Statistical Estimation: Data & Non-data Information

Roger J-B Wets
University of California, Davis
& M.Casey @ Raytheon
G.Pflug @ U. Vienna,
X. Dong @ EpiRisk,
G-M You @ EpiRisk.

a little background

- Decision making under uncertainty:

$$\min f_{01}(x) + E\{Q(\boldsymbol{\xi}, x)\} \text{ so that } x \in C_1 \subset \mathbb{R}^n$$

$$Q(\xi, x) = \inf\left[f_{02}(\xi, y) \middle| y \in C_2(\xi, x) \right]$$

$$C_2(\xi, x) = \left\{ y \in \mathbb{R}^d_+ \middle| W_\xi y = d_\xi - T_\xi x \right\}$$

- = <u>stochastic programming problems</u>.
- Issue: realizations of $(d_\xi, T_\xi, W_\xi) = \boldsymbol{\xi} \in \mathbb{R}^N, N \gg 2!$

  data: 43 samples,… never reaches the

  asymptotic range

Formulation

- Find F^{est}, an estimate of the distribution of a random (phenomena) variable X given **all** the information available about this random phenomena,
- i.e. such that

$$\forall x : F^{est}(x) \simeq F^{true}(x) = \text{prob.}[X \leq x].$$

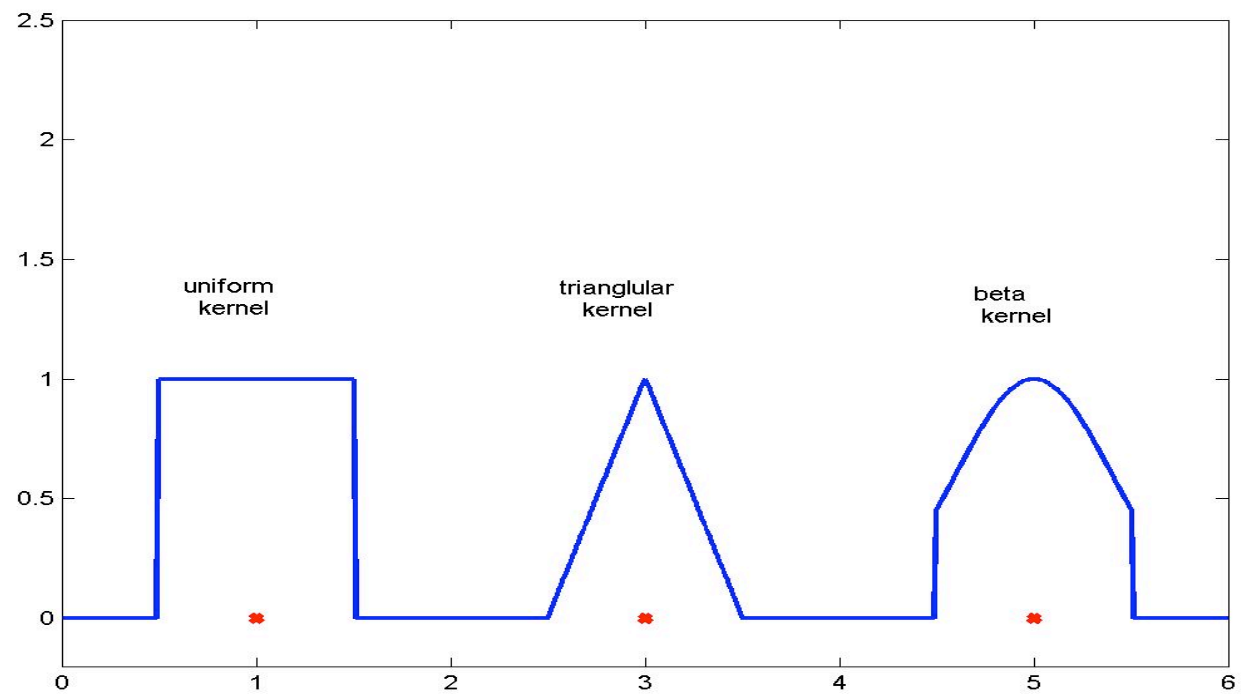
Information

- Observations (data): $x_1, x_2, \ldots, x_\nu$
- Non-data facts:
  - density or discrete distribution,
  - bounds on expectation, moments,
  - shape: unimodal, decreasing, parametric class
- Non-data modeling assumptions:
  - see above + …
  - density is smooth, (un)bounded support, ..

Applications:

- Estimating (cum.) distribution functions
- Estimating coefficient of time series
- Estimating coefficients of stochastic differential equation (SDE)
- Estimating financial curves (zero-curves)
- Dealing with lack of data: ***few*** *observations*
- Estimating density functions: h^{est}

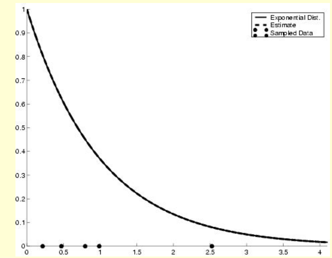
Kernel-choice

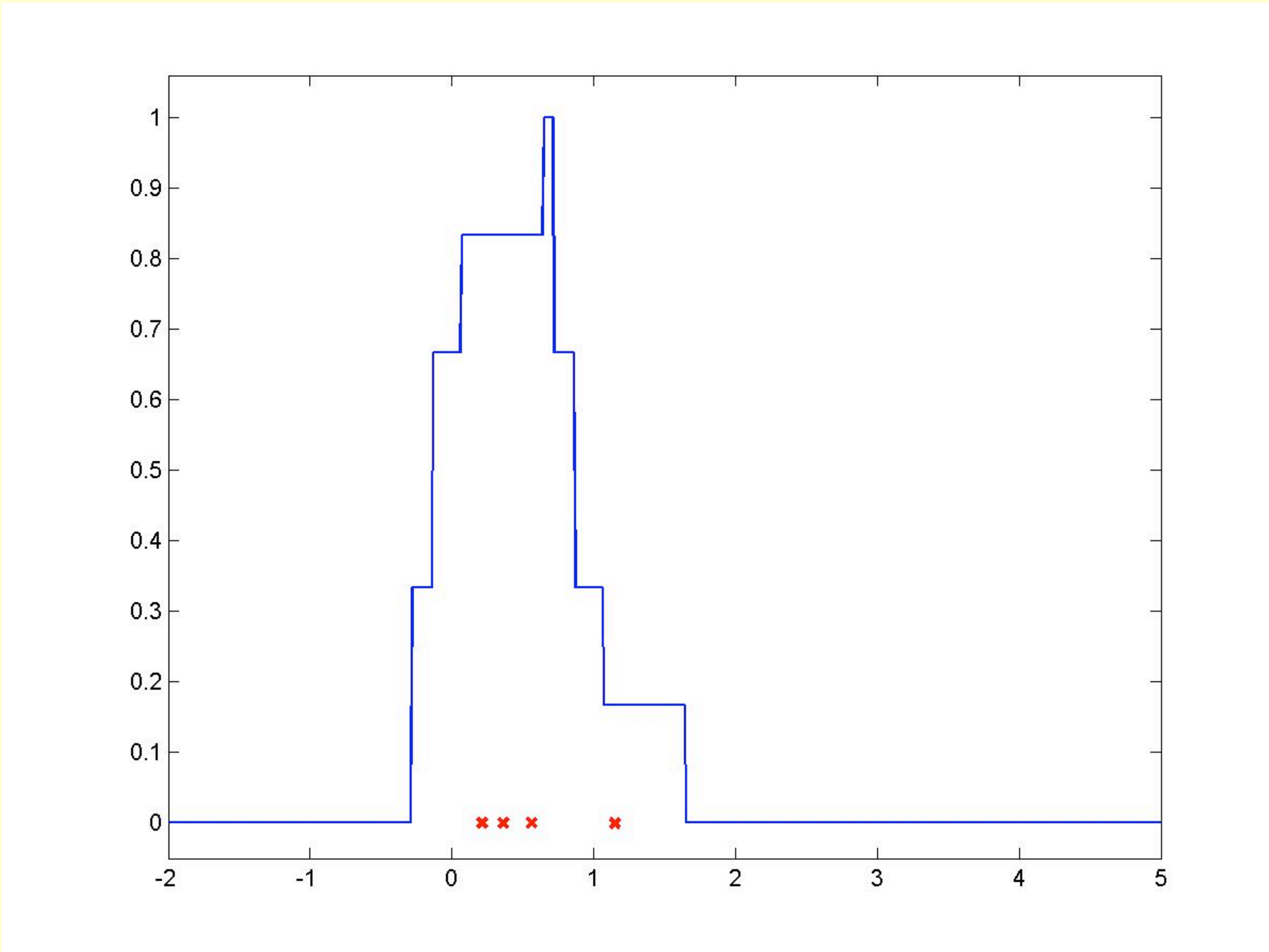


Information = Observations: $x_1, x_2, \ldots, x_\nu$

Optimal bandwidth = kernel support ?

