# Local polynomial estimators

(notes by Pavel Chigansky)

## 1. The heuristics

**1.1. Regression.** Consider the nonparametric regression model

$$Y_j = f(X_j) + \varepsilon_j, \quad j = 1, ..., n, \tag{1.1}$$

where $f$ is the unknown function and $\varepsilon_j$'s are i.i.d. random variables with zero mean and unit variance. Given $n$ pairs $(X_j, Y_j)$, the Nadaraya-Watson regression estimator is

$$\widehat{f}(x) = \frac{\sum_{j=1}^n Y_j K\Big(\frac{x-X_j}{h}\Big)}{\sum_{j=1}^n K\Big(\frac{x-X_j}{h}\Big)}, \tag{1.2}$$

where $K(\cdot)$ is a kernel and $h$ is the bandwidth tuning parameter. If a positive kernel is chosen, (1.2) can be viewed as a solution to the minimization problem

$$\widehat{f}(x) = \operatorname{argmin}_{\theta\in\mathbb{R}} \sum_{j=1}^n (Y_j - \theta)^2 K\Big(\frac{x-X_j}{h}\Big).$$

Hence it is a *local* least squares fit by a constant, which at a point $x$ is influenced mostly by the samples $X_j$ in a neighbourhood of $x$, when $h$ is small.

This suggests the following generalization. Suppose $f$ has $\ell$ derivatives, then for all $z$ in a vicinity of $x$

$$f(z) = f(x) + f'(x)(z-x) + ... + \frac{1}{\ell!} f^{(\ell)}(x)(z-x)^\ell + o\big((z-x)^\ell\big) =: \\ \theta(x)^\top U\Big(\frac{z-x}{h}\Big) + o\big((z-x)^\ell\big),$$

where we defined

$$\theta(x) = \big(f(x), f'(x)h, ..., f^{(\ell)}(x)h^\ell\big)^\top, \quad U(t) = \Big(1, t, \frac{1}{2!}t^2, ..., \frac{1}{\ell!}t^\ell\Big)^\top. \tag{1.3}$$

Hence for $X_j$'s close to $x$

$$Y_j = \theta(x)^\top U\Big(\frac{X_j - x}{h}\Big) + o\big((X_j - x)^\ell\big) + \varepsilon_j,$$

---

lecture notes for "Advanced Statistical Models B" course.

and $\theta(x)$ can be estimated by solving the parametric regression problem

$$\widehat{\theta}(x) = \mathrm{argmin}_{\theta\in\mathbb{R}^{\ell+1}} \sum_{j=1}^{n} \Big(Y_j - \theta^\top U\Big(\frac{X_j - x}{h}\Big)\Big)^2 K\Big(\frac{x - X_j}{h}\Big), \tag{1.4}$$

where $K(\cdot)$ is a positive kernel with most of its mass concentrated around zero. When $h$ is small, the summation effectively includes only the terms with $X_j$'s close to $x$. The value of $f(x)$ is estimated by the first entry of $\widehat{\theta}(x)$,

$$\widehat{f}(x) := \widehat{\theta}(x)^\top U(0). \tag{1.5}$$

This statistic is known as the Local Polynomial (LP) estimator of order $\ell$. The N-W estimator (1.2) is such an estimator of order zero. Note that the other entries of $\widehat{\theta}(x)$ are reasonable estimators for the derivatives of $f$, after a suitable normalization (Problem 3).

The quadratic minimization problem (1.4) has a closed form solution

$$\widehat{\theta}(x) = B_n(x)^{-1} a_n(x) \tag{1.6}$$

where

$$a_n(x) = \frac{1}{nh} \sum_{j=1}^{n} Y_j U\Big(\frac{X_j - x}{h}\Big) K\Big(\frac{x - X_j}{h}\Big),$$

if the matrix

$$B_n(x) = \frac{1}{nh} \sum_{j=1}^{n} U\Big(\frac{X_j - x}{h}\Big) U\Big(\frac{X_j - x}{h}\Big)^\top K\Big(\frac{x - X_j}{h}\Big) \tag{1.7}$$

