AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 2&3, 365–370

Discrete Kernels

František Vávra, Pavel Nový, Lucie Reismüllerová, Kateřina Vokáčová, and Martina Neumanová
University of West Bohemia, Pilsen, Czech Republic

Abstract: The submitted paper deals with ”kernel approximations” of discrete probability distributions derived for cardinal variables with finite alphabet of elementary events. The work is concentrated on one type of the kernel:

$$\kappa(x, x_i, a, b) = \frac{a^{e(x_i,x)} b^{1-e(x_i,x)}}{a + b(|X| - 1)},$$

where $|X|$ is finite and $a, b > 0$, $a = a(n)$, $b = \{b(n)e(x, y = 1 \Leftrightarrow x = y,$ $e(x, y) = 0 \Leftrightarrow (x \neq y)\}$. Then the approximation of probability distribution has the form:

$$\widehat{p_n}(x) = \frac{1}{n} \sum_{i=1}^{n} \kappa(x, x_i, a, b),$$

where n is range of sample. Possible options of kernel’s parameters and their effects are analyzed in the text. The asymptotical behavior is studied as well. Further the context among some of classical and non classical frequency probability estimates is presented.

Keywords: Kernel Approximation, Kernel´s Parameters, Frequency Estimate, Smoothness.

1 Introduction and Assumptions

We assume probability distribution defined on finite set X (set of elementary events) and that we have $n > 1$ samples $\{x_1, \dots, x_n\}$. Kernel is considered in a form:

$$\kappa : X \times X \to \langle 0, 1 \rangle, \qquad \kappa(x, x_i) \mapsto \frac{a^{e(x_i,x)} b^{1-e(x_i,x)}}{a + b(|X| - 1)},$$

where $e(x_i,^x) = 1 \Leftrightarrow x_i = x$ and $e(x_i, x) = 0 \Leftrightarrow x_i \neq x$ for $x_i, x \in X$, and $i = 1, \dots, n$ and $a, b > 0$, $a = a(n)$, $b = b(n)$. With these assumptions we obtain

$$\begin{aligned}
\sum_{x \in X} \kappa(x, x_i, a, b) &= \sum_{x \in X} \frac{a^{e(x_i,x)} b^{1-e(x_i,x)}}{a + b(|X| - 1)} \\
&= \frac{1}{a + b(|X| - 1)} \sum_{x \in X} a^{e(x_i,x)} b^{1-e(x_i,x)} \\
&= \frac{1}{a + b(|X| - 1)} \left(a^{e(x_i,x)} + \sum_{x \in X; x \neq x_i} b^{1-e(x_i,x)} \right) \\
&= \frac{1}{a + b(|X| - 1)} (a + b(|X| - 1)) = 1 .
\end{aligned}$$

As probability estimate is understood

$$\widehat{p_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{a^{e(x_i,x)} b^{1-e(x_i,x)}}{a+b(|X|-1)} = \frac{1}{n(a+b(|X|-1))}\sum_{i=1}^{n} a^{e(x_i,x)} b^{1-e(x_i,x)}\,, \quad \text{for } x \in X\,.$$

Let $n(x) = |\{x = x_i, i = 1,\ldots,n\}|$ denote the number of observed values x in the sample $\{x_1,\ldots,x_n\}$ and $p(x)$ the probability of observing the event x. Then we obtain

$$\begin{aligned}
\widehat{p_n}(x) &= \frac{1}{n(a+b(|X|-1))}\sum_{i=1}^{n} a^{e(x_i,x)} b^{1-e(x_i,x)} \\
&= \frac{1}{n(a+b(|X|-1))}\left(an(x) + b(n-n(x))\right) \\
&= \frac{n(x)(a-b)+nb}{n(a+b(|X|-1))} \\
&= \frac{n(x)}{n}\frac{a-b}{a+b(|X|-1)} + \frac{b}{a+b(|X|-1)}
\end{aligned}$$

with

$$\begin{aligned}
\mathrm{E}\{\widehat{p_n}(x)\} &= \frac{\mathrm{E}\{n(x)\}}{n}\frac{a-b}{a+b(|X|-1)} + \frac{b}{a+b(|X|-1)} \\
&= \frac{np(x)}{n}\frac{a-b}{a+b(|X|-1)} + \frac{b}{a+b(|X|-1)} \\
&= p(x)\frac{a-b}{a+b(|X|-1)} + \frac{b}{a+b(|X|-1)} \\
&= p(x)\frac{a-b}{a-b+b|X|} + \frac{b}{a-b+b|X|}\,.
\end{aligned}$$