Kernel estimates





Statistical Estimation

from an optimization viewpoint

Criterion: Maximum Likelihood

- Find $h \in H$=*class-fcns*($\mathbb{R}$) that maximizes the probability of observing $x_1, x_2, \ldots, x_\nu$

$$\boxed{\max \prod_{l=1}^{\nu} h(x_l)}$$

- equivalently:

$$\max \frac{1}{\nu} \sum_{l=1}^{\nu} \ln h(x_l) =$$

$$\max E^{\nu}\left\{\ln h(x)\right\} = \max \int \ln h(x) P^{\nu}(dx)$$

leads to:

∞–dim. optimization problem

$$\max E^{\nu}\left\{\ln h(x)\right\} = \frac{1}{\nu}\sum_{l=1}^{\nu} \ln h(x_l)$$

$$\text{so that } \int h(x)\,dx = 1,$$

$$h(x) \geq 0,\ \forall x \in \mathbb{R}$$

$$h \in A^{\nu} \subset H$$

A^{ν}: non-data information constraints

$H = C^2(\mathbb{R}), L^p(S), H^1(S), \ \dots \quad S \subset \mathbb{R}$

Non-data information constraints

support: $S=[\alpha,\beta],\ S=[\alpha,\infty),\ldots$

bounds on moments: $\mathrm{a}\le\int xh(x)dx\le b$

unimodal: $h(x)=e^{-Q(x)},Q:$ convex

decreasing: $h'(x)\le 0,\forall x\in S$

'smoothness': $h\in H_0^1,\quad \int\frac{h'(x)^2}{h(x)}dx\le\kappa$ (Fisher info.)

\`Bayesian' : $\left|h-h_a\right|\le\beta$ or $\theta\left(\left|h-h_a\right|\right)$, Total variation, …

Questions (M^{+}-estimators)

- Consistency: as $\nu \to \infty$ does $h^{est} \to h^{true}$?

  hypo-convergence: “a law of large numbers”

- Convergence rate: how large does ν have to be so that $h^{est} \sim h^{true}$?

  quantitative hypo-convergence: dist(h^{est}, h^{true})

- Solving the ∞-dimensional optimization problem?

  finite dimensional approximations
  + again convergence issues.

When is $(\mathcal{P})$ "near" $(\mathcal{P}^a)$

$$(\mathcal{P}) \quad \max f_0(x) \quad \text{s.t. } f_i(x) \le 0,\ i \in I_f, \quad x \in X_f \subset E$$

$$(\mathcal{P}^a) \quad \max g_0(x) \quad \text{s.t. } g_i(x) \le 0,\ i \in I_g, \quad x \in X_g \subset E$$

Example:

$$(\mathcal{P}) \quad \max f_0(x) \quad \text{s.t. } f_i(x) \le 0,\ i = 1, \ldots, m$$

$$(\mathcal{P}^a) \quad \max f_0(x) + \nu \sum_{i=1}^{m} \max[0, f_i(x)]^2$$

From: when is $(\mathcal{P})$ “near” $(\mathcal{P}^a)$

$(\mathcal{P})$ max $f(x)$ where

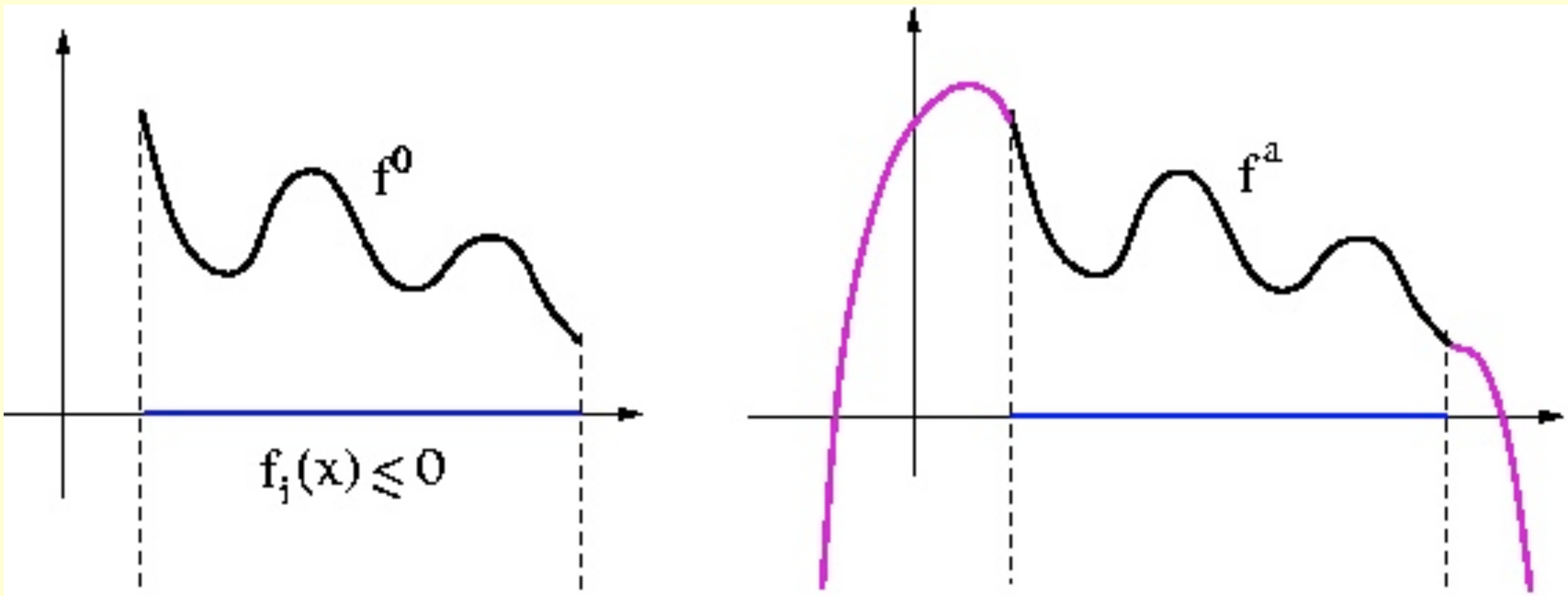
$$f(x) = f_0(x) \text{ when } f_i(x) \le 0,\ i \in I_f,\ x \in X_f \subset E$$
$$-\infty \text{ otherwise}$$

$(\mathcal{P}^a)$ max $g(x)$ where

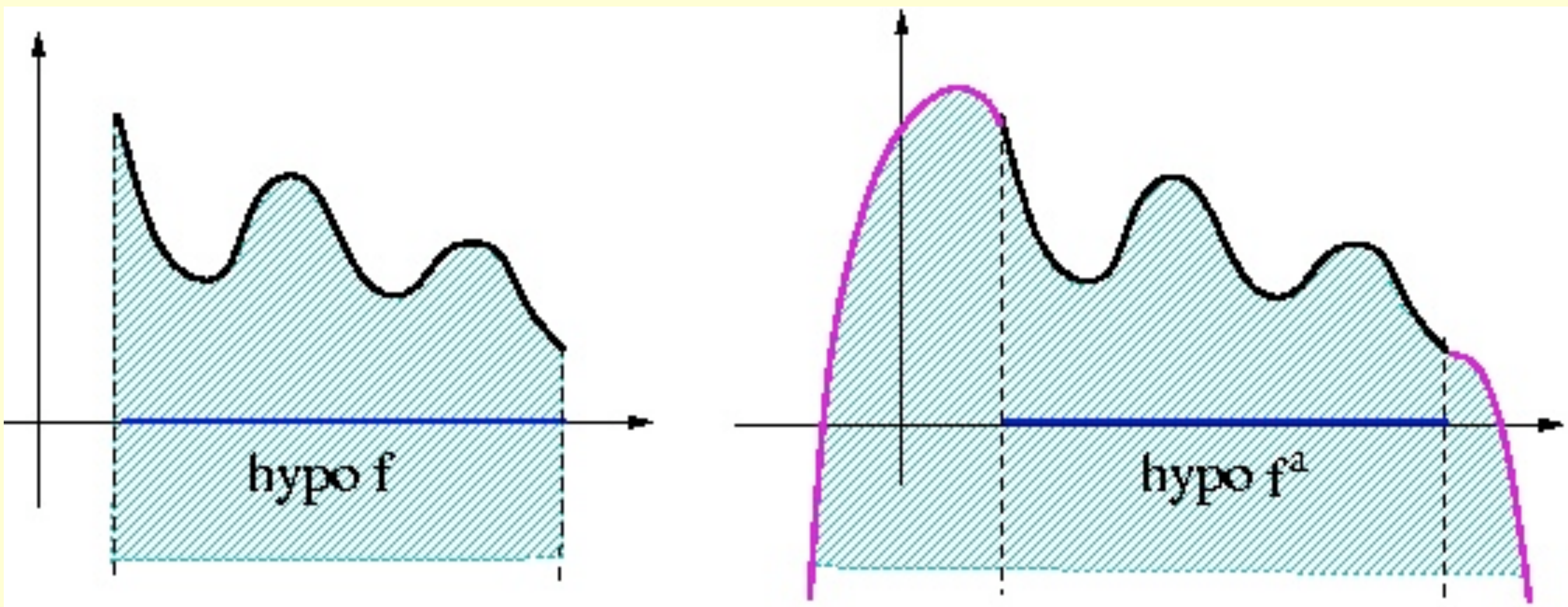
$$g(x) = \mathrm{g}_0(x) \text{ when } \mathrm{g}_i(x) \le 0,\ i \in I_g,\ x \in X_g \subset E$$
$$-\infty \text{ otherwise}$$

To: when is f “near” g

Example (with penalty)



Hypographs of f & f^a



hypo f^a “near” hypo f

➔ argmax ($\mathcal{P}^a$) ~ argmax ($\mathcal{P}$)

$$f = h\text{-}\lim_\nu f^\nu, \quad f^\nu \xrightarrow[h]{} f$$

- $$x^\nu \in \arg\max f^\nu, x^\nu \underset{cluster}{\longrightarrow} x \Rightarrow x \in \arg\max f$$

aw-topology for usc functions (Attouch-Wets, 1992)

- $\exists\, dl$ such that $dl(f^\nu, f) \to 0 \Leftrightarrow f^\nu \underset{h}{\to} f$

$$dl(f^\nu, f) \le \eta \Rightarrow \ \mathrm{dist}(\arg\max f^\nu, \arg\max f) \le \kappa \psi^{-1}(\eta)$$

