Electronic Journal of Statistics
Vol. 16 (2022) 4247–4274
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2045

Stacked Grenander and rearrangement estimators of a discrete distribution*

Vladimir Pastukhov

Department of Computer Science and Engineering,
Chalmers University of Technology
e-mail: vlapas@chalmers.se

Abstract: In this paper we consider the stacking of isotonic regression and the method of rearrangement with the empirical estimator to estimate a discrete distribution with an infinite support. The estimators are proved to be strongly consistent with $\sqrt{n}$-rate of convergence. We obtain the asymptotic distributions of the estimators and construct the asymptotically correct conservative global confidence bands. We show that stacked Grenander estimator outperforms the stacked rearrangement estimator. The new estimators behave well even for small sized data sets and provide a trade-off between goodness-of-fit and shape constraints.

MSC2020 subject classifications: 62E20, 62G07, 62G20.
Keywords and phrases: Constrained inference, cross-validation, discrete distribution, Grenander estimator, isotonic regression, model stacking, rearrangement, smoothing.

Received August 2021.

Contents

arXiv: 2106.00560
*This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundataion.

1. Introduction

This work is largely inspired by recent papers in the estimation of discrete distributions with shape constraints. The first paper in this area is [19], where the authors studied the method of rearrangement and maximum likelihood estimator (MLE) of probability mass function (p.m.f.) under monotonicity constraint. The MLE under monotonicity constraint is also known as Grenander estimator. Next, in the paper [13] the authors introduced the least squares estimator of a discrete distribution under the constraint of convexity and, further, its limiting distribution was obtained in [1]. Furthermore, the MLE of log-concave p.m.f. was studied in detail in [4], and in [20] the problem was generalised to the case of multidimensional discrete support. Next, in paper [3] the authors introduced the MLE of unimodal p.m.f. with unknown support, proved the consistency and obtained the asymptotic distribution. The problem of least squares estimation of a completely monotone p.m.f. was considered in papers [2, 5].

In most of the papers listed above the authors considered both the well- and the mis-specified cases and studied the asymptotic properties of the estimators in both cases. In this work we do not have the mis-specified case in a sense that we assume that the true p.m.f. can be non-monotone and our estimators are strongly consistent even if the true p.m.f. is not decreasing.

The estimators introduced and studied in this paper are in some sense similar to nearly-isotonic regression approach, cf. [33] and [23] for multidimensional case. Nearly-isotonic regression is a convex optimisation problem, which provides intermediate less restrictive solution and the isotonic regression is included in the path of the solutions.

At the same time, our approach is in some sense opposite to liso (lasso-isotone), cf. [14], and to bounded isotonic regression, cf. [22]. The liso is a combination of isotonic regression and lasso penalties, and bounded isotonic regression imposes additional penalisation to the range of the fitted model.

In this paper we combine Grenander estimator and the method of rearrangement with cross-validation-based model-mix concept, cf. [31]. The estimator is constructed as a convex combination of the empirical estimator and Grenander estimator or the empirical estimator and rearrangement estimator. Following the terminology for regression and classification problems in [9, 21, 35], we call the resulting estimators as *stacked Grenander estimator* and *stacked rearrangement estimator*, respectively. Therefore, we do not impose the strict monotonic restriction and let the data decide.

There are several papers where the authors studied a convex combination of the empirical estimator with a prescribed probability vector, cf. [15, 16, 31, 34]. In particular, in [31] the authors proposed the combination of the empirical estimator and a constant p.m.f. with a mixture parameter selected by cross-validation. Also, the minimax estimator of a p.m.f. with respect to ℓ_2-loss with a fixed known finite support and sample size n is given by a convex combination of the empirical estimator and the uniform distribution with a mixture parameter equal to $\frac{\sqrt{n}}{n+\sqrt{n}}$, cf. [34]. Furthermore, in [16] the authors provide a geometrical explanation on the gain from stacking the empirical estimator with

a fixed probability vector and show that the improvement of the estimation increases as the size of the support becomes larger.

In the case of continuous support the first paper on the density estimation via stacking is [30], where it is shown that the method of stacking performs better than selecting the best model by cross-validation. Next, in [27] the authors studied the approach of linear and convex aggregation of density estimators and, in particular, proved that the aggregation of two estimators allows to combine the advantages of both. To the authors' knowledge the constrained stacked estimators have not been investigated for the case of continuous density.

To the authors' knowledge, the problem of staking the shape constrained estimators has not been studied much even in a regression setup, except for the paper [36]. In the paper [36] the author used a convex combination of linear regression with isotonic regression to obtain a strictly monotonic solution. Also, it is worth to mention the paper [18], where it was shown that in terms of prediction accuracy the simplified relaxed lasso (which is stacking of least squares estimator and lasso) performs almost equally to the lasso in low signal-to-noise ratio regimes, and nearly as well as the best subset selection in high signal-to-noise ratio scenarios.

The paper is organised as follows. In Section 2 we state the problem and introduce notation. The derivation of cross-validation based mixture parameter is given in Section 3. Section 4 is dedicated to the theoretical properties of the estimators such as consistency, rate of convergence and asymptotic distribution. Also, in Section 4 we construct asymptotic confidence bands. In Section 5 we do simulation study to compare the performance of the estimators with empirical, minimax, rearrangement and Grenander estimators. The article closes with a conclusion and a discussion of possible generalisations in Section 6. The ancillary results and the proofs of some statements are given in Appendix. The R code for the simulations is available upon request.

2. Statement of the problem and notation

First, let us introduce notation and several definitions. Assume that $z_1, z_2, \dots, z_n$ is a sample of n i.i.d. random variables with values in $\mathbb{N}$ and generated by a p.m.f. $\boldsymbol{p}$. For a given data sample let us create the frequency data $\boldsymbol{x} = (x_0, \dots, x_{t_n})$, where $x_j = \sum_{i=1}^{n} 1\{z_i = j\}$ and $t_n = \sup\{j : x_j > 0\}$ denotes the largest order statistic for the sample.

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

$$\hat{p}_{n,j} = \frac{x_j}{n}, \quad j \in \mathbb{N},$$

and it is strongly consistent, unbiased and asymptotically normal in ℓ_2-space.

The rearrangement estimator studied in [19] is defined as

$$\hat{\boldsymbol{r}}_n = rear(\hat{\boldsymbol{p}}_n), \tag{2.1}$$

where $rear(\boldsymbol{w})$ denotes the reversed-ordered vector. Also, equivalently, the rearrangement estimator can be written as $\hat{r}_{n,j} = \sup\{u : Q_n(u) \leq j\}$, where $Q_n(u) = \#\{k : \hat{p}_{n,k} \geq u\}$.

The MLE of decreasing p.m.f., or Grenander estimator, which we denote by $\hat{\boldsymbol{g}}_n$, is equivalent to the isotonic regression of the empirical estimator, cf. [6, 19, 28], i.e.

$$\hat{\boldsymbol{g}}_n = \Pi(\hat{\boldsymbol{p}}_n|\mathcal{F}^{decr}) := \operatorname*{argmin}_{\boldsymbol{f}\in\mathcal{F}^{decr}} \sum_j [\hat{p}_{n,j} - f_j]^2, \tag{2.2}$$

where $\mathcal{F}^{decr}$ is the monotonic cone in ℓ_2, i.e. $\mathcal{F}^{decr} = \left\{\boldsymbol{f} \in \ell_2 : f_0 \geq f_1 \geq \dots\right\}$, $\hat{\boldsymbol{p}}_n$ is the empirical estimator and $\Pi(\hat{\boldsymbol{p}}_n|\mathcal{F}^{decr})$ denotes the ℓ_2-projection of $\hat{\boldsymbol{p}}_n$ onto $\mathcal{F}^{decr}$ [1].

In our work we construct the estimator in the following way:

$$\hat{\boldsymbol{\phi}}_n = \beta\hat{\boldsymbol{h}}_n + (1-\beta)\hat{\boldsymbol{p}}_n, \tag{2.3}$$

where

$$\hat{\boldsymbol{h}}_n = \begin{cases} \hat{\boldsymbol{r}}_n, & \text{for the stacked rearrangement estimator,} \\ \hat{\boldsymbol{g}}_n, & \text{for the stacked Grenander estimator,} \end{cases}$$

with the data-driven selection of β:

$$\hat{\beta}_n = \operatorname*{argmin}_{\beta\in[0,1]} CV(\beta),$$

where $CV(\beta)$ is a cross-validation criterion, which we introduce and study below.

We associate each component x_j of the frequency vector $\boldsymbol{x}$ with multinomial indicator $\boldsymbol{\delta}^{[j]} \in \mathbb{R}^{t_n+1}$, given by

$$\boldsymbol{\delta}^{[j]} = (0, \dots, 0, 1, 0, \dots, 0) \tag{2.4}$$

for $j = 0, \dots, t_n$, cf. [31]. All elements of $\boldsymbol{\delta}^{[j]}$ are zeros, except for the one with index j.

Next, let $\hat{\boldsymbol{p}}_n^{\setminus[j]}$ for $j = 0, \dots, t_n$ denote the leave-one-out version of the empirical estimator $\hat{\boldsymbol{\phi}}_n$ for the frequency data $\boldsymbol{x} = (x_0, \dots, x_{t_n})$, i.e. for j such that $x_j > 0$ let

$$\hat{\boldsymbol{p}}_n^{\setminus[j]} = \frac{\boldsymbol{x} - \boldsymbol{\delta}^{[j]}}{n-1}.$$

Next, for the rearrangement estimator, the leave-one-out version is given by

$$\hat{\boldsymbol{r}}_n^{\setminus[j]} = rear(\hat{\boldsymbol{p}}_n^{\setminus[j]}),$$

[1] The notion of "isotonic regression" in (2.2) might be confusing. Though, for historical reasons, it is a standard notion in the subject of constrained inference, cf. the monographs [28, 29] and also papers [7, 32], dedicated to the computational aspects, where the notation "isotonic regression" is used for the isotonic projection of a general vector.

and for Grenander estimator:

$$\hat{\boldsymbol{g}}_n^{\setminus[j]} = \Pi\big(\hat{\boldsymbol{p}}_n^{\setminus[j]} | \mathcal{F}^{decr}\big).$$

Therefore, for j such that $x_j > 0$ the leave-one-out versions of stacked rearrangement and stacked Grenander estimators for a fixed misture parameter β are given by

$$\hat{\boldsymbol{\phi}}_n^{\setminus[j]} = \beta\, \hat{\boldsymbol{h}}_n^{\setminus[j]} + (1-\beta)\hat{\boldsymbol{p}}_n^{\setminus[j]}, \tag{2.5}$$

with $\hat{\boldsymbol{h}}_n^{\setminus[j]} = \hat{\boldsymbol{r}}_n^{\setminus[j]}$ for the case of stacked rearrangement estimator, and $\hat{\boldsymbol{h}}_n^{\setminus[j]} = \hat{\boldsymbol{g}}_n^{\setminus[j]}$ for the case of stacked Grenander estimator, respectively.

