Nonparametrics,

or

"Can I really assume that distribution is Gaussian?"

Chad Schafer

Department of Statistics, Carnegie Mellon University

cschafer@stat.cmu.edu

Overview

Many of the "standard" statistical inference procedures are based on assumptions regarding the distribution of the observed data.

- Gaussian Errors in Regression
- Small sample hypothesis tests for the population mean
- Any likelihood-based inference: The rise of systematic errors

Example, Motivation

The luminosity function gives the number of (quasars, galaxies, galaxy clusters, . . .) as a function of luminosity (L) or absolute magnitude (M)

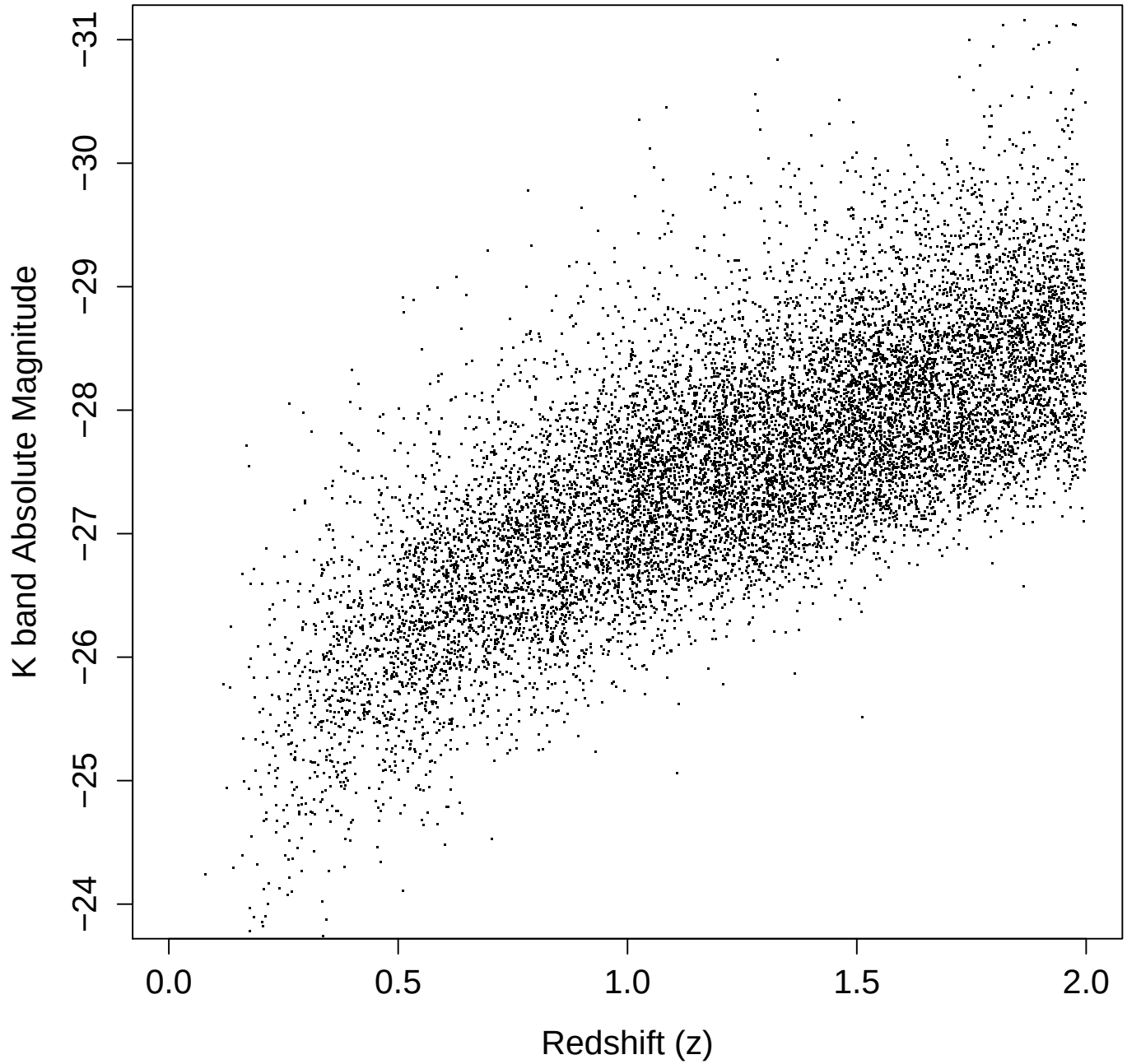
The 2-d luminosity function allows for dependence on redshift (z)

Schechter Form (Schechter, 1976):

$$\phi_S(L)\, dL = n_\star \left(\frac{L}{L^\star}\right) \exp\left(-\frac{L}{L^\star}\right)^\alpha d\left(\frac{L}{L^\star}\right)$$

or

$$\phi_S(M)\, dM \propto n_\star 10^{0.4(\alpha+1)(M^\star - M)} \exp\left(-10^{0.4(M^\star - M)}\right) dM$$



12,626 quasars from Peth, et al. (2011) catalog.