is nonsingular. We will assume that this is the case and formulate sufficient conditions for which this assumption holds later on. Hence the LP estimator is linear with respect to $Y_j$'s, i.e. has the form

$$\widehat{f}(x) = \sum_{j=1}^{n} Y_j W_{nj}(x),$$

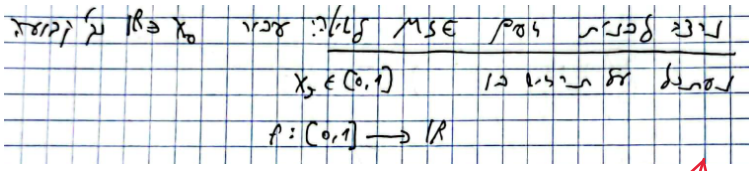
where the weights are given by

$$W_{nj}(x) = \frac{1}{nh} U^\top(0) B_n(x)^{-1} U\Big(\frac{X_j - x}{h}\Big) K\Big(\frac{x - X_j}{h}\Big). \tag{1.8}$$

**1.2. Density estimation.** The LP method can be applied to the density estimation problem in a number of ways. One possibility is to approximate the density by a histogram and regard its bins heights and positions as regression data to which the LP estimator is applied as above. If the number of bins, their widths and the LP kernel and bandwidth are chosen in a suitable way, depending on the sample size, such an estimator can be shown rate optimal on the usual smoothness classes.

Another closely related possibility is to estimate the distribution function by the local polynomial fit and estimate the density as its derivative. More precisely, compute $\widehat{\theta}(x)$ by means of (1.4), where $Y_j$ is replaced with $\widehat{F}_n(X_j)$ and estimate the density by (c.f. Problem 3),

$$\widehat{p}_n(x) = \frac{1}{h} U'(0)^\top \widehat{\theta}_n(x).$$





This natural approach is surprisingly recent [2].

## 2. Upper bound for MSE

In this section we will show that LP estimator (1.4)-(1.5) satisfies the upper bound for the minimax MSE risk

$$\overline{\lim_{n}} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f\left(n^{\frac{\beta}{2\beta+1}}(\widehat{f}_n(x_0)-f(x_0))\right)^2 \le C(\beta,L,K), \tag{2.1}$$

if the design points are sufficiently dense and the kernel $K(\cdot)$ is chosen appropriately. Here $\Sigma(\beta,L)$ is the Holder class and the constant $C(\beta,L,K)$ depends only on this class and the kernel in use. Since the domain of the regression function $f$ is bounded, the integrated MSE risk satisfies the same bound multiplied by the domain diameter.

**2.1. The polynomial reproducing property.** An important property of the LP($\ell$) estimator is that it reproduces polynomials of all degrees not greater than $\ell$. This means that when applied to noiseless observations of the form $Y_j = Q(X_j)$ where $Q$ is a polynomial with $\deg(Q) \le \ell$, it returns exactly $Q(x)$ for all $x$.

LEMMA 2.1. *For any sample $X_1,\dots,X_n$ and a polynomial $Q$ with* $\deg(Q)\le\ell$,

$$\sum_{j=1}^{n} Q(X_j)W_{nj}(x) = Q(x), \quad x\in[0,1]. \tag{2.2}$$

*In particular,*

$$\sum_{j=1}^{n} W_{nj}(x) = 1 \quad \textit{and} \quad \sum_{j=1}^{n}(X_j-x)^k W_{nj}(x) = 0, \quad 1\le k\le \ell. \tag{2.3}$$

PROOF. For a polynomial $Q$ with $\deg(Q)\le\ell$, the Taylor approximation is exact

$$Q(X_j) = Q(x)+Q'(x)(X_j-x)+\dots+\frac{1}{\ell!}Q^{(\ell)}(x)(X_j-x)^\ell = q^\top(x)U\left(\frac{z-x}{h}\right)$$

where $q(x) = \left(Q(x),Q'(x)h,\dots,Q^{(\ell)}(x)h^\ell\right)$. Hence the expression in (1.4) with $Y_j = Q(X_j)$ is minimized to zero by $\widehat{\theta}(x) = q(x)$, i.e. the fit is exact.

