npt 6
entle

Estimation of Functions

An interesting problem (& often difficult) in stat. is the estimation of continuous functions.
- probability density
- function we wish to integrate.

Basic Problem: (1) - we have observations of the function at specific points
-or- we have indirect measurements of the function; obs. related to derivative or integral.
(2) we wish to estimate the function with the goals: (i) providing a good fit to the obs. data
(ii) predicting values at other points.

There are a variety of approaches to estimating functions - some of which we have already discussed: - MLE for parametric family
(in general assuming underlying form & fitting parameters)

this chpt. - representation as a linear comb of basis funct.

future. - filtering a kernel est.

before we begin:
Notation: In general
We wish to estimate $f: \mathbb{R}^d \to R$ by an estimator $\hat{f}$. Just as $\hat{\Theta}$ estimates Θ, $\hat{f}$ is a r.v. with an underlying prob. distribution.
(and hence $\hat{f}$ differs from an approx of f).
How useful $\hat{f}$ is as an estimator depends on its distribution - we will look at exp val, variance, bias, etc. & will be clearer when we discuss measures by which we evaluate p.d. estimators.

Before we proceed to methods of estimation, let us develop methods of comparing functions. We will use these ideas throughout next few chapters.

Compare to vector spaces.

Inner products & Norms (a crash course).

Def (1): Let f & g be real valued functions over the domain D, then the inner product of f & g denoted $\langle f, g \rangle$ is

$$\langle f, g \rangle = \int_D f(x) g(x) \, dx$$

if the integral exists.

Notes:

(1) $\langle f, g \rangle_D$ is used if concern about ambiguity of domain

(2) this is the Lebesgue integral, but most of the time Riemann integrability suffices.

(3) The def holds (w/ slight variation) for complex functions - outside the scope of this course

(4) To avoid integrability issues, we restrict our discussion to functions whose inner product with themselves exist.

$$\langle f, f \rangle = \int_D |f(x)|^2 \, dx$$

(5) Cauchy - Schwarz holds

$$\langle f, g \rangle \leq \langle f, f \rangle^{1/2} \langle g, g \rangle^{1/2}$$

Pg 130 prob.

(6) Sometimes we define I.P.S. in terms of a weight function, $w(x)$, w.r.t. measure μ where $d\mu = w(x) \, dx$

$$\langle f, g \rangle_{(\mu; D)} = \int_D f(x) g(x) w(x) \, dx$$

(7) Inner products of functions are linear

$$\langle af + g, h \rangle = a \langle f, h \rangle + \langle g, h \rangle .$$

Def 2: The <u>norm</u> of a function f, denoted $\|f\|$, is a mapping into the nonneg. reals s.t.

- If $f \neq 0$, then $\|f\| > 0$,
- $\|0\| = 0$
- $\|af\| = |a| \, \|f\| \quad \forall a \in \mathbb{R}$
- $\|f+g\| \leq \|f\| + \|g\|$

Notes: 1. Often defined in terms of some inner prod of f with itself → NOT ALL NORMS DEFINED THIS WAY!

2. Lp norm

$$\|f\|_p = \left(\int_D |f(x)|^p w(x)\, dx \right)^{1/p}$$

where $w(x)$ is some weight function. (often $w(x) = 1$ over finite domains).

fairly common

3. The L_2 norm is

$$\|f\|_2 = \langle f, f \rangle^{1/2}$$

4. The L_∞ norm

$$\|f\|_\infty = \sup |f(x) w(x)|$$

5. Norms are used to measure the difference between functions $\|f-g\|$ norm of error.

6. A <u>normal function</u> is one whose $\|f\| = 1$.

7. For some methods (esp. iterative) it is important to know if a seq of functions converges & this is discussed on pg 132, we will assume necessary convergence in this course.

Banach Hilbert Spaces etc.

Now that we have a way of comparing functions let's discuss the stat. prop. we use to determine the usefulness of an estimation

Two types 1. Pointwise Prop
2. Global. Prop.

Skip ahead to 6.2. Wendle.

Pointwise Prop. of Function Estimators.

* $\hat{f}$ estimates $f: D \to \mathbb{R}$

The pointwise prop. of estimators of functions at a given point are analogous to those of estimators of parameters and involve determining exp. values & variances of the r.v. $\hat{f}$ wrt its pdf.

