# A short notes on Nonparametric regression

Tong Zhou*

November 1, 2020

This notes is mainly based on the following materials:

- All of Nonparametric Statistics, by *Larry Wasserman*
- A Primer on Regression Splines, by *Jeffrey S. Racine*
- Lecture 3: Regression: Nonparametric Approaches, STAT 535 SML Autumn 2019, by *Yen-chi Chen*
- Lecture notes: Nonparametric Regression and Classification, Statistical Machine Learning, Spring 2017, by *Ryan Tibshirani and Larry Wasserman*

Given pairs of data $(x_1, Y_1), \cdots, (x_n, Y_n)$, in which $Y$ is called **response variable** and $x$ is called **covariate**, our goal is to relate them by:

$$Y_i = r(x_i) + \varepsilon_i, \quad \mathbf{E}[\varepsilon_i] = 0, i = 1, \cdots, n.$$

where $r()$ is called the **regression function**. In the machine learning community, the goal is to "learn" $r$ under weak assumptions. The estimator of $r$ is denoted by $\widehat{r}_n(x)$, which is often referred to as a **smoother**.

Note that at this point, the interpretations regarding this model vary depending on different assumptions imposed on the data. Let's discuss them:

1. Assume the random pair $(X, Y) \in \mathcal{P}$ and $\left(x_i, y_i\right) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \cdots, n$ are i.i.d. . The regression function is **defined** as $r(x) = \mathbf{E}[Y|X = x]$. Then $\varepsilon_i$ can be perceived as the error and equal to $\varepsilon_i = y_i - r(x_i)$. Under this setup, it is natural to get $\mathbf{E}[\varepsilon_i] = 0$ by the *Law of Iterated Expectations* and $\varepsilon_i$ are i.i.d., for any $i$. This interpretation agrees with the one in Bruce's Hansen's Econometric textbooks.

*Johns Hopkins University, tzhou11@jhu.edu

2. Assume $\varepsilon_i \perp\!\!\!\perp X_i$. This independence assumption is not free. Bear in mind that throughout the discussion, we maintain the strong assumption. Why this assumption is strong? Think about a multiple regression example, if a relevant regressor is omitted which is also correlated with some existing regressor, then this assumption fails to hold. Specifically, if $y_i = \beta X_i + \alpha Z_i + \varepsilon_i$ and $\varepsilon_i \perp\!\!\!\perp (X_i, Z_i)$. If we adhere to the regression $y_i = \beta X_i + \widetilde{\varepsilon}_i$ and $X_i$ and $Z_i$ happen to not be independent, then there is no hope $X_i \perp\!\!\!\perp \widetilde{\varepsilon}_i$.

3. **Fixed $x$ world**. $\mathbf{E}[\varepsilon_i] = 0$ and $\mathbf{E}[\varepsilon_i^2] = \sigma^2$. In this scenario, the *Fixed* inputs assumption is equivalent to the *Random* inputs with the independence assumption maintained.

We have mentioned $r$ is called a smoother. Before plunging into the formal treatment of nonparametric regression, we first define the notion of **linear smoother**.

**Definition 1.** *An estimator $\widehat{r}_n$ of $r$ is a **linear smoother** if, for each $x$, there exists a vector $\ell(x) = (\ell_1(x), \cdots, \ell_n(x))^\top$ such that*

$$\widehat{r}_n(x) = \sum_{i=1}^{n} \ell_i(x) Y_i.$$

Some related quantities are:

- **fitted values:** $\boldsymbol{r} = (\widehat{r}_n(x_1), \cdots, \widehat{r}_n(x_n))^\top$
- **Smoothing matrix / hat matrix:** $L = (\ell(x_1), \cdots, \ell(x_n))^\top$ and $L_{ij} = \ell_j(x_i)$.
- **effective kernel** for $r(x_i)$ : $\ell(x_i)^\top$ the $i$-th row of $L$ with $\sum_{i=1}^{n} \ell_i(x) = 1$.
- **effective degrees of freedom:** $\nu = \operatorname{tr}(L)$.
- $\boldsymbol{r} = LY$.

