Lecture 11: Smoothing using Orthogonal Functions
11.1 Density estimation using orthogonal functions

- Let $X_1, X_2, ..., X_n$ be IID observations from a distribution on $[0, 1]$ with density $f$. If we assume that $f \in L^2$, we can write

$$f(x) = \sum_{j=0}^{\infty} \beta_j \phi_j(x)$$

where $\{\phi_j\}$ is an orthonormal basis of $L^2[0, 1]$.

- If we know $f(x)$, the coefficient $\beta_j$ is given by

$$\beta_j = \int_{[0,1]} f(x) \phi_j(x) \, dx.$$  

- The above formula looks similar to the Kernel density estimation. But the basis function $\phi_j(x)$ does not necessarily have measure 1 in contrast to the Kernel.

- Without knowing $f(x)$, how can we calculate the coefficient $\beta_j$? We need to estimate it using the data.
11.1 Density estimation using orthogonal functions

The estimate $\hat{\beta}_j$ of $\beta_j$ is given by

$$\hat{\beta}_j = \frac{1}{n} \sum_{i}^{n} \phi_j(x_i)$$

**Theorem.** The mean and variance of $\hat{\beta}_j$ are

$$E[\hat{\beta}_j] = \beta_j, \quad \text{Var}(\hat{\beta}_j) = \frac{\sigma^2_j}{n}$$

where $\sigma^2_j = \text{Var}(\phi_j(X_i)) = \int (\phi_j(x) - \beta_j)^2 f(x) dx$.

**Proof.**

$$E[\hat{\beta}_j] = \frac{1}{n} \sum_{i}^{n} E[\phi_j(X_i)] = E[\phi_j(X_1)] = \int \phi_j(x)f(x) dx = \beta_j.$$
11.1 Density estimation using orthogonal functions

► The estimate \( \hat{\beta}_j \) of \( \beta_j \) is given by

\[
\hat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(x_i)
\]

**Theorem.** The mean and variance of \( \hat{\beta}_j \) are

\[
E[\hat{\beta}_j] = \beta_j, \quad \text{Var}(\hat{\beta}_j) = \frac{\sigma_j^2}{n}
\]

where \( \sigma_j^2 = \text{Var}(\phi_j(X_i)) = \int (\phi_j(x) - \beta_j)^2 f(x)dx \).

**Proof.**

\[
E[\hat{\beta}_j] = \frac{1}{n} \sum_{i=1}^{n} E[\phi_j(X_i)] = E[\phi_j(X_1)] = \int \phi_j(x) f(x)dx = \beta_j.
\]

**Exercise.** Prove the variance.
11.1 Density estimation using orthogonal functions

► For a given $f(x)$, we know that

$$J \sum_j \beta_j \phi_j(x)$$

is more accurate if $J \in \mathbb{N}$ increases.

► This is not true anymore with the estimates $\{\hat{\beta}_j\}$. Think about the regression. A higher order polynomial regression function is not always better than a lower order polynomial regression function (bias and variance tradeoff).

► $J$ is called the **smoothing parameter**. It is typically chosen between 1 and $\sqrt{n}$ where $n$ is the sample size. $J$ is chosen so that it minimizes the **risk** (or **mean integrated squared error**).
11.1 Density estimation using orthogonal functions

Let \( \hat{f}(x) \) is an estimate of \( f(x) \) given by

\[
\hat{f}(x) = \sum_{j=1}^{J} \hat{\beta}_j \phi_j(x).
\]

Remember that the risk of \( \hat{f} \) using a smoothing parameter \( J \) is the expected value of the \( L^2 \) error, that is

\[
R(J) = E \left[ \int (\hat{f}(x) - f(x))^2 \, dx \right] = \sum_{j=1}^{J} \frac{\sigma_j^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2.
\]
11.1 Density estimation using orthogonal functions

**Theorem.** An estimate of the risk $R(J)$ is

$$
\hat{R}(J) = \sum_{j=1}^{J} \frac{\hat{\sigma}_{j}^{2}}{n} + \sum_{j=J+1}^{\infty} \left( \hat{\beta}_{j}^{2} - \frac{\hat{\sigma}_{j}^{2}}{n} \right)
$$

where $a_{+} = \max\{a, 0\}$ and

$$
\hat{a}_{j}^{2} = \frac{1}{n - 1} \sum_{i}^{n} \left( \phi_{j}(X_{i}) - \hat{\beta}_{j} \right)^{2}.
$$

- Using the $J^{*}$ that minimizes $\hat{R}(J)$, the estimate of the density $\hat{f}(x)$ is given by

$$
\hat{f}(x) = \sum_{j}^{J^{*}} \hat{\beta}_{j} \phi_{j}(x)
$$

- Note that $\hat{f}(x)$ can be negative!! If so, take $\hat{f}^{*} = \max(\hat{f}, 0)$ and normalize it.
11.2 Regression

For a data set \( \{X_i, Y_i\} \),

- Remember that the regression function \( r(x) \) is defined as the expected value of \( Y \) given \( x \)
  \[
  r(x) = E[Y|X = x].
  \]

- We studied parametric and nonparametric regressions. In particular, for nonparametric regression, we know a kernel density estimation based regression method.

