Statistics

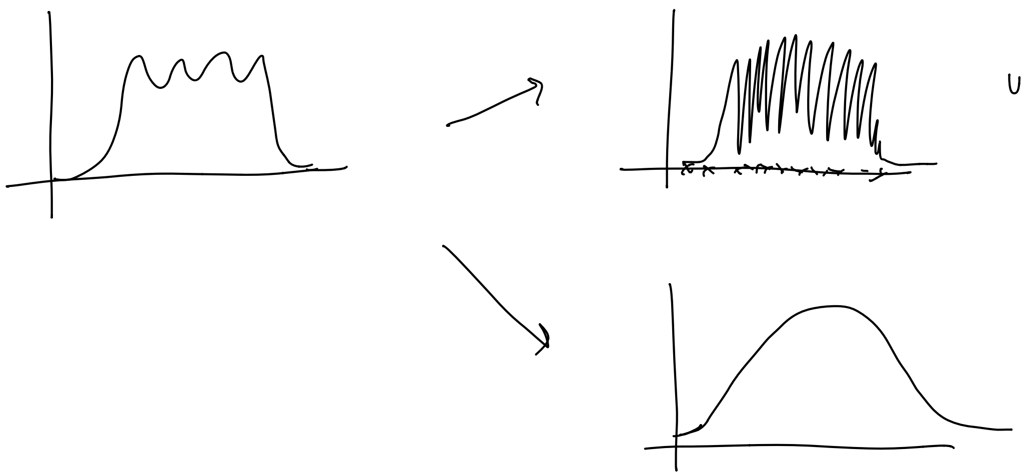
April 28, 2021

Density Estimation

Setup: Observe data $X_1, \dots, X_n \sim F$, and the density is $\underline{f} = F'$.

Goal: Estimate f using as few assumptions as possible.

$\Rightarrow$ Still a smoothing problem:



One possible measure of the error is the L^2 error:

$$\text{Loss} = L = \int \left(\hat{f}(x) - f(x) \right)^2 dx$$

$$= \| \hat{f} - f \|^2$$

$$= \underbrace{\| \hat{f} \|^2 - 2 (\hat{f}, f)}_{J} + \| f \|^2 = J + \| f \|^2 = J + C.$$

$$(\hat{f}, f) = \text{inner product of } \hat{f} \text{ with } f$$

$$= \int \hat{f}(x) f(x)\, dx$$

$$= \int \hat{f}(x)\, dF(x)$$

$$= \mathbb{E}(\hat{f}(x))$$

Goal is to estimate J.

As before, denote by $\hat{f}_{(-i)}$ the estimator obtained by leaving out X_i:

<u>Def:</u> CV estimate of the risk:

$$\hat{J} = \|\hat{f}\|^2 - \frac{2}{n} \sum \hat{f}_{(-i)}(X_i)$$

Histograms

Assume we are estimating f on $[0,1]$, set $h = \frac{1}{m}$, then we have bins $B_1 = [0,h)$, $B_2 = [h, 2h)$, ..., $B_j = [(j-1)h, jh)$.

Denote by $Y_j = \#\, X_i$'s in bin j.

$\hat{p}_j = Y_j / n$. $\leftarrow$ estimate probability of ending up in bin j

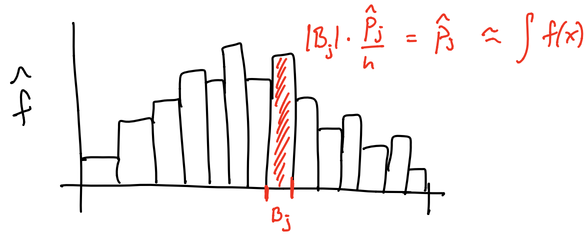
$p_j = \int_{B_j} f(x)\,dx$ $\leftarrow$ true probability of landing in bin j.

$= \mathbb{P}(X \in B_j)$.

Histogram estimator: $\displaystyle \hat{f}(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} \mathbb{1}(x \in B_j)$.

Maybe a surprising factor?

