Maximum likelihood estimation of a unimodal probability mass function

Fadoua Balabdaoui and Hanna Jankowski

October 13, 2015

Abstract

We develop an estimation procedure of a discrete probability mass function (pmf) with unknown support. We derive the maximum likelihood estimator of this pmf under the mild and natural shape-constraint of unimodality. Shape-constrained estimation is a powerful and robust technique, which additionally provides smoothing of the empirical distribution yielding thereby gains in efficiency. We show that our unimodal estimator is consistent when the model is specified, and that it converges to the best projection of the true pmf on the unimodal class under model misspecification. Furthermore, we derive the limiting distribution of the estimator, and use the obtained result to build asymptotic confidence bands for the unknown pmf when the latter is unimodal. We illustrate our approach using time-to-onset data of the Ebola virus during the 1976 outbreak in the former republic of Zaire.

1 Introduction

Discrete or discretized data show up in many practical instances, see Harlan et al. (2014); Chowell et al. (2013, 2009); Laskowski et al. (2011); Breman and Johnson (2014). If computing the empirical distribution requires no assumptions on the unknown law, gains in efficiency can be made by imposing additional constraints. Such a constraint is unimodality, which is a natural and mild assumption in many real statistical applications.

Nonparametric estimation of a unimodal density has been treated in many research papers. In case the mode is known, the problem boils down to fitting the well-known Grenander estimator (Grenander, 1956). However, as noted by Birgé (1997), it is unrealistic in practice to assume that the location of the mode is known. The main consequence of not making such an assumption is that the maximum likelihood estimator (MLE) fails to exist. To address this problem, several estimators have been proposed, see Wegman (1968, 1969); Prakasa Rao (1969); Wegman (1970a,b); Reiss (1973, 1976) in which the Grenander estimator has been additionally constrained.
More recent work appears in Birgé (1997), where the proposed estimator is chosen among all the possible unimodal Grenander estimators as the one with cumulative distribution function closest to the empirical distribution. Durot et al. (2013) consider the estimation of a discrete convex distribution, using the least squares criterion. Recently, Dümbgen and Rufibach (2009); Cule et al. (2010) proposed to use the maximum likelihood estimator of a log-concave density in lieu of the unimodal assumption, partially due to the inherent problems faced when estimating a unimodal density using maximum likelihood. As opposed to the continuous setting, existence of the unimodal MLE when the data are discrete is guaranteed, even when the mode is unknown. On the other hand, uniqueness is not always true, but this problem is rather marginal, as a rule for selecting from among the finite options is immediate, making our estimator fully automatic and easy to compute. Furthermore, if the pmf is not unimodal, the MLE is still consistent, in the sense that it approaches the best unimodal pmf among a finite number of choices. Further details of this behavior are provided in Section 4.

In the recent work of Balabdaoui et al. (2013), the discrete MLE under the constraint of log-concavity was studied. One important consequence of this work is that we can evaluate the loss when data exhibit unimodality but at the same time log-concavity would not be a valid assumption. The unimodal MLE seems to be a more natural estimator to consider when additional features of the true distribution besides unimodality are lacking or hard to obtain. On the other hand, it is expected the log-concave MLE to be more efficient than the unimodal one in case log-concavity is a correct assumption about the model. This is studied via simulations in Section 3. Although restricted to discrete distributions, our results may be interesting to those studying the continuous setting as well.

The manuscript is organized as follows. In Section 2, we provide the technical details required to define and compute the MLE of a discrete unimodal distribution. In our set-up, the support is assumed to be unknown, and is also estimated empirically from the data. In Section 3, we consider the finite sample size behavior of our estimator via simulations. As mentioned previously, we compare here our estimator with the discrete log-concave MLE, but also we assess the loss of efficiency when the support is unknown and must be estimated from the data. Sections 4 and 5 we establish consistency and global asymptotic theory for the estimator. One of our key contributions is the application of these to develop global confidence bands for a unimodal pmf, see Section 6. Finally, we illustrate the estimator on a data set for the 1976 Ebola outbreak in Zaire; see Section 7. The data clearly shows a drastic difference in the time from infection to onset of symptoms depending on the type of infection: whether the individual was infected from person-to-person contact or from injection with an unsterilized needle. R (R Core Team, 2014) code for this analysis (along with all simulations) is available online at www.math.yorku.ca/~hkj/Research/. All proofs and additional
details are left to the Appendices of the online supplementary material.

2 Maximum likelihood estimation

2.1 Discrete unimodal distributions

In this work, we consider estimation of a unimodal pmf of a discrete real-valued random variable. We denote the support of such a pmf as $S = \{s_i\}_{i \in K}$, where $K$ is a subset of $\mathbb{Z}$. Without loss of generality, we take $s_i \in \mathbb{R}$ for all $i \in K$, and we assume that $s_i < s_{i+1}$.

We say that a pmf $p$ is unimodal if there exists an integer $m$ such that

$$p(s_i) \geq p(s_{i+1}), \quad \text{for all } i \geq m,$$

and

$$p(s_{i-1}) \leq p(s_i), \quad \text{for all } i \leq m.$$  \hspace{1cm} (2.1)

The element $s_m$ is thus a mode of the pmf $p$, but is not necessarily unique.

In general, we can define the modal region, denoted here by $\mathcal{M}$,

$$\mathcal{M} = \{s_\kappa \in S: \ p \text{ satisfies (2.1) at } m = \kappa\}$$  \hspace{1cm} (2.2)

$\mathcal{M}$ is necessarily a finite set and we have that $p(s) = p(s')$ for all $s, s' \in \mathcal{M}$.

Next, let $\mathcal{U}^1(S)$ denote the space of unimodal pmfs with the same fixed support $S$. For the purpose of estimating such a $p$, it is most convenient to decompose the space of unimodal pmfs as

$$\mathcal{U}^1(S) = \bigcup_{\kappa \in K} \mathcal{U}^1|_{\kappa}(S),$$  \hspace{1cm} (2.3)

where $\mathcal{U}^1|_{\kappa}(S)$ is the space of pmfs which are increasing on $\{s_i: i \leq \kappa - 1\}$ and decreasing on $\{s_i: i \geq \kappa\}$. Note that a pmf in $\mathcal{U}^1|_{\kappa}(S)$ is unimodal either at $s_{\kappa-1}$ or $s_{\kappa}$ depending on the order of its values at these points. It may seem, at first, that it would be more natural to decompose $\mathcal{U}^1(S)$ into the spaces of pmfs that are unimodal at $\kappa$. However, it turned out that the decomposition (2.3) is much more convenient. In addition, the MLE will always “decide” between these two possibilities by choosing the one that yields the largest value of the likelihood. Note also that if $\kappa = \min K$, then $\mathcal{U}^1|_{\kappa}(S)$ is simply the space of non-increasing pmfs on $S$. Notably, each space $\mathcal{U}^1|_{\kappa}(S)$ is convex, whereas $\mathcal{U}^1(S)$ is not.

Known as Khintchine’s Theorem, a density with respect to Lebesgue measure is unimodal if and only if it can be written as a mixture of uniform densities, see for example Olshen and Savage (1970). Hence, it is expected that such a representation exists also in the discrete setting.