- $$F^\nu(x) = \int f(\xi, x) P^\nu(d\xi) \ \& \ F(x) = \int f(\xi, x) P(d\xi)$$

$$P^\nu \underset{n}{\to} P \ \ \& f\text{-tight} \ \Rightarrow F^\nu \underset{h}{\to} F$$

Quantitative hypo-convergence

- $\psi : \mathbb{R}_+ \to \left[0,1\right]$ continuous, strictly increasing
  - for example: $\psi(a) = (2/\pi)\arctan a$
- $dl_\rho(C,D) = \sup_{|x|\le\rho} |\psi\left(d(x,C\right) - \psi\left(d(x,D)\right)|$
- $dl\left(C,D\right) = \int_0^\infty e^{-\rho} dl_\rho\left(C,D\right) d\rho$

  metric on space of closed sets: τ_{aw}-topology
- $dl(f,g) = dl\left(\text{hypo}\, f, \text{hypo}\, g\right)$

  metric on space of usc-fcns: τ_{aw}-topology

Quantitative hypo-convergence II

- $dl(f^\nu, f) \to 0 \Leftrightarrow f^\nu \underset{aw\text{-hypo}}{\longrightarrow} f$
- Theorem:

$f, g : X \to \overline{\mathbb{R}}$, usc, concave,

$\rho_0 : \arg\max f \cap \rho_0 B \neq \emptyset,\ \sup f < \rho_0$ & for g

then $\forall \rho > \rho_0,\ \varepsilon > 0:$

$$dl_\rho\left(\varepsilon\text{-}\arg\max f, \varepsilon\text{-}\arg\max g\right) \leq (1 + \frac{4\rho}{\varepsilon})\, dl_\rho\left(f, g\right)$$

“Opt”-Formulation

$$\max E^{\nu}\left\{\ln h(x)\right\} = \frac{1}{\nu}\sum_{l=1}^{\nu} \ln h(x_l)$$

$$\text{so that } \int h(x)\,dx = 1,$$

$$h(x) \geq 0,\ \forall x \in \mathbb{R}$$

$$h \in A^{\nu} \subset H$$

A^{ν}: non-data information constraints

$H = C^2(\mathbb{R}),\ L^p(S),\ H^1(S),\ \ldots \quad S \subset \mathbb{R}$

Convergence "rate"

$$L^\nu(h,x) = \ln h(x) \text{ if } h \in A^\nu \subset H, \int h(x)\,dx = 1,\ h \geq 0$$

$$= -\infty \text{ otherwise}$$

$L = L^\nu$ except for $A = aw\text{-}\lim A^\nu \quad (aw = dl)$

then $\int L^\nu(\bullet, x)\,dP^\nu = \boxed{E^\nu L^\nu \underset{aw\text{-hypo}}{\longrightarrow} EL} = \int L(\bullet, x)\,dP^{true}$

when H Hilbert, $H \xrightarrow[embedding]{cont.compact} \breve{H}$, $\breve{H}$ RKHS

constraint set: H-bounded, $\breve{H}$-closed

$\breve{H}$ lin. subspace, $\forall x : x \mapsto h(x)$ cont.,

relies on the Hilbert-Schmidt Embedding Theorem

Numerical Procedures

1. $h \simeq \sum_{k=1}^{q} u_k \varphi_k(\bullet)$

   Fourier coefficients, wavelets, "kernels-basis"

2. $h = \exp\left(s(\bullet)\right)$

   $s(\bullet)$ constrained cubic (or quadratic) spline

Basis: Fourier coefficients

$$\max \ \frac{1}{v}\sum_{l=1}^{v} \ln\left[\frac{1}{\sqrt{\theta}}u_0 + \sqrt{\frac{2}{\theta}}\sum_{k=1}^{q}\cos(\frac{w_k \pi x_l}{\theta})u_k\right]$$

$$\text{so that } \ u_0 + \sqrt{2}\sum_{k=1}^{q}\frac{\sin(w_k\pi)}{w_k\pi}u_k = \sqrt{\theta},$$

$$u_0 + \sqrt{2}\sum_{k=1}^{q}\cos(\frac{w_k\pi x}{\theta}) \geq 0, \ \forall x \in [0,\theta],$$

$$u_k \in \mathbb{R}, k = 0,\ldots,q, \quad w_k \in \mathbb{R}_+, k = 1,\ldots,q$$

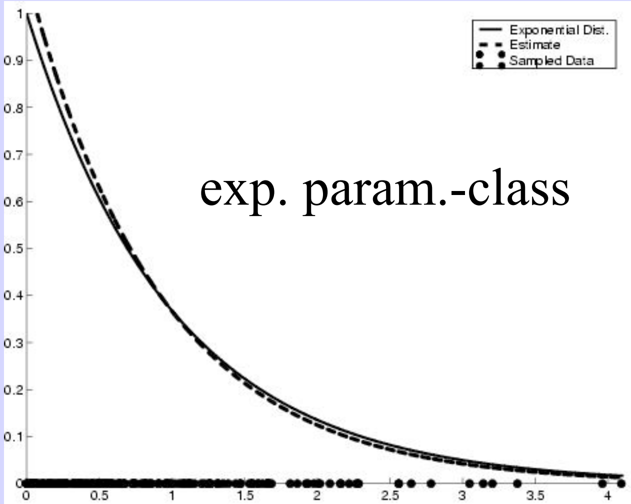
$$\sum_{k=1}^{q}\left[\cos\left(\frac{w_k\pi x}{\theta}\right) - \cos\left(\frac{w_k\pi x'}{\theta}\right)\right]u_k \geq 0, \forall 0 \leq x \leq x' \leq \theta$$

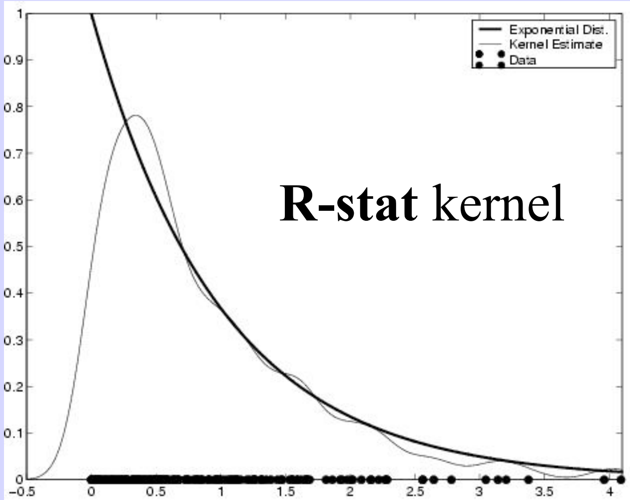
i.e., h is decreasing

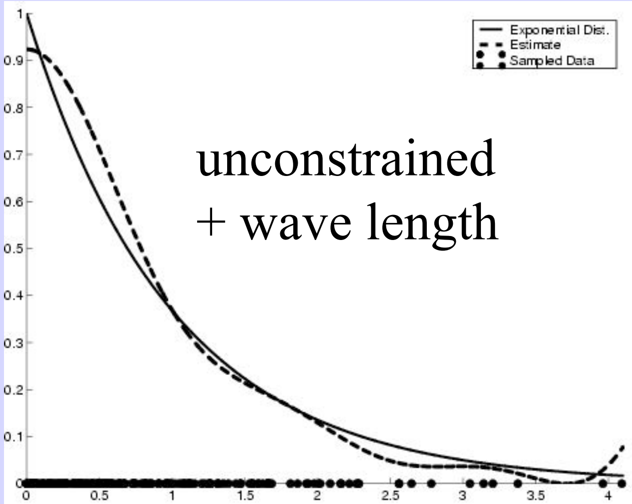
Test Case: exponential

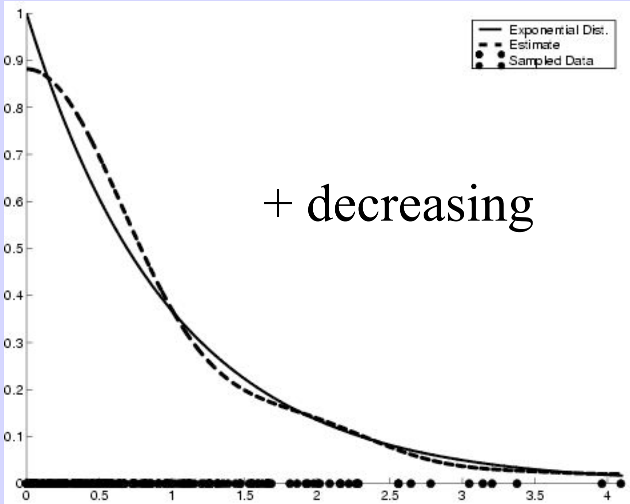
- $h^{true}(x) = \lambda e^{-\lambda x}$ if $x \geq 0; \quad = 0$ if $x < 0 \qquad (\lambda = 1)$
- "empirical" estimate
- kernel estimate from **R-stat**
- unconstrained with support $[0,\infty)$
- unconstrained with adaptive wave length
- constrained (*h* decreasing)
- parametric, i.e., $h \in$ exp-*class*

200-observations!