- It is also possible to calculate a regression function using density estimation with orthogonal functions.

- Assume that \( r(x) \) is in \( L^2(0, 1) \) and \( x_i \) is uniformly distributed.

- \( r(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x) \) where \( \beta_j = \int_0^1 r(x) \phi_j(x) dx \) for an orthonormal basis \( \{\phi_j\} \) of \( L^2(0, 1) \).
11.2 Regression

The estimate of $\beta_j$, $\hat{\beta}_j$ is given by

$$\hat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(x_i), \quad j = 1, 2, \ldots$$

Theorem.

$$\hat{\beta}_j \sim N(\beta_j, \frac{\sigma^2}{n})$$

where $\sigma^2$ is the variance of the measurement error $e_i$

$$Y_i = r(x_i) + e_i$$

Idea of Proof. For the mean,

$$E[\hat{\beta}_j] = \frac{1}{n} \sum_{i=1}^{n} E[Y_i] \phi_j(x_i) = \frac{1}{n} \sum_{i=1}^{n} r(x_i) \phi_j(x_i)$$

$$\sim \int r(x) \phi_j(x) dx = \beta_j.$$
11.3 Wavelets

- Suppose that a regression function \( r(x) \) has a sharp jump but that \( r(x) \) is otherwise very smooth. That is, \( r(x) \) is spatially inhomogeneous.

- Doppler function \( \sqrt{x(1-x)} \sin \left( \frac{2.1\pi}{x+.05} \right) \)
11.3 Wavelets

Wavelets are local orthogonal functions.

**Harr wavelet.**

- Harr father wavelet (or Harr scaling function)

\[
\phi(x) = \begin{cases} 
1 & \text{if } 0 \leq x < 1 \\
0 & \text{otherwise.}
\end{cases}
\]

- Haar mother wavelet

\[
\psi(x) = \begin{cases} 
-1 & \text{if } 0 \leq x \leq 1/2 \\
1 & \text{if } 1/2 < x \leq 1
\end{cases}
\]

- For any integers \(j\) and \(k\) define

\[
\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k).
\]

- Let \(W_j = \{\psi_{j,k}, k = 1, 2, ..., 2^j - 1\}\) be the set of rescaled and shifted mother wavelets at resolution \(j\).
11.3 Wavelets

**Theorem.** The set of functions

\[ \{ \phi, W_0, W_1, \ldots \} \]

is an orthonormal basis for \( L^2(0, 1) \).

**Corollary.** For any \( f \in L^2(0, 1) \),

\[
f(x) = \alpha \phi(x) + \sum_{j} \sum_{k=0}^{\infty} 2^{j-1} \beta_{j,k} \psi_{j,k}(x)
\]

where

\[
\alpha = \int_{0}^{1} f(x) \phi(x) \, dx, \quad \beta_{j,k} = \int_{0}^{1} f(x) \psi_{j,k}(x) \, dx.
\]

- \( \alpha \) is called **scaling coefficient**.
- \( \beta_{j,k} \) are called **detail coefficients**.
- In a finite sum approximation of \( f \) using \( J \) different scales

\[
f(x) = \alpha \phi(x) + \sum_{j} \sum_{k=0}^{J} 2^{j-1} \beta_{j,k} \psi_{j,k}(x)
\]

\( J \) represents the resolution of the approximation.
11.3 Wavelets

Regression.

- Consider the regression model $Y_i = r(x_i) + \sigma e_i$ where $e \sim N(0, 1)$ and $x_i = i/n$.
- For simplicity, assume that $n = 2^J$ for some $J$.
- Smoothing with wavelets requires thresholding instead of truncation. That is, instead of choosing a smoothing parameter that determines the number of terms to keep, thresholding keeps coefficients that are sufficiently large.
- One example of thresholding is hard, universal thresholding.
11.3 Wavelets

Hard, universal thresholding.

1. Calculate

\[ \hat{\alpha} = \frac{1}{n} \sum_i \phi_k(x_i) Y_i, \quad \text{and} \quad D_{j,k} = \frac{1}{n} \sum_k \psi_{j,k}(x_i) Y_i \]

for \( 0 \leq j \leq J - 1 \) where \( J = \log_2(n) \).

2. Apply universal thresholding

\[ \hat{\beta}_{j,k} = \begin{cases} D_{j,k} & \text{if } |D_{j,k}| > \text{threshold value} \\ 0 & \text{otherwise} \end{cases} \]

3. Set

\[ \hat{r}(x) = \hat{\alpha} \phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \hat{b}_{j,k} \psi_{j,k}(x). \]
Homework

For $n = 10,000$, set $x_i = i/n$ and $y_i = \text{doppler}(x_i) + e_i$ where $e_i \sim N(0, 0.05^2)$.

1. Use the trigonometric functions to estimate the regression function.

2. Use the Legendre polynomials to estimate the regression function.

3. Use the Harr wavelets to estimate the regression function.

For 1-3, try to use a small number of terms. You are okay to use any programming libraries (that is, you do not need to make your own code; just use standard libraries) but specify all parameters to get your estimates.