For an arbitrary vector $\boldsymbol{f} \in \ell_k$ we define ℓ_k-norm

$$||\boldsymbol{f}||_k = \begin{cases} \Big(\sum_{j=0}^{\infty} |f_j|^k\Big)^{1/k}, & \text{if } k \in \mathbb{N}\backslash\{0\}, \\ \sup_{j\in\mathbb{N}} |f_j|, & \text{if } k = \infty, \end{cases}$$

and for $\boldsymbol{v} \in \ell_2$ and $\boldsymbol{w} \in \ell_2$ let $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \sum_{j=0}^{\infty} v_j w_j$ denote the inner product on ℓ_2.

For a random sequence $b_n \in \mathbb{R}$ we will use the notation $b_n = O_p(n^q)$ if for any $\varepsilon > 0$ there exists a finite $M > 0$ and a finite $N > 0$ such that

$$\mathbb{P}[n^{-q}|b_n| > M] < \varepsilon,$$

for any $n > N$.

3. Data-driven selection of the mixture parameter β

Let us consider squared ℓ_2-distance between the true p.m.f. $\boldsymbol{p}$ and the stacked estimator $\hat{\boldsymbol{\phi}}_n$:

$$L_n = ||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_2^2 :\equiv L_n^{(1)} - 2L_n^{(2)} + L_n^{(3)}, \tag{3.1}$$

where $L_n^{(1)} = \sum_{j=0}^{t_n} \hat{\phi}_{n,j}^2$, $L_n^{(2)} = \sum_{j=0}^{t_n} \hat{\phi}_{n,j} p_j$ and $L_n^{(3)} = \sum_{j=0}^{t_n} p_j^2$.

We aim to minimise L_n. Obviously, $\boldsymbol{p}$ is unknown, and we will use the approach introduced in [24] to estimate L_n. First, note that $L_n^{(3)}$ is a constant and can be omitted. Next, note that for a given n we have for L_2 we have

$$L_2 = \sum_{j=0}^{t_n} \hat{\phi}_{n,j} p_j = \mathbb{E}[\hat{\boldsymbol{\phi}}_n],$$

and following [24] we estimate $L_n^{(2)}$ by

$$\hat{L}_n^{(2)} = \sum_{j=0}^{t_n} \hat{p}_{n,j} \hat{\phi}_{n,j}^{\setminus[j]},$$

with $\hat{\boldsymbol{\phi}}_n^{\setminus[j]}$ defined in (2.5). Therefore, we select the mixture parameter β to minimise

$$CV(\beta) = L_n^{(1)} - 2\hat{L}_n^{(2)}, \tag{3.2}$$

i.e.

$$\hat{\beta}_n = \operatorname*{argmin}_{\beta\in[0,1]} CV(\beta).$$

This cross-validation approach for estimation of discrete distributions was first introduced in [24] for smoothing kernel estimator and was also used in, for example, [11, 12, 25]. The mixture parameter $\hat{\beta}_n$ is given in the following theorem.

Theorem 1. *The leave-one-out least-squares cross-validation mixture parameter $\hat{\beta}_n$ is given by*

$$\hat{\beta}_n = \begin{cases} \frac{b_n}{a_n}, & \text{if } a_n \neq 0 \text{ and } 0 \leq b_n \leq a_n, \\ 1, & \text{if } 0 < a_n \leq b_n, \\ 0, & \text{otherwise}, \end{cases}$$

where

$$a_n = \sum_{j=0}^{t_n} (\hat{h}_{n,j} - \hat{p}_{n,j})^2,$$

and

$$b_n = \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{h}^{\setminus[j]}_{n,j} - \hat{p}^{\setminus[j]}_{n,j}) - \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{h}_{n,j} - \hat{p}_{n,j}),$$

with $\hat{\boldsymbol{h}}^{\setminus[j]}_n = \hat{\boldsymbol{r}}^{\setminus[j]}_n$ for the case of stacked rearrangement estimator, and $\hat{\boldsymbol{h}}^{\setminus[j]}_n = \hat{\boldsymbol{g}}^{\setminus[j]}_n$ for the case of stacked Grenander estimator, respectively.

In the sequel of the paper we always assume that both $\hat{\boldsymbol{\phi}}_n$ and $\hat{\boldsymbol{\phi}}^{\setminus[j]}_n$ are constructed with the leave-one-out least-squares cross-validation mixture parameter $\hat{\beta}_n$.

4. Theoretical properties of the estimator

In this section we study theoretical properties of stacked rearrangement and stacked Grenander estimators. First, let us assume that $\boldsymbol{p} \in \mathcal{F}^{decr}$, i.e. the underlying p.m.f. is decreasing. Note that from the subadditivity of the norms for $||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_k$, with $1 \leq k \leq \infty$, we have

$$\begin{aligned} ||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_k = ||\hat{\beta}_n \hat{\boldsymbol{h}}_n + (1-\hat{\beta}_n)\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_k \leq \\ \hat{\beta}_n ||\hat{\boldsymbol{h}}_n - \boldsymbol{p}||_k + (1-\hat{\beta}_n)||\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_k. \end{aligned}$$

From the error reduction property of the rearrangement and Grenander estimators, i.e. $||\hat{\boldsymbol{h}}_n - \boldsymbol{p}||_k \leq ||\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_k$, with $1 \leq k \leq \infty$, cf. Theorem 2.1 in [19], we have

$$||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_k \leq ||\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_k \tag{4.1}$$

for all $1 \leq k \leq \infty$. Therefore, in the case of a decreasing true p.m.f. both the stacked rearrangement and stacked Grenander estimators also provide the error reduction.

Assume that the true p.m.f. is not decreasing. Let $\boldsymbol{r} = rear(\boldsymbol{p})$ and $\boldsymbol{g} = \Pi\big(\boldsymbol{p}|\mathcal{F}^{decr}\big)$. Note that $\boldsymbol{r} \neq \boldsymbol{p}$ nor $\boldsymbol{g} \neq \boldsymbol{p}$, if $\boldsymbol{p} \notin \mathcal{F}^{decr}$, i.e. the vector $\boldsymbol{r}$ is reversed ordered vector $\boldsymbol{p}$ and $\boldsymbol{g}$ is decreasing vector in ℓ_2 which is closest in ℓ_2-norm to the true p.m.f. $\boldsymbol{p}$.

Then, since the isotonic regression and the rearrangement, viewed as a mapping from ℓ_2 into ℓ_2, are continuous in the case of a finite support, and the empirical estimator is strongly consistent, then

$$\hat{\boldsymbol{r}}_n \overset{a.s.}{\to} \boldsymbol{r}, \text{ and } \hat{\boldsymbol{g}}_n \overset{a.s.}{\to} \boldsymbol{g},$$

pointwise. Note that from the statements (i), (ii) and (iv) of Lemma 3 in Appendix it follows that $\hat{\boldsymbol{g}}_n$ always exists, and it is a probability vector for all n. Clearly, the same result holds for the rearrangement estimator $\hat{\boldsymbol{r}}_n$ for all n. The almost sure convergence in ℓ_k-norm, for $1 \leq k \leq \infty$, of $\hat{\boldsymbol{r}}_n$ and $\hat{\boldsymbol{g}}_n$ to $\boldsymbol{r}$ and $\boldsymbol{g}$, respectively, now follows from Lemma C.2 in the supporting material of [4].

4.1. Consistency

First, let us study the leave-one-out versions of the empirical, rearrangement and Grenander estimators. Recall that

$$\hat{\boldsymbol{p}}_n^{\setminus[j]} = \frac{\boldsymbol{x} - \boldsymbol{\delta}^{[j]}}{n-1}, \ \hat{\boldsymbol{r}}_n^{\setminus[j]} = rear(\hat{\boldsymbol{p}}_n^{\setminus[j]}) \text{ and } \hat{\boldsymbol{g}}_n^{\setminus[j]} = \Pi\big(\hat{\boldsymbol{p}}_n^{\setminus[j]}|\mathcal{F}^{decr}\big),$$

for j such that $x_j > 0$.

Let us define vectors $\hat{\boldsymbol{\pi}}_n \in \ell_1$, $\hat{\boldsymbol{\rho}}_n \in \ell_1$, and $\hat{\boldsymbol{\gamma}}_n \in \ell_1$ as

$$\begin{aligned}
\hat{\pi}_{n,j} &= \begin{cases} \hat{p}_{n,j}^{\setminus[j]}, & \text{if } x_j > 0, \\ 0, & \text{otherwise}, \end{cases} \\
\hat{\rho}_{n,j} &= \begin{cases} \hat{r}_{n,j}^{\setminus[j]}, & \text{if } x_j > 0, \\ 0, & \text{otherwise}, \end{cases} \\
\hat{\gamma}_{n,j} &= \begin{cases} \hat{g}_{n,j}^{\setminus[j]}, & \text{if } x_j > 0, \\ 0, & \text{otherwise}. \end{cases}
\end{aligned} \tag{4.2}$$

Lemma 1. *The sequences of vectors $\hat{\boldsymbol{\pi}}_n$, $\hat{\boldsymbol{\rho}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ converge pointwise a.s. to $\boldsymbol{p}$, $\boldsymbol{r}$, and $\boldsymbol{g}$, respectively.*

Proof. The proof is given in Appendix. □

Next, we prove the following important lemma.

Lemma 2. *For the vectors $\hat{\boldsymbol{\pi}}_n$ we have*

$$\hat{\pi}_{n,j} \leq \hat{p}_{n,j}$$

for all j, and for $\hat{\boldsymbol{\rho}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ we have

$$\hat{\rho}_{n,j} \leq \frac{n}{n-1}\hat{r}_{n,j} \ \text{ and } \ \hat{\gamma}_{n,j} \leq \frac{n}{n-1}\hat{g}_{n,j}$$

for all j.

Proof. The proof is given in Appendix. □

In Lemma C.2 in the supporting material of [4] it was proved that for probability mass functions the pointwise convergence and the convergence in ℓ_k for $1 \leq k \leq \infty$ are all equivalent. Note, in our case the sequences $\hat{\boldsymbol{\pi}}_n$, $\hat{\boldsymbol{\rho}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ are not probability vectors. Nevertheless, as we prove below, all $\hat{\boldsymbol{\pi}}_n$, $\hat{\boldsymbol{\rho}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ converge a.s. to $\boldsymbol{p}$, $\boldsymbol{r}$ and $\boldsymbol{g}$, respectively, in ℓ_k-norm for $1 \leq k \leq \infty$.

Theorem 2. *For the vectors $\hat{\boldsymbol{\pi}}_n$, $\hat{\boldsymbol{\rho}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ we have*

$$\hat{\boldsymbol{\pi}}_n \overset{a.s.}{\to} \boldsymbol{p},$$

$$\hat{\boldsymbol{\rho}}_n \overset{a.s.}{\to} \boldsymbol{r},$$

and

$$\hat{\boldsymbol{\gamma}}_n \overset{a.s.}{\to} \boldsymbol{g}$$

in ℓ_k-norm for $1 \leq k \leq \infty$.