NOTE

For the regression model:

$$Y_i = r(x_i) + \varepsilon_i$$

The estimator $\widehat{r}_n$ of $r$ is to minimize the sum of squares:

$$\sum_{i=1}^{n} (Y_i - \widehat{r}_n(x_i))^2 .$$

There are two extreme cases:

1. Minimizing over *ALL linear functions* $\Longrightarrow$ **least square estimators**.
2. Minimizing over *ALL functions* $\Longrightarrow$ **a function that interpolates the data**.

There are several solutions:

1. Local regression: use a locally weighted sums of squares.

$$\sum_{i=1}^{n} w_i(x) (Y_i - a)^2 .$$

2. Penalized sums of squares :

$$M(\lambda) = \sum_{i} (Y_i - \widehat{r}_n(x_i))^2 + \lambda J(r)$$

**Example 1** (Local Regression)**.** *Suppose $x_i \in \mathbb{R}$ is scalar, and consider the regression model $Y_i = r(x_i) + \varepsilon_i$. Idea of estimating $r(x)$ is by taking a weighted average of $Y_i$, giving higher weight to those points near $x$.*

– ***Nadaraya-Watson kernel estimator:***

$$\widehat{r}_n(x) = \sum_{i=1}^{n} \ell_i(x) Y_i.$$

– *$K$ is a kernel and*

$$\ell_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}$$

– *Example* 5.26 (*boxcar kernel*:) *Suppose that* $x_i \in [a, b], i = 1, \cdots, n$. *Divide* $(a, b)$ *into* $m$ *equally spaced bins denoted by* $B_1, B_2, \cdots, B_m$. *Define* $\widehat{r}_n(x)$ *by*

$$\widehat{r}_n(x) = \frac{1}{k_j} \sum_{i:\, x_i \in B_j} Y_i, \ \text{for} \ \ x \in B_j$$

*where* $k_j$ *is the number of points in* $B_j$. *So for* $x \in B_j$, $\ell_i(x) = \frac{1}{k_j}$ *if* $x_i \in B_j$ *and* $\ell_i = 0$ *otherwise.*

**Example 2** (Penalized Regression)**.** *Goal*: *to minimize the* ***penalized sums of squares:***

$$M(\lambda) = \sum_i (Y_i - \widehat{r}_n(x_i))^2 + \lambda J(r)$$

*where* $J(r)$ *is some* ***roughness penalty****,* $\lambda$ *controls the amount of smoothing.*

1. $\lambda \longrightarrow 0$: *the interpolating function.*
2. $\lambda \longrightarrow \infty$: *the least squares line.*
3. $0 < \lambda < \infty$: *use* ***splines,*** *a special piecewise polynomial.*

*A special case on the penalty term is such that*

$$J(r) = \int |r''(x)|^2 \, \mathrm{d}x \tag{1}$$

Before diving into details, we first look at an important theorem:

**Theorem 1.** *The function* $\widehat{r}_n(x)$ *that minimizes* $M(\lambda)$ *with penalty defined in* 1 *is a natural cubic spline with knots at the data points. The estimator* $\widehat{r}_n$ *is called a* ***smoothing spline****.*

Now let us get back to some notions underlined in this theorem:

– Let $\xi_1 < \xi_2 < \cdots < \xi_p$ be set of ordered points-called **knots**.

– **cubic spline:** is a continuous function $r$ such that (i) $r$ is a cubic piecewise polynomial over each bins $(\xi_1, \xi_2), \cdots$ and (ii) $r$ has continuous first and second derivatives at the knots.

