AUSTRIAN JOURNAL OF STATISTICS
Volume 32 (2003), Number 1&2, 85–98

Estimating the Structural Distribution Function of Cell Probabilities

Bert van Es, Chris A.J. Klaassen
University of Amsterdam
and
Robert M. Mnatsakanov
West Virginia University, Morgantown
A. Razmadze Mathematical Institute, Tbilisi

Abstract: We consider estimation of the structural distribution function of the cell probabilities of a multinomial sample in situations where the number of cells is large. We review the performance of the natural estimator, an estimator based on grouping of the cells, and a kernel type estimator. Inconsistency of the natural estimator and weak consistency of the other two estimators is derived by Poissonization and other, new, technical devices.

Keywords: Multinomial Distribution, Poissonization, Kernel Smoothing, Cell Probabilities, Parent Density.

1 The Structural Distribution Function

Let the vector $X = (X_1, \ldots, X_M)$ denote a mult(n, p_M) distributed random vector, where $p_M = (p_{M1}, p_{M2}, \ldots, p_{MM})$ is the vector of cell probabilities. Hence, the components of p_M are nonnegative and satisfy $p_{M1} + \ldots + p_{MM} = 1$.

We will consider situations where $M = M_n$ is large with respect to n, i.e.

$$M/n \not\to 0, \text{ as } n \to \infty. \tag{1}$$

In these cases X/n does not estimate p_M accurately. For instance, for the average mean squared error in estimating $Mp_{Mi}, i = 1, \ldots, M$, we have

$$\frac{1}{M}\sum_{i=1}^{M} \mathrm{E}\left(M\,\frac{X_i}{n} - Mp_{Mi}\right)^2 = \frac{M}{n}\sum_{i=1}^{M} p_{Mi}(1-p_{Mi}) = \frac{M}{n}\Big(1 - \sum_{i=1}^{n} p_{Mi}^2\Big) \not\to 0,$$

unless $\sum_{i=1}^{M} p_{Mi}^2 \to 1$ holds, i.e. unless the vector p_M comes close to a unit vector $(0, \ldots, 0, 1, 0, \ldots, 0)$.

However, there are characteristics of p_M that can be estimated consistently. Here we will study the *structural distribution function* of p_M. This is a special case of the more general concept of a structural function introduced in Khmaladze (1988) and Khmaladze and Chitashvili (1989). The structural distribution function is defined as the empirical distribution function of the $Mp_{Mi},\ i = 1, \ldots M$, and it is given by

$$F_M(x) = \tfrac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{[Mp_{Mi} \le x]},\ x \in \mathbb{R}.$$

Our basic assumption will be that F_M converges weakly to a limit distribution function F, i.e.

$$F_M \xrightarrow{w} F, \text{ as } n \to \infty. \tag{2}$$

The basic estimation problem is how to estimate F_M (or F) from an observation of X.

A rule of thumb in statistics is to replace unknown probabilities by sample fractions. This yields the so called *natural estimator*. This estimator, denoted by $\hat{F}_M$, is equal to the empirical distribution function based on M times the cell fractions X_i/n, so

$$\hat{F}_M(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}_{[\frac{M}{n} X_i \leq x]}, \ x \in \mathbb{R}.$$

This estimator has often been used in linguistics, but turns out to be inconsistent for estimating F; see Section 5.1, Khmaladze (1988), Khmaladze and Chitashvili (1989) and Klaassen and Mnatsakanov (2000).

Our estimation problem is related to estimation in sparse multinomial tables. For recent results on the estimation of cell probabilities in this context see Aerts et al. (2000).

In Section 2 we present a small simulation study of a typical multinomial sample and the behavior of the natural estimator. It turns out that smoothing is required to obtain weakly consistent estimators. An estimator based on grouping and an estimator based on kernel smoothing are presented in Section 3. Section 4 deals with the technique of Poissonization and with the relation between weak and L_1 consistency. These basic results are used in the weak consistency proofs in Section 5. Section 6 contains a discussion.

2 A Simulation

We have simulated a sample with $M = 1000$ and $n = 2000$. The cell probabilities are generated via

$$p_{Mi} = G(i/M) - G((i-1)/M), i = 1, \ldots, M.$$

The distribution function G and its density g have been chosen equal to the functions

$$g(x) = 30x^2(1-x)^2 \quad \text{and} \quad G(x) = 10x^3 - 15x^4 + 6x^5, 0 \leq x \leq 1.$$

In Section 3 we show that for these cell probabilities, the limit structural distribution function F from (2) is equal to the distribution function of $g(U)$. Here it is given by

$$F(x) = 1 - \sqrt{1 - \sqrt{\tfrac{8}{15}x}}, \quad 0 \leq x \leq \tfrac{15}{8}.$$

These functions are drawn in Figure 1.