Bias

The bias of $\hat{f}$ at x is given by

$$\text{Bias}(\hat{f}(x)) = E[\hat{f}(x)] - f(x).$$

If $\text{Bias}(\hat{f}(x)) = 0$, then $\hat{f}$ is unbiased at x. If $\hat{f}$ is unbiased at every point x in D, then we say the estimator is pointwise unbiased.

Variance

The variance of $\hat{f}$ at x is given by

$$V[\hat{f}(x)] = E[(\hat{f}(x) - E[\hat{f}(x)])^2].$$

Meas Squared Error (MSE)

The MSE of $\hat{f}$ at x is

$$\begin{aligned} MSE[\hat{f}(x)] &= E[(\hat{f}(x) - f(x))^2] \\ &= V(\hat{f}(x)) + (\text{Bias}(\hat{f}(x)))^2. \end{aligned}$$

Comment: Ideally one would look for an unbiased estimator w/ small variance. However, in func est. often an est. w/ small bias has a much lower variance (or no unbiased est. exists) and hence MSE is a good measure of comparison.

Mean Absolute Error (MAE)

The MAE of $\hat{f}$ at x is given by

$$MAE[\hat{f}(x)] = E(|\hat{f}(x) - f(x)|)$$

However, does not decompose & can be difficult to work with.

Consistency

Similarly to ~~[illegible]~~ an est of a parameter, $\hat{f}$ is said to be pointwise consistent if

$$E[\hat{f}(x)] \to f(x)$$

for each x as the sample size $n \to \infty$.
Convergence is usually in terms of prob.
(See sec 1.6).

Sec 6.3 Gentle

Global Properties of Est. of Functions

Rather than focusing on pointwise prop, it is often of greater interest to study the stat. prop of $\hat{f}$ over the whole domain D of f. These measures are defined in terms of f & $\hat{f}$ w/ no indication on x. Often they are ① integration of point wise prop.
② defined in terms of a norm the function.

The main tool for comparing f & $\hat{f}$ is the L_p norm of the difference $\hat{f} - f$:

$$\|\hat{f} - f\|_p = \left(\int_D |\hat{f}(x) - f(x)|^p \, dx\right)^{1/p}$$

which requires ① $\hat{f}$ defined over all of D
② The integral exists.

Other useful measures:

- L_1: Integrated absolute error (IAE)
$$IAE(\hat{f}) = \int_D |\hat{f}(x) - f(x)| dx$$

- L_2: Integrated squared error (ISE)
$$ISE(\hat{f}) = \int_D (\hat{f}(x) - f(x))^2 dx$$

- L_∞: Sup Absolute error (SAE)
$$SAE(\hat{f}) = \sup_{x \in D} |\hat{f}(x) - f(x)|$$

! & more pg 146 Gentle.

Now we wish to extend the ideas of bias & variance to global concepts. Most obvious extension is to integrate the pointwise measure over the domain

Bias: must be careful for regions where Bias in neg

most common IAB: Integrated Absolute Bias $IAB(\hat{f}) = \int_D |E(\hat{f}(x)) - f(x)| dx$

ISB: Integrated squared Bias $ISB(\hat{f}) = \int_D (E(\hat{f}(x)) - f(x))^2 dx$

Here if $\hat{f}$ is unbiased, then $IAB(\hat{f}) = ISB(\hat{f}) = 0$ and $Bias(\hat{f}(x)) = 0$ almost everywhere.

Comment: While it is not uncommon for a parameter estimator to be unbiased, it is unlikely for function estimators.

Variance: Integrated Variance (IV)
$$IV(\hat{f}) = \int_D V(\hat{f}(x)) = \int_D E[(\hat{f}(x) - E[\hat{f}(x)])^2] dx$$

Again, due to this lack of global unbiassedness an important measure is
Integrated Mean Squared error (IMSE)

$$IMSE(\hat{f}) = \int_D E((\hat{f}(x) - f(x))^2)dx$$

$$= IV(\hat{f}) + ISB(\hat{f})$$

Finally, if the expectation & integration can interchanged we have

$$IMSE(\hat{f}) = E\left[\int_D (\hat{f}(x) - f(x))^2\right] = E\left[ISE(\hat{f})\right] = MISE_{(\hat{f})}$$

and is called the mean integrated squared error.

