04 Series Estimators

1 Local, global, penalized modeling

Györfi et al. [2002] distinguish three types of regression estimators: local averaging/modeling, global modeling, and penalized modeling. Prime examples of local modeling include the methods we have discussed so far: kNN, kernel estimators, and local polynomials, all of which take the form

$$\widehat{\mu}(x) = \sum_{i=1}^{n} w_i(x; X^n) Y_i$$

for weights $w_i(x; X^n)$ that are large near x and small otherwise.

Global modeling is not explicitly motivated by local averaging, and instead works by setting up a sequence of function classes $\mathcal{F}_n$ growing in complexity with sample size n, and then optimizing

$$\widehat{\mu} = \arg\min_{f \in \mathcal{F}_k} \sum_{i=1}^{n} \Big\{ Y_i - f(X_i) \Big\}^2.$$

Here the tuning parameter is not a number of neighbors, or bandwidth, but instead the index k in the function class $\mathcal{F}_k$. This is typically something like the number of regressors/features included in the function class, and just as with other tuning parameters we have discussed, is chosen in some data-dependent way. When $\mathcal{F}_k = \{\theta^{\mathrm{T}} b(x) : \theta \in \mathbb{R}^k, b(x) = \{b_1(x), ..., b_k(x)\}^{\mathrm{T}}\}$ is linear combinations of basis functions, the corresponding global modeling approach is typically called series or linear sieve estimation.

Penalized modeling, on the other hand, picks a large function class $\mathcal{F}$ and then optimizes a penalized estimate of risk over the class, i.e.,

$$\widehat{\mu} = \arg\min_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} \Big\{ Y_i - f(X_i) \Big\}^2 + J_n(f) \right]$$

for some penalty function J_n, which will also involve some tuning parameter controlling the amount of penalization.

In these notes we discuss series estimation, mostly based on discussion and analysis in Chen [2007] and Belloni et al. [2015].

2 Series construction

As mentioned above, series estimation amounts to using $\mathcal{F}_k = \{\theta^{\mathrm{T}} b(x) : \theta \in \mathbb{R}^k, b(x) = \{b_1(x), ..., b_k(x)\}^{\mathrm{T}}\}$ and thus approximating $\mu(x)$ with linear functions $b(x)^{\mathrm{T}}\beta$ where

$$b(x) = \{b_1(x), ..., b_k(x)\}^{\mathrm{T}}$$

is a sequence of vector basis functions whose dimension can change with n. Due to the linearity the series estimator has a simple closed-form expression

$$\begin{aligned}\widehat{\mu}(x) &= b(x)^{\mathrm{T}}\mathbb{P}_n\{b(X)b(X)^{\mathrm{T}}\}^{-1}\mathbb{P}_n\{b(X)Y\} \\ &\equiv b(x)^{\mathrm{T}}\widehat{Q}^{-1}\mathbb{P}_n\{b(X)Y\} \equiv b(x)^{\mathrm{T}}\widehat{\beta}\end{aligned}$$

matching that of standard least squares estimation. (Note however that the fact the dimension k can change with n makes this a nonparametric estimator, and thus crucially different from standard parametric finite-dimensional regression estimation).

The choice of basis often depends on the underlying assumptions about the regression function μ. When μ is assumed to be Holder smooth, then it is common to use polynomial, trigonometric, spline, or wavelet series. First we describe these series in the univariate case and then later discuss how one can use tensor products to extend them to arbitrary dimension.

The polynomial basis is

$$b(x) = (1, x, x^2, ..., x^{k-1})^{\mathrm{T}}$$

but it is useful to use the Legendre polynomials

$$b(x) = \{1, \sqrt{3}x, \sqrt{5/4}(3x^2 - 1), ..., \}^{\mathrm{T}}$$

instead, which are orthonormal on $[0, 1]$ and so satisfy $\int_0^1 b(x)b(x)^{\mathrm{T}}\, dx = I_k$ and so can avoid collinearity issues. The trigonometric series is

$$b(x) = \{1, \cos(2\pi jx), \sin(2\pi jx), ...\}^{\mathrm{T}}, j = 1, ..., (k-1)/2$$

and is also orthonormal on $[0, 1]$. The linear spline series is

$$b(x) = \{1, x, (x - \ell_1)_+, ..., (x - \ell_{k-2})_+\}^{\mathrm{T}}$$

where $(\ell_1, ..., \ell_{k-2})$ are $k-2$ equally spaced knots on the support of X. The cubic spline series is

$$b(x) = \{1, x, x^2, x^3, (x - \ell_1)_+^3, ..., (x - \ell_{k-4})_+^3\}^{\mathrm{T}}$$

where $(\ell_1, ..., \ell_{k-4})$ are $k-4$ equally spaced knots. This ensures the function is a cubic polynomial within each knotpoint and that there are no discontinuities. B-splines are popular transformed versions of the above splines that reduce collinearity. The wavelet series is a bit cumbersome to describe but it is also orthonormal on $[0, 1]$.