Proof. The proof starts in a similar way as the one for Lemma C.2 in [4]. Let us, first, study the case of $\hat{\boldsymbol{\pi}}_n$. Fix some $\varepsilon > 0$. Then, we can choose K such that

$$\sum_{j \leq K} p_j \geq 1 - \frac{\varepsilon}{4}.$$

Since both $\boldsymbol{\pi}_n$ and the empirical estimator $\boldsymbol{p}_n$ converge to $\mathbf{p}$ pointwise, then there exists random n_0 such that for all $n \geq n_0$

$$\sup_{j \leq K} |\hat{p}_{n,j} - p_j| \leq \frac{\varepsilon}{4(K+1)},$$

$$\sup_{j \leq K} |\hat{\pi}_{n,j} - p_j| \leq \frac{\varepsilon}{4(K+1)},$$

almost surely.

This implies that for all $n \geq n_0$ we have $\sum_{j \leq K} \hat{p}_{n,j} \geq 1 - \frac{\varepsilon}{2}$ and $\sum_{j \leq K} |\hat{\pi}_{n,j} - p_j| \leq \frac{\varepsilon}{4}$, almost surely.

Next, for any n

$$\sum_{j=0}^{\infty} |\hat{\pi}_{n,j} - p_j| = \sum_{j \leq K} |\hat{\pi}_{n,j} - p_j| + \sum_{j > K} |\hat{\pi}_{n,j} - p_j| \leq \sum_{j \leq K} |\hat{\pi}_{n,j} - p_j| + \sum_{j > K} \hat{\pi}_{n,j} + \sum_{j > K} p_j.$$

Furthermore, $\sum_{j>K} \hat{\pi}_{n,j} \leq \sum_{j>K} \hat{p}_{n,j}$ since $0 < \hat{\pi}_{n,j} \leq \hat{p}_{n,j}$. Then, for all $n > n_0$ we have proved that

$$\sum_{j=0}^{\infty} |\hat{\pi}_{n,j} - p_j| \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{2} + \frac{\varepsilon}{4} = \varepsilon,$$

almost surely. This means that for any $\varepsilon > 0$ there exists random n_0, such that for all $n > n_0$

$$||\hat{\boldsymbol{\pi}}_n - \mathbf{p}||_1 \leq \varepsilon,$$

almost surely.

Furthermore, since $\ell_1 \subset \ell_k$, for all $k > 1$, then a.s. convergence holds in ℓ_k, for all $1 \leq k \leq \infty$.

Let us prove the convergence for $\hat{\boldsymbol{\gamma}}_n$. First, from Lemma 2 it follows that

$$\frac{n-1}{n}\hat{\gamma}_{n,j} \leq \hat{g}_{n,j}.$$

Then, since both $\frac{n-1}{n}\hat{\boldsymbol{\gamma}}_n$ and $\hat{\boldsymbol{g}}_n$ converge to $\boldsymbol{g}$ a.s., we can use the same approach as for $\hat{\boldsymbol{\pi}}$ above, and prove that

$$\frac{n-1}{n}\hat{\boldsymbol{\gamma}}_n \overset{a.s.}{\to} \boldsymbol{g},$$

in ℓ_k, for $1 \leq k \leq \infty$, which means that

$$\hat{\boldsymbol{\gamma}}_n \overset{a.s.}{\to} \boldsymbol{g},$$

in ℓ_k, for $1 \leq k \leq \infty$.

Now, using the result of Lemma 2, we can prove the result for $\hat{\boldsymbol{\rho}}_n$ in the same way as we did for $\hat{\boldsymbol{\gamma}}_n$. □

Now we can summarize the above results in the following theorem.

Theorem 3. *For any underlying distribution $\boldsymbol{p}$, both the stacked rearrangement and stacked Grenander estimators are strongly consistent:*

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

in ℓ_k-norm for $1 \leq k \leq \infty$.

Proof. Firs, let us assume that $\boldsymbol{p}$ is decreasing. Then the result of the theorem follows from the strong consistency of $\hat{\boldsymbol{g}}_n$, $\hat{\boldsymbol{r}}_n$ and $\hat{\boldsymbol{p}}_n$.

Next, assume that $\boldsymbol{p}$ is not decreasing. From Theorem 2 it follows that for the case of stacked rearrangement estimator we have

$$a_n \overset{a.s.}{\to} ||\boldsymbol{r} - \boldsymbol{p}||_2^2,$$

and

$$b_n \overset{a.s.}{\to} \langle \boldsymbol{p}, (\boldsymbol{r} - \boldsymbol{p})\rangle - \langle \boldsymbol{p}, (\boldsymbol{r} - \boldsymbol{p})\rangle = 0,$$

and for the case of stacked Grenander estimator we have

$$a_n \overset{a.s.}{\to} ||\boldsymbol{g} - \boldsymbol{p}||_2^2,$$

and

$$b_n \overset{a.s.}{\to} \langle \boldsymbol{p}, (\boldsymbol{g} - \boldsymbol{p})\rangle - \langle \boldsymbol{p}, (\boldsymbol{g} - \boldsymbol{p})\rangle = 0.$$

Therefore,

$$\hat{\beta}_n \overset{a.s.}{\to} 0.$$

Next, since

$$||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_k \leq \hat{\beta}_n||\hat{\boldsymbol{h}}_n - \boldsymbol{p}||_k + (1 - \hat{\beta}_n)||\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_k$$

for all $1 \leq k \leq \infty$, it follows

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

in ℓ_k-norm for $1 \leq k \leq \infty$. □

4.2. Rate of convergence

In this section we study the rate of convergence of stacked estimator. In the case of bounded support the $\sqrt{n}$-rate of convergence follows from pointwise convergence of the vectors $\hat{\boldsymbol{\pi}}_n$, $\hat{\boldsymbol{\rho}}_n$ and $\hat{\boldsymbol{\gamma}}_n$. In this work we assume that the support can be infinite.

Theorem 4. *Stacked rearrangement and Grenander estimators have $\sqrt{n}$-rate of convergence for any underlying p.m.f. $\boldsymbol{p}$:*

$$\sqrt{n}||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_k = O_p(1)$$

for $1 < k \leq \infty$. Next, if $\sum_{j=0}^{\infty} \sqrt{p_j} < \infty$, then

$$\sqrt{n}||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_1 = O_p(1).$$

Proof. Assume that $\boldsymbol{p}$ is decreasing. Then the result follows from (4.1) and Corollaries 4.1 and 4.2 in [19].

Next, assume that $\boldsymbol{p}$ is not decreasing. Let us, first, prove the case of stacked Grenander estimator. Recall that

$$\beta_n = \begin{cases} \frac{b_n}{a_n}, & \text{if } a_n \neq 0 \text{ and } 0 \leq b_n \leq a_n, \\ 1, & \text{if } 0 < a_n \leq b_n, \\ 0, & \text{otherwise,} \end{cases}$$

where

$$a_n = \sum_{j=0}^{t_n} (\hat{g}_{n,j} - \hat{p}_{n,j})^2,$$

and in the notation introduced in 4.2, we can write b_n as

$$b_n = \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{\gamma}_{n,j} - \hat{\pi}_{n,j}) - \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{g}_{n,j} - \hat{p}_{n,j}).$$

First, as we proved in Theorem 3

$$a_n \overset{a.s.}{\to} ||\boldsymbol{g} - \boldsymbol{p}||_2^2 > 0. \tag{4.3}$$

Second, note that from Lemma 2 it follows that for all n we have

$$b_n = \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{\gamma}_{n,j} - \hat{g}_{n,j}) + \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{p}_{n,j} - \hat{\pi}_{n,j}) \leq$$
$$\frac{n}{n-1} \sum_{j=0}^{t_n} \hat{p}_{n,j}\hat{g}_{n,j} - \sum_{j=0}^{t_n} \hat{p}_{n,j}\hat{g}_{n,j} + \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{p}_{n,j} - \hat{\pi}_{n,j}).$$

Next,

$$\frac{n}{n-1} \sum_{j=0}^{t_n} \hat{p}_{n,j}\hat{g}_{n,j} - \sum_{j=0}^{t_n} \hat{p}_{n,j}\hat{g}_{n,j} = \frac{\sum_{j=0}^{t_n} \hat{p}_{n,j}\hat{g}_{n,j}}{n-1}.$$

Recall that

$$\hat{\pi}_{n,j} = \begin{cases} \frac{x_j - 1}{n-1} = \frac{n}{n-1}\hat{p}_{n,j} - \frac{1}{n-1}, & \text{if } x_j \neq 0, \\ 0, & \text{otherwise,} \end{cases}$$

which leads to

$$\sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{p}_{n,j} - \hat{\pi}_{n,j}) = \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{p}_{n,j} - \hat{\pi}_{n,j}) = \frac{1 - \sum_{j=0}^{t_n} \hat{p}_{n,j}^2}{n-1}.$$

Therefore, the upper bound for b_n is given by

$$b_n \leq \frac{\sum_{j=0}^{t_n} \hat{p}_{n,j}\hat{g}_{n,j}}{n-1} + \frac{1 - \sum_{j=0}^{t_n} \hat{p}_{n,j}^2}{n-1},$$

and, consequently,

$$\sqrt{n} b_n \xrightarrow{a.s.} 0, \tag{4.4}$$

since both sequences $\sum_{j=0}^{t_n} \hat{p}_{n,j}\hat{g}_{n,j}$ and $\hat{p}_{n,j}^2$ are bounded.

Next, since $\beta_n \geq 0$, from (4.3) and (4.4) it follows that

$$\sqrt{n}\hat{\beta}_n \xrightarrow{a.s.} 0. \tag{4.5}$$

Then, from (4.5) for any $\boldsymbol{p}$ and all $1 \leq k \leq \infty$ the following holds

$$\hat{\beta}_n \sqrt{n} ||\hat{\boldsymbol{g}}_n - \boldsymbol{p}||_k \xrightarrow{a.s.} 0,$$

for all $1 \leq k \leq \infty$. Further, as it follows from Corollary 4.2 in [19], if $\sum_{j=0}^{\infty} \sqrt{p_j} < \infty$, then

$$\sqrt{n}||\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_1 = O_p(1).$$

Therefore, for all $2 \leq k \leq \infty$ and all $\boldsymbol{p}$ we have

$$(1 - \hat{\beta}_n)\sqrt{n}||\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_k = O_p(1),$$

and, if $\sum_{j=0}^{\infty} \sqrt{p_j} < \infty$, then we have

$$(1 - \hat{\beta}_n)\sqrt{n}||\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_1 = O_p(1).$$

Finally, recall that

$$\sqrt{n}||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_k \leq \hat{\beta}_n\sqrt{n}||\hat{\boldsymbol{g}}_n - \boldsymbol{p}||_k + (1 - \hat{\beta}_n)\sqrt{n}||\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_k,$$

which finishes the prove of theorem for the case of Grenander estimator.