Schechter (1976)

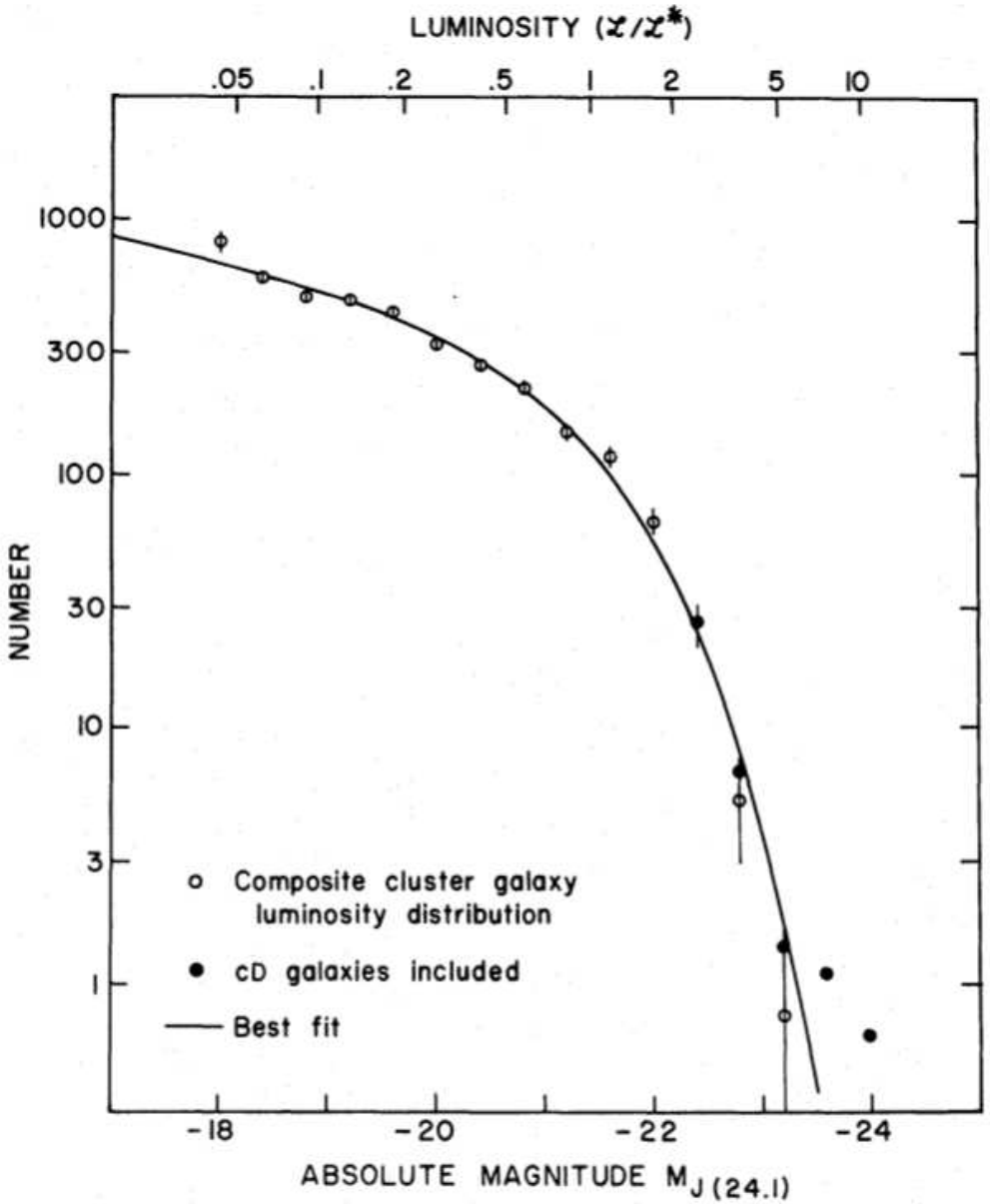


FIG. 2.—Best fit of analytic expression to observed composite cluster galaxy luminosity distribution. Filled circles show the effect of including cD galaxies in composite.

From Binney and Merrifield (1998), pages 163-164:

> "This formula was initially motivated by a simple model of galaxy formation (Press and Schecter, 1974), but it has proved to have a wider range of application than originally envisaged . . .
>
> With larger, deeper surveys, the limitations of the simple Schechter function start to become apparent."

Particular problems on faint end, and with luminosity function evolution with redshift.

Blanton, et al. (2003): Model 2-d galaxy luminosity function as

$$\Phi(M,z) = \overline{n} 10^{0.4(z-z_0)P} \sum_{k=1}^{K} \frac{\Phi_k}{\sqrt{2\pi\sigma_M^2}} \exp\left[-\frac{1}{2}\frac{\left(M - M_k + (z-z_0)\,Q\right)^2}{\sigma_M^2}\right]$$

Motivation

Such challenges are arising in a wide range of situations in cosmology and astronomy

Seek tools for statistical inference that do not rely upon an assumed form for the distribution of the data – These are nonparametric procedures

First, we will consider classic nonparametric tools – procedures that do not rely upon distributional assumptions

These are largely based on ranks, and hence discard some amount of useful information in the data – the assumption versus power tradeoff

An Exercise

Suppose I have a sample of size n from a single population. I want to test the null hypothesis that the median of this population equals μ, versus the alternative that it is not equal to μ. How could I do this nonparametrically (i.e., without making any assumption regarding the population distribution)?

The Sign Test

Setup: One sample, $X_1, X_2, \ldots, X_n$, drawn from a distribution with median μ.

Null Hypothesis: The median μ equals μ_0.

Alternative Hypothesis: The median μ does not equal μ_0.

The Test Statistic: Let T equal the number of the observations larger than μ_0.

Under the null, T has the binomial(n, p) distribution. This leads directly to a p-value for the test: If T is very large or small, one should reject H_0.

