Chapter 1

Non-parametric Density Estimation

1.1 The Empirical Distribution Function

In the course *Statistics*, the *empirical distribution function* was discussed. Let F be a cumulative distribution function (c.d.f.) and $X_1, \ldots, X_n$ an i.i.d. sample from F, then $\widehat{F}_n(x) := \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{(-\infty,x]}(X_j)$ is an unbiased estimator of F and the sequence $(\widehat{F}_n)_{n\geq 1}$ is *consistent* in the sense that

$$\lim_{n\to+\infty} \sup_x |\widehat{F}_n(x) - F(x)| = 0$$

in probability. These results were proved in the course *Statistics*.

Let $D_n = \sup_x |\widehat{F}_n(x) - F(x)|$. It was also proved that the distribution of D_n is independent of the generating c.d.f. F and the *Kolmogorov-Smirnov* theorem was stated (without proof); let K be a random variable with c.d.f.

$$F_K(x) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}$$

and let K_α denote the number such that $F_K(K_\alpha) = \alpha$, then

$$\lim_{n\to+\infty} \mathbb{P}(\sqrt{n}D_n > K_\alpha) = \alpha.$$

This may be used to test whether or not data comes from a distribution with c.d.f. F.

One important application of this is the *normal probability plot*. To test whether or not data comes from a $N(\mu, \sigma^2)$ distribution, firstly the data is *centred*, by subtracting $\overline{x}$ (the sample average) from each point and then *standardised* by divided through by the sample standard deviation s, where $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$. This gives a standardised column of data (mean 0, standard deviation 1). The standardised column $(z_1, \ldots, z_n)$ is then ordered, to give $z_{(1)} < \ldots < z_{(n)}$. Let $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2} dy$, then the points $(\Phi(z_{(j)}), \frac{j}{n})$ are plotted. If the data comes from i.i.d. normal

sampling, then these points should lie on a straight line. The Kolmogorov-Smirnov statistic may be used to give confidence intervals.

Suppose F is continuous, with density f. The empirical ditribution cannot be used, at least not directly, for estimating the *density*, since $\widehat{F}_n$ is not differentiable (it increases in jumps sized $\frac{1}{n}$).

1.2 Density Estimators

Suppose we want to estimate a continuous probability density function $p(\underline{x}) : \underline{x} \in \mathbb{R}^r$. That is, $p(\underline{x}) \geq 0$ and $\int_{\mathbb{R}^r} p(\underline{x}) d\underline{x} = 1$, based on a random sample $X_1, \ldots, X_n$ generated by p. Recall that an estimator $\widehat{p}(x)$ of $p(x)$ is *unbiased* if $\mathbb{E}_p[\widehat{p}(x)] = p(x)$. Notation: $\mathbb{E}$ denotes expectation, the subscript indicates the distribution; in other words the expectation of $\widehat{p}(x)$ when $\widehat{p}(x)$ is generated by a random sample from p. The notation is clear from the context. Although in some special cases unbiased estimators do exist, these are the exception. If $\widehat{p}_n$ is an estimator based on a sample size n, the sequence of estimators is said to be *pointwise consistent* if $\mathbb{E}_p[\widehat{p}_n(x)] \overset{n\to+\infty}{\longrightarrow} p(x)$ and $\mathrm{Var}_p(\widehat{p}_n(x)) \overset{n\to+\infty}{\longrightarrow} 0$ for each $x \in \mathbb{R}^r$.

1.3 Measures of Efficiency of Density Estimators

1.3.1 Consistency

Weak Pointwise Consistency

Let $(\widehat{p}_n)_{n\geq 1}$ be a sequence of estimators of p. The simplest notion of consistency is *weak-pointwise consistency*. This simply means that $\widehat{p}_n(x) \longrightarrow_{n\to+\infty} p(x)$ in probability for each $x \in \mathbb{R}^r$.

Strong Pointwise Consistency

A sequence of estimators is *strongly pointwise consistent* if convergence holds almost surely for each $x \in \mathbb{R}^r$.

Consistent in Quadratic Mean

The *mean squared error* at point $x \in \mathbb{R}^r$ is defined as:

$$\mathrm{MSE}(x) = \mathbb{E}_p\left[(\widehat{p}(x) - p(x))^2\right] = \mathrm{Var}_p(\widehat{p}(x)) + (\mathrm{bias}_p(\widehat{p}(x))^2 .$$

If $\mathrm{MSE}(x) \overset{n\to+\infty}{\longrightarrow} 0$ for each $x \in \mathbb{R}^r$, then $\widehat{p}$ is said to be *pointwise consistent in quadratic mean*. Note that for some x values the convergence may be much slower than for others; the fact that we have pointwise consistency does not imply that $\lim_{n\to+\infty} \sup_x \mathrm{MSE}(x) = 0$.