From the previous expression we see, that if $\lim_{n\to\infty} b(n) = 0$, $\lim_{n\to\infty} a(n) = \alpha > 0$ then $\lim_{n\to\infty} \mathrm{E}\{\widehat{p_n}(x)\} = p(x)$. Because of that the estimate $\widehat{p_n}(x)$ is asymptotically unbiased. Furthermore,

$$\widehat{p_n}(x) - \mathrm{E}\{\widehat{p_n}(x)\} = \frac{a-b}{a+b(|X|-1)}\left(\frac{n(x)}{n} - p(x)\right)\,.$$

With this result the variance of the estimate $\hat{p}_n(x)$ is

$$\begin{aligned}
\sigma^2(\widehat{p_n}(x)) &= \left(\frac{a-b}{n[a+b(|X|-1)]}\right)^2 \mathrm{E}\{(n(x)-np(x))^2\} \\
&= \left(\frac{a-b}{n[a+b(|X|-1)]}\right)^2 np(x)(1-p(x)) \\
&= \frac{1}{n}\left(\frac{a-b}{a+b(|X|-1)}\right)^2 p(x)(1-p(x))\,.
\end{aligned}$$

With these deductions we get

$$\mathrm{E}\{\widehat{p_n}(x)\} = p(x)\frac{a-b}{a+b(|X|-1)} + \frac{b}{a+b(|X|-1)}$$

$$
\begin{aligned}
&= p(x)\frac{a-b}{(a-b+b|X|} + \frac{b}{a-b+b|X|}\,,\\
\sigma^2(\widehat{p_n}(x)) &= \frac{1}{n}\left(\frac{a-b}{a+b(|X|-1)}\right)^2 p(x)(1-p(x))\\
&= \frac{1}{n}\left(\frac{a-b}{a-b+b|X|}\right)^2 p(x)(1-p(x))\,.
\end{aligned}
$$

2 Reparametrization

For $a = b$ we have $\mathrm{E}\{\widehat{p_n}(x)\} = 1/|X|$ and $\sigma^2(\widehat{p_n}(x)) = 0$. This is nothing else than assuming for probability estimate the uniform distribution regardless of the observed data. For this reason we will further assume $a \neq b$. In this case we can use the new parameter $c = b/(a-b)$. Then

$$
\begin{aligned}
\widehat{p_n}(x) &= \frac{n(x)}{n}\frac{1}{1+c|X|} + \frac{c}{1+c|X|}\,,\\
\mathrm{E}\{\widehat{p_n}(x)\} &= p(x)\frac{1}{1+c|X|} + \frac{c}{1+c|X|}\,,\\
\sigma^2(\widehat{p_n}(x)) &= \frac{1}{n}\left(\frac{1}{1+c|X|}\right)^2 p(x)(1-p(x))\,.
\end{aligned}
$$

Of course in accordance with the assumptions mentioned above, this means that $a = a(n)$, $b = b(n) \Rightarrow c = c(n)$. It is clear, that for $\lim_{n\to\infty} c(n) = 0$ we obtain $\mathrm{E}\{\widehat{p_n}(x)\} \overset{n\to\infty}{\to} p(x)$ and $\sigma^2\{\widehat{p_n}(x)\} \overset{n\to\infty}{\to} 0$. Dealing with the mentioned conditions, the estimate $\widehat{p_n}(x)$ is asymptotically unbiased and efficient. Criterion $\sigma^2\{\widehat{p_n}(x)\} \overset{n\to\infty}{\to} 0$ can be sometimes distorting, because in the previously mentioned case

$$
\widehat{p_n}(x) = \frac{1}{|X|} \quad \text{is} \quad \sigma^2(\widehat{p_n}(x)) = 0 \quad \text{as well.}
$$

For this reason a more appropriate criterion to describe the quality of $\widehat{p_n}(x)$ would be

$$
\delta^2\{\widehat{p_n}(x)\} = \mathrm{E}\{(\widehat{p_n}(x) - p(x))^2\}\,.
$$

Its minimization do not guarantee (asymptotically) to be unbiased but it guarantees ”convergence” to the true probability $p(x)$ in the sense of

$$
\lim_{n\to\infty} \mathrm{E}\{(\widehat{p_n}(x) - p(x))^2\} = 0\,, \qquad \text{if the ”zero is accessible”.}
$$