The first identity in (2.3) is obtained by taking $Q(x)=1$ in (2.2). To see why the second identity holds, fix any $x\in[0,1]$ and let $Q(u)=(u-x)^k$, then by (2.2),

$$\sum_{j=1}^{n}(X_j-x)^k W_{nj}(u) = (u-x)^k, \quad \forall u\in[0,1].$$

The desired identity is obtained by taking $u:=x$.

□

**2.2. Some properties of the weights.** Further analysis will be carried out under the following additional assumptions.

(LP1) There exists an integer $n_0$ and a real number $\lambda_0 > 0$, such that

$$\lambda_{\min}(B_n(x)) \geq \lambda_0, \quad \forall n \geq n_0, \quad x \in [0,1],$$

where $\lambda_{\min}(\cdot)$ is the smallest eigenvalue.

(LP2) There exists a real number $a_0 > 0$ such that

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{X_j \in A\}} \leq a_0 \max(1/n, \mathrm{Leb}(A)) \tag{2.4}$$

for *any* interval $A \subseteq [0,1]$ and all $n \geq 1$.

(LP3) $\mathrm{supp}(K) \subseteq [-1,1]$ and $\|K\|_\infty < \infty$.

Assumption (LP3) is obviously satisfied by numerous kernels. Assumption (LP1) is a uniform non-degeneracy condition. It can be shown that for the uniform design $X_j = j/n$ it holds for all $n$ large enough if the kernel $K$ is uniformly positive near the origin, i.e., satisfies the condition

$$K(u) \geq K_{\min} \mathbf{1}_{\{|u| \leq \Delta\}}, \quad u \in \mathbb{R}, \tag{2.5}$$

for some positive constants $K_{\min}$ and $\Delta$ (see Problem 6).

Assumption (LP2) implies that the design points $X_j$ do not accumulate in any subinterval of the domain, or in other words, are well spread over the domain. In technical terms it guarantees that the scaled sum

$$\frac{1}{nh}\sum_{j=1}^{n} \mathbf{1}_{\{x-h \leq X_j \leq x+h\}}$$

remains bounded uniformly over $x$, if $h$ does not decrease with $n$ to zero too fast. For example, if all $X_j$'s are equal, then at $x := X_1$, this quantity is unbounded as $h \to 0$ and hence such a degenerate design violates (LP2). On the other hand, for the uniform design $X_j = j/n$, (LP2) holds with $a_0 = 2$. Indeed, if $A$ satisfies $\mathrm{Leb}(A) < 1/n$, it contains at most one design point, the sum in (2.4) consists of a single summand and hence the bound holds for any $a_0 \geq 1$. Otherwise, if $\mathrm{Leb}(A) \geq 1/n$, $A$ contains at most $1 + n\mathrm{Leb}(A)$ points and the bound holds with any $a_0 \geq 2$. Combining the two cases verifies (2.4) for $a_0 = 2$.

These assumptions guarantee the following properties of the weights.

LEMMA 2.2. *Assume* (LP1)-(LP3) *hold, then for all* $n \geq n_0$, $h \geq 1/(2n)$ *and* $x \in [0,1]$

(a) $|W_{nj}(x)| \leq \dfrac{C}{nh}$

(b) $\sum_{j=1}^{n} |W_{nj}(x)| \leq C$

(c) $W_{nj}(x) = 0$ *for* $|X_j - x| > h$

*where $C$ is a constant, which depends only on $\lambda_0$, $a_0$ and $\|K\|_\infty$.*

PROOF.

(a) Using properties (LP1) and (LP3) and the definition (1.8),

$$\begin{aligned}\big|W_{nj}(x)\big| &= \frac{1}{nh}\Big|U^\top(0)B_n(x)^{-1}U\Big(\frac{X_j-x}{h}\Big)K\Big(\frac{x-X_j}{h}\Big)\Big| \le \\ &\frac{1}{nh}\lambda_0^{-1}\|K\|_\infty\Big\|U\Big(\frac{X_j-x}{h}\Big)\mathbf{1}_{\{|(x-X_j)/h|\le 1\}}\Big\| = \\ &\frac{1}{nh}\lambda_0^{-1}\|K\|_\infty\sqrt{1+1+\frac{1}{(2!)^2}+\ldots+\frac{1}{(\ell!)^2}} \le \frac{2}{nh}\lambda_0^{-1}\|K\|_\infty.\end{aligned}$$