For this simulated sample we have plotted the cell counts, multiplied by M/n, and the natural estimate in Figure 2. Comparison with the real F in Figure 1 clearly illustrates the inconsistency of the natural estimator.

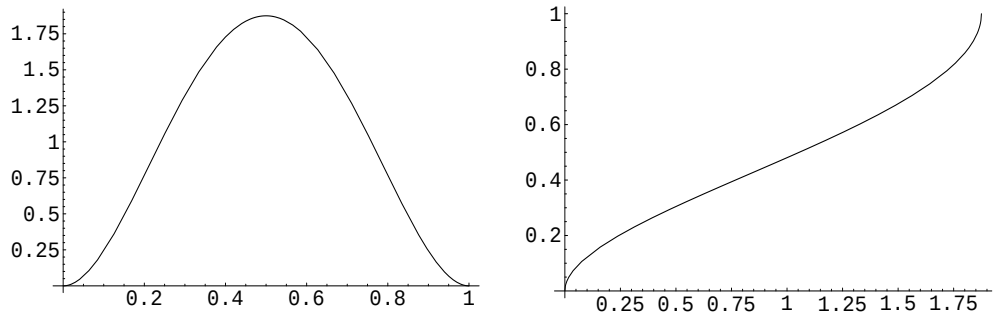

Figure 1: The function g and the corresponding structural distribution function F.

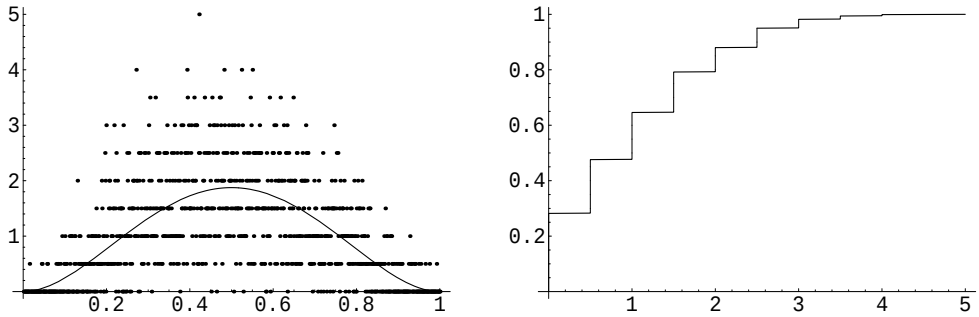

Figure 2: The function g, M/n times the cell counts, and in the second figure the natural estimator of F.

3 Estimators Based on Smoothing Techniques

Up to now we have only assumed that the structural distribution function F_M converges weakly to a limit distribution function F.

From now on we will assume more structure.

Consider the function

$$g_M(u) = \sum_{i=1}^{M} M p_{Mi} \mathbf{1}_{(\frac{i-1}{M}, \frac{i}{M}]}(u),\ u \in \mathbb{R}.$$

This step function is a density representing the cell probabilities and we shall call it the *parent density*. The relation between this parent density g_M and the structural distribution function F_M is given by the fact that if U is a uniform(0,1) random variable then F_M is the distribution function of $g_M(U)$. Note that

$$\mathrm{E}\, g_M(U) = \int_{-\infty}^{\infty} g_M(u)du = \sum_{i=1}^{M} p_{Mi} = 1,$$

so g_M is a probability density indeed.

We will assume that there exists a limiting parent density g on [0,1] such that, as $n \to \infty$,

$$\sup_{0<u\leq 1} |g_M(u) - g(u)| \to 0. \tag{3}$$

Consequently we have $g_M(U) \to g(U)$, almost surely, and hence $F_M \xrightarrow{w} F$.

The inconsistency of the natural estimator can be lifted by first smoothing the cell counts X_i. We consider two smoothing methods, grouping, which is actually some kind of histogram smoothing, and a method based on kernel smoothing of the counts.

3.1 Grouping

Let $m, k_j, j = 0, 1, \ldots, m$, be integers, all depending on n, such that $0 = k_0 < k_1 < \ldots < k_m = M$. Define the group frequencies $\bar{X}_j$ as

$$\bar{X}_j = \sum_{i=k_{j-1}+1}^{k_j} X_i, \quad j = 1, \ldots, m.$$

Then the vector of grouped counts $\bar{X}$ is again multinomially distributed,

$$\bar{X} = (\bar{X}_1, \ldots, \bar{X}_m) \sim mult(n, q_m)$$

with $q_m = (q_{m1}, \ldots, q_{mm})$ and

$$q_{mj} = \sum_{i=k_{j-1}+1}^{k_j} p_{Mi}, \quad j = 1, \ldots, m.$$

The grouped cells estimator, introduced in Klaassen and Mnatsakanov (2000), is defined by

$$\hat{F}_M(x) = \frac{1}{M} \sum_{j=1}^{m} (k_j - k_{j-1}) \mathbf{1}_{[\frac{M}{n(k_j - k_{j-1})} \bar{X}_j \leq x]}, \; x \in \mathbb{R}.$$

