-2} & u_{s-3} & \cdots & u_1 \end{pmatrix} \tag{D.11}$$

We next obtain a preliminary result saying that, for large enough n, the random matrix $\boldsymbol{C}(\hat{\boldsymbol{a}}_n)$ is non-singular, almost surely.

Lemma 3.1. *There exists n_1 such that for any $n > n_1$*

$$\mathbb{P}[\det(\boldsymbol{C}(\hat{\boldsymbol{a}}_n)) \neq 0] \quad = \quad 1.$$

Proof. First, note that from the strong law of large numbers one has

$$\boldsymbol{C}(\hat{\boldsymbol{a}}_n) \quad \stackrel{as}{\to} \quad \boldsymbol{C}(\boldsymbol{a}),$$

where $\boldsymbol{a}$ denotes the a.s. limit of the sequence $\hat{\boldsymbol{a}}_n$ and, therefore, the matrix $\boldsymbol{C}(\boldsymbol{a})$ is given by

$$\boldsymbol{C} = \begin{pmatrix} \sum_{r=1}^s p_r^{s-1} q_r & \sum_{r=1}^s p_r^{s-2} q_r & \sum_{r=1}^s p_r^{s-3} q_r & \cdots & \sum_{r=1}^s q_r \\ \sum_{r=1}^s p_r^{s} q_r & \sum_{r=1}^s p_r^{s-1} q_r & \sum_{r=1}^s p_r^{s-2} q_r & \cdots & \sum_{r=1}^s p_r q_r \\ \sum_{r=1}^s p_r^{s+1} q_r & \sum_{r=1}^s p_r^{s} q_r & \sum_{r=1}^s p_r^{s-1} q_r & \cdots & \sum_{r=1}^s p_r^2 q_r \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{r=1}^s p_r^{2s-2} q_r & \sum_{r=1}^s p_r^{2s-3} q_r & \hat{a}_n^{(2s-3)} & \cdots & \sum_{r=1}^s p_r^{s-1} q_r \end{pmatrix} \tag{D.12}$$

with $q_1, \ldots, q_s, p_1, \ldots, p_s$ the true values of the parameters $\boldsymbol{q}$ and $\boldsymbol{p}$. Next, note that $\boldsymbol{C}(\hat{\boldsymbol{a}}_n)$ can be diagonalized as

$$\boldsymbol{C}(\hat{\boldsymbol{a}}_n) = \boldsymbol{V}\boldsymbol{Q}\boldsymbol{V}^T,$$

where

$$\boldsymbol{V} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ p_1 & p_2 & p_3 & \cdots & p_s \\ p_1^2 & p_2^2 & p_3^2 & \cdots & p_s^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_1^{s-1} & p_2^{s-1} & p_3^{s-1} & \cdots & p_s^{s-1} \end{pmatrix}$$

and $\boldsymbol{Q}$ is a diagonal matrix with diagonal elements given by the vector $\boldsymbol{q}$. Since $\boldsymbol{V}$ is a square Vandermonde matrix and $p_1 > p_2 > ... > p_s$, it is full-rank i.e. $rank(C) = s$. Therefore $\det(\boldsymbol{C}(\hat{\boldsymbol{a}}_n)) \neq 0$. The continuous mapping theorem then implies

$$\det(\boldsymbol{C}(\hat{\boldsymbol{a}}_n)) \quad \rightarrow \quad \det(\boldsymbol{C}(\boldsymbol{a})),$$

almost surely, which implies the statement of the lemma. □

Next, we study the system of equations in (D.8). We will follow closely Ramanujan's derivation of the solution, cf. [9]. Define the function

$$\varphi(\theta) = \frac{d_1 + d_2\theta + d_3\theta^2 + \cdots + d_s\theta^{s-1}}{1 + c_1\theta + c_2\theta^2 + \cdots + c_s\theta^s},$$

with the vectors $\boldsymbol{c} = (c_1, \ldots, c_s), \boldsymbol{d} = (d_1, \ldots, d_s)$ given by

$$\begin{aligned} \boldsymbol{c} &= \boldsymbol{C}(\hat{\boldsymbol{a}}_n)^{-1}[\hat{\boldsymbol{a}}_n]^{(s+1,2s)}, \\ \boldsymbol{d} &= [\hat{\boldsymbol{a}}_n]^{(1,s)} + \boldsymbol{D}(\hat{\boldsymbol{a}}_n)\boldsymbol{c}, \end{aligned}$$

and where $[\hat{\boldsymbol{a}}_n]^{(i,j)}$ denotes the restriction of the vector $\hat{\boldsymbol{a}}_n$ in $\mathbb{R}^{k+1}$ to the index set (i, j). The next result says that if $\boldsymbol{C}(\hat{\boldsymbol{a}}_n)$ is nonsingular and if we have a certain relation between the number of layers and the support of the wavelength distribution, then the mle exists, and is unique, up to permutations of the indices.