Why not just $\hat{p}_j$?



$$\mathbb{E}(\hat{f}(x)) = \frac{\mathbb{E}(\hat{p}_j)}{h} = \frac{p_j}{h} = \frac{1}{h}\int_{B_j} f(x)\,dx \approx \frac{1}{h} f(x) \cdot h = f(x).$$

<u>Thm</u> $\mathbb{E}(\hat{f}(x)) = \frac{p_j}{h}$ for $x \in B_j$

$$\mathrm{Var}(\hat{f}(x)) = \frac{p_j(1-p_j)}{n h^2}$$

and the risk can be computed as:

Thm Assume that f' is "absolutely continuous" and $\int (f')^2 < \infty$, then

$$R(\hat{f}, f) = \frac{h^2}{12} \int (f'(x))^2 dx + \frac{1}{nh} + O(h^2) + O\left(\frac{1}{n}\right)$$

and for fixed n, the minimum occurs at

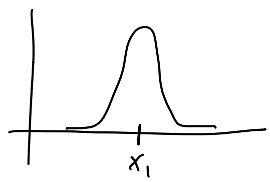
$$h_* = \frac{1}{n^{1/3}} \left(\frac{6}{\int f'^2 dx} \right)^{1/3} \sim \frac{1}{n^{1/3}}$$ and

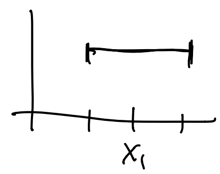
then $R(\hat{f}, f) \sim C \frac{1}{n^{2/3}}$.

Remember: Since f is not known, minimize $\hat{J}$ instead since it can actually be computed.

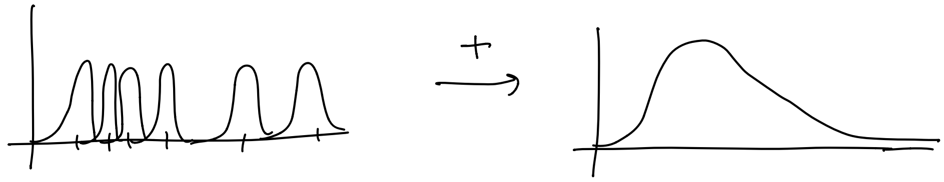
Kernel Density Estimators

If you had only one data point, x_1, what would you do?





Idea: Place a local kernel at each data point, and sum:



How wide should the kernel be?

Def: Kernel density estimator:

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

$\int K = 1$

$\int x K = 0$

$\sigma_k^2 = \int x^2 K < \infty$

Thm If f is continuous at x, and $h \to 0$, $nh \to \infty$, then $\hat{f}(x) \xrightarrow{\mathbb{P}} f(x)$:

$$\lim_{n\to\infty} \mathbb{P}\left(| \hat{f}(x) - f(x) | > \varepsilon \right) = 0$$

Thm (6.28) Let $R(x) = \mathbb{E}\left((f(x) - \hat{f}(x))^2 \right)$ be the risk at x. Then

$$R(x) = \frac{1}{4} \sigma_k^4 h^4 f''(x)^2 + \frac{f(x)}{nh} \int K^2(x)\, dx + O\left(\frac{1}{n}\right) + O(h^6).$$

Proof: $\mathbb{E}(\hat{f}(x)) = \mathbb{E}\left(\frac{1}{n} \sum_i \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \right)$

$$= \mathbb{E}\left(\frac{1}{h} K\left(\frac{x-x_i}{h} \right) \right)$$

$$= \frac{1}{h} \int K\left(\frac{x-t}{h} \right) f(t)\, dt$$

$$= \int K(u) f(x - hu)\, du$$, expand f around $hu = 0$.

$$= \int K(u) \left(f(x) - \cancel{hu f'(x)} + \frac{h^2 u^2}{2} f''(x) + \dots \right) du$$

$$= f(x) + \frac{1}{2} h^2 f''(x) \underbrace{\int u^2 K(u)\, du}_{\sigma_k^2} + \dots$$

assuming K is even.

$\Rightarrow$ bias $= \mathbb{E}\left(\hat{f}(x) - f(x) \right) = \frac{1}{2} h^2 \sigma_k^2 f''(u) + O(h^4)$

Similarly, $\operatorname{var}(\hat{f}(x)) = \frac{f(x)}{nh} \int K^2(u)\, du + O\left(\frac{1}{n}\right)$.