Proposition 2.1. A pmf $p$ satisfies $p \in \mathcal{U}^1|_{\kappa}(S)$ if and only if

$$p(s_i) = \sum_{j \geq 0} \frac{1_{i \in (\kappa, \kappa+j)} q(s_j)}{j+1} + \sum_{j \leq -1} \frac{1_{i \in (\kappa+j, \kappa-1)} q(s_j)}{|j|} q(s_j),$$  \hspace{1cm} (2.4)

for some pmf $q$ with support $S$. 

3
A proof of Proposition 2.1 can be found in the Appendix. Using (2.3), a unimodal \( p \in \mathcal{U}^1(S) \) admits a representation (2.4) for some \( \kappa \in K \).

**Remark 2.2.** Suppose that \( K \) is finite, and write \( K = \{0, 1, \ldots, k\} \). Then \( \mathcal{U}^1|_0(S) \subset \mathcal{U}^1|_1(S) \).

### 2.1.1 Relationship with unimodal densities

Given a probability mass function \( p \) with support on \( S = \mathbb{Z} \), one can define a density function \( f \) on \( \mathbb{R} \) by

\[
f(x) = p(z) \quad \text{for } x \in (z-1, z], \ z \in \mathbb{Z}.
\]

The mass function \( p \) is unimodal iff the (piecewise constant) density \( f \) is unimodal.

Given a general unimodal density with support on \( \mathbb{R} \), one can also define a unimodal pmf \( p \) via

\[
p(z) = \int_{z-1}^{z} f(x) \, dx, \quad z \in \mathbb{Z}.
\]

Here, the choice of the discretization on \( [z-1, z) \) is arbitrary. Indeed, any choice of \( a \in \mathbb{R} \), with \( [z + a - 1, z + a) \) is possible. One could also consider intervals of length other than one, as long as the length is fixed.

In this sense, discrete distributions provide a useful way to analyze data which has been “discretized” in such a manner. One such example is considered in Section 7. This relationship with unimodal densities is particularly noteworthy, since, although the MLE of a unimodal density does not exist, the MLE of its discretized version does.

### 2.1.2 When the true support is unknown

The discussion above relating unimodal densities and pmfs implies that one natural assumption on the support \( S \) is that it is a connected subset of \( a + \delta \mathbb{Z} \), for some \( a \in \mathbb{R} \) and \( \delta > 0 \). However, we believe that in certain instances additional generality may be required. For this reason, the only assumption we make about the support \( S \) is that it is an ordered subset of \( \mathbb{R} \). This assumption provides additional flexibility to our approach: unimodality of a pmf is preserved under scalar transformations (if the pmf of a random variable \( X \) is unimodal, then so is the pmf of \( aX \)), and under removal of elements of the support.

In order to reflect this flexibility in our estimation approach, we do not assume that the true support is known a priori. Instead, we estimate both the unimodal pmf and its support from the collected observations. However, if the true support is known a priori, then it is expected that more efficiency would be gained by including this information to the estimation procedure. Some simulations studying this are given in Section 3.1.
Figure 1: The same empirical observations (shown in grey) yield two different solutions maximizing the likelihood.

Notably, our consistency and asymptotic results developed later apply to both versions of the MLE (either when the support is known or when it is unknown). All theoretical results are proved and stated for the unknown support version; the proofs are only simplified when the support is known.

2.2 The unimodal maximum likelihood estimator

Let $X_1, \ldots, X_n$ be $n$ independent observations from a discrete pmf $p_0$. Also, let $p_n(z) = n^{-1} \sum_{j=1}^{n} I_{\{z\}}(X_j)$ and $F_n(z) = n^{-1} \sum_{j=1}^{n} I_{((-\infty,z])}(X_j)$ denote the empirical pmf and the associated empirical cumulative distribution function (cdf), respectively. Finally, let $S_n$ denote the observed support of $p_n$, that is $S_n = \{z_0, \ldots, z_{J-1}\}$ is the set of distinct values in the sample $\{X_1, \ldots, X_n\}$. We assume in our notation that $z_0 < z_1 < \ldots < z_{J-1}$.

2.2.1 Definition

Recall that we do not assume that the support $S$ is known. Thus, we define the maximum likelihood estimator (MLE) as

$$\hat{p}_n = \arg\max_{p \in \mathcal{U}(S_n)} L_n(p),$$
where the log-likelihood is given by

$$L_n(p) = \int \log p(z) \, dF_n(z) = \sum_{j=0}^{J-1} \log(p(z_j)) \, \overline{p}_n(z_j).$$

This maximization is done in two steps: (1) we maximize $L_n$ over the space $\mathcal{U}^1|_\kappa(S_n)$ for each $\kappa$, and (2) we find $\overline{p}_n$ and the corresponding estimator at which the overall maximum is attained.

### 2.2.2 The shape operators iso, anti, and uni

To describe and compute the MLE, we find it convenient to first define several shape operators. For any $z \in \mathbb{R}^d$, we denote $z_{st}$ the sub-vector $(z_s, \ldots, z_t)$ where $1 \leq s \leq t \leq d$. Consider the following sets of constrained vectors

$$\mathcal{I}_d = \{ u = (u_1, \ldots, u_d) \in \mathbb{R}^d : u_1 \leq \cdots \leq u_d \},$$

$$\mathcal{D}_d = \{ w = (w_1, \ldots, w_d) \in \mathbb{R}^d : w_1 \geq \cdots \geq w_d \}.$$

Also, for $\kappa \in \{1, \ldots, d\}$, let

$$\mathcal{U}_{d|\kappa} = \{ z = (z_1, \ldots, z_d) \in \mathbb{R}^d : z_1:<(\kappa-1) \in \mathcal{I}_{\kappa-1} \text{ and } z_{\kappa:d} \in \mathcal{D}_{d-\kappa+1} \},$$

$$\mathcal{U}_d = \bigcup_{\kappa=1}^d \mathcal{U}_{d|\kappa}.$$

Lastly, we denote the $\ell_2$ distance by $\|v - u\|_2^2 = \sum_{j=1}^{d} (v_j - u_j)^2$.

We can now define the first two operators $\text{iso} : \mathbb{R}^d \to \mathcal{I}_d$ and $\text{anti} : \mathbb{R}^d \to \mathcal{D}_d$ as

$$\text{iso}[v] = \arg\min_{u \in \mathcal{I}_d} \|v - u\|_2,$$

$$\text{anti}[v] = \arg\min_{u \in \mathcal{D}_d} \|w - u\|_2.$$

In other words, $\text{iso}[v]$ and $\text{anti}[v] = -\text{iso}[-v]$ are the well-known least squares projections of $v$ on the spaces $\mathcal{I}_d$ and $\mathcal{D}_d$ respectively; cf. Barlow et al. (1972); Sen and Meyer (2013). Note also that the operator anti is the same as the gren operator discussed in Jankowski and Wellner (2009); Jankowski (2014).