Because for each probability value y, for which such a mentioned expected values exists, we have

$$
E\{(y-m)^2\} = \sigma^2(y) + (\mathrm{E}\{y\} - m)^2
$$

thus we can write

$$
\begin{aligned}
\delta^2(\widehat{p_n}(x)) &= \frac{1}{n}\left(\frac{1}{1+c|X|}\right)^2 p(x)(1-p(x)) + \left(p(x)\frac{1}{1+c|X|} + \frac{c}{1+c|X|} - p(x)\right)^2\\
&= \frac{1}{n}\left(\frac{1}{1+c|X|}\right)^2 p(x)(1-p(x)) + \left(\frac{c}{1+c|X|}\right)^2 (1-p(x)|X|)^2\,.
\end{aligned}
$$

Discussion

1. $c = 0 \Rightarrow \widehat{p_n}(x) = n(x)/n$, thus for the classical frequency estimate we have
$$\mathrm{E}\{\widehat{p_n}(x)\} = p(x) \quad \text{and} \quad \sigma^2\{\widehat{p_n}(x)\} = \frac{1}{n}p(x)(1-p(x))\,.$$

2. Further, we analyze the case $c > 0$. In order to get
$$\lim_{n\to\infty} \mathrm{E}\{(\widehat{p_n}(x) - p(x))^2\} = 0\,,$$
$\lim_{n\to\infty} c(n) = 0$ is sufficient. To hold respecting this condition is enough $|c(n)| \leq An^{-\alpha}$, $A \geq 0$, $\alpha > 0$. Because we will exclude the case of $b \leq 0$ from practical reasons, we have $a - b > 0 \Leftrightarrow a > b$. Then
$$[b > 0] \wedge [|c(n)| \leq An^{-\alpha}] \Leftrightarrow 0 < \frac{b}{a-b} \leq An^{-\alpha} \Leftrightarrow 0 < b(n) \leq \frac{n^{-\alpha}}{\frac{1}{A} + n^{-\alpha}}a(n)\,,$$
if $A > 0$. From these conditions then we can choose for practical purposes choice $b(n) = n^{-\alpha}/(B + n^{-\alpha}) = 1/(Bn^{\alpha} + 1)$ and $a(n) = 1$ for some $B = 1/A > 0$ and $\alpha > 0$. Thus by this choice $c(n) = An^{-\alpha}$.

3. We will not discuss the case $c < 0$.

3 Comparison with Some Estimates

Classical Frequency Estimate: $\widehat{p_n^0}(x) = \frac{n(x)}{n}$, in this case $b = 0$ and $a \neq 0 \Rightarrow c = 0$, then
$$\mathrm{E}\{\widehat{p_n}(x)\} = p(x) \quad \text{and} \quad \sigma^2\{\widehat{p_n}(x)\} = \frac{1}{n}p(x)(1-p(x))\,.$$

Posterior Bayes Estimate with Uniform Prior: $0 \leq p(x) \leq 1$, $\sum_{x\in X} p(x) = 1$, $\widehat{p_n^a}(x) = \frac{n(x)+1}{n+|X|} = \frac{n(x)}{n}\frac{1}{1+\frac{|X|}{n}} + \frac{\frac{1}{n}}{1+\frac{|X|}{n}}$, hence $c = 1/n$ and $b(n)(n+1) = a(n)$, then
$$\mathrm{E}\{\widehat{p_n}(x)\} = p(x)\frac{n}{n+|X|} + \frac{1}{n+|X|} \quad \text{and} \quad \sigma^2\{\widehat{p_n}(x)\} = \frac{1}{n}\left(\frac{n}{n+|X|}\right)^2 p(x)(1-p(x))\,.$$

Mix of Frequency and Uniform Estimate: $\widehat{p_n^\lambda}(x) = (1-\lambda)\frac{n(x)}{n} + \frac{\lambda}{|X|}$. By comparison with $\widehat{p_n}(x) = \frac{n(x)}{n}\frac{1}{1+c|X|} + \frac{c}{1+c|X|}$ we will get $\lambda = \frac{1}{1+c|X|}$ or $c = \frac{\lambda}{|X|(1-\lambda)}$ for $0 \leq \lambda < 1$, then
$$\mathrm{E}\{\widehat{p_n}(x)\} = p(x)(1-\lambda) + \frac{\lambda}{|X|} \quad \text{and} \quad \sigma^2\{\widehat{p_n}(x)\} = \frac{1}{n}(1-\lambda)^2 p(x)(1-p(x))\,.$$

4 Smoothness

In the case of a mix of the frequency and uniform estimate is interesting parameter λ. We can use it to set the smoothness of the estimate. Situation $\lambda = 1$ is said to be absolutely smooth (we use uniform distribution for the estimate) and $\lambda = 0$ we can call non-smoothed estimate (then it is classical frequency estimate). Thus, $0 \leq \lambda \leq 1$ we can think as smoothness in previous terms.