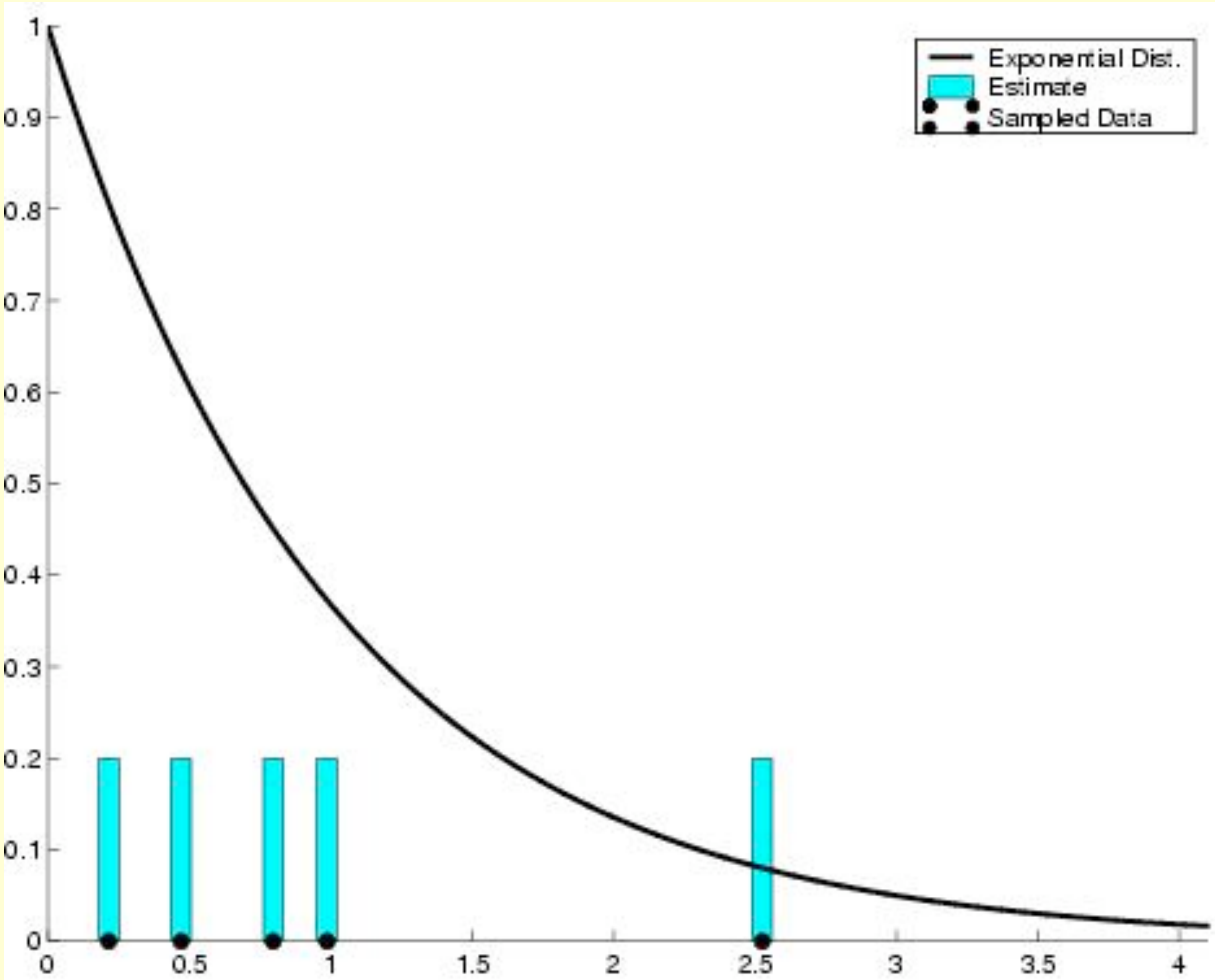




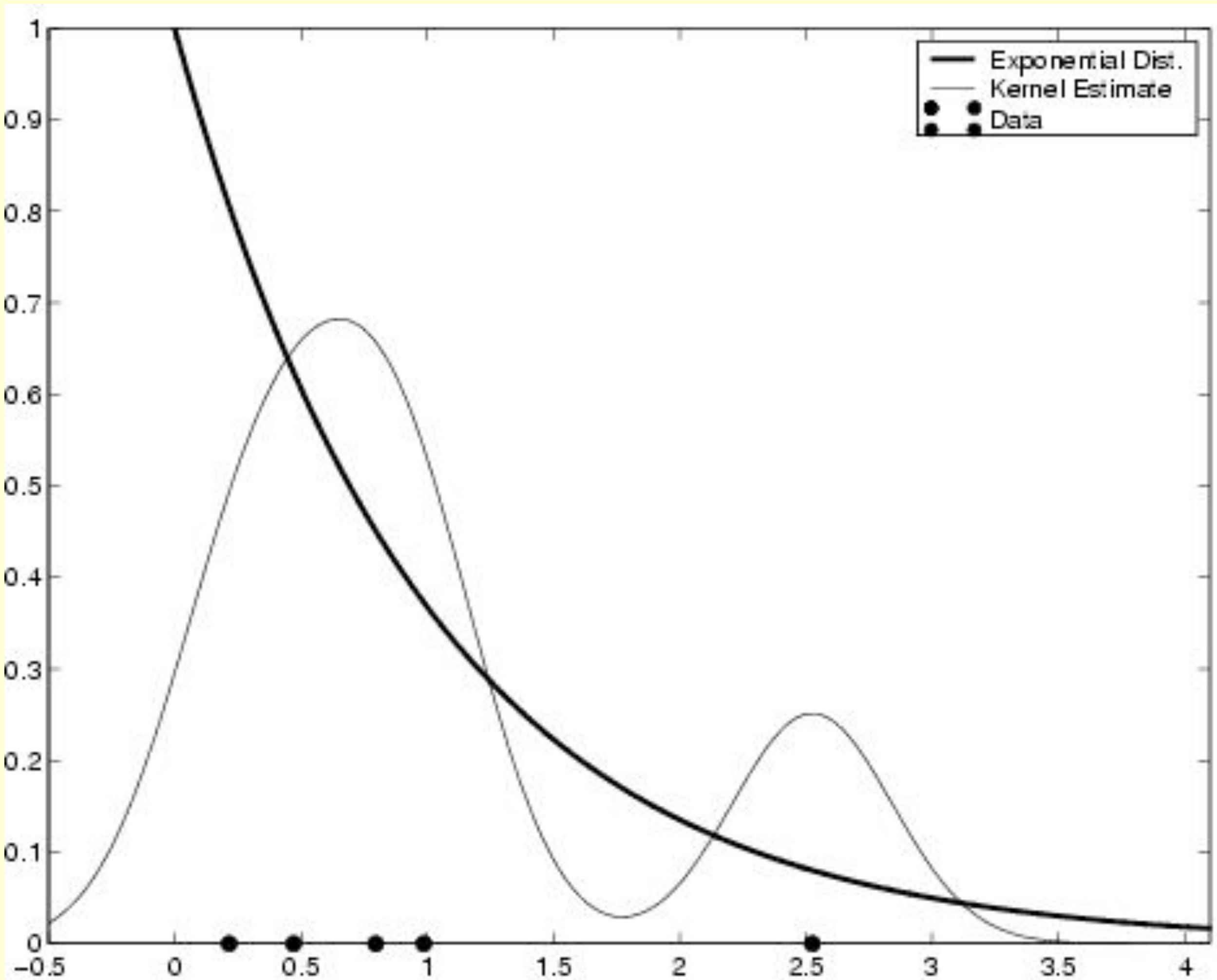




h^{est} : empirical (five observations)