This estimator may be viewed as a structural distribution function with parent density

$$\hat{g}_M(u) = \sum_{i=1}^{m} \frac{M}{n(k_i - k_{i-1})} \bar{X}_i \mathbf{1}_{[\frac{k_{i-1}}{M} < u \leq \frac{k_i}{M}]}, \; u \in \mathbb{R}.$$

This histogram is an estimator of the limiting parent density g in (3). We will prove weak consistency of the corresponding estimator $\hat{F}_M$ in Section 5.2.

For our simulated example the estimates of g and F resulting from grouping with equal group size $k = 50$ are given in Figure 3.

3.2 A Kernel Type Estimator

Now that we have seen that the estimator based on the grouped cells counts is in fact based on a histogram estimate of the parent density g we might also use kernel smoothing

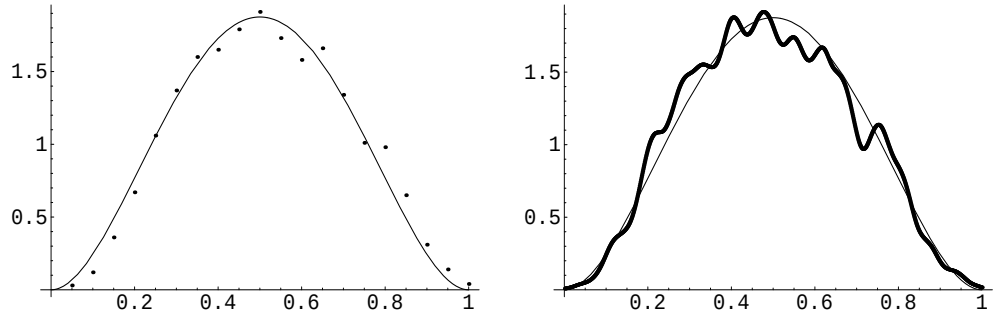

Figure 3: g, F, and estimates $\hat{g}_M$ and $\hat{F}_M$ by grouping with equal cell size.

to estimate g and proceed in a similar manner. If we choose a probability density w as *kernel function* and a *bandwidth* $k \geq 0$, we get the following estimator for the parent density g

$$\hat{g}_M(u) = \frac{M}{nk} \sum_{i=1}^{M} w\Big(\frac{\lceil Mu \rceil - i}{k}\Big) X_i, \ u \in \mathbb{R}.$$

As an estimator for the structural distribution function of the function F we take the empirical distribution function of $\hat{g}_M(U)$ with U uniform, namely

$$\hat{F}_M(x) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}_{[\frac{M}{nk} \sum_{i=1}^{M} w(\frac{j-i}{k}) X_i \leq x]}, \ x \in \mathbb{R}.$$

Weak consistency of this estimator will be derived in Section 5.3.

For our simulated example, kernel estimates $\hat{g}_M$ and $\hat{F}_M$ of g and F respectively, with k equal to 50 are given in Figure 4.

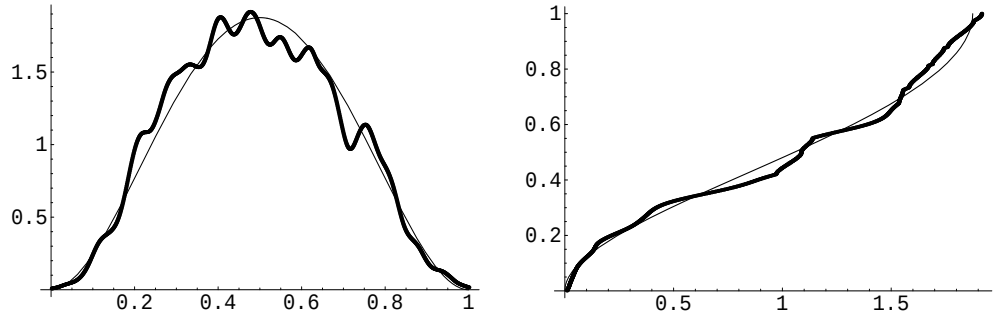

Figure 4: g, F, and estimates $\hat{g}_M$ and $\hat{F}_M$ by kernel smoothing.

4 Relevant Techniques

In our proofs we shall use repeatedly the powerful method of Poissonization and a device involving L_1 convergence.

