6. Density Estimation

1. Cross Validation
2. Histogram
3. Kernel Density Estimation
4. Local Polynomials
5. Higher Dimensions
6. Mixture Models
7. Converting Density Estimation Into Regression

6.1 Cross Validation

Suppose we observe $X_1, \ldots, X_n$ from an unknown density $f$. Our goal is to estimate $f$ nonparametrically. Finding the best estimator $\hat{f}_n$ in some sense is equivalent to finding the optimal smoothing parameter $h$.

**How to measure the performance of $\hat{f}_n$?**

- Risk/Integrated Mean Square Error (IMSE)
  
  $$R(\hat{f}_n, f) = \mathbb{E}(L(\hat{f}_n, f))$$

  where $L(\hat{f}_n, f) = \int (\hat{f}_n(x) - f(x))^2 dx$.

- One can find an optimal estimator that minimizes the risk function:
  
  $$\hat{f}_n^* = \arg\min_{\hat{f}_n} R(\hat{f}_n, f),$$

  but the risk function is unknown!

**Finding an optimal estimator $\hat{f}_n$**

- $E(\hat{J}(h)) = E(J(h)) = R(h) + c$

- The optimal smoothing parameter
  
  $$h^* = \arg\min_h J(h)$$

- Nonparametric estimator $\hat{f}$ can be expressed as a function of $h$ and the best estimator $\hat{f}_n^*$ can be obtained by Plug-in the optimal smoothing parameter
  
  $$\hat{f}_n^* = \hat{f}_{h^*}.$$
6.2 Histogram

Histogram Estimator

- Without loss of generality, we assume that the support of \( f \) is \([0,1]\). Divide the support into \( m \) equally sized bins
  \[ B_1 = \left[ 0, \frac{1}{m} \right), B_2 = \left[ \frac{1}{m}, \frac{2}{m} \right), \ldots, B_m = \left[ \frac{m-1}{m}, 1 \right] \]
- Let \( h = \frac{1}{m} \), \( p_j = \int_{B_j} f(x) dx \) and \( Y_j = \sum_{i=1}^{n} I(X_i \in B_j) \)
- The histogram estimator is defined by
  \[ \hat{f}_n(x) = \sum_{j=1}^{n} \frac{Y_j}{n} \]

where \( \hat{p}_j = \frac{Y_j}{n} \).

**Proof.** For any \( x, u \in B_j \),

\[
f(u) = f(x) + (u - x)f'(x) + \frac{(u - x)^2}{2}f''(\bar{x}).
\]

for some \( \bar{x} \) between \( x \) and \( u \). Hence,

\[
p_j = \int_{B_j} f(u) du
= \int_{B_j} \left( f(x) + (u - x)f'(x) + \frac{(u - x)^2}{2}f''(\bar{x}) \right) du
= f(x)h + hf'(x) \left( h \left( j - \frac{1}{2} \right) - x \right) + O(h^3).
\]

**Theorem 1.** Suppose that \( f' \) is absolutely continuous and \( \int (f')^2 du < \infty \). Then

\[
R(\hat{f}_n, f) = \frac{h^2}{12} \int (f')^2 du + \frac{1}{nh} + o(h^2) + o\left( \frac{1}{n} \right).
\]

The optimal bandwidth is

\[
h^* = \frac{1}{n^{1/3}} \left( \frac{6}{\int (f'(u))^2 du} \right)^{1/3}.
\]

With the optimal binwidth,

\[
R(\hat{f}_n, f) \approx \frac{C}{n^{2/3}}
\]

where \( C = \left( \frac{2}{3} \right)^{2/3} \left( \int (f'(u))^2 du \right)^{1/3} \).