Obviously: $[\lambda = \frac{c|X|}{1+c|X|}] \Rightarrow [(c \rightarrow 0) \Rightarrow (\lambda \rightarrow 0)]$. Therefore, with an increasing number of observations the degree of smoothness decreases (by the condition $\lim_{n\to\infty} c(n) = 0$). As we choose $c(n) = An^{-\alpha}$ we get: $\lambda = [1 + n^{\alpha}(A|X|)^{-1}]^{-1}$. Hence, $\partial\lambda/\partial\alpha < 0$ and $\partial\lambda/\partial A > 0$ follow. So the bigger exponent α means the smaller smoothness and bigger parameter A means bigger smoothness.

5 Future Developments

Theory and usage of kernel approximations for continuous problems is quite sophisticated. It has own tasks and research fields. Work as well as applications of this theory is relatively frequented. Main task for continuous distributions is the selection of smoothing parameter (bandwidth). And there is only small number of works which use this theory for discrete distributions. It´s clear that the formulation and solving task of estimation of discrete distribution by means of kernel theory and classical estimation techniques are good transferable at each other. That illustrates this paper as well. But theory of kernel definitions has interpretative values. Connection of kernel and observation enables to suppose observation error. We award relatively high weight (probability) to the registered event. But we do not forget award small but not zero probability to the others. In our inspected case equal to all. It was considered that do not exist instrument (except-probability) to measure diversity of particular events, hence equal. If there would be available some rate of diversity the task becomes more interesting. As rate of diversity (discrimination) we understand any finite and non-negative function $\varphi : X \times X \rightarrow R_1$, that $\varphi(x_1, x_2) \geq 0$ and $\varphi(x_1, x_1) = 0$ for all $x_1, x_2 \in X$. Then it is possible to leave the uncertainty principle (to all non-registered equally) and to more diverse values (from the registered event) ascribe minor weight than to less diverse. Kernel model for such idea can looks like

$$\kappa(x, x_i, D, E) = D^{\varphi(x_i,x)} E^{-\varphi(x_i,x)} / F \,,$$

for suitable selected F. This and the smoothness problems including the convergence speed will be objectives for next research.

6 Sources and Conclusions

The discrete probability estimation problem is quite complicated, especially in nonparametric case. For the continuous case wide developed theory is at disposal (see ?, ?, ?, ?). The situation in discrete case is not the same. The survey of different methods and conceptions of the discrete kernel estimation can be found in ? (?). Discrete kernel estimations are very useful for the interpretation of observed frequencies. This approach may serve as an acceptable uncertainty model.

Discrete nonparametric probability estimation method publication activity is characterized by many bringing back changes and reviving. This phenomenon essence was mentioned in Chapter 5. Sources for reusing the discrete kernels method are some practical tasks (zero frequency problem ? (?), nonstandard time series models, Markovian switching descriptions, mathematical insurance processes, Bayesian approaches ? (?) and so one).

Our paper contribution is mainly concentrated in Chapter 2-4. This part offers description of the model convergence in the language of the second order probability moments. That fact is very clear for practical usage. The rate of smoothness set in Chapter 4 is important too and has simple interpretation (but in nature is very trivial). The presented generalization in Chapter 5 may be found as interesting for future development (it is inspired by ?, ?).

Authors' address:

František Vávra, Pavel Nový, Lucie Reismüllerová, Kateřina Vokáčová, Martina Neumanová
University of West Bohemia
Faculty of Applied Sciences
Department of Computer Science and Engineering
Univerzitní 22
306 14 Plzeň, Czech Republic

E-mail: vavra@kiv.zcu.cz, novyp@kiv.zcu.cz, reis@kiv.zcu.cz, vokac@kiv.zcu.cz,
mneumano@kiv.zcu.cz