4.1 Poissonization

Consider the random vectors X and Y, with

$$X = (X_1, \ldots, X_M) \sim \text{mult}(n, p_M) \tag{4}$$

and

$$Y = (Y_1, \ldots, Y_M),\ Y_i \sim \text{Poisson}(np_{Mi}), \tag{5}$$

where $Y_1, \ldots, Y_M$ are independent. Note

$$N = \sum_{i=1}^{M} Y_i \sim \text{Poisson}(n).$$

Given $N = k$ the random vector Y has a mult(k, p_M) distribution.

Based on an infinite sequence of $mult(1, p_{M1}, \ldots, p_{MM})$ random vectors one can construct vectors X and Y, the cell counts over n and N of these vectors respectively, with the distributions (4) and (5). Given $N = k$ they are coupled as follows

$$\begin{aligned} &k \leq n : X = Y + \text{mult}(n - k, p_M), \\ &k > n : Y = X + \text{mult}(k - n, p_M). \end{aligned} \tag{6}$$

Note that this shows that either $X_i \leq Y_i$ for all i or $X_i \geq Y_i$ for all i.

4.2 Convergence in L_1 and Weak Convergence

The focus of this paper is on consistency of estimators of a distribution function. This concept will be defined in Section 5 in terms of weak convergence in probability, which we will define first.

Definition 4.1 *Let F be a distribution function and let F_n be a possibly random distribution function. We say that F_n converges weakly to F in probability,*

$$F_n \xrightarrow{w} F, \quad \textit{in probability},$$

if for all $\epsilon > 0$ and all continuity points x_0 of F

$$P(|F_n(x_0) - F(x_0)| > \epsilon) \to 0,$$

holds.

An important step in the (in)consistency proofs is to show that "Poissonization is allowed", i.e. that we can transfer the limit result for the estimator based on the Poissonized sample, the "Poissonized version", to the original estimator. The following proposition is used repeatedly, also if no Poissonized version is involved.

Definition 4.2 *Let F be a distribution function and let F_n be a possibly random distribution function. We say that F_n converges to F in probability,*

$$F_n \xrightarrow{w} F, \quad \textit{in probability},$$

if for all $\epsilon > 0$ and all continuity points x_0 of F

$$P(|F_n(x_0) - F(x_0)| > \epsilon) \to 0.$$

Proposition 4.1 *Let F be a distribution function and let $\hat{F}_n$ and $\tilde{F}_n$ be possibly random distribution functions. If*

$$\tilde{F}_n \xrightarrow{w} F, \quad \textit{in probability}, \tag{7}$$

and

$$\int |\hat{F}_n - \tilde{F}_n| \xrightarrow{P} 0 \tag{8}$$

hold, then we have

$$\hat{F}_n \xrightarrow{w} F, \quad \textit{in probability}.$$

In the special case where $\tilde{F}_n$ equals F, the proposition states that L_1 convergence implies weak convergence.

Proof. Note that for all x_0 and all $\delta > 0$ we have

$$\int_{x_0-\delta}^{x_0+\delta} |\hat{F}_n - F| \le \int_{-\infty}^{\infty} |\hat{F}_n - \tilde{F}| + \int_{x_0-\delta}^{x_0+\delta} |\tilde{F}_n - F|. \tag{9}$$

Let x_0 denote an arbitrary continuity point of F and ϵ an arbitrary positive number. Choose $\delta > 0$ such that $F(x_0 + \delta) - F(x_0 - \delta) \le \epsilon$ and such that $x_0 - \delta$ and $x_0 + \delta$ are continuity points of F. Then

$$|\tilde{F}_n(x_0 - \delta) - F(x_0 - \delta)| < \epsilon \text{ and } |\tilde{F}_n(x_0 + \delta) - F(x_0 + \delta)| < \epsilon$$

imply

$$\int_{x_0-\delta}^{x_0+\delta} |\tilde{F}_n - F| < 4\delta\epsilon.$$

Hence, we have

$$\begin{aligned} &P\Big(\int_{x_0-\delta}^{x_0+\delta} |\tilde{F}_n - F| \ge 4\delta\epsilon\Big) \\ &\le P(|\tilde{F}_n(x_0 - \delta) - F(x_0 - \delta)| \ge \epsilon) + P(|\tilde{F}_n(x_0 + \delta) - F(x_0 + \delta)| \ge \epsilon) \end{aligned}$$

and, by (7),

$$\int_{x_0-\delta}^{x_0+\delta} |\tilde{F}_n - F| \xrightarrow{P} 0.$$

Consequently, by (8) and (9) we get

$$\int_{x_0-\delta}^{x_0+\delta} |\hat{F}_n - F| \xrightarrow{P} 0.$$

Choose $0 < \delta' < \delta$ such that $F(x_0+\delta') \le F(x_0) + \frac{1}{2}\epsilon$ and $F(x_0-\delta') \ge F(x_0) - \frac{1}{2}\epsilon$. Then we see

