6 Multivariate density estimation over $\mathbb{R}$

Let $X(\omega) : \Omega \to \mathbb{R}^D$ be a random vector defined on $(\Omega, \mathcal{F}, P)$ for $D \in \mathbb{N}$ and consider a random sample of observations $\{X_i\}_{i=1}^n$ on X. We assume that $F(x) = \int_{X \leq x} f(u)du$, where $x \in \mathbb{R}^D$ and $X \leq x$ is taken element-wise. As in the univariate case f is called the density function of X.

Definition 9. *Let $H_n = diag\{h_{d,n}\}_{d=1}^D$ be a nonstochastic matrix with $0 < h_{d,n}$ for all d and n, $x \in \mathbb{R}^D$, $K : \mathbb{R}^D \to \mathbb{R}$ be a measurable function. Then, we define*

$$\hat{f}(x) = \frac{1}{ndet(H_n)} \sum_{i=1}^{n} K\left(H_n^{-1}(X_i - x)\right)$$

where $det(H_n)$ denotes the determinant of the matrix H_n and H_n^{-1} denotes its inverse.

Remark. Since $0 < h_{d,n}$ for all d, $det(H_n) > 0$ and H_n^{-1} exists. Under the assumption that $K : \mathbb{R}^D \to \mathbb{R}$ is a measurable function and that $E\left|K\left(H_n^{-1}(X_i - x)\right)\right| < \infty$ and $E\left(K\left(H_n^{-1}(X_i - x)\right)^2\right) < \infty$ it is straightforward to prove a multivariate version of Theorem 6. Hence, we have

$$E(\hat{f}(x)) = \int K(\gamma) f(x + H_n\gamma) d\gamma$$

and

$$V(\hat{f}(x)) = \frac{1}{ndet(H_n)} \int K^2(\gamma) f(x + H_n\gamma) d\gamma - \frac{1}{n}\left(\int K(\gamma) f(x + H_n\gamma) d\gamma\right)^2.$$

The integral sign in both expression are D-fold Riemann integrals. We now state an extension of Theorem 7. Throughout let $||x||_1 = \sum_{i=1}^{D} |x_i|$ be the norm of the vector x and recall that by the Cauchy-Schwarz Inequality $||x||_1 \leq D^{1/2}||x||_E$.

Theorem 25. *Assume that a) $\int_{\mathbb{R}^D} |K(\phi)| d\phi < \infty$; b) $||\phi||_1 |K(\phi)| \to 0$ as $||\phi||_1 \to \infty$; c) $0 < h_{d,n} \to 0$ for all d as $n \to \infty$. Let $f(x) : \mathbb{R}^D \to \mathbb{R}$ such that d) $\int_{\mathbb{R}^D} |f(\phi)| d\phi < \infty$. Then, for every continuity point x of f,*

$$det(H_n)^{-1} E\left(K\left(H_n^{-1}(X_i - x)\right)\right) \to f(x) \int K(\phi) d\phi \text{ as } n \to \infty. \tag{40}$$

The remark that follows Theorem 7 applies to the multivariate case and therefore if $|K(\phi)| < C$

$$det(H_n)^{-1} E\left(K^r\left(H_n^{-1}(X_i - x)\right)\right) \to f(x) \int K^r(\phi) d\phi$$

as $n \to \infty$ for $r > 1$. Hence, we have the following multivariate version of Theorem 8.

Theorem 26. *Let $\{X_i\}_{i=1}^n$ be a sequence of IID random vectors with density f. Assume that a) $|K(\gamma)| \leq M$ for all $\gamma \in \mathbb{R}^D$; b) $\int |K(\gamma)|d\gamma < \infty$ and $\int K(\gamma)d\gamma = 1$; c) $||\gamma||_1 |K(\gamma)| \to 0$ as $||\gamma||_1 \to \infty$; d) $h_{d,n} > 0$ for all n and $h_{d,n} \to 0$ as $n \to \infty$. Then, $E(\hat{f}(x)) - f(x) \to 0$ as $n \to \infty$ and if $n\, det(H_n) \to \infty$ as $n \to \infty$, $V(\hat{f}(x)) \to 0$ as $n \to \infty$ for every x which is a point of continuity of f.*