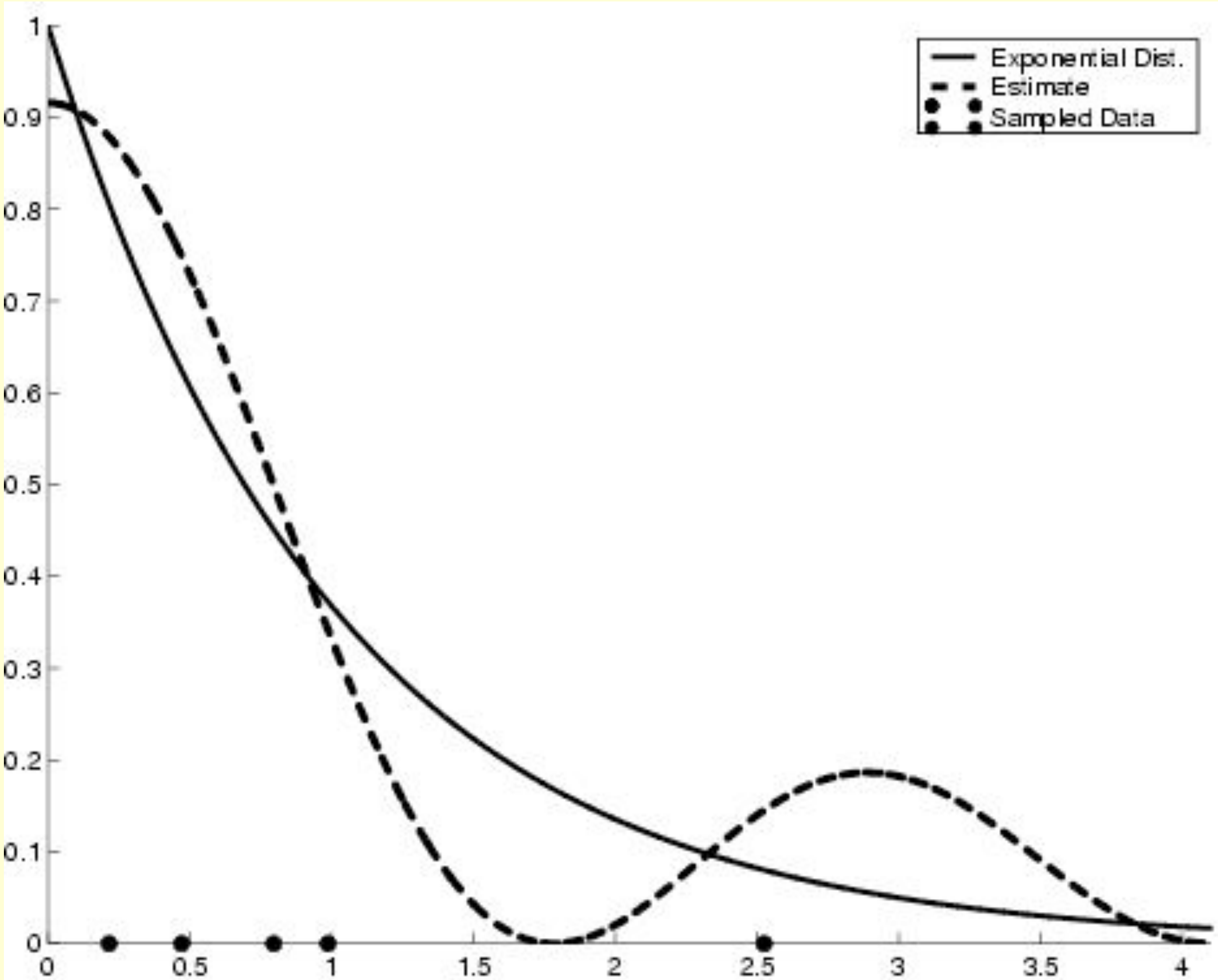


h^{est} : kernel

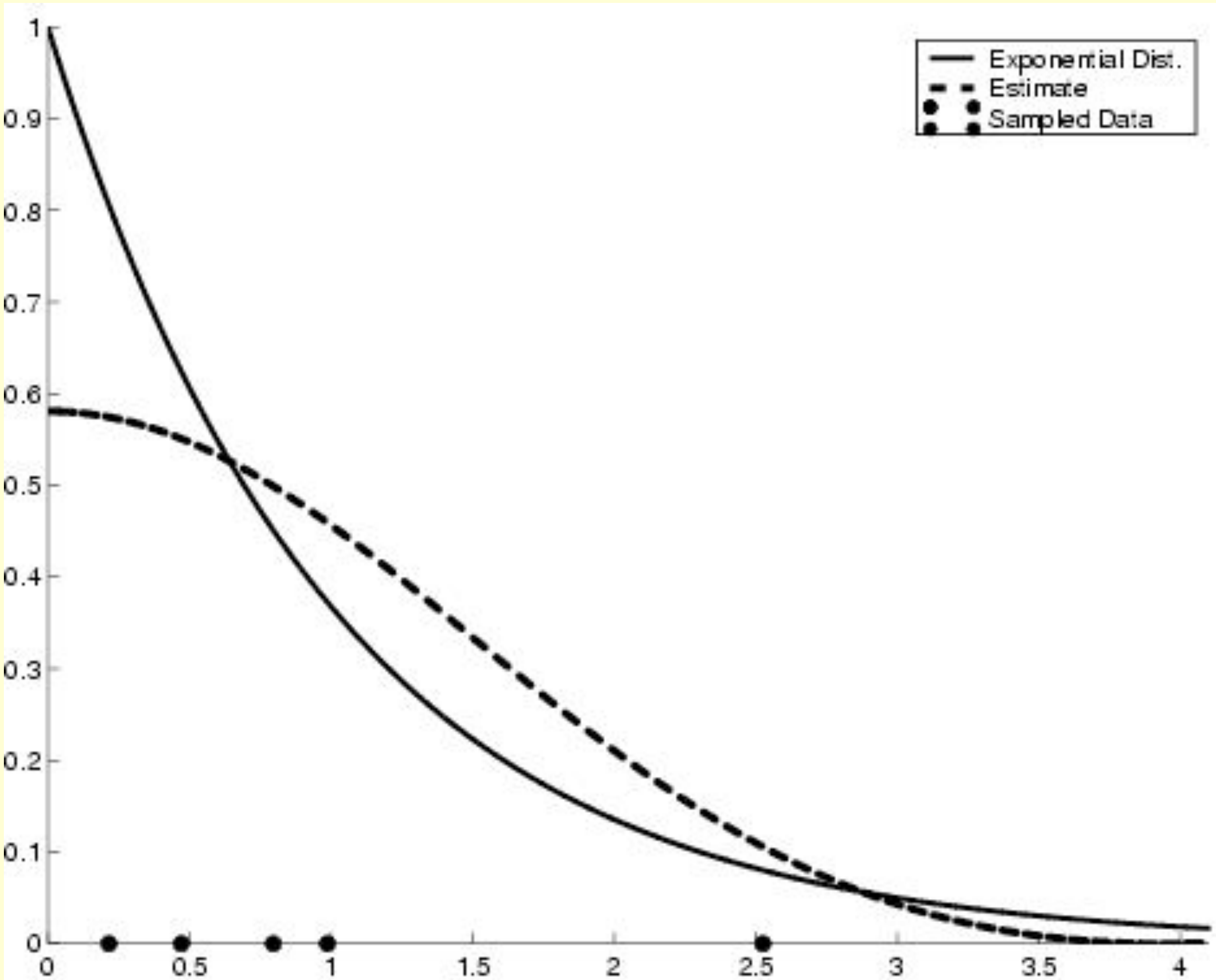


Exponential Dist.
Kernel Estimate
Data
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−0.5
0
0.5
1
1.5
2
2.5
3
3.5
4

