12 Series Methods

12.1 General Approach

A model has parameters (β, η) where β is finite-dimensional and η is nonparametric. (Sometimes, there is no β.) We will focus on regression.

The function η is approximated by a series – a finite dimensional model which depends on an integer K and a K dimensional parameter θ. Let $\eta_K(\theta)$ denote this approximating function.

Typically, the parameters (β, θ) are estimated by a conventional parametric technique $(\hat{\beta}, \hat{\theta})$. Then $\hat{\eta} = \eta_L(\hat{\theta})$

Tasks:

- To find a class of functions $\eta_K(\theta)$ which are good approximations to η.
- Study the bias (due to the finite dimensional approximation) and variance of the estimators
- Find optimal rates for K to diverge to infinity
- Find rules for selection of K
- Show that $\hat{\beta}, \hat{\eta}$ are asymptotically normal.
- Asymptotic variance computation, and standard error calculation.

Data Transformation: Typically the methods are applied after transforming the regressors X to lie in a specific compact space, such as $[0, 1]$.

12.2 Regression and Splines

Take the univariate regression

$$y_i = g(X_i) + e_i$$

In this case, $\eta = g$.

Series Approximations:

- power series (polynomial)
  - works for low order polynomials
  - unstable for high order polynomials
- trigonometric (sin and cos functions)
  - bounded functions
  - can produce "wiggly" implausible nonparametric function estimates
- splines

- piecewise polynomial of order r
- continuous derivatives up to $r-1$
- cubic splines popular
- join points (knots) can be selected evenly, or estimated

12.3 Splines

It is useful to define the "positive part" function

$$\begin{aligned}(a)_+ &= \max[0,a] \\ &= \begin{cases} 0 & a<0 \\ a & a\geq 0 \end{cases}\end{aligned}$$

Linear, quadratic and cubic splines with knots at $t_1 < t_2 < \cdots < t_{J-1}$ are

$$g_K(x) = \theta_0 + \theta_1 x + \sum_{j=1}^{J-1} \theta_{1+j}(x-t_j)_+$$

$$g_K(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \sum_{j=1}^{J-1} \theta_{2+j}(x-t_j)_+^2$$

$$g_K(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \sum_{j=1}^{J-1} \theta_{3+j}(x-t_j)_+^3$$

This model is set up so that it is everywhere a polynomial of order s, with continuous derivatives of order up to s, and the s''th derivative changing discontinuously at the knots. Cubic splines are smooth approximating functions, flexible, and popular. The approximation improves as the number of knots increases. The dimension of θ is $K = J + s$.

For a given set of knots the function g_K is linear in the parameters. Define

$$z = z(x) = \begin{pmatrix} 1 & x & x^2 & x^3 & (x-t_1)_+^3 & \cdots & (x-t_{J-1})_+^3 \end{pmatrix}',$$

then

$$g_K(x) = \theta_K' z$$

12.4 B Splines

Another popular class of series approximation are called B-splines. These are basis functions which are bounded, integrable and density-shaped. They can be constructed from a variety of basic shapes. Polynomials are common.

Let $X \in [0,1]$ and divide the support into J equal subintervals, with knots are $t_j = j/J$, $j = 0, 1, ..., J$. We also need knots outside of $[0,1]$ so let $t_j = j/m$ for all integers j.

An r'th order B-spline is a piecewise $(r-1)$-order polynomial.

A linear ($r = 2$) B-spline base functions are linear on two adjacent subintervals, zero elsewhere. They take the form

$$B_2(x \mid t_j, t_{j+1}, t_{j+2}) = (x - t_j)_+ - 2(x - t_{j+1})_+ + (x - t_{j+2})_+ .$$

A quadratic ($r = 3$) B-spline base function is piecewise quadratic over three subintervals

$$B_3(x \mid t_j, t_{j+1}, t_{j+2}, t_{j+3}) = (x - t_j)_+ - 3(x - t_{j+1})_+ + 3(x - t_{j+2})_+ - (x - t_{j+3})_+ .$$

For general r

$$B_r(x \mid t_j, \cdots, t_{j+r}) = \sum_{s=0}^{r} (-1)^s \binom{r}{s} (x - t_{j+s})_+ .$$

The B-spline is a linear combination of these basis functions.

$$\begin{aligned} g_K(x) &= \sum_{j=1-r}^{J-1} \theta_j B_r(x \mid t_j, \cdots, t_{j+r}) \\ &= \theta_K' z \end{aligned}$$

where $z = z(x)$ is the vector of the basic functions. The dimension of θ is $K = J + r + 1$

12.5 Estimation

For all of the examples, the function g_K is linear in the parameters (at least if the knots are fixed). Define the vector $Z_i = z(X_i)$ as the sample base function transformations. For example, in the case of a cubic spline

$$Z_i = \begin{pmatrix} 1 & X_i & X_i^2 & X_i^3 & (X_i - t_1)_+^3 & \cdots & (X_i - t_{J-1})_+^3 \end{pmatrix}' .$$

From Z_i, construct the regressor matrix Z. The LS estimate of θ_K is $\hat{\theta}_K = Z(Z'Z)^{-1} Z'y$. The estimate of $g(x)$ is $\hat{g}(x) = z'\hat{\theta}_K$, that of $g(X_i)$ is $\hat{g}(X_i) = z_i'\hat{\theta}_K$ and that of the vector $g = (g(X_1), ..., g(X_n))'$ is

$$\hat{g} = Z\hat{\theta}_K = Py$$

where

$$P = Z\left(Z'Z\right)^{-1} Z'$$

is a projection matrix.