Therefore, the bias of \( \hat{f}_n \) is

\[
b(x) = E(\hat{f}_n(x)) - f(x) = \frac{1}{h} E(\hat{p}_j) - f(x)
= \frac{p_j}{h} - f(x)
= \frac{1}{h} \left( f(x)h + hf'(x) \left( h \left( j - \frac{1}{2} \right) - x \right) + O(h^3) \right) - f(x)
= f'(x) \left( h \left( j - \frac{1}{2} \right) - x \right) + O(h^2).
\]

By the mean value theorem, for some \( \bar{x}_j \in B_j \),

\[
\int_{B_j} b^2(x) dx = \int_{B_j} (f'(x))^2 \left( h \left( j - \frac{1}{2} \right) - x \right)^2 dx + O(h^4)
= (f'(\bar{x}_j))^2 \int_{B_j} \left( h \left( j - \frac{1}{2} \right) - x \right)^2 dx + O(h^4)
= (f'(\bar{x}_j))^2 \frac{h^3}{12} + O(h^4).
\]
Hence
\[
\int_0^1 b^2(x)dx = \sum_{j=1}^m \int_{B_j} b^2(x)dx + O(h^3)
\]
\[
= \sum_{j=1}^m (f''(\bar{x}_j))^2 \frac{h^3}{12} x + O(h^3)
\]
\[
= \frac{h^2}{12} \int_0^1 (f''(x))^2dx + o(h^2).
\]

For the variance of \(\hat{f}_n\):
\[
v(x) \equiv \text{Var}(\hat{f}_n(x)) = \frac{1}{h^2} \text{Var}(\hat{p}_j) = \frac{p_j(1-p_j)}{nh^2}.
\]

By the mean value theorem, for some \(x_j \in B_j\),
\[
p_j = \int_{B_j} f(x)dx - hf(x_j).
\]

Therefore,
\[
\int_0^1 v(x)dx = \sum_{j=1}^m \int_{B_j} v(x)dx = \sum_{j=1}^m \int_{B_j} p_j(1-p_j)dx
\]
\[
= \frac{1}{nh^2} \sum_{j=1}^m \int_{B_j} p_j - \frac{1}{nh^2} \sum_{j=1}^m \int_{B_j} p_j^2
\]
\[
= \frac{1}{nh} - \frac{1}{nh^2} \sum_{j=1}^m p_j^2 = \frac{1}{nh} - \frac{1}{nh} \sum_{j=1}^m h^2f^2(x_j)
\]
\[
= \frac{1}{nh} - \frac{1}{nh} \left(\int_0^1 f^2(x)dx + o(1)\right) = \frac{1}{nh} + o\left(\frac{1}{n}\right)
\]

\(\square\)

**Theorem 2.** The cross-validation estimator of risk for the histogram is
\[
\hat{f}(h) = \frac{2}{h(n-1)} - \frac{(n+1)}{h(n-1)} \sum_{j=1}^m \hat{p}_j^2.
\]

**Theorem 3.** Let \(m = m(n)\) be the number of bins in the histogram \(\hat{f}_n\). Assume that \(m(n) \to \infty\) and \(m(n) = \log n/n \to 0\) as \(n \to \infty\). Define
\[
l_n(x) = \left(\max\left\{\sqrt{\hat{f}_n(x)} - c, 0\right\}\right)^2, \quad u_n(x) = \left(\sqrt{\hat{f}_n(x)} - c, 0\right)^2
\]
where \(c = \frac{z_{n/2m^2}}{2} \sqrt{\frac{m}{n}}\). Then
\[
\Pr\left(l_n(x) \leq E(\hat{f}_n) \leq u_n(x) \forall x\right) \geq 1 - \alpha.
\]