All of the above bases were described for one-dimensional X. To extend to d-dimensional features we can simply use tensor products, i.e., apply each basis to each feature individually, then take all interactions. If one uses J terms for each feature, this amounts to $k = J^d$ features total, which is exponential in dimension. This can present challenges in finite samples. For example, suppose we have $d = 3$ features and use a polynomial basis with $J = 2$ terms for each feature. Then the tensorized basis

$$b(x) = (1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)^{\mathrm{T}}$$

has $k = 2^3 = 8$ terms total.

3 L_2 risk bound

Here we describe the analysis from Belloni et al. [2015], which gives in-probability bounds on the L_2 error

$$\|\widehat{\mu} - \mu\| = \sqrt{\int \left\{\widehat{\mu}(x) - \mu(x)\right\}^2 d\mathbb{P}(x)}.$$

Theorem 1. *Assume:*

1. $var(Y \mid X = x) \leq \sigma^2$,
2. $\mu(x)$ *is* s*-smooth,*
3. $\sup_{f \in \mathcal{H}(s)} \inf_\theta \|b^{\mathrm{T}}\theta - f\| \lesssim k^{-s/d}$,
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*

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

Before proving the result, we first discuss the assumptions and some implications. The third assumption, that $\sup_{f \in \mathcal{H}(s)} \inf_\theta \|b^{\mathrm{T}}\theta - f\| \lesssim k^{-s/d}$ is only a property of the basis, and characterizes how well it approximates Holder-smooth functions. Indeed all four bases described above (polynomial, trigonometric, spline, wavelet) satisfy this condition (for spline and wavelet series one need to use splines/wavelets of order at least s). Useful references for these kinds of bounds are the approximation theory textbooks Timan [1963] and DeVore and Lorentz [1993].

The assumption that the eigenvalues of Q are bounded is standard, and is needed so the features are not too correlated. Proposition 2.1 of Belloni et al. [2015] shows that this will hold for example if $b(x)$ is orthonormal on $[0, 1]$ and the covariate density is bounded above and below away from zero.

For the polynomial basis we have $\sup_x \|b(x)\| \lesssim k$, and for the other three bases we have $\sup_x \|b(x)\| \lesssim \sqrt{k}$, so the condition on number of features k is either $k^2 \log k/n \to 0$ or $k \log k/n \to 0$. Note that the latter can allow one to use nearly as many features as samples. This condition can be eliminated by adjusting the series estimator slightly, via a method called Forster-Warmuth regression by Yang et al. [2023].

Now we discuss the overall rate. The first $k^{-s/d}$ term is akin to bias, and decreases as we include more features, i.e., as we increase k. The second term $\sqrt{k/n}$ is akin to variance, and increases with more features. If we balance the two terms we have

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

which gives the overall rate

$$\|\widehat{\mu} - \mu\| = O_{\mathbb{P}}\left(n^{-1/(2+d/s)}\right)$$

i.e., the minimax optimal rate for estimating s-smooth functions.

4 Proof

Now we give the proof in some detail. Define

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

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

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

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

Note the approximation error satisfies

$$\begin{aligned}\|b^{\mathrm{T}}\beta - \mu\| = \min_\theta \|b^{\mathrm{T}}\theta - \mu\| &\leq \sup_{f\in\mathcal{H}(s)} \inf_\theta \|b^{\mathrm{T}}\theta - f\| \\ &\leq \sup_{f\in\mathcal{H}(s)} \inf_\theta \|b^{\mathrm{T}}\theta - f\|_\infty \lesssim k^{-s/d}\end{aligned}$$

where the first inequality follows by Assumption 2 and the last by Assumption 3.

Therefore it only remains to bound the estimation error $\|b^{\mathrm{T}}(\widehat{\beta} - \beta)\|$. In fact we have

$$\begin{aligned}\|b^{\mathrm{T}}(\widehat{\beta} - \beta)\|^2 &= (\widehat{\beta} - \beta)^{\mathrm{T}} \int b(x)b(x)^{\mathrm{T}} d\mathbb{P}(x)\ (\widehat{\beta} - \beta) \\ &= (\widehat{\beta} - \beta)^{\mathrm{T}} Q (\widehat{\beta} - \beta) \\ &\leq \|\widehat{\beta} - \beta\|^2 \|Q\|\end{aligned}$$

so really it is sufficient to bound the error $\|\widehat{\beta} - \beta\|$.

For it we have two terms

$$\begin{aligned}\|\widehat{\beta} - \beta\| &= \left\|\widehat{Q}^{-1}\mathbb{P}_n\Big[b(X)\{Y - b(X)^{\mathrm{T}}\beta\}\Big]\right\| \\ &\leq \left\|\widehat{Q}^{-1}\mathbb{P}_n\Big[b(X)\{Y - \mu(X)\}\Big]\right\| + \left\|\widehat{Q}^{-1}\mathbb{P}_n\Big[b(X)\{\mu(X) - b(X)^{\mathrm{T}}\beta\}\Big]\right\|\end{aligned}$$

For the first term we will argue that $\|\widehat{Q}^{-1}\|$ is bounded in probability. This is implied by the following two lemmas, together with Assumption 4 and the condition $\sup_x \|b(x)\|^2 \log k/n \to 0$.