Finally, for $\kappa \in \{1, \ldots, d\}$, define the operators $\text{uni}_\kappa : \mathbb{R}^d \to \mathcal{U}_{d|\kappa}$ and $\text{uni} : \mathbb{R}^d \to \mathcal{U}_d$ as

$$\text{uni}_\kappa[v] = (\text{iso}[v_1:<(\kappa-1)], \text{anti}[v_{\kappa:d}]) = \arg\min_{u \in \mathcal{U}_{d|\kappa}} \|v - u\|_2,$$

$$\text{uni}[v] = \arg\min_{u \in \mathcal{U}_d} \|v - u\|_2.$$

Note that, as before, we have that

$$\text{uni}[v] = \text{uni}_{\kappa,\overline{\kappa}}[v], \text{ where } \overline{\kappa} \in \arg\min_{\kappa} \|v - \text{uni}_\kappa[v]\|_2.$$

The operators iso and anti are unique. However, the operator uni may yield more than one solution, much like the operator yielding the MLE. Properties of these operators are discussed in detail in Appendix C.3 of the supplementary material.
2.2.3 Existence and characterization of the MLE

Using these operators, we may now state some facts about the MLE.

**Proposition 2.3.** The restricted MLE $\hat{\lambda}_n|_\kappa$ exists and is unique. Furthermore, it is characterized by

$$\hat{\lambda}_n|_\kappa = \text{uni}_\kappa[\lambda_n].$$

The (unrestricted) unimodal MLE $\hat{\lambda}_n$ exists, but is not necessarily unique. For $\{\hat{\kappa}_n\} = \arg\max_{1 \leq \kappa \leq J-1} L_n(\hat{\lambda}_n|_\kappa)$, the (finite) collection of solutions to the maximization problem, $\{\hat{\lambda}_n\}$, is characterized as

$$\{\hat{\lambda}_n\} = \{\hat{\lambda}_n|_\kappa; \kappa \in \{\hat{\kappa}_n\}\}. \quad (2.5)$$

Note that the MLE is not defined in terms of the operator uni, however, the operator does show up in its limiting distribution.

**Remark 2.4.** One of the key conclusions of Proposition 2.3 is that the size of the set $\{\hat{\lambda}_n\}$ may be greater than one (see, for example, Figure 1). To overcome the computational difficulties that would result from this non-uniqueness, we simply take the MLE to be equal to the maximizer with the smallest mode. That is, let $\hat{\kappa}_n$ denote the smallest integer $\kappa$ such that

$$L_n(\hat{\lambda}_n|_\kappa) = \max_{1 \leq \kappa \leq J-1} L_n(\hat{\lambda}_n|_\kappa).$$

Then $\hat{\lambda}_n = \hat{\lambda}_n|_{\hat{\kappa}_n}$. Note the slight abuse of notation: we denote $\hat{\kappa}_n$ as the smallest element of $\{\hat{\kappa}_n\}$. Also, note that in order to find $\hat{\kappa}_n$ we can search only over $1 \leq \kappa \leq J-1$ using Remark 2.2.

The characterization in (2.5) along with our convention provides a straightforward way to compute $\hat{\lambda}_n$. Namely, we first find the restricted MLE $\hat{\lambda}_n|_\kappa$ as the right slopes of the greatest convex minorant of $\{(0, 0), (z_j, F_n(z_j)), 0 \leq j \leq \kappa-1\}$ and the left slopes of the least concave majorant of $\{(0, 0), (z_j, F_n(z_j)), \kappa \leq j \leq J-1\}$. The MLE $\hat{\lambda}_n$ will be then taken to be equal to $\hat{\lambda}_n|_{\hat{\kappa}_n}$ which maximizes the overall likelihood for the smallest integer $\kappa$. Proposition 2.3 follows immediately from the more general result of Theorem 4.2 as well as Lemma C.3 given in Appendix C of the online supplementary material.

3 Finite sample performance of the MLE

Here we compare three maximum likelihood estimators for small and medium samples sizes. The three estimators are

1. the MLE under no assumption on the pmf; i.e., the empirical MLE,
2. the MLE assuming the pmf is unimodal ($\hat{\lambda}_n$ as defined in this work),
(3) the log-concave MLE assuming the pmf is log-concave. Theoretical and computational aspects of this estimator have been studied in Balabdaoui et al. (2013).

In our simulations, we consider six different distributions:

- The negative binomial distribution with parameters \( r = 6, p = 0.3 \). This is a distribution is both strictly unimodal and strictly log-concave.

- The double logarithmic distribution with \( S = Z \), which we define as

\[
p(z) = \begin{cases} 
p^{|z|} & z \leq -1 \\
p^{|z|}(p - \log(1-p)) & z = 0, \\
\frac{p^{|z|}}{\log(1-p)} & z \geq 1. 
\end{cases}
\]

This distribution is strictly unimodal but not log-concave. In the simulations, we take \( p = 0.9 \).

- The uniform pmf with \( S = \{0, \ldots, 9\} \). This is an example of a pmf which is neither strictly unimodal nor strictly log-concave.

- The mixture of uniform distributions with support on \( \{0, \ldots, 49\} \), with pmf given by taking \( S = Z, \kappa = 0 \) and

\[
q(z) = \begin{cases} 
1/3 & z = 9, 39, 49 \\
0 & \text{otherwise.}
\end{cases}
\]

in decomposition (2.1). This distribution is unimodal (though not strictly unimodal), and is not log-concave.

- The Poisson with rate \( \lambda = 2 \), a strictly log-concave and unimodal distribution.

- A mixture of Poisson distributions: Letting \( p_\lambda \) denote the pmf of a Poisson distribution with rate \( \lambda \), then the mixture we consider is given by

\[
\frac{1}{4} \cdot p_1(\cdot) + \frac{1}{8} \cdot p_3(\cdot) + \frac{5}{8} \cdot p_8(\cdot).
\]

This distribution is (strictly) bimodal, and is therefore neither unimodal nor log-concave.

<table>
<thead>
<tr>
<th></th>
<th>unimodal</th>
<th>log-concave</th>
<th>finite support</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative binomial</td>
<td>yes (strict)</td>
<td>yes (strict)</td>
<td>no</td>
</tr>
<tr>
<td>double logarithmic</td>
<td>yes (strict)</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>uniform</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>uniform mixture</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Poisson</td>
<td>yes (strict)</td>
<td>yes (strict)</td>
<td>no</td>
</tr>
<tr>
<td>Poisson mixture</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>
Figure 2: Boxplots of the $\ell_2$ distance of the estimated pmf from the true pmf under each of three estimators: the empirical MLE (1), the unimodal MLE (2), the log-concave MLE (3). Each boxplot is the result of $B = 1000$ simulations. Properties of these distributions are summarized in Table 1.
Figure 3: Boxplots for the six distributions of Figure 2 of the $\ell_2$ distance of the estimated pmf from the true pmf for the unimodal MLE when the support is known (a) and unknown (b). Each boxplot is the result of $B = 1000$ simulations.
Properties of the six distributions are summarized in Table 1 for convenience.