Integrated Mean Squared Error

A more standard estimator of performance (which does not require the very strong condition that $\lim_{n\to+\infty} \sup_x |\widehat{p}_n(x) - p(x)| = 0$) is the IMSE (Integrated Mean Squared Error)

$$\text{IMSE} = \int_{\mathbb{R}^r} \text{MSE}(x)dx.$$

This may be written as:

$$\text{IMSE} = \mathbb{E}_p\left[\int \widehat{p}(x)^2 dx\right] - 2\int_{\mathbb{R}^r} p(x)\mathbb{E}_p\left[\widehat{p}(x)\right]dx + \int p(x)^2 dx.$$

Notation We use $R(g) = \int g(x)^2 dx$. Also, we use $\widehat{p} := \int_{\mathbb{R}^r} p(x)\widehat{p}(x)dx$. Then

$$\text{IMSE} - R(p) = \mathbb{E}_p\left[R(\widehat{p}) - 2\widehat{p}\right].$$

Then $R(\widehat{p}) - 2\widehat{p}$ is an unbiased estimator of $\text{IMSE} - R(p)$.

$\text{IMSE} = \mathbb{E}_p[\text{ISE}]$ where

$$\text{ISE} = \int_{\mathbb{R}^r} \left(\widehat{p}(x) - p(x)\right)^2 dx$$

ISE means *Integrated Squared Error*.

For *bona fide* density estimates (i.e. those that satisfy $\widehat{p}(x) \geq 0$ and $\int \widehat{p}(x)dx = 1$), it turns out that the best possible asymptotic rate of convergence of the IMSE is $O(n^{-4/5})$; we'll show this later.

Alternative Measures

The L_1 approach considers the IAE (*integrated absolute error*)

$$IAE = \int |\widehat{p}(x) - p(x)|dx.$$

The Kullback-Leibler divergence between two probability densities f and g is defined as

$$KL(f|g) = \int f(x)\log\frac{f(x)}{g(x)}dx;$$

the Kullback-Leibler between $\widehat{p}$ and p is therefore:

$$KL = \int \widehat{p}(x)\log\left(\frac{\widehat{p}(x)}{p(x)}\right)dx.$$

The *Hellinger Distance*(HD) is defined as:

$$HD(m) = \left(\int (\widehat{p}(x)^{1/m} - p(x)^{1/m})^m dx\right)^{1/m}.$$

1.4 The Histogram

A standard method of estimating p from an i.i.d. sample and giving a visual representation is the histogram. A set of non-overlapping bins is constructed,

$$a = t_{n,0} < t_{n,1} < \ldots < t_{n,L} = b$$

the bins are $T_l = [t_{n,l}, t_{n,l+1})$ for $l = 0, \ldots, L-1$ and the histogram is defined by:

$$\widehat{p}(x) = \frac{1}{n} \sum_{l=0}^{L-1} \frac{N_l}{t_{n,l+1} - t_{n,l}} \mathbf{1}_{T_l}(x)$$

where $N_l = \sum_{j=1}^{n} \mathbf{1}_{T_l}$, the number of observations in bin T_l. Often a common bin width is fixed, $h_n = t_{n,l+1} - t_{n,l}$ for each l and then the density estimator is:

$$\widehat{p}(x) = \frac{1}{nh_n} \sum_{l=0}^{L-1} N_l \mathbf{1}_{T_l}(x). \tag{1.1}$$

1.4.1 The Histogram as a ML Estimator

Let $H(\Omega)$ be a specified class of real valued functions defined on Ω. For a random sample of observations $X_1, \ldots, X_n$, the problem is to find a $p \in H(\Omega)$ which maximises

$$L(p) = \prod_{j=1}^{n} p(X_j)$$

(Here we are treating p as the parameter).

Of course, $\int p(x)dx = 1$ and $p(x) \geq 0$ for all $x \in \Omega$. If the bins are fixed and of equal length, this reduces to finding a function of the form $\sum_{l=0}^{L-1} y_l \mathbf{1}_{T_l}(x)$ subject to $\sum_{l=0}^{L-1} y_l = 1$ and $y_l \geq 0$. It is left as an exercise to show that the ML estimator is (1.1).

1.4.2 Asymptotics

Let $p_l = \int_{T_l} p(x)dx$. Clearly, by the mean value theorem, there exists a $\xi_l \in T_l$ such that $p_l = h_n p(\xi_l)$. Then:

$$\mathbb{P}((N_0, \ldots, N_{L-1}) = (k_0, \ldots, k_{L-1})) = \frac{n!}{k_0! \ldots k_{L-1}!} \prod_{j=0}^{L-1} p_j^{k_j} \qquad k_0 + \ldots + k_{L-1} = n.$$

This is the multinomial distribution. It follows that $\mathbb{E}[N_l] = np_l$ and $\mathrm{Var}(N_l) = np_l(1 - p_l)$. Also, for $i \neq j$, $\mathrm{Cov}(N_i, N_j) = -np_i p_j$. Hence, for $x \in T_l$,

$$\mathbb{E}[\widehat{p}(x)] = p(\xi_l) \qquad \text{Var}(\widehat{p}(x)) = \frac{p_l(1-p_l)}{nh_n^2} \leq \frac{p(\xi_l)}{nh_n}.$$

The integrated squared bias (ISB) is:

$$\text{ISB} = \sum_{l=0}^{L-1} \int_{T_l} (p(x) - p(\xi_l))^2 dx.$$

Approximating, by taking a Taylor expansion, assuming ξ_l occurs at the half-way point and only considering first order terms, $p(x) \simeq p(\xi_l) + (x - \xi_l)p'(\xi_l)$ so that the Asymptotic Integrated Squared Bias (AISB) is:

$$\text{AISB} \simeq \sum_{l=0}^{L-1} (2 \int_0^{h_n/2} y^2 dy) p'^2(\xi_l),$$

so that (using $2\int_0^x y^2 dy = \frac{2x^3}{3}$):

$$AISB = \frac{1}{12} h_n^2 R(p').$$

The assumption is that $h_n \to 0$ as $n \to +\infty$.