Theorem 3.2. *Assume that $k = 2s - 1$ and $\det(\boldsymbol{C}(\hat{\boldsymbol{a}}_n)) \neq 0$. Then the solution to (D.8) is unique, up to permutations of the indices, and is given by*

$$(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n) = (\boldsymbol{y}, \boldsymbol{z}),$$

where $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^s$ are the coefficients in the following representation of $\varphi(\theta)$,

$$\varphi(\theta) = \frac{y_1}{1 - z_1\theta} + \frac{y_2}{1 - z_2\theta} + \cdots + \frac{y_s}{1 - z_s\theta}. \tag{D.13}$$

Proof. First, recall that the system in (D.8) is given by

$$\begin{cases} q_1(1 - p_1) + q_2(1 - p_2) + \cdots + q_s(1 - p_s) = \hat{b}_n^{(1)}, \\ q_1(1 - p_1)p_1 + q_2(1 - p_2)p_2 + \cdots + q_s(1 - p_s)p_s = \hat{b}_n^{(2)}, \\ \cdots \\ q_1(1 - p_1)p_1^{k-1} + q_2(1 - p_2)p_2^{k-1} + \cdots + q_s(1 - p_s)p_s^{k-1} = \hat{b}_n^{(k)}. \end{cases} \tag{D.14}$$

□

Note, that (D.14) can be simplified as

$$\begin{cases} q_1 + q_2 + \cdots + q_s = \hat{a}_n^{(1)} \\ q_1 p_1 + q_2 p_2 + \cdots + q_s p_s = \hat{a}_n^{(2)} \\ q_1 p_1^2 + q_2 p_2^2 + \cdots + q_s p_s^2 = \hat{a}_n^{(3)} \\ \cdots \\ q_1 p_1^k + q_2 p_2^k + \cdots + q_s p_s^k = \hat{a}_n^{(k+1)} \end{cases} \tag{D.15}$$

The system of equations in (D.15) for $k = 2s - 1$ was studied and solved by Ramanujan in his third paper, published in the Journal of Indian Mathematical Society cf. [9]. From the results in [9] it follows that if $\det(\boldsymbol{C}(\hat{\boldsymbol{a}}_n)) \neq 0$, then the solution of (D.15) exist, it is unique, up to permutations of the indices $\{1, \ldots, s\}$, and given by $(\boldsymbol{y}, \boldsymbol{z})$, the coefficients in the parametrisation (D.13). □

Since the solution is invariant under permutation of the indices, we may choose any permutation as the correct, and since we know the order for the wavelength values, cf. (D.1), the choice is simple: we choose the known and correct order.

3.2 Asymptotic properties of the mle

Before we obtain the asymptotic distribution of the estimator we prove an auxiliary lemma. Assume that $k = 2s - 1$. We may rewrite the system of equations (D.15) as

$$\begin{cases} F_1(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{u}) = 0 \\ F_2(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{u}) = 0 \\ \cdots \\ F_{2s}(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{u}) = 0 \end{cases}, \tag{D.16}$$

with $\boldsymbol{u} = \hat{\boldsymbol{a}}_n$, where the functions $F_i : \mathbb{R}^{3s} \to \mathbb{R}$ are given by

$$F_i(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{u}) = q_1 p_1^{i-1} + q_2 p_2^{i-1} + \cdots + q_s p_s^{i-1} - u_i,$$

for $i = 1, \ldots, 2s$. We see that the system of equations in (D.16) gives an implicit definition of a function $\boldsymbol{\psi}(\boldsymbol{u}) : \mathbb{R}^{2s} = \mathbb{R}^{k+1} \ni \boldsymbol{u} \to (\boldsymbol{q}, \boldsymbol{p}) \in \mathbb{R}^{2s}$. The Jacobian