Remark 14. *1. Theorem 26 implies that $\hat{f}(x)$ converges in quadratic mean to $f(x)$ and consequently $\hat{f}(x) \xrightarrow{p} f(x)$.*

2. If f satisfies a Lipschitz condition of order 1 in $\mathbb{R}^D$, i.e., for all $x, x' \in \mathbb{R}^D$ with $x \neq x'$ we have that for some $C < \infty$, $|f(x) - f(x')| \leq C\|x - x'\|_1$. Then

$$\begin{aligned}
|E(\hat{f}(x)) - f(x)| &\leq \int |f(x + H_n\gamma) - f(x)||K(\gamma)|d\gamma \leq C \int \|H_n\gamma\|_1 |K(\gamma)|d\gamma \\
&= C \int \left(\sum_{d=1}^{D} h_{n,d}|\gamma_d| \right) |K(\gamma)|d\gamma \\
&= C \sum_{d=1}^{D} h_{n,d} \int |\gamma_d||K(\gamma)|d\gamma \leq C\ trace(H_n) \text{ if } \int |\gamma_d||K(\gamma)|d\gamma < C.
\end{aligned}$$

The condition that $\int |\gamma_d||K(\gamma)|d\gamma < C$ is verified if $K \geq 0$ and $\int \gamma_d K(\gamma)d\gamma < C$.

Let $D_{d_1,\cdots,d_m} f(x) = \frac{\partial^m}{\partial x_{d_1} \cdots \partial x_{d_m}} f(x)$ for $d_1, \cdots, d_m = 1, \cdots, D$ whenever the partial derivatives on the righthand side of the equality exist. We now define

Definition 10. *Let $T \subseteq \mathbb{R}^D$ and let β and C be two positive numbers. The Hölder class $\Sigma(\beta, C)$ on T is defined as the set of $\lfloor\beta\rfloor$ times differentiable functions $f : T \to \mathbb{R}$ whose partial derivatives of order $\lfloor\beta\rfloor$ satisfy*

$$|D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x) - D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x')| \leq C||x - x'||_1^{\beta - \lfloor\beta\rfloor}, \forall\, x \neq x',\ x, x' \in T.$$

Theorem 27. *Assume that $f \in \Sigma(\beta, C)$ for $\beta \geq 2$ and K is a multivariate kernel of order $\lfloor\beta\rfloor$, i.e., $\int \gamma_{d_1} \cdots \gamma_{d_{\lfloor\beta\rfloor}} K(\gamma)d\gamma = 0$ for $d_1, \cdots, d_{\lfloor\beta\rfloor} = 1, \cdots, D$. Then, if $\int |K(\gamma)||\gamma_d|^\beta d\gamma < C$ for $d = 1, \cdots, D$*

$$E(\hat{f}(x)) - f(x) = O\left(\sum_{d=1}^{D} h_{n,d}^{\beta} \right) = O\left(trace(H_n^{\beta}) \right). \tag{41}$$

Proof. Since $f \in \Sigma(\beta, C)$, by Taylor's Theorem, for some $\lambda \in (0,1)$

$$
\begin{aligned}
f(x+H_n\gamma) - f(x) &= \sum_{d_1=1}^{D} D_{d_1} f(x) h_{n,d_1}\gamma_{d_1} + \frac{1}{2}\sum_{d_1=1}^{D}\sum_{d_2=1}^{D} D_{d_1,d_2} f(x) h_{n,d_1}h_{n,d_2}\gamma_{d_1}\gamma_{d_2} + \cdots \\
&\quad + \frac{1}{(\lfloor\beta\rfloor-1)!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor-1}=1}^{D} D_{d_1,\cdots,d_{\lfloor\beta\rfloor-1}} f(x) h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor-1}}\gamma_{d_1}\cdots\gamma_{d_{\lfloor\beta\rfloor-1}} \\
&\quad + \frac{1}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D} D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x+\lambda H_n\gamma) h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}\gamma_{d_1}\cdots\gamma_{d_{\lfloor\beta\rfloor}}
\end{aligned}
$$