For the Integrated Variance (IV), using $\text{Var}(\widehat{p}(x)) = \frac{1}{nh_n^2} p_l(1-p_l)^2$, the Asymptotic Integrated Variance (AIV) is:

$$\text{AIV} = \sum_{l=0}^{L-1} h_n \frac{1}{nh_n^2} (p_l - p_l^2) = \frac{1}{nh_n} - \frac{1}{n} R(p).$$

Since $h_n \to 0$, it follows that only the first of these is important and hence the AIMSE is:

$$AIMSE = \frac{1}{nh_n} + \frac{1}{12} h_n^2 R(p').$$

If $h_n \to 0$ and $nh_n \to +\infty$, then $IMSE \to 0$.

Differentiating the expression for the AIMSE in h_n and setting the result equal to zero gives an optimal bin width of:

$$h_n^* = \left(\frac{6}{R(p')n} \right)^{1/3}.$$

1.4.3 Estimating Bin Width

Optimal bin width is often estimated under the assumption of normality; for $N(\mu, \sigma^2)$, the bin width is:

$$h_n^* \simeq 3.4809\sigma n^{-1/3}.$$

This is a straightforward computation:

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{|z-\mu|^2}{2\sigma^2}\} \Rightarrow p'(z) = p(z)\frac{(\mu - z)}{\sigma^2}$$

so that (using $x = \sqrt{2}\frac{z-\mu}{\sigma}$)

$$R(p') = \int_{-\infty}^{\infty}\frac{1}{2\pi\sigma^2}\exp\{-\frac{(z-\mu)^2}{\sigma^2}\}\frac{(\mu-z)^2}{\sigma^4}dz = \frac{1}{4\sqrt{\pi}\sigma^3}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}x^2dx = \frac{1}{4\sqrt{\pi}\sigma^3}.$$

Since σ is unknown, Scott's rule, of using the sample variance s in place of σ is often used, which gives $h_n^* \simeq 3.5sn^{-1/3}$.

If the data has heavier tails than Gaussian, the *interquartile range* IQR is used. The *quartiles* are q_1, q_2 and q_3 where q_2 is the median and $F(q_1) = 0.25$, $F(q_3) = 0.75$, where F is the c.d.f.. The interquartile range is defined as IQR $= q_3 - q_1$. The bin width is:

$$h_n^* = 2(IQR)n^{-1/3}.$$

1.4.4 Multivariate Histograms

The univariate results transfer easily to the multivariate setting. For a random sample of r-vectors from the multivariate density $p(x) : x \in \mathbb{R}^r$, the space is partitioned into hyper rectangles, each with volume $h_{1n}\dots h_{rn}$. Suppose N_l of the multivariate observations fall into the lth hyperrectangle, where $\sum_l N_l = n$, then the density estimate is

$$\widehat{p}(x) = \frac{1}{nh_{1n}\dots h_{rn}}\sum_l N_l \mathbf{1}_{B_l}(x)$$

where B_l denotes hyperrectangle with index l. The asymptotically optimal bin width can be computed similarly to the univariate setting:

$$h_{ln}^* = \frac{1}{\sqrt{R(p_l)}}\left(6\prod_{j=1}^{r}(R(p_j))^{1/2}\right)^{1/(2+r)}\frac{1}{n^{1/(2+r)}}.$$

The asymptotically optimal IMSE is

$$AIMSE^* = \frac{1}{4}6^{2/(r+2)}\left(\prod_{j=1}^{r}R(p_j)\right)^{1/(2+r)}\frac{1}{n^{2/(2+r)}},$$

where $p_j = \frac{\partial}{\partial x_j} p(x)$.

In the multivariate Gaussian setting, $N(0, \Sigma)$ where $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_r^2)$, this gives

$$h_{ln}^* = 2 \times 3^{1/(2+r)} \pi^{r/(4+2r)} \sigma_l n^{-1/(2+r)}.$$

1.5 Kernel Density Estimation

A popular way of estimating a density is the so-called *kernel density estimation*. Take a function K satisfying $K \geq 0$ and $\int K(x)dx = 1$. For a random sample $X_1, \ldots, X_n$ from a distribution with density p over $\mathbb{R}^r$, the estimator is:

$$\widehat{p}(x) = \frac{1}{n|H|} \sum_{j=1}^{n} K(H^{-1}(x - X_j)) \qquad x \in \mathbb{R}^r$$

where H is a $r \times r$ matrix. Usually, $H = hA$, where A is a fixed $r \times r$ matrix satisfying $|A| = 1$ and $h \in \mathbb{R}_+$ is the band width parameter. The parameter h determines the size and $A^t A$ determines the elliptical shape. If $A = I_r$, then

$$\widehat{p}(x) = \frac{1}{nh^r} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right).$$