matrix for the system (D.16) is then given by

$$\boldsymbol{J}(\boldsymbol{q},\boldsymbol{p}) = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ p_1 & \cdots & p_s & q_1 & \cdots & q_s \\ p_1^2 & \cdots & p_s^2 & 2q_1p_1 & \cdots & 2q_sp_s \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p_1^{2s-1} & \cdots & p_s^{2s-1} & (2s-1)q_1p_1^{2s-2} & \cdots & (2s-1)q_sp_s^{2s-2} \end{pmatrix} \quad \text{(D.17)}$$

The next lemma shows that the function $\boldsymbol{\psi}(\boldsymbol{u})$, implicitly defined by the equations (D.16), is differentiable.

Lemma 3.3. *Assume that $\boldsymbol{u}$ is such that $\det(\boldsymbol{C}(\boldsymbol{u})) \neq 0$. Then the function $\boldsymbol{\psi}$, implicitly defined by (D.16), is differentiable at the point $\boldsymbol{u}$.*

Proof. The statement of the lemma will follow from the implicit function theorem, for which we now check the conditions.

First we note that (D.16) is a rewriting of (D.15) which is a simplification of (D.14) which is identical to (D.8). Theorem 1 says that if $\det(\boldsymbol{C}(\boldsymbol{u})) \neq 0$, at some $\boldsymbol{u}$, then there are unique $(\boldsymbol{q},\boldsymbol{p})$ which satisfy (D.8), which implies that $(\boldsymbol{q},\boldsymbol{p})$ are unique solutions to (D.16).

Second, the functions $F_i(\boldsymbol{q},\boldsymbol{p},\boldsymbol{u})$, for $i = 1,\ldots,2s$, are differentiable and continuous.

It remains to prove that the Jacobian $\boldsymbol{J}$ in (D.17) is a non-singular matrix, i.e. to show that $\det(\boldsymbol{J}) \neq 0$. In fact, we note that $\boldsymbol{q}$ can be factored out of the determinant, i.e.

$$\det(\boldsymbol{J}(\boldsymbol{q},\boldsymbol{p})) = q_1 \cdots q_s \cdot \det(\boldsymbol{W}(\boldsymbol{p})),$$

where

$$\boldsymbol{W}(\boldsymbol{p}) = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ p_1 & \cdots & p_s & 1 & \cdots & 1 \\ p_1^2 & \cdots & p_s^2 & 2p_1 & \cdots & 2p_s \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p_1^{2s-1} & \cdots & p_s^{2s-1} & (2s-1)p_1^{2s-2} & \cdots & (2s-1)p_s^{2s-2} \end{pmatrix}. \quad \text{(D.18)}$$

We rewrite $\boldsymbol{W}(\boldsymbol{p})$ on column matrix form as

$$\boldsymbol{W}(\boldsymbol{p}) = [\boldsymbol{w}(p_1), \boldsymbol{w}(p_2), \ldots, \boldsymbol{w}(p_s), \boldsymbol{w}^{(1)}(p_1), \boldsymbol{w}^{(1)}(p_2), \ldots, \boldsymbol{w}^{(1)}(p_s)], \quad \text{(D.19)}$$

where $\boldsymbol{w}(p) = (1, p, p^2, \ldots, p^{2s})^T$, and $\boldsymbol{w}^{(1)}(p)$ denotes the vector of componentwise first derivatives of the column $\boldsymbol{w}(p)$.

Consider $\rho(x) = \det(\boldsymbol{W}(x, p_2, \ldots, p_s))$, which is a polynomial of order $(4s-4)$ in x. Let us show that the multiplicity of the component p_2 of the root $(\boldsymbol{p}, \boldsymbol{q})$ is equal to 4. The third derivative $\rho^{(3)}(x)$ of the polynomial is equal to

