06 Additive & Partially Linear Models

We saw previously how higher-order Holder smoothness can help ameliorate the curse of dimensionality, e.g., by reducing error rates from $n^{-1/(2+d)}$ to $n^{-1/(2+d/s)}$. Here we consider other kinds of structure that can yield faster rates of convergence in regression problems, namely additivity and partial linearity.

1 Additive models

In an additive model, we assume the regression function decomposes into a sum of individual functions of each feature (plus an intercept), i.e.,

$$\mathbb{E}(Y \mid X) \equiv \mu(X) = \mu_0 + \sum_{j=1}^{d} \mu_j(X_j).$$

We typically constrain

$$\mathbb{E}\{\mu_j(X_j)\} = 0, \quad j = 1, ..., d$$

so that the intercept is identifiable (note otherwise we could shift each μ_j to $\mu_j + \alpha_j$ and change the intercept to $\mu_0 - \sum_j \alpha_j$ without changing the model). Importantly, this model assumes away interactions. More specifically, an additive model ensures that when we look at slices of $\mu(x)$ at any $x_{-j} = (x_1, ..., x_d) \setminus x_j$ across different x_j values, then the shape of the function stays the same regardless of x_{-j}, with x_{-j} only changing the intercept (i.e., shifting the function up or down). Similarly, the difference

$$\mu(x_1, ..., x_j, ..., x_d) - \mu(x_1, ..., x'_j, ..., x_d) = \mu(x_j) - \mu(x'_j)$$

only depends on (x_j, x'_j), and not on any values of other features x_k, for $k \neq j$. This makes it so that all of the information about the function $\mu(x)$ can be conveyed via the d simple plots of x_j against $\mu_j(x_j)$.

Additive models are very popular in practice, and they come with a number of important advantages. First, they are interpretable. As mentioned above, the entire function $\mu(x)$ can be conveyed with d single plots, as opposed to trying to visualize a d-dimensional surface. Second, as we will see in the next section, there is a simple but powerful algorithm for fitting these models, which allows one to utilize generic regression estimators and only requires one-dimensional regressions at each step. Third, additive models arguably represent a nice middle ground / compromise between, say, parametric models at one extreme and very non-smooth or unstructured nonparametric models at the other extreme.

Of course, there is no free lunch in statistics/ML, and additive models also come with some disadvantages. For example, ruling out interactions a priori risks bias. In some settings these interactions may be substantively important and disregarding them could lead to improper conclusions. We will also see how there can be some gaps between theory and practice when it comes to analyzing a popular algorithm for fitting additive models.

1.1 Backfitting

There is a very popular and conceptually simple algorithm for fitting additive models, known as backfitting. It is by far the most common procedure used in most software for computing additive model estimates. The main heuristic is as follows. Suppose we knew all $\mu_j(x)$, $j = 1, ..., d$, except for $j = 1$. Then we could regress $Y - \mu_0 - \sum_{j>1} \mu_j(X_j)$ on X_1 to estimate $\mu_1(x_1)$. The same holds for all $j > 1$ as well. This motivates the following algorithm:

1. Set initial estimates $\widehat{\mu}_0 = \overline{Y}$, $\widehat{\mu}_j(x_j)$, $j = 1, ..., d$.
2. $\ell = 1$: Regress $Y - \widehat{\mu}_0 - \sum_{j \neq 1} \widehat{\mu}_j(X_j)$ on X_1 to estimate $\widetilde{\mu}_1(x_1)$, then center by setting $\widehat{\mu}_1(x_1) = \widetilde{\mu}_1(x_1) - \frac{1}{n} \sum_i \widetilde{\mu}_1(x_{1i})$.
3. Repeat step (2) for $\ell = 2, ..., d$.
4. Repeat step (2–3) until the functions $\widehat{\mu}_j$ do not change by more than some small ϵ, in some metric.

The final estimate then takes the form

$$\widehat{\mu}(x) = \widehat{\mu}_0 + \sum_{j=1}^{d} \widehat{\mu}_j(x_j).$$

Some of the analyses of backfitting get a bit complicated due to its iterative construction.

