STAT 391 5/2/23

# Lecture 11

Kernel density estimation

selecting $h$ by CV

# Lecture Notes V – Non-parametric density estimation

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2023

Kernels ✓

Kernel density estimators ✓

Choosing $h$ by Cross-Validation and the Bias-Variance trade-off

Model Selection

**Reading**: Ch. 7

# Kernel density estimators

$$f_{h,K,\mathcal{D}}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x^i}{h}\right)$$

smoothing

Non-parametric

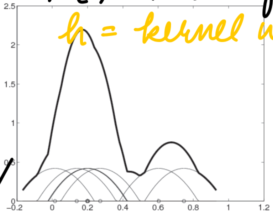
$K, h, \mathcal{D}$ are parameters but have no separate meaning

- no fixed number

$\mathcal{D}$ data $|\mathcal{D}| = n$

$K(\,)$ kernel fction

$h$ = kernel width




$$f_{h,K,\mathcal{D}_1} \neq f_{h,K,\mathcal{D}_2}$$

$f_{h,K,\mathcal{D}}$ ≡ a "python" function

kde(double h, @K, double[] D, double x)

returns density at x

Non-parametric $\rightarrow$ $\infty$ dimensional $\Rightarrow$ # parameters cannot be bounded by any constant

grows with $n$

Parametric $\rightarrow$ finite dimensions

Ex: $(\mu, \sigma^2)$ for Normal $\Rightarrow$ dim $= 2$

$\gamma$ for exponential $\Rightarrow$ dim $= 1$

$\lambda$ for Poisson $\Rightarrow$ — " —

clear meaning

dim independent of $n$

shape constrained (e.g. bell shape)

any shape

Model family

KDE → Model

Given data $\mathcal{D} = \{x^{1:n}\}$

Model family $K$, $h$ → $\{ f_{h,K,\mathcal{D}}$ for any $\mathcal{D} \}$

Estimation/Training "$f_{h,K,\mathcal{D}}$ for given $\mathcal{D}$"

store $\mathcal{D}$, $h$, $(K)$ ↑ estimated model (fixed)

Using estimated model / "Testing"

given new $x$ : calculate $f_{h,K,\mathcal{D}}(x)$

Test set $\mathcal{D}' = \{x^1, \dots x^{n'}\}$ → calculate $f_{h,K,\mathcal{D}}(x^{1:n'})$

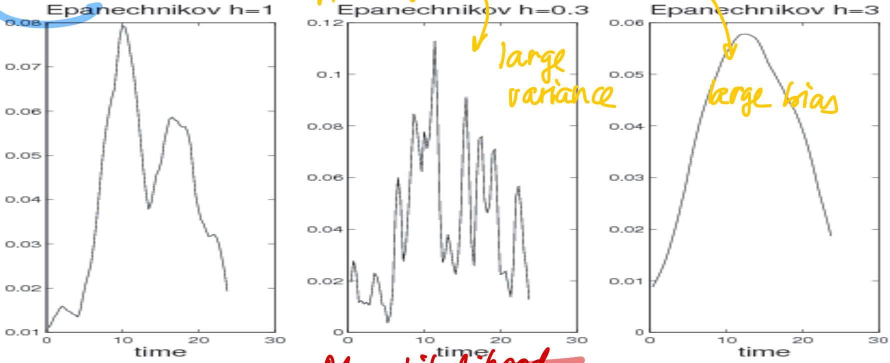
not the same as $\mathcal{D}$

Model → (value) for $x$

Output

## Choosing $h$

effects of h

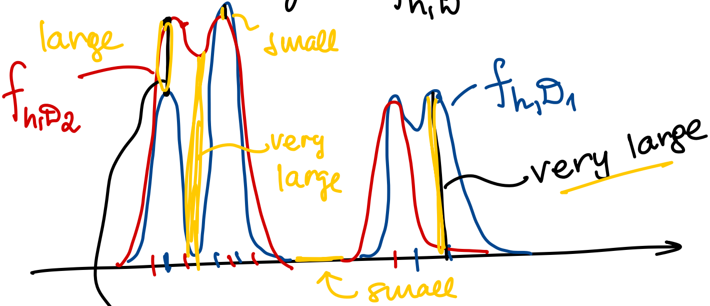



- Principle
  - ~~Max Likelihood~~
  - Bias-variance tradeoff
- How to do it → Cross Validation ⇒ selection of h

# The Bias-Variance trade-off

← KDE $f_{h,\mathcal{D}}$ [K fixed]

↳ change in $f_{h,\mathcal{D}}$ with $\mathcal{D}$




$$|f_{h,\mathcal{D}_1}(x) - f_{h,\mathcal{D}_2}(x)|$$

h small

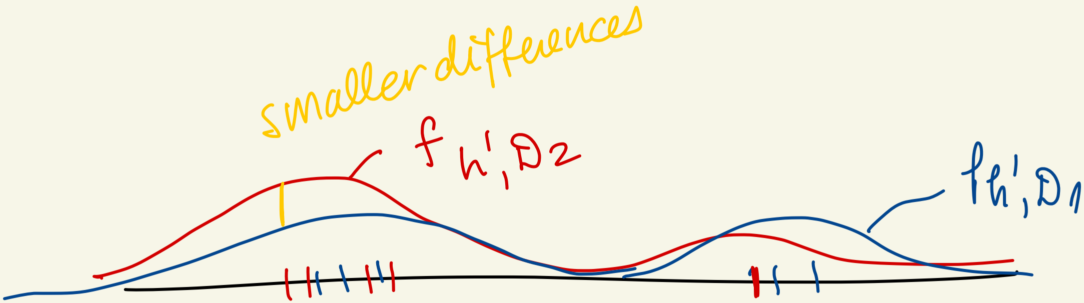
$\mathcal{D}_1, \mathcal{D}_2 \sim P_{unknown}$

