POISSONIZATION METHODS AND APPLICATIONS

David M. Mason

Department of Food and
Resource Economics

University of Delaware

The Kernel Density Estimator

Let X, $X_1, X_2, \ldots,$ be a sequence of independent and identically distributed random variables in $\mathbb{R}$ with Lebesgue density f.

Further let $\{h_n\}_{n\geq 1}$ be a sequence of positive constants such that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. The classical kernel estimator is defined as

$$f_{n,K}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right), \text{ for } x \in \mathbb{R},$$

where K is a kernel satisfying $K(u) = 0$ for $|x| > 1/2$ and

$$\int_{\mathbb{R}} K(u)du = 1.$$

The L_1-norm Distance

Devroye and Györfi have long advocated that the natural distance to measure the error in estimation between a density estimator and the density is the L_1-distance

$$||f_{n,K} - f||_1 = \int_{\mathbb{R}} |f_{n,K} - f(x)|dx.$$

Devroye and Györfi (1984), in their book, Nonparametric Density Estimation: The L_1 View, posed the challenging problem to find the asymptotic distribution of the L_1 distance

$$||f_{n,K} - f||_1.$$

The L_1-consistency

Devroye (1983) had shown that whenever $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$ then with probability one as $n \to \infty$.

$$||f_{n,K} - f||_1 \to 0.$$

Later on we shall discuss conditions under which

$$(nh_n)^{1/2} E \|f_{n,K} - f\|_1 \to m(1, f, K).$$

and

$$(nh_n)^{1/2} \|f_{n,K} - f\|_1 \to_p m(1, f, K),$$

where

$$m(1, f, K) = \|K\|_2 E|Z| \int_{\mathbb{R}} f^{1/2}(y)\, dy,$$

with Z being a standard normal random variable.

Central Limit Theorem

In 2001 I applied the Poissonization methods in Beirlant and Mason (1995) to the special case of the L_1-norm of the kernel density estimator to show that whenever K is bounded, $h_n \to 0$ and $\sqrt{n}h_n \to \infty$ then

$$\xi_n(K) := \sqrt{n}\{||f_{n,K} - Ef_{n,K}||_1 - E||f_{n,K} - Ef_{n,K}||_1\},$$

converges in distribution to a normal random variable with mean zero and variance

$$\sigma^2(K)/||K||_2^2 =$$

$$\int_{-1}^{1} cov\left(\left|\sqrt{1-\rho^2(K,K,t)}Z_1 + \rho(K,K,t)Z_2\right|, |Z_2|\right) dt,$$

where Z_1 and Z_2 are independent standard normal random variables and

$$\rho(K,K,t) = \frac{\int_{-1}^{1} K(u)K(u+t)du}{||K||_2^2}.$$

The proof appeared in the 2001 Eggermont and LaRiccia book on penalized maximum likelihood.

The L1-norm Density Estimator Process

Later Evarist Giné, Andre Zaitsev and I (2003) extended this result to show that under suitable assumptions on a class of kernels $\mathcal{K}$ the sequence of processes

$$\{\xi_n(K) : K \in \mathcal{K}\}_{n\geq 1}$$

converges weakly to a mean zero Gaussian process

$$\{\xi(K) : K \in \mathcal{K}\}$$

with covariance function defined for $K_1, K_2 \in \mathcal{K}$ by

$$\frac{\sigma(K_1, K_2)}{||K_1||_2||K_2||_2} :=$$

$$\int_{-1}^{1} \text{Cov}\left(\left|\sqrt{1-\rho^2(K_1,K_2,t)}Z_1 + \rho(K_1,K_2,t)Z_2\right|, |Z_2|\right) dt$$

and where for $t \in \mathbb{R}$

$$\rho(K_1, K_2, t) := \frac{\int_{\mathbb{R}} K_1(u)K_2(u+t)du}{||K_1||_2||K_2||_2}.$$

In the process we developed a number of very useful Poissonization tools.

In this talk I describe two interesting problems that I have been working on with Wolfgang Polonik and Boris Levit, whose solution relies heavily on these Poissonization methods.

Part 1: Level Set Estimation

In this part of my talk I discuss work in progress with Wolfgang Polonik on level set estimation.

Let f be a bounded Lebesgue density on $\mathbb{R}^2$. (We consider only the $\mathbb{R}^2$-version of the problem for now.) Define the level set

$$C(c) = \{x : f(x) \geq c\}.$$

Assume that

$$\inf_{x \in \mathbb{R}^2} f(x) < c < \sup_{x \in \mathbb{R}^2} f(x).$$

Let $X_1, X_2, \ldots$ be i.i.d. with density f and consider the kernel density estimator of f based on $X_1, \ldots, X_n$, $n \geq 1$,

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n^{1/2}}\right), \; x \in \mathbb{R}^2,$$

where K is a kernel having support contained in the closed ball of radius $1/2$ centered at zero and is bounded by a constant κ.