1.5.1 Popular Kernels

The most popular kernels are as follows:

1. Rectangular: $K(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)$
2. Triangular: $K(x) = (1 - |x|)\mathbf{1}_{[-1,1]}(x)$
3. Bartlett-Epanechnikov: $K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{[-1,1]}(x)$
4. Biweight $K(x) = \frac{15}{16}(1 - x^2)^2\mathbf{1}_{[-1,1]}(x)$
5. Triweight $K(x) = \frac{35}{32}(1 - x^2)^3\mathbf{1}_{[-1,1]}(x)$
6. Cosine $K(x) = \frac{\pi}{4}\cos\left(\frac{\pi x}{2}\right)\mathbf{1}_{[-1,1]}(x)$.

Note that if K is a kernel, then $K_2(x) := \int K(y)K(x - y)dy$ is also a kernel. In fact, a simple computation shows that the triangular kernel can be obtained from the rectangular in this way.

1.5.2 Asymptotic Variance and Bias

In the univariate case, the asymptotic pointwise variance and bias are:

$$\begin{cases} \mathrm{Var}(\widehat{p}_h(x)) = \frac{R(K)p(x)}{nh_n} - \frac{p(x)^2}{n} \\ \mathrm{bias}(\widehat{p}_h(x)) = \frac{1}{2}\sigma_K^2 h_n^2 p''(x). \end{cases} \tag{1.2}$$

Here $\sigma_K := \int x^2 K(x)dx$ and (as before) $R(g) = \int g(x)^2 dx$. These are computed by taking a Taylor expansion; the details are left as an exercise (in the tutorial exercises).

Since $h_n \to 0$, the second term for the variance can be ignored when computing the asymptotic. The AMISE is therefore:

$$\mathrm{AMISE}(h_n) = \frac{R(K)}{nh_n} + \frac{1}{4}\sigma_K^4 h_n^4 R(p''). \tag{1.3}$$

This gives an optimal window width of:

$$h_n^* = \left(\frac{R(K)}{\sigma_K^4 R(p'')}\right)^{1/5} \frac{1}{n^{1/5}}.$$

With this wiindow width, the AMISE is:

$$\mathrm{AMISE}^* = \frac{5}{4}\left(\sigma_K R(K)\right)^{4/5} \left(R(p''\right)^{1/5} \frac{1}{n^{4/5}}.$$

In the univariate case, where K is the standard Gaussian kernel and p is Gaussian with variance σ^2, this may be computed to give:

$$h_n = 1.06\sigma n^{-1/5}. \tag{1.4}$$

1.5.3 Estimating Window Width

Consider the univariate setting. Since σ is (in general) unavailable, the sample standard deviation s is used. If the distribution is non-Gaussian, this may be misleading. The interquartile range may be used. For the Gaussian, IQR $\simeq 1.34s$, so replacing σ by $\min(s, \frac{\mathrm{IQR}}{1.34})$ in (1.4) is a standard rule-of-thumb.

Rule-of-thumb estimators are generally considered unsatisfactory; the window width may be too large, leading to over-smoothing. The presence of an important node may be unknowingly removed.

1.5.4 Unbiased Cross Validation

Recall that the ISE is:

$$\mathrm{ISE} = \int \widehat{p}(x)^2 dx - 2\int \widehat{p}(x)p(x)dx + \int p(x)^2 dx.$$

The aim is to find the value for h which minimises the ISE. Since the third term does not depend on $\widehat{p}$, this term may be ignored. The second term is $-2\mathbb{E}\left[\widehat{p}(X)\right]$. This may be estimated by a leave-one-out method. The density estimator based on the sample without point X_i is:

$$\widehat{p}_{h,-i}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{X_j - x}{h}\right)$$

and $\mathbb{E}[\widehat{p}(X)]$ is estimated by $\frac{1}{n}\sum_{i=1}^{n}\widehat{p}_{h,-i}(X_i)$. The *unbiased cross-validation* UCV choice is the window width that minimises

$$\mathrm{UCV}(h) = R(\widehat{p}_h) - \frac{2}{n}\sum_{i=1}^{n}\widehat{p}_{h,-i}(X_i).$$

It is *unbiased* in the sense that

$$\mathbb{E}\left[\mathrm{UCV}(h)\right] = \mathrm{MISE}(h) - R(p).$$

1.5.5 Biased Cross Validation

Biased cross validation (BCV) comes from choosing h to minimise AMISE(h) (Equation 1.3) directly. Consider the univariate case. The AMISE depends on the unknown $R(p'')$, which needs to be estimated.

