# Nonparametric Regression

Consider observing

$$Y_i = r(x_i) + \epsilon_i,$$

where the $\epsilon_i$'s are $iid$ with $E[\epsilon_i] = 0$ and $V[\epsilon_i] = \sigma^2$. The function $r$ is unknown, and we want to estimate it under minimal assumptions.

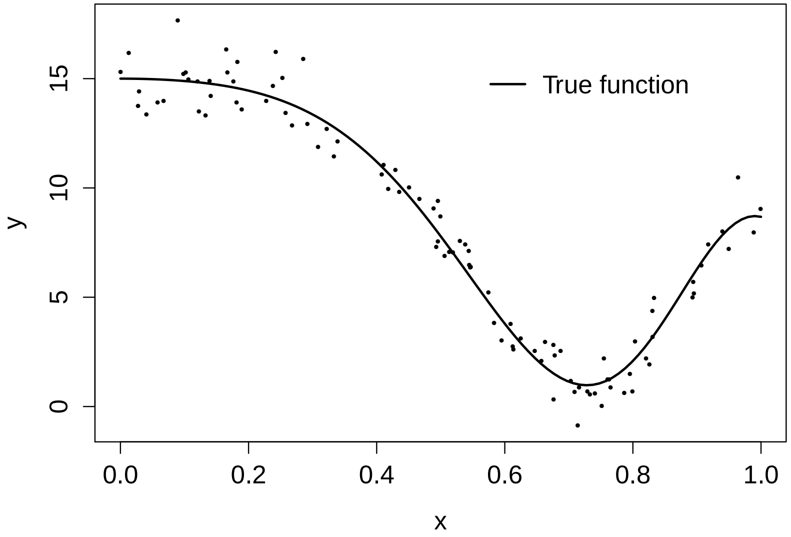
When $r(x) = E[Y|X = x]$, this extends the linear regression model we looked at recently.

(For simplicity we'll only consider the univariate case, with $x_i \in \mathbb{R}$).

Finding a good estimate of $r(x)$ will involve understanding the role of *smoothing* in the *bias-variance tradeoff*. The methods we'll consider fall into two categories:

1. "Localize" the problem: use observations close to $x$ to estimate $r(x)$.
2. Turn the problem back into something we know how to solve: multiple regression, using orthonormal functions in $x$.

Example:



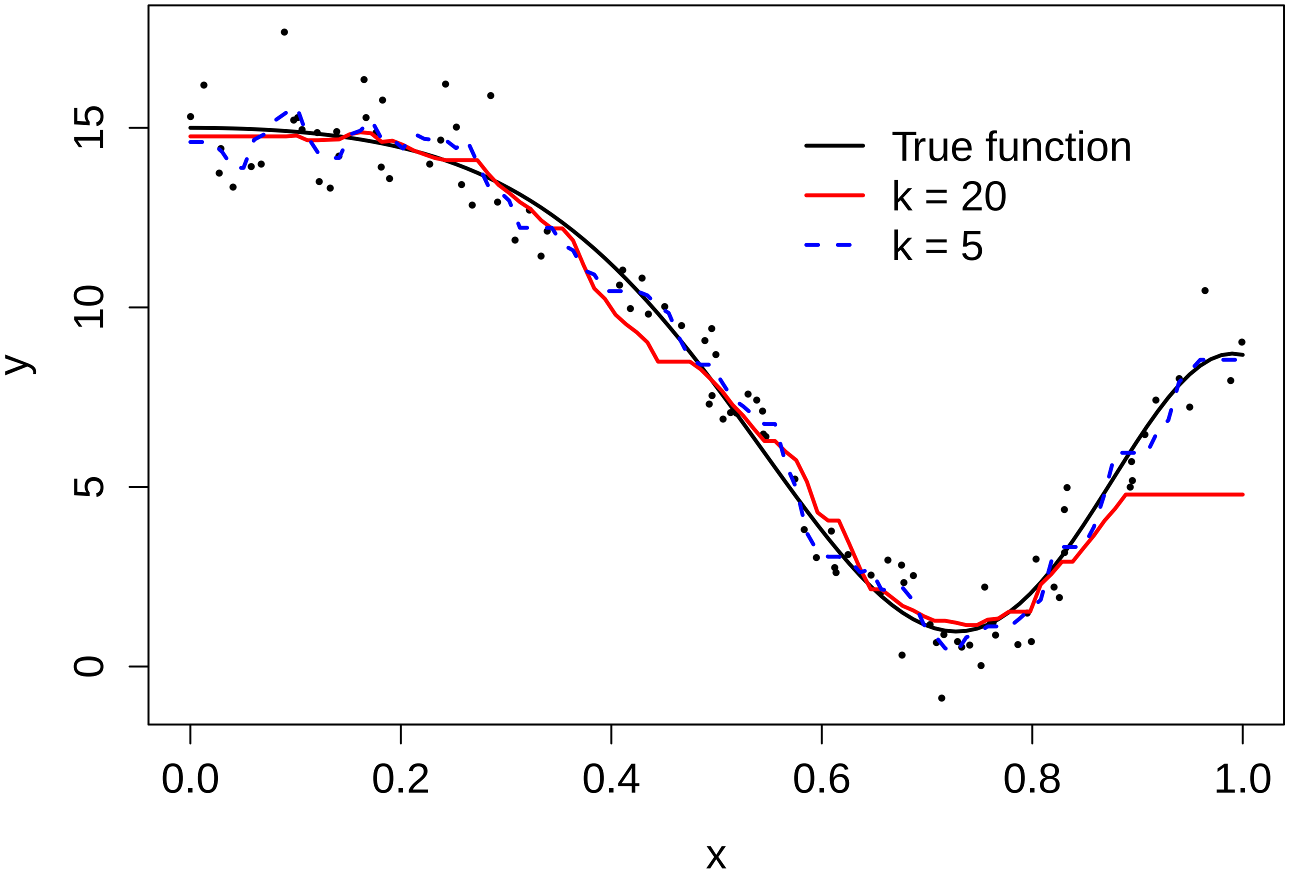
One simple but popular method for estimating $r$ is called $k-$nearest neighbors. Let $N_k(x)$ denote the $k$ values of $x_1, \ldots, x_n$ that are closest to $x$. Then the estimator is just the mean over the corresponding $Y_i$ values:

$$\hat{r}(x) = \frac{1}{k} \sum_{i: x_i \in N_k(x)} Y_i$$

The motivation for this estimator is that nearby observations "stand in" for repeated samples at a particular $x$. The function $r(x)$ is treated as if it were constant over the neighborhood $N_k(x)$.

Although this is not usually true, it may be a good approximation if $r$ is smooth and the neighborhood is small.

The degree of smoothing depends on $k$.



Let $f$ be the function we're trying to estimate. As with the ECDF $\hat{F}_n(x)$, in studying a particular estimator $\hat{f}_n(x)$, we need to keep track of two things that can vary: the observed data used to construct $\hat{f}_n$, and the value of $x$ at which we evaluate the function.

We will (primarily for tractability) use the integrated squared error (ISE) as our loss function:

$$L(f, \hat{f}_n) = \int [f(x) - \hat{f}_n(x)]^2 g(x) dx,$$

where $g(x)$ is pdf for $X$.