**6.3 Kernel Density Estimation**

- Given a kernel \(K\) and a positive number \(h\), called the bandwidth, the kernel density estimator is:
\[
\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x-X_i}{h}\right).
\]
- The choice of kernel \(K\) is not crucial but the choice of bandwidth \(h\) is important.
- We assume that \(K\) satisfies
\[
\int K(x)dx = 1, \quad \int x K(x)dx = 0 \text{ and } \sigma^2_K \equiv \int x^2 K(x) > 0.
\]
Theorem 4. Let \( R_x = \mathbb{E}((f(x) - \hat{f}(x))^2) \) be the risk at point \( x \) and let \( R = \int R_x dx \) denote the integrated risk. Assume that \( f'' \) is absolutely continuous and that \( \int (f'''(x))^2 dx < \infty \). Also, assume that \( K \) satisfies
\[
\int K(x) dx = 1, \quad \int xK(x) dx = 0 \quad \text{and} \quad \sigma_K^2 \equiv \int x^2 K(x) > 0.
\]
Then,
\[
R = \left\{ \begin{array}{ll}
\frac{1}{4} \sigma_K^4 h_n^4 (f''(x))^2 + \frac{f(x) \int K^2(x) dx}{nh_n} + O(n^{-1}) + O(h_n^4), \\
\frac{1}{4} \sigma_K^4 h_n^4 \int (f''(x))^2 dx + \frac{\int K^2(x) dx}{nh_n} + O(n^{-1}) + O(h_n^4).
\end{array} \right.
\]

Proof. Write \( K_h(x, X) = \frac{1}{h} K \left( \frac{x - X}{h} \right) \) and \( \hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x, X_i) \).
\[
\mathbb{E}(\hat{f}_n(x)) = \mathbb{E}(K_h(x, X)) = \int \frac{1}{h} K \left( \frac{x - t}{h} \right) f(t) dt
\]
\[
= \int K(u)f(x - hu) du
\]
\[
= \int K(u) \left( f(x) - hu f'(x) + \frac{h^2 u^2}{2} f''(x) + \cdots \right) du
\]
\[
f(x) + \frac{1}{2} h^2 f''(x) \int u^2 K(u) du.
\]
Hence
\[
\text{bias}(\hat{f}_n(x)) = \frac{1}{2} \sigma_K^4 h_n^4 f''(x) + O(h^4).
\]

Similarly,
\[
\text{Var}(\hat{f}_n(x)) = \frac{\int f(x) \int K^2(x) dx}{nh_n} + O(n^{-1}).
\]

Bandwidth selection
- The optimal smoothing bandwidth is
\[
h^* = \left( \frac{c_2}{c_1^2 A(f)} \right)^{1/5}
\]
where \( c_1 = \int x^2 K(x) dx, c_2 = \int (K(x))^2 dx \) and \( A(f) = \int (f''(x))^2 dx \).
- The only unknown quantity in \( h^* \) is \( A(f) = \int (f''(x))^2 dx \)
- \( A(f) = \int (f''(x))^2 dx = \int (f^{(4)}(x)) f(x) dx = \mathbb{E}(f^{(4)}) \) where \( f^{(r)} \) denote the \( r \)th derivative of \( f \)

The Normal reference rule
- Assume that \( f \) is Normal, one can compute the \( h^* \) with the Gaussian kernel.
\[
h^* = 1.06 \sigma_n^{-1/5}
\]
- \( \sigma \) is estimated by \( \hat{\sigma} = \min\{s, \text{IQR}/1.34\} \) where \( s \) is the sample standard deviation and IQR is the interquartile range.
- The selected bandwidth is:
\[
h_n = \frac{1.06 \hat{\sigma}}{n^{1/5}}.
\]
Plug-in method

Let \( \psi_r = E(f^{(r)}) \). To estimate \( \psi_4 \) by using the Kernel method, one need to choose the optimal bandwidth which is a functional of \( \psi_6 \).

1. Estimate \( \psi_8 \) with the bandwidth chosen the normal reference rule.
2. Estimate \( \psi_6 \) with the bandwidth depending on \( \hat{\psi}_8 \)
3. Estimate \( \psi_4 \) with the bandwidth depending on \( \hat{\psi}_6 \)
4. The selected bandwidth is
   \[
   h^* = \left( \frac{c_2}{c_1 \psi_4 n} \right)^{1/5}
   \]

Cross-validation

- Cross-validation score function:
  \[
  \hat{J}(h) = \int \hat{f}^2(x) dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)
  \]
- The selected bandwidth is
  \[
  h_* = \arg\min_h \hat{J}(h)
  \]