Relationship: ISE - performance of $\hat{f}$ base on sample x. MISE average value w.r.t. sampling density. IMSE = MISE - accumulation of local mean squared error at every x.

Similarly
Integrated mean Absolute error (IMAE)

$$IMAE(\hat{f}) = \int_D E(|\hat{f}(x) - f(x)|)dx = E\left[\int_D |\hat{f}(x) - f(x)|dx\right]$$

$$= MIAE(\hat{f})$$ Mean integrated ab. error.

We can also extend the ideas of consistency (pg 148-149) Gentle.

Also other global Properties discussed pg 149-150. Gentle.

Now → lets learn how to estimate.

A few more defs:

Def: If each function in a linear space H can be expressed as a linear comb of functions in a set G, then G is a basis, generating set, spanning set of H. EF (8)

Back to P8 133 Gentle.

Def: A set of functions $\{g_i(x)\}$ is said to be <u>orthoganal</u> (over the domain D w.r.t. the nonneg. weight function $w(x)$) if

$$\langle g_i, g_j \rangle = \int_D g_i(x)\, g_j(x) w(x)\, dx = \begin{cases} 0 & i \neq j \\ \lambda_i > 0 & i = j \end{cases}$$

If in addition

$$\langle g_i, g_i \rangle = \int_D g_i^2(x)\, w(x)\, dx = 1 \qquad \forall i$$

the functions are called orthonormal.

Called the Fourier trig. family

Ex/ $\{1, \cos x, \sin x, \cos 2x, \sin 2x, \cos 3x, \sin 3x, \ldots\}$ over $0 \leq x \leq 2\pi$ is an orthoganal family. Pf/H.W. (3 cases)

- Also, if $\{g_i(x)\}$ is orthoganal w/ $\lambda_i \neq 1$, then $\{g_i(x)/\sqrt{\lambda_i}\}$ is orthonormal.

- If $\{g_i(x)\}$ is a set of orthoganal functions, then it is a linearly independent set.

Pf/ H.W?

- From any linearly independent set $\{g_i(x)\}$ we can always construct an orthoganal set. $\{\tilde{g}_i(x)\}$

Pf/ Gram-Schmidt Orthonormalization

$$\tilde{g}_1 = g_1$$

$$\tilde{g}_2 = g_2 - \frac{\langle \tilde{g}_1, g_2 \rangle}{\langle \tilde{g}_1, \tilde{g}_1 \rangle} \tilde{g}_1$$

$$\tilde{g}_3 = g_3 - \frac{\langle \tilde{g}_1, g_3 \rangle}{\langle \tilde{g}_1, \tilde{g}_1 \rangle} \tilde{g}_1 - \frac{\langle \tilde{g}_2, g_3 \rangle}{\langle \tilde{g}_2, \tilde{g}_2 \rangle} \tilde{g}_2$$

$$\vdots$$

$$\tilde{g}_K = g_K - \sum_{i=1}^{K-1} \frac{\langle \tilde{g}_i, g_K \rangle}{\langle \tilde{g}_i, \tilde{g}_i \rangle} \tilde{g}_i$$

Clearly $\{\tilde{g}_i\}$ is orthoganal & $\left\{\frac{\tilde{g}_i}{\|\tilde{g}_i\|}\right\}$ is orthonormal.

In our goal of estimating functions, often a 1st step is to represent the function of interest $f(x)$ as a linear combination of "simpler functions" $g_0(x), g_1(x), \ldots$ ie.

$$f(x) = \sum_{K=0}^{\infty} c_K \, g_K(x)$$

There are a variety of ways to do this, however a set $\{g_i(x)\}$ of orthoganal basis functions is often the best because they have nice properties that facilitate computations & a large body of theory.

Q: So. given $\{g_K(x)\}$ how to find c_K?

A: If f is cont & integrable over D, then

$$c_K = \langle f, g_K \rangle$$

and $\{c_K\}$ are called the Fourier coeff. of f w.r.t $\{g_K\}$.

In practice we approx f w/

$$\hat{f} \approx \sum_{K=0}^{K} c_K g_K(x)$$

which has error $f - \sum_{K=0}^{K} c_K g_K(x)$ = error = $f - \tilde{f}$

and often the MSE (mean squared err) $\|f - \tilde{f}\|^2 = \frac{1}{|D|}\|f - \sum_{K=0}^{K} c_K g_K\|^2$ is used to measure error.