Similarly, using the results of Lemma 2, for the case of stacked rearrangement estimator we can show that

$$b_n \leq \frac{\sum_{j=0}^{t_n} \hat{p}_{n,j}\hat{r}_{n,j}}{n-1} + \frac{1 - \sum_{j=0}^{t_n} \hat{p}_{n,j}^2}{n-1},$$

for all n. Then, the rest of the proof is the same as for Grenander estimator with $\hat{\boldsymbol{g}}_n$ and $\boldsymbol{g}$ suitably changed to $\hat{\boldsymbol{r}}_n$ and $\boldsymbol{r}$, respectively. □

4.3. Asymptotic distribution and global confidence band

In this section we study the asymptotic distribution of stacked rearrangement and Grenander estimators and discuss calculation of global confidence band for $\boldsymbol{p}$. The limit distribution of rearrangement and Grenanader estimators were obtained in [19]. The asymptotic distribution of stacked Grenander estimator for the case when true p.m.f. $\boldsymbol{p}$ is either not decreasing with a countable support or strictly decreasing with a finite support is given in the next theorem.

Theorem 5. *Assume that $\boldsymbol{p}$ is either not decreasing with a countable support or strictly decreasing with a finite support. Then stacked rearrangement and Grenander estimators are asymptotically normal*

$$\sqrt{n}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}) \xrightarrow{d} \boldsymbol{Y}_{\boldsymbol{0},C},$$

in ℓ_2, where $\boldsymbol{Y}_{\boldsymbol{0},C}$ is a Gaussian process in ℓ_2 with mean zero and the covariance operator C such that $\langle C\boldsymbol{e}_i, \boldsymbol{e}_{i'}\rangle = p_i\delta_{i,i'} - p_ip_{i'}$, with $\boldsymbol{e}_i \in \ell_2$ the orthonormal basis in ℓ_2 such that in a vector $\boldsymbol{e}_i$ all elements are equal to zero but the one with the index i is equal to 1, and $\delta_{i,j} = 1$, if $i = j$ and 0 otherwise, cf. [19].

Proof. The proof is given in Appendix. □

For the case of a general decreasing underlying p.m.f. with some constant regions the limit distribution of the stacked estimator remains an open problem. Figure 1 illustrates the difference of the asymptotic distributions of the empirical estimator, monotonically constrained estimators and the stacked estimators. Let $U(s)$ denote the uniform distribution over $\{0,\dots,s\}$ and $T^d(s)$ be strictly decreasing triangular function with the support $\{0,\dots,s\}$ (for the definition of triangular function see e.g. [13]). Figure 1 shows standard normal QQ-plots of 1000 samples of $\sqrt{n}(\hat{p}_{n,1} - p_1)$, $\sqrt{n}(\hat{g}_{n,1} - p_1)$, $\sqrt{n}(\hat{r}_{n,1} - p_1)$ and $\sqrt{n}(\hat{\phi}_{n,1} - p_1)$ for both $\hat{\boldsymbol{h}}_n = \hat{\boldsymbol{g}}_n$ and $\hat{\boldsymbol{h}}_n = \hat{\boldsymbol{r}}_n$, with $n = 1000$ for the following distributions:

(a) (left) $\boldsymbol{p} = U(11)$,
(b) (middle) $\boldsymbol{p} = 0.15U(3) + 0.1U(7) + 0.75U(11)$,
(c) (right) $\boldsymbol{p} = T^d(11)$.

From Figure 1 we can conclude that, first, in the case of a decreasing p.m.f. the distributions of stacked estimators asymptotically are not equivalent to the distribution of the empirical estimator, and, second, stacked estimators and constrained estimators have different asymptotic distribution if the underlying p.m.f. has constant regions.

For the process $\boldsymbol{Y}_{\boldsymbol{0},C}$ defined in Theorem 5 let q_α denote the α-quantile of its ℓ_∞-norm, i.e.

$$\mathbb{P}[||\boldsymbol{Y}_{\boldsymbol{0},C}||_\infty > q_\alpha] = \alpha.$$

Then, if $\boldsymbol{p}$ is not decreasing or strictly decreasing, from Theorem 5 for stacked estimator we have

$$\lim_n \mathbb{P}[\sqrt{n}||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_\infty \le q_\alpha] = 1 - \alpha.$$

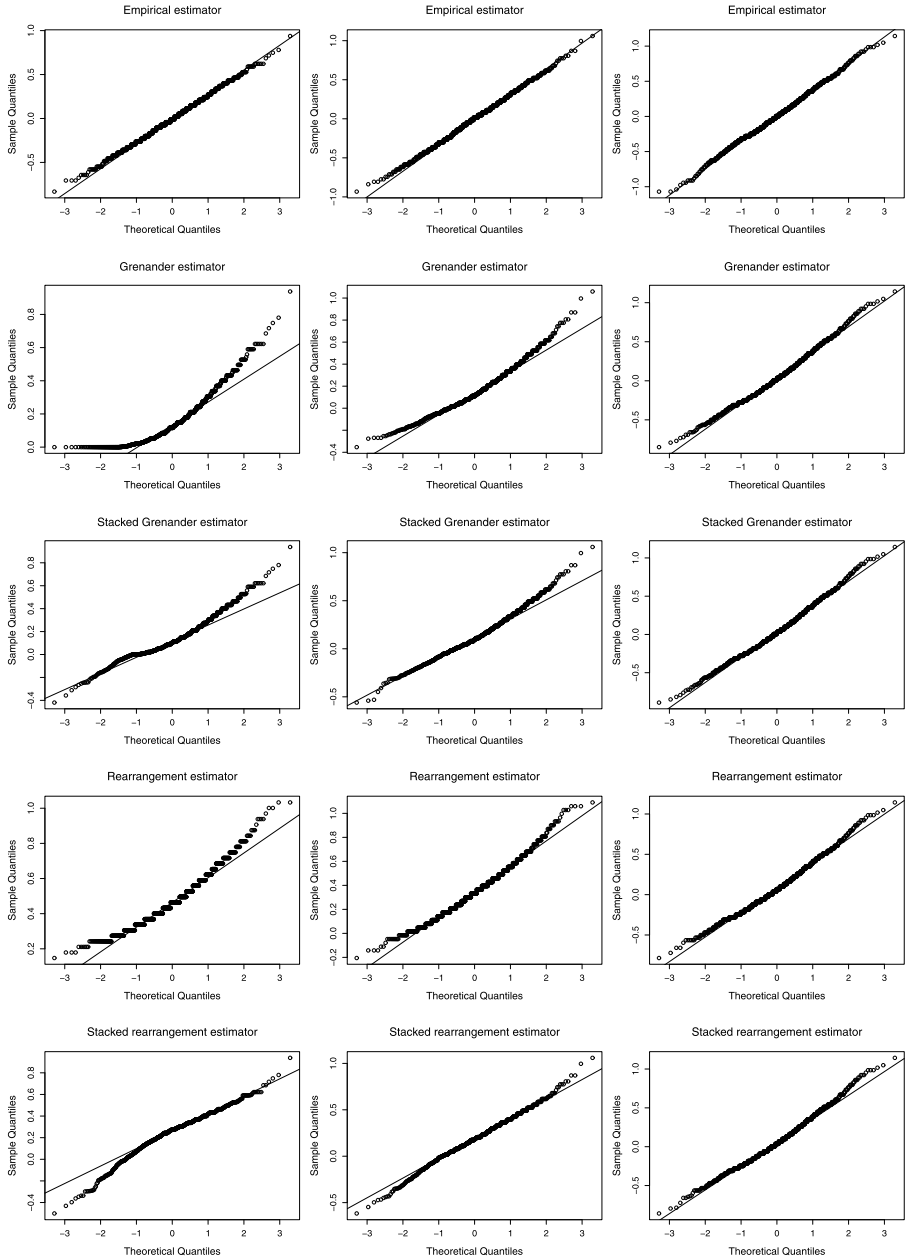

FIG 1. *Standard normal QQ-plots of 1000 samples of* $\sqrt{n}(\hat{p}_{n,1}-p_1)$, $\sqrt{n}(\hat{g}_{n,1}-p_1)$, $\sqrt{n}(\hat{r}_{n,1}-p_1)$ *and* $\sqrt{n}(\hat{\phi}_{n,1}-p_1)$ *for both* $\hat{\boldsymbol{h}}_n=\hat{\boldsymbol{g}}_n$ *and* $\hat{\boldsymbol{h}}_n=\hat{\boldsymbol{r}}_n$, *with* $n=1000$ *for uniform distribution (left), decreasing distribution (middle) and strictly decreasing distribution (right).*

Next, note that in the case of a decreasing p.m.f. $\boldsymbol{p}$ from (4.1) it follows

$$\mathbb{P}[\sqrt{n}||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_\infty \leq q_\alpha] \geq \mathbb{P}[\sqrt{n}||\hat{\boldsymbol{p}}_n - \boldsymbol{p}||_\infty \leq q_\alpha]$$

for all n. Therefore, in the case of a decreasing $\boldsymbol{p}$ we have

$$\liminf_n \mathbb{P}[\sqrt{n}||\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}||_\infty \leq q_\alpha] \geq 1 - \alpha.$$

In the same way as in [3], to estimate q_α we can use the stacked estimator $\hat{\boldsymbol{\phi}}_n$ in place of $\boldsymbol{p}$ in $\boldsymbol{Y}_{\mathbf{0},C}$, and then each quantile can be estimated using Monte-Carlo method. In Proposition B.7 in the supplementary material of [3] it was proved that $\hat{q}_\alpha \overset{a.s.}{\to} q_\alpha$. Therefore, the following confidence band

$$\Big[\max\Big((\hat{\phi}_{n,j} - \frac{\hat{q}_\alpha}{\sqrt{n}}), 0\Big), \hat{\phi}_{n,j} + \frac{\hat{q}_\alpha}{\sqrt{n}}\Big], \text{ for } j \in \mathbb{N}$$

is asymptotically correct global confidence band if $\boldsymbol{p}$ is either not decreasing or strictly decreasing, and it is asymptotically correct conservative global confidence band if $\boldsymbol{p}$ is decreasing with some constant regions.

5. Simulation study of performance of the stacked estimators

In this section we do simulation study to compare the performance of stacked estimators with the empirical, Grenander, rearrangement and the minimax estimators. For the p.m.f. with finite support $\{0, \dots, s\}$ and for a given sample size n the minimax estimator of $\boldsymbol{p}$ with respect to ℓ_2-loss is given by

$$\hat{\boldsymbol{p}}_n^{mm} = \alpha_n^{mm}\boldsymbol{\lambda} + (1 - \alpha_n^{mm})\hat{\boldsymbol{p}}_n, \tag{5.1}$$

with $\boldsymbol{\lambda} = (\frac{1}{s+1}, \dots, \frac{1}{s+1})$ and $\alpha_n^{mm} = \frac{\sqrt{n}}{n+\sqrt{n}}$, cf. [34]. To the authors' knowledge, the minimax estimation with respect to ℓ_2-loss for infinitely supported p.m.f. is an open problem. With some abuse of notation, in this and next sections for infinitely supported distributions we refer the estimator defined in (5.1) with $s = t_n$ as "minimax".

