A note on nonparametric additive models

Yen-Chi Chen
University of Washington
September 19, 2024

All nonparametric regression suffers from the curse of dimensionality; namely, when the number of covariates d is large, the convergence rate could be extremely slow. For instance, in the kernel regression, the optimal rate under a standard smoothness (2-Hölder) condition is $O_P\left(n^{-\frac{4}{4+d}}\right)$. When d is greater than 6, this rate is very slow.

To deal with this problem, a common solution is the additive model. Namely, we assume that the regression model

$$\mathbb{E}(Y|X=x) \equiv m(x) = \mu_0 + \mu_1(x_1) + \cdots + \mu_d(x_d) \tag{1}$$

with the condition that

$$\mathbb{E}(m_j(X_j)) = \int m_j(x_j) p_j(x_j) dx_j = 0, \qquad j = 1, 2, \cdots, d \tag{2}$$

to avoid identification problem. Note that $p_j(x_j)$ is the marginal PDF of X_j. This is called the additive model.

Here we will introduce three common methods for estimating the additive model.

1 Direct approach

From equations (1) and (2), we immediately have the following result:

$$\mathbb{E}(m(x_1, X_2, \cdots, X_d)) = m_0 + \mu_1(x_1) + \sum_{j=2}^{d} \mathbb{E}(\mu_j(X_j)) = m_0 + \mu_1(x_1).$$

The same result holds for any $\mu_j(x_j)$. Note that $m_0 = \mathbb{E}(Y)$ can be estimated by the simple sample mean $\bar{Y}_n$. Thus, all we need is a multivariate regression estimator $\widehat{m}(x)$ and then construct the estimator

$$\widehat{\mu}_1(x_1) = -\bar{Y}_n + \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(x_1, X_{i,2}, \cdots, X_{i,d}).$$

A similar idea can be applied to other $\widehat{m}_j(x_j)$.

However, this idea may not give an estimator with a fast convergence rate because we are still estimating the full-dimensional regression model $\widehat{m}$. To obtain a fast rate, we consider a 'partial' local polynomial regression. Let

$$\left(\widehat{\alpha}_1(x), \widehat{\beta}_1(x)\right) = \mathsf{argmin}_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \beta(X_{i1} - x_1))^2 K\left(\frac{X_{i1} - x_1}{h}\right) \prod_{j\neq 1}^{d} K\left(\frac{X_{ij} - x_j}{b}\right),$$

where h, b are smoothing bandwidth that may not necessarily be the same. Note that the above local linear model is linear only in x_1 but the kernel (localization) is on *all* variables. The constant term $\widehat{\alpha}_1(x)$ is an estimator of the regression model. To obtain the estimator $\widehat{\mu}_1(x_1)$, we average out other variables:

$$\widehat{\mu}_1(x_1) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}_1(x_1, X_{i2}, \cdots, X_{id}). \tag{3}$$

We can apply the same idea to other coordinates. It can be shown that estimator in equation (3) has a convergence rate $O(h^2) + O_P\left(\sqrt{\frac{1}{nh}}\right)$, which can recover the convergence rate to $n^{-4/5}$; see the following paper[1]:

> [FHM1998] Fan, J., Härdle, W., & Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. The Annals of Statistics, 26(3), 943-971.

We provide a high-level derivation on the convergence rate in Section 4.

2 Least square approach

A second approach to the additive model is the least square method. The high-level idea is that we want to construct estimators $\widehat{\mu}_1, \cdots, \widehat{\mu}_d$ from the minimizing the following criterion

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{d} \mu_j(X_{ij})\right)^2.$$

While this minimization could be challenge, we may restrict our model to a particular form such as the orthonormal basis or spline (with penalization on the smoothness) to make it easier.