It is a straightforward computation to show that:

$$\mathbb{E}_p[R(\widehat{p}'')] = R(p'') + \frac{R(K'')}{nh^5} + O(h^2)$$

Indeed,

$$\begin{aligned}
\mathbb{E}[R(\widehat{p}'')] &= \frac{1}{nh^6}\int\int K''(\frac{x-y}{h})^2 p(y)dydx + (1-\frac{1}{n})\frac{1}{h^6}\int\int\int K''(\frac{x-y}{h})K''(\frac{x-z}{h})dydzdx \\
&= \frac{1}{nh^5}\int\int K''(z)^2 p(x-hz)dzdx \\
&\quad +(1-\frac{1}{n})\frac{1}{h^4}\int\int\int K''(z_1)K''(z_2)p(x-hz_1)p(x-hz_2)dz_1dz_2dx \\
&\simeq \frac{1}{nh^5}R(K'') + (1-\frac{1}{n})R(p'') + O(h^2)
\end{aligned}$$

In the second term, the derivatives (integration by parts) are put on the p's. The fact that the other terms are small come from the fact that $\int p^{(n)}(x)dx = 0$ for any order of derivative $n \geq 1$ (in the first term) and that $\int p''p''' = 0$ (second term).

It follows that $R(\widehat{p}'')$ asymptotically overestimates $R(p'')$, hence the estimator:

$$\widehat{R}(p'') = R(\widehat{p}''_h) - \frac{R(K'')}{nh^5}.$$

Set $K_h(x) = \frac{1}{h}K(\frac{x}{h})$ so that $K''(\frac{x}{h}) = h^3 K''_h(x)$. Starting from

$$\widehat{p}''_h(x) = \frac{1}{n}\sum_{j=1}^{n} K''_h(x - X_j)$$

gives:

$$R(\widehat{p}''_h) = \frac{1}{n^2}\sum_{i,j}\int K''_h(x - X_i)K''_h(x - X_j)dx =: \frac{1}{n^2}\sum_{ij} K''_h * K''_h(X_i - X_j)$$

where $f * g(x) = \int f(y)g(x - h)dy$ denotes the convolution. Dealing with $i = j$ and $i \neq j$ separately yields:

$$R(\widehat{p}''_h) = \frac{R(K'')}{nh^5} + \frac{1}{n^2h^5}\sum_{i\neq j} K''_h * K''_h(X_i - X_j)$$

from which

$$\widehat{R}(p''_h) = \frac{1}{n^2h^5}\sum_{i\neq j} K''_h * K''_h(X_i - X_j)$$

Substituting into the AMISE and setting $h = h_n$ gives the BCV

$$BCV(h_n) = \frac{R(K)}{nh_n} + \frac{\sigma_K^4}{2n^2h_n}\sum_{i<j} K''_{h_n} * K''_{h_n}(X_i - X_j).$$

UCV tends to undersmooth while BCV tends to oversmooth.

1.5.6 Sheather - Jones (SJ)

The plug-in idea is as follows:

1. Choose a window width g_n as a pilot density estimate $\widehat{p}_{g_n}$ and use this to estimate $R(p'')$ as $R(\widehat{p}''_{g_n})$.

2. Plug this estimate for $R(p'')$ into (1.3).

Sheather-Jones propose the following: estimating $R(p'')$ is different from estimating p, so the window-widths g_n and h_n are different, but related, through a function g; $g_n = g(h_n)$. The optimal choices of windows are derived by mean squared criteria in each case and estimating the unknown terms.

1.6 Kernel Regression

Consider a regression problem

$$Y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \qquad i.i.d. N(0, \sigma^2)$$

where the function f is unknown. Using

$$\mathbb{E}[Y|X = x] = f(x)$$

we can take the weighted average of the observed responses $(y_1, \ldots, y_n)$ and estimate $f(x)$ by:

$$\widehat{f}(x) := \frac{\sum_{i=1}^n y_i \frac{1}{h} K\left(\frac{x-x_i}{h}\right)}{\frac{1}{h}\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}.$$

This estimator is derived in the following way. Let us use the notation $K_h(z) = \frac{1}{h}K(\frac{z}{h})$. Let $e(x, y) = \frac{1}{n}\sum_{j=1}^n \delta_{x_j}(x)\delta_{y_j}(y)$ denote the empirical density for pairs (x_i, y_i). We replace $\delta_{x_j}(x)\delta_{y_j}(y)$ by $K_h(x - x_j)\widetilde{K}_h(y - y_j)$ (the kernels for the x and y variables may be different). Marginalising over y gives a 'density estimate' for the x-variable:

$$\widehat{p}(x) = \frac{1}{n}\sum_{j=1}^n K_h(x - x_j).$$

Now,

$$f(x) = \mathbb{E}[Y|X = x] = \int y p_{y|X}(y|x)dy = \int y \frac{p_{X,Y}(x, y)}{p_X(x)}dx$$

which we approximate by

$$\widehat{f}(x) = \frac{\frac{1}{n}\sum_{j=1}^n K_h(x - x_j)\int y\widetilde{K}_h(y - y_j)dy}{\frac{1}{n}\sum_{j=1}^n K_h(x - x_j)} = \frac{\sum_{j=1}^n y_j K_h(x - x_j)}{\sum_{j=1}^n K_h(x - x_j)}.$$

The kernel regression method will be illustrated in the tutorial.