**Theorem 5 (Stone’s Theorem).** Suppose \( f \) is bounded. Let \( \hat{f}_h \) denote the kernel estimator with bandwidth \( h \) and let \( h_* \) denote the bandwidth chosen by cross-validation. Then
  \[
  \frac{\int (f(x) - \hat{f}_h)^2 dx}{\inf_h \int (f(x) - \hat{f}_h)^2 dx} \xrightarrow{a.s.} 1.
  \]

Computation

- The cross-validation score function \( \hat{J}(h) \) can be approximated by
  \[
  \hat{J}(h) = \frac{1}{nh^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K^*(\frac{X_i - X_j}{h}) + \frac{2}{nh} K(0) + O\left(\frac{1}{n^2}\right)
  \]
  where \( K^*(x) = K^{(2)}(x) - 2K(x) \) and \( K^{(2)} z - y K(y) dy \).
- For the computation of \( \hat{f}_n \) and \( \hat{J}(h) \), use:
  - fast Fourier transform (FFT)
  - binning strategy

Optimal Convergence rate

From Theorem 4, the optimal convergence rate is \( O(n^{-4/5}) \) if the optimal bandwidth is used.

**Theorem 6.** Let \( F \) be the set of all pdfs and let \( f^{(m)} \) denote the \( m \) derivative of \( f \). Define
  \[
  \mathcal{F}_m(c) = \left\{ f \in \mathcal{F} : \int |f^{(m)}(x)|^2 dx \leq c^2 \right\}.
  \]
  For any estimator \( \hat{f}_n \),
  \[
  \sup_{f \in \mathcal{F}_m(c)} \mathbb{E}_f \int (\hat{f}_n(x) - f(x))^2 dx \geq bn^{-2m/(2m+1)}.
  \]
  where \( b > 0 \) is a universal constant that depends only on \( m \) and \( c \).
6.4 Local Polynomials

**Smoothed log-likelihood**

- Recall that the nonparametric MLE of the pdf is
  \[ p = (p_1, \ldots, p_n) \equiv \arg \min_p L(p) \]
  where \[ L(p) = \sum_{i=1}^{n} \log p_i - n \left( \sum_{i=1}^{n} p_i - 1 \right). \]
- A smoothed version of the log-likelihood at \( x \) (up to constant) is
  \[ L_x(p) = \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) \log p_i - n \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) p_i. \]

**Local likelihood density estimator**

- Approximate \( \log f(u) \) with a polynomial
  \[ P_x(a, u) = \sum_{j=0}^{p} a_j (x - u)^j. \]
- The local polynomial log-likelihood is
  \[ L_x(a) = \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) P_x(a, X_i) - n \int K \left( \frac{u - X_i}{h} \right) e^{P_x(a, u)} du. \]
- The local likelihood density estimator is
  \[ \hat{f}_n(x) = e^{P_x(\hat{a}, x)} = e^{\hat{a}_0}, \]
  where
  \[ \hat{a} = (\hat{a}_0, \ldots, \hat{a}_p)^T \equiv \arg \max_a L_x(a) \]

**Local log-likelihood**

- The kernel density estimator:
  \[ \hat{f}_K(x) = \arg \max_p L_x(p). \]
- The log-likelihood function of \( f \):
  \[ L(f) = \sum_{i=1}^{n} \log f(X_i) - n \left( \int f(u) du - 1 \right). \]
- The local log-likelihood at target value \( x \) is
  \[ L_x(f) = \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) \log f(X_i) - n \int K \left( \frac{x - u}{h} \right) f(u) du. \]

6.5 Higher Dimensions

Suppose \( X_i = (X_{i1}, \ldots, X_{id}) \). The multivariate kernel estimator is
\[ \hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|H|} K(H^{-1}(x - X)). \]
We often assume a simple form of the bandwidth matrix or kernel.
For example, \( H = \text{diag}(h_1, \ldots, h_d) \). Then
\[ \hat{f}_n(x) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^{d} \left\{ \prod_{j=1}^{d} K \left( \frac{x_j - X_{ij}}{h_j} \right) \right\}. \]
The risk function:
\[
R \approx \frac{1}{4} \sigma_K^4 \left[ \sum_{j=1}^{d} h_j^2 \int f_j^2(x) dx + \sum_{j \neq k} h_j^2 h_k^2 \int f_{jk} f_{kk} dx \right] + \frac{\int K^2(x) dx}{nh_1 \cdots h_d}.
\]

- The optimal bandwidth \( h^*_i = O(n^{-1/(4+\theta)}) \).
- In practice, one can choose the optimal bandwidth by cross-validation and often assume a simple form of the bandwidth matrix. \( \mathbf{H} = h \cdot I \).
- The curse of dimensionality.