1.2 Series estimation

Here we discuss a simple additive model estimator based on series estimation, where we can use essentially all the same analysis tools from the non-additive model setting.

The estimator is exactly the same as the series estimator we discussed previously, with an important exception: in the basis expansions we omit all interaction terms. Specifically, we construct a series estimator

$$\widehat{\mu}(x) = b(x)^{\mathrm{T}} \mathbb{P}_n \{b(X) b(X)^{\mathrm{T}}\}^{-1} \mathbb{P}_n \{b(X) Y\}$$

that looks very similar to before, except now we take

$$b(X) = \{\widetilde{b}(X_1)^{\mathrm{T}}, ..., \widetilde{b}(X_d)^{\mathrm{T}}\}^{\mathrm{T}}$$

where $\widetilde{b}(x)$ is a univariate basis expansion (e.g., using a polynomial or legendre basis) of order k/d, so that the total number of basis terms in $b(x)$ is $d(k/d) = k$. Recall creating these univariate basis expansions was the first step in constructing our standard series estimator, but now we omit the tensor product step where we also include all interactions. This makes the estimator much more practically feasible and easy to compute.

The idea now is that the approximation error term in the analysis will come from approximating d one-dimensional functions with bases of order k/d, rather than approximating one d-dimensional function with a basis of order k; the estimation error arguments are all the same as before.

Theorem 1. *Assume:*

1. *var*$(Y \mid X = x) \le \sigma^2$,
2. $\mu(x)$ *is additive, with each component* s*-smooth,*
3. $\sup_{f \in \mathcal{H}(s)} \inf_\theta \|\widetilde{b}^{\mathrm{T}}\theta - f\| \lesssim \ell^{-s}$ *for* $\dim(\widetilde{b}) = \ell$,
4. *the eigenvalues of* $Q \equiv \mathbb{E}\{b(X)b(X)^{\mathrm{T}}\}$ *are bounded above and below away from zero.*

Then as long as $\sup_x \|b(x)\|^2 \log k/n \to 0$ *the series estimator* $\widehat{\mu}$ *satisfies*

$$\|\widehat{\mu} - \mu\| = O_{\mathbb{P}}\left(k^{-s} + \sqrt{\frac{k}{n}}\right).$$

Note these are the same assumptions that we used before, except now we are using the additivity of $\mu(x)$. Assumption 3 is just saying that the basis has the right approximation error for Holder functions, as usual, but now we are approximating 1d functions.

Proof. Now we describe the basics of the proof, focusing on approximation error as that is the crucial piece. As before, define

$$\beta = \arg\min_\theta \mathbb{E}\left[\left\{Y - b(X)^{\mathrm{T}}\theta\right\}^2\right] = \mathbb{E}\{b(X)b(X)^{\mathrm{T}}\}^{-1}\mathbb{E}\{b(X)Y\}$$

as the population projection of μ onto b, and note we can decompose the error as

$$\|\widehat{\mu} - \mu\| \le \|b^{\mathrm{T}}(\widehat{\beta} - \beta)\| + \|b^{\mathrm{T}}\beta - \mu\|.$$

The first term is estimation error and the second is approximation error. Note the approximation error satisfies

$$\begin{aligned}\|b^{\mathrm{T}}\beta - \mu\| &= \inf_\theta \|b^{\mathrm{T}}\theta - \mu\| = \inf_\theta \left\|\sum_j \left(b_j^{\mathrm{T}}\theta_j - \mu_j\right)\right\| \\ &\le \inf_\theta \sum_j \|b_j^{\mathrm{T}}\theta_j - \mu_j\| \le d \inf_\theta \max_j \|b_j^{\mathrm{T}}\theta_j - \mu_j\| \lesssim d(k/d)^{-s} = d^{1+s}k^{-s}\end{aligned}$$

where the first inequality follows by Assumption 2 and the last by Assumption 3. Therefore it only remains to bound the estimation error $\|b^{\mathrm{T}}(\widehat{\beta} - \beta)\|$. This follows essentially the same logic as in the non-additive model case. □

Note the above rate of convergence mimics of that of the $d = 1$ setting, i.e., when we balance (squared) bias and variance we take

$$k \sim n^{1/(2s+d)}$$

which gives the one-dimensional rate

$$n^{-1/(2+1/s)}.$$

This rate can be seen to be minimax optimal, disregarding dependence on dimension d. In fact, getting the right dependence on dimension d can be a bit more nuanced.