5.1. Performance of the estimators

We study the cases of decreasing and not decreasing true p.m.f. $\boldsymbol{p}$ separately.

5.1.1. True p.m.f. is decreasing

Let us consider the following uniform and decreasing p.m.f.:

$$\begin{aligned} \boldsymbol{M1} : \boldsymbol{p} &= U(11), \\ \boldsymbol{M2} : \boldsymbol{p} &= 0.15U(3) + 0.1U(7) + 0.75U(11), \end{aligned}$$

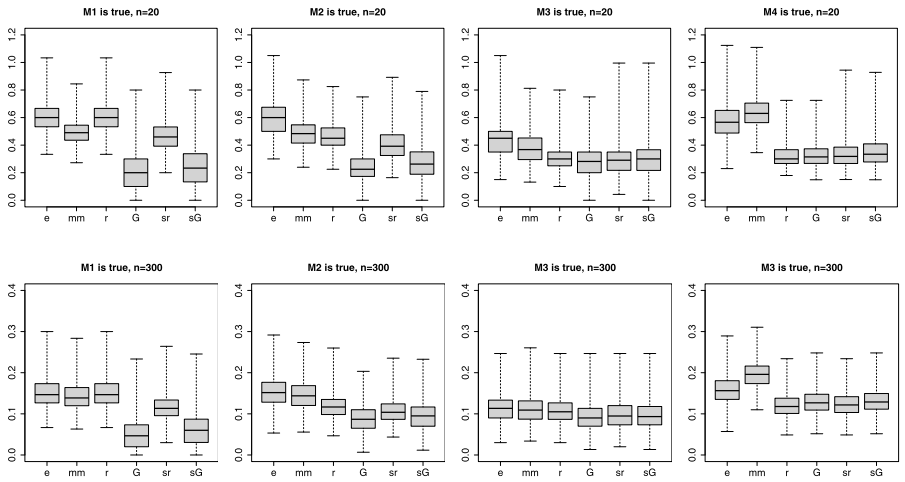

FIG 2. *The boxplots for ℓ_1-distances of the estimators: the empirical estimator (e), minimax estimator (mm), rearrangement estimator (r), Grenander estimator (G), the stacked rearrangement estimator (sr) and the stacked Grenander estimator (sG) for the models* ***M1****,* ***M2****,* ***M3*** *and* ***M4****.*

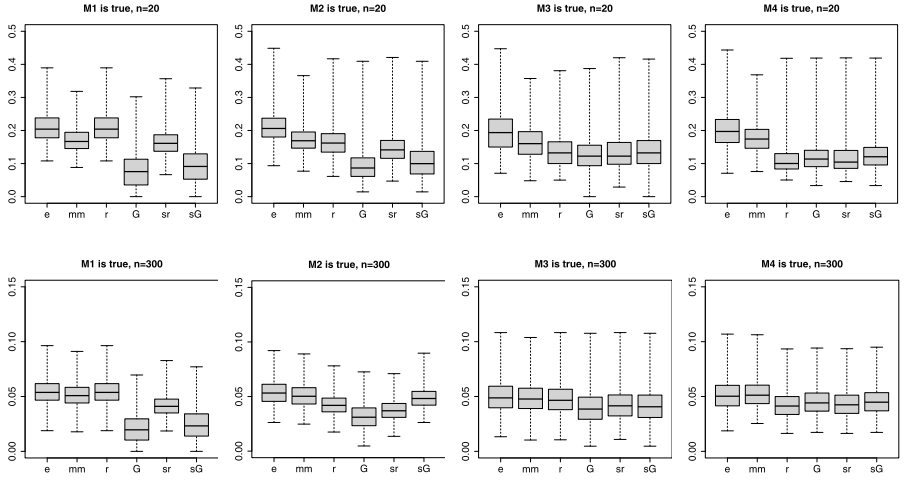

FIG 3. *The boxplots for ℓ_2-distances of the estimators: the empirical estimator (e), minimax estimator (mm), rearrangement estimator (r), Grenander estimator (G), the stacked rearrangement estimator (sr) and the stacked Grenander estimator (sG) for the models* ***M1****,* ***M2****,* ***M3*** *and* ***M4****.*

$$\begin{aligned} \boldsymbol{M3}: \boldsymbol{p} &= 0.25U(1) + 0.2U(3) + 0.15U(5) + 0.4U(7), \\ \boldsymbol{M4}: \boldsymbol{p} &= Geom(0.25), \end{aligned}$$

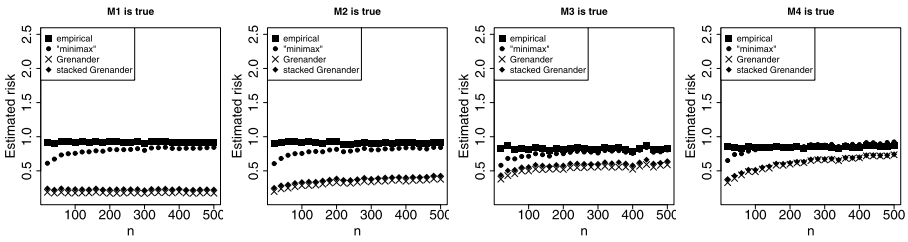

FIG 4. *The estimates of the scaled risk for the models* ***M1****,* ***M2****,* ***M3*** *and* ***M4****.*

where $Geom(\theta)$ is Geometric distribution, i.e. $p_j = (1-\theta)\theta^j$ for $j \in \mathbb{N}$ with $0 < \theta < 1$.

The models $\boldsymbol{M2}$, $\boldsymbol{M3}$ and $\boldsymbol{M4}$ were used in [19] to assess the performance of Grenander estimator and compare its performance with empirical and rearrangement estimators. First, we compare the performance of the estimators in ℓ_1 (Figure 2) and ℓ_2 (Figure 3) distances for small $n = 20$ and moderate $n = 300$ sample sizes with 1000 Monte Carlo simulations.

From the boxplots at Figure 2 and Figure 3 we can conclude that for both small and moderate sized data sets stacked Grenander estimator outperforms in ℓ_1 and ℓ_2 norms both the empirical estimator and minimax estimator ("minimax" for the case of Geometric distribution). Further, stacked Grenander estimator outperforms stacked rearrangement estimator when the underlying distribution has constant regions and it performs almost the same in the case of strictly decreasing p.m.f. The superiority of Grenander estimator over the rearrangement estimator was proved in [19].

Next, in order to summarise the results and demonstrate the superiority of stacked Grenander estimator we plot the estimates of scaled risk $n\mathbb{E}[||\hat{\boldsymbol{\xi}}_n - \boldsymbol{p}||_2^2]$ (with $\hat{\boldsymbol{\xi}}_n$ one of the following estimators: empirical, minimax Grenander or stacked Grenander estimator) versus the sample size n, based on 1000 Monte Carlo simulations, cf. Figure 4. We can conclude that in the case of a decreasing underlying distribution stacked Grenander estimator performs almost as good as Grenander estimator and it performs significantly better than the empirical and the minimax estimators.

5.1.2. True p.m.f. is not decreasing

Now let us consider the case when the underlying distributions are not decreasing:

$$\begin{aligned} \boldsymbol{M5} : \boldsymbol{p} &= T^i(11), \\ \boldsymbol{M6} : \boldsymbol{p} &= NBin(7, 0.4), \\ \boldsymbol{M7} : \boldsymbol{p} &= \frac{3}{8}Pois(2) + \frac{5}{8}Pois(15), \end{aligned}$$

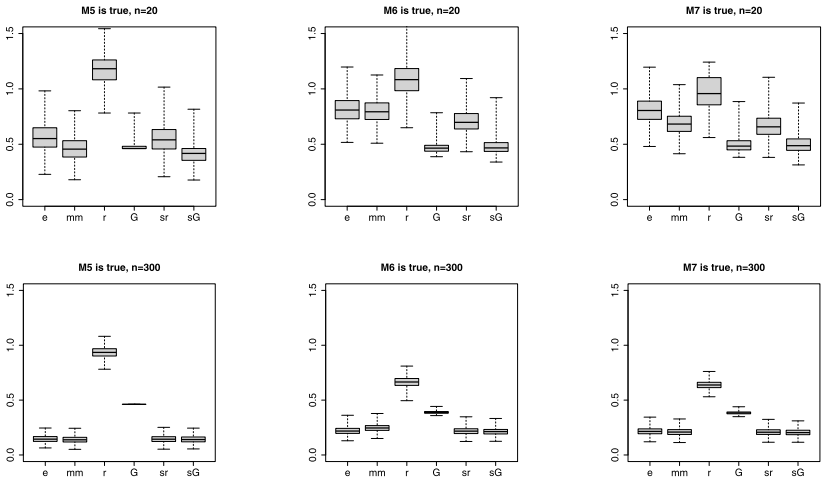

FIG 5. *The boxplots for ℓ_1-distances of the estimators: the empirical estimator (e), minimax estimator (mm), rearrangement estimator (r), Grenander estimator (G), the stacked rearrangement estimator (sr) and the stacked Grenander estimator (sG) for the models* ***M5****,* ***M6*** *and* ***M7****.*

where $T^i(s)$ stands for strictly increasing triangular function; $NBin(r,\theta)$ is the negative binomial distribution with r the number of failures until the experiment is stopped and θ the success probability; $Pois(\lambda)$ is Poisson distribution with rate λ. Therefore, we consider very non-monotonic distributions. Indeed, model $\boldsymbol{M5}$ is a strictly increasing p.m.f., $\boldsymbol{M6}$ is a unimodal distribution, and $\boldsymbol{M7}$ is bimodal.

From Figure 5 and Figure 6 we can conclude that stacked Grenander estimator outperforms in ℓ_1 and ℓ_2 norms the empirical, rearrangement and minimax estimators ("minimax" for the cases of Negative Binomial and Poisson mixture).

Next, it is interesting to note that even if the underlying distribution is not monotone, Grenander estimator can still outperform the empirical estimator in both ℓ_1 and ℓ_2 norms for small sample size. This happens because the isotonisation decreases the variance of the estimator though bias becomes larger.

Let us summarise the results at Figure 7 by plotting the estimates of the scaled risk $n\mathbb{E}[||\hat{\boldsymbol{\xi}}_n - \boldsymbol{p}||_2^2]$ (with $\hat{\boldsymbol{\xi}}_n$ one of the following estimators: empirical, minimax or stacked Grenander estimator). Note that in the case of non-decreasing true p.m.f. we do not plot the risk for Grenander estimator, because, obviously, in the miss-specified case the scaled risks of the constrained estimators are worse than the risk of consistent estimators. Based on the simulations we can conclude that stacked Grenander estimator performs better than empirical and minimax estimators even when the underlying distribution is not decreasing.