Since K is of order $\lfloor\beta\rfloor$ and $f \in \Sigma(\beta, C)$,

$$
\begin{aligned}
E(\hat{f}(x)) - f(x) &= \int (f(x+H_n\gamma) - f(x))d\gamma \\
&= \frac{1}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D}\int D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x+\lambda H_n\gamma) h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}\gamma_{d_1}\cdots\gamma_{d_{\lfloor\beta\rfloor}} K(\gamma)d\gamma \\
&= \frac{1}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D}\int \Big(D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x+\lambda H_n\gamma) - D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x)\Big) \\
&\times h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}\gamma_{d_1}\cdots\gamma_{d_{\lfloor\beta\rfloor}} K(\gamma)d\gamma
\end{aligned}
$$

By the Triangle Inequality, $f \in \Sigma(\beta, C)$ and the fact that $\lambda \in (0,1)$, we have

$$
\begin{aligned}
|E(\hat{f}(x)) - f(x)| &\le \frac{C}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D}\int \|\lambda H_n\gamma\|_1^{\beta-\lfloor\beta\rfloor} h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}|\gamma_{d_1}|\cdots|\gamma_{d_{\lfloor\beta\rfloor}}||K(\gamma)|d\gamma \\
&= \frac{C}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D}\int \sum_{d=1}^{D} h_{n,d}^{\beta-\lfloor\beta\rfloor}|\gamma_d|^{\beta-\lfloor\beta\rfloor} h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}|\gamma_{d_1}|\cdots|\gamma_{d_{\lfloor\beta\rfloor}}||K(\gamma)|d\gamma
\end{aligned}
$$

where the last inequality follows from the fact that $h_{d,n}, |\gamma_d| > 0$ and $\beta - \lfloor\beta\rfloor \le 1$. The largest power on $h_{n,d}$ occurs when $d_1 = \cdots = d_{\lfloor\beta\rfloor} = d$ and we have

$$
|E(\hat{f}(x)) - f(x)| \le \frac{CD}{\lfloor\beta\rfloor!}\sum_{d=1}^{D} h_{n,d}^{\beta}\int |\gamma_d|^{\beta}|K(\gamma)|d\gamma = O\left(\sum_{d=1}^{D} h_{n,d}^{\beta}\right)
$$

provided $\int |K(\gamma)||\gamma|^{\beta}d\gamma < C$ and $h_n > 0$. $\square$

The following theorem is an extension of Theorem 10 to the multivariate setting.

Theorem 28. *Assume that: a)$\{X_i\}_{i=1}^n$ forms an IID sequence of random vectors with density f; b) $|K(\gamma)| \leq M$ for all $\gamma \in \mathbb{R}^D$; c) $\int |K(\gamma)|d\gamma < \infty$; d) $||\gamma||_1|K(\gamma)| \to 0$ as $||\gamma||_1 \to \infty$; e) $\int K(\psi)^{2+\delta}d\psi < \infty$ for some $\delta > 0$; f) $h_{d,n} > 0$ for all n, $h_{d,n} \to 0$ and $ndet(H_n) \to \infty$ as $n \to \infty$. Then,*

$$(ndet(H_n))^{\frac{1}{2}}(\hat{f}(x) - E(\hat{f}(x))) \xrightarrow{d} N\left(0, f(x)\int K^2(\psi)d\psi\right),$$

where x is a point of continuity of f.

Remark 15.

1. *Theorem 28 shows that one of the consequences of estimating multivariate densities is that the rate of convergence in distribution slows (exponentially) as D grows. This is known as the "curse" of dimensionality.*

2. *Note that under the conditions of Theorem 27*

$$\begin{aligned}(n\det(H_n))^{\frac{1}{2}}(\hat{f}(x) - f(x)) &= (n\det(H_n))^{\frac{1}{2}}(\hat{f}(x) - E(\hat{f}(x))) \\ &\quad + (n\det(H_n))^{\frac{1}{2}}(E(\hat{f}(x)) - f(x)) \\ &= (n\det(H_n))^{\frac{1}{2}}(\hat{f}(x) - E(\hat{f}(x))) + (n\det(H_n))^{\frac{1}{2}}O\left(\sum_{d=1}^{D} h_{n,d}^{\beta}\right).\end{aligned}$$

Hence, if $(n\det(H_n))^{\frac{1}{2}}\sum_{d=1}^{D} h_{n,d}^{\beta} \to 0$ as $n \to \infty$,

$$(ndet(H_n))^{\frac{1}{2}}(\hat{f}(x) - f(x)) \xrightarrow{d} N\left(0, f(x)\int K^2(\psi)d\psi\right).$$