1.7 Projection Pursuit Density Estimation

Projection pursuit density estimation constructs a density p over $\mathbb{R}^d$. At the kth interation the formula is:

$$\widehat{p}^{(k)}(x) = \widehat{p}^{(k-1)}(x)g_k((a_k, x))$$

where the a_ks are unit vectors in $\mathbb{R}^d$; $(a, x) = \sum_{j=1}^d a_j x_j$.

For a density p over $\mathbb{R}^d$, let $p_a : \mathbb{R} \to \mathbb{R}_+$ denote the density (i.e. $\int_{-\infty}^{\infty} p_a(z)dz = 1$, $p_a(z) \geq 0$ for all $z \in \mathbb{R}_+$) such that $p_a(z) \propto p(x)$ for $\{x : (a, x) = z\}$. This is the *marginal density* for direction a (a is a unit vector). The function g_j satisfies:

$$g_j((a_j, x)) = \frac{p_{a_j}((a_j, x))}{\widehat{p}_{a_j}((a_j, x))} \qquad j = 1, 2, \ldots, k$$

The projection pursuit density estimation algorithm is as follows:

1. The input is $\mathcal{I} = \{x_i : i = 1, \ldots, x_n\}$, an observed random sample of d-variate observations. Firstly centre the data to have mean 0 and sphere the data to have covariance I_r (Take principle components and divide through so that each PC has unit variance).

2. Initialise: choose $\widehat{p}^{(0)}$ as the initial density estimate, usually standard Gaussian.

3. Do $j = 1, 2, \ldots$

   (a) Find the unit vector $a_j \in \mathbb{R}^d$ for which the model marginal p_{a_j} differs most from the current *estimated* marginal $\widehat{p}_{a_j}$ along a_j. The choice of direction will not, in general, be unique.

   (b) Having found a suitable a_j, define a univariate 'augmenting function'

   $$g_j((a_j, x)) = \frac{p_{a_j}((a_j, x))}{\widehat{p}_{a_j}((a_j, x))}$$

   (c) Update the previous estimate so that

   $$\widehat{p}^{(j)}(x) = \widehat{p}^{(j-1)}(x) g_j((a_j, x))$$

1.7.1 Projection Index

The original aim of projection pursuit was to automate the search for interesting one dimensional marginals. These could be encoded by a *projection index*, of the form:

$$I(p) = \int J(p(z))p(z)dz = \mathbb{E}_p[J(p)]$$

The choice of J depends on the specific context. For example (recall that we first centre and sphere the data), the Gaussian distribution maximises entropy, so that taking $J(p) = \log p$ is a good way to detect departures from Gaussianity.

If we are interested in a one-dimensional projection, then let z_i denote the projected values; I can be estimated (along the ray) as:

$$\widehat{I}(p) = \frac{1}{n} \sum_{i=1}^{n} J(\widehat{p}(z_i)).$$

Exercises: Non-parametric Density Estimation

1. Let

$$p(x) = \sum_{l=0}^{L-1} y_l \mathbf{1}_{T_l}(x)$$

where $T_l = [t_l, t_{l+1})$, $t_{l+1} - t_l = h$ for $l = 0, 1, \ldots, L-1$, $h\sum_{l=0}^{L-1} y_l = 1$ and $y_0, \ldots, y_{L-1}$ are unknown paramters, subject to the constraint.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from $p(x)$ and show that the histogram

$$\widehat{p}(x) = \frac{1}{nh} \sum_{l=0}^{L-1} N_l \mathbf{1}_{T_l}(x)$$

is the unique MLE of $p(x)$, where $N_l = \sum_{j=1}^{n} \mathbf{1}_{T_l}(X_j)$.

2. The *average shifted histogram (ASH)* is constructed by taking m histograms $\widehat{p}_1, \ldots, \widehat{p}_m$, constructed from the same data, each with the same bin width h_n, but with different bin origins $0, \frac{h_n}{m}, \frac{2h_n}{m}, \ldots, \frac{(m-1)h_n}{m}$ respectively and then averaging them:

$$\widehat{p}_{\text{ASH}}(x) = \frac{1}{m} \sum_{k=1}^{m} \widehat{p}_k(x).$$

The resulting estimator is piecewise constant over intervals $[k\delta, (k+1)\delta)$ where $\delta = \frac{h_n}{m}$.

   (a) Derive the integrated variance and integrated squared bias of the average shifted histogram.

   (b) (For the enthusiast) Show that the asymptotic IMSE is

$$\text{AIMSE} = \frac{2}{3nh_n}\left(1 + \frac{1}{2m^2}\right) + \frac{h_n^2}{12m^2} R(p') + \frac{h_n^4}{144}\left(1 - \frac{2}{m^2} + \frac{3}{5m^2}\right) R(p''),$$

   where $R(g) = \int g^2$.