$$|\hat{F}_n(x_0) - F(x_0)| \ge \epsilon \Rightarrow \int_{x_0-\delta'}^{x_0+\delta'} |\hat{F}_n - F| \ge \tfrac{1}{2}\delta'\epsilon$$

and hence

$$\begin{aligned} P(|\hat{F}_n(x_0) - F(x_0)| \ge \epsilon) &\le P\Big(\int_{x_0-\delta'}^{x_0+\delta'} |\hat{F}_n - F| \ge \tfrac{1}{2}\delta'\epsilon\Big) \\ &\le P\Big(\int_{x_0-\delta}^{x_0+\delta} |\hat{F}_n - F| \ge \tfrac{1}{2}\delta'\epsilon\Big) \to 0. \end{aligned}$$

Since this holds for arbitrary continuity points x_0 and arbitrary $\epsilon > 0$ we have established $\hat{F}_n \xrightarrow{w} F$, in probability. □

5 Consistency

We define consistency as follows.

Definition 5.1 *An estimator $\hat{F}_n$ of a distribution function F is called consistent iff $\hat{F}_n$ converges weakly to F in probability as n tends to infinity, i.e.*

$$\hat{F}_n \xrightarrow{w} F, \quad \textit{in probability}, \ n \to \infty.$$

5.1 The Natural Estimator

The basic trick in dealing with the difference of the natural estimator and its Poissonized version,

$$\tilde{F}_M(x) = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{[M\frac{Y_i}{n} \le x]}, \ x \in \mathbb{R},$$

uses the coupling as in (6) and is given by the following string of inequalities

$$\begin{aligned} |\hat{F}_M(x) - \tilde{F}_M(x)| &\le \frac{1}{M}\sum_{i=1}^{M} |\mathbf{1}_{[M\frac{X_i}{n} \le x]} - \mathbf{1}_{[M\frac{Y_i}{n} \le x]}| \\ &\le \tfrac{1}{M}\textstyle\sum_{i=1}^{M} \mathbf{1}_{[X_i \ne Y_i]} \le \frac{|N-n|}{M} = O_P\Big(\frac{\sqrt{n}}{M}\Big). \end{aligned} \tag{10}$$

By (1) the right hand side converges to zero in probability and this shows that Poissonization is allowed.

Because of the independence of the Poisson counts Y_i we can easily bound the variance of the Poissonized estimator. We get

$$\text{Var}\,\tilde{F}_M(x) = \text{Var}\,\frac{1}{M}\sum_{i=1}^{M}\mathbf{1}_{[M\frac{Y_i}{n}\le x]} \le \frac{1}{4M} \to 0.$$

We also have

$$\text{E}\,\tilde{F}_M(x) = \frac{1}{M}\sum_{i=1}^{M} P(\frac{M}{n}Y_i \le x) \ne \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}_{[Mp_{Mi}\le x]} = F_M(x)$$

and

$$\begin{aligned}\text{E}\int x^2 d\tilde{F}_M(x) &= \text{E}\,\frac{1}{M}\sum_{i=1}^{M}\Big(\frac{M}{n}Y_i\Big)^2 \\ &= \tfrac{1}{M}\textstyle\sum_{i=1}^{M}\Big(\frac{M}{n}\Big)^2\{np_{Mi} + (np_{Mi})^2\} = \frac{M}{n} + \int x^2 d\,F_M(x).\end{aligned}$$

Together with (1) this gives two reasons why $\tilde{F}_M(x)$ is probably not a consistent estimator of F. Then, by (10) the natural estimator has to be inconsistent too.

The inconsistency of the structural distribution function has previously been established in Khmaladze (1988), Khmaladze and Chitashvili (1989), Klaassen and Mnatsakanov (2000) and van Es and Kolios (2002). In these papers the situation of a *large number of rare events* is considered, i.e. $n/M \to \lambda$ for some constant λ. The explicit limit in probability of $\hat{F}_M(x)$ turns out to be a Poisson mixture of F then.

5.2 Grouping

Under the additional assumption $n/M \to \lambda$, for some constant λ, weak consistency of the estimator based on grouped cells has been proved, without using Poissonization, by Klaassen and Mnatsakanov (2000) and by the Poissonization method for the simpler case of equal group size, i.e. $k_j = k$, by van Es and Kolios (2002). We shall prove the following generalization without using Poissonization.