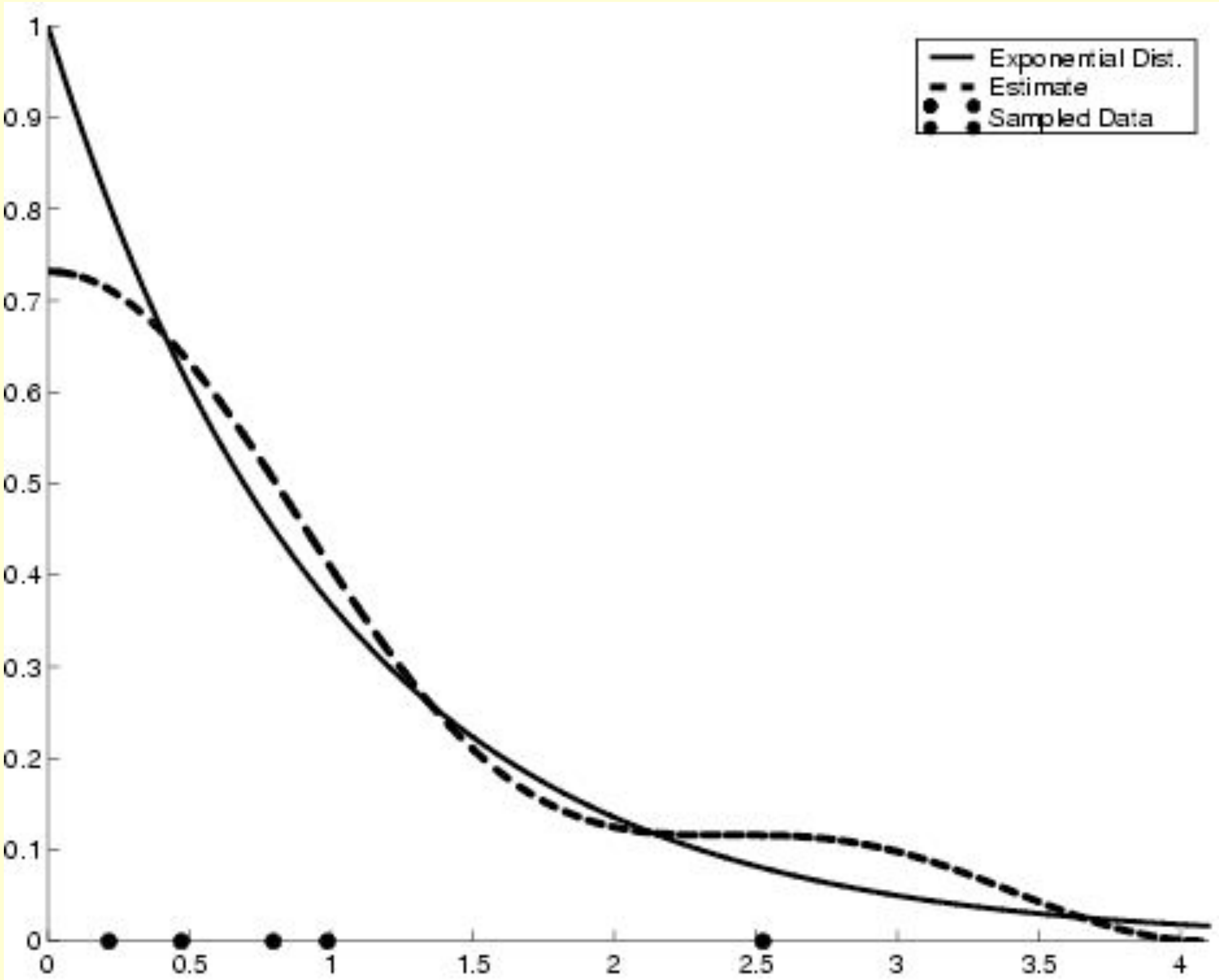
h^{est} : unconstrained



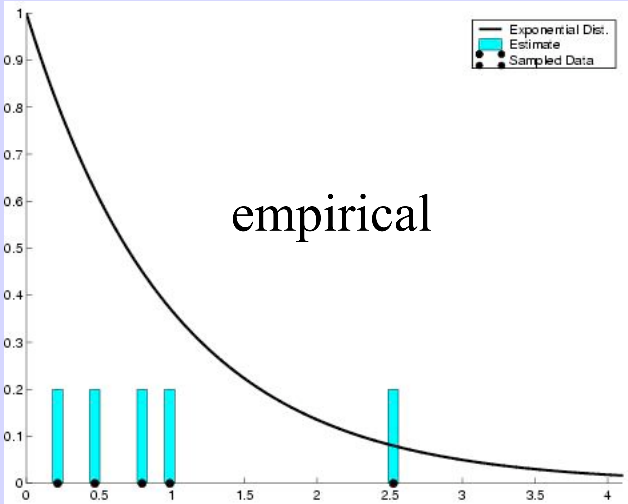
h^{est} : ..with adaptive wave

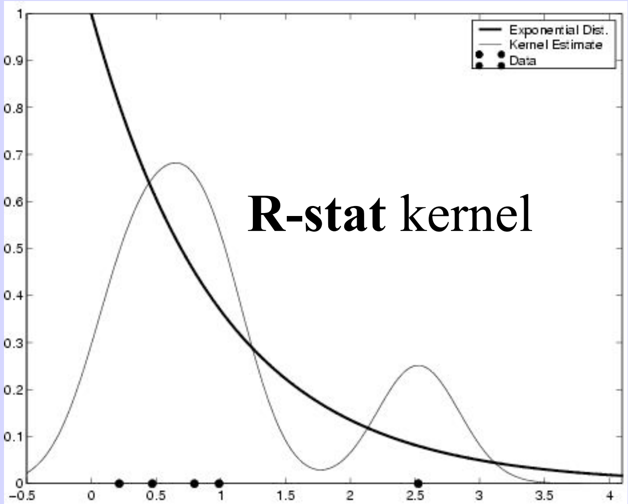


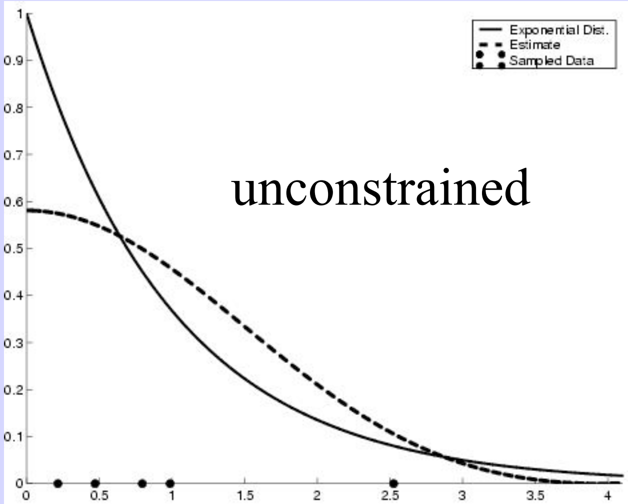
h^{est} : with decreasing constraint

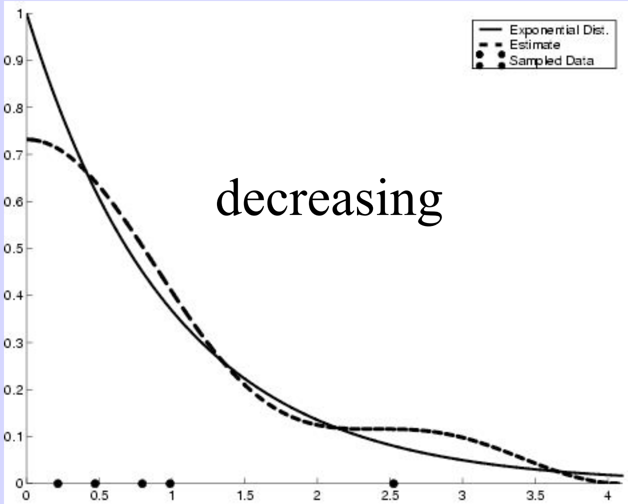


5-observations!

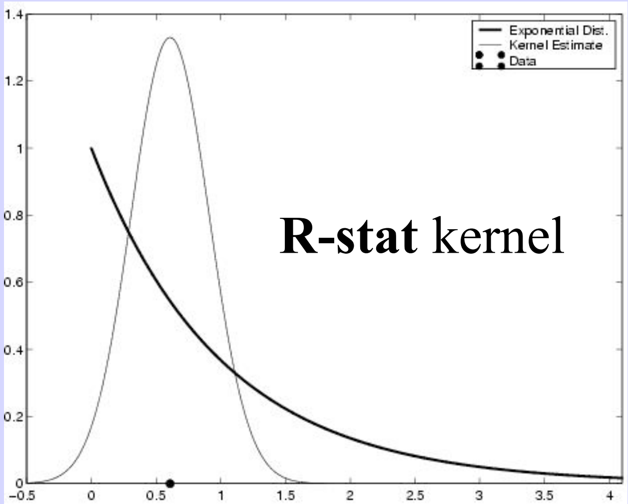


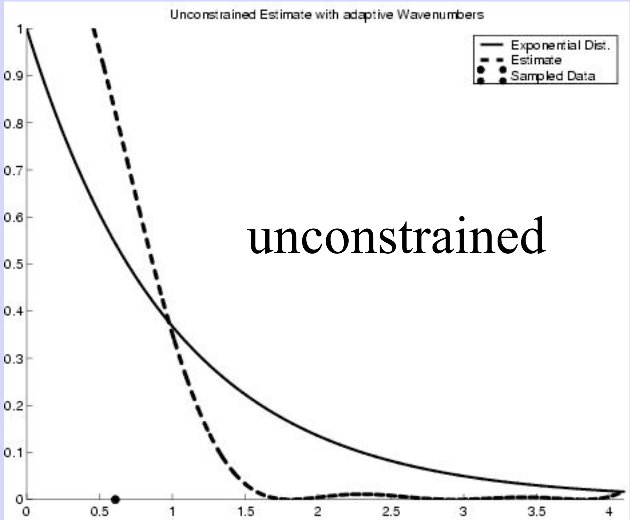


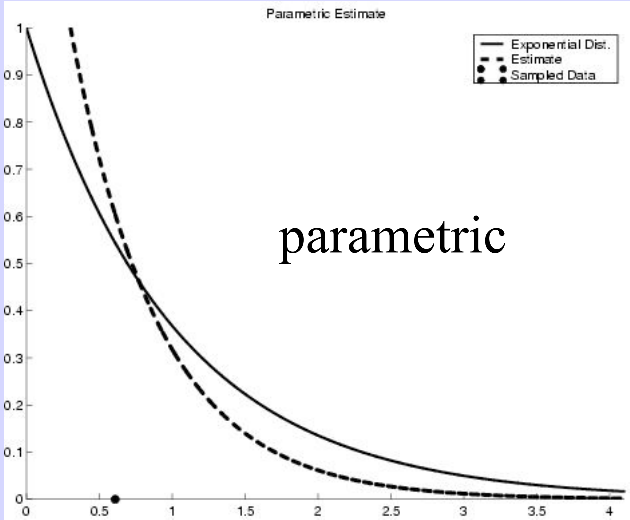


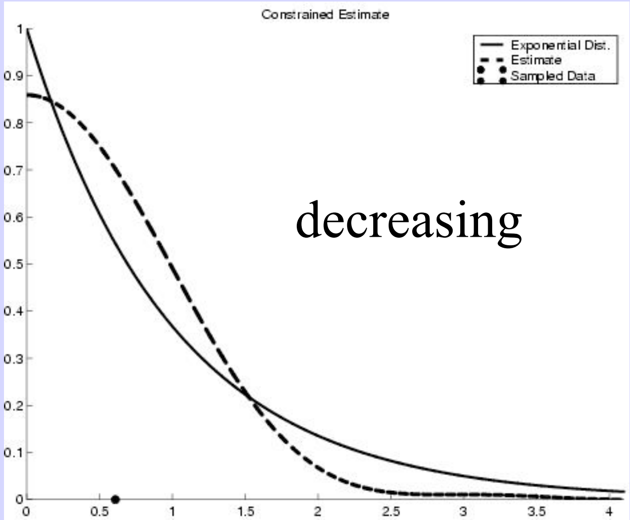


1-observation!









Exponential (quadratic) spline

$$h(x) = e^{-s(x)}$$

$$s(x) = s_0 + v_0 x + \int_0^x dr \int_0^r dt\, z(t),\ \ z(t) \equiv z_k \ \text{ on } \ \left(x_k, x_{k+1}\right]$$

$$= s_0 + v_0 x + \sum\nolimits_{j=1}^{k} a_{kj} z_k \ \text{ when } x \in \left(x_k, x_{k+1}\right]$$

$$\max E^{\nu} \left\{ \ln h(x) \right\} \sim \min \frac{1}{\nu} \sum_{l=1}^{\nu} s(x_l)$$