We shall assume that for some $0 \leq \tau < \infty$,

(H) $\quad nh_n^2 \to \tau, \;$ as $n \to \infty$.

Random Level Sets

Consider the random level set

$$\mathbb{C}_n(c) = \{x : f_n(x) \geq c\}.$$

Our interest is to derive the exact asymptotic distribution of the Lebesgue measure of the symmetric difference betweeen $C_n(c)$ and $C(c)$, that is, the quantity

$$d(\mathbb{C}_n(c),\, C(c)\,) := Leb(\mathbb{C}_n(c)\,\Delta\, C(c))$$

$$= \int_{\mathbb{R}^2} |\, I\{f_n(x) \geq c\,\} - I\{\, f(x) \geq c\}\,|\, dx.$$

Convergence of Random Sets

It is well-known that under mild conditions

$$d(\mathbb{C}_n(c),\, C(c)) \to_P \; 0.$$

Even more is known.

Cadre (2006) derived assumptions under which for some $\mu > 0$ we have

$$\sqrt{n\, h_n}\, d(\mathbb{C}_n(c),\, C(c)) \;\to_P \mu.$$

Asymptotic Normality

Our aim is to show that with the normalizing sequence $a_n = \left(\frac{n}{h_n}\right)^{\frac{1}{4}}$ and a suitable centering sequence b_n we have

$$a_n \, \{d(\mathbb{C}_n(c),\, C(c)) - b_n \} \to_d \mathcal{N}(0, \sigma^2)$$

for some $\sigma^2 > 0$. The following heuristics indicate why

$$a_n = \left(\frac{n}{h_n}\right)^{1/4}$$

is the correct normalizing factor.

Heuristics

First notice that the boundary of the set $\mathbb{C}_n(c)$ can be expected to fluctuate around a band B around the set

$$\partial C(c) = \{x : f(x) = c\}.$$

It is well-known that under certain assumptions we have ignoring a log term

$$\sqrt{n\,h_n}\,\sup_x |f_n(x) - f(x)| = O_P(1) \quad \text{as } n \to \infty.$$

This indicates that under appropriate smoothness assumptions on f the set B will have a 'width' of order $O_P\left(\frac{1}{\sqrt{n\,h_n}}\right)$.

We can cover the band B by

$$N = O\left(\frac{1}{\sqrt{n\,h_n}\,h_n}\right) = O\left(\frac{1}{\sqrt{n\,h_n^3}}\right)$$

disjoint regions $R_i, i = 1, \ldots, N$, of area h_n.

Approximation

We can approximate $d(\mathbb{C}_n(c),\, C(c))$ as

$$d(\mathbb{C}_n(c),\, C(c)) \approx \int_B |\, I\{f_n(x) \geq c\,\} - I\{\, f(x) \geq c\}\,|\, dx$$

$$\approx \sum_{k=1}^{N} \int_{R_i} |\, I\{f_n(x) \geq c\,\} - I\{\, f(x) \geq c\}\,|\, dx$$

$$=: \sum_{k=1}^{N} Y_{n,k}$$

Writing

$$Y_{n,k} = \int_{R_k} \Delta_n(x)\, dx$$

and

$$\Delta_n(x) = |\, I\{f_n(x) \geq c\,\} - I\{\, f(x) \geq c\}\,|,$$

we see that

$$\mathrm{Var}(Y_{n,k}) = \int_{R_k} \int_{R_k} \mathrm{cov}(\Delta_n(x), \Delta_n(y))\, dx\, dy$$

$$= O\,\left(Leb(R_k)^2\right) = O(h_n^2).$$

Further Heuristics

The O-term turns out to be exact. Further, after a Poissonization, which will be soon described, due to the choice of a kernel with a compact support and the fact that we choose the regions R_k so as not to overlap, the random variables $Y_{n,k}$ will be m-dependent.

Hence, the variance of

$$d(\mathbb{C}_n(c), C(c))$$

can be expected to be of the order

$$N\, h_n^2 = \sqrt{\frac{h_n}{n}},$$

which motivates the normalizing factor $a_n = \left(\frac{n}{h_n}\right)^{\frac{1}{4}}$.

