# Statistical Data Analysis, Lecture 4

dr. Dennis Dobler

Vrije Universiteit Amsterdam

24 February 2021

# Topics in this course

1. Summarizing data
2. Exploring distributions
3. Density estimation
4. Bootstrap methods
5. Nonparametric tests
6. Analysis of categorical data
7. Multiple linear regression

# Chapter 4: Density estimation

Contents of Chapter 4:

1. Kernel density estimators
2. Choice of kernel and bandwidth
3. Cross-validation
4. Other density estimators
5. Multivariate density estimation

# Kernel density estimation (continued)

# Choice of kernel and bandwidth (4)

Objective criterion for choosing $K$, $h$:
minimizers of mean integrated squared error (MISE):

$$MISE(\hat{f}) = \int MSE(\hat{f}(t))dt = \int var(\hat{f}(t))dt + \int (E\hat{f}(t) - f(t))^2 dt.$$

# Choice of kernel and bandwidth (4)

Objective criterion for choosing $K$, $h$:
minimizers of mean integrated squared error (MISE):

$$MISE(\hat{f}) = \int MSE(\hat{f}(t))dt = \int var(\hat{f}(t))dt + \int (E\hat{f}(t) - f(t))^2 dt.$$

## Lemma

*For all* $n, h$*,*

$$\int var(\hat{f}(t))dt \leq \frac{1}{nh} \int K(x)^2 dx.$$

# Choice of kernel and bandwidth (4)

Objective criterion for choosing $K$, $h$:
minimizers of mean integrated squared error (MISE):

$$MISE(\hat{f}) = \int MSE(\hat{f}(t))dt = \int var(\hat{f}(t))dt + \int (E\hat{f}(t) - f(t))^2 dt.$$

### Lemma

*For all $n, h$,*

$$\int var(\hat{f}(t))dt \leq \frac{1}{nh} \int K(x)^2 dx.$$

### Lemma

*Assume $f$ twice continuously differentiable. As $h \downarrow 0$,*

$$\int (E\hat{f}(t) - f(t))^2 dt \approx \frac{h^4}{4} \int (f''(t))^2 dt.$$

# Choice of kernel and bandwidth (5)

$$MISE(\hat{f}) \lesssim \frac{1}{nh}\int K(x)^2 dx + \frac{h^4}{4}\int (f''(t))^2 dt.$$

**Theorem**

*Assume $f$ twice continuously differentiable. Optimal bandwidth:*

$$h_{opt} = \Big\{ \int K(x)^2 dx \Big\}^{1/5} \Big\{ \int (f''(t))^2 \Big\}^{-1/5} n^{-1/5}.$$

Re-inserting $h_{opt}$, $MISE(\hat{f})$ minimized by minimizing $\int K(x)^2 dx$.

# Choice of kernel and bandwidth (5)

$$MISE(\hat{f}) \lesssim \frac{1}{nh}\int K(x)^2 dx + \frac{h^4}{4}\int (f''(t))^2 dt.$$

**Theorem**

*Assume f twice continuously differentiable. Optimal bandwidth:*

$$h_{opt} = \Big\{ \int K(x)^2 dx \Big\}^{1/5} \Big\{ \int (f''(t))^2 \Big\}^{-1/5} n^{-1/5}.$$

Re-inserting $h_{opt}$, $MISE(\hat{f})$ minimized by minimizing $\int K(x)^2 dx$.
Minimizing kernel: Epanechnikov kernel

$$K_e(x) = \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}x^2) \quad \text{if } -\sqrt{5} \le x \le \sqrt{5}; \qquad 0 \text{ otherwise.}$$

# Choice of kernel and bandwidth (6)

$$h_{opt} = \Big\{ \int K(x)^2 dx \Big\}^{1/5} \Big\{ \int (f''(t))^2 \Big\}^{-1/5} n^{-1/5}.$$

$\Rightarrow$ Upper bound:

$$MISE(\hat{f}) \lesssim \frac{5}{4} \cdot n^{-4/5} \Big\{ \int K(x)^2 dx \Big\}^{4/5} \Big\{ \int (f''(t))^2 \Big\}^{1/5}$$