3. *If, in the previous item, $h_{n,d} = h_n$ for all d, then*

$$(n\det(H_n))^{\frac{1}{2}}\sum_{d=1}^{D} h_{n,d}^{\beta} = D(nh_n^{D+2\beta})^{1/2} = o(1) \text{ if } nh_n^{D+2\beta} = o(1).$$

If $h \propto n^{\delta}$, since $n\det H_n \to \infty$ and $nh_n^{D+2\beta} = o(1)$ as $n \to \infty$ we need

$$-\frac{1}{D} < \delta < -\frac{1}{2\beta + D}$$

to conclude that

$$(nh_n^D)^{\frac{1}{2}}(\hat{f}(x) - f(x)) \xrightarrow{d} N\left(0, f(x)\int K^2(\psi)d\psi\right).$$

7 Bandwidth Choice

7.1 Minimizing Integrated Squared Error

It should be apparent by now that bandwidth choice is an important part of nonparametric density or distribution estimation. Intuitively, we would like to select a sequence $h_n := h$ so that a suitably defined distance between $\hat{f}(x)$ and $f(x)$ is as small as possible over the set where the random variable X takes values. We are here searching for a global criterion for the choice of h_n.

Distances in function spaces can be defined in various ways, subject to some mathematical technicalities. Here, we search for h that will minimize

$$\int (f(x) - \hat{f}(x))^2 dx \tag{42}$$

In the statistical literature this integral is called the Integrated Square Error (ISE) of $\hat{f}$, and we write $ISE(\hat{f})$. Note that $ISE(\hat{f})$ does not depend on where the estimator $\hat{f}$ is being evaluated, but it does depend on the sample $\{X_i\}_{i=1}^n$, hence $ISE(\hat{f})$ is itself a random variable. It is clear that minimization of $ISE(\hat{f})$ with respect to h_n is equivalent to the minimization of $\int \hat{f}^2(x)dx - 2\int f(x)\hat{f}(x)dx$.

From the definition of $\hat{f}(x)$ and taking K to be an even function,

$$\int \hat{f}^2(x)dx = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{h_n}K * K\left(\frac{X_i - X_j}{h_n}\right)$$

where $K * K(u) = \int K(u-t)K(t)dt$.

The integral $\int f(x)\hat{f}(x)dx$ clearly depends on f, which is unknown. Hence, to render the criterion for choosing h_n operational we need a suitable approximation for $\int f(x)\hat{f}(x)dx$. Note that,

$$\begin{aligned} \int f(x)\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{h_n}K\left(\frac{X_i - x}{h_n}\right)f(x)dx \\ &= \frac{1}{n}\sum_{i=1}^{n} Z_{ni} \end{aligned}$$

and that

$$E\left(Z_{ni}\right) = \int\int \frac{1}{h_n}K\left(\frac{\alpha - x}{h_n}\right)f(x)f(\alpha)dxd\alpha = E\left(\frac{1}{h_n}K\left(\frac{\alpha - x}{h_n}\right)\right),$$

where the expectation is taken with respect to the joint density of two independent and identically distributed random variables. Note also that

$$E\left(\frac{1}{h_n}K\left(\frac{\alpha-x}{h_n}\right)\right)\approx\frac{1}{n(n-1)}\frac{1}{h_n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K\left(\frac{X_i-X_j}{h_n}\right),$$

suggesting the following approximation for $\int\hat{f}^2(x)dx-2\int f(x)\hat{f}(x)$ which can be minimized with respect to h_n,

$$M(\hat{f})=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{h_n}K*K\left(\frac{X_j-X_i}{h_n}\right)-2\frac{1}{n(n-1)}\frac{1}{h_n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K\left(\frac{X_i-X_j}{h_n}\right).$$

Since K is known, h_n that minimizes $M(\hat{f})$ can be chosen by a numerical optimization procedure.