(b) Simialrly, by (LP2),

$$\begin{aligned}\sum_{j=1}^n |W_{nj}(x)| &\le \frac{1}{nh}\lambda_0^{-1}\|K\|_\infty\sum_{j=1}^n\Big\|U\Big(\frac{X_j-x}{h}\Big)\Big\|\mathbf{1}_{\{|(x-X_j)/h|\le 1\}} \le \\ &2\lambda_0^{-1}\|K\|_\infty\frac{1}{nh}\sum_{j=1}^n\mathbf{1}_{\{x-h\le X_j\le x+h\}} \le \\ &2\lambda_0^{-1}\|K\|_\infty\frac{1}{h}a_0\max\big(1/n,2h\big) \le 4\lambda_0^{-1}\|K\|_\infty a_0.\end{aligned}$$

(c) This inequality holds since $K$ vanishes outside $[-1,1]$.

□

**2.3. The upper risk bound.** As usual, the starting point is the bias-variance decomposition

$$\mathbb{E}_f(\widehat{f}(x_0)-f(x_0))^2 = \mathrm{Var}_f(\widehat{f}(x_0)) + (\mathbb{E}_f\widehat{f}(x_0)-f(x_0))^2.$$

The variance term satisfies

$$\begin{aligned}&\mathrm{Var}_f(\widehat{f}(x_0)) = \mathbb{E}_f\Big(\sum_{j=1}^n\big(Y_jW_{nj}(x_0)-\mathbb{E}_fY_jW_{nj}(x_0)\big)\Big)^2 = \\ &\mathbb{E}_f\Big(\sum_{j=1}^n\varepsilon_jW_{nj}(x_0)\Big)^2 = \sum_{j=1}^n W_{nj}(x_0)^2 \le \max_j|W_{nj}(x_0)|\sum_{j=1}^n|W_{nj}(x_0)| \le C_1\frac{1}{nh},\end{aligned}$$

where we used the bounds (a) and (b) from Lemma 2.2. To bound the bias, note that by the first identity in (2.3),

$$\mathbb{E}_f\widehat{f}(x_0)-f(x_0) = \sum_{j=1}^n f(X_j)W_{nj}(x_0)-f(x_0) = \sum_{j=1}^n\big(f(X_j)-f(x_0)\big)W_{nj}(x_0).$$

For any $f\in\Sigma(\beta,L)$,

$$\begin{aligned}f(X_j)-f(x_0) = &\sum_{k=1}^\ell\frac{1}{k!}f^{(k)}(x_0)(X_j-x_0)^k+ \\ &\frac{1}{\ell!}\Big(f^{(\ell)}(x_0+\xi_j(X_j-x_0))-f^{(\ell)}(x_0)\Big)(X_j-x_0)^\ell,\end{aligned}$$

where $0 \le \xi_j \le 1$. Hence by the second identity in (2.3),

$$\begin{aligned}
&\left|\mathbb{E}_f \widehat{f}(x_0) - f(x_0)\right| = \\
&\left|\frac{1}{\ell!}\sum_{j=1}^{n}\left(f^{(\ell)}(x_0+\xi_j(X_j-x_0)) - f^{(\ell)}(x_0)\right)(X_j-x_0)^{\ell} W_{nj}(x_0)\right| \le \\
&\frac{L}{\ell!}\sum_{j=1}^{n}|X_j-x_0|^{\beta}\left|W_{nj}(x_0)\right| \overset{(c)}{=} \frac{L}{\ell!}\sum_{j=1}^{n}\mathbf{1}_{\{|X_j-x_0|\le h\}}|X_j-x_0|^{\beta}\left|W_{nj}(x_0)\right| \le \\
&\frac{L}{\ell!}\sum_{j=1}^{n}\mathbf{1}_{\{|X_j-x_0|\le h\}}h^{\beta}\left|W_{nj}(x_0)\right| \overset{(b)}{\le} C_2 h^{\beta}.
\end{aligned}$$

Thus we obtained the already familiar bound for the MSE

$$\mathbb{E}_f(\widehat{f}(x_0) - f(x_0))^2 \le C_1\frac{1}{nh} + C_2^2 h^{2\beta}$$

and, optimizing over $h$, yields (2.1).