$\Rightarrow$ $R = \text{bias}^2 + \text{variance}$

Then for the optimal bandwidth,

solve $\frac{dR}{dh} = 0 \quad \Rightarrow \quad h \sim \frac{\sigma_K^2}{n^{1/5}}$

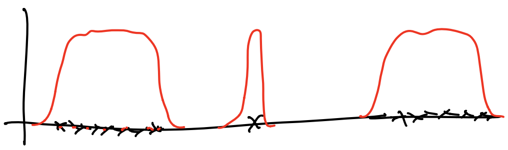
$$\Rightarrow \quad R \sim O\left(\frac{1}{n^{4/5}}\right)$$

vs for the histogram: $R \sim O\left(\frac{1}{n^{2/3}}\right)$

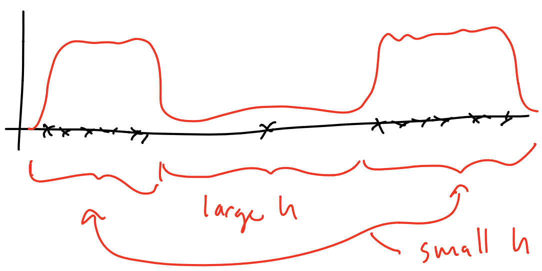
<u>Note</u> Assuming only that $\int (f'')^2 < \infty$, the rate of $\frac{1}{n^{4/5}}$ is <u>the best</u> that can be obtained (see Thm 6.31 in AoNPS.)

<u>Adaptive Methods</u> Choose h locally depending on clustering of data and other considerations.

<u>Ex</u>:



uniformly small h



<u>Multivariate Version</u>

Same mathematical idea as in the one-dimensional case.

One option is to use what is known as a product kernel:

$$K_h^d \underbrace{(x_1, \dots, x_d)}_{\vec{x}} = \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x_j - x_j'}{h_j}\right)$$

K_h^d must satisfy kernel properties.

then $\hat{f}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h^d(\vec{x}-\vec{x}_i)$

$$= \frac{1}{n}\sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x_j - x_{ij}}{h_j}\right)$$

Risk can be estimated in the same way using multivariable Taylor series for f.

Curse of Dimensionality

If we want the risk $R \sim 0.1$ at $\vec{x}=0$ for $f \sim \text{Normal}(0,1) \in \mathbb{R}^d$, using the optimal bandwidth then $n \sim c^d$!

d	n
1	4
2	19
4	223
8	43,700
9	187,000
10	842,000

Bootstrap

(Ch. 8 in AoS) (C&B 10.1.4, 10.6.5) (AoNPS Ch. 3).

Goal: Estimate standard errors and confidence sets for statistics.

Outline: Given $X_1, \dots, X_n \sim F$, statistic $T = g(X_1, \dots, X_n)$, want $\text{Var}_F(T)$.

dependence on unknown distribution F.

Ex: $T = \bar{X}$

$\text{Var}_F = \frac{\sigma^2}{n}$ if $\text{var}\, X_i = \sigma^2$

$= \int (x-\mu)\, dF(x)$

function of F.

The idea of the bootstrap:

① Estimate $Var_F(T)$ with $Var_{\hat{F}}(T)$.

$\hat{F}$ put $1/n$ mass at every X_i.

② Use simulation to approximate $Var_{\hat{F}}(T)$.

In this example, step 2 is not needed because

$$Var_{\hat{F}}(T) = \frac{\hat{\sigma}^2}{n} \quad \text{when } \hat{\sigma}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2.$$

What is simulation? Drawing samples from some distribution, and computing averages.

Ex: Draw $Y_1, \dots, Y_m$ from a distribution G, by the law of large numbers

$$\bar{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j \xrightarrow{\mathbb{P}} \mathbb{E}(Y) = \int y \, dG(y) \quad \text{as } m \to \infty.$$

Choosing m large enough means that $\bar{Y} \approx \mathbb{E}(Y)$, use this as an estimate for $\mathbb{E}(Y)$.