12.6 Bias

Since $y = g + e$ then

$$\begin{aligned} E\left(\hat{\theta}_K \mid X\right) &= \left(Z'Z\right)^{-1} Z'E\left(y \mid X\right) \\ &= \left(Z'Z\right)^{-1} Z'g \\ &= \theta_K^* \end{aligned}$$

the coefficient from a regression of g on Z. This is the effective projection or pseudo-true value.

Similarly,

$$E\left(\hat{g} \mid X\right) = Pg = g_K^*$$

is the projection of g on Z.

The bias in estimation of g is

$$E\left(\hat{g} - g \mid X\right) = g_K^* - g.$$

If the series approximation works well, the bias will decrease as K gets increases. If g is α-times differentiable, then for splines and power series

$$\sup_x \left|g_K^*(x) - g(x)\right| \leq O\left(K^{-\alpha}\right).$$

The integrated squared bias is

$$ISB_K = \int \left(g_K^*(x) - g(x)\right)^2 dF(x) \leq O\left(K^{-2\alpha}\right)$$

where $F(x)$ is the marginal distribution of X.

This is approximately the same as the empirical average

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \left(g_K^*(X_i) - g(X_i)\right)^2 &= \frac{1}{n}\left(g_K^* - g\right)'\left(g_K^* - g\right) \\ &= \frac{1}{n} g'\left(I - P'\right)\left(I - P\right) g \\ &= \frac{1}{n} g'\left(I - P\right) g \end{aligned}$$

12.7 Integrated Squared Error

The integrated squared error of $\hat{g}(x)$ for $g(x)$ is

$$\begin{aligned} ISE &= \int (\hat{g}(x) - g(x))^2 \, dF(x) \\ &\simeq \frac{1}{n} \sum_{i=1}^{n} (\hat{g}(X_i) - g(X_i))^2 \\ &= \frac{1}{n} (\hat{g} - g)' (\hat{g} - g) \end{aligned}$$

Since

$$\begin{aligned} \hat{g} - g &= P(g + e) - g \\ &= Pe - (I - P) g \end{aligned}$$

then

$$\begin{aligned} ISE_K &= \frac{1}{n} (Pe - (I - P) g)' (Pe - (I - P) g) \\ &= \frac{1}{n} e'PPe + \frac{1}{n} g' \left(I - P'\right) (I - P) g - \frac{2}{n} e'P (I - P) g \end{aligned}$$

and when P is a projection matrix (as for LS estimation) then this simplifies to

$$ISE_K = \frac{1}{n} e'Pe + ISB_K \tag{1}$$

The first part represents estimation variance, the second is the integrated squared bias.

If the error is conditionally homoskedastic, then the conditional expectation of the first part is

$$\begin{aligned} E\left(\frac{1}{n} e'Pe \mid X\right) &= \frac{1}{n} \operatorname{tr}\left(PE\left(ee' \mid X\right)\right) \\ &= \frac{1}{n} \operatorname{tr}(P) \sigma^2 \\ &= \frac{K}{n} \sigma^2 \end{aligned}$$

In general, it can be shown that

$$\frac{1}{n} e'Pe = O_p\left(\frac{K}{n}\right)$$

Put together with the analyis of the ISB, we have

$$ISE_K \leq O_p\left(\frac{K}{n}\right) + O\left(K^{-2\alpha}\right).$$

The optimal rate for K is $K = n^{1/(2\alpha+1)}$ yielding a MSE convergence $n^{-2\alpha/(2\alpha+1)}$. This is the

same as the best rate attained by kernel regression using higher-order kernels or local polynomials.

12.8 Asymptotic Normality

The dimension of $\hat{\theta}_K$ grows with n, so we do not discuss its asymptotic distribution.

At any x, the estimate of $g(x)$ is $\hat{g}(x) = z'\hat{\theta}_K$, a linear function of the OLS estimator $\hat{\theta}_K$. Let $\hat{V}_K$ be the conventional (White) asymptoticcovariance matrix estimator for $\hat{\theta}_K$, so that for $z'\hat{\theta}_K$ is $z'\hat{V}_K z$. Applying the CLT we can find

$$\frac{\sqrt{n}\left(\hat{g}(x) - g_K^*(x)\right)}{\sqrt{z'\hat{V}_K z}} \to_d N(0,1)$$

Since the estimator is nonparametric, it is biased, so the estimator should be centered at the projection or pseudo-true value rather than the true $g(x)$. Alternative, if K is larger than optimal, so the estimator is "undersmoothed", then the squared bias will be of smaller order than the variance and it can be omitted from the asymptotic expression.