Theorem 5.1 *If* $m/n \to 0$,

$$\sup_{1\le j\le m}\frac{k_j - k_{j-1}}{M} \to 0,$$

and

$$\sup_{0<u\le 1}|g_M(u) - g(u)| \to 0$$

are valid for some limiting parent density g *that is continuous on* $[0,1]$, *then*

$$\hat{F}_M \xrightarrow{w} F, \quad \textit{in probability},$$

holds with

$$\hat{F}_M(x) = \frac{1}{M}\sum_{j=1}^{m}(k_j - k_{j-1})\mathbf{1}_{[\frac{M}{n(k_j-k_{j-1})}\sum_{i=k_{j-1}+1}^{k_j}X_i\le x]},\ x\in\mathbb{R}.$$

Proof. The estimator $\hat{F}_M$ behaves asymptotically as

$$\bar{F}_M(x) = \frac{1}{M}\sum_{j=1}^{m}(k_j - k_{j-1})\mathbf{1}_{[\frac{Mq_{mj}}{k_j-k_{j-1}}\le x]}.$$

Indeed, in view of $\int |\mathbf{1}_{[a\le x]} - \mathbf{1}_{[b\le x]}|dx = |b-a|$ we have

$$\begin{aligned}
&\int |\hat{F}_M(x) - \bar{F}_M(x)|dx \\
&\le \int \sum_{j=1}^{m} \frac{k_j - k_{j-1}}{M}\Big|\mathbf{1}_{[\frac{M\bar{X}_j}{n(k_j-k_{j-1})}\le x]} - \mathbf{1}_{[\frac{Mq_{mj}}{k_j-k_{j-1}}\le x]}\Big|dx \\
&= \sum_{j=1}^{m} \frac{k_j - k_{j-1}}{M}\Big|\frac{M\bar{X}_j}{n(k_j-k_{j-1})} - \frac{Mq_{mj}}{k_j-k_{j-1}}\Big|.
\end{aligned}$$

Consequently, we obtain

$$\begin{aligned}
&\mathrm{E}\int |\hat{F}_M(x) - \bar{F}_M(x)|dx \le \frac{m}{n}\frac{1}{m}\sum_{j=1}^{m}\mathrm{E}\,|\bar{X}_j - nq_{mj}| \\
&\le \frac{m}{n}\sqrt{\frac{1}{m}\sum_{j=1}^{m}\mathrm{E}\,(\bar{X}_j - nq_{mj})^2} = \frac{m}{n}\sqrt{\frac{1}{m}\sum_{j=1}^{m} nq_{mj}(1-q_{mj})} \\
&\le \sqrt{\frac{m}{n}} \to 0
\end{aligned}$$

and hence

$$\int |\hat{F}_M(x) - \bar{F}_M(x)|dx \xrightarrow{P} 0.$$

In order to prove $\hat{F}_M \xrightarrow{w} F$ in probability, by Proposition 4.1 it remains to show $\bar{F}_M \xrightarrow{w} F$.

Consider the function

$$\bar{g}_M(u) = \sum_{j=1}^{m}\frac{1}{k_j - k_{j-1}}\sum_{i=k_{j-1}+1}^{k_j} Mp_{Mi}\mathbf{1}_{(\frac{k_{j-1}}{M},\frac{k_j}{M}]}(u),\ u\in\mathbb{R}.$$

For $k_{j-1}/M < u \le k_j/M$ we have

$$\begin{aligned}
&|\bar{g}_M(u) - g(u)| \le \frac{1}{k_j - k_{j-1}}\sum_{i=k_{j-1}+1}^{k_j}|Mp_{Mi} - g(u)| \\
&\le \sup_{k_{j-1}/M<v\le k_j/M}|g_M(v) - g(u)| \\
&\le \sup_v |g_M(v) - g(v)| + \sup_{|u-v|\le \sup_j(k_j-k_{j-1})/M}|g(v) - g(u)|.
\end{aligned}$$

By assumption, the function g is uniformly continuous and hence $\sup_j(k_j - k_{j-1})/M \to 0$ implies $\bar{g}_M(U) \to g(U)$, almost surely, and in distribution, i.e. $\bar{F} \xrightarrow{w} F$, which completes the proof of the theorem. □

5.3 The Kernel Type Estimator

Weak consistency of the kernel type estimator is established by the next theorem.