Lemma 1 (Matrix LLN)**.** *For $\widehat{Q} = \mathbb{P}_n\{b(X)b(X)^{\mathrm{T}}\}$, $Q = \mathbb{E}(\widehat{Q})$, and $\xi_k = \sup_x \|b(x)\|$ we have*

$$\|\widehat{Q} - Q\| = O_{\mathbb{P}}\left(\sqrt{\frac{\|Q\|\xi_k^2 \log k}{n}}\right).$$

Proof. See Lemma 6.2 of Belloni et al. [2015]. □

Lemma 2. *For any matrix $\widehat{Q}$, if $\|\widehat{Q} - Q\| \xrightarrow{p} 0$ and $\|Q^{-1}\| \leq C < \infty$ then $\|\widehat{Q}^{-1}\| = O_{\mathbb{P}}(1)$.*

Proof. See previous lecture notes. □

Now note that

$$\mathbb{E}\|\mathbb{P}_n\{b(Y-\mu)\}\|^2 = \mathbb{E}\sum_{j=1}^{k}\Big[\mathbb{P}_n\{b_j(Y-\mu)\}\Big]^2 = \frac{1}{n}\sum_{j=1}^{k}\mathbb{E}\Big[b_j(X)^2\{Y-\mu(X)\}^2\Big]$$
$$\leq \frac{\sigma^2}{n}\sum_{j=1}^{k}\mathbb{E}\{b_j(X)^2\} = \frac{\sigma^2}{n}\mathrm{tr}(Q) \leq \frac{\sigma^2 k\|Q\|}{n}$$

so this, together with $\|\widehat{Q}^{-1}\| = O_{\mathbb{P}}(1)$ and Markov's inequality, implies

$$\Big\|\widehat{Q}^{-1}\mathbb{P}_n\Big[b(X)\{Y-\mu(X)\}\Big]\Big\| = O_{\mathbb{P}}\left(\sqrt{\frac{k}{n}}\right).$$

The second term in the estimation error

$$\Big\|\widehat{Q}^{-1}\mathbb{P}_n\Big[b(X)\{\mu(X)-b(X)^{\mathrm{T}}\beta\}\Big]\Big\|$$

is really approximation error again, and is $O_{\mathbb{P}}(k^{-s/d})$ by Assumptions 2–3, together with $\|\widehat{Q}^{-1}\| = O_{\mathbb{P}}(1)$. Specifically, letting $r = r(X) = \{\mu(X) - b(X)^{\mathrm{T}}\beta\}$, we have

$$\begin{aligned}
\|\widehat{Q}^{-1/2}\mathbb{P}_n(br)\|^2 &= \mathbb{P}_n(rb^{\mathrm{T}})\widehat{Q}^{-1/2}\widehat{Q}^{-1/2}\mathbb{P}_n(br) \\
&= \mathbb{P}_n(rb^{\mathrm{T}})\widehat{Q}^{-1}\widehat{Q}\widehat{Q}^{-1}\mathbb{P}_n(br) \\
&= \mathbb{P}_n\left[\left\{b^{\mathrm{T}}\widehat{Q}^{-1}\mathbb{P}_n(br)\right\}^2\right] \\
&\leq \mathbb{P}_n(r^2) = O_{\mathbb{P}}\left(\|\mu - b^{\mathrm{T}}\beta\|^2\right) = O_{\mathbb{P}}(k^{-2s/d})
\end{aligned}$$

where the inequality follows since $0 \leq \mathbb{P}_n\{(r - b^{\mathrm{T}}\widehat{\theta})^2\} = \mathbb{P}_n(r^2) - \mathbb{P}_n\{(b^{\mathrm{T}}\widehat{\theta})^2\}$ for $\widehat{\theta} = \widehat{Q}^{-1}\mathbb{P}_n(br)$, and the last equality by Assumptions 2–3. The fact that $\|\widehat{Q}^{-1}\| = O_{\mathbb{P}}(1)$ then yields the result.

References

A. Belloni, V. Chernozhukov, D. Chetverikov, and K. Kato. Some new asymptotic theory for least squares series: pointwise and uniform results. *Journal of Econometrics*, 186(2): 345–366, 2015.

X. Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

A. F. Timan. *Theory of approximation of functions of a real variable*. Macmillan Company, 1963.

Y. Yang, A. K. Kuchibhotla, and E. T. Tchetgen. Forster-warmuth counterfactual regression: A unified learning approach. *arXiv preprint arXiv:2307.16798*, 2023.