This gives us a frequentist risk function

$$R(f, \hat{f}_n) = E[L(f, \hat{f}_n)]$$

We can rewrite the risk as

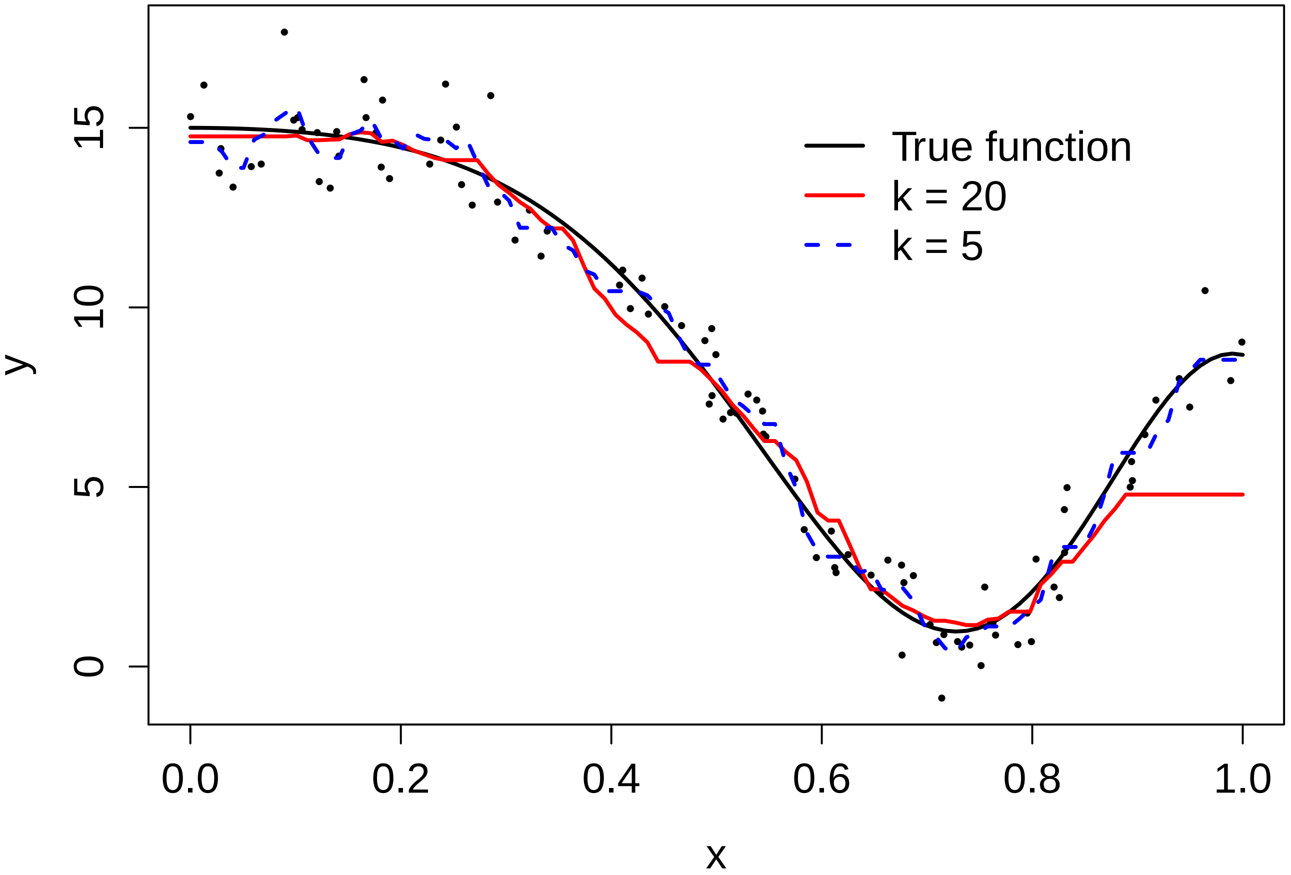
$$R(f,\hat{f}_n) = \int b^2(x)g(x)dx + \int v(x)g(x)dx$$