The Sign Test – Simple Example

Testing $H_0 : \mu = 0.5$ versus $H_1 : \mu \neq 0.5$, with $n = 7$.

i	1	2	3	4	5	6	7
X_i	2.26	0.67	2.33	2.27	1.41	-0.54	0.07
$I_{\{X_i > \mu_0\}}$	1	1	1	1	1	0	0

So, $T = 5$.

The p-value will be

$$P(T = 0) + P(T = 1) + P(T = 2) + P(T = 5) + P(T = 6) + P(T = 7)$$

under the assumption that T has the binomial$(n = 7, p = 0.5)$ distribution. This equals 0.45.

Wilcoxon Signed Ranks Test

Setup: One sample, $X_1, X_2, \ldots, X_n$, drawn from a distribution which is symmetric about μ.

Null Hypothesis: The "center" μ equals μ_0.

The Test Statistic: Rank the values of $|X_i - \mu_0|$ from smallest to largest; let R_i denote the rank of the i^{th} observation. Define

$$V = \sum_{i=1}^{n} I_{\{X_i > \mu_0\}} R_i$$

If V is large or small relative to that expected under the null hypothesis, there is evidence against the null.

Wilcoxon Signed Rank Test – Simple Example

Testing $H_0 : \mu = 0.5$ versus $H_1 : \mu \neq 0.5$, with $n = 7$.

i	1	2	3	4	5	6	7
X_i	2.26	0.67	2.33	2.27	1.41	-0.54	0.07
$\lvert X_i - \mu_0 \rvert$	1.76	0.17	1.83	1.77	0.91	1.04	0.43
R_i	5	1	7	6	3	4	2
$I_{\{X_i > \mu_0\}}$	1	1	1	1	1	0	0

So,

$$V = 5 + 1 + 7 + 6 + 3 = 22$$

Wilcoxon Signed Rank Test

Implemented in R as `wilcox.test()`

```
> wilcox.test(x, conf.int=T, mu = 0.5)

Wilcoxon signed rank test

data:  x
V = 22, p-value = 0.2188
alternative hypothesis: true location is not equal to 0.5
95 percent confidence interval:
 0.065 2.295
sample estimates:
(pseudo)median
         1.185
```

Mann-Whitney-Wilcoxon Test

Setup: Two samples, drawn independently from two populations. Samples need not be of equal sizes.

Null Hypothesis: The two population distributions are the same

Alternative Hypothesis: The two population distributions are the same except for a shift by a constant μ

The Test Statistic: Jointly sort all observations from smallest to largest. Sum the ranks of the observations from population one.

If this sum is large or small relative to that expected under the null hypothesis, there is evidence against the null.

Mann-Whitney-Wilcoxon Test – Example

From "Ice Mineralogy Across and Into the Surfaces of Pluto, Triton, and Eris" by Tegler, et al. (2012):

"For Pluto, we find bulk, hemisphere-averaged, methane abundances of $9.1 \pm 0.5\%, 7.1 \pm 0.4\%,$ and $8.2 \pm 0.3\%$ for sub-Earth longitudes of $10°$, $125°$, and $257°$. Application of the Wilcoxon rank sum test to our measurements finds these small differences are statistically significant."

Table 3. Methane Abundances For Pluto

Band (μm)	%CH_4 (10°)	%CH_4 (125°)	%CH_4 (194°)	%CH_4 (257°)
0.8897			7.9[a]	
1.1645	9.7	6.9		8.8
1.3355	9.2	7.2		7.9
1.7245	8.8	7.6		8.1
1.7968	8.5	6.9		8.3
2.2081	9.2	6.7		8.1
avg	9.1	7.1		8.2
std	0.5	0.4		0.3

[a]Bok Observations, Tegler et al. (2010)

Mann-Whitney-Wilcoxon Test – Example

```
> samp1 = c(9.7,9.2,8.8,8.5,9.2)
> samp2 = c(6.9,7.2,7.6,6.9,6.7)
> wilcox.test(samp1,samp2)

Wilcoxon rank sum test with continuity correction

data:  samp1 and samp2
W = 25, p-value = 0.01167
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(samp1, samp2) :
  cannot compute exact p-value with ties
```

Mann-Whitney-Wilcoxon Test – Example

```
> samp1 = c(9.7,9.2,8.8,8.5,9.19)
> samp2 = c(6.9,7.2,7.6,6.89,6.7)
> wilcox.test(samp1,samp2)

Wilcoxon rank sum test

data:  samp1 and samp2
W = 25, p-value = 0.007937
alternative hypothesis: true location shift is not equal to 0
```

Tegler (2012): "We found a 0.8% probability that the abundances at sub-Earth longitudes of 10° and 125° have the same mean. In other words, the difference is statistically significant."

Two-Sample Kolmogorov-Smirnov Test

Setup: Two samples, drawn independently from two populations. Samples need not be of equal sizes.