3. By considering m shifted histograms (as in the previous question), let $B_k = [k\delta, (k+1)\delta)$ be the kth bin of the ASH, where $\delta = \frac{h_n}{m}$ and let ν_k denote the bin count of bin B_k. Note that the ASH bin count for B_k is the average of the bin counts of the m shifted histograms, each of width δ, in bin B_k. Show that for $x \in B_k$ and m large, the ASH can be expressed as a kernel density estimator with triangular kernel on $(-1, 1)$ (that is

$$K(x) = \frac{1}{2}(1+x)\mathbf{1}_{[-1,0]} + \frac{1}{2}(1-x)\mathbf{1}_{(0,1]}.$$

4. Verify the asymptotic results stated in lectures for the variance and bias of a kernel density estimator:

$$\begin{cases} \text{Var}(\widehat{p}(x)) \simeq \frac{R(K)p(x)}{nh_n} - \frac{p(x)^2}{n} \\ \text{bias}(\widehat{p}(x)) \simeq \frac{1}{2}\sigma_K^2 h_n^2 p''(x) \end{cases}$$

the notation given in the lecture.

5. Let $\widehat{F}_n$ denote the empirical distribution function based on a random sample of size n. Rosenblatt's density estimator is:

$$\widehat{p}_n(x) = \frac{1}{h}\left(\widehat{F}_n\left(x + \frac{h}{2}\right) - \widehat{F}_n\left(x - \frac{h}{2}\right)\right).$$

   (a) Show that this estimator is a kernel density estimator and specify the type of kernel that corresponds to this estimator.

   (b) Compute the asymptotic bias and variance of Rosenblatt's estimator and hence compute the AMISE of this estimator.

Answers

1. The likelihood function is:

$$L(y_0, \ldots, y_{L-1}) = \prod_{j=1}^{n} p(X_j) = \prod_{i=0}^{L-1} y_i^{N_i}$$

$$\log L = \sum_{i=0}^{L-1} N_i \log y_i \qquad \sum y_i = 1$$

$$\frac{\partial}{\partial y_i} \log L = \frac{N_i}{y_i} - \frac{N_{L-1}}{y_{L-1}} = 0 \qquad \text{for MLE}$$

$$y_i = \alpha N_i \qquad \alpha = \frac{1}{n} \qquad \widehat{y}_{i,MLE} = \frac{N_i}{n}.$$

2. For the squared bias, the bins are of length $\delta = \frac{h_n}{m}$ and the same argument as in lectures gives:

$$AISB \simeq \frac{h_n^2}{12m^2} R(p')$$

For a histogram, with bins T_l and $x \in T_l = [t_l, t_{l+1})$,

$$\widehat{p}(x) = \frac{1}{nh} \sum_{j=1}^{n} \mathbf{1}_{[t_l, t_{l+1})}(Y_j)$$

To compute the variance of the average shifted histogram,

$$\begin{aligned} \mathrm{Var}(\widehat{p}_{ASH}(x)) &= \frac{1}{m^2} \sum_{k_1,k_2=1}^{m} \mathrm{Cov}(\widehat{p}_{k_1}(x), \widehat{p}_{k_2}(x)) \\ &= \frac{1}{m^2 n h_n^2} \mathrm{Cov}\left(\mathbf{1}_{T_{k_1,x}}(Y), \mathbf{1}_{T_{k_2,x}}(Y)\right) \end{aligned}$$

where $T_{k,x}$ denotes the bin in histogram k which contains the point x and Y is a random variable with density p. Now,

$$\mathbb{P}(Y \in T_{k,x}) = \int_{T_{k,x}} p(x)dx \simeq p(x)h_n.$$

Also, $|T_{k_1,x} \cap T_{k_2,x}| = h_n \left(1 - \frac{|k_1 - k_2|}{m}\right)$ for $|k_1 - k_2| \leq m$ and 0 otherwise. It follows that

$$\mathrm{Var}(\widehat{p}_{ASH}(x)) \simeq \left(\frac{p(x)}{mnh_n} - \frac{p(x)^2}{mn}\right) + \frac{2}{m^2} \sum_{k_1=1}^{m-1} \sum_{k_2=k_1+1}^{m} \left(1 - \frac{k_2 - k_1}{m}\right) \frac{p(x)}{nh_n} - \frac{p(x)^2}{n})$$

Since $h_n \to 0$, it follows that the terms with p^2 do not contribute. For standard sums, look up wikipedia:

$$\sum_{j=1}^{m} j = \frac{1}{2}m(m-1) \qquad \sum_{j=1}^{k} j^2 = \frac{m^3}{3} + \frac{m^2}{2} + \frac{m}{6}$$

giving:

$$\begin{aligned}\sum_{k_1=1}^{m-1}\sum_{k_2=1}^{m-k_1} k_2 &= \frac{1}{2}\sum_{j=1}^{m-1} j(j-1) = \frac{1}{2}\left(\frac{(m-1)^3}{3} + \frac{(m-1)^2}{2} + \frac{m-1}{6}\right) - \frac{1}{4}(m-1)(m-2) \\ &= \frac{(m-1)^3}{6} + \frac{5(m-1)}{12}.\end{aligned}$$

From this, it follows that:

$$\mathrm{Var}(\widehat{p}_{ASH}(x)) \simeq \frac{2}{3}(1 + \frac{1}{2m})\frac{p(x)}{nh_n}$$