Example: Suppose $K(\psi)=\frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}\psi^2)$. Then, $K*K(\psi)=\frac{1}{\sqrt{2\pi}\sqrt{2}}exp(-\frac{1}{4}\psi^2)$, and

$$M(\hat{f})=\frac{1}{h_n n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}\sqrt{2}}exp\left(-\frac{1}{4}\left(\frac{X_i-X_j}{h_n}\right)^2\right)-2(n(n-1)h_n)^{-1}\sum_{t=1}^{n}\sum_{i=1,i\neq t}^{n}K\left(\frac{X_t-X_i}{h_n}\right).$$

7.2 Minimizing Mean Integrated Squared Error

One of the remarks made regarding ISE was that it depends on the sample $\{X_i\}_{i=1}^n$. One can imagine a process of repeated sampling that induces a distribution on ISE. Hence, we might consider,

$$\begin{aligned}
MISE(\hat{f}) &= E\left(\int(\hat{f}(x)-f(x))^2dx\right)\\
&= E\left(\int(\hat{f}(x)-E(\hat{f}(x))+E(\hat{f}(x))-f(x))^2dx\right)\\
&= E\left(\int\{(\hat{f}(x)-E(\hat{f}(x)))^2+(E(\hat{f}(x))-f(x))^2\right.\\
&+ \left.2(\hat{f}(x)-E(\hat{f}(x)))(E(\hat{f}(x))-f(x))\}dx\right)\\
&= \int\{V(\hat{f}(x))+E(\hat{f}(x)-f(x))^2\}dx
\end{aligned}$$

Note that $V(\hat{f}(x))=\frac{1}{nh_n}f(x)\int K^2(\psi)d\psi+o((nh_n)^{-1})$ and if $f\in\Sigma(2,C)$ we have that $E(\hat{f}(x))-f(x))=\frac{h_n^2}{2}\mu_2 f^{(2)}(x)+o(h_n^2)$ with $\mu_2=\int x^2K(x)dx$. Hence, $MISE(\hat{f})$ can be written as

$$MISE(\hat{f})=(nh_n)^{-1}\int K^2(\psi)d\psi+\frac{h_n^4}{4}\mu_2^2\int f^{(2)}(x)^2dx+o((nh_n)^{-1})+o(h_n^4).$$

Let $\int K^2(\psi)d\psi = \lambda_2$, $\mu_2^2 \int f^{(2)}(x)^2 dx = \lambda_1$ so that

$$MISE(\hat{f}) = (nh_n)^{-1}\lambda_2 + \frac{h_n^4}{4}\lambda_1 + o((nh_n)^{-1}) + o(h_n^4).$$

Let the h_n that minimizes the first two terms of $MISE(\hat{f})$ be denoted by $\hat{h}_n$. It satisfies

$$-n^{-1}(\hat{h}_n)^{-2}\lambda_2 + \hat{h}_n^3\lambda_1 = 0$$

or $\hat{h}_n^5\lambda_1 = \frac{1}{n}\lambda_2$ which implies

$$\hat{h}_n = \left(\frac{\lambda_2}{n\lambda_1}\right)^{\frac{1}{5}} = n^{-\frac{1}{5}}\left(\frac{\int K^2(\psi)d\psi}{(\int \psi^2 K(\psi)d\psi)^2 \int (f^{(2)}(x))^2 dx}\right)^{\frac{1}{5}}$$

Remark 16. *1. Although λ_2 can be obtained from knowledge of K, λ_1 depends on $f^{(2)}(x)$. Hence, in practice $\hat{h}_n$ depends on obtaining an estimate for $\int (f^{(2)}(x))^2 dx$.*

2. Taking $(\frac{\lambda_2}{\lambda_1})^{\frac{1}{5}} = c$ we see that $\hat{h}_n \propto n^{-\frac{1}{5}}$. Hence, according to the AMISE criterion, there is an optimal rate of convergence of the bandwidth to zero, namely, $n^{-1/5}$. Recall that

$$(nh_n)^{1/2}(\hat{f}(x) - f(x)) = (nh_n)^{1/2}(\hat{f}(x) - E(\hat{f}(x)) + (nh_n)^{1/2}(E(\hat{f}(x)) - f(x))$$

and since the first term after the equality converges in distribution to a normal distribution, normality of the left side of the equality results if $(nh_n)^{1/2}(E(\hat{f}(x)) - f(x))$ approaches zero as $n \to \infty$. But

$$(nh_n)^{1/2}(E(\hat{f}(x)) - f(x)) = n^{1/2}h_n^{5/2}\frac{1}{2}\mu_2 f^{(2)}(x) + n^{1/2}\frac{h_n^{5/2}}{h_n^2}R(n)$$

where $R(n) = O(h_n^2)$. Hence, for $(nh_n)^{1/2}(E(\hat{f}(x)) - f(x))$ to vanish asymptotically, we need $nh_n^5 \to 0$ as $n \to \infty$ which implies that $h_n \propto n^{\delta}$ where $-1 < \delta < -1/5$. Consequently, asymptotic normality of $(nh_n)^{1/2}(\hat{f}(x) - f(x))$ cannot result when the optimal choice h_n is based on AMISE, oversmoothing is necessary.