Also if h is some function with $\int h(y)\,dy < \infty$,

then $\frac{1}{m}\sum h(Y_j) \xrightarrow{\mathbb{P}} \mathbb{E}(h(Y)) = \int h(y)\, dG(y)$.

Ex:
$$\frac{1}{m}\sum (Y_i - \bar{Y})^2 = \frac{1}{m}\sum Y_i^2 - \left(\frac{1}{m}\sum Y_i\right)^2$$
$$\xrightarrow{\mathbb{P}} \int y^2\, dG(y) - \left(\int y\, dG(y)\right)$$
$$= Var(Y).$$

Bootstrap Variance Estimate

If we have data X_i, but F is unknown, then estimate F with $\hat{F}$, and draw from $\hat{F}$.

$\Rightarrow$ Draw $X_1^*, \ldots, X_n^*$ from $X_1, \ldots, X_n$ with replacement.

$\Rightarrow$ Compute $T^* = g(X_1^*, \ldots, X_n^*)$

Note Some of the X_i^* will be duplicates.

Variance Algorithm

DO $i = 1, \ldots, m$

Draw $X_1^*, \ldots, X_n^*$ from $\hat{F}$

Compute $T_i^* = g(X_1^*, \ldots, X_n^*)$

COMPUTE $v_{boot} = \frac{1}{m} \sum_{j=1}^{m} (T_j^* - \bar{T}^*)^2$

bootstrap estimate of the variance $\Rightarrow$ $\hat{se} = \sqrt{v_{boot}}$

We can use exactly the same algorithm to estimate the variance of median, mode, or any other integrable statistic. $\int g < \infty$.

Bootstrap Confidence Intervals

Method 1 If T is approximately normal, e.g. an MLE, the T^* is also approximately normal (and so is v_{boot})

$\Rightarrow$ a $1-\alpha$ confidence interval for T is:

$$CI = \left(T - z_{\alpha/2}\,\hat{se}_{boot} \,,\; T + z_{\alpha/2}\,\hat{se}_{boot} \right)$$

actual T from data

$\hat{se}_{boot} = \sqrt{v_{boot}}$

<u>Method 3</u> **<u>Percentile Intervals</u>** (obvious idea)

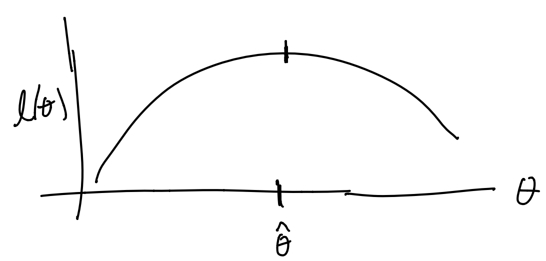
Generate $T_1^*, \dots, T_m^*$ using simulation, and let $T^*_{\alpha/2}$ be the $\alpha/2$ percentile from $T_1^*, \dots, T_m^*$

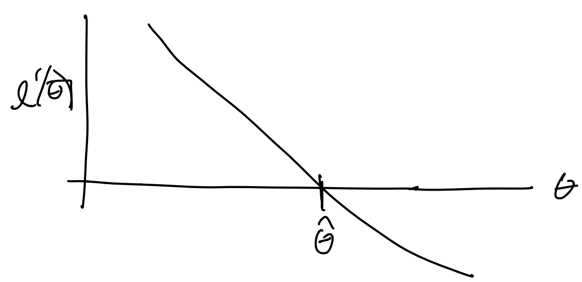
$$\Rightarrow CI = \left(T^*_{\alpha/2} \,,\; T^*_{1-\alpha/2} \right)$$

Requires some justification, see appendix.

<u>Numerically Finding the MLE</u>

Often $\ell'(\theta) = 0$ cannot be solved analytically, ($\ell(\theta) = \log \mathcal{L}(\theta)$).
however, we can solve it numerically using Newton's Method.