Poissonization

First replace $f_n(x)$ by its Poissonized version

$$\pi_n(x) = \frac{1}{nh_n} \sum_{i=1}^{N_n} K\left(\frac{x - X_i}{h_n^{1/2}}\right),$$

where N_n is a mean n Poisson random variable independent of $X_1, X_2, \ldots$

Notice that

$$E\pi_n(x) = Ef_n(x).$$

Define

$$\Pi_n(c) = \int_B |I\{\pi_n(x) \geq c\} - I\{Ef_n(x) \geq c\}| \, dx.$$

The idea is to infer a central limit theorem for

$$d(\mathbb{C}_n(c), C(c))$$

from a central limit theorem for $\Pi_n(c)$.

The idea works!

After Poissonization, one must then de-Poissonize.

De-Poissonizatioin

To de-Poissonize we need a version of a result in Beirlant and M (1995).

LEMMA Let $N_{1,n}$ and $N_{2,n}$be independent Poisson random variables with $N_{1,n}$ being Poisson$(n\beta_n)$ and $N_{2,n}$ being Poisson$(n\,(1-\beta_n))$ where $\beta_n \in (0,1)$.

Denote $N_n = N_{1,n} + N_{2,n}$ and set

$$U_n = \frac{N_{1,n} - n\beta_n}{\sqrt{n}} \text{ and } V_n = \frac{N_{2,n} - n\,(1-\beta_n)}{\sqrt{n}}.$$

Let $\{S_n\}_{n=1}^{\infty}$ be a sequence of random variables such that

(i) for each $n \geq 1$, the random vector (S_n, U_n) is independent of V_n,

(ii) for some $\sigma^2 < \infty$, $\qquad S_n \to_d \sigma Z$, as $n \to \infty$,

(iii) $\beta_n \to 0$, as $n \to \infty$.

Then, for all x,

$$\Pr\{S_n \leq x \mid N_n = n\} \to \Pr\{\sigma Z \leq x\}.$$

Part 2: Lp-Risk Bounds for Kernel Density Estimators

Let $X, X_1, X_2, \ldots$ be i.i.d. with density f and consider the kernel density estimator of f based on $X_{1,\ldots,} X_n$, $n \geq 1$,

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \; x \in \mathbb{R},$$

where h_n are positive constants such that

$$h_n \to 0 \text{ and } nh_n \to \infty \text{ as } n \to \infty$$

and K is an $L_1(\mathbb{R})$ kernel bounded by some constant $\kappa > 0$ and satisfying

$$\int_{\mathbb{R}} \Psi_K(x)dx < \infty,$$

with

$$\Psi_K(x) = \sup_{|y| \geq |x|} |K(y)|, \; x \in \mathbb{R}.$$

and

$$f * K_h(z) := h^{-1} \int_{\mathbb{R}} f(x) \, K\left(\frac{z - x}{h}\right) \, dx.$$

Lp–Risk

Motivated by the work of Ibragimov and Hasminskii (1990) and earlier, Boris Levit and I are interested in finding good finite sample and asymptotic bounds for the Lp risk

$$E\left(\int |f_n(x) - Ef_n(x)|^p \, dx\right)^{1/p}, \; p \geq 1.$$

One of our essential tools is the following Banach space moment bound.

Fact (Corollary 1 of Pinelis (1995)) If $\mathbf{B}$ is a separable Banach space with norm $\|\cdot\|$, Z_i, $i \in \mathbf{N}$, are independent mean zero random vectors taking values in $\mathbf{B}$ and $r \geq 1$, then for all $n \geq 1$,

$$E\left(\|S_n\|^r\right) \leq 2^{r-1} r^{r/2} e^r \left(E\|S_n\|\right)^r + 2^{r-1} r^r E \max_{1 \leq i \leq n} \|Z_i\|^r,$$

where $S_n = Z_1 + \cdots + Z_n$.

Application of Bound

We apply this bound to the random functions

$$Z_i(\cdot) = \frac{K\left(\frac{\cdot - X_i}{h_n}\right) - EK\left(\frac{\cdot - X}{h_n}\right)}{nh_n}, \ i = 1, \ldots, n,$$

which by the assumptions on K are in $L_p(\mathbb{R})$ for any $p \geq 1$.

Eventually we get the following finite sample bound

$$(nh_n)^{r/2} E \|f_n - Ef_n\|_p^r \leq A^r \left[r^{r/2} + \frac{r^r}{(nh_n)^{r/2}} \right],$$

which leads to:

Corollary For any $p \geq 1$ under suitable conditions on K and f, for every $t > 0$ there exists an n_t such that for all $n \geq n_t$,

$$E \exp\left(t\sqrt{nh_n} E \|f_n - Ef_n\|_p \right) < \infty.$$

Exact Asymptotic Risk

Proposition Under suitable conditions on K and f, for $p \geq 1$, as $n \to \infty$,

$$(nh_n)^{p/2} E \left\| f_n - E f_n \right\|_p^p \to m\left(p, f, K\right),$$

and

$$(nh_n)^{p/2} \left\| f_n - E f_n \right\|_p^p \to_p m\left(p, f, K\right),$$

where

$$m\left(p, f, K\right) = \left\| K \right\|_2^p E\left|Z\right|^p \int_{\mathbb{R}} f^{p/2}\left(y\right) dy,$$

with Z denoting a standard normal random variable.