3. It should be clear that both minimization of the cross validation function and AMISE can be extended to multivariate density estimation with little conceptual effort. For a random vector $X : \Omega \to \mathbb{R}^D$ we have

$$M(\hat{f}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{det(H_n)}K * K\left(H_n^{-1}(X_j - X_i)\right)$$
$$- 2\frac{1}{n(n-1)}\frac{1}{det(H_n)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} K\left(H_n^{-1}(X_i - X_j)\right).$$

Also, in the case where $f \in \Sigma(2, C)$ *and* $\int \gamma_d K(\gamma) d\gamma = 0$ *and* $\int \gamma_{d_1}\gamma_{d_2} K(\gamma) d\gamma = 0$ *when* $d_1 \neq d_2$ *we can write*

$$\begin{aligned} AMISE(\hat{f}) &= (ndet(H_n))^{-1} \int K^2(\psi) d\psi + \frac{\mu_2^2}{4} \int \left(\sum_{i=1}^{D} D_{ii} f(x) h_{i,n}^2 \right)^2 dx \\ &= (ndet(H_n))^{-1} \int K^2(\psi) d\psi + \frac{\mu_2^2}{4} \sum_{i=1}^{D} \sum_{j=1}^{D} h_{i,n}^2 h_{j,n}^2 \int D_{ii} f(x) D_{jj} f(x) dx. \end{aligned}$$

Taking derivatives with respect to each of the elements in H_n *we have that* $\hat{H}_n$ *must satisfy,*

$$\mu_2^2 \left(\hat{h}_{d,n}^3 \int (D_{dd})^2 f(x) dx + \hat{h}_{d,n} \sum_{i=1, i \neq d}^{D} \hat{h}_{i,n}^2 \int D_{dd} f(x) D_{ii} f(x) dx \right) = \frac{\lambda_2}{n \hat{h}_{d,n} det(\hat{H}_n)}$$

for all d. When $H_n = h_n I_D$ *we have*

$$\hat{h}_n^{D+4} = \frac{\lambda_2}{n\mu_2^2 \sum_{i=1}^{D} \sum_{d=1}^{D} \int D_{dd} f(x) D_{ii} f(x) dx} \text{ which implies that } \hat{h}_n = \left(\frac{\lambda_2}{n\mu_2^2 \lambda_1} \right)^{\frac{1}{D+4}}.$$

4. *The simplest procedure to estimate* $\int (f^{(2)}(x))^2 dx$ *is to assume, at this stage, that* $f(x)$ *belongs to a parametric family of densities. For example, if we assume that* $X_i \sim N(\mu, \sigma^2)$, *then for* $r \in \mathbb{N}$

$$\int (f^{(r)}(u))^2 du = \frac{(2r)!}{2^{2r+1} r! \sqrt{\pi} \sigma^{2r+1}}.$$

For $r = 2$ *and plugging back into the optimal bandwidth formula gives*

$$h_{AMISE} = \begin{cases} n^{-1/5} 1.06\sigma & \text{if } K \text{ is a Gaussian kernel} \\ n^{-1/5}(1.06)(2.214)\sigma & \text{if } K \text{ is an Epanechnikov kernel} \\ n^{-1/5}(1.06)(1.740)\sigma & \text{if } K \text{ is an uniform kernel} \end{cases}$$

where σ *can be estimated by the sample standard deviation. The resulting bandwidths are normally referred to as Silverman's rule-of-thumb bandwidths.*

Appendix

Lemma 6. *Let $C \in \mathbb{R}$, $n \in \mathbb{N}$ and $\{X_1, \cdots, X_n\}$ be a collection of random variables defined in the same probability space. Then,*

$$P\left(\max\{X_1, \cdots, X_n\} > C\right) \leq \sum_{i=1}^{n} P(\{X_i > C\}).$$