REMARK 2.3. Note that the role of the kernel in LP estimators is completely different from that for the kernel estimators and thus so are the assumptions. In particular, the MSE bound (2.1) does not require that the kernel has any specific order.

## Computer assignment

Implement LP regression estimator and $C_p$ cross validation bandwidth tuning, as detailed in Problem 5. Choose a smooth regression function $f : [0,1] \mapsto \mathbb{R}$ and generate a data set of size $n = 100$ as in (1.1) with $\varepsilon_j \overset{\text{i.i.d.}}{\sim} N(0,1)$ and uniform design $X_j = j/n$. Apply your code with the kernel $K(u) = (1-u^2)\mathbf{1}_{\{|u|\le 1\}}$ to generate the LP($\ell$) regression estimate for $\ell = 1$ at the design points and plot against $Y_j$'s and the true function. For the same data, add to your plot the estimates for $\ell \in \{2,5\}$. Increase the sample size to $n = 500$ and repeat the experiments. Comment on your results.

## Exercises

PROBLEM 1. Check formula (1.6).

PROBLEM 2. Specify the LP regression estimator for $\ell = 1$, derive the closed form formulas for the regression function and its derivative.

PROBLEM 3 (Exercise 1.4, [1]). Define the LP($\ell$) estimators for the derivatives $f^{(s)}(x)$, $s = 1,\dots,\ell$ by

$$\widehat{f}_n^{(s)}(x) = U^{(s)}(0)^{\top}\widehat{\theta}_n(x)h^{-s}$$

where $U^{(s)}(u)$ is the vector, whose coordinates are the $s$-th derivatives of the corresponding coordinates of $U(u)$ defined in (1.3).

(1) Prove that if $B_n(x) > 0$, then the estimator $\widehat{f}_n^{(s)}(x)$ is linear and it reproduces (derivatives of) polynomials of degree less or equal $\ell - s$.

(2) Prove that, under the assumptions similar to those made above, the maximum of MSE of $\widehat{f}_n^{(s)}(x)$ over $\Sigma(\beta,L)$ is of order $O(n^{-\frac{2(\beta-s)}{2\beta+1}})$ as $n\to\infty$ if the bandwidth $h=h_n$ is chosen optimally.

PROBLEM 4. Consider the regression model (1.1) with unknown function $f$ and let $T(f)$ be a functional of the form

$$T(f)=\int \psi(x)f(x)dx,$$

where $\psi$ is some bounded function[1]. Define the plug-in estimator

$$T(\widehat{f}_n)=\int \psi(x)\widehat{f}_n(x)dx,$$

where $\widehat{f}_n(x)$ is the LP regressor (1.5). Assuming that (LP1)-(LP3) hold, show that the kernel $K(\cdot)$ and the bandwidth $h$ can be chosen so that

$$\sup_{f:\|f\|_\infty\le M}\mathbb{E}_f\big(n^{1/2}\big(T(\widehat{f}_n)-T(f)\big)\big)^2\le C(K,M)<\infty.$$

In other words, LP plug-in estimator can be tuned to estimate $T(f)$ at rate $n^{-1/2}$, regardless of the smoothness properties of $f$.

PROBLEM 5 ($C_p$ cross-validation). In this problem we will derive the so called $C_p$ cross-validation criteria for the choice of bandwidth of any linear estimator of the form

$$\widehat{f}_h(x)=\sum_{j=1}^n Y_jW_{nj}(x,h),$$

in the regression problem (1.1).