Theorem 5.2 *If $k \to \infty, k/M \to 0, M/(nk) \to 0$ hold, if w is a density that is Riemann integrable on bounded intervals, that is also Riemann square integrable on bounded intervals, and that has bounded support or is ultimately monotone in its tails, and if*

$$\sup_{0<u\le 1} |g_M(u) - g(u)| \to 0 \tag{11}$$

holds with g continuous on $[0,1]$, then

$$\hat{F}_M \xrightarrow{w} F, \quad \textit{in probability},$$

is valid for

$$\hat{F}_M(x) = \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}_{[\frac{M}{nk}\sum_{i=1}^{M} w(\frac{j-i}{k})X_i \le x]}, \ x \in \mathbb{R}.$$

Proof. Let

$$\tilde{F}_M(x) = \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}_{[\frac{M}{nk}\sum_{i=1}^{M} w(\frac{j-i}{k})Y_i \le x]}, \ x \in \mathbb{R},$$

be the Poissonized version of $\hat{F}_M(x)$. Note that by the coupling argument $X_i \ge Y_i$ for all i or $X_i \le Y_i$ for all i. Since w is a Riemann integrable density we thus get

$$\begin{aligned}
&\mathrm{E}\int |\hat{F}_M(x) - \tilde{F}_M(x)|dx \\
&\le \mathrm{E}\,\frac{1}{M}\sum_{j=1}^{M}\int |\mathbf{1}_{[\frac{M}{nk}\sum_{i=1}^{M} w(\frac{j-i}{k})X_i \le x]} - \mathbf{1}_{[\frac{M}{nk}\sum_{i=1}^{M} w(\frac{j-i}{k})Y_i \le x]}|dx \\
&= \mathrm{E}\,\frac{1}{M}\sum_{j=1}^{M}\Big|\frac{M}{nk}\sum_{i=1}^{M} w\Big(\frac{j-i}{k}\Big)(X_i - Y_i)\Big| = \mathrm{E}\,\frac{1}{M}\sum_{j=1}^{M}\frac{M}{nk}\sum_{i=1}^{M} w\Big(\frac{j-i}{k}\Big)|X_i - Y_i| \\
&= \mathrm{E}\,\frac{1}{n}\sum_{i=1}^{M}\Big(\sum_{j=1}^{M}\frac{1}{k}\,w\Big(\frac{j-i}{k}\Big)\Big)|X_i - Y_i| \le \sum_{\ell\in\mathbb{Z}}\frac{1}{k}\,w\Big(\frac{\ell}{k}\Big)\,\mathrm{E}\,\frac{|N-n|}{n} = O\Big(\frac{1}{\sqrt{n}}\Big).
\end{aligned}$$

Consequently, by Proposition 4.1 it suffices to prove

$$\tilde{F}_M \xrightarrow{w} F, \quad \text{in probability}. \tag{12}$$

Define

$$\bar{F}_M(x) = \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}_{[\frac{1}{k}\sum_{i=1}^{M} w(\frac{j-i}{k})Mp_{Mi} \le x]}, \ x \in \mathbb{R}.$$

To prove (12), by Proposition 4.1, it suffices to prove

$$\mathrm{E}\int |\tilde{F}_M(x) - \bar{F}_M(x)|dx \xrightarrow{P} 0 \quad \text{and} \quad \bar{F}_M \xrightarrow{w} F, \text{ in probability}. \tag{13}$$

Indeed, since the Y_i are independent and w is square Riemann integrable, we have

$$\begin{aligned}
&\mathrm{E}\int |\tilde{F}_M(x)-\bar{F}_M(x)|dx \le \frac{1}{M}\sum_{j=1}^{M}\mathrm{E}\Big|\frac{M}{nk}\sum_{i=1}^{M} w\Big(\frac{j-i}{k}\Big)(Y_i-np_{Mi})\Big| \\
&\le \sqrt{\frac{1}{M}\sum_{j=1}^{M}\mathrm{Var}\Big\{\frac{M}{nk}\sum_{i=1}^{M} w\Big(\frac{j-i}{k}\Big)(Y_i-np_{Mi})\Big\}} \\
&= \sqrt{\frac{M}{n^2k^2}\sum_{j=1}^{M}\sum_{i=1}^{M} w^2\Big(\frac{j-i}{k}\Big)np_{Mi}} \le \sqrt{\frac{M}{nk^2}\sum_{i=1}^{M}\sum_{\ell\in\mathbb{Z}} w^2\Big(\frac{\ell}{k}\Big)p_{Mi}} \\
&= \sqrt{\frac{M}{nk}\sum_{\ell\in\mathbb{Z}}\frac{1}{k}\,w^2\Big(\frac{\ell}{k}\Big)} = O\Big(\sqrt{\frac{M}{nk}}\Big)\to 0,
\end{aligned}$$

because of $k\to\infty$ and $M/(nk)\to 0$. This proves the first statement of (13).