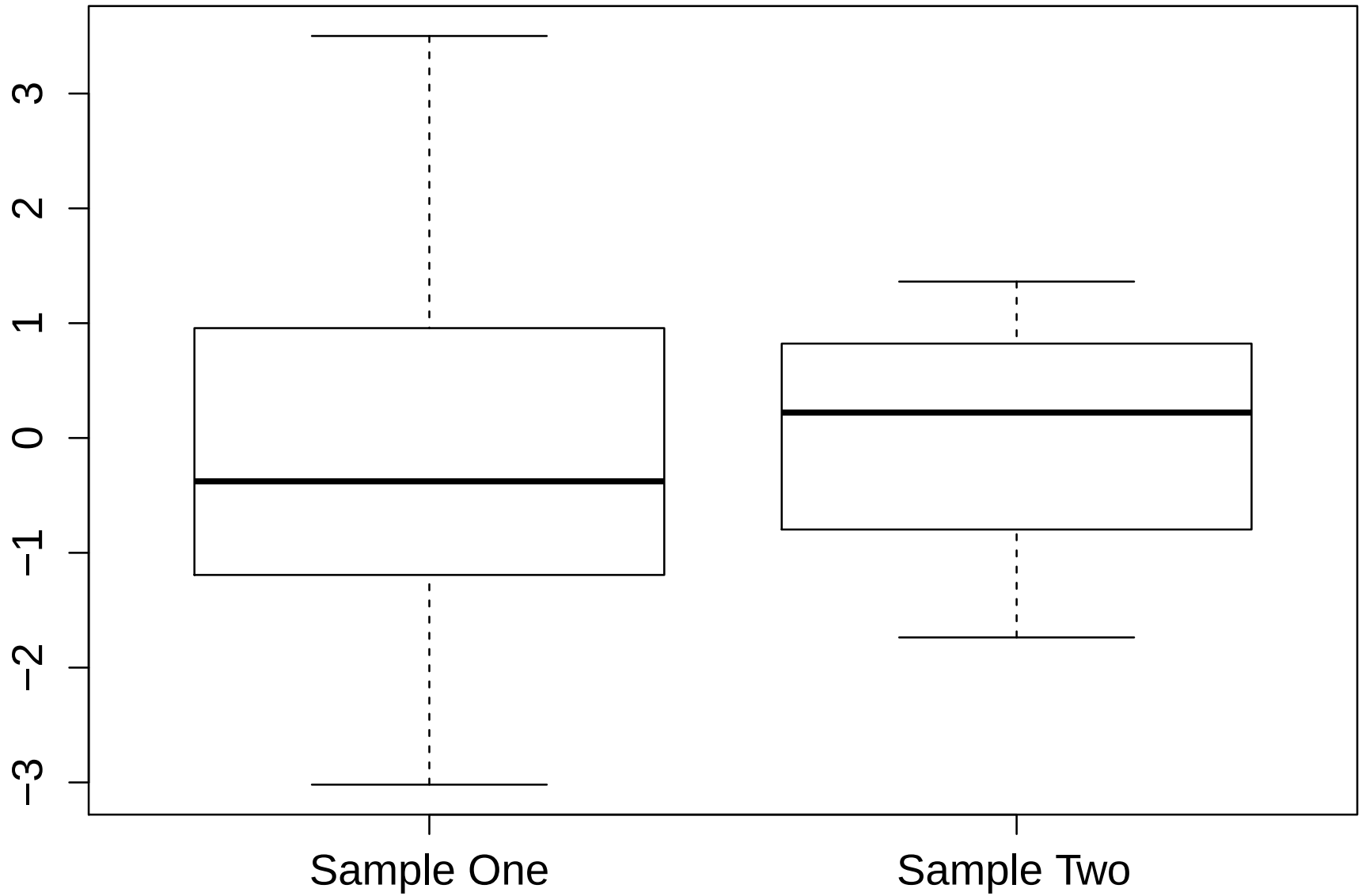
Null Hypothesis: The two population distributions are the same

The Test Statistic: Calculate the empirical CDF for each sample, denoted $\widehat{F}_1(x)$ and $\widehat{F}_2(x)$. Then find

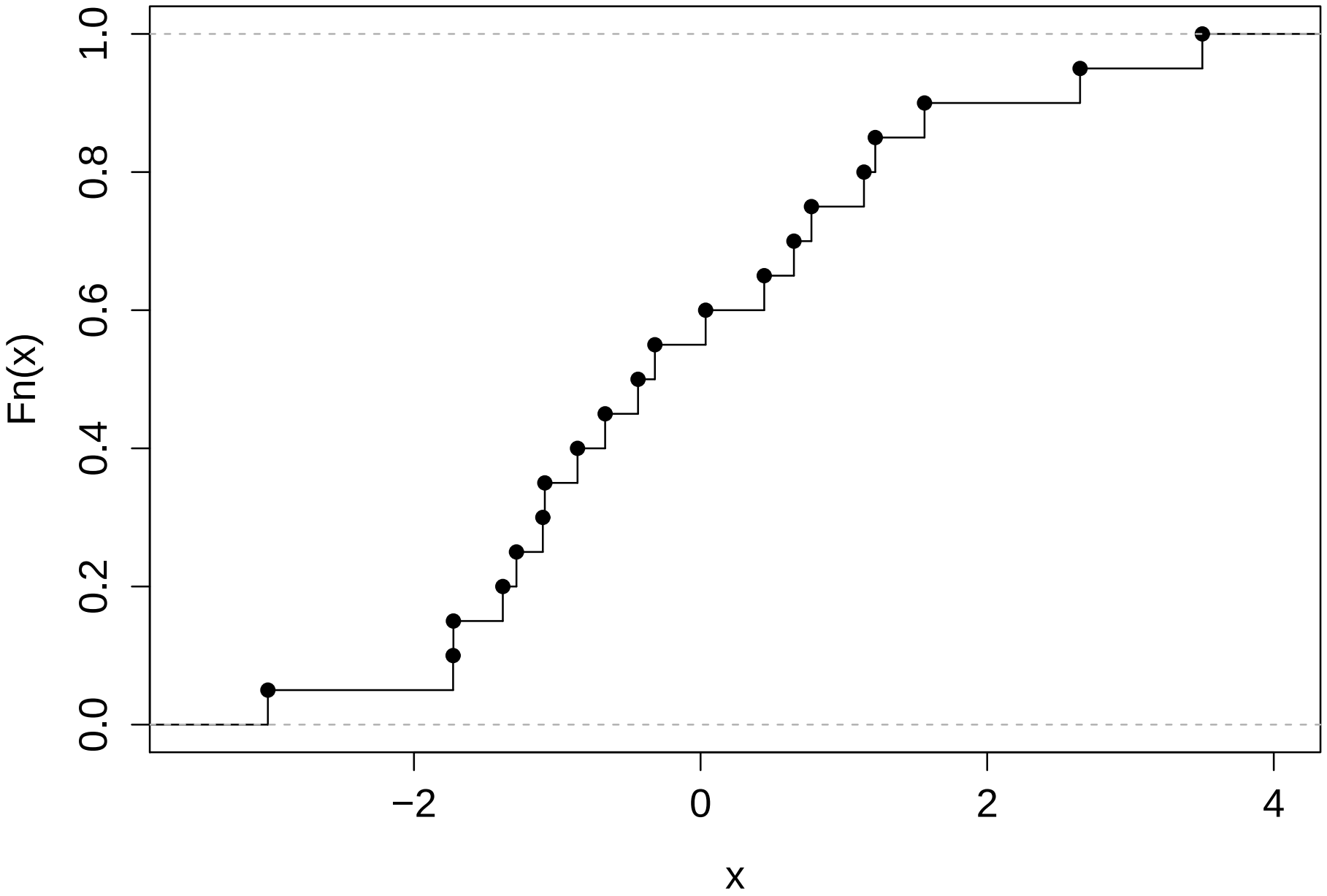
$$D = \max_x \left| \widehat{F}_1(x) - \widehat{F}_2(x) \right|$$