Comment: ① The Fourier coef $\{c_K\}$ minimize this when $\{g_K\}$ are orthonormal.

② In statistical analysis, we will form approx $\tilde{f}$ & then estimate the coef. (more to come).

Q: What basis to use in Practice?

A: - Fourier trignometric family
- Orthoganal polynomials (Examples)
- Splines (Examples).
- Wavelets

⋮

we cover these

Orthoganal Polynomials

Pg 135 Gentle 5.3.1 pg 136 G-H

These are useful for a wide range of functions. Several widely used systems, each can be developed by starting w/ $1, x, x^2, x^3, \ldots$ and applying G-S w/ approp weight.

(See Table pg 136 Gentle)

Ex/ Legendre Range $[-1,1]$ $w(x) = 1$.

unnormalized: $g_0 = 1$ $\qquad \langle g_i, g_j \rangle = \int_{-1}^{1} g_i(x) g_j(x)\, dx$.

$$q_1 = x - \underbrace{\frac{\langle 1,x\rangle}{\langle 1,1\rangle}}_{\to 0} \cdot 1 \qquad \langle 1,x\rangle = \int_{-1}^{1} x\,dx = \frac{x^2}{2}\Big|_{-1}^{1} = 0$$

$$\langle 1,1\rangle = \int_{-1}^{1} 1\,dx = x\Big|_{-1}^{1} = 2$$

$$\boxed{q_1 = x}$$

$$q_2 = x^2 - \frac{\langle 1,x^2\rangle}{\langle 1,1\rangle}\,1 - \underbrace{\frac{\langle x,x^2\rangle}{\langle x,x\rangle}}_{\to 0}\,x \qquad \langle 1,x^2\rangle = \int_{-1}^{1} x^2\,dx = \frac{2}{3}$$

$$\langle x,x\rangle \;''$$

$$\langle x,x^2\rangle = \int_{-1}^{1} x^3\,dx = 0$$

$$\boxed{q_2 = x^2 - \frac{1}{3}}$$

and so on.

Note: ① These differ by a constant from those given in the book.
② can normalize by dividing by $\sqrt{\langle q_i, q_i\rangle}$

Orthog. poly recurrances:

For the kth poly in an orthog. set $\exists\, a_k$ (no dep on x) s.t. $q_k(x) - a_k x\, q_{k-1}(x)$ is a poly of deg $k-1$.

$$\therefore\ q_k(x) - a_k x\, q_{k-1}(x) = \sum_{i=0}^{k-1} c_i q_i(x)$$

$$q_k(x) = (a_k x + c_{k-1})q_{k-1}(x) + c_{k-2}q_{k-2}(x) + \cdots$$

& due to orthog. $c_0 = c_1 = \cdots = c_{k-3} = 0$

$\therefore$ for some a_k, b_k, c_k.

$$q_k(x) = (a_k x + b_k)q_{k-1}(x) - c_k q_{k-2}(x).$$

Ex/ Legendre

$$q_k(x) = \frac{2k-1}{k}\, x\, P_{k-1}(x) - \frac{k-1}{k} P_{k-2}(x).$$

Example pg 137 Bentle.
$f(x) = e^{-x}$ on $[-1,1]$ Fig 6.2 pg 138
shows approx using $P_0 \dots P_j$ for $j = 0, \dots 5$.

Q How to apply to a dataset $X = \{x_1, \dots, x_n\}$ from unknown density p.

A: As we did w/ MC methods, let's suppose that our function of interest can be written as

$$f(x) = g(x)\,p(x)$$

where $p(x)$ is a prob. density function.

Then for any orthog. set $\{q_i\}$ we can approx

$$f(x) = \sum_{k=0}^{\infty} c_k q_k(x)$$

where

$$\begin{aligned} c_k &= \langle f, q_k \rangle \\ &= \int_D q_k(x)\, g(x)\, p(x)\, dx \\ &= E[q_k(X)\, g(X)] \end{aligned}$$

where X is a r.v. w/ density p. $\therefore$ given $x_1, \dots, x_n$ we can unbiasedly estimate c_i by

$$\hat{c}_k = \frac{1}{n} \sum_{i=1}^{n} q_k(x_i)\, g(x_i)$$

$\therefore$ an estimator of f is

$$\hat{f}(x) = \frac{1}{n} \sum_{k=0}^{j} \sum_{i=1}^{n} q_k(x_i)\, g(x_i)\, q_k(x)$$

Not unbiased $\rightarrow$ truncation.