Expand ℓ' in a Taylor series about some guess θ^0:

$$\ell'(\theta) \approx \ell'(\theta^0) + \ell''(\theta^0)(\theta - \theta^0)$$

at the MLE, $\hat{\theta}$, $\ell'(\hat{\theta}) = 0$

$$\Rightarrow 0 \approx \ell'(\theta^0) + \ell''(\theta^0)(\hat{\theta} - \theta^0)$$

$$\Rightarrow \hat{\theta} \approx \theta^0 - \frac{\ell'(\theta^0)}{\ell''(\theta^0)}$$

Iterate: $\theta^{j+1} = \theta^j - \dfrac{l'(\theta^j)}{l''(\theta^j)}$

Can show that if θ^j is "close enough" to root, then $|\theta^{j+1} - \hat{\theta}| \sim |\theta^j - \hat{\theta}|^2$

quadratic convergence.

$\lvert\theta^j - \hat{\theta}\rvert$	$\lvert\theta^{j+1} - \hat{\theta}\rvert$
10^{-1}	10^{-2}
10^{-2}	10^{-4}
10^{-4}	10^{-8}
10^{-8}	10^{-16}

Notes

1. θ^0 can usually be estimated by the method of moments estimator.
2. l'' must be computed or approximated.

In the multivariate case, $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$

and

$$l'(\vec{\theta}) = \nabla l(\vec{\theta}) \approx \begin{pmatrix} \frac{\partial l}{\partial \theta_1}(\theta_1^0) \\ \vdots \\ \frac{\partial l}{\partial \theta_k}(\theta_k^0) \end{pmatrix} + \underbrace{\begin{pmatrix} \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \cdots \\ \vdots & & \ddots \end{pmatrix}}_{H = \text{Hessian}} \begin{pmatrix} \theta_1 - \theta_1^0 \\ \theta_2 - \theta_2^0 \\ \vdots \\ \theta_2 - \theta_k^0 \end{pmatrix}$$

At $\hat{\theta}$, $\nabla l(\hat{\theta}) = \vec{0}$, so

$$\vec{\theta}^{j+1} = \vec{\theta}^j - H^{-1}(\vec{\theta}^j)\left(\nabla l(\vec{\theta}^j)\right)$$

Multivariate Newton's Method.

Stochastic Processes

or $X(t) = X_t$
X_t is a r.v.

A stochastic process $\{X_t : t \in T\}$ is a collection of random variables indexed by t

- X_t takes values in the state space $\mathcal{X}$
- T is the index set (i.e. $\mathbb{R}$, $\mathbb{N}$, ..)
- Ex include stock prices, weather, IID ~~sequence~~ $X_1, \dots, X_n \dots$
- Recall: for $X_1, \dots, X_n$ the joint density is given by

$$f(x_1, \dots, x_n) = f(x_1)\, f(x_2|x_1)\, f(x_3|x_1, x_2) \cdots f(x_n|x_1, \dots, x_{n-1})$$

$$= \prod_{i=1}^{n} f(x_i \mid \text{past } i\text{'s})$$

Markov Chains

Def $\{X_n : n \in T\}$ is a Markov Chain

if $\mathbb{P}(X_n = x \mid X_0, \dots, X_{n-1}) = \mathbb{P}(X_n = x \mid X_{n-1})$

for all $n \in T$ and $x \in \mathcal{X}$.

$$\Rightarrow f(x_n \mid x_{n-1} \dots x_0) = f(x_n \mid x_{n-1})$$

$$\Rightarrow f(x_1, x_2, \dots x_n) = f(x_1)\, f(x_2|x_1)\, f(x_3|x_2) \cdots f(x_n|x_{n-1})$$

Questions to answer:

1. When does a MC achieve "equilibrium"? Does it at all?
2. Estimate parameters controlling the MC
3. Can we construct a MC that converges to a specified equilibrium? i.e. $X_n \rightsquigarrow F$, some given distribution

Transition Probability

Def: $p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i)$ are the transition probabilities.

If p_{ij} does not depend on n, called a homogeneous MC.

The matrix P with elements $P_{ij} = p_{ij}$ is known as the transition matrix.

Two properties of the transition probabilities:

① $p_{ij} \geq 0$

② $\sum_j p_{ij} = 1$ (Typo in book).

Each row of P is a prob. mass function.