However, Theorem 1 only asserts that the minimizer of the penalty regression problems exists and belongs to the class of cubic splines. It does not provide guidance on what the cubic spline looks like. The next theorem fills a gap:

**Theorem 2.** *Let $\xi_1 < \xi_2 < \cdots < \xi_p$ be knots contained in an interval $(a, b)$. Define $h_1(x) = 1, h_2(x) = x, h_3(x) = x^2, h_4(x) = x^3, h_{j+4}(x) = (x - \xi_j)_+^3$ for $j = 1, \cdots, p$. The functions $\{h_1, \cdots, h_{k+4}\}$ form a basis for the set of cubic splines at these knots, called the* ***truncated power basis****. Thus, any cubic spline $r(x)$ with these knots can be written as*

$$r(x) = \sum_{j=1}^{k+4} \beta_j h_j(x).$$

*Remark* 1.
- The space of the 3-order splines with knots $\xi_1, \cdots, \xi_p$ has dimension $p + 4$.
- The basis of functions is called **truncated power basis**.
- More general, the $k$-order spline $r$ is a piecewise polynomial function of degree $k$ that is continuous and has continuous derivatives of orders $1, \cdots, k - 1$, at its knot points. Specifically, its corresponding *truncated power basis* is $g_1, g_2, \cdots, g_{p+k+1}$ such that

$$g_1(x) = 1, g_2(x) = x, \cdots, g_{k+1}(x) = x^k,$$
$$g_{k+1+j}(x) = (x - \xi_j)_+^k, j = 1, \cdots, p.$$

  and its dimension is $p + k + 1$.

To gain more computational edge, another more popular spline is **B-spline**, where its basis functions are defined recursively.

Let $\xi_0 = a$ and $\xi_{p+1} = b$. Define the appended $2M$ knots by:

$$\tau_1 \leqslant \tau_2 \leqslant \tau_3 \leqslant \cdots \leqslant \tau_M \leqslant \xi_0,$$

and

$$\xi_{p+1} \leqslant \tau_{k+M+1} \leqslant \cdots \leqslant \tau_{p+2M}$$

This appending is needed due to the recursive nature of the B-spline. Now for the $j$-th order of spline, the basis functions can be defined as:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \leqslant x \leqslant \tau_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

for each $i$. Next, for $m \leqslant M$, we define

$$B_{i,m}(x) = \alpha_{i,m-1} B_{i,m-1}(x) + (1 - \alpha_{i+1,m}) B_{i+1,m}(x)$$

where

$$\alpha_{i,m-1}(x) = \begin{cases} \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i}, & \text{if } \tau_{i+m-1} \neq \tau_i \\ 0, & \text{otherwise} \end{cases}$$

Now all relevant notations are squared away, the following theorem dictates the role of B-spline:

**Theorem 3.** *The function* $\{B_{i,4}, i = 1, \cdots, k+4\}$ *are a basis for the set of cubic splines. The are called the* ***B-spline basis functions****.*

Let's switch back to the original setting: $(x_1, Y_1), \cdots, (x_n, Y_n)$. Then $\widehat{r}_n$ is a natural cubic spline:

$$\widehat{r}_n(x) = \sum_{j=1}^{N} \widehat{\beta}_j B_j(x)$$

where $N = n + 4$ and $\widehat{\beta} = \left(\widehat{\beta}_1, \cdots, \widehat{\beta}_N\right)^\top$.

The minimization problem then can be expressed as

Minimization of Penalized Regression Estimator

$$\text{minimize: } \left(Y - B\beta\right)^\top \left(Y - B\beta\right) + \lambda \beta^\top \Omega \beta$$

where $B_{ij} = B_j(X_i)$ and $\Omega_{jk} = \int B_j''(x) B_k''(x)\, \mathrm{d}x$

The value of $\beta$ that minimizes the problem is:

$$\widehat{\beta} = \left(B^\top B + \lambda \Omega\right)^{-1} B^\top Y.$$

# Curse of dimensionality

If $f_0$ is $L$-Lipschitz continuous, the local polynomial estimator's error rate is $n^{-2/(2+d)}$, i.e.

$$\mathbf{E}\left[\|\widehat{f} - f_0\|_2^2\right] \lesssim n^{-2/(2+d)}.$$

It exhibits a very discouraging result: the error rate depends heavily on the dimension $d$. To be more concrete, fix a tolerance rate $\varepsilon > 0$, how large is the sample size $n$ that can ensure $n^{-2/(2+d)} \leqslant \varepsilon$ ? By simple rearrangement, we have

$$n \geqslant \varepsilon^{-(2+d)/2}.$$

That is, as $d$ is increasing, $n$ is required to exponentially increase to achieve an error bound of $\varepsilon$. Recall in the linear regression models, we have a more reasonable requirement for the sample size $n \geqslant \frac{d}{\varepsilon}$.