If D is large relative to the expected under the null hypothesis, there is evidence against the null.

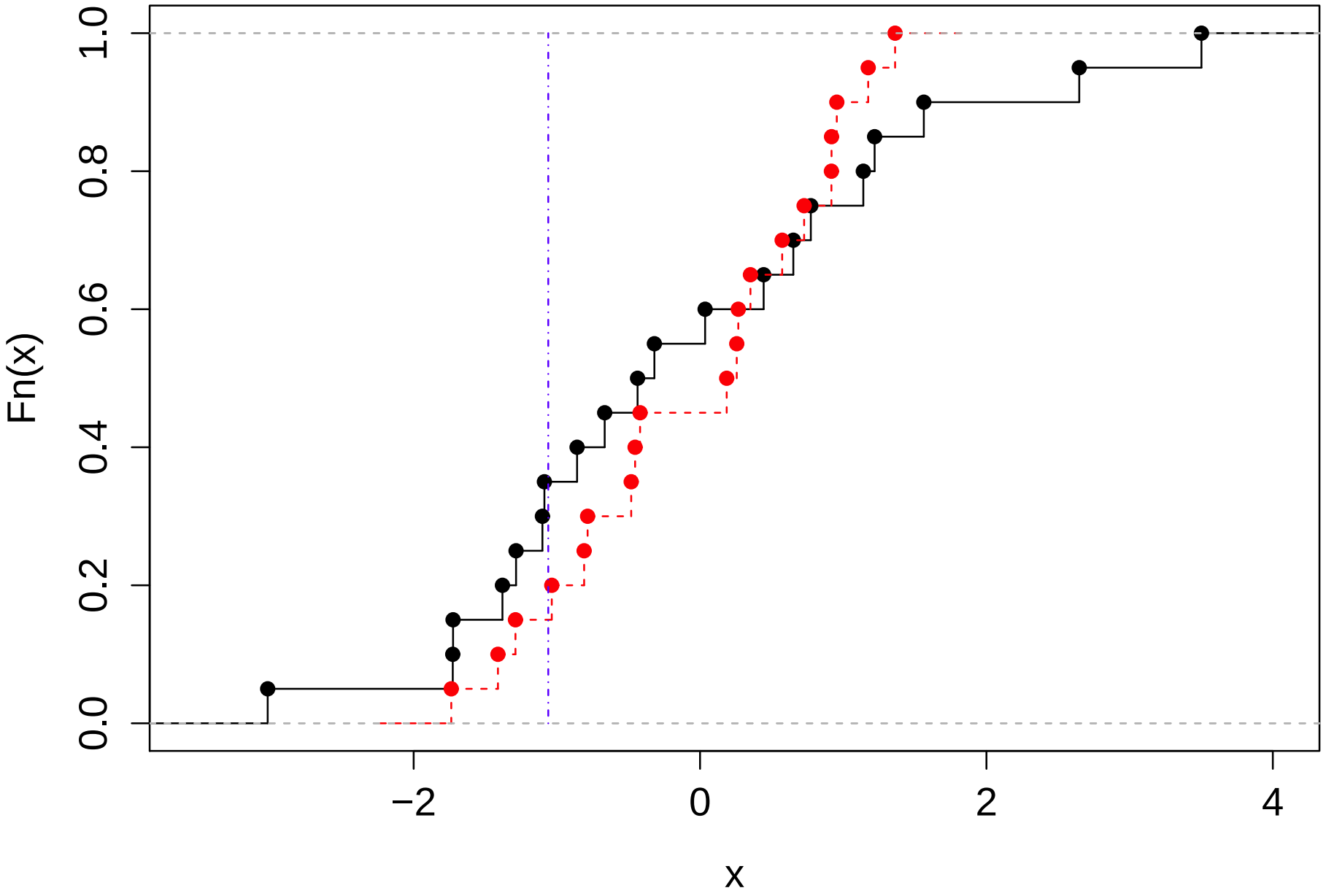
Implemented in R as `ks.test()`



Two simulated samples, each of size 20.