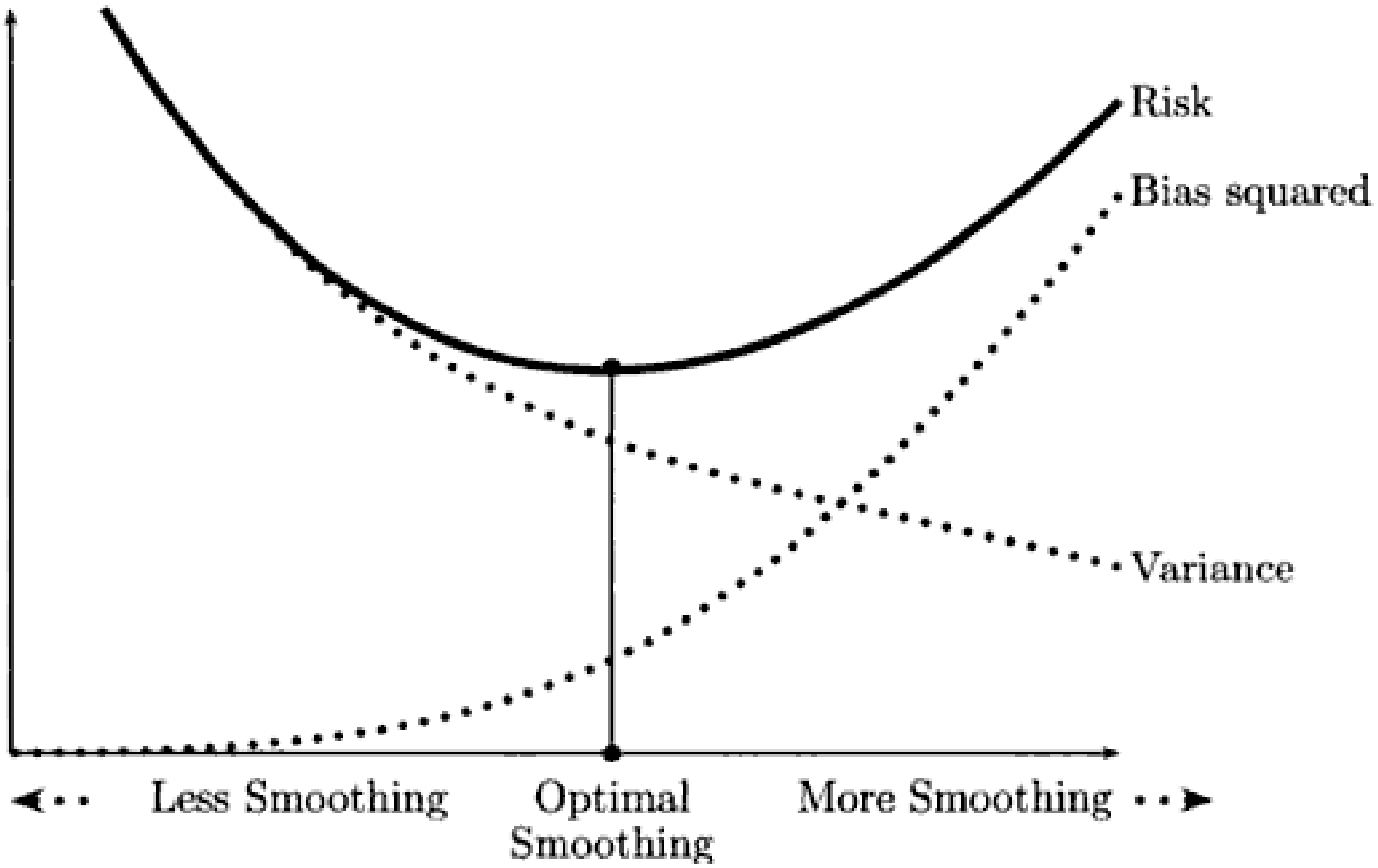
where

$$\begin{aligned} b(x) &= E[\hat{f}_n(x)] - f(x) \\ v(x) &= V[\hat{f}_n(x)] \end{aligned}$$

In the cases we will study, $\hat{f}_n$ will depend not only on the data, but also on a choice of *smoothing parameter*. We want to choose the smoothing parameter to minimize $R(f,\hat{f}_n)$. Typically we have that $b(x)$ increases with more smoothing and $v(x)$ decreases, so that the sum $R(f,\hat{f}_n)$ involves a *tradeoff* between bias and variance.

The degree of smoothing depends on $k$.




Risk
Bias squared
Variance
Less Smoothing
Optimal Smoothing
More Smoothing

One drawback to the $k-$nearest neighbor estimator is that it is discontinuous. We can think of this estimator as a weighted average, where each $Y_i$ is given a weight of either zero or $1/k$, depending on whether it's in the neighborhood.

The Nadaraya-Watson kernel estimator allows the weights to decay smoothly with distance.