Finally, we prove the second statement of (13). As parent density for the distribution function $\bar{F}_M$ we choose

$$\bar{g}_M(u)=\sum_{j=1}^{M}\ \frac{1}{k}\sum_{i=1}^{M} w\Big(\frac{j-i}{k}\Big)Mp_{Mi}\,\mathbf{1}_{(\frac{j-1}{M},\frac{j}{M}]}(u),\ u\in\mathbb{R}.$$

Note that g_M vanishes outside (0,1]. Fix $u\in(0,1)$. For $u\in(\frac{j-1}{M},\frac{j}{M}]$, and $K>0$ fixed, we have

$$\begin{aligned}
&|\bar{g}_M(u)-g(u)| \\
&\le \sum_{\ell\in\mathbb{Z}}\frac{1}{k}\,w\Big(\frac{\ell}{k}\Big)\Big|g_M\Big(\frac{j-\ell}{M}\Big)-g\Big(\frac{j-\ell}{M}\Big)\Big| \\
&+\sum_{|\ell|\le Kk}\frac{1}{k}\,w\Big(\frac{\ell}{k}\Big)\Big|g\Big(\frac{j-\ell}{M}\Big)-g(u)\Big| \qquad (14)\\
&+\Big\{\sum_{|\ell|>Kk}\frac{1}{k}\,w\Big(\frac{\ell}{k}\Big)+\Big|\sum_{\ell\in\mathbb{Z}}\frac{1}{k}\,w\Big(\frac{\ell}{k}\Big)-1\Big|\Big\}\sup_u g(u).
\end{aligned}$$

Note that the conditions imposed on w guarantee that

$$\sum_{|\ell|>Kk}\frac{1}{k}\,w\Big(\frac{\ell}{k}\Big)$$

is arbitrarily small for K sufficiently large, that

$$\sum_{|\ell|\le Kk}\frac{1}{k}\,w\Big(\frac{\ell}{k}\Big)\to\int_{-K}^{K} w(u)du,$$

which is arbitrarily close to one for K large enough, and hence that

$$\sum_{\ell\in\mathbb{Z}}\frac{1}{k}\,w\Big(\frac{\ell}{k}\Big)\to 1,$$

as $k \to \infty$. Consequently, in view of (11), and in view of the uniform continuity and boundedness of g, all three terms at the right hand side tend of (14) to zero as $k \to \infty$ and subsequently $K \to \infty$. So, $\bar{g}_M(U) \to g(U)$, almost surely and in distribution, which implies $\bar{F}_M \xrightarrow{w} F$.

□

6 Discussion

The key assumption in the consistency proofs of the grouping and kernel estimators is the existence of a limiting parent density. This is a reasonable assumption only, if there is a natural ordering of the cells and neighboring cells have approximately the same cell probabilities. In applications like e.g. linguistics this need not be the case. Consider a text of n words of an author with a vocabulary of M words. Here the words in the vocabulary correspond to the cells of the multinomial distribution and the existence of a limiting or approximating parent density is rather unrealistic. To a lesser extent this might be the case in biology, where cells correspond to species and n is the number of individuals found in some ecological entity.

An estimator that is consistent even if our key assumption does not hold, has been constructed in Klaassen and Mnatsakanov (2000). However, it seems to have a logarithmic rate of convergence only. The rates of convergence of our grouping and kernel estimators will depend on the rate at which the assumed limiting parent density can be estimated. This issue is still to be investigated, but under the assumption $n/M \to \lambda$, for some constant λ, van Es and Kolios (2002) show that, for the relatively simple case of equal group sizes, an algebraic rate of convergence can be achieved by the estimator based on grouping.

Since the estimators studied here are based on smoothing of the cell frequencies an important open problem is the choice of the smoothing parameter. For the estimator based on grouping this is the choice of the sizes of the groups and for the kernel type estimator the choice of the bandwidth. By studying convergence rates these choices may be optimized.

Acknowledgement

This paper has been prepared under INTAS-97-Georgia-1828.

References

M. Aerts, I. Augustyns, and P. Janssen. Central limit theorem for the total squared error of local polynomial estimators of cell probabilities. *J. Statist. Plann. Inference*, 91: 181–193, 2000.

E.V. Khmaladze. The statistical analysis of a large number of rare events. Technical report, CWI, Amsterdam, Report MS-R8804, 1988.

E.V. Khmaladze and R.Ya. Chitashvili. Statistical analysis of a large number of rare events and related problems (Russian). *Proc. A. Razmadze Math. Inst. Georgian Acad. Sci., Tbilisi*, 92:196–245, 1989.

C.A.J. Klaassen and R.M. Mnatsakanov. Consistent estimation of the structural distribution function. *Scand. J. Statist.*, 27:733–746, 2000.

B. van Es and S. Kolios. Estimating a structural distribution function by grouping. Technical report, University of Amsterdam, Mathematics ArXiv PR/0203080, 2002.

Corresponding author's address:

Dr. A.J. van Es
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
The Netherlands

Tel. +20 5255365
Fax +20 5255101
E-mail: vanes@science.uva.nl
http://www.turing.science.uva.nl/∼vanes