2 Partially linear model

Another common way to impose some non-trivial structure is through a partially linear model

$$\mathbb{E}(Y \mid X, A) \equiv \mu(X, A) = \beta^{\mathrm{T}} A + \mu_0(X).$$

Here we partition the features into two components X and A, and the hope is we know the regression function depends on A in some simple parametric way, while the dependence on X can be essentially arbitrarily complex. This is an example of a *semiparametric model* since the dependence via features A is known up to a finite-dimensional parameter $\beta \in \mathbb{R}^p$, while the dependence through features X comes through the function μ_0 and is left unspecified and infinite-dimensional. Note also that this model is neither stronger nor weaker than the additive model, since stronger assumptions are made on the A component, but weaker assumptions are made on the X component.

The partially linear model is particularly useful when the subcomponent of features A represents some variable(s) whose relationship with outcome is particularly well-understood. This comes up often in causal inference, econometrics, and conditional independence testing, but can also arise in standard semiparametric regression problems, depending on context. Semiparametric models bring with them a lot of interesting nuances and mysteries; some useful references include Chapter 25 of van der Vaart [2000], Tsiatis [2006], Kennedy [2022], etc.

Note this model can equivalently be written as

$$Y = \beta^{\mathrm{T}} A + \mu_0(X) + \epsilon \tag{1}$$

where $\mathbb{E}(\epsilon \mid X, A) = 0$. One classic motivating example comes from causal inference where $A \in \{0, 1\}$ is a binary treatment, and so

$$\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0) = \beta,$$

i.e., the difference in expected outcomes among those with the same features X under treatment versus control does not actually depend on features X, and is simply a constant.

2.1 Plug-in estimation

A first stab at estimating β might be motivated as follows. Rearranging (1) to read $(Y - \mu_0(X)) = \beta A + \epsilon$, one might consider regressing $Y - \widehat{\mu}_0(X)$ on A to give

$$\widehat{\beta} = \frac{\widehat{\mathrm{cov}}(A, Y - \widehat{\mu}_0)}{\widehat{\mathrm{var}}(A)} = \frac{\mathbb{P}_n\{A(Y - \widehat{\mu}_0)\}}{\widehat{\mathrm{var}}(A)}$$

where we omit the centering of A (i.e., assuming $\mathbb{P}_n(A) = 0$) for simplicity. Let D^n denote training data on which $\widehat{\mu}_0$ is estimated, and consider the bias of the numerator estimator

$$\mathbb{E}\Big[\mathbb{P}_n\{A(Y - \widehat{\mu}_0)\} \mid D^n\Big] - \mathrm{cov}(A, Y - \mu_0) = \mathbb{E}\Big[\pi(X)\{\mu_0(X) - \widehat{\mu}_0(X)\}\Big]$$

$$\implies \mathbb{E}\Big[\mathbb{P}_n\{A(Y - \widehat{\mu}_0)\}\Big] - \mathrm{cov}(A, Y - \mu_0) = \int \pi(x)\mathbb{E}\Big\{\widehat{\mu}_0(x) - \mu_0(x)\Big\}\, d\mathbb{P}(x)$$

In general this bias would be of the same order as the error in estimating μ_0. This means that with this analysis of this estimator, one would get hit by slow nonparametric rates for estimating μ_0 even when estimating the finite-dimensional parameter β.

2.2 Bias-corrected regression on residuals

Are the above slow rates the best we can do for estimating β? In fact, no, they are not. Robinson [1988] proposed the following approach.