The bottom line is that for series estimation, we calculate standard errors using the conventional formula, as if the model were parametric. However, it is not constructive to focus on standard errors for individual coefficients, as they do not have individual meaning. Rather, standard errors should be for identifiable parameters, such as the conditional mean $g(x)$.

12.9 Selection of Series Terms

The role of K is similar to that of the bandwidth in kernel regression. Automatic data-dependent procedures are necessary for implementation.

As we worked out before, the integrated squared error is

$$ISE_K = \frac{1}{n} e'Pe + ISB_K$$

The optimal K minimizes this expression, but it is unknown.

We can estimate it using the sum-of-squared residuals from a model. For a given K, there regressors define a projection matrix P, fitted value $\hat{g} = Py$ and residual vector $\hat{e}_K = y - Py$. Note that

$$\begin{aligned} \hat{e}_K &= (I - P)\, y \\ &= (I - P)\, g + (I - P)\, e \end{aligned}$$

Thus the SSE is

$$\begin{aligned}\frac{1}{n}\hat{e}_K'\hat{e}_K &= \frac{1}{n}g'(I-P)g + \frac{2}{n}g'(I-P)e + \frac{1}{n}e'(I-P)e \\ &= ISB_K - \frac{1}{n}e'Pe + \frac{2}{n}g'(I-P)e + \frac{1}{n}e'e \\ &= ISE_K - \frac{2}{n}e'Pe + \frac{1}{n}2g'(I-P)e + \frac{1}{n}e'e\end{aligned}$$

Taking expectations conditional on X,

$$\begin{aligned}E\left(\frac{1}{n}\hat{e}_K'\hat{e}_K \mid X\right) &= E(ISE_K \mid X) - E\left(\frac{2}{n}e'Pe \mid X\right) + \sigma^2 \\ &= E(ISE_K \mid X) - \frac{2K\sigma^2}{n} + \sigma^2\end{aligned}$$

where the second line holds under conditional homoskedasticity.

Thus $\frac{1}{n}\hat{e}'\hat{e}$ is biased for ISE_K, but this can be corrected if we correct for the bias. This leads to Mallows (1973) criteria

$$C_K = \hat{e}_K'\hat{e}_K + 2K\hat{\sigma}^2$$

where $\hat{\sigma}^2$ is a preliminary estimate of σ^2. The scale doesn't matter, so I have multiplied through by n as is conventional, and the final σ^2 term doesn't matter, as it is independent of K.

The Mallows estimate $\hat{K}$ is the value which minimizes C_K.

A method which does not require homoskedasticity is cross-validation. The CV criterion is

$$CV_K = \sum_{i=1}^{n}\left(y_i - \hat{g}_{-i}^K(X_i)\right)^2$$

where $\hat{g}_{-i}^K$ is a K-th order series estimator omitting observation i. The CV estimate $\hat{K}$ is the value which minimizes CV_K.

Li (1987, Annals of Statistics) showed under quite minimal conditions that Mallows, GCV, and CV are asymptotically optimal for selection of K, in the sense that

$$\frac{ISE_{\hat{K}}}{\inf_k ISE_K} \to_p 1$$

Andrews (1991, JoE) showed that this optimality only extends to the heteroskedastic case if CV is used for selection. The reason is that the Mallows criterion uses homoskedasticity to calculate the bias adjustment, as we showed above, and this is not needed under CV.

12.10 Partially Linear and Additive Models

Suppose

$$y_i = W_i'\gamma + g(X_i) + e_i$$

with g nonparametric. A series approximation for g is $z'\theta_K$ yielding the model for estimation

$$y_i = W_i'\gamma + z_i'\theta_K + error_i$$

which is estimated by least-squares. The estimate for γ is similar to that from the Robinson kernel estimator, which had a residual-regression interpretation.

The asymptotic distribution for $\hat{\gamma}$ is the same as for the Robinson estimator, under the condition that the nonparametric component has MSE converging faster than $n^{-1/2}$, e.g. if $K/n + K^{-2\alpha} = o\left(n^{-1/2}\right)$. This is similar to the requirement for the Robinson estimator.

You can easily generalize this idea to multiple additive nonparametric components

$$y_i = W_i'\gamma + g_1\left(X_{1i}\right) + g_2\left(X_{2i}\right) + e_i$$

In practice, the components X_{1i} and X_{2i} are real-valued.

As discussed in Li-Racine, W_i can contain nonlinear interaction effects between X_{1i} and X_{2i}, such as $X_{1i}X_{2i}$. The main requirement is that the components of W_i cannot be additively separable in X_{1i} and X_{2i}. So in this sense the additive model can allow for simple interaction effects.