The result might look surprising at the first sight. Nevertheless, the explanation of the effect of ℓ_2-risk reduction by stacking empirical estimator with some

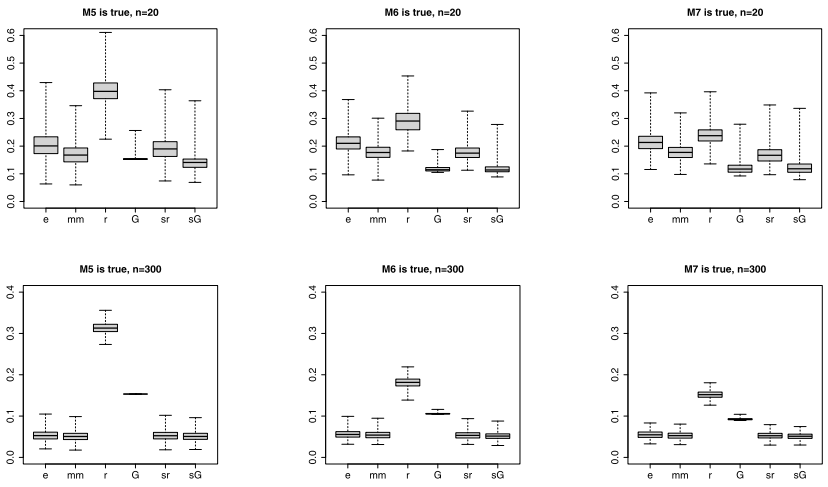

FIG 6. *The boxplots for ℓ_2-distances of the estimators:the empirical estimator (e), minimax estimator (mm), rearrangement estimator (r), Grenander estimator (G), the stacked rearrangement estimator (sr) and the stacked Grenander estimator (sG) for the models **M5**, **M6** and **M7**.*

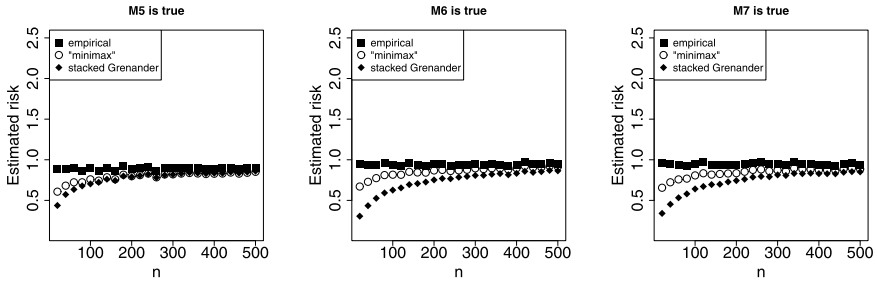

FIG 7. *The estimates of the scaled risk for the models **M5**, **M6** and **M7**.*

fixed probability vector was explained in [16]. Further, let us consider the case of model **M5**, i.e. very non-decreasing case when the underlying distribution is strictly increasing. Then, since the empirical estimator is strongly consistent there exist a random n_1 such that for all $n > n_1$ the vector is $\hat{\boldsymbol{p}}_n$ is strictly increasing almost surely. Next, note that from Lemma 4 it follows that for all $n > n_1$ we have $\hat{g}_j = 1/12$, for all $j = 0, \dots 11$, almost surely. Therefore, for $n > n_1$ stacked Grenander estimator becomes the stacking of the empirical estimator with a uniform distribution $U(11)$ almost surely, which is similar to what, for example, minimax estimator in (5.1) does. One can also see from Figure 7 that in the case of model **M5** stacked Grenander estimator performs very similarly to the minimax estimator in a sense of ℓ_2-risk.

5.2. Coverage probabilities for the confidence bands

The Table 1 presents the proportion of times that

$$\max\left((\hat{\phi}_{n,j} - \frac{\hat{q}_\alpha}{\sqrt{n}}), 0\right) \leq p_j \leq \hat{\phi}_{n,j} + \frac{\hat{q}_\alpha}{\sqrt{n}}, \text{ for all } j \in \mathbb{N}$$

among 1000 runs for the models **M1**–**M7**. The quantiles $\hat{q}_\alpha$ are estimated based on 100000 Monte-Carlo simulations.

First, one can see that the proposed global confidence band performs well. Second, note that for the decreasing p.m.f (models **M1**-**M4**) the coverage probabilities mostly larger than 0.95, while for non-decreasing p.m.f (models **M5**-**M7**) the coverage probabilities are closer to 0.95 when n becomes large, because in the former case the confidence band is asymptotically conservative, while in the later case it is asymptotically correct.

TABLE 1

Empirical coverage probabilities for the confidence bands for $\alpha = 0.05$ of the empirical estimator (e), stacked rearrangement estimator (sr) and stacked Grenander estimator (sG).

Estimator	n	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**
e	100	0.961	0.961	0.957	0.956	0.963	0.973	0.971
	1000	0.945	0.945	0.952	0.949	0.953	0.964	0.956
	5000	0.955	0.943	0.95	0.955	0.945	0.953	0.951
sr	100	0.994	0.994	0.981	0.982	0.969	0.996	0.996
	1000	0.994	0.985	0.972	0.952	0.95	0.973	0.959
	5000	0.996	0.981	0.97	0.959	0.945	0.954	0.949
sG	100	0.996	0.994	0.979	0.981	0.989	0.999	0.997
	1000	0.998	0.984	0.971	0.951	0.953	0.976	0.963
	5000	0.997	0.984	0.97	0.959	0.945	0.954	0.953

5.3. Computational times

First, note, that in general the complexity of the solution for the mixture parameter $\hat{\beta}_n$ depends on the largest order statistic t_n. In Table 2 we provide the "worst case" computational times, i.e. we compute $\hat{\beta}_n$ for the estimator based on the following strictly increasing frequency data vector $\boldsymbol{x}' = (x'_0, \ldots, x'_s)$, with $x'_j = j + 1$ for the different values of s, averaged over 10 runs for every s.

TABLE 2

The "worst case" averaged over 10 runs computational times of the mixture parameter $\hat{\beta}_n$ for stacked rearrangement (SR) and stacked Grenander (SG) estimators for different sizes s of the frequency data vector $\boldsymbol{x}$.

Estimator	**s=500**	**s=1000**	**s=3000**	**s=5000**
SR	0.4 s	2.6 s	3.1 m	14.1 m
SG	0.3 s	1.6 s	3.0 m	14.0 m

Second, recall that in order to compute the confidence band for a given estimated distribution $\hat{\boldsymbol{\theta}}_n$ for estimation of the coverage probability in Table 1 we

performed 100000 Monte-Carlo simulations of the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}(\hat{\boldsymbol{\theta}}_n))$, with $\Sigma_{i,j}(\hat{\boldsymbol{\theta}}_n) = \hat{\theta}_{n,j}\delta_{i,j} - \hat{\theta}_{n,i}\hat{\theta}_{n,j}$ $(i, j = 0, \dots, t_n)$ to estimate the quantile $\hat{q}_\alpha$. The Table 3 shows the averaged over 10 runs computational times of the estimation of $\hat{q}_\alpha$ of $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}(\boldsymbol{\theta}))$ for a fixed non-random p.m.f. vector $\boldsymbol{\theta} = T^d(s)$ (recall that $T^d(s)$ is a strictly decreasing triangular function), for different values of s based on 100000 Monte-Carlo simulations.

TABLE 3
The averaged over 10 runs computational times of the quantile $\hat{q}_\alpha$ for different values of the support size s.

s=500	**s=1000**	**s=3000**	**s=5000**
14.9 s	49.6 s	7.8 m	22.0 m

All the computations were performed on MacBook Air (Apple M1 chip), 16 GB RAM. We can conclude that both stacked rearrangement and stacked Grenander estimators are computationally feasible.

6. Conclusion and discussion

In this paper we introduced and studied estimation of a discrete infinitely supported distribution by stacking the empirical estimator with Grenander estimator and the empirical estimator with rearrangement estimator.

The main results of the paper: the stacked Grenander estimator is computationally feasible, it outperforms the empirical estimator, and it is almost as good as Grenander estimator for the case of decreasing true p.m.f. Also, stacked Grenander estimator outperforms the stacked rearrangement estimator, except for the case of a strictly decreasing p.m.f. The same effect was shown in [19] for rearrangement and Grenander estimators in the case when underlying p.m.f. is decreasing. We proved that even when the true distribution is not decreasing, the estimator remains strongly consistent with $\sqrt{n}$-rate of convergence. Therefore, the stacked Grenander estimator provides a trade-off between goodness of fit and monotonicity.

The first natural generalisation of stacked Grenander estimator could be stacking with isotonic regression for a general isotonic constraint (cf. Appendix for the definition). Throughout the paper, in almost all the proofs we used properties of a general isotonic regression, cf. Lemma 3. However, the proof of Lemma 2 is based on the maximum upper sets algorithm, which is given in Lemma 4 in Appendix, and this algorithm is valid only for one dimensional monotonic case. Therefore, the generalisation of stacked Grenander estimator to the general isotonic case for finite support is straightforward, though the case of an infinite support remains an open problem.

Second, it is also important to consider other shape constraints, such as unimodal, convex and log-concave cases. Stacking these estimators is, in effect, similar to the generalisation of nearly-isotonic regression to the nearly-convex regression in [33].

Third, in this work we studied the case of discrete distribution with infinite support. The empirical estimator is closely related to estimation of probability density functions via histograms. Therefore, another direction is stacking the histogram estimators with isotonised histogram.

Forth, as mentioned in the introduction, the constrained stacked estimators have not been investigated for the case of continuous density. The interesting property of Grenander estimator in a continuous case is that the distributional pointwise rate of convergence depends on the local behaviour of the underlying distribution: if the true distribution is flat, the Grenander estimator has $n^{1/2}$-rate of convergence cf. [10], and $n^{1/3}$-rate otherwise, cf. [26]. Therefore, in the case of a continuous support it would be interesting to study stacking, for example, Grenander estimator and kernel density estimator.

Another interesting direction of research concerns the stacking with a cross-validation based on other loss functions. For the overview and theoretical properties of different loss functions for evaluation of discrete distributions we refer to the paper [17].

Finally, as we mentioned in the introduction, the problem of stacking shaped constrained regression estimators has not been studied much. Therefore, since stacked Grenander estimator performs quite well, it would be interesting to explore, for example, the prediction performance of stacked isotonic regression.

Appendix A: Appendix

We start with the definition of a general isotonic regression. Let $\mathcal{J} = \{j_1, \ldots, j_s\}$, with $s \leq \infty$, be some index set. Next, let us define the following binary relations on $\mathcal{J}$:

A binary relation $\preceq$ on $\mathcal{J}$ is a simple order if

(i) it is reflexive, i.e. $j \preceq j$ for $j \in \mathcal{J}$;
(ii) it is transitive, i.e. $j_1, j_2, j_3 \in \mathcal{J}$, $j_1 \preceq j_2$ and $j_2 \preceq j_3$ imply $j_1 \preceq j_3$;
(iii) it is antisymmetric, i.e. $j_1, j_2 \in \mathcal{J}$, $j_1 \preceq j_2$ and $j_2 \preceq j_1$ imply $j_1 = j_2$;
(iv) every two elements of $\mathcal{J}$ are comparable, i.e. $j_1, j_2 \in \mathcal{X}$ implies that either $j_1 \preceq j_2$ or $j_2 \preceq j_1$.