First note that, letting

$$\pi(X) = \mathbb{E}(A \mid X)$$
$$m(X) = \mathbb{E}(Y \mid X) = \mathbb{E}\{\mathbb{E}(Y \mid X, A) \mid X\} = \beta^{\mathrm{T}}\pi(X) + \mu_0(X)$$

then another way to write the model is

$$Y - m(X) = \beta\{A - \pi(X)\} + \epsilon$$

Therefore we can also write

$$\beta = \frac{\mathrm{cov}\{(A-\pi)(Y-m)\}}{\mathrm{var}\{(A-\pi)\}} = \frac{\mathbb{E}\{\mathrm{cov}(A, Y \mid X)\}}{\mathbb{E}\{\mathrm{var}(A \mid X)\}}.$$

It is also important to note that we can define β in this way regardless of whether the partially linear model actually holds or not. This gives a model-agnostic interpretation of β, as an X-adjusted association between A and Y. Further, this formulation suggests another estimator

$$\widehat{\beta} = \frac{\widehat{\mathrm{cov}}\{(A-\widehat{\pi}), (Y-\widehat{m})\}}{\widehat{\mathrm{var}}\{(A-\widehat{\pi})\}}$$
$$= \frac{\mathbb{P}_n\{(A-\widehat{\pi})(Y-\widehat{m})\}}{\mathbb{P}_n\{(A-\widehat{\pi})^2\}} = \frac{\widehat{\beta}_{num}}{\widehat{\beta}_{den}}$$

which has some surprising and special properties, as detailed by the result below.

Theorem 2. *Let $\varphi = (A-\pi)(Y-m)$ and let $\widehat{\varphi}$ be the estimated version. Then:*

$$\mathbb{E}(\widehat{\beta}_{num} - \beta_{num} \mid D^n) \leq \|\widehat{\pi} - \pi\|\|\widehat{m} - m\|$$
$$\sqrt{var}(\widehat{\beta}_{num} \mid D^n) \leq \frac{\sqrt{var}(\varphi) + \|\widehat{\varphi} - \varphi\|}{\sqrt{n}}$$

Proof. Consider the numerator expectation

$$\mathbb{E}\Big[\mathbb{P}_n\{(A-\widehat{\pi})(Y-\widehat{m})\} \mid D^n\Big] = \mathbb{E}\Big\{(A-\pi+\pi-\widehat{\pi})(Y-m+m-\widehat{m}) \mid D^n\Big\}$$
$$= \mathbb{E}\{(A-\pi)(Y-m)\} + \int \Big\{\pi(x) - \widehat{\pi}(x)\Big\}\Big\{m(x) - \widehat{m}(x)\Big\} d\mathbb{P}(x)$$

and so note the bias is bounded above by

$$\|\widehat{\pi} - \pi\|\|\widehat{m} - m\|$$

For n times the variance we have

$$\mathrm{var}\{(A-\widehat{\pi})(Y-\widehat{m}) \mid D^n\} = \mathrm{var}\{\varphi + (\widehat{\varphi} - \varphi) \mid D^n\}$$
$$= \mathrm{var}(\varphi) + \mathrm{var}(\widehat{\varphi} - \varphi \mid D^n) - 2\mathrm{cov}(\varphi, \widehat{\varphi} - \varphi \mid D^n)$$
$$\leq \mathrm{var}(\varphi) + \|\widehat{\varphi} - \varphi\|^2 + 2\sqrt{\mathrm{var}(\varphi)\mathrm{var}(\widehat{\varphi} - \varphi \mid D^n)}$$
$$\leq \Big(\sqrt{\mathrm{var}}(\varphi) + \|\widehat{\varphi} - \varphi\|\Big)^2$$

□

The crucial thing about this estimator and the result shown above is that its bias is a *product* of errors, rather than simply inheriting the error from the regression estimator $\widehat{\mu}_0$, which is what occured for the simple plug-in estimator described in 2.1 This product error phenomenon is referred to as semiparametric, or double robust, or targeted, or more recently, double machine learning. The important point is that one can attain a faster rate for estimating β than for estimating the more complicated regression function μ_0.

The construction of this estimator may have seemed ad hoc, but there is actually a deep theory of underlying semiparametrics that provides guidance on how to construct estimators with these kinds of properties. We will return to these topics later in the semester when we talk about causal inference.

References

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. 2022.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.