$$\hat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

where $K$ is a kernel and

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

A kernel is a function $K$ such that

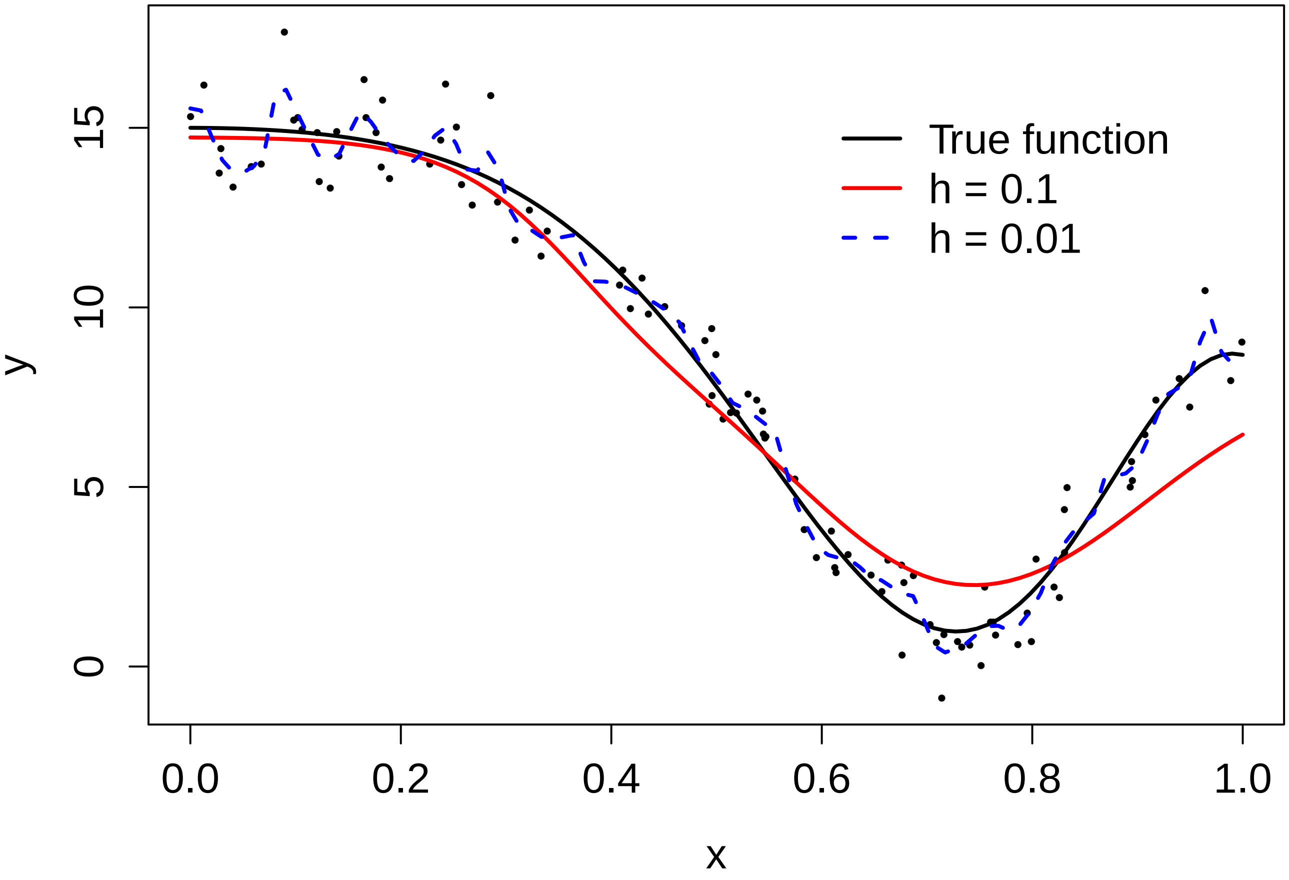
- $K(x) \geq 0$
- $\int K(x)dx = 1$
- $\int xK(x)dx = 0$
- $\int x^2 K(x)dx \equiv \sigma_K^2 > 0$

The Nadaraya-Watson kernel estimator

$$\hat{r}(x) = \sum_{i=1}^{n} \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} Y_i,$$

where $h$ is called the bandwidth. The choice of $h$ matters more than the functional form of $K$. One popular kernel function is Gaussian kernel.

The degree of smoothing depends on $h$.



We can choose the bandwidth $h$ to minimize the cross-validation estimate of the risk (using squared error loss)

$$\begin{aligned}
\hat{J}(h) &= \sum_{i=1}^{n} (Y_i - \hat{r}_{-i}(x_i))^2 \\
&= \sum_{i=1}^{n} \frac{(Y_i - \hat{r}(x_i))^2}{\left(1 - \frac{K(0)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)}\right)^2}
\end{aligned}$$

(See `npreg.R` for implementation.)

Now consider a completely different idea, appropriate for functions $r$ in

$$L_2(a,b) = \left\{ f : [a,b] \to \mathbb{R}, \quad \int_a^b f(x)^2 dx < \infty \right\}$$

These functions may be written as

$$r(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x), \quad \text{where } \beta_j = \int_a^b r(x)\phi_j(x)dx$$

and the functions $\phi_1, \phi_2, \ldots$ form an *orthonormal basis* for $L_2(a,b)$.