$$\begin{aligned}
\rho^{(3)}(x) &= \det([\boldsymbol{w}(x)^{(3)}, \boldsymbol{w}(p_2), \ldots, \boldsymbol{w}(p_s), \boldsymbol{w}^{(1)}(x), \boldsymbol{w}^{(1)}(p_2), \ldots, \boldsymbol{w}^{(1)}(p_s)]) \\
&+ 3\det([\boldsymbol{w}(x)^{(2)}, \boldsymbol{w}(p_2), \ldots, \boldsymbol{w}(p_s), \boldsymbol{w}^{(2)}(x), \boldsymbol{w}^{(1)}(p_2), \ldots, \boldsymbol{w}^{(1)}(p_s)]) \\
&+ 3\det([\boldsymbol{w}(x)^{(1)}, \boldsymbol{w}(p_2), \ldots, \boldsymbol{w}(p_s), \boldsymbol{w}^{(3)}(x), \boldsymbol{w}^{(1)}(p_2), \ldots, \boldsymbol{w}^{(1)}(p_s)]) \\
&+ \det([\boldsymbol{w}(x), \boldsymbol{w}(p_2), \ldots, \boldsymbol{w}(p_s), \boldsymbol{w}^{(4)}(x), \boldsymbol{w}^{(1)}(p_2), \ldots, \boldsymbol{w}^{(1)}(p_s)]).
\end{aligned}$$

It follows that for $x = p_2$, each term contains two equal columns. Therefore, we have proved that $\rho^{(3)}(x) = 0$ at $x = p_2$, which implies that the multiplicity of p_2 is at least 4. Now since $\det(\boldsymbol{W}(p_1, p_2, \ldots, p_s))$ is symmetric (with no sign change, since flipping two of the arguments p_i, p_j means flipping four columns in the matrix at once), then any p_i, for $i = 2, \ldots, s$ is also a root of $\rho(x) = \det(\boldsymbol{W}(x, p_2, \ldots, p_s))$, and the same argument as above shows that they all have multiplicity at least 4. Since there are $s-1$ roots and $\rho(x)$ is of order $(4s-4)$, the multiplicity is exactly 4, for each root. Therefore, we have shown that

$$\det(\boldsymbol{W}(x, p_2, \ldots, p_s)) = c\prod_{j=2}^{s}(x - p_i)^4,$$

where c is a leading coefficient. Using the symmetry of the determinant, we may replace any of the p_i's with x and study the upcoming polynomial, to obtain

$$\det(\boldsymbol{J}(\boldsymbol{q}, \boldsymbol{p})) = c_1 q_1 \cdots q_s \prod_{p_i \neq p_j} (p_i - p_j)^4,$$

where c_1 is a constant.

Thus, we have shown that $\det(\boldsymbol{J}) \neq 0$, provided $p_i \neq p_j$ for all $i \neq j$. The fact that the unique solution $(\boldsymbol{p}, \boldsymbol{q})$ to $(D.8)$ satisfies $p_i \neq p_j$ for all $i \neq j$ follows from a refinement of Ramanujan's theorem, given in [8]. □

Theorem 3.4. *Let $k = 2s - 1$. Then the mle $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n)$ in (D.5) is strongly consistent*

$$(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n) \xrightarrow{a.s.} (\boldsymbol{q}, \boldsymbol{p}),$$

and asymptotically normal

$$\sqrt{n}((\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n) - (\boldsymbol{q}, \boldsymbol{p})) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^2),$$

as $n \to \infty$.

Proof. From Lemma 3.3 it follows that for $\boldsymbol{u}$, such that $\det(\boldsymbol{C}(\boldsymbol{u}) \neq 0$, the system of equations in (D.16) gives an implicit definition of a differentiable function $\boldsymbol{\psi} : \mathbb{R}^{2s} \to \mathbb{R}^{2s}$. Let $\boldsymbol{a}$ denote the a.s. limit of the sequence $\hat{\boldsymbol{a}}_n$, and recall that, because of the definition of the matrix $\boldsymbol{C}(a)$ in (D.12), and of the function $\boldsymbol{\psi}$,

$$(\boldsymbol{q}, \boldsymbol{p}) = \boldsymbol{\psi}(\boldsymbol{a}). \tag{D.20}$$

Combining Theorem 3.2 and Lemma 3.1, it follows that there exists an n_1, such that for all $n > n_1$, $(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$ is the solution to (D.8), so that furthermore one has

$$(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n) \overset{a.s.}{=} \boldsymbol{\psi}(\hat{\boldsymbol{a}}_n). \tag{D.21}$$

Recall that we can claim that $(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$ is equal to the mle $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n)$ only when the restrictions in $\mathcal{F}$, cf. (D.4), are satisfied for $(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$, which we will check below. Then, from (D.20), (D.21) and since $\hat{\boldsymbol{a}}_n \xrightarrow{a.s.} \boldsymbol{a}$, using the continuous mapping theorem we obtain the consistency result

$$(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n) \xrightarrow{a.s.} (\boldsymbol{q}, \boldsymbol{p}).$$

for $(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$.