Finally: when $f(x)$ is itself a density

$c_K = E[q_K(X)]$ & $\hat{f}(x) = \frac{1}{n} \sum_{K=0}^{j} \sum_{i=0}^{n} q_K(x_i) q_K(x)$.

---

pg 139

Splines

Thus far we have discussed methods that use a finite subset of an infinite basis set ~~to~~ of poly to approx f. These approaches yield a smooth f(x) - cont w/ con't derivatives - however it may have a high # of oscillations (1 - deg of poly). ~~conden~~.

also if the shape of f varies greatly over the Domain these can be bad approx.

A new approach: subdivide D & use polynomials of low degree.

- New $\hat{f}(x)$ is sum of piecewise poly.
- even w/ low degree, enough subint gives good approx.
- can force smoothness by imposing cont. conditions.

This is called spline approximation.

More formally:

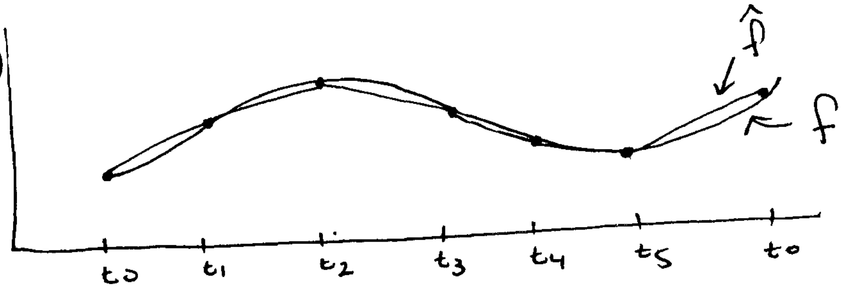
A spline function consists of poly pieces on subintervals joined together w/ certain continuity conditions.

Def: Suppose n+1 points (called knots), $t_0 < t_1 < \cdots < t_n$ subdivide our interval. Then a spline function of degree $k \geq 0$ having knots $t_0, \ldots, t_n$ is a func. S such that

1. On each interval $[t_{i-1}, t)$, S is a poly of deg $\leq k$.
2. S has a con't $(k-1)$st derivative on $[t_0, t_n]$.

Ex

$$S(x) = \begin{cases} a_0 x + b_0 & [t_0, t_1) \\ a_1 x + b_1 \\ \vdots \\ a_{n-1} x + b_{n-1} \end{cases}$$



degree 1

Cubic Splines

One of the most common spline methods uses cubic splines.

Suppose we have values

x	t_0	t_1	$\cdots$	t_n
y	y_0	y_1		y_n

Then the cubic spline on $[t_0, t_n]$ will have the form

$$S(x) = \begin{cases} S_0(x) & x \in [t_0, t_1] \\ S_1(x) & x \in [t_1, t_2] \\ \vdots & \\ S_{n-1}(x) & x \in [t_{n-1}, t_n] \end{cases}$$

where $\forall i$

$S_i(x)$ is a poly of deg ≤ 3 & continuity

1. $S_{i-1}(t_i) = y_i = S_i(t_i)$
2. $S'_{i-1}(t_i) = S'_i(t_i)$
3. $S''_{i-1}(t_i) = S''_i(t_i) = z_i$

Also, since $S_i(x)$ is cubic, $S''_i(x)$ is a linear function w/ $S''_i(t_i) = z_i$ & $S''_i(t_{i+1}) = z_{i+1}$.

$\therefore$ it can be shown that

$$(*) \quad S''_i(x) = \frac{z_i}{h_i}(t_{i+1} - x) + \frac{z_{i+1}}{h_i}(x - t_i)$$

where $h_i = t_{i+1} - t_i$. (Have 2 pts to determine line)

So we can twice integrate S''_i & use cond 1 to obtain

$$S_i(x) = \frac{z_i}{6h_i}(t_{i+1} - x)^3 + \frac{z_{i+1}}{6h_i}(x - t_i)^3 + \left(\frac{y_{i+1}}{h_i} - \frac{z_{i+1}h_i}{6}\right)(x - t_i) + \left(\frac{y_i}{h_i} - \frac{z_i h_i}{6}\right)(t_{i+1} - x).$$

And if we knew $z_0, \dots, z_n$ we would know $S(x)$.