Proof. First, note that

$$\begin{aligned} P\left(\max\{X_1, \cdots, X_n\} > C\right) &= 1 - P\left(\max\{X_1, \cdots, X_n\} \leq C\right) \\ &= 1 - P\left(\cap_{i=1}^{n}\{X_i \leq C\}\right) \\ &= P\left((\cap_{i=1}^{n}\{X_i \leq C\})^c\right). \end{aligned}$$

Second, by De Morgan's Laws

$$\begin{aligned} P\left(\max\{X_1, \cdots, X_n\} > C\right) &= P\left(\cup_{i=1}^{n}\{X_i \leq C\}^c\right) \\ &\leq \sum_{i=1}^{n} P(\{X_i \leq C\}^c) = \sum_{i=1}^{n} P(\{X_i > C\}). \end{aligned}$$

□

Lemma 7. *(Bernstein's Inequality) Let $\{W_i\}_{i=1}^{n}$ be a sequence of independent random variables satisfying*

1. *$|W_i - E(W_i)| \leq m_i$ for all i,*
2. *$V(W_i) = \sigma_i^2 <$ for all i*

Let $M_n := \max\limits_{1 \leq i \leq n}\{m_i\}$, $S_n := \sum_{i=1}^{n} W_i$ and $\sigma_n^2 := \sum_{i=1}^{n} \sigma_i^2$. Then, for all $\Delta > 0$ we have

$$P\left(|S_n| \geq \Delta\sigma_n\right) < 2\exp\left(-\frac{\Delta^2}{2 + \frac{2}{3}\frac{M_n}{\sigma_n}\Delta}\right)$$

Proof. See Bernstein (1924) or Bennett (1962). □

Lemma 8. *(Loève's c_r Inequality) For $r > 0$,*

$$E\left(\left|\sum_{i=1}^{n} X_i\right|^r\right) \leq c_r \sum_{i=1}^{n} E(|X_i|^r)$$

where $c_r = 1$ when $r \leq 1$and $c_r = n^{r-1}$ when $r > 1$.

Proof. We will prove that

$$\left|\sum_{i=1}^{n} X_i\right|^r \leq c_r \sum_{i=1}^{n} |X_i|^r \tag{43}$$

and then take expectations. By the Triangle Inequality $|\sum_{i=1}^{n} X_i|^r \leq (\sum_{i=1}^{n} |X_i|)^r$. Thus, it suffices to consider the case where each $X_i \geq 0$. Let $0 < r \leq 1$ and note that since $\sum_{i=1}^{n} X_i > 0$ (if $X_i = 0$ for all i then inequality (43) holds with equality) it suffices to show that

$$1 \leq \sum_{i=1}^{n} Z_i^r$$

where $Z_i = \frac{X_i}{\sum_{i=1}^{n} X_i}$, $\sum_{i=1}^{n} Z_i = 1$ and $0 \leq Z_i \leq 1$. But since, $Z_i^r \geq Z_i$ and $\sum_{i=1}^{n} Z_i = 1$, the results follows immediately. For $r > 1$, convexity of x^r immediately gives

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^r \leq \frac{1}{n}\sum_{i=1}^{n} X_i^r \Leftrightarrow \left(\sum_{i=1}^{n} X_i\right)^r \leq n^{r-1}\sum_{i=1}^{n} X_i^r.$$

Taking expectations in both cases gives the desired inequality. □

References

Bennett, G., 1962. Probability inequalities for the sum of independent random variables. Journal of the American Statistical Association 57, 33–45.

Bernstein, S., 1924. Sur une modification de l'inéqualité de tchebichef. Annals Science Institute Sav. Ukraine Sect. Math. I.

Davidson, J., 1994. Stochastic limit theory. Oxford University Press, New York.

Hestenes, M. R., 1941. Extension of the range of a differentiable function. Duke Mathematical Journal 8, 183–192.

Marron, J. S., Ruppert, D., 1994. Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society - Series B 56, 653–671.

Natanson, I. P., 1955. Theory of functions of a real variable. Vol. 1. Frederick Ungar Publishing Co, New York, NY.

Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics 27, 832–837.

Schuster, E. F., 1985. Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics - Theory and Methods, 14, 1123–1136.