To make it more precise, consider a minimax problem, which asserts that we cannot hope to do better than the error rate $n^{-2/(2+d)}$ over all $L$-Lipschitz function in $d$ dimensions, denoted by $H_d(1, L)$. That is

$$\inf_{\widehat{f}} \sup_{f_0 \in H_d(1,L)} \mathbf{E}\left[\|\widehat{f} - f_0\|_2^2\right] \gtrsim n^{-2/(2+d)}$$

So to circumvent this curse, more assumptions about it are to be made. One such example is the additive model.

# Exercises

**a.**

Assume that we observe i.i.d. samples $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}, i = 1, \cdots, n$ from a model

$$y_i = f_0(x_i) + \varepsilon_i, i = 1, \cdots, n$$

where $\varepsilon_i, i = 1, \cdots, n$ are independent with $\mathbf{E}[\varepsilon_i] = 0$ and $\mathbf{E}[\varepsilon_i^2] = \sigma^2$. For simplicity, we treat the input points $x_i \in [0, 1], i = 1, \cdots, n$ as fixed, satisfying the condition

$$\mathbf{P}_n(I) \geqslant c\,|I|\,, \quad \text{for any interval } I \subset [0, 1] \text{ with } |I| \geqslant 1/n,$$

where $\mathbf{P}_n$ is the empirical distribution of the input points. Also assume that $x \in [0, 1]$. We also assume that the underlying regression function $f_0$ has a continuous, bounded derivative. Consider $\widehat{f}$, the kernel smoothing estimate with a boxcar kernel and bandwidth $h$.

1. Prove that the squared bias and variance of $\widehat{f}$, at an arbitrary point $x$, satisfy

$$\left(\mathbf{E}\left[\widehat{f}(x)\right] - f_0(x)\right)^2 \lesssim h^2 \text{ and } \mathbf{E}\left[\left(\widehat{f}(x) - \mathbf{E}[\widehat{f}(x)]\right)^2\right] \lesssim \frac{1}{nh}$$

   (Hint: use a Taylor expansion of $f_0$ around $x$. )

2. Derive the rate for the optimal choice of bandwidth $h$, and the give the error rate for the corresponding kernel smoothing estimator.

**b.**

Consider the univariate $k$-th order local polynomial regression estimate, trained on the points $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}, i = 1, \cdots, n$, which we know can be expressed as

$$\widehat{f}(x) = \sum_{i=1}^{n} w_i(x) y_i,$$

for some weights $w_i(x), i = 1, \cdots, n$.

1. Prove that at any point $x \in \mathbb{R}$, we have

$$\sum_{i=1}^{n} w_i(x) = 1 \text{ and } \sum_{i=1}^{n} w_i(x)(x_i - x)^j = 0, \text{ for } j = 1, \cdots, k.$$

2. Now assume the same model for the data as in part (a), and further assume that $x_i = 1/n, i = 1, \cdots, n$and $f_0 \in H_1(k+1, L)$ for a positive integer $k$ and a constant $L > 0$. Also assume that $x \in [0, 1]$. Take $\widehat{f}$ to be the local polynomial regression estimate of order $k$, and compute the bias of $\widehat{f}$ at an arbitrary point $x$, using the results in the last part. (Hint: use a Taylor expansion, and the results of the last question. You can use the fact that $\sum_{i=1}^{n} |w_i| \leqslant C$ for some constant $C$ that does not depend on $n$ or $h$.) What do you conclude about the bias of local polynomial regression, compared that of kernel regression?

3. Why don't we just keep increasing the polynomial order $k$ without end? You can answer this either with some theory, or a quick simulation. (Hint: consider the variance of $\widehat{f}$. You can use the fact that $\sum_{i=1}^{n} w_i^2(x)$ is an increasing function of $k$.)

**C.**

Suppose that, in backfitting, we choose our univariate smoother just to be standard univariate linear regression. Prove that backfitting converges in one pass, and the resulting estimate is just standard multivariate linear regression.