Now, let us consider the vector $\hat{\boldsymbol{a}}_n$, defined in (D.9). Note that $[\hat{\boldsymbol{a}}_n]^{(2,2s)}$ can be written as

$$[\hat{\boldsymbol{a}}_n]^{(2,2s)} = \mathbf{1} - \boldsymbol{L}\boldsymbol{b}_n,$$

where $\boldsymbol{L}$ is a lower triangular $(2s-1) \times (2s-1)$ matrix of ones. Using a central limit theorem one can show that

$$\sqrt{n}([\hat{\boldsymbol{a}}_n]^{(2,2s)} - [\boldsymbol{a}]^{(2,2s)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_A^2), \tag{D.22}$$

as $n \to \infty$, where

$$\boldsymbol{\Sigma}_A^2 = \boldsymbol{L}\boldsymbol{\Sigma}_m^2\boldsymbol{L}^T,$$

with $\boldsymbol{\Sigma}_m^2 = diag([\boldsymbol{m}]^{(1,2s-1)})$. Recall that the first element of $\hat{\boldsymbol{a}}_n$ is deterministic and equals 1, cf. (D.9), and thus we do not include it in the limit result (D.22).

Let $\boldsymbol{\partial\psi}(\boldsymbol{u})$ be the matrix of partial derivatives of $\psi(\boldsymbol{u})$, i.e.

$$\boldsymbol{\partial\psi}(\boldsymbol{u}) = \begin{pmatrix} \frac{\partial\psi_1}{\partial u_1}(\boldsymbol{u}) & \frac{\partial\psi_1}{\partial u_2}(\boldsymbol{u}) & \frac{\partial\psi_1}{\partial u_3}(\boldsymbol{u}) & \cdots & \frac{\partial\psi_1}{\partial u_{2s}}(\boldsymbol{u}) \\ \frac{\partial\psi_2}{\partial u_1}(\boldsymbol{u}) & \frac{\partial\psi_2}{\partial u_2}(\boldsymbol{u}) & \frac{\partial\psi_2}{\partial u_3}(\boldsymbol{u}) & \cdots & \frac{\partial\psi_2}{\partial u_{2s}}(\boldsymbol{u}) \\ \frac{\partial\psi_3}{\partial u_1}(\boldsymbol{u}) & \frac{\partial\psi_3}{\partial u_2}(\boldsymbol{u}) & \frac{\partial\psi_3}{\partial u_3}(\boldsymbol{u}) & \cdots & \frac{\partial\psi_3}{\partial u_{2s}}(\boldsymbol{u}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\psi_{2s}}{\partial u_1}(\boldsymbol{u}) & \frac{\partial\psi_{2s}}{\partial u_2}(\boldsymbol{u}) & \frac{\partial\psi_{2s}}{\partial u_3}(\boldsymbol{u}) & \cdots & \frac{\partial\psi_{2s}}{\partial u_{2s}}(\boldsymbol{u}) \end{pmatrix} \tag{D.23}$$

The values of $\boldsymbol{\partial\psi}(\boldsymbol{u})$ can be found using the implicit function theorem. In fact, the j-th column $\boldsymbol{\partial\psi}(\boldsymbol{u})[j]$ of $\boldsymbol{\partial\psi}(\boldsymbol{u})$ is the solution of the following system of linear equations

$$\boldsymbol{J}\boldsymbol{\partial\psi}(\boldsymbol{u})[j] = \mathbf{1}^{(j)},$$

where $\mathbf{1}^{(j)} \in \mathbb{R}^{2s}$ is defined by $\mathbf{1}_j^{(j)} = -1$ and $\mathbf{1}_l^{(j)} = 0$ for $l \neq j$, cf. (D.16). The solution exists and it is unique, when $\det(\boldsymbol{J}) \neq 0$, which is true for $\boldsymbol{u} = \hat{\boldsymbol{a}}_n$ for all $n \geq n_1$, and for $\boldsymbol{u} = \boldsymbol{a}$. Thus the matrices $\boldsymbol{\partial\psi}(\hat{\boldsymbol{a}}_n)$ are (uniquelly) given for all $n \geq n_1$, and so is the matrix $\boldsymbol{\partial\psi}(\boldsymbol{a})$.