# Choice of kernel and bandwidth (6)

$$h_{opt} = \left\{ \int K(x)^2 dx \right\}^{1/5} \left\{ \int (f''(t))^2 \right\}^{-1/5} n^{-1/5}.$$

$\Rightarrow$ Upper bound:

$$MISE(\hat{f}) \lesssim \frac{5}{4} \cdot n^{-4/5} \left\{ \int K(x)^2 dx \right\}^{4/5} \left\{ \int (f''(t))^2 \right\}^{1/5}$$

$\Rightarrow$ Choice of $K$ not too important unless

$$\left\{ \int K(x)^2 dx \right\} / \left\{ \int K_e(x)^2 dx \right\} \gg 1.$$

E.g. $\left\{ \int K_{Gauss}(x)^2 dx \right\} / \left\{ \int K_e(x)^2 dx = \frac{\frac{1}{2\sqrt{\pi}}}{\sqrt{5} \cdot \frac{3}{25}} \right\} \approx 1.05$
$\Rightarrow$ $K_{Gauss}$ almost as good as $K_e$.

# Choice of kernel and bandwidth (7)

$$h_{opt} = \left\{ \int K(x)^2 dx \right\}^{1/5} \left\{ \int (f''(t))^2 \right\}^{-1/5} n^{-1/5}.$$

What to do with $\int (f''(t))^2 dt$? ($f''$ unknown!)

# Choice of kernel and bandwidth (7)

$$h_{opt} = \Big\{ \int K(x)^2 dx \Big\}^{1/5} \Big\{ \int (f''(t))^2 \Big\}^{-1/5} n^{-1/5}.$$

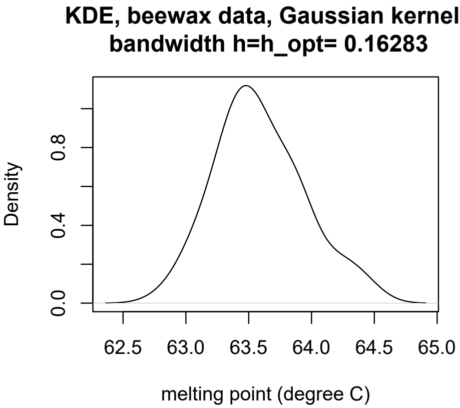
What to do with $\int (f''(t))^2 dt$? ($f''$ unknown!)

Assume $f$ belongs to parametric class of distributions.
E.g., $f_{\mu,\sigma^2}(t) = \frac{1}{\sqrt{2\pi\sigma}} \exp\Big( - \frac{(t-\mu)^2}{2\sigma^2} \Big)$.
$\Rightarrow \int (f''(t))^2 dt \approx 0.212\sigma^{-5}$; estimate $\sigma$ by sample standard deviation.
Using normal kernel, $h_{opt} \approx 1.06\hat{\sigma} n^{-1/5}$.

Adjust if true density multimodal/strongly fluctuating (see syllabus).

# Choice of kernel and bandwidth (8)

Recall Example Melting points (°C) of 59 samples of beewax.
Using Gaussian kernel, $h_{opt} \approx 1.06 \cdot 0.442 \cdot 0.347 \approx 0.163$.
($> 0.1383$, automatic choice by R.)



# Cross-validation

# Cross-validation (1)

Another way to bandwidth: cross-validation (or out-of-sample testing). Objectively, without assumptions on distributions.

Choose minimizer $h^*$ of integrated squared error (ISE):

$$ISE(\hat{f}) = \int (\hat{f}(t) - f(t))^2 dt = \int \hat{f}(t)^2 dt - 2\int \hat{f}(t) f(t) dt + \int f(t)^2 dt,$$

i.e. minimize

$$R(\hat{f}) = \int \hat{f}(t)^2 dt - 2\int \hat{f}(t) f(t) dt.$$

Depends on $f$ ...

# Cross-validation (2)

$$R(\hat{f}) = \int \hat{f}(t)^2 dt - 2 \int \hat{f}(t) f(t) dt.$$

Replace $R(\hat{f})$ by estimate $\hat{R}(\hat{f})$ independent of $f$.