(1) Consider the discrete version of the MISE risk

$$\mathrm{dMISE}(\widehat{f}_h,f):=\mathbb{E}_f\frac{1}{n}\sum_{j=1}^n\big(\widehat{f}_h(X_j)-f(X_j)\big)^2=:\mathbb{E}_f\|\widehat{f}_h-f\|_{2,n}^2 \tag{2.6}$$

Show that this expression is minimized by the same $h$ as

$$J(h):=\mathbb{E}_f\Big(\big\|\widehat{f}_h\big\|_{2,n}^2-\frac{2}{n}\sum_{j=1}^n\widehat{f}_h(X_j)f(X_j)\Big). \tag{2.7}$$

(2) Define the statistic

$$\widehat{T}(h)=\frac{2}{n}\sum_{j=1}^n Y_j\widehat{f}_h(X_j)-\frac{2\sigma^2}{n}\sum_{j=1}^n W_{nj}(X_j,h),$$

where $\sigma^2=\mathrm{Var}(\varepsilon_1)$ and the bandwidth variable $h$ is added to the notation for the weights in (1.8). Show that this statistic is an unbiased estimator for the second term in (2.7).

(3) Find an unbiased estimator for $J(h)$.

[1] this is a mild assumption, since the domain of $\psi$ is bounded; in particular, any continuous function is bounded, etc.

(4) Define the $C_p$-criterion

$$C_p(h) := \frac{1}{n}\sum_{j=1}^{n}\big(Y_i - \widehat{f}(X_j)\big)^2 + \frac{2\sigma^2}{n}\sum_{j=1}^{n} W_{nj}(X_j,h).$$

Show that $C_p(h) - \sigma^2$ is an unbiased estimator for the risk (2.6). Explain why the minimizer $\widehat{h} := \mathrm{argmin}_{h>0} C_p(h)$ is a reasonable data-driven choice for the bandwidth.

PROBLEM 6. ([1, Lemma 1.5, p. 41]) Following the steps below, prove that the matrix $B_n(x)$ defined in (1.7) satisfies assumption (LP1) for the uniform design $X_j = j/n$ and kernels with the property (2.5).

(1) Let $G$ be a nonnegative function, such that

$$\mathrm{Leb}\{u \in \mathbb{R} : G(u) > 0\} > 0. \tag{6a}$$

Show that the matrix

$$B = \int_{\mathbb{R}} U(x)U(x)^\top G(x)dx,$$

where $U(x)$ is defined in (1.3), is positive definite.

**Hint**: show first that $B$ is nonnegative definite and argue by contradiction, that in fact it is positive definite.

(2) Let $h := h_n$ depend on $n$, so that $h_n n \to \infty$, and define $z_j := (X_j - x)/h_n$. Assuming uniform design $X_j = j/n$, $j \in \{1,\dots,n\}$, argue that for $\Delta > 0$,

$$\frac{1}{nh_n}\sum_{j=1}^{n}\big(v^\top U(z_j)\big)^2 \mathbf{1}_{\{0<z_j<\Delta\}} \xrightarrow[n\to\infty]{} \int_0^\Delta \big(v^\top U(s)\big)^2 ds, \quad \forall x \in [0, 1-h_n\Delta],$$

$$\frac{1}{nh_n}\sum_{j=1}^{n}\big(v^\top U(z_j)\big)^2 \mathbf{1}_{\{-\Delta<z_j<0\}} \xrightarrow[n\to\infty]{} \int_0^\Delta \big(v^\top U(s)\big)^2 ds, \quad \forall x \in [\tfrac{1}{n}+h_n\Delta, 1].$$

(3) Under the assumption of the previous question and condition (2.5), prove that $B_n(x)$ satisfies (LP1), if $h_n \to 0$.

**Hint:** use the variational characterization of the least eigenvalue of a nonnegative symmetric matrix $S$

$$\lambda_{\min}(S) = \min_{v:\|v\|=1} v^\top S v.$$

## References

[1] AB Tsybakov, Introduction to Nonparametric Estimation, Springer Science & Business Media, 22 Oct 2008

[2] M.D. Cattaneo, M. Jansson, X. Ma, Simple Local Polynomial Density Estimators, Journal of the American Statistical Association, 2019, in press

DEPARTMENT OF STATISTICS, THE HEBREW UNIVERSITY, MOUNT SCOPUS, JERUSALEM 91905, ISRAEL

*E-mail address*: Pavel.Chigansky@mail.huji.ac.il