The empirical CDF of the first sample.



Comparing the empirical CDFs of both samples. Note that $D = 0.2$.

R output for this example:

```
> ks.test(x,y)

Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.2, p-value = 0.832
alternative hypothesis: two-sided
```

One-Sample Kolmogorov-Smirnov Test

There is a one-sample version of the K-S Test: Simply replace one of the two empirical CDFs with the CDF of a distribution $F(x)$.

$$D = \max_x \left| \widehat{F}(x) - F(x) \right|$$

Testing $H_0 : X_1, X_2, \ldots, X_n \sim F$

This is also available using `ks.test()` in R: For example,

```
> ks.test(x, pnorm, mean = 10, sd = 2)
```

tests if the sample in `x` comes from the Gaussian distribution with mean 10 and SD 2.

BUT, the p-values reported by `ks.test()` are not valid if the distribution F has parameters that were estimated from the data.

Rank-Based Correlation

Kendall's Tau is a measure of the strength of relationship between two variables X and Y, but not restricted to linear relationships

$$\tau = \frac{\text{number of concordant pair} - \text{number of discordant pairs}}{\text{total number of pairs}}$$

A pair of observations, (X_i, Y_i) and (X_j, Y_j), is concordant if both $X_i < X_j$ and $Y_i < Y_j$. Otherwise, it is discordant.

It holds that $-1 \leq \tau \leq 1$.

In R: `cor(x, y, method = "kendall")`

Rank-Based Correlation

Spearman Rank Correlation is another measure of the strength of relationship between two variables X and Y not restricted to linear relationships.

ρ is the standard Pearson correlation coefficient calculated on the ranks of the X and Y, instead of on the original variables.

Again, it holds that $-1 \leq \rho \leq 1$.

In R: `cor(x, y, method = "spearman")`

If there is a perfect increasing (decreasing) relationship between X and Y, then $\rho = 1$ and $\tau = 1$ ($\rho = -1$ and $\tau = -1$).

Nonparametric Denstiy Estimation

Now, we will shift into a discussion of how to estimate distributions

Recall our earlier discussion of luminosity function estimation

Luminosity Functions as Probability Density Functions

Integrating a luminosity function gives the count in that bin

$$\text{count with } -27 \leq M \leq -26 = \int_{-27}^{-26} \phi(M)\, dM$$

Proportional to the probability a randomly chosen such object is such that $-27 \leq M \leq -26$.