Minimize $\hat{R}(\hat{f})$ w.r.t. $h$.

# Cross-validation (2)

$$R(\hat{f}) = \int \hat{f}(t)^2 dt - 2 \int \hat{f}(t) f(t) dt.$$

Replace $R(\hat{f})$ by estimate $\hat{R}(\hat{f})$ independent of $f$.

Minimize $\hat{R}(\hat{f})$ w.r.t. $h$.

If $X_1, \ldots, X_n, Y \overset{i.i.d.}{\sim} F$ with density $f$,

$$E(\hat{f}(Y) \mid X_1, \ldots, X_n) = \int \hat{f}(t) f(t) dt.$$

# Cross-validation (2)

$$R(\hat{f}) = \int \hat{f}(t)^2 dt - 2 \int \hat{f}(t) f(t) dt.$$

Replace $R(\hat{f})$ by estimate $\hat{R}(\hat{f})$ independent of $f$.
Minimize $\hat{R}(\hat{f})$ w.r.t. $h$.

If $X_1, \ldots, X_n, Y \overset{i.i.d.}{\sim} F$ with density $f$,

$$E(\hat{f}(Y) \mid X_1, \ldots, X_n) = \int \hat{f}(t) f(t) dt.$$

$\Rightarrow$ Use $Y = X_i$; estimate $f$ by $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n \rightsquigarrow$ "$\hat{f}_{-i}$".

# Cross-validation (2)

$$R(\hat{f}) = \int \hat{f}(t)^2 dt - 2 \int \hat{f}(t) f(t) dt.$$

Replace $R(\hat{f})$ by estimate $\hat{R}(\hat{f})$ independent of $f$.

Minimize $\hat{R}(\hat{f})$ w.r.t. $h$.

If $X_1, \ldots, X_n, Y \overset{i.i.d.}{\sim} F$ with density $f$,

$$E(\hat{f}(Y) \mid X_1, \ldots, X_n) = \int \hat{f}(t) f(t) dt.$$

$\Rightarrow$ Use $Y = X_i$; estimate $f$ by $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n \rightsquigarrow$ "$\hat{f}_{-i}$".
Repeat for each $i = 1, \ldots, n$, then average:

$$\frac{1}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i) \approx \int \hat{f}(t) f(t) dt.$$

## Cross-validation (3)

Leads to $\hat{R}(\hat{f}) = \int \hat{f}(t)^2 dt - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)$.

Minimize $h \mapsto \hat{R}(\hat{f}) \Rightarrow h^*$.

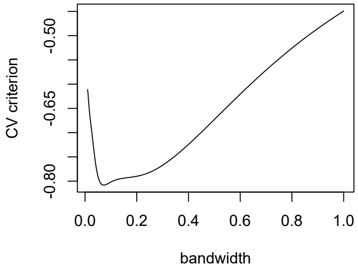
Caution: always verify that $h^*$ is reasonable! ($h^* = 0$ is possible.)

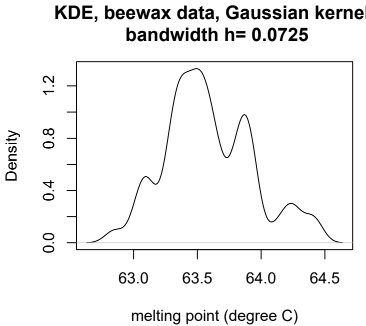
# Cross-validation (3)

Leads to $\hat{R}(\hat{f}) = \int \hat{f}(t)^2 dt - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)$.

Minimize $h \mapsto \hat{R}(\hat{f}) \Rightarrow h^*$.

Caution: always verify that $h^*$ is reasonable! ($h^* = 0$ is possible.)




Cross-validation: not always good results!

# Other density estimators

# Other density estimators (1)

Problem: kernel density estimates possibly positive in undesirable regions.
E.g. positive random variables, Gaussian kernel $\Rightarrow \hat{f}(t) > 0$ for all $t < 0$.