Suppose that each $X_j \in [0,1]$. Let $\{\phi_\ell(z) : \ell = 1, \cdots, \}$ be an orthonormal basis (e.g., cosine basis). We then consider M basis functions $\phi_1(z), \cdots, \phi_M(z)$ and approximate each function

$$\mu_j(x_j) \approx \sum_{\ell=1}^{M} \theta_{j\ell}\phi_\ell(x_j).$$

All we need is to estimate the coefficients $\theta \in \mathbb{R}^{d\times M}$. Under the least-square criterion, we may estimate the coefficients by

$$\widehat{\theta} = \mathsf{argmin}_\theta \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{d}\sum_{\ell=1}^{M} \theta_{j\ell}\cdot\phi_\ell(X_{ij})\right)^2.$$

The estimator

$$\widehat{\mu}_j(x_j) = \sum_{\ell=1}^{M} \widehat{\theta}_{j\ell}\cdot\phi_\ell(x_j).$$

[1] A caveat is that we still need $nhb^{d-1} \to \infty$ and $b/h \to 0$. To obtain the optimal rate $h \asymp n^{-1/5}$, we need $d < 5$, so there is still a restriction on the dimension.

Under the regular smoothness (2-Soblev), the bias will be $O(M^{-2})$ and the variance is $O(Md/n)$, so the optimal rate will be $O(d \cdot n^{-4/5})$ with $M \asymp n^{1/5}$, which does not suffer too much from the curse of dimensionality.

The above method has a limitation that the asymptotic distribution is difficult to characterize. To resolve this problem, people recommend to perform an additional step that for each j, we compute a pseudo-outcome

$$\widehat{Y}_{ij} = Y_i - \bar{Y}_n - \sum_{k \neq j} \widehat{\mu}_k(X_{ik})$$

by leaving out the j-th coordinate. Then we use a marginal model of regressing $\widehat{Y}_{ij}$ against X_{ij} such as a kernel regression:

$$\widetilde{\mu}_j(x_j) = \frac{\sum_{i=1}^n K\left(\frac{X_{ij}-x_j}{h}\right)\widehat{Y}_{ij}}{\sum_{i=1}^n K\left(\frac{X_{ij}-x_j}{h}\right)}.$$

The estimator $\widetilde{\mu}_j(x_j)$ has a nice asymptotic distribution (asymptotically normal).

See the following papers for the use of this idea

1. Wang, L., & Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model.
2. Horowitz, J. L., & Mammen, E. (2004). Nonparametric estimation of an additive model with a link function.

3 Backfitting approach

The backfitting is perhaps the most popular method for the additive model. Note that the additive model in equation (1) can be written as

$$Y = \mu_0 + \mu_1(X_1) + \cdots + \mu_d(X_d) + \varepsilon.$$

Now we take conditional expectation $\mathbb{E}(\cdot|X_j = x_j)$ in both sides, leading to

$$\mathbb{E}(Y|X_j = x_j) = \mu_0 + \mu_j(x_j) + \sum_{k \neq j} \mathbb{E}(\mu_k(X_k)|X_j = x_j).$$

By rearrangements and using the fact that $\mu_0 = \mathbb{E}(Y)$,

$$\begin{aligned}\mu_j(x_j) &= \mathbb{E}(Y|X_j = x_j) - \mathbb{E}(Y) - \sum_{k \neq j} \mathbb{E}(\mu_k(X_k)|X_j = x_j) \\ &= \mathbb{E}(Y|X_j = x_j) - \mathbb{E}(Y) - \sum_{k \neq j} \int \mu_k(x_k) p(x_k|x_j) dx_k.\end{aligned} \tag{4}$$

Equation (4) is the famous backfitting equation.

The function $\mathbb{E}(Y|X_j = x_j)$ can be easily estimated by any marginal nonparametric regression model and $E(Y)$ can be estimated by the simple sample mean $\bar{Y}_n$. Thus, a good estimator $\widehat{mu}_j(x_j)$ should satisfies the following empirical equation

$$\widehat{\mu}_j(x_j) = \widehat{m}_j(x_j) - \bar{Y}_n - \sum_{k\neq j} \int \widehat{\mu}_k(x_k)\widehat{p}(x_k|x_j)dx_k, \tag{5}$$

where $\widehat{m}_j(x_j)$ is an estimator of the marginal model $\mathbb{E}(Y|X_j = x_j)$ and $\widehat{p}(x_k|x_j)$ is the conditional PDF estimator. Our goal is to find estimators solving equation (5).

