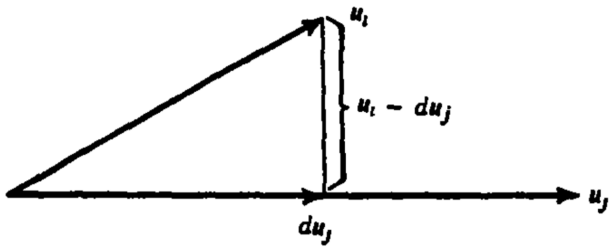
1 A geometric interpretation of $\hat{\rho}_{i,j}$

The sample (Pearson) correlation coefficient:

$$\hat{\rho}_{i,j} = \frac{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)^2}\sqrt{\sum_{\alpha=1}^{N}(x_{j\alpha} - \bar{x}_j)^2}}$$

1.1 The cosine of the angle between vectors u_i and u_j

- Choose the scalar d so the vector du_j is orthogonal to $u_i - du_j$.



- du_j is the projection of u_i on the vector u_j
- $0 = du_j^T(u_i - du_j) = d(u_j^T u_i - du_j^T u_j)$.
- Thus, we have $d = \frac{u_j^T u_i}{u_j^T u_j}$.
- The projection of u_i on the vector u_j is $du_j = \frac{u_j^T u_i}{u_j^T u_j} u_j$.
- The absolute value of the cosine of the angle between u_i and u_j is the length of du_j divided by the length of u_i.

$$\frac{\sqrt{du_j^T(du_j)}}{\sqrt{u_i^T u_i}} = \sqrt{\frac{du_j^T u_j d}{u_i^T u_i}} = \sqrt{\frac{du_j^T u_j}{u_i^T u_i}\frac{u_j^T u_i}{u_j^T u_j}} = \sqrt{\frac{du_j^T u_i}{u_i^T u_i}}$$
$$= \sqrt{\frac{u_j^T u_i}{u_i^T u_i}\frac{u_j^T u_i}{u_j^T u_j}} = \sqrt{\frac{(u_j^T u_i)(u_j^T u_i)}{(u_i^T u_i)(u_j^T u_j)}}.$$

- $u_i^T u_j = u_j^T u_i$.
- The cosine of the angle between vectors u_i and u_j is

$$\frac{u_i^T u_j}{\sqrt{(u_i^T u_i)(u_j^T u_j)}}.$$

1.2 A geometric interpretation of $\hat{\rho}_{i,j}$

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \dots & \dots & \dots & \dots \\ x_{p1} & x_{p2} & \dots & x_{pN} \end{pmatrix} = \begin{pmatrix} u_1^T \\ u_2^T \\ \dots \\ u_p^T \end{pmatrix}$$

- $u_i = (x_{i1}, x_{i2}, \dots, x_{iN})^T$: the transpose of the ith row of observations matrix X.
- We introduce a vector e such that $e = (1, 1, \dots, 1)^T \in \mathbb{R}^{N\times 1}$. (ϵ)
- Recall that the projection of u_i on the vector u_j is $\frac{u_j^T u_i}{u_j^T u_j} u_j$.
- The projection of u_i on the vector e is

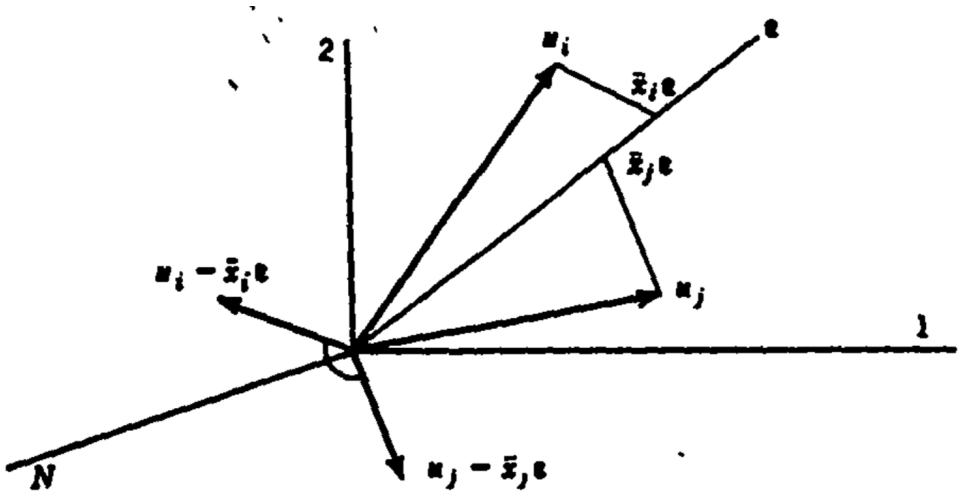
$$\frac{e^T u_i}{e^T e} e = \frac{\sum_{\alpha=1}^{N} x_{i\alpha}}{\sum_{\alpha=1}^{N} 1} = \frac{1}{N} \sum_{\alpha=1}^{N} x_{i\alpha} e = \bar{x}_i e = (\bar{x}_i, \bar{x}_i, \dots, \bar{x}_i)^T.$$