Before we define an orthonormal basis, note what we've done: we've given ourselves a way to approximate

$$\begin{aligned} r(x) &= \sum_{j=1}^{\infty} \beta_j \phi_j(x) \\ &\approx \sum_{j=1}^{J} \beta_j \phi_j(x) \end{aligned}$$

This approximation is accurate when $r$ is smooth, since in this case $\beta_j$ is small for large $j$.

Since we don't know the true function, we can't calculate $\beta_j$ exactly, but we can estimate it, treating $\phi_1(x), \ldots, \phi_J(x)$ as covariates in a multiple regression.

A sequence of functions $\phi_1, \phi_2, \phi_3, \ldots$ forms an *orthonormal basis* for $L_2(a,b)$ if

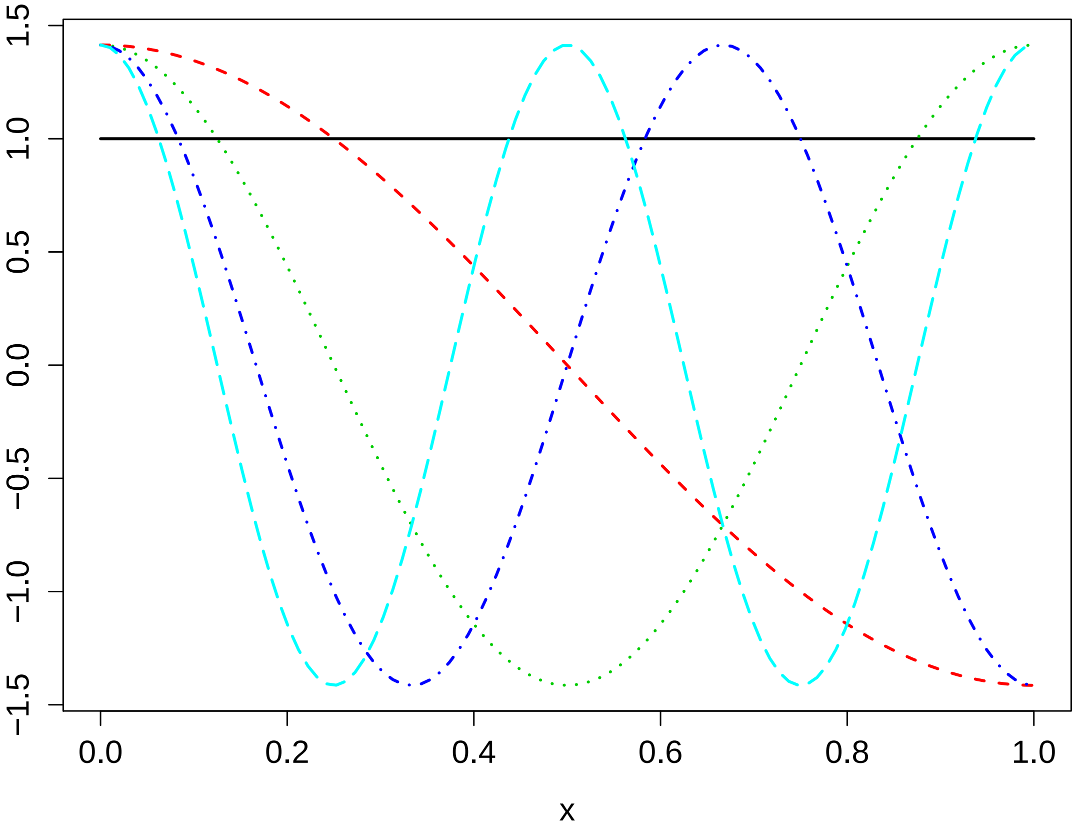
- $\int_a^b \phi_i(x)\phi_j(x)dx = 0$ for $i \neq j$ (orthogonal)
- $\int_a^b \phi_i(x)dx = 1$ for all $i$ (normal)
- The only function orthogonal to each $\phi_j(x)$ is the zero function (complete)

For example, the *cosine basis* for $L_2(0,1)$ takes $\phi_0(x) = 1$ and

$$\phi_j(x) = \sqrt{2}\cos(j\pi x)$$

for $j \geq 1$.