Hence, estimating $\phi(M)$ analogous to density estimation

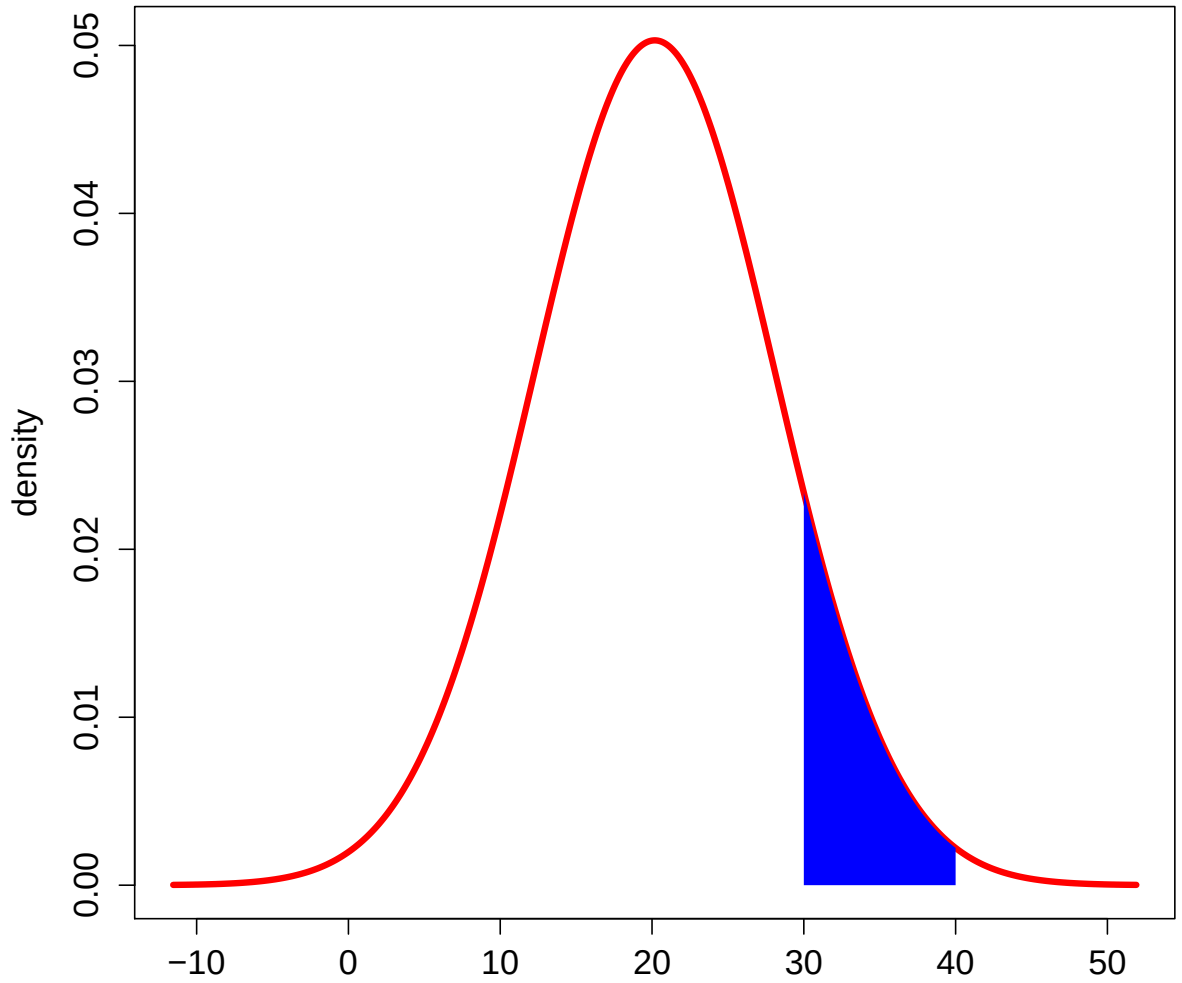
Density Estimation

Basic Problem: Estimate f, where assume that $X_1, X_2, \ldots, X_n \sim f$, i.e., for $a \leq b$,

$$P(a \leq X_i \leq b) = \int_a^b f(x)\, dx$$

Parametric Approach: Assume f belongs to some family of densities (e.g., Gaussian), and estimate the unknown parameters via maximum likelihood

Parametric Estimator



Finding the probability between 30 and 40 under Gaussian assumption. $(\mu = 20, \sigma = 8)$.

Advantages of Parametric Form

Relatively simple to fit, via maximum likelihood

Functional form

Easy calculation of probabilities

Smaller errors, on average, provided assumption regarding density is correct

Density Estimation

Basic Problem: Estimate f, where assume that $X_1, X_2, \ldots, X_n \sim f$, i.e., for $a \leq b$,

$$P(a \leq X_i \leq b) = \int_a^b f(x)\, dx$$

Parametric Approach: Assume f belongs to some family of densities (e.g., Gaussian), and estimate the unknown parameters via maximum likelihood

Nonparametric Approach: Estimate built on smoothing available data, slower rate of convergence, but less bias

Histograms

Create bins $B_1, B_2, \ldots, B_m$ of width h. Define

$$\widehat{f}_n(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} I(x \in B_j).$$

where $\widehat{p}_j$ is proportion of observations in B_j. Note that

$$\int \widehat{f}_n(x) dx = 1.$$

A histogram is an example of a nonparametric density estimator. Note that it is controlled by the tuning parameter h, the bin width.