In Figure 2, we can see that the unimodal MLE performs better than the empirical MLE for all six distributions. It is, however, outperformed by the log-concave MLE for the distributions which are log-concave, although they seem to have very comparable errors in the case of the Poisson distribution with rate \( \lambda = 2 \). On the other hand, the unimodal MLE outperforms the log-concave MLE when the distribution is not log-concave, at least for sample sizes which are “large enough”. Our simulations show that this sample size is related to how far away the true pmf is from the set of log-concave distributions. In Figure 2, the \( \ell_2 \) distance to the corresponding log-concave Kullback-Leibler projection (cf. Balabdaoui et al. (2013)) is approximatively 0.363 for the double logarithmic and 0.050 for the uniform mixture. Overall, we expect that when log-concavity fails to hold, the unimodal MLE will be the better estimator for larger sample sizes. Moreover, this behavior will hold also for smaller sample sizes for pmfs that are further away from the log-concave class. The bimodal Poisson mixture model is the only example in which neither the log-concave nor unimodal classes are correct. Notably, although the empirical pmf is the only well-specified MLE in this case, it outperforms the other two estimators only for the largest sample size.

### 3.1 Comparison of known versus unknown support

It seems self-evident that some efficiency will be lost by assuming that the support is unknown. Here, we briefly consider the question of “how much efficiency is lost?” via simulations. To be precise, when we say that the support is known, the MLE is defined as

\[
\hat{p}_n = \arg\max_{p \in \mathcal{U}_1(S)} L_n(p),
\]

unlike in the definition of \( \bar{p}_n \), where \( S \) is replaced by its estimate \( S_n \). In order to avoid existence issues of the estimator defined above, the class \( \mathcal{U}_1(S) \) should be viewed as the set of probability mass functions \( p \) with support contained in \( S \). Our simulations show that although some difference is seen for small a sample size, the cost is not great, and the difference disappears with increased sample size. As mentioned previously, our consistency and asymptotic results developed later apply to both versions of the MLE.

Figure 4 gives an example of the two approaches (known vs. unknown support in the unimodal MLE) for a sample from the negative binomial distribution for \( n = 50 \). Both unimodal MLE approaches provide considerable “smoothing” to the empirical pmf. However, when the support is unknown, the MLE will only place mass on \( S_n \), whereas the MLE with known support will place mass on the entire range \( \{X_{(1)}, \ldots, X_{(n)}\} \). This is clearly seen in Figure 4. Thus, the potential loss of efficiency will most likely occur in the tails of the true distribution, and this will be particularly true for distributions with a fatter tail.
Figure 4: Example comparing the unimodal MLE when the support is known vs. unknown. The true distribution is the negative binomial with sample size $n = 50$.

We compared the two approaches via simulations, the results of which are shown in Figure 3. The distributions considered are exactly the same as those described on page 8. The loss is small for the uniform distribution which has support on only ten points, and also for both Poisson distributions, where the tails converge to zero very quickly. For the other distributions, which slower rate of decay in the tails, some efficiency is lost for the small sample size $(n = 50)$. However, the loss appears almost negligible for the medium sample size $(n = 200)$.

Remark 3.1. When $S = S_n$, the known support and unknown support MLE versions will be the same. For the case when $|S|$ is finite, the probability that this does not happen for a given $n$ decreases exponentially with $n$. Furthermore, with probability one, there exists an $n_0$, such that for all $n \geq n_0$, $S = S_n$ in this case.

4 The Kullback-Leibler projection and consistency of the unimodal MLE

Let $p_0$ denote a fixed probability mass function on $S_0$ with distribution function $P_0$. We let

$$\rho(p|p_0) = \int \log \frac{p_0}{p} dP_0,$$
denote the Kullback-Leibler (KL) divergence. In this section, we seek the KL projection \( \overline{p}_0 \in \mathcal{U}^1(S_0) \) of a given pmf \( p_0 \). The KL projection has been considered extensively for the log-concave shape constraint for densities on \( \mathbb{R}^d \) in Cule and Samworth (2010) and Dümbgen et al. (2011) and for probability mass functions in Balabdaoui et al. (2013). As in Cule and Samworth (2010); Cule et al. (2010); Balabdaoui et al. (2013), we can define such a projection as

\[
\overline{p}_0 = \arg\min_{p \in \mathcal{U}^1(S_0)} \int_{S_0} \log \frac{p_0}{p} \, dP_0 = \arg\min_{p \in \mathcal{U}^1(S_0)} \rho(p|p_0), \tag{4.8}
\]

which is the element of \( \mathcal{U}^1(S_0) \) closest to the unknown pmf \( p_0 \) in the sense of Kullback-Leibler divergence. From a practical point of view, this allows us to view the shape constrained estimator as the closest approximation within a class of distributions.

Alternatively, Patilea (2001) uses the definition

\[
\int \log \frac{\overline{p}_0}{p} \, dP_0 \geq 0, \quad \text{for all } p \in \mathcal{U}^1(S_0), \tag{4.9}
\]

and refers to the pmf \( \overline{p}_0 \) satisfying (4.9) as the pseudo-true pmf. If the integrals involved are finite, one can re-arrange (4.9) into (4.8) and vice versa. In particular, if \( \inf_{q \in \mathcal{U}^1(S_0)} \rho(q|p_0) = \rho(\overline{p}_0|p_0) < \infty \) for some \( \overline{p}_0 \) then (4.8) is equivalent to (4.9), since then

\[
0 \leq \int \log \frac{p_0}{\overline{p}_0} \, dP_0 \leq \int \log \frac{p_0}{p} \, dP_0 = \int \log \frac{p_0}{\overline{p}_0} \, dP_0 = \int \log \frac{p_0}{\overline{p}_0} \, dP_0 + \int \log \frac{\overline{p}_0}{p} \, dP_0.
\]

Alternatively, as in Dümbgen et al. (2011), one could also consider

\[
\int \log \overline{p}_0 \, dP_0 \geq \int \log p \, dP_0, \quad \text{for all } p \in \mathcal{U}^1(S_0), \tag{4.10}
\]

which is akin to maximizing the likelihood. If \( p_0 \) admits a finite entropy, that is \( \int \log p_0 \, dP_0 > -\infty \), then (4.10) is equivalent to (4.8). Furthermore, (4.9) is equivalent to (4.10) whenever \( \sup_{p \in \mathcal{U}^1(S_0)} \int \log p \, dP_0 > -\infty \) and is attained.

In what follows, we work with the formulation of Patilea (2001) in (4.9), although we continue to refer to it as the KL projection. Before stating our first theorem, we recall that \( \mathcal{U}^1|\kappa(S_0) \) is the space of unimodal pmfs with support \( S_0 \) and mode at either \( s_{\kappa-1} \) or \( s_{\kappa} \).

**Theorem 4.1.** Let \( p_0 \) be a discrete pmf with support \( S_0 \). Let \( \overline{p}_0|\kappa \) denote the greatest convex majorant of the cumulative sum of \( p_0(s_i), i \leq \kappa - 1 \) and the least concave minorant of the cumulative sum of \( p_0(s_i), i \geq \kappa \), and let \( \overline{p}_0|\kappa \) denote the pmf corresponding to \( \overline{p}_0|\kappa \). Then

\[
\int \log \frac{\overline{p}_0|\kappa}{p} \, dP_0 \geq 0, \quad \text{for all } p \in \mathcal{U}^1|\kappa(S_0). \tag{4.11}
\]
Furthermore, when \( p_0 \in U^1(S_0) \), or when \( \sum_{j \neq 0} \log |j| p_0(s_j) < \infty \), \( q = \widehat{p}_0 \) is the unique pmf which satisfies \( \int \log(q/p) dP_0 \geq 0 \) for all \( p \in U^1(S_0) \).