Since the derivatives $\boldsymbol{\partial\psi}$ are continuous, using (D.22) and the delta method, cf. [11], we derive the limit distribution for $(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$,

$$\sqrt{n}((\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n) - (\boldsymbol{q}, \boldsymbol{p})) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^2),$$

as $n \to \infty$, with

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

where we use the notation $[\cdot]_{1:2s,2:2s}$ for denoting a matrix without the first column.

Finally, $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n) = (\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$ if and only if $(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n) \in \mathcal{F}$, with $\mathcal{F}$ defined in (D.4). Since $(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$ is strongly consistent, there exists an $n_2 > n_1$ such that for all $n > n_2$

$$\mathbb{P}[(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n) \in \mathcal{F}] = 1.$$

Thus, the mle $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n)$ is consistent, almost surely, and has the same asymptotic distribution as $(\tilde{\boldsymbol{q}}_n, \tilde{\boldsymbol{p}}_n)$, which ends the proof. □

We note the having obtained an estimator of $\boldsymbol{p}$ one can use a functional relation between a thinning probability and a wavelength value, cf. [12], similarly to as in Corollaries 1 and 2 in [3]. The derivation is straightforward and is omitted.

4 Order restricted estimation of the parameters

We note that the components in the mle are not necessarily ordered vectors, and that we have order restrictions on both the wavelength distribution $\boldsymbol{q}$ and the thinning probabilities $\boldsymbol{p}$.

We therefore address order restricted problems in this section. In Subsection 4.1 we treat order restricted inference for the wavelength distribution $\boldsymbol{q}$. The estimator in that subsection is obtained as the l^2 projection of the mle of $\boldsymbol{q}$ on the set of decreasing vectors. When projecting on this space we note that the order of the wavelengths $\mu_1, \ldots, \mu_s$ is assumed to be known. We note that since the system of algebraic equations used to obtained the mle is symmetric with respect to permutation of the s pairs $(p_i, q_i), i = 1, \ldots, s$, any order that we choose for the solution is ok; we choose however the the order that we know to be correct, cf. the comment after Theorem 3.2.

4.1 Estimating a decreasing wavelength distribution

In this subsection we assume that it is known that the wavelength distribution $\boldsymbol{q}$ is a decreasing vector and construct an appropriate estimator, based on the mle defined previously. In fact our estimator is the l^2 projection of the mle of $\boldsymbol{q}$ on the space of positive and decreasing vectors, i.e. the isotonic regression of the vector $\hat{\boldsymbol{q}}_n$.

We define the set $\mathcal{Q}^* \subset \mathbb{R}^s$

$$\mathcal{Q}^* = \{\boldsymbol{q} \in \mathbb{R}^s : q_1 \geq q_2 \geq \cdots \geq q_s \text{ for } r = 1, \ldots, s, \} \tag{D.24}$$

and assume that the true value satisfies $\boldsymbol{q} \in \mathcal{Q}^*$. Note first that since $\boldsymbol{q}$ is supposed to be a probability mass function, we should really demand that $\mathcal{Q}^*$ is a subset of positive s-dimensional vectors and furthermore that there should be a linear constraint. This is however not necessary when projecting a vector

that already is a probability mass function, since isotonic regression preserves linear constraints as well as upper and lower bounds of the vector, cf. [10] for these results and a general overview of order restricted inference.

We define the monotone constrained estimator of $\boldsymbol{q}$ as

$$\hat{\boldsymbol{q}}_n^* = \underset{q \in \mathcal{Q}^*}{\operatorname{argmin}} \sum_{r=1}^{s} (q_r - \hat{q}_{n,r})^2, \tag{D.25}$$

i.e. $\hat{\boldsymbol{q}}_n^*$ is the isotonic regression of the mle $\hat{\boldsymbol{q}}_n$. We note that from the error reduction property of the isotonic regression we have

$$||\hat{\boldsymbol{q}}_n^* - \boldsymbol{q}||_\alpha \leq ||\hat{\boldsymbol{q}}_n - \boldsymbol{q}||_\alpha \tag{D.26}$$

for all $\alpha \geq 1$, cf. [10].