# First five cosine basis functions on $[0, 1]$



The Legendre polynomials on $[-1, 1]$ are given by the recursive relationship $P_0(x) = 1$, $P_1(x) = x$, and

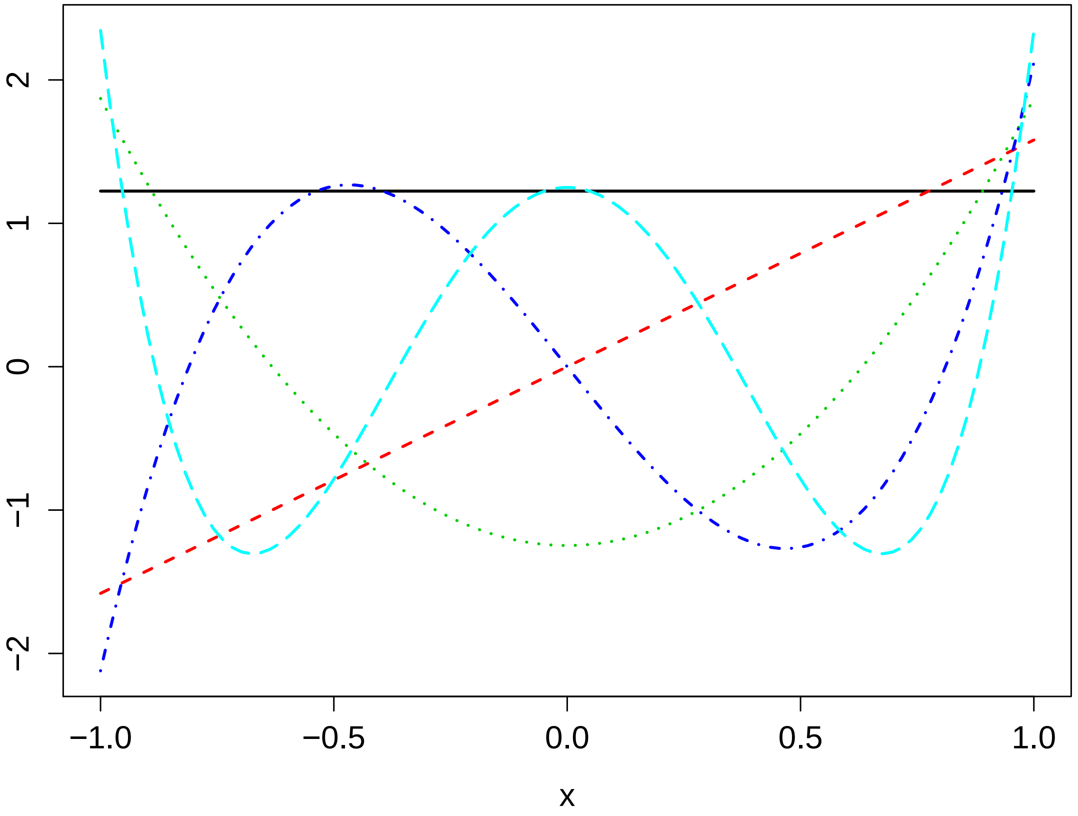
$$P_{j+1}(x) = \frac{(2j+1)xP_j(x) - jP_{j-1}(x)}{j+1}$$

They can be used to form an orthonormal basis for $L_2(-1, 1)$, setting

$$\phi_j(x) = \sqrt{(2j+1)/2}P_j(x)$$

(Note: The particular interval doesn't really matter, since it's easy to rescale the $x$'s.)

# First five Legendre polynomial basis functions on $[-1, 1]$



Recall our approximation $r(x) \approx \sum_{j=1}^{J} \beta_j \phi_j(x)$. For a particular choice of $J \leq n$, define

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_J(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_J(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \cdots & \phi_J(x_n) \end{pmatrix}$$

We have recast the nonparametric regression problem $Y_i = r(x_i) + \epsilon_i$ as the multivariate regression

$$Y = \Phi\beta + \eta$$

where $\eta_i = \epsilon_i$ plus some error from the approximation, hopefully small.

The least squares estimator of $\beta$ is just

$$\hat{\beta} = (\Phi'\Phi)^{-1}\Phi'Y$$