We next consider the larger class \( U^1(S_0) \).

**Theorem 4.2.** Let \( p_0 \) be a discrete pmf with support \( S_0 \).

1. Suppose that \( p_0 \in U^1(S_0) \). Then \( \widehat{p}_0 = p_0 \) is the unique unimodal pmf satisfying

\[
\int \log \frac{\widehat{p}_0}{p} dP_0 \geq 0 \quad \text{for all } p \in U^1(S_0).
\]

2. Suppose that \( p_0 \notin U^1(S_0) \) and \( \sum_{j \neq 0} \log |j| p_0(s_j) < \infty \). Then there exists a \( \widehat{p}_0 \in U^1(S_0) \) such that

\[
\int \log \frac{\widehat{p}_0}{p} dP_0 \geq 0 \quad \text{for all } p \in U^1(S_0).
\]

When \( \widehat{p}_0 \) is not unique, we shall denote by \( \{\widehat{p}_0\} \) the (finite) collection of all such projections.

Theorem 4.2 says that in case the model is well-specified, then the KL projection of \( p_0 \) is unique and equal to the true pmf itself under no additional assumptions. However, if the model is misspecified, there may exist several different KL projections. These are collected in the set \( \{\widehat{p}_0\} \) which is necessarily finite. Examples of such non-uniqueness are given later in Figure 5. We believe that this lack of uniqueness is due to the fact that the space of unimodal densities is not convex. Although the condition \( \sum_{j \neq 0} \log |j| p_0(s_j) < \infty \) may seem a bit unnatural at first, one can express it in a more transparent form thanks to the next proposition.

**Proposition 4.3.** Let \( p_0 \) be a discrete pmf with support \( S_0 \). Then

\[
\sum_{j \neq 0} \log |j| \ p_0(s_j) < \infty \quad \text{if and only if} \quad \sup_{p \in U^1(S_0)} \int \log p \ dP_0 \in (-\infty, 0].
\]

Recall that under this condition, (4.9) is equivalent to (4.10). Furthermore, if we assume in addition that

\[
0 < \delta_1 \leq \inf (s_{j+1} - s_j) \leq \sup (s_{j+1} - s_j) \leq \delta_2 < \infty,
\]

then one can show that the condition \( \sum_{j \neq 0} \log |j| p_0(s_j) < \infty \) is equivalent to \( \int \log |x - a| dP_0(x) \in \mathbb{R} \) for some \( a \notin S_0 \). Therefore, this condition gives a bound on the speed of decay of \( p_0 \). Also, it is weaker than the assumption of having a finite mean required by Cule and Samworth (2010); Dümbgen et al. (2011). Our assumption is also weaker than that made by Patilea (2001, Corollary 5.6), although the latter is a condition in order to derive
rates of convergence. In our setting, Patilea’s assumption boils down to existence of an $\epsilon > 0$ such that
\[
\int \hat{p}_0^\epsilon dP_0 < \infty
\]
where $P_0$ is the cumulative distribution function of $p_0$. By the inequality $\log(x) \leq x^\epsilon/\epsilon$ for $x \in (0, \infty)$, we find
\[
\int \log(1/\hat{p}_0) dP_0 \leq \frac{1}{\epsilon} \int \hat{p}_0^\epsilon dP_0 < \infty,
\]
implying our condition in Proposition 4.3 (since then $\int \log \hat{p}_0 dP_0 > -\infty$).

### 4.1 Consistency

For two pmfs $p$ and $q$ defined on $S$, let $\ell_k(p,q)$ and $h(p,q)$ denote the $\ell_k$ and Hellinger distances between $p$ and $q$, respectively. That is,
\[
\ell_k(p,q) = \left\{ \begin{array}{ll}
\left( \sum_{x \in S} |p(x) - q(x)|^k \right)^{1/k}, & \text{if } 1 \leq k < \infty \\
\sup_{x \in S} |p(x) - q(x)|, & \text{if } k = \infty
\end{array} \right.
\]
and
\[
h(p,q) = \frac{1}{2} \sum_{x \in S} \left( \sqrt{p(x)} - \sqrt{q(x)} \right)^2.
\]

In the following, we establish almost sure consistency of the unimodal MLE under a mild condition on the true pmf $p_0$. Let us fix a discrete pmf $p_0$ with support $S_0$, and assume that we observe i.i.d. data $X_1, \ldots, X_n \sim p_0$. Here, we do not necessarily assume that $p_0$ is itself unimodal. Let $\hat{p}_n$ denote again the unimodal MLE based on the sample $(X_1, \ldots, X_n)$. Recall that in the well-specified model, the KL projection $\hat{p}_0$ in the sense of (4.9) is $p_0$ itself.

When the model is misspecified and $p_0$ satisfies $\sum_i \log |p_0(s_i)| < \infty$, then the KL projection $\hat{p}_0$ exists in the sense of (4.10) but may not be unique. In this situation, we denote by $\{\hat{p}_0\}$ the set of all such KL projections.

**Theorem 4.4.** Suppose that $\sum_i \log |p_0(s_i)| < \infty$, and let $d \equiv \ell_k$ or $h$. Then
\[
d(\hat{p}_n, \{\hat{p}_0\}) \equiv \inf_{\hat{p} \in \{\hat{p}_0\}} d(\hat{p}_n, \hat{p}) \to 0
\]
almost surely. If $p_0$ is unimodal, then
\[
d(\hat{p}_n, p_0) \to 0
\]
amost surely.

**Remark 4.5.** Pointwise convergence and convergence in $\ell_k, 1 \leq k \leq \infty$ and Hellinger distance $h$ are all equivalent for probability mass functions. This follows for example from Lemma C.2 in the online supporting material of Balabdaoui et al. (2013).
Figure 5: Convergence of $\hat{p}_n$ to $\{\hat{p}_0\}$ for $p_0$ as in (4.12). The boxplots show the $d = \ell_2$ distance for $B = 1000$ Monte Carlo samples with a sample size of $n = 1000000$. The three columns give (a) $d(\hat{p}_n, \hat{p}_{01})$, (b) $d(\hat{p}_n, \hat{p}_{02})$, and (c) $d(\hat{p}_n, \{\hat{p}_0\})$. The plot on the right differs from the plot on the left in that, on the right, in (a) and (b) the boxplots have been split into large/small values to show the bimodal nature of the data. For reference, the dashed horizontal line gives $d(\hat{p}_{01}, \hat{p}_{02})$.

The fact that $\{\hat{p}_0\}$ is not necessarily a singleton means that the MLE does not necessarily converge to a particular element of $\{\hat{p}_0\}$. Rather, our proof shows instead that the MLE is sequentially compact: there exists an element $\hat{q} \in \{\hat{p}_0\}$ and a subsequence $n_k$ such that $d(\hat{p}_{n_k}, \hat{q}) \to 0$. We illustrate this behaviour via the following example. Let $S_0 = \{-2, -1, 0, 1, 2\}$ and define

$$p_0(s_i) = \begin{cases} 
1/6 & s_i = -2, 0, 2 \\
1/4 & s_i = -1, 1. 
\end{cases} \tag{4.12}$$

In this case, $\{\hat{p}_0\}$ has two elements, which we denote by $\hat{p}_{01}^0$ and $\hat{p}_{02}^0$. Straightforward calculations show that

$$\hat{p}_{01}^0(s_i) = \begin{cases} 
1/6 & s_i = -2, 2 \\
1/4 & s_i = -1, \\
5/24 & s_i = 0, 1 
\end{cases}$$

with mode at $-1$ and $\hat{p}_{02}^0(s_i) = \hat{p}_{01}^0(-s_i)$ (with mode at 1). Simulations for a very large sample size are shown in Figure 5, where the convergence in set distance is clearly visible.