So, we use cond. 2 $S_i'(t_i) = S_{i-1}'(t_i)$ to show that

$$(**) \quad h_{i-1} z_{i-1} + 2(h_i + h_{i-1}) z_i + h_i z_{i+1} = \frac{6}{h_i}(y_{i+1} - y_i) - \frac{6}{h_{i-1}}(y_i - y_{i-1})$$

is a system of $n-1$ eq. & we can select z_0 & z_n & solve.

Often $z_0 = z_n = 0$ chosen (natural splines) & ** becomes

$$\begin{bmatrix} u_1 & h_1 & & & & 0 \\ h_1 & u_2 & h_2 & & & \\ & h_2 & u_3 & h_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & h_{n-3} & u_{n-2} & h_{n-2} \\ 0 & & & & h_{n-2} & u_{n-1} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{n-1} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \end{bmatrix}$$

where

$$h_i = t_{i+1} - t_i, \quad u_i = 2(h_i + h_{i+1})$$

$$b_i = \frac{6}{h_i}(y_{i+1} - y_i), \quad v_i = b_i - b_{i-1}.$$

Ex/
$$S(x) = \begin{cases} 7(x-2)^2 + 2(x-1)^3 & (-\infty, 1] \\ 7(x-2)^2 & [1, 3] \\ 7(x-2)^2 + 3(x-3)^3 & [3, \infty) \end{cases}$$

is a cubic spline that interpol.

x	0	1	4
y	26	7	25

Other common spline basis.

- truncated power functions: $1, x, \dots, x^p$, $((x-z_1)_+)^p, \dots, ((x-z_k)_+)^p$
- B-splines – in comp packages, specially developed set.

Most packages have Spline applications built in

EF (16)

Spline Usage.

(1) Interpolating points. - Each point is a knot & $S(x)$ goes exactly through.

more in Chpt 11.

(2) Smoothing - if points are subject to error, the spines are eval at each abscissa (x) & fitted to the ordinate (least squares).

Biggest Problem - choosing knots.

pg 294

Example Sec 10.3.1 G-H.

Estimating a density f via logspline.

Kooperberg & Stone's density est.

Basic idea:

- m knots on $[L, U]$
- $\mathcal{S}$ the m-dim space of cubic splines w/ knots $t_1, \dots, t_m$.
- $\mathcal{B} = \{1, B_1, \dots, B_{m-1}\}$ a basis for $\mathcal{S}$
- parameters $\theta = \{\theta_1, \dots, \theta_{m-1}\}$.

Assume f can be modeled by a para. density $f_{X|\theta}$ defined by

$$\log f_{X|\theta}(X|\theta) = \theta_1 B_1(x) + \cdots + \theta_{m-1} B_{m-1}(x) - c(\theta)$$

where

$$\exp(c(\theta)) = \int_L^U \exp\{\theta_1 B_1(x) + \cdots + \theta_{m-1} B(x)\}\, dx$$

Then the log likelihood of θ is

$$l(\theta | x_1, \dots, x_n) = \sum_{i=1}^{n} f_{X|\theta}(x_i|\theta)$$

for observed $x_1, \dots, x_n$. Then use MLE to find $\hat{\theta}$ and $\hat{f}(x) = f_{X|\theta}(x|\hat{\theta})$.

Then the question is where & how many knots.

The answer is each interval must contain enough data points to allow estimation & maximization & they suggest a approximate quantile method.

In practice – use software & add & delete knots is ~~[illegible]~~ to try to improve the estimation. Other strategies exist as well.

Density Estimation

Often in statistics, the function that we wish to estimate is a probability density. i.e. we are given a sample $X_1, \dots, X_n$ of iid observations from unknown density f on D and and we construct $\hat{f}$ with

- $\hat{f}(x) \geq 0 \quad \forall \; x \in D$
- $\int_D \hat{f}(x)\,dx = 1$

hoping to find $\hat{f}$ with

- small error (ex M.S.E).
- $E[\hat{f}_n(x)] \to f(x) \quad \forall \; x \in D$ as $n \to \infty$.

If it is believed that f is a parametric density $f(x|\theta)$ there are a variety of techniques to est. f.

- MLE
- MOM
- Fitting by matching quantiles.
- mixtures.
- logspline

So we assume NOT.