Eventually our goal is to obtain asymptotic minimax results of the form

$$\inf_{f_n} \sup_{f \in \Sigma} Ew \left(\frac{\sqrt{nh_n} \left\| f_n - f \right\|_p}{\left\| K \right\|_2 \left(E\left|Z\right|^p \int_{\mathbb{R}} f^{p/2}\left(y\right) dy \right)^{1/p}} \right) \to w\left(1\right)$$

for a general class of functions w, the infimum taken over all estimators f_n of f and the supremum over a subclass Σ of the L_p densities.

An analoguous result was proved by Guerre and Tsybakov (1998) in a Gaussian regression setting. A step in this direction is the following result.

Exact Asymptotic Risk for General Loss Functions

Corollary Under the conditions of the Proposition for $p \geq 1$, for any loss function w continuous at 1 such that for some $\lambda > 0$ and $C > 0$

$$0 \leq w(x) \leq C \exp(\lambda |x|), \ x \in \mathbb{R},$$

we have as $n \to \infty$,

$$Ew\left(\frac{\sqrt{nh_n}\,\|f_n - Ef_n\|_p}{\|K\|_2 \left(E|Z|^p \int_{\mathbb{R}} f^{p/2}(y)\,dy\right)^{1/p}}\right) \to w(1).$$

Remark To replace $\|f_n - Ef_n\|_p$ by $\|f_n - f\|_p$ requires additional smoothness conditions

In order to obtain these exact results we needed the following two basic Poissonization bounds for random sums. They were used to show that

$$Var\left((nh_n)^{p/2}\,\|f_n - Ef_n\|_p^p\right) \to 0.$$

In addition, a general Berry-Esseen result of Sweeting (1977) was essential to obtain the exact rates. It had been also crucial in Giné, Mason and Zaitsev (2003) to derive an exact asymptotic expression for the variance in their L1-norm CLT.

The following fact is a special case of Lemma 2.1 of Giné, Mason and Zaitsev (2003).

Fact Let X, $X_i, i \in \mathbf{N}$ be a sequence of i.i.d. real valued random variables and independent of them let η be a Poisson random variable with mean n. Let $\mathbb{L}$ be a measurable bounded function equal to zero off of a compact interval $[-L, L]$, C a measurable set and A the Lh-neighborhood of C, $h > 0$. Also set

$$b(x) = E\mathbb{L}\left(\frac{x - X}{h}\right).$$

and with $p \geq 1$, set

$$c(x) = E\left(\left|\sum_{i=1}^{\eta} \mathbb{L}\left(\frac{x - X_i}{h}\right) - b(x)\right|^p\right).$$

Define the real valued measurable function on $\mathbb{R}^n$

$$H\left(\sum_{i=1}^{n} I(x_i \in A)\delta_{x_i}\right)$$

$$= \left(\int_C \left\{\left|\sum_{i=1}^{n} \mathbb{L}\left(\frac{x - x_i}{h}\right) - b(x)\right|^p - c(x)\right\} dx\right)^2,$$

Then if $\gamma = P(X_i \in A) < 1$, we have for some $C_\gamma > 0$

$$EH\left(\sum_{i=1}^{n} I(X_i \in A)\delta_{X_i}\right) \leq C_\gamma EH\left(\sum_{i=0}^{\eta} I(X_i \in A)\delta_{X_i}\right).$$

The next fact is Lemma 2.3 of Giné, Mason and Zaitsev (2003). It is a Poissonized version of Rosenthal's Inequality.

Fact If, for each $n \in N$, $\zeta, \zeta_1, \zeta_2, \ldots, \zeta_n, \ldots$, are independent identically distributed random variables, $\zeta_0 = 0$, and η is a Poisson random variable with mean $\gamma > 0$ and independent of the variables $\{\zeta_i\}_{i=1}^{\infty}$, then, for every $p \geq 2$,

$$E\left|\sum_{i=0}^{\eta} \zeta_i - \gamma E\zeta\right|^p \leq \left(\frac{15p}{\log p}\right)^p \max\left[\left(\gamma E\zeta^2\right)^{p/2}, \gamma E|\zeta|^p\right].$$

Moreover, specializing to $\zeta \equiv 1$, we have for every $p \geq 2$,

$$E\left|\eta - \gamma\right|^p \leq \left(\frac{15p}{\log p}\right)^p \max\left[\gamma^{p/2}, \gamma\right].$$