On the other hand, if $|\{\hat{p}_0\}| = 1$, then the unimodal MLE converges to the unique element of $\{\hat{p}_0\}$. We also note that if we consider the restricted MLE $\hat{p}_n|\kappa$, then a similar result to the above holds. A proof may be provided using, for example, Marshall’s lemma as in Patilea (2001, Lemma 5.5, page 114), without any restrictions on $p_0$. 

16
Let $\overline{M}_n$ be the modal region of the unimodal MLE $\widehat{p}_n$, cf. (2.2). An immediate corollary of the preceding theorem is the following statement about convergence of $\overline{M}_n$. For simplicity, we now assume that the KL projection is unique. However, one may state the following results with some additional generality, albeit in a less clear manner.

**Corollary 4.6.** Assume $\sum_{i \neq 0} \log |i| p_0(s_i) < \infty$ and that $|\{\overline{p}_0\}| = 1$, and let $\mathcal{M}$ denote the modal region of $p_0$. Then with probability one, there exists a sufficiently large $n_0$, such that for all $n \geq n_0$, $\overline{M}_n \subset \mathcal{M}$.

If $|\mathcal{M}| = 1$, then Corollary 4.6 implies that, with probability one, there exists a sufficiently large $n_0$, such that the mode of the MLE coincides with the true mode. In the case $|\mathcal{M}| > 1$, then this is no longer true, and all we can say is that eventually the estimated mode will be in $\mathcal{M}$. Note that the latter has nothing to do with the fact that we make the convention of taking the smallest mode to define the MLE; any such convention (or no convention) would result in the same behavior.

From Theorem 4.4 we immediately obtain the following result.

**Corollary 4.7.** Assume $\sum_{i \neq 0} \log |i| p_0(s_i) < \infty$ and that $|\{\overline{p}_0\}| = 1$. Let $\overline{F}_n$ and $\overline{F}_0$ denote the cdfs of $\widehat{p}_n$ and $\overline{p}_0$ respectively. Then

$$\lim_{n \to \infty} \sup_{s \in S_0} |\overline{F}_n(s) - \overline{F}_0(s)| = 0$$

almost surely.

## 5 Global asymptotics

The asymptotic behaviour of the unimodal MLE, as well as the proof thereof, share many similarities with those given in Jankowski and Wellner (2009) for the Grenander estimator of decreasing pmf on $\mathbb{N}$. Our main interest here is to derive the weak limit of the estimator when $p_0$ is unimodal, and therefore we do not consider the misspecified setting. One could, however, mimic the work in Jankowski (2014) to obtain the asymptotic distributions in this case under some further restrictions on $p_0$. Despite the similarity mentioned earlier with the monotone problem, some technical details need special attention due to the fact that (1) the mode of the true pmf is unknown, and (2) we do not assume that the true support is known.

To describe the asymptotic theory, we first need to define an operator, denoted here as $\varphi$. Recall that $\mathcal{M}$ denotes the modal region of $p_0$ as defined in (2.2). Let us write

$$D = \left\{ s_i : s_i \notin \mathcal{M} \text{ and } p_0(s_i) \geq p_0(s_{i+1}) \right\}, \quad \text{and}$$

$$I = \left\{ s_i : s_i \notin \mathcal{M} \text{ and } p_0(s_{i-1}) \leq p_0(s_i) \right\}$$
as the decreasing and increasing regions of $S_0$ respectively. We will write $\mathcal{M} = \{\tau_0^I, \ldots, \tau_0^D\}$ (where $\tau_0^I \leq \tau_0^D$), and let $\{\tau_i^D\}_{i \geq 1}$ enumerate the points in $D$ such that $p_0(s_i) > p_0(s_{i+1})$, where $\tau_i^D < \tau_{i+1}^D$. Similarly, let $\{\tau_i^I\}_{i \geq 1}$ enumerate the points in $I$ such that $p_0(s_{i-1}) < p_0(s_i)$, where $\tau_{i+1}^I < \tau_i^I$. We will write $D_j = \{s \in S_0, \tau_{j-1}^D < s \leq \tau_j^D\}$ for $j \geq 1$, and $I_j = \{s \in S_0, \tau_j^I \leq s < \tau_{j-1}^I\}$ for $j \geq 1$. Notice that each of these regions is necessarily finite, and that $p_0$ is constant on each subset $I_j, D_j$ and $\mathcal{M}$. We therefore have that

$$I = \cup I_j, \quad D = \cup D_j, \quad \text{and} \quad S_0 = I \cup \mathcal{M} \cup D.$$  \hfill (5.13)

We also denote the collection of knots as

$$\mathcal{T} = \{\tau_j^I, j \geq 1\} \cup \{\tau_0^D, \tau_0^I\} \cup \{\tau_j^D, j \geq 1\}.$$  \hfill (5.14)

Note that our definition of a knot, as well as the collection of knots, depends on the underlying pmf $p_0$. Finally, let $q$ be an element of $\ell_2(S_0)$, and for a subset $C \subset S_0$ we write the vector $q_C = \{q(s_j), s_j \in C\}$ to denote the sequence $q$ restricted to $C$. We may now define $\varphi$:

$$\varphi[q](s) = \begin{cases} \text{iso}[q_{I_j}](s) & s \in I_j, \\ \text{uni}[q_{\mathcal{M}}](s) & s \in \mathcal{M}, \\ \text{anti}[q_{D_j}](s) & s \in D_j. \end{cases} \hfill (5.15)$$

Note that the definition of $\varphi$ technically depends on $p_0$, although we omit this dependence in addition. In addition, $\varphi$ satisfies $\varphi[p_0] = p_0$.

**Theorem 5.1.** Suppose that $p_0$ is unimodal and that $\sum_{i=0}^{\infty} \log |i| p_0(s_i) < \infty$. Let $\mathbb{W}$ denote the discrete white noise process: That is, $\mathbb{W}$ is the mean zero Gaussian process defined on $S_0$ such that $\text{cov}(\mathbb{W}(s_i), \mathbb{W}(s_j)) = p_0(s_i) \delta_{i,j} - p_0(s_i) p_0(s_j)$. Then

$$\sqrt{n}(\hat{p}_n - p_0) \Rightarrow \varphi[\mathbb{W}],$$

in $\ell_k(S_0)$, where $2 \leq k \leq \infty$.