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

Next, a vector $\boldsymbol{v}$ with the elements indexed by $\mathcal{J}$ is isotonic if $j_1 \preceq j_2$ implies $v_{j_1} \leq v_{j_2}$. We denote the set of all isotonic square summable vectors by $\boldsymbol{\mathcal{F}}^{is}$, which is also called isotonic cone.

Furthermore, a vector $\boldsymbol{v}^* \in \mathbb{R}^s$, with $s \leq \infty$, is the isotonic regression of an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^s$ (or $\boldsymbol{v} \in \ell_2$, if $s = \infty$) over the pre-ordered index set $\mathcal{J}$

if

$$\boldsymbol{v}^* = \operatorname*{argmin}_{\boldsymbol{f}\in\boldsymbol{\mathcal{F}}^{is}} \sum_{j\in\mathcal{J}} (f_j - v_j)^2.$$

In Lemma 3 we provide properties of a general isotonic regression which are referred to in the paper.

Lemma 3. *[Properties of a general isotonic regression] Let $\boldsymbol{v}_n^* \in \ell_2$ be the isotonic regressions of some set of vectors $\boldsymbol{v}_n \in \ell_2$, for $n = 1, 2 \dots$. Then, the following holds.*

(i) $\boldsymbol{v}_n^*$ *exists and it is unique.*
(ii) $\sum_j v_{n,j} = \sum_j v_{n,j}^*$, *for all* $n = 1, 2, \dots$.
(iii) $\boldsymbol{v}_n^*$, *viewed as a mapping from* ℓ_2 *into* ℓ_2, *is continuous.*
(iv) $\boldsymbol{v}_n^*$ *satisfies the same bounds as the basic estimator, i.e.* $a \le v_{n,j}^* \le b$, *for all* $n = 1, 2, \dots$ *and* $j = 1, 2, \dots$.
(v) $\Pi(a\boldsymbol{v}_n|\mathcal{F}^{is}) = a\Pi(\boldsymbol{v}_n|\mathcal{F}^{is})$ *for all* $a \in \mathbb{R}_+$.

Proof. 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 [28], statements (iv), (v) and (vi) follow from Corollary B of Theorem 7.9, Theorems 7.5, respectively, in [6]. □

In the next lemma we describe the maximum upper sets algorithm for the solution to the isotonic regression in the monotone case.

Lemma 4. *[Maximum upper sets algorithm] For a given $\boldsymbol{x} \in \mathbb{R}_+^{t+1}$ the solution $\boldsymbol{x}^*$ of a simple order isotonic regression, i.e.*

$$\boldsymbol{x}^* = \operatorname*{argmin}_{f_0 \ge f_1 \ge \dots \ge f_{t_n}} \sum_{j=0}^{t_n} [x_j - f_j]^2$$

is given by the following algorithm. First, let us define $m(-1) = -1$. Second, we choose $m(0) > m(-1)$ to be the largest integer which maximizes the following mean

$$\frac{\sum_{k=m(-1)+1}^{m(0)} x_k}{m(0) - m(-1)}.$$

Next, let us choose $m(1) > m(0)$ to be the largest integer which maximizes

$$\frac{\sum_{k=m(0)+1}^{m(1)} x_k}{m(1) - m(0)}.$$

We continue this process and get

$$-1 = m(-1) < m(1) < \dots < m(l) = t_n.$$

The solution $\boldsymbol{x}^*$ *(i.e. the isotonic regression of* $\boldsymbol{x}$*) is given by*

$$x_j^* = \frac{\sum_{k=m(r-1)+1}^{m(r)} x_k}{m(r) - m(r-1)}$$

for $j \in [m(r-1)+1, m(r)]$ *and* $r \in [0, l]$.

Proof. The proof is given on p. 77 in [6] and p. 26 in [28], and, also, for simpler explanation of the algorithm we refer to [37]. □

Proof of Lemma 1. In order to prove the statement of the lemma, we show that the pointwise convergence almost surely of $\hat{\boldsymbol{p}}_n^{\setminus[j]}$, $\hat{\boldsymbol{r}}_n^{\setminus[j]}$ and $\hat{\boldsymbol{g}}_n^{\setminus[j]}$ for a fixed j holds. First, note that for j such that $p_j = 0$ the statement holds, since in this case we have

$$\hat{\pi}_{n,j} = \hat{\rho}_{n,j} = \hat{\gamma}_{n,j} = 0$$

for all n almost surely.

Second, let us fix some $0 \le j \le t_n$, such that $p_j \neq 0$. Next, clearly,

$$\hat{\boldsymbol{p}}_n^{\setminus[j]} \xrightarrow{a.s.} \boldsymbol{p} \tag{A.1}$$

in ℓ_k-norm for $1 \le k \le \infty$. Next, from (A.1) for the sequence $\hat{\boldsymbol{g}}_n^{\setminus[j]}$ we have

$$\hat{\boldsymbol{g}}_n^{\setminus[j]} = \Pi\big(\hat{\boldsymbol{p}}_n^{\setminus[j]} | \mathcal{F}^{decr}\big) \xrightarrow{a.s.} \boldsymbol{g}$$

in ℓ_2-norm, since the isotonic regression is a continuous map (cf. statement (iii) in Lemma 3). Therefore, we have proved the statement of the lemma for the sequences $\hat{\boldsymbol{\pi}}_n$ and $\hat{\boldsymbol{\gamma}}_n$.

Next, we prove the statement for $\hat{\boldsymbol{\rho}}_n$. Let us fix some $s > j$ such that $p_k < p_j$ for all $k > s$. Next, let

$$\hat{\mathfrak{p}}_{(k)} = \text{the } k\text{th largest of } \{\hat{p}_{n,0}^{\setminus[j]}, \dots, \hat{p}_{n,t_n}^{\setminus[j]}\}.$$

Further, from (A.1) it follows that there exists n_1 such that for all $n > n_1$

$$[\hat{\boldsymbol{r}}_n^{\setminus[j]}]^{(0,j)} = \{\hat{\mathfrak{p}}_{(1)}, \dots, \hat{\mathfrak{p}}_{(j)}\} \subset \{\hat{p}_{n,0}^{\setminus[j]}, \dots, \hat{p}_{n,s}^{\setminus[j]}\},$$

almost surely, where $[\cdot]^{(0,j)}$ denotes the first $(j+1)$ elements of the vector. Finally, since the rearrangement operator is continuous map in a finite dimensional case (Lemma 6.1 in [19]), the result of the lemma follows from continuous mapping theorem. □

Proof of Theorem 1. Recall that the least-squares cross-validation criterion is given by

$$CV(\beta) = \sum_{j=0}^{t_n} \hat{\phi}_{n,j}^2 - 2\sum_{j=0}^{t_n} \hat{p}_{n,j} \hat{\phi}_{n,j}^{\setminus[j]} =$$
$$\sum_{j=0}^{t_n} (\beta\, \hat{h}_{n,j} + (1-\beta)\hat{p}_{n,j})^2 - 2\sum_{j=0}^{t_n} \hat{p}_{n,j} (\beta\, \hat{h}_{n,j}^{\setminus[j]} + (1-\beta)\hat{p}_{n,j}^{\setminus[j]}).$$

Then, after simplification we get

$$CV(\beta) = a_n\beta^2 - 2b_n\beta + c_n,$$

where the term c_n does not depend on β, and

$$a_n = \sum_{j=0}^{t_n} (\hat{h}_{n,j} - \hat{p}_{n,j})^2,$$

and

$$b_n = \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{h}_{n,j}^{\setminus[j]} - \hat{p}_{n,j}^{\setminus[j]}) - \sum_{j=0}^{t_n} \hat{p}_{n,j}(\hat{h}_{n,j} - \hat{p}_{n,j}).$$

Assume, that $a_n \neq 0$. Then, $CV(\beta)$ is minimised by

$$\beta_n = \begin{cases} \frac{b_n}{a_n}, & \text{if } 0 \leq b_n \leq a_n, \\ 1, & \text{if } a_n \leq b_n, \\ 0, & \text{if } b_n \leq 0. \end{cases}$$

Next, note that if $\hat{\boldsymbol{p}}_n = \hat{\boldsymbol{h}}_n$, then $\hat{\boldsymbol{\phi}}_n = \hat{\boldsymbol{p}}_n = \hat{\boldsymbol{h}}_n$ for any $0 \leq \beta_n \leq 1$, and, therefore, for consistency of notation we define $\hat{\beta}_n = 0$ when $a_n = 0$. □

Proof of Lemma 2. First, we prove the statement for $\hat{\boldsymbol{\pi}}_n$. Assume that for some j we have $\hat{p}_{n,j}^{\setminus[j]} \neq 0$ and recall that

$$\hat{\pi}_{n,j} = \hat{p}_{n,j}^{\setminus[j]} = \frac{x_j - 1}{n-1}.$$

Next, note that

$$\frac{x_j - 1}{n-1} - \frac{x_j}{n} = \frac{-n + x_j}{n(n-1)} < 0.$$

Let us study the case of $\hat{\boldsymbol{\gamma}}_n$. To prove the statement of the lemma we will use maximum upper sets algorithm, which is given in Lemma 4 in the Appendix. Let $\boldsymbol{x} = (x_0, \ldots, x_{t_n})$ be frequency data from $\boldsymbol{p}$. Next, let $\boldsymbol{x}^* = (x_0^*, \ldots, x_{t_n}^*)$ be the isotonic regression of $\boldsymbol{x}$ and assume that $\boldsymbol{x}^*$ has $(l+1)$ constant regions. Let

$$m(0) < \cdots < m(l) = t_n$$

be the indices of the last elements in the constant regions of $\boldsymbol{x}^*$ and $m(-1) = -1$. Therefore, we have

$$x_j^* = \frac{\sum_{k=m(r-1)+1}^{m(r)} x_k}{m(r) - m(r-1)}$$

for $j \in [m(r-1)+1, m(r)]$ and $r \in [0, l]$.

Let us consider the first constant region of $\boldsymbol{x}^*$ and for some integer $q \in [0, m(0)]$ define vector $\boldsymbol{y} \in \mathbb{R}_+^{t+1}$

$$y_j = \begin{cases} x_j - 1, & \text{if } j = q, \\ x_j, & \text{otherwise,} \end{cases}$$

and let $\boldsymbol{y}^*$ be isotonic regression of $\boldsymbol{y}$.