**EM algorithm**

For \( K = 2 \),
- Define latent variables \( Z_i \):
  \[
  Z_i = \begin{cases} 
  1 & \text{if } Y_i \text{ is from group 1} \\
  0 & \text{if } Y_i \text{ is from group 2} 
  \end{cases}
  \]
- The log-likelihood function is
  \[
  \mathcal{L}(y, z, \theta) = \sum_{i: z_i = 0} \log[p\phi(y_i | \mu_1, \sigma_1^2)] + \sum_{i: z_i = 1} \log[(1 - p)\phi(y_i | \mu_2, \sigma_2^2)]
  \]
  \[
  = \sum_{i=1}^{n} \log[p\phi(y_i | \mu_1, \sigma_1^2)] + (1 - z_i) \log[(1 - p)\phi(y_i | \mu_2, \sigma_2^2)]
  \]

**6.6 Mixture Models**

**The Normal Mixture Model**

\[
f(x, \theta) = \sum_{j=1}^{K} p_j \phi(x | \mu_j, \sigma_j^2),
\]
where \( \phi_j(x | \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi} \sigma_j} e^{-\frac{(x - \mu_j)^2}{2\sigma_j^2}} \).

- Define \( \theta = (p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2) \)
- Given \( K \), how to estimate \( \theta \)?
  \( \rightarrow \text{EM algorithm} \)

**E-step** Given \( \theta \), compute \( \mathbb{E}(Z_i | Y, \theta) \).

\[
E(Z_i | Y = y_i, \theta) \equiv w_i = \frac{\hat{p}\phi(y_i)}{\phi(y) + (1 - \hat{p})\phi(y)}
\]

**M-step**

\[
\hat{\theta} = \arg\max_{\theta} \tilde{\mathcal{L}}(\theta),
\]
where

\[
\tilde{\mathcal{L}}(\theta) = \sum_{i=1}^{n} \left( w_i \log \hat{p}\phi(y_i) + (1 - w_i) \log(1 - p)\phi(y_i) \right).
\]

Then
\[
\hat{\rho} = \frac{\sum_{i=1}^{n} w_i}{n}, \quad \hat{\mu}_1 = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \quad \hat{\mu}_2 = \frac{\sum_{i=1}^{n} (1 - w_i) y_i}{\sum_{i=1}^{n} (1 - w_i)},
\]
\[
\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n} w_i (y_i - \hat{\mu}_1)^2}{\sum_{i=1}^{n} w_i}, \quad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{n} (1 - w_i) (y_i - \hat{\mu}_1)^2}{\sum_{i=1}^{n} (1 - w_i)}.
\]
How to estimate the number of components $k$?

- AIC (Akaike Information Criterion): find most predictive model
  \[ \text{AIC} = L - q \]
  where $L$ is the loglikelihood and $p$ is the number of parameters.

- BIC (Bayesian Information Criterion): find the true model with high probability
  \[ \text{BIC} = L - q \log n \]