An immediate corollary of our result is that if $s$ is such that $s \in C$ where $C = I_j, \mathcal{M}$, or $D_j$ and $|C| = 1$, then $\sqrt{n}(\hat{p}_n(s) - p_0(s)) \Rightarrow \mathbb{W}(s)$, since in such cases $\varphi[q](s) = q(s)$. Namely, this says that in regions where $p_0$ is strictly unimodal, the asymptotics of $\hat{p}_n$ are the same as those of $\hat{p}_n$. Similar observations have been made in Jankowski and Wellner (2009) for the Grenander estimator and Balabdaoui et al. (2013) for the log-concave MLE. In addition, we note that $\ell_2(S_0)$ is the smallest space, of those considered above, where one can prove the asymptotics. In other words, convergence in a smaller space such as $\ell_1(S_0)$ cannot be considered without additional assumptions on $p_0$. We refer to Jankowski and Wellner (2009) for additional details.

The next result follow immediately from the definition of the operator $\varphi$ as well as Jankowski and Wellner (2009, Theorem 2.1).
Figure 6: 95\% constant-width ($\beta = 0$) confidence bands for the true pmf when sampling from the double logarithmic distribution $p = 0.9$. The sample size is $n = 100$ on the top and $n = 1000$ on the bottom.

**Proposition 5.2.** For $2 \leq k \leq \infty$, we have that $||\varphi[\mathbb{W}]||_k \leq ||\mathbb{W}||_k$.

In addition, it is also possible to develop a Marshall’s lemma type result in our setting. The (asymptotically negligible) error term not seen in the usual type of result here is due to estimation of the support in our approach.

**Proposition 5.3** (Marshall’s Lemma). Suppose that $\sum_{s \in S_0} p_0^{1/2}(s) < \infty$ and that the true pmf $p_0$ is unimodal with associated cumulative distribution function $F_0$. Then, with probability one, there exists an $n_0$ such that for all $n \geq n_0$

$$\sup_{s \in S_0} |\widehat{F}_n(s) - F_0(s)| \leq \sup_{s \in S_0} |F_n(s) - F_0(s)| + o_P(n^{-1/2}).$$

6 Global confidence bands for $p_0$

The key application of the previous section is the calculation of confidence bands for the true pmf $p_0$, which we assume to be unimodal. To this end, let $q_{0,\alpha}$ be such that $P(||\mathbb{W}||_\infty > q_{0,\alpha}) = \alpha$. Then, it follows that

$$\lim_{n} P \left( \sqrt{n}||\widehat{p}_n - p_0||_\infty \leq q_{0,\alpha} \right) \geq 1 - \alpha.$$
This follows since
\[ \sqrt{n}|\hat{p}_n - p_0|_\infty \Rightarrow \|\varphi[\hat{W}]\|_\infty \leq \|W\|_\infty. \]

It is important to note that if \( p_0 \) is strictly monotone then \( \varphi[\hat{W}] = W \), and then the last inequality above becomes an equality, resulting in an asymptotically exact confidence band.

In order to estimate \( q_{0,\alpha} \), we use \( \hat{p}_n \) in place of \( p_0 \). In Proposition B.7 of the online supplementary material, we show that this yields an almost surely consistent method of estimating \( q_{0,\alpha} \). Also, we estimate each quantile using Monte Carlo simulations. Thus, let \( \tilde{q}_{0,\alpha} \) denote the Monte Carlo estimate of the quantile of \( \|\hat{W}\|_\infty \).

It follows that an asymptotically correct conservative confidence band is given by
\[
\left\{ \left( \hat{p}_n(s_i) - \tilde{q}_{0,\alpha} \sqrt{n} \right) \vee 0, \hat{p}_n(s_i) + \tilde{q}_{0,\alpha} \sqrt{n} \right\}, s_i \in \text{supp}(\hat{p}_n) \}
\]
where \( \text{supp}(\hat{p}_n) \) denotes the support of \( \hat{p}_n \). When the support of \( p_0 \) is estimated from the data, then \( \text{supp}(\hat{p}_n) = S_n \), the support of the empirical distribution.

In Figure 6, we show an example of confidence bands thus constructed, when the true pmf is the double logarithmic distribution with \( p = 0.9 \). We found the constant width of the confidence bands, particularly for the smaller sample size, somewhat visually jarring. For this reason, we also create confidence bands which are visually more appealing in that they do not have uniform width. Define, for \( \beta \geq 0 \),
\[
\tilde{W}_n^\beta(s) = \begin{cases} \frac{\sqrt{n}(\hat{p}_n - p_0)(s)}{\hat{p}_n(s)}, & s \in \text{supp}(\hat{p}_n) \\ 0, & s \notin \text{supp}(\hat{p}_n). \end{cases}
\]
If \( \beta = 0 \), then \( \tilde{W}_n^0 = \sqrt{n}(\hat{p}_n - p_0) \), and we are in the situation of constant-width confidence bands discussed above.

**Proposition 6.1.** Fix \( \beta > 0 \) and assume that the support of \( p_0 \) is finite. Then
\[
\|\tilde{W}_n^\beta\|_\infty \Rightarrow \left\| \frac{\varphi[\tilde{W}]}{P_0^\beta} \right\|_\infty \leq \left\| \frac{\tilde{W}}{P_0} \right\|_\infty.
\]

In this case, an asymptotically correct conservative confidence band is given by
\[
\left\{ \left( \hat{p}_n(s) - \tilde{q}_{0,\alpha}^\beta \sqrt{n} \right) \vee 0, \hat{p}_n(s) + \tilde{q}_{0,\alpha}^\beta \sqrt{n} \right\}, s \in \text{supp}(\hat{p}_n),
\]
where \( \tilde{q}_{0,\alpha}^\beta \) is an estimate of \( q_{\beta,\alpha} \) where \( P\left( \|p_0 - \tilde{W}\|_\infty > q_{\beta,\alpha} \right) = \alpha \). Estimation of this quantile can be done using a Monte Carlo approach, as before.
Figure 7: 95% confidence bands for the true pmf when sampling from the mixture of uniforms distribution with mixing distribution given in (3.7). The sample size is $n = 100$ and we chose $\beta = 0$ (left, constant width) and $\beta = 0.5$ (right, varying width).

Remark 6.2. When $p_0$ has infinite support, the limiting distribution $p_0^{-\beta} \mathcal{W}$ exists in $\ell_2$ provided that $\sum p_0^{1-2\beta} < \infty$, which adds the restriction that $\beta < 1/2$. We conjecture that the above result continues to hold for distributions with infinite support with the restriction that $\beta \in [0,1/2)$, although we do not pursue the proof here. We do note that the assumption of finite support may be highly plausible in certain practical situations, whereas the (weaker) assumption of $\sum p_0^{1-2\beta} < \infty$, may not be as easy to motivate.

In Figure 7 we compare the constant-width confidence band to the varying width confidence band (with $\beta = 0.5$) when the true distribution is the mixture of uniforms, whose mixing distribution is given in (3.7). Visually, we find the choice of $\beta = 0.5$ preferable in that the values, where $\hat{p}_n$ is smaller, express slightly more accuracy, as one would expect. In this particular example, the difference is not great, but is still eye-pleasing. For $\beta = 0$, the width of the confidence bands varies from 0.13 to 0.07 (median 0.08), while for $\beta = 0.5$, the width of the confidence bands varies from 0.17 to 0.04 (median 0.06). Note that, although for $\beta = 0$ the confidence bands have constant width, we have to cut off the lower bound at a maximum value of zero, and hence the bands end up being non-constant in reality. Without this cutoff, the width would be constant at 0.13.