Recall, $m(0)$ is the largest non-negative integer which maximizes the following mean

$$S_1 = \frac{\sum\limits_{k=0}^{m(0)} x_k}{m(0)+1}.$$

Further, let $m'(0)$ be the largest non-negative integer which maximizes the following mean for the vector $\boldsymbol{y}$

$$S_2 = \frac{\sum\limits_{k=0}^{m'(0)} y_k}{m'(0)+1}.$$

Let us prove that $S_2 \leq S_2$. First, assume that $m'(0) = m(0)$, then, clearly, $S_2 \leq S_1$ since $y_j \leq x_j$. Second, let us assume that $m'(0) \neq m(0)$. Then, from the definitions of $m(0)$ and $m'(0)$ it follows

$$S_2 = \frac{\sum\limits_{k=0}^{m'(0)} y_k}{m'(0)+1} \leq \frac{\sum\limits_{k=0}^{m'(0)} x_k}{m'(0)+1} \leq \frac{\sum\limits_{k=0}^{m(0)} x_k}{m(0)+1} = S_1.$$

Next, assume that q is not in the first constant region. Then in this case from maximum upper sets algorithm it follows that the constant regions in the isotonic regressions $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ are the same up to the region which contains element with index m. Then, we can use the same approach as for the first region. Therefore, we have proved that $y_q^* \leq x_q^*$.

Next, from statement (v) of Lemma 3 for $\hat{\boldsymbol{g}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ we have

$$\hat{g}_{n,j} = \frac{x_j^*}{n},$$

and

$$\hat{\gamma}_{n,j} = \frac{y_j^*}{n-1},$$

therefore, we proved that

$$\hat{\gamma}_{n,j} \leq \frac{n}{n-1}\hat{g}_{n,j}.$$

Finally, we prove the inequality for $\hat{\boldsymbol{\rho}}_n$. Analogously to the case of $\hat{\boldsymbol{\gamma}}_n$, let us consider the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, discussed above. Note that $y_j \leq x_j$ for all

j, therefore, the same componentwise inequality holds for the sorted vectors $rear(\boldsymbol{x})$ and $rear(\boldsymbol{y})$. Next, using the definition of $\hat{\boldsymbol{r}}_n$ and $\hat{\boldsymbol{\rho}}_n$ we prove that

$$\hat{\rho}_{n,j} \leq \frac{n}{n-1}\hat{r}_{n,j}.$$

□

Proof of Theorem 5. Assume that the p.m.f. $\boldsymbol{p}$ is not decreasing. Note that

$$||\sqrt{n}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{p}) - \sqrt{n}(\hat{\boldsymbol{p}}_n - \boldsymbol{p}))||_2 = \sqrt{n}||\hat{\boldsymbol{\phi}}_n - \hat{\boldsymbol{p}}_n||_2 \leq$$
$$\hat{\beta}_n\sqrt{n}||\hat{\boldsymbol{h}}_n - \hat{\boldsymbol{p}}_n||_2 + (1-\hat{\beta}_n)\sqrt{n}||\hat{\boldsymbol{p}}_n - \hat{\boldsymbol{p}}_n||_2 = \hat{\beta}_n\sqrt{n}||\hat{\boldsymbol{h}}_n - \hat{\boldsymbol{p}}_n||_2.$$

Then, since

$$||\hat{\boldsymbol{r}}_n - \hat{\boldsymbol{p}}_n||_2 \overset{a.s.}{\to} ||\boldsymbol{r} - \boldsymbol{p}||_2 < \infty,$$
$$||\hat{\boldsymbol{g}}_n - \hat{\boldsymbol{p}}_n||_2 \overset{a.s.}{\to} ||\boldsymbol{g} - \boldsymbol{p}||_2 < \infty,$$

and using (4.5) we have

$$\hat{\beta}_n\sqrt{n}||\hat{\boldsymbol{r}}_n - \hat{\boldsymbol{p}}_n||_2 \overset{a.s.}{\to} 0,$$
$$\hat{\beta}_n\sqrt{n}||\hat{\boldsymbol{g}}_n - \hat{\boldsymbol{p}}_n||_2 \overset{a.s.}{\to} 0,$$

which leads to

$$\sqrt{n}||\hat{\boldsymbol{\phi}}_n - \hat{\boldsymbol{p}}_n||_2 \overset{a.s.}{\to} 0.$$

The statement of the theorem now follows from Theorem 3.1 in [8].

Assume that $\boldsymbol{p}$ is a strictly decreasing p.m.f. over $\{0,\dots,s\}$, with $s < \infty$. Next, let $\varepsilon = \inf\{|p_j - p_{j+1}| : j = 0,\dots,s-1\}$ and note that

$$\{\sup_j\{|\hat{p}_j - p_j| < \varepsilon/2\} \subseteq \{\hat{h}_{n,j} = \hat{p}_{n,j}\}$$

for both $\hat{\boldsymbol{h}}_n = \hat{\boldsymbol{g}}_n$ and $\hat{\boldsymbol{h}}_n = \hat{\boldsymbol{r}}_n$. Therefore, this implies that for any $j = \{0,\dots,s$ we have

$$\mathbb{P}[\hat{\phi}_{n,j} = \hat{p}_{n,j}] \geq \mathbb{P}[\sup_j\{|\hat{p}_{n,j} - p_j| < \varepsilon/2] \to 1,$$

since the empirical estimator is strongly consistent. The statement of the theorem follows from Theorem 3.1 in [8]. □

References

[1] Balabdaoui, F., Durot, C., Koladjo, F. (2017). On asymptotics of the discrete convex LSE of a p.m.f. *Bernoulli* **23**, 1449–1480. MR3624867

[2] Balabdaoui, F. and de Fournas-Labrosse, G.(2020). Least squares estimation of a completely monotone pmf: From Analysis to Statistics. *Journal of Statistical Planning and Inference*, **204**, 55–71. MR3961929

[3] Balabdaoui, F. and Jankowski, H. (2016). Maximum likelihood estimation of a unimodal probability mass function. *Statistica Sinica* **26**, 1061–1086. MR3559943

[4] Balabdaoui, F., Jankowski, H., Rufibach, K., and Pavlides, M.(2013). Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *Journal of the Royal Statistical Society: SERIES B: Statistical Methodology*, **75**, 769–790. MR3091658

[5] Balabdaoui, F. and Kulagina, Y.(2020). Completely monotone distributions: Mixing, approximation and estimation of number of species. *Computational Statistics & Data Analysis*, **150**, 107014. MR4101998

[6] 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.

[7] Best, M. J. and Nilotpal C. (1990). Active set algorithms for isotonic regression; A unifying framework. *Mathematical Programming*, **47**, 425–439. MR1068274

[8] Billingsley, P. (2013). *Convergence of probability measures.*. John Wiley & Sonsc. MR0233396

[9] Breiman, L. (1995). Stacked regressions. *Machine Learning*, **24**, 49–64.

[10] Carolan, C. and Dykstra, R. (1999). Asymptotic behavior of the Grenander estimator at density flat regions. *Canadian Journal of Statistics*, **27**, 557–566. MR1745821

[11] Chu, C. Y., Henderson, D. J. and Parmeter, C. F. (2015). Plug-in bandwidth selection for kernel density estimation with discrete data. Econometrics, **3**, 199–214.

[12] Chu, C. Y., Henderson, D. J. and Parmeter, C. F. (2017). On discrete Epanechnikov kernel functions. Computational Statistics & Data Analysis, **116**, 79–105. MR3692306

[13] Durot, C., Huet, S., Koladjo, F. and Robin, S. (2014). Least-squares estimation of a convex discrete distribution. *Computational Statistics & Data Analysis*, **67**, 282–298. MR3079603

[14] Fang, Z., Meinshausen, N. (2012). Liso isotone for high-dimensional additive isotonic regression. *Journal of Computational and Graphical Statistics*, **21**, 72–91. MR2913357

[15] Fienberg, S. E. and Holland, P. W. (1972). On the choice of flattening constants for estimating multinomial probabilities. *Journal of Multivariate Analysis*, **2**, 127–134. MR0303651

[16] Fienberg, S. E. and Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, **68**, 683–691. MR0359153

[17] Haghtalab, N., Musco, M. and Waggoner, B. (2019). Toward a Characterization of Loss Functions for Distribution Learning. Tech. rep., arXiv:1906.02652v2.

[18] Hastie, T., Tibshirani, R., and Tibshirani, R. (2020) Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science* **35**, 579–592. MR4175382

[19] Jankowski, H. K. and Wellner, J. A. (2009). Estimation of a discrete monotone distribution. *Electronic journal of statistics*, **39**, 125–153. MR2578839

[20] Jankowski, H. and Tian, Y. H. (2018). Estimating a discrete log-concave distribution in higher dimensions. *Statistica Sinica*, **28**, 2697–2712. MR3840011

[21] LeBlanc, M. and Tibshirani, R. (1996). Combining estiamates in regression and lassification. *Journal of the American Statistical Association*, **91**, 1641–1650. MR1439105

[22] Luss, R. and Rosset, S. (2017). Bounded isotonic regression. *Electronic Journal of Statistics*, **11**, 4488–4514. MR3724487

[23] Minami, K. (2020). Estimating piecewise monotone signals. *Electronic Journal of Statistics*, **14**, 1508–1576. MR4082476

[24] Ouyang, D., Li, Q., and Racine, J. (2006). Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics*, **18**, 69–100. MR2214066

[25] Racine, J. S., Li, Q. and Yan, K. X. (2020). Kernel smoothed probability mass functions for ordered datatypes. *Journal of Nonparametric Statistics*, **32**, 563–586. MR4136583

[26] Rao, B. P. (1969). Estimation of a unimodal density. *Sankhyā: The Indian Journal of Statistics*, **31**, 23–36. MR0267677

[27] Rigollet, P., and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, **16**, 260–280. MR2356821

[28] Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. John Wiley & Sons, Ltd., Chichester. MR0961262

[29] Silvapulle, M. J. and Sen, P. K. (2005). *Constrained Statistical Inference*. John Wiley & Sons, Ink., Hoboken, New Jersey. MR2099529

[30] Smyth, P. and Wolpert, D. (1999). Linearly combined density estimators via stacking. *Machine Learning*, **36**, 59–83.

[31] Stone, M. (1974). Cross-Validation and Multinomial Prediction. *Biometrika*, **61**, 509–515. MR0415896

[32] Stout, Q. F. (2013). Isotonic Regression via Partitioning. *Algorithmica*, **66**, 93–112. MR3023808

[33] Tibshirani, R. J., Hoefling, H. and Tibshirani, R. (2011). Nearly-isotonic regression. *Technometrics*, **53**, 54–61. MR2791946

[34] Tribula, S. (1958). Some Problems of Simultaneous Minimax Estimation. *The Annals of Mathematical Statistics*, **29**, 245–253. MR0093856

[35] Wolpert, D. (1992). Stacked Generalization. *Neural Networks*, **5**, 241–259.

[36] Wright, F. T. (1978). Estimating strictly increasing regression functions. *Journal of the American Statistical Association*, **73**, 636–639.

[37] Wright, F. T. (1982). Monotone regression estimates for grouped observations. *The Annals of Statistics*, **10**, 278–286. MR0642739