- The cosine of the angle between vectors $u_i - \bar{x}_i e$ and $u_j - \bar{x}_j e$ is

$$\frac{(u_i - \bar{x}_i e)^T (u_j - \bar{x}_j e)}{\sqrt{[(u_i - \bar{x}_i e)^T (u_i - \bar{x}_i e)][(u_j - \bar{x}_j e)^T (u_j - \bar{x}_j e)]}}.$$

- Note that $\alpha^T \beta = \sum \alpha_i \beta_i$.
- The numerator is $(u_i - \bar{x}_i e)^T (u_j - \bar{x}_j e) = \sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)$.
- The denominator is the product of the length.
- The squared length of $u_i - \bar{x}_i e$ is $(u_i - \bar{x}_i e)^T (u_i - \bar{x}_i e) = \sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)^2$.
- The squared length of $u_j - \bar{x}_j e$ is $(u_j - \bar{x}_j e)^T (u_j - \bar{x}_j e) = \sum_{\alpha=1}^{N} (x_{j\alpha} - \bar{x}_j)^2$.
- The cosine of the angle between vectors $u_i - \bar{x}_i e$ and $u_j - \bar{x}_j e$ is

$$\begin{aligned} &\frac{(u_i - \bar{x}_i e)^T (u_j - \bar{x}_j e)}{\sqrt{[(u_i - \bar{x}_i e)^T (u_i - \bar{x}_i e)][(u_j - \bar{x}_j e)^T (u_j - \bar{x}_j e)]}} \\ &= \frac{\sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)^2} \sqrt{\sum_{\alpha=1}^{N} (x_{j\alpha} - \bar{x}_j)^2}} \\ &= \hat{\rho}_{i,j} \end{aligned}$$



2 Nonparametric density estimation

Textbook: All of Nonparametric Statistics, Larry Wasserman.

- Parameter Estimation: the process of using sample data to estimate the parameters of the selected distribution.
- Nonparametric Estimation: the process of using prior knowledge to estimate the model without assuming the distribution.

Let F be a distribution with probability density $f = F'$ and let $X_1, \ldots, X_n \sim F$ be an i.i.d sample from F. The goal of nonparametric density estimation is to estimate f with as few assumptions about f as possible. We denote the estimator by $\hat{f}_n$.

We will evaluate the quality of an estimator $\hat{f}_n$ with the risk, or integrated mean squared error, $R = \mathbb{E}(L)$ where

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

is the integrated squared error loss function. The estimators will depend on some smoothing parameter h and we will choose h to minimize an estimate of the risk.

2.1 Kernels

The word kernel refers to any function K such that $K(x) \geq 0$ and

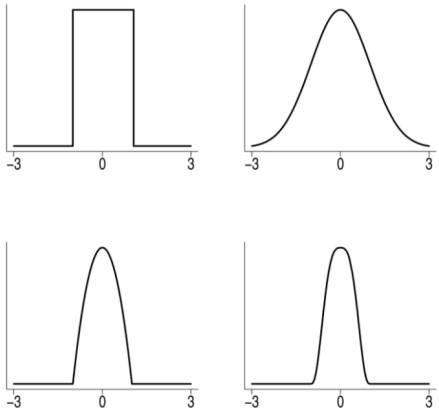
$$\int K(x)dx = 1, \ \int xK(x)dx = 0, \ \sigma_K^2 \equiv \int x^2 K(x)dx > 0. \tag{1}$$

Let function $I(x)$ satisfy

$$I(x) = \begin{cases} 1 & \text{if } |x| \leq 1, \\ 0 & \text{if } |x| > 1. \end{cases}$$

Some commonly used kernels are the following:

- **the boxcar kernel:** $K(x) = \frac{1}{2}I(x)$,
- **the Gaussian kernel:** $K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$,
- **the Epanechinikov kernel:** $K(x) = \frac{3}{4}(1-x^2)I(x)$,
- **the tricube kernel:** $K(x) = \frac{70}{81}(1-|x|^3)^3 I(x)$.



2.2 Kernel Density Estimator

Given a kernel K and a positive number h, called the bandwidth, the kernel density estimator is defined to be

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-x_i}{h}).$$

In fact, the choice of kernel K is not crucial, but the choice of h is important. In general we will let the bandwidth depend on the sample size n so we write h_n seriously.

Theorem 1. *Assume that f is continuous at x and that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Then $\hat{f}_n(x) \xrightarrow{P} f(x)$.*

That is, $\hat{f}_n(x)$ converges to $f(x)$ in probability, for every $\epsilon > 0$,

$$P(|\hat{f}_n(x) - f(x)| > \epsilon) \to 0, \text{ as } n \to \infty.$$