Numerically, the backfitting method is the following iterative procedure:

1. Start with initial estimates
$$\widehat{\mu}_j^{(0)}(x_j), \qquad j = 1, \cdots, d.$$
2. For $t = 1, \cdots$, do the following until a stopping criterion is met:
   (a) For $j = 1, \cdots, d$, do:
$$\widehat{\mu}_j^{(t)}(x_j) = \widehat{m}_j(x_j) - \bar{Y}_n - \sum_{k<j} \int \widehat{\mu}_k^{(t)}(x_k)\widehat{p}(x_k|x_j)dx_k + \sum_{k>j} \widehat{\mu}_k^{(t-1)}(x_k)\widehat{p}(x_k|x_j)dx_k.$$

Namely, we sequentially update the estimator $\widehat{\mu}_j$ according to equation (5).

Theoretical properties of the backfitting method can be found in the following paper:

> Mammen, E., Linton, O., & Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. The Annals of Statistics, 27(5), 1443-1490.

A very common conditional PDF estimator is the KDE:

$$\widehat{p}(x_k|x_j) = \frac{\sum_{i=1}^n K\left(\frac{X_{ik}-x_k}{h}\right) K\left(\frac{X_{ij}-x_j}{h}\right)}{h \cdot \sum_{i=1}^n K\left(\frac{X_{ij}-x_j}{h}\right)}.$$

Note that we may use a kernel CDF approach to replace the PDF estimator in equation (5) in the sense that $\widehat{p}(x_k|x_j)dx_k$ can be replaced by $d\widehat{P}(x_k|x_j)$, where

$$\widehat{P}(x_k|x_j) = \frac{\sum_{i=1}^n I(X_{ik} \leq x_k) K\left(\frac{X_{ij}-x_j}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_{ij}-x_j}{h}\right)}.$$

With this,

$$\int \widehat{\mu}_k(x_k)d\widehat{P}(x_k|x_j) = \sum_{i=1}^n W_{ji}(x_j) \cdot \widehat{\mu}_k(X_{ik}),$$

where $W_{ji}(x_j) \geq 0, \sum_{i=1}^n W_{ji}(x_j) = 1$, and

$$W_{ji}(x_j) = \frac{K\left(\frac{X_{ij}-x_j}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_{ij}-x_j}{h}\right)}$$

which is the kernel weight of j-th coordinate for each observation. Thus, backfitting equation can be reexpressed as

$$\widehat{\mu}_j(x_j) = \widehat{m}_j(x_j) - \bar{Y}_n - \sum_{k\neq j}\sum_{i=1}^{n} W_{ij}(x_j)\cdot\widehat{\mu}_k(X_{ik}).$$

4 A high-level idea of the rate in the direct approach

Here we illustrate the high-level idea on how the direct approach in Section 1 in the additive model can improve the convergence rate. The original work in [FHM1998] is on local polynomial regression and the derivation is a lot more involved. To simplify the problem, we use the kernel regression as an example.

Suppose $X \in \mathbb{R}^2$ and let

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_{i1}-x_1}{h_1}\right) K\left(\frac{X_{i2}-x_2}{h_2}\right)}{\sum_{i=1}^{n} K\left(\frac{X_{i1}-x_1}{h_1}\right) K\left(\frac{X_{i2}-x_2}{h_2}\right)}$$

be the kernel regression estimator. When using the direct approach, the estimator of the first component $\mu_1(x_1)$ will be

$$\widehat{\mu}_1(x_1) = -\bar{Y}_n + \frac{1}{n}\sum_{i=1}^{n}\widehat{m}(x_1, X_{i,2}) = -\bar{Y}_n + \int \widehat{m}(x_1,x_2)d\widehat{P}(x_2),$$

where $\widehat{P}(x_2) = \frac{1}{n}\sum_{i=1}^{n} I(X_{i2} \leq x_2)$ is the empirical distribution.

Clearly, the convergence rate of $\widehat{\mu}_1(x_1)$ is dominated by the rate in the second term $\frac{1}{n}\sum_{i=1}^{n}\widehat{m}(x_1, X_{i,2})$. So we focus on deriving its rate.