In order to obtain the asymptotic distribution of $\hat{\boldsymbol{q}}_n^*$, we need to specify the exact shape of the pmf $\boldsymbol{q}$, since the shape of $\boldsymbol{q}$ will determine the limit distribution. In particular we need to specify the regions where $\boldsymbol{q}$ is constant. Thus we assume that the true vector $\boldsymbol{q} \in \mathbb{R}^s$ has the following structure

$$\begin{aligned} q_{t_1} = \cdots = q_{t_1+v_1-1} > q_{t_2} &= \cdots = q_{t_2+v_2-1} > \cdots > \\ q_{t_m} &= \cdots = q_s, \end{aligned} \tag{D.27}$$

where t_j for $j = 1, \ldots m$ is the index of the first element in the j-th flat region, $q_{t_1} = q_1$, m is the total number of flat regions of $\boldsymbol{q}$, $\boldsymbol{v} = (v_1, \ldots, v_m)$ is the vector of the lengths (the numbers of points) of the flat regions of $\boldsymbol{q}$, so that $\sum_{j=1}^{m} v_j = s$.

We define the map $\varphi = \varphi_{\mathbf{q}} : \mathbb{R}^s \to \mathbb{R}^s$ by specifying that for any $Y \in \mathbb{R}^s$, for all constant regions $(t_j, t_j + v_j - 1)$ of $\boldsymbol{q}$,

$$[\varphi(Y)]^{(t_j, t_j+v_j-1)} = \operatorname{argmin}_{\boldsymbol{y} \in \{\boldsymbol{y} \in \mathbb{R}^{v_j} : y_1 \geq \ldots \geq y_{v_j}\}} ||Y^{(t_j, t_j+v_j-1)} - \boldsymbol{y}||^2,$$

where $||\cdot||^2$ denotes the l^2-norm in $\mathbb{R}^{v_j}$, so that the values of $[\varphi(Y)]^{(t_j, t_j+v_j-1)}$ are given as the separate isotonic regression of Y over the region of constancy $(t_j, t_j + v_j - 1)$. Note that if the region of constancy is of length 1 then the isotonic regression of Y at that region (point) is equal to the value of Y at that point. With this definition, we see that $\varphi(Y)$ is a concatenation of separate isotonic regressions over each region of constancy of the true $\boldsymbol{q}$, cf. [7] and [4] for a more detailed description of the map (operator).

Finally we obtain consistency and the asymptotic distribution of the estimator $\hat{\boldsymbol{q}}_n^*$.

Theorem 4.1. *Suppose $\boldsymbol{q}$ satisfies (D.27), and let $k = 2s - 1$. Then the order restricted estimator $\hat{\boldsymbol{q}}_n^*$ defined in (D.25) is strongly consistent*

$$\hat{\boldsymbol{q}}_n^* \xrightarrow{a.s.} \boldsymbol{q},$$

and has the asymptotic distribution

$$\sqrt{n}(\hat{\boldsymbol{q}}_n^* - \boldsymbol{q}) \xrightarrow{d} \varphi(\boldsymbol{Q}_{\boldsymbol{q}}),$$

as $n \to \infty$, where $\boldsymbol{Q}_{\boldsymbol{q}}$ is the limit distribution of $\hat{\boldsymbol{q}}_n$, i.e. $\boldsymbol{Q}_{\boldsymbol{q}} = \mathcal{N}(\boldsymbol{0}, [\boldsymbol{\Sigma}^2]_{1:s,1:s})$, with $\boldsymbol{\Sigma}^2$ defined in Theorem 3.4.