- The number of clusters is not always equal to the number of component
- Gaussian is sensitive to outliers → Replace the normal density with $t$-distribution density.
- If $\sigma_j^2 = \sigma^2 > 0$ is fixed and $K \to n$, then MLE of the mixture model approaches the kernel estimate where $p_j = \frac{1}{n}$ and $\hat{\mu}_j = x_j$.

```r
source("ch6.r")
library(MASS)
library(sm)
library(locfit)
library(mclust)

### Read data
faithful <- read.table("faithful.dat", header=T)
eruptions <- faithful$eruptions
n <- length(eruptions)

### Select the optimal number of bins by CV
hist.h <- cv.hist.fun(eruptions)$mbest

### Select optimal bandwidth of kernel estimators by normal reference rule/Cross-validation/plug-in
sigma.hat <- min(sd(eruptions), IQR(eruptions)/1.34)
h.normal <- 1.06*sigma.hat/n^0.2

h.cv <- ucv(eruptions)
h.plugin <- width.SJ(eruptions)

f1 <- density(eruptions, width=h.normal) # normal reference
f2 <- density(eruptions, width=h.cv) # cross-validation
f3 <- density(eruptions, width=h.plugin) # plug-in
f4 <- density(eruptions, method="mclust") # finite mixture model
f5 <- locfit(~ eruptions, alpha=c(0.1,0.8), flim=c(1,6))
f6 <- locfit(~ eruptions, alpha=c(0.1,0.6), flim=c(1,6), link="ident")

postscript("density.ps")
par(mfrow=c(2,2))
truehist(eruptions, xlim=c(1,6), ymax=0.8) # histogram
truehist(eruptions, nbins=hist.h, xlim=c(1,6), ymax=0.8)
plot(f1, ylim=c(0,0.8)); plot(f2, ylim=c(0,0.8))
plot(f3, ylim=c(0,0.8)); plot(f4, ylim=c(0,0.8))
plot(f5, ylim=c(0,0.8), main="local linear")
plot(f6, ylim=c(0,0.8), main="local quadratic")
dev.off()
```
6.7 Converting Density Estimation Into Regression

Description of Methodette

1. Suppose that $X_1, \ldots, X_n$ are from a density $f$ on $[0, 1]$
2. Create $k$ equal width bins where $k \approx n/10$
3. Define
   \[ Y_j = \sqrt{\frac{k}{n}} \times \sqrt{N_j + \frac{1}{4}}, \]
   where $N_j = \sum_{i=1}^{n} I(X_i \in B_j)$ and $B_j$ is the $j^{th}$ bin.
4. Let $r(x) = \sqrt{f(x)}$ and $t_j$ is the mid point of $B_j$. Then
   \[ Y_j \approx r(t_j) + \sigma \epsilon_j, \]
   where $\epsilon_j \sim N(0, 1)$ and $\sigma = \sqrt{\frac{k}{4n}}$.
5. Apply your favorite nonparametric regression method to $(t_j, Y_j)$ to get an estimate $\hat{r}_n$.
6. Calculate $(\hat{r}_n(x))^2$ and normalize to be a density.
   \[ \hat{f}_n(x) = \frac{(\hat{r}^+(x))^2}{\frac{1}{k} \sum_{j=1}^{k} (\hat{r}^+(t_j))^2}, \]
   where $\hat{r}^+(x) = \max\{\hat{r}_n(x), 0\}$. 
How does it work?

- Poissonization
  \[ N_j \approx \text{Poisson} \left( n \int_{B_j} f(x)dx \right) \approx \text{Poisson} \left( \frac{nf(t_j)}{k} \right) \]

- \( E(N_j) = \text{Var}(N_j) \approx nf(t_j)/k \).

- Variance stabilization
  \[ Y_j = \sqrt{\frac{k}{n}} \times \sqrt{N_j + \frac{1}{4}}, \]

- \( E(Y_j) \approx \sqrt{f(t_j)} \) and \( \text{Var}(Y_j) \approx k/(4n) \)

source("ch6.r")
library(sm)

tmp <- den.to.reg(eruptions)
temp2 <- smooth.spline(tmp$x, temp.y, cv=TRUE)
temp3 <- temp2$y
temp3[temp3 < 0] <- 0
k <- length(temp3)
normal.const <- sum(temp3^2)/k
segs <- seq(0,1,by=0.01)
temp4 <- predict(temp2, segs)$y
temp4[temp4 < 0] <- 0
f.hat <- (temp4)^2/normal.const*k
postscript("den_reg.ps")
plot(segs, f.hat, main="Density to Regression")
dev.off()

REFERENCES