(dropping terms of $o(\frac{1}{m})$). Hence

$$AIV \simeq \frac{2}{3nh_n}.(1 + \frac{1}{2m})$$

3. We may write

$$\widehat{p}_{ASH}(x) = \frac{1}{nmh}\sum_{j=1}^{n}\sum_{i=1}^{m} \mathbf{1}_{T_{i,x}}(Y_j)$$

where $T_{i,x}$ is a bin of width h which contains x for histogram i, $i = 1, \ldots, m$ and $Y_1, \ldots, Y_n$ are the observations. Therefore

$$\widehat{p}_{ASH}(x) = \frac{1}{nh}\sum_{j=1}^{n} K_{x,h}(Y_j)$$

where

$$K_{x,h}(y) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{T_{i,x}}(y).$$

$T_{i,x}$ denotes the bin for histogram i which contains x, Now

$$x \in \left[jh + \frac{k}{m}, (j+1)h + \frac{k}{m}\right) \Rightarrow \left\lfloor \frac{x - \frac{k}{m}}{h} \right\rfloor = j$$

so that

$$K_{x,h}(y) = \frac{1}{m}\sum_{k=1}^{m} \mathbf{1}_{[(\lfloor \frac{x-\frac{k}{m}}{h} \rfloor + \frac{k}{m})h, (\lfloor \frac{x-\frac{k}{m}}{h} \rfloor + 1 + \frac{k}{m})h)}(y)$$

Now it is a counting argument: add $\frac{1}{m}$ to the sum for each k such that

$$\lfloor \frac{x-\frac{k}{m}}{h} \rfloor - \frac{x}{h} + \frac{k}{m} \leq \frac{y-x}{h} < \lfloor \frac{x-\frac{k}{m}}{h} \rfloor - \frac{x}{h} + 1 + \frac{k}{m}$$

giving

$$K_{x,h}(y) \simeq \frac{j}{m} \qquad 1 - \frac{j}{m} \leq \frac{|y-x|}{h} < 1 - \frac{(j+1)}{m} \qquad j = 0, \ldots, m-1$$

from which the result follows.

4. For the Kernel density estimator:

$$\widehat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_j}{h}\right).$$

$$\begin{aligned}
\mathbb{E}[\widehat{p}(x)] - p(x) &= \frac{1}{nh}\sum_{i=1}^{n} \int \frac{1}{h} K(\frac{x-y}{h})(p(y) - p(x))dy \\
&= -\int K(z)(p(x - zh) - p(x))dz \simeq hp'(x)\int zK(z)dz - \frac{h^2}{2}p''(x)\int z^2 K(z)dz
\end{aligned}$$

Since a kernel is symmetric, it follows that $\int zK(z)dz = 0$, hence

$$\text{bias}(\widehat{p}(x)) = p(x) - \mathbb{E}[\widehat{p}(x)] = \frac{h^2}{2}p''(x)\int z^2 K(z)dz = \frac{h^2}{2}p''(x)\sigma_K^2.$$

For the variance,

$$\text{Var}(\widehat{p}(x)) = \frac{1}{nh^2}\text{Var}\left(K(\frac{X-x}{h})\right)$$

$$\begin{aligned}
\mathbb{E}[K(\frac{X-x}{h})] &= \int p(y)K(\frac{y-x}{h})dy \\
&= h\int p(x + hz)K(z)dz \simeq hp(x) + h^2 p'(x)\int zK(z)dz + \frac{h^3}{2}p''(x)\int z^2 K(z)dz \\
&= hp(x) + \frac{h^3}{2}p''(x)\int z^2 K(z)dz.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[K(\frac{X-x}{h})^2] &= \int p(y)K(\frac{y-x}{h})^2 dy \\
&= h\int p(x+hz)K(z)^2 dz \\
&\simeq hp(x)\int K^2(z)dz + h^2 p'(x)\int zK(z)^2 dz + \frac{h^3}{2}p''(x)\int z^2 K(z)^2 dz \\
&= hp(x)R(K) + \frac{h^3}{2}p''(x)\int z^2 K(z)^2 dz.
\end{aligned}$$

From this, it follows that

$$\mathrm{Var}(\widehat{p}(x)) = \frac{1}{nh}p(x)R(K) - \frac{1}{n}p(x)^2.$$

5. (a)

$$\widehat{p}_n(x) = \frac{1}{nh}\sum_{j=1}^{n}\mathbf{1}_{[x-\frac{h}{2},x+\frac{h}{2}]}(X_j) = \frac{1}{nh}\sum_{j=1}^{n}\mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}(\frac{X_j-x}{h})$$

so it has kernel function $K(z) = \mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}$.

(b) From previous exercise:

$$R(K) = \int K(x)^2 dx = 1 \qquad \sigma_K^2 = \int_{-1/2}^{1/2} x^2 dx = \frac{1}{12}$$

Plug them in to get the pointwise variance and bias.

$$\mathrm{AMISE} = \frac{1}{nh} - \frac{1}{n}R(p) + \frac{1}{576}h^4 R(p'').$$