$$\text{so that } \int e^{-s(x)}\, dx \le 1, \qquad (h \ge 0)$$

$$z_k \in \left[-\kappa_l, \kappa_u\right] \text{ 'constrained'-spline}$$

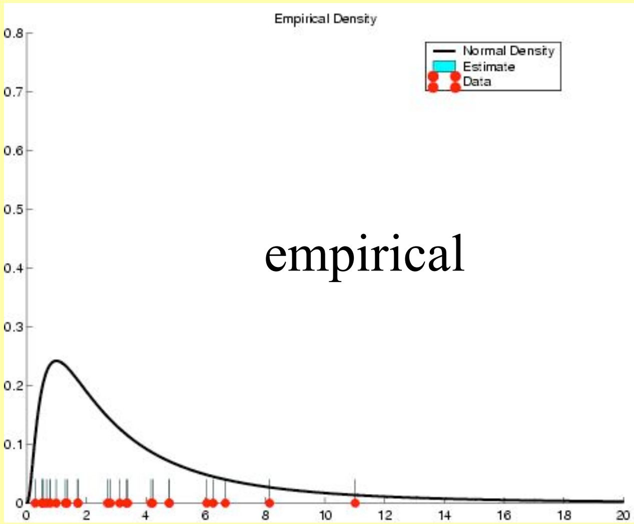
$$\text{unimodal: } \kappa_l = 0$$

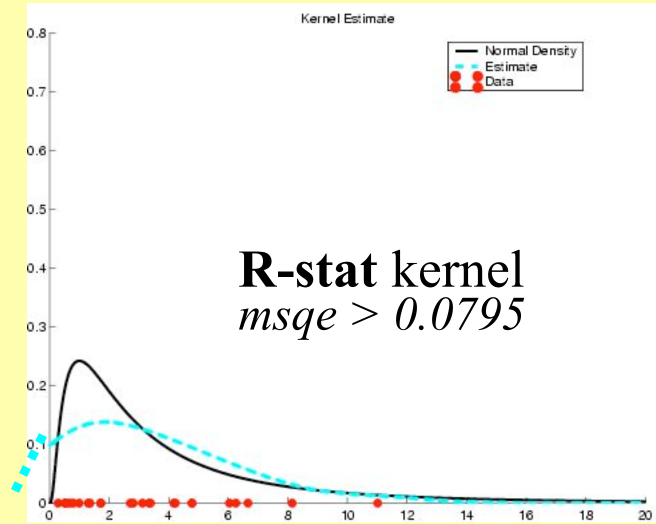
Test Case: log-normal, $\nu = 25$

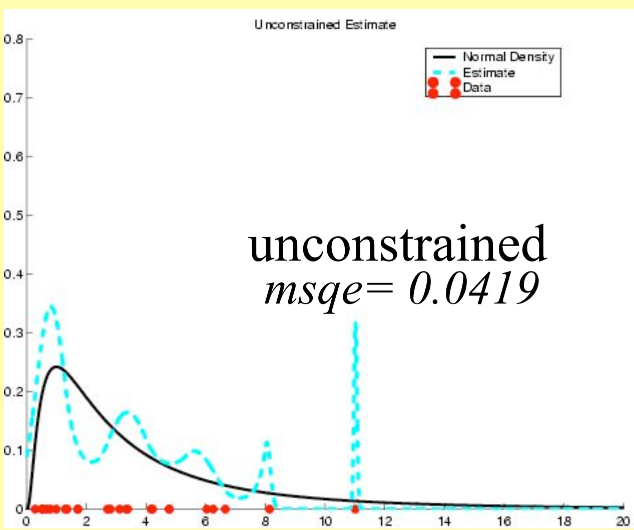
- $h^{true}(x) = \left(x\tau\sqrt{2\pi}\right)^{-1} e^{-(\ln x - \theta)^2/2\tau^2}$ for $x \geq 0$.
- "empirical" estimate
- kernel estimate from **R-stat**
- unconstrained with support $[0,\infty)$
- constrained (*h* unimodal)

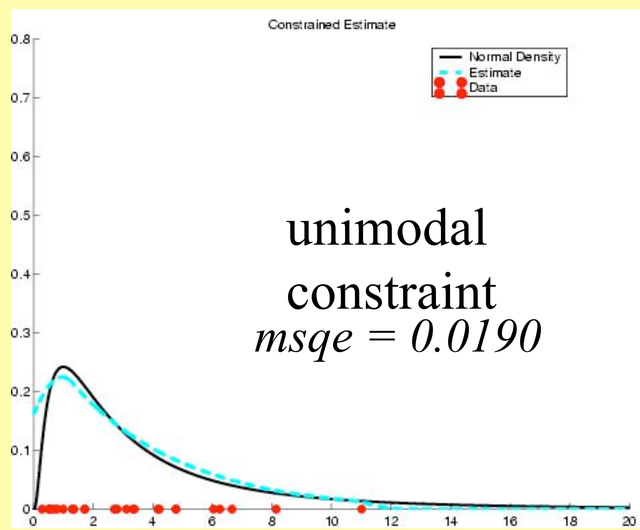
25 observations!

msqe = mean square error





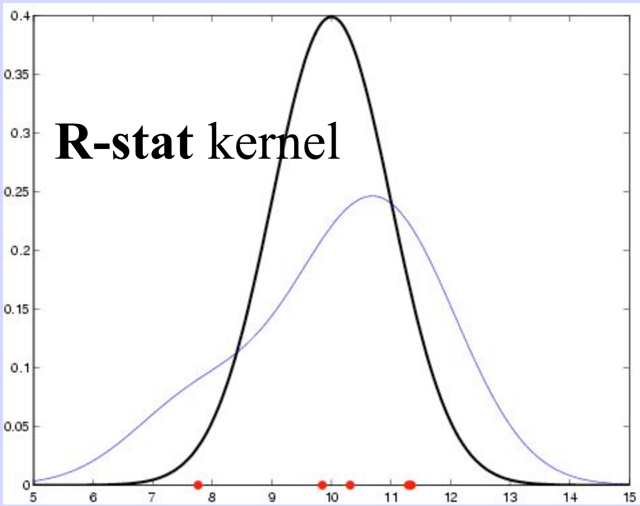


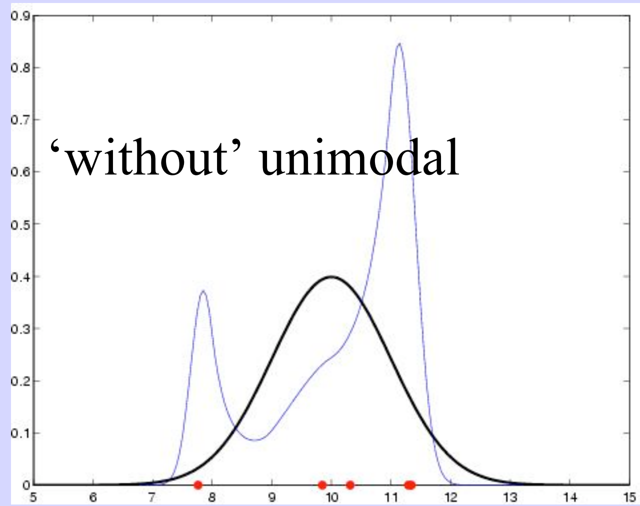


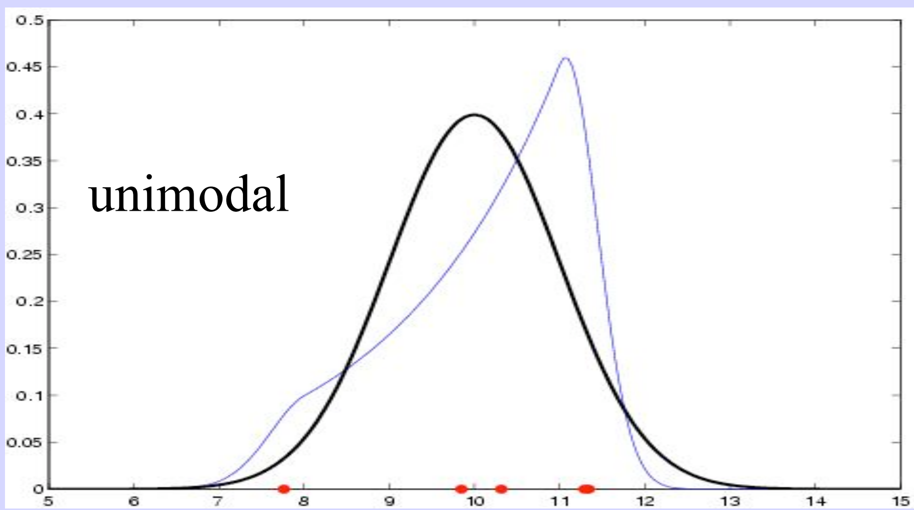
Test Case: $\mathcal{N}(10,1)$, $\nu = 5$

- $h^{true}(x) = (1/\sqrt{2\pi})\, e^{-(x-10)^2/2},\ x \in \mathbb{R}$
- kernel estimate from **R-stat**
- unconstrained with support $(-\infty,\infty)$
- light-unimodality constraint on *h*
- stricter-unimodality constraint on *h*
- ….
- *h*: unimodal