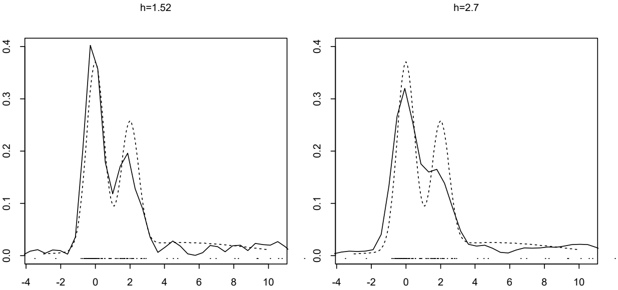
Possible solutions (for positive sample $x_1, \ldots, x_n$):

1. Transform data: $y_i = \log(x_i)$, derive KDE $\hat{f}_y$ for $y$-sample. Transform back: $\hat{f}_x(t) = \frac{1}{t}\hat{f}_y(\log t)$.
2. Symmetrize: $\hat{f}_s$ KDE based on sample $x_1, -x_1, \ldots, x_n, -x_n$. Then use $t \mapsto 2 \cdot \hat{f}_s(t)$ for $t > 0$, and 0 otherwise.
3. Not good: just set $\hat{f}(t)$ to 0 for $t < 0$, and rescale to density.

# Other density estimators (2)

Problem for multimodal densities with heavy tails:
resulting KDE could be oversmoothed, could make estimate unimodal.

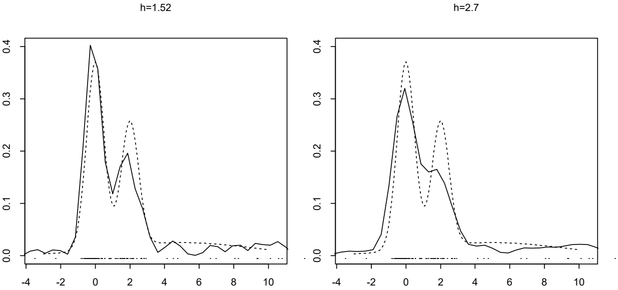
Example: Mixture of 3 different normal distributions



# Other density estimators (2)

Problem for multimodal densities with heavy tails:
resulting KDE could be oversmoothed, could make estimate unimodal.

Example: Mixture of 3 different normal distributions




Possible solution: variable KDE $\hat{f}_v(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{hd_i} K\left(\frac{t-X_i}{hd_i}\right)$.

$d_i$: measure of degree of isolation of $X_i$,
e.g. $k$-th nearest neighbor distance.

Multivariate density estimators

# Multivariate density estimators (1)

Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$ be random vectors, density $f$.

KDE: $\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{\det H}} K(H^{-1/2}(t - X_i))$;

$K$: multivariate density, e.g. multivariate standard normal,

$H$: positive definite bandwidth matrix.

# Multivariate density estimators (1)

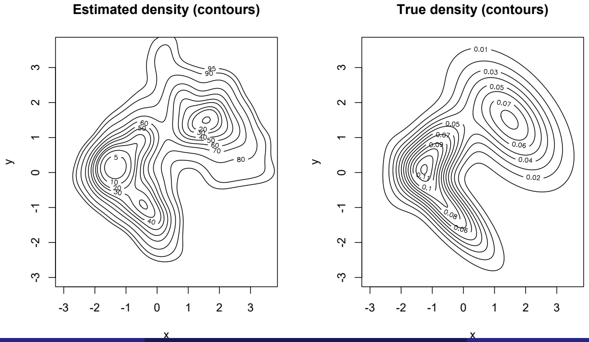
Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$ be random vectors, density $f$.

KDE: $\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{\det H}} K(H^{-1/2}(t - X_i))$;

$K$: multivariate density, e.g. multivariate standard normal,

$H$: positive definite bandwidth matrix.

Example: Mixture of 3 different multivariate normal distributions



# to finish

# To summarize

Today we discussed

- Kernel density estimators
- Choice of kernel and bandwidth
- Cross-validation
- Other density estimators
- Multivariate density estimation

Next week bootstrap!