Using the fact that the denominator of $\widehat{m}(x)$ is the 2-D KDE, we have the following approximation of the kernel regression:

$$\begin{aligned}
\widehat{m}(x) &= \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_{i1}-x_1}{h_1}\right) K\left(\frac{X_{i2}-x_2}{h_2}\right)}{\sum_{i=1}^{n} K\left(\frac{X_{i1}-x_1}{h_1}\right) K\left(\frac{X_{i2}-x_2}{h_2}\right)} \\
&= \frac{\frac{1}{nh_1h_2}\sum_{i=1}^{n} Y_i K\left(\frac{X_{i1}-x_1}{h_1}\right) K\left(\frac{X_{i2}-x_2}{h_2}\right)}{\frac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{X_{i1}-x_1}{h_1}\right) K\left(\frac{X_{i2}-x_2}{h_2}\right)} \\
&= \frac{R_n(x_1,x_2)}{\widehat{p}_{h_1,h_2}(x_1,x_2)} \\
&\approx \frac{R_n(x_1,x_2)}{p(x_1,x_2)} - \frac{\bar{R}(x_1,x_2)}{p(x_1,x_2)}\frac{\widehat{p}_{h_1,h_2}(x_1,x_2) - p(x_1,x_2)}{p(x_1,x_2)},
\end{aligned}$$

where $R_n(x_1,x_2) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} Y_i K\left(\frac{X_{i1}-x_1}{h_1}\right) K\left(\frac{X_{i2}-x_2}{h_2}\right)$ and $\bar{R}(x_1,x_2) = \int yp(y,x_1,x_2)dy$ is the asymptotic limit of R_n and $p(x_1,x_2)$ is the joint PDF and $p_{h_1,h_2}(x_1,x_2)$ is the 2-D KDE.

Applying this into $\widehat{\mu}_1(x_1)$, we obtain

$$\begin{aligned}\widehat{\mu}_1(x_1) &= \int \widehat{m}(x_1,x_2)d\widehat{P}(x_2) \\ &\approx \underbrace{\int \frac{R_n(x_1,x_2)}{p(x_1,x_2)}d\widehat{P}(x_2)}_{(I)} - \underbrace{\int \frac{\bar{R}(x_1,x_2)}{p(x_1,x_2)}\frac{\widehat{p}_{h_1,h_2}(x_1,x_2)-p(x_1,x_2)}{p(x_1,x_2)}d\widehat{P}(x_2)}_{(II)}.\end{aligned}$$

Clearly, the bias in both (I) and (II) will be $O(h_1^2+h_2^2)$. So we now focus on the variance/stochastic variation in both terms.

Variance in (I). A direct calculation shows that

$$\begin{aligned}(I) &= \int \frac{R_n(x_1,x_2)}{p(x_1,x_2)}d\widehat{P}(x_2) \\ &= \frac{1}{nh_1h_2}\sum_{i=1}^n Y_i K\left(\frac{X_{i1}-x_1}{h_1}\right)\int K\left(\frac{X_{i2}-x_2}{h_2}\right)/p(x_1,x_2)d\widehat{P}(x_2) \\ &= \frac{1}{nh_1}\sum_{i=1}^n Y_i K\left(\frac{X_{i1}-x_1}{h_1}\right)\frac{1}{nh_2}\sum_{j=1}^n K\left(\frac{X_{i2}-X_{j2}}{h_2}\right)/p(x_1,X_{j2}).\end{aligned}$$

The quantity $\frac{1}{nh_2}\sum_{j=1}^n K\left(\frac{X_{i2}-X_{j2}}{h_2}\right)/p(x_1,X_{j2})$ is essentially a 1D weighted KDE centered at X_{i2} with a weight $\frac{1}{p(x_1,X_{j2})}$ and asymptotically,

$$\frac{1}{nh_2}\sum_{j=1}^n K\left(\frac{X_{i2}-X_{j2}}{h_2}\right)/p(x_1,X_{j2}) = \frac{p(X_{i2})}{p(x_1,X_{i2})}+O(h_2^2)+O_P\left(\sqrt{\frac{1}{nh_2}}\right)\approx\frac{1}{p(x_1|X_{i2})}.$$

Thus,

$$(I)\approx\frac{1}{nh_1}\sum_{i=1}^n\frac{Y_i}{p(x_1|X_{i2})}K\left(\frac{X_{i1}-x_1}{h_1}\right).$$

Clearly, the variance of (I) will be of the order of $O(\frac{1}{nh_1})$, which is the desired result.