5-observations: $\mathcal{N}(10,1)$







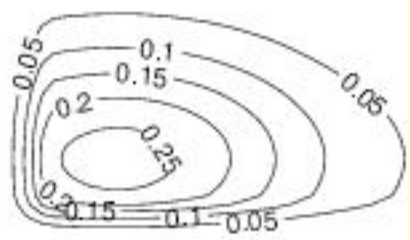
Higher Dimensions

- 2-dim. (& dim. >2) $h(x,y) = \exp(z(x,y))$

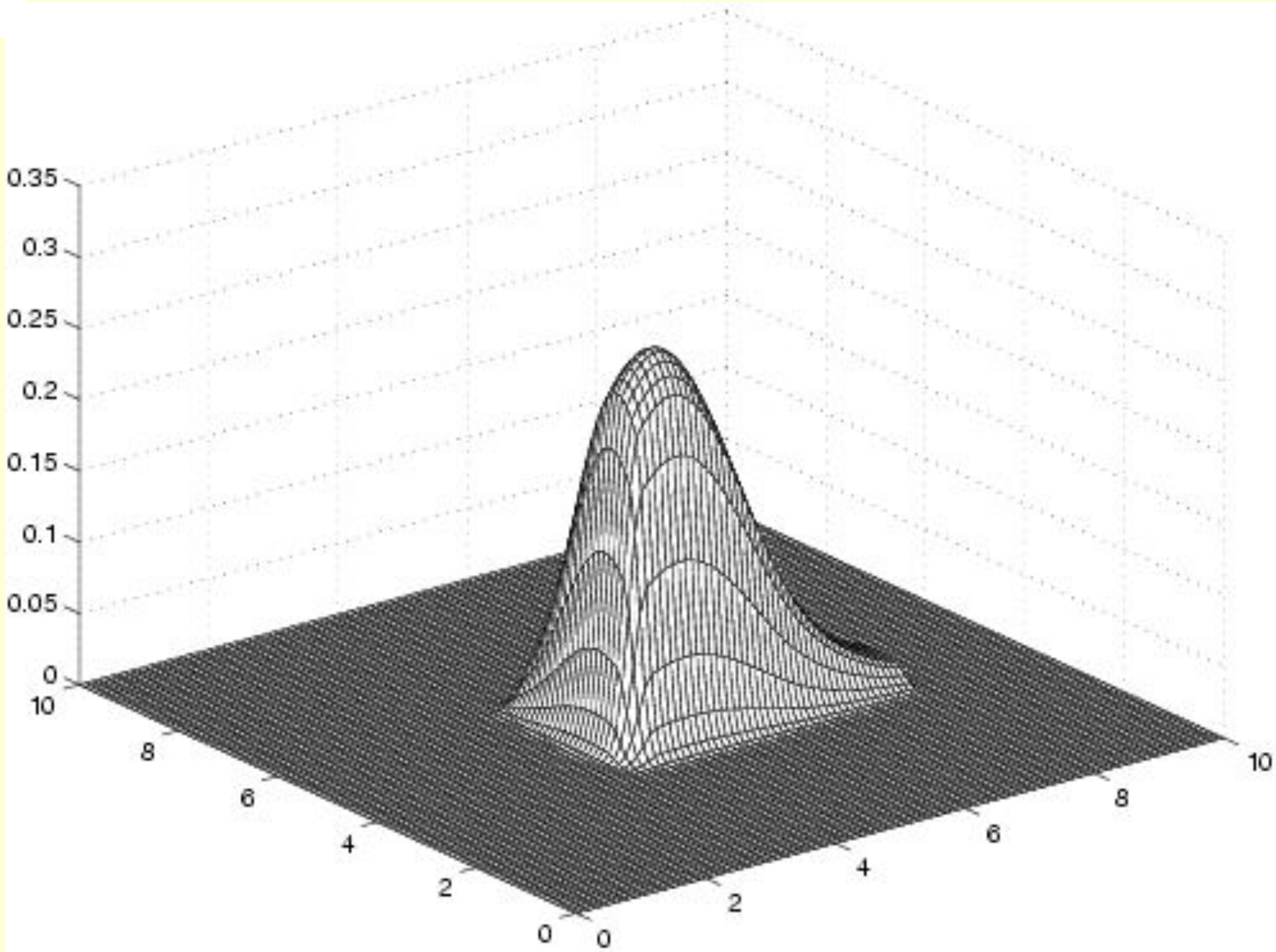
$$z(x,y) = z_0 + v_0^x x + \int_0^x d\tau \int_0^\tau ds\, a^x(s)$$
$$+ \int_0^y d\tau \int_0^\tau dr\, a^y(r)$$
$$+ \int_0^x ds \int_0^y dr\, a^{x,y}(s,r).$$

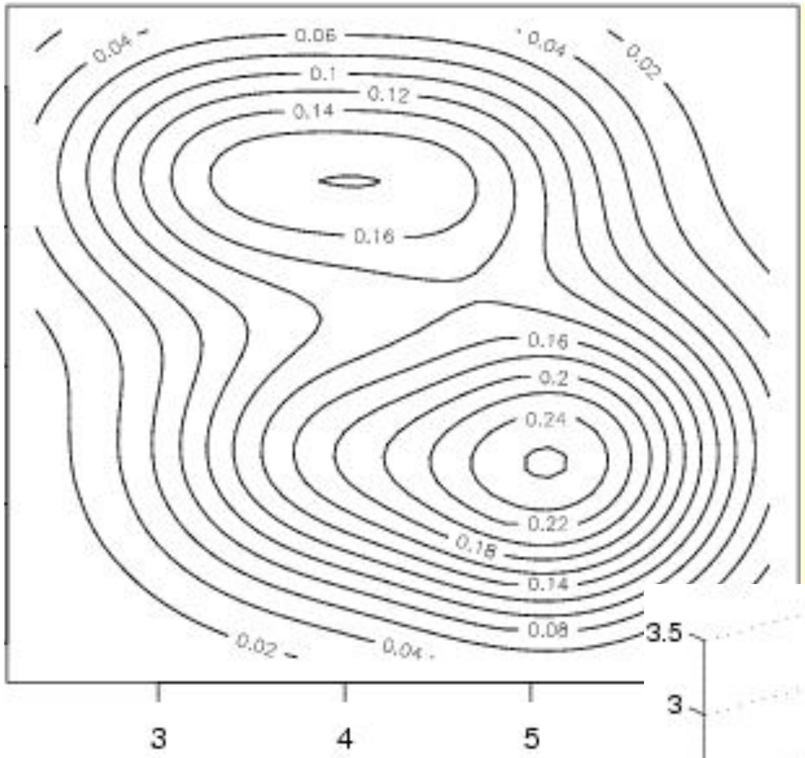
- Set $a^x(t) = a_k$ on (s_{k-1}, s_k), $a^y(t) = a_l$ on (r_{l-1}, r_l)

$$a^{x,y}(t) = a_{k,l} \text{ on } (s_{k-1}, s_k) \times (r_{l-1}, r_l)$$

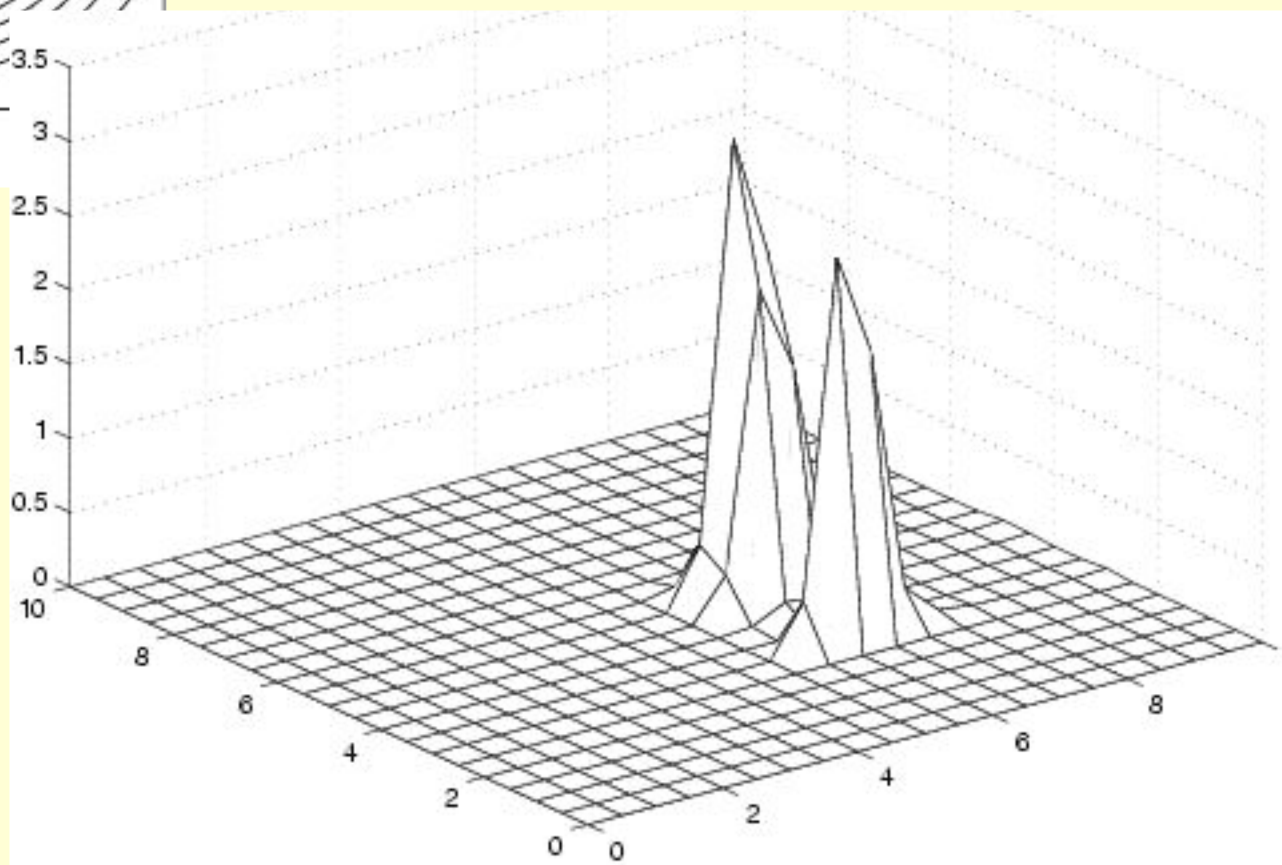


20 samples & unimodal




0.04
0.06
0.04
0.02
0.1
0.12
0.14
0.16
0.16
0.2
0.24
0.22
0.18
0.14
0.08
0.02
0.04
3
4
5

15 samples


3.5
3
2.5
2
1.5
1
0.5
0
10
8
6
4
2
0
0
2
4
6
8


THE • UNIVERSITY • OF • CALIFORNIA
LET THERE BE LIGHT
•1868•