$|\mathcal{D}_1| = |\mathcal{D}_2| = n$

Variance = variation w.r.t sample

decreases when $n \uparrow$, $h \uparrow$

- Var $\to 0$ for $n \to \infty$




$h \nearrow \Rightarrow f_{h,D}$ less sensitive to locations of data points

$\Downarrow$

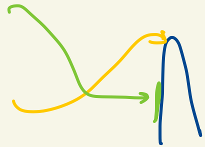
**Bias** $\hookrightarrow$ can't model true $f$ with sharp changes

$\Updownarrow$

can't adapt to all shapes

$|f'(x)|$ large

$|f''(x)|$ large



$=$ inability to model training data

Bias decreases when:

- $h \searrow$
- $n \searrow$

# The Bias-Variance trade-off

→ need to choose $h$ to balance Bias with Variance

Method = Cross validation

Idea evaluate $f_{h,D}$ on new data

$h$: must be chosen

$D$: fixed

$f_{h,D}$ better $\iff$ likelihood $(D' \mid f_{h,D}) \uparrow$

NEW

**CV Algorithm** Given $\mathcal{D}$, $\mathcal{D}'$ ← validation set

↑ training set

**Try** $h_1 < h_2 < h_3 < \cdots < h_M$

"Estimate" $f_{h_1,\mathcal{D}}$ $f_{h_2,\mathcal{D}}$ ...

→ try some models estimate them all then compare

**Evaluate** for $h \in \{h_1, h_2, \ldots h_M\}$

$$\ell^{CV}(h \mid \mathcal{D}') = \sum_{i'=1}^{n'} \ln f_h(x^{i'})$$

← score of $h$

$\mathcal{D}'$

log $\ell$ on new $\mathcal{D}'$

**Select**

$$h^* = \operatorname*{argmax}_{h_{1:M}} \ell^{CV}(h_k \mid \mathcal{D}')$$

← pick max from set pre-selected

# Practical CV - how to split data ?

**Given** $\mathcal{D}_0$, $|\mathcal{D}_0| = n_0$ data $\sim$ iid $P^{\text{unkn}}$

**Wanted**

$\mathcal{D}$ = training data  $\mathcal{D} \cap \mathcal{D}' = \emptyset$ → for estimating $f_{h,\mathcal{D}}$ ← "many params" a shape

$\mathcal{D}'$ = validation data  $\mathcal{D} \cup \mathcal{D}' = \mathcal{D}_0$ → for estimating $\ell^{cv}(f_{h,\mathcal{D}})$ ← single number mean of R.V.

$n + n' = n_0$

↳ has variance too

**How to choose $n$ ?**

- If $n_0$ large ⇒ $n' \approx 1000$ ← sufficient

  $n = n_0 - n'$

- Otherwise ⇒ do **k-fold CV**  {want most data for training}

1. $K = 5, \dots 10$, split $\mathcal{D}_0$ into $\mathcal{D}^{(1)}, \dots \mathcal{D}^{(K)}$, with $|\mathcal{D}^{(k)}| = \frac{n_0}{K} < 1000$

2. for $k = 1 : K$

   $\mathcal{D}' \leftarrow \mathcal{D}^{(k)}$

   $\mathcal{D} \leftarrow \mathcal{D}_0 \setminus \mathcal{D}^{(k)}$

   $\ell^{cv,k}(f_{h,\mathcal{D}}; \mathcal{D}')$ ← for all $h$'s

3. $\ell^{cv}(f_{h,\cancel{\mathcal{D}}}) = \frac{1}{K} \sum_{k=1}^{K} \ell^{cv,k}(f_{h,\mathcal{D}}; \mathcal{D}')$ ← averaging

   $\mathcal{D}_0 \setminus \mathcal{D}^{(k)}$  $\mathcal{D}^{(k)}$

4. **Select** $h^* = \arg\max_h \ell^{cv}(f_h)$

# Model Selection in Statistics

Given $\mathcal{D}$

Model ~~family~~ families

(try them all)

| | estimation | best in family |
|---|---|---|
| $\mathcal{F}_1$ (gaussian) | $\rightarrow$ | $f_1 \in \mathcal{F}_1$ |
| $\mathcal{F}_2$ (logistic) | $\rightarrow$ | $f_2$ |
| $\mathcal{F}_3$ (KDE $h_1$) | $\rightarrow$ | |
| $\mathcal{F}_4$ (— " — $h_2$) | $\rightarrow$ | $f_4$ |
| . . . | . . . | |
| | $\rightarrow$ | $f_M$ |

Select "best" $f_1, f_2, \cdots f_M$

$$f^* = \underset{k=1:M}{\operatorname{argmax}} \; \underline{\text{score}(f_k)}$$

CV

AIC, BIC (coming next)