In Table 2, we examine the empirical performance of the proposed confidence bands. We consider two different unimodal distributions: the mixture of uniforms as above, and the double logarithmic with $p = 0.9$ from (3.6). Our simulations span various samples sizes and values of $\beta$. Note that when $\beta = 0.5$, and the true pmf is double logarithmic, the conditions for convergence are violated (see Proposition 6.1 and Remark 6.2), and we include this example for comparison only (seeing as the condition that $\sum p_0^{1-2\beta} < \infty$ may
be difficult to verify without additional information about \( p_0 \).

Table 2: Empirical coverage probabilities for the proposed confidence bands with \( \alpha = 0.05 \).

<table>
<thead>
<tr>
<th>( \beta )</th>
<th>( n = 100 )</th>
<th>( n = 1000 )</th>
<th>( n = 5000 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>mixture of uniforms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.972</td>
<td>0.963</td>
<td>0.959</td>
</tr>
<tr>
<td>0.25</td>
<td>0.991</td>
<td>0.971</td>
<td>0.970</td>
</tr>
<tr>
<td>0.5</td>
<td>0.959</td>
<td>0.953</td>
<td>0.991</td>
</tr>
<tr>
<td>double logarithmic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.956</td>
<td>0.949</td>
<td>0.949</td>
</tr>
<tr>
<td>0.25</td>
<td>0.970</td>
<td>0.950</td>
<td>0.948</td>
</tr>
<tr>
<td>0.5</td>
<td>0.980</td>
<td>0.989</td>
<td>0.989</td>
</tr>
</tbody>
</table>

Define, for \( \beta \geq 0 \),

\[
\hat{c}_{n,u}(s) = \hat{p}_n(s) + \hat{p}_n^\beta(s) \frac{\bar{q}_{\beta,\alpha}}{\sqrt{n}}, \quad s \in \text{supp}(\hat{p}_n),
\]

\[
\hat{c}_{n,l}(s) = 0 \vee \left( \hat{p}_n(s) - \hat{p}_n^\beta(s) \frac{\bar{q}_{\beta,\alpha}}{\sqrt{n}} \right), \quad s \in \text{supp}(\hat{p}_n),
\]

\[
\hat{c}_{n,u}(s) = \hat{c}_{n,l}(s) = 0, \quad s \notin \text{supp}(\hat{p}_n).
\]

The results in Table 2 give the empirical coverage on the set \( S_n \) as indicated in the third column. That is, we report the proportion of times that

\[
\hat{c}_{n,l}(s) \leq p_0(s) \leq \hat{c}_{n,u}(s), \quad \text{for all } s \in S_n \quad (6.17)
\]

was observed.

Overall, we find that the confidence bands perform rather well. Note that the double logarithmic case \( \beta < 0.5 \) we would expect to obtain asymptotically correct bands, whereas in both uniform mixture scenarios, we expect an asymptotically conservative result. In Appendix A of the online supplementary material, we provide some additional results where we study the cost of defining the bands on \( \text{supp}(\hat{p}_n) = S_n \) in the simulations.

7 Time-to-onset of the Ebola virus

In a recent article, Breman and Johnson (2014) describe their experiences during the 1976 Ebola virus outbreak in Zaire (currently, the Democratic Republic of the Congo). The figure in the article shows histograms of the time of onset of the disease based on the transmission route: patients became infected either with an unsterilized needle or through person-to-person contact. This data was also previously published in Breman et al. (1978). Here, we use the histograms in Breman and Johnson (2014) to transcribe the
Figure 8: Time to onset of symptoms of the Ebola virus based on transmission type. The sample size is $n = 57$ for those infected from unsterilized needles and $n = 108$ for person to person contact.

data and perform a brief analysis. We note that transcribing the histograms resulted in samples sizes of $n = 57$ and $n = 108$, which differs slightly from that presented in Breman and Johnson (2014).

Figure 8 shows the empirical observations and the fitted unimodal MLEs for the two types of transmission routes. The 95% asymptotic global confidence band are also included, where we used the version with constant width; i.e. those corresponding to $\beta = 0$. Note that the time-to-onset is measured in days, and it therefore makes sense to assume that the support of the true pmf is either equal to the natural numbers, or is a connected subset thereof. Thus, we use the version of the likelihood maximization where the support is not estimated from the data. As mentioned earlier, our results apply also
to this (easier) case. Visually, there is no glaring reason that the assumption of unimodality is not appropriate in these two cases. On the other hand, the fitted MLE provides a slight smoothing to the empirical distribution, which is appealing.

We note also that the confidence bands in Figure 8 appear somewhat wide at first glance. However, this is due to the smaller sample sizes observed in both distributions. The average width for the injection infection was found to be 0.18, and 0.12 for infection from person-to-person contact. As a crude benchmark, the average widths of 95% pointwise confidence intervals were calculated for the true pmf

\[ p_n \pm 1.96 \sqrt{p_n(1-p_n)}, \]

based on Theorem 5.1 and under the (untested) assumption that the true pmf is strictly unimodal. Here, the average width for the injection infection was found to be 0.12, and 0.08 for infection from person-to-person contact. These are also rather wide, but smaller than the global confidence bands, as expected.

It is quite interesting how different the two distributions appear to be. The standard Kolmogorov-Smirnov test does not yield exact \( p \)-values in this setting because the data is discretized, and hence we used a permutation test (Jöckel, 1986). This modified approach yielded a \( p \)-value of 0.0014 for the hypothesis that the two distributions are the same (incorrectly applying the regular Kolmogorov-Smirnov test also rejected the null hypothesis). This is in line with what we observe in Figures 8 and 9. R (R Core Team, 2014) code for performing this analysis is available online at www.math.yorku.ca/~hkj/Research.

A biological explanation for the difference between the two distributions was provided to us by Jane Heffernan (2014, private communication): “Injection gets the pathogen into the blood stream. Person-to-person contact provides exposure to the mucosa (innate immunity) first, so the pathogens that ultimately make it to the blood will be different in fitness distribution than the injection method. Also, the amount of pathogen ultimately making it to the blood could be smaller compared to the injection method. Both of these variables will affect the incubation period.” In the data, we see this difference not only through a mean comparison (the mean time-to-onset is 6.3 days for transmission via injection and 9.4 days for person-to-person infections) but also in the stochastic dominance observed via the fitted and empirical CDFs in Figure 9. The latter suggest that \( T_{inj} \leq T_{ptp} \) stochastically, where \( T_{inj} \) and \( T_{ptp} \) denote the times to onset for injection and person-to-person infections, respectively. Repeating the permutation test for the hypothesis that the two distributions are equal against the alternative that \( F_{inj} > F_{ptp} \) yields a \( p \)-value of 0.0008.
Figure 9: Time to onset of symptoms of the Ebola virus based on transmission type: a comparison of cumulative distribution functions

8 Supplementary Material

In this supplement we present some additional proofs, and discuss further the assumptions of the main manuscript.

Acknowledgments. The authors would like to thank Hélène Massam and Jon Wellner for their help with the proof of Lemma B.5 and Lemma B.6 of the online supplementary material. We also thank the anonymous referees for useful comments which helped improve the paper.

References