Proof. The strong consistency follows from the consistency of the mle $\hat{\boldsymbol{q}}_n$ and the error reduction property of the isotonic regression. The asymptotic distribution of $\hat{\boldsymbol{q}}_n^*$ follows by Theorem 2 in [4], see also Theorem 5.2.1 in [10], and [7]. □

5 Discussion

In this paper we have derived the mle $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n)$ of the distribution of events of differnet types $\boldsymbol{q}$ of a multimode Poisson process and the thinning probabilities $\boldsymbol{p}$, based on data from sequential thinning of the Poisson process. We have established that the number, k, of sequential thinnings needed in order to solve a system of algebraic equations that determine the mle is $k = 2s - 1$, where s is the length of the vector $\boldsymbol{q}$, cf. Theorem 3.2. In Theorem 3.4 we derived the strong consistency and asymptotic normality of the mle $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{p}}_n)$. We have constructed an order restricted estimator $\hat{\boldsymbol{q}}_n^*$ of $\boldsymbol{q}$, and in Theorem 4.1 we derived the consistency and asymptotic distribution of $\hat{\boldsymbol{q}}_n^*$.

A possible way to improve the efficiency for the order restricted estimator may to use model selection to choose the appropriate class of probability mass functions $\boldsymbol{q}$, or of vectors of probabilities $\boldsymbol{p}$. The model class may be determined by the regions of constancy, as defined (D.27). One advantage with having knowledge about the specific sets of regions for the unknown vector is that one can then use the knowledge to construct an order restricted estimator that outperforms the regular isotonic regression estimator, as shown in [5] . In [5] we have introduced an information criterion which can be used

for model selection in order restricted inference and also we have provided a post model selection estimator, and derived asymptotic properties for it. An attempt to adapt the methods developped in [5] to the problem treated in this paper may be of interest.

In the assumptions for the experiment that we perform we state that although the values of the wavelengths are assumed to be unknown, we however do know their order, and this is given in (D.1). Thus the indices $1, 2, \ldots, s$ correspond to an ordered set of wavelengths and one goal has been to estimate their values. When estimating $\boldsymbol{q}$ a possibly reasonable loosening of this assumption in a real world physics experiment may be to not know the order of the wavelengths. Then one may assume that there is an order (D.1) for the unknown wavelengths but that one does not know that the indices, or labels, $1, 2, \ldots, s$ is the correct ordering. The problem would then be to estimate $\boldsymbol{q}$, under the assumption of an order on the values of $\boldsymbol{q}$ (which is ordered in the reverse way to the wavelengths) but in which one does not know the correct order. A problem which is reminiscent to this, but then in a simpler setting, was treated in [1], in which one derived a likelihood based estimator for an unknown ordered probability mass function in which one does not know the correct order.

6 Acknowledgments

VP's research is fully supported by the Swedish Research Council (SRC), and the research of DA is partially supported by the SRC. The authors gratefully acknowledge the SRC's support. We would also like to thank Victor Ufnarovski and Andrey Gulchak for their kind help with Lemma 3.

7 Bibliography

[1] Anevski D., Gill R.D., Zohren S., (2017) Annals of Statistics, Volume 45, Number 6, 2708–2735.

[2] Anevski D., Hall-Wilton R. (2012). Statistical methods for energy determination in neutron detector systems. Technical report, Mathematical Sciences Lund University and ESS, 2012.

[3] Anevski D., Hall-Wilton R., Kanaki K., Pastukhov V. (2018). A stochastic process approach to multilayer neutron detectors Tech. rep., arXiv.org.

[4] Anevski D.,Pastukhov V. (2018). The asymptotic distribution of the isotonic regression estimator over a general countable pre ordered set, Tech report. arXiv.org

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

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

[7] Jankowski, H. K. and Wellner, J. A. (2009). Estimation of a discrete monotone distribution. *Electronic journal of statistics* **39**, 125–153.

[8] Lyubich Y. I., (2004). The Sylvester-Ramanujan system of equations and the complex power moment problem. The Ramanujam Journal, **8**, 23–45

[9] Ramanujan S. (1912). Note on a set of simultaneous equations. Journal of the Indian Mathematical Society, **IV**, 94–96.

[10] Robertson, T., Wright, F. T. and Dykstra R. L. (1988). *Order restricted statistical inference*. John Wiley & Sons, Ltd., Chichester.

[11] van der Vaart A.W. (1998). Asymptotic Statistics, Cambridge Univ. Press, New York.

[12] Willis B. T. M., Carlile C. J. (1999). Experimental neutron scattering, Oxford University Press.

Printed by Media-Tryck, Lund 2018 NORDIC SWAN ECOLABEL 3041 0903

Doctoral Theses in Mathematical Sciences 2018:8

ISBN 978-91-7753-808-0
ISRN LUNFMS-1025-2018
ISSN 1404-0034