Variance in (II). The variance of the second term can be derived from essentially the same approach. We now focus only on $\widehat{p}_{h_1,h_2}(x_1,x_2)$ since the other quantity is non-random.

$$\begin{aligned}(II') &= \int \frac{\bar{R}(x_1,x_2)}{p(x_1,x_2)}\frac{\widehat{p}_{h_1,h_2}(x_1,x_2)}{p(x_1,x_2)}d\widehat{P}(x_2) \\ &= \frac{1}{nh_1h_2}\sum_{i=1}^n K\left(\frac{X_{i1}-x_1}{h_1}\right)\int\frac{\bar{R}(x_1,x_2)}{p^2(x_1,x_2)}K\left(\frac{X_{i2}-x_2}{h_2}\right)d\widehat{P}(x_2) \\ &= \frac{1}{nh_1}\sum_{i=1}^n K\left(\frac{X_{i1}-x_1}{h_1}\right)\frac{1}{nh_2}\sum_{j=1}^n\frac{\bar{R}(x_1,X_{j2})}{p^2(x_1,X_{j2})}K\left(\frac{X_{i2}-X_{j2}}{h_2}\right) \\ &\approx \frac{1}{nh_1}\sum_{i=1}^n K\left(\frac{X_{i1}-x_1}{h_1}\right)\cdot\frac{\bar{R}(x_1,X_{i2})p(X_{i2})}{p^2(x_1,X_{i2})}.\end{aligned}$$

This term clearly has an asymptotic variance of the order of $O(\frac{1}{nh_1})$.

Finally, using the fact that $\mathsf{Var}(X+Y) \leq 2\mathsf{Var}(X)+2\mathsf{Var}(Y)$, we conclude that the variance of $\widehat{\mu}_1(x_1)$ is of the order of $O(\frac{1}{nh_1})$.

Formally, the rate should be written as

$$\widehat{\mu}_1(x_1) - \mu_1(x_1) = O(h_1^2) + O(h_2^2) + O_P\left(\sqrt{\frac{1}{nh_1}}\right)$$

when $h_1 \to 0, h_2 \to 0, nh_1h_2 \to \infty$. We still need $nh_1h_2 \to \infty$ to ensure the 2-D KDE can approximate $p(x_1,x_2)$ well. In some paper, we add an additional condition $\frac{h_2}{h_1} \to 0$, so that we can drop $O(h_2^2)$ in the rate, making it $O(h_1^2) + O_P\left(\sqrt{\frac{1}{nh_1}}\right)$, the usual 1-D rate.

Remark.

1. The key to improve the rate is the integral $\int \widehat{m}(x_1,x_2)d\widehat{P}(x_2)$ that removes the effect of the second variable. This integral converts the kernel into a weight at each observation. Without this integral, we will still be in the usual 2D rate.

2. While we only consider $d = 2$, the whole derivation remains the same when we have more variables. Suppose we have d variables, then we still have

$$\widehat{\mu}_1(x_1) - \mu_1(x_1) = O\left(\sum_{\ell=1}^{d} h_\ell^2\right) + O_P\left(\sqrt{\frac{1}{nh_1}}\right)$$

   under the condition that $nh_1h_2\cdots h_d \to \infty$.

3. In fact, this derivation holds if we are considering the additive model in the form of

$$m(x) = \mu_1(x_1) + \eta(x_2,\cdots,x_d).$$

   We will still obtain the same convergence rate using the estimator $\widehat{\mu}_1(x_1)$! In [FHM1998], they even consider a more general setup that

$$m(x) = \mu_1(x_1) + \mu_2(x_2)$$

   with $x_1 \in \mathbb{R}^p$ and $x_2 \in \mathbb{R}^d$. Let h be the smoothing bandwidth for x_1 and b be the smoothing bandwidth for x_2. The convergence rate will be

$$\widehat{\mu}_1(x_1) - \mu_1(x_1) = O\left(h^2 + b^2\right) + O_P\left(\sqrt{\frac{1}{nh^p}}\right),$$

   under the constraint $nh^pb^d \to \infty$.