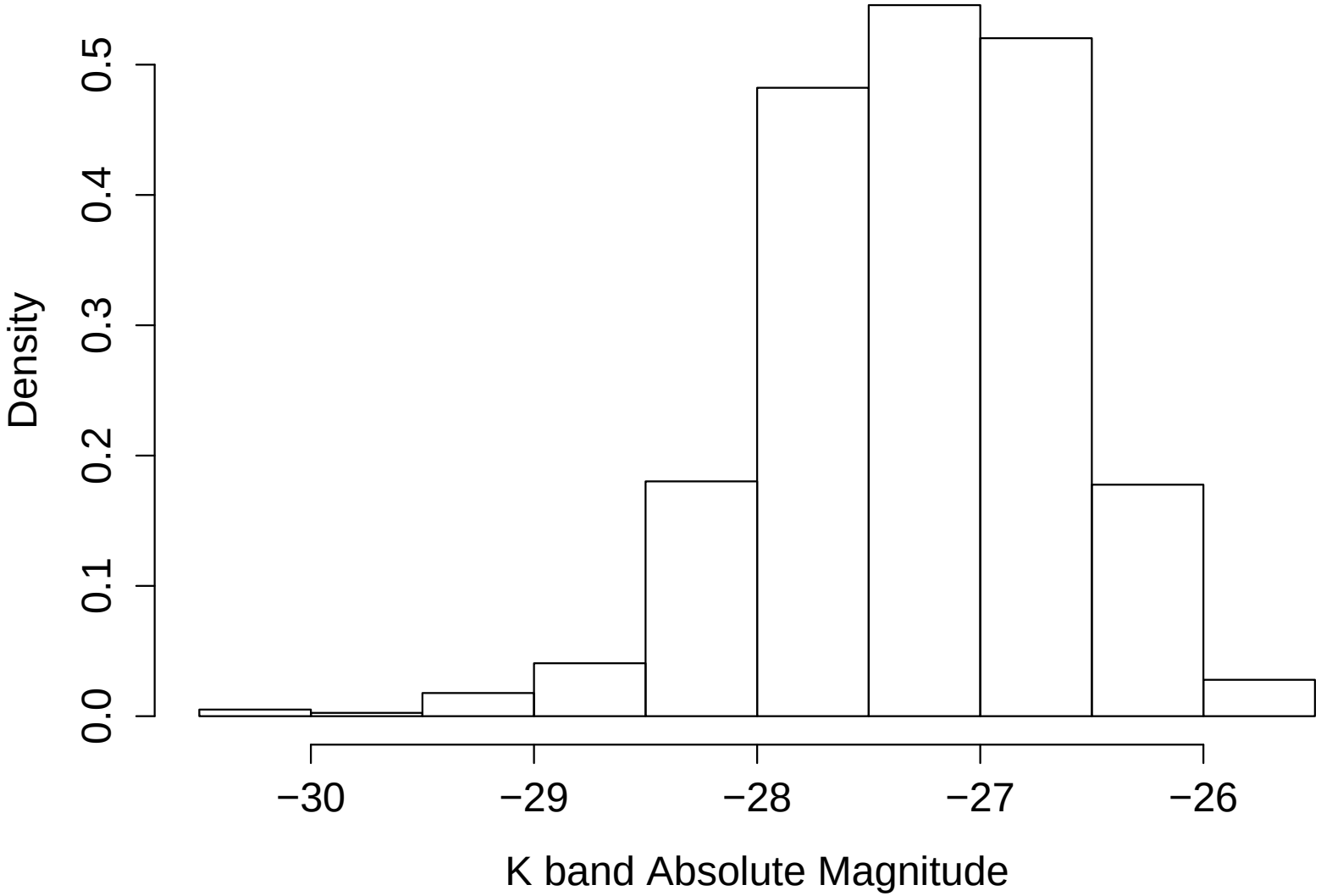
Density Estimation

Sample of 788 absolute magnitudes of quasars with $0.95 < z < 1.05$:

```
-30.3550, -30.1197, -29.7318, -29.4338, -29.3156, -29.1965, -29.1557,
-29.0450, -29.0266, -29.0019, -28.9094, -28.8384, -28.8189, -28.8157,
-28.8025, -28.7646, -28.7486, -28.7220, -28.6730, -28.6572, ...
```

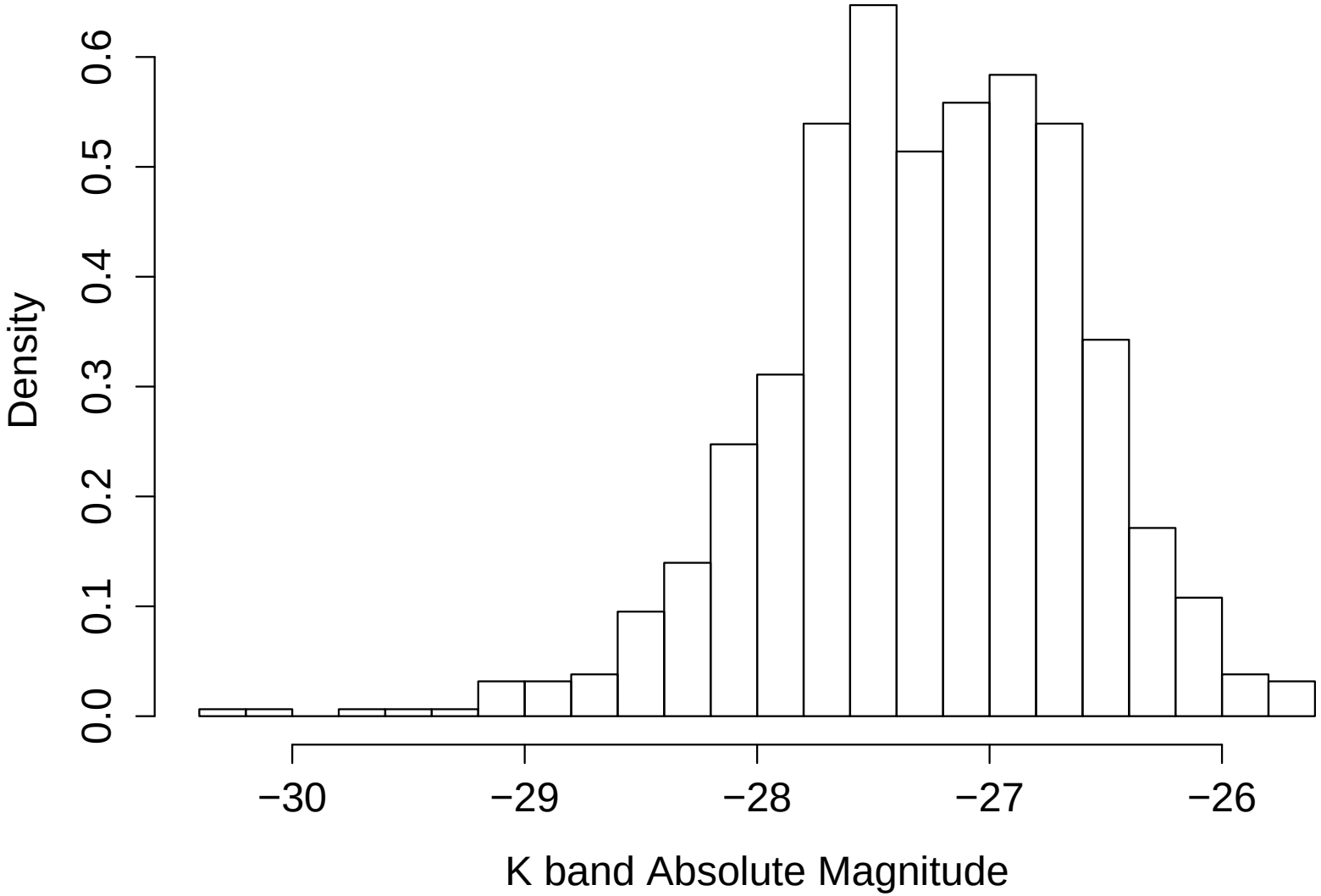
Note that $\overline{x} = -27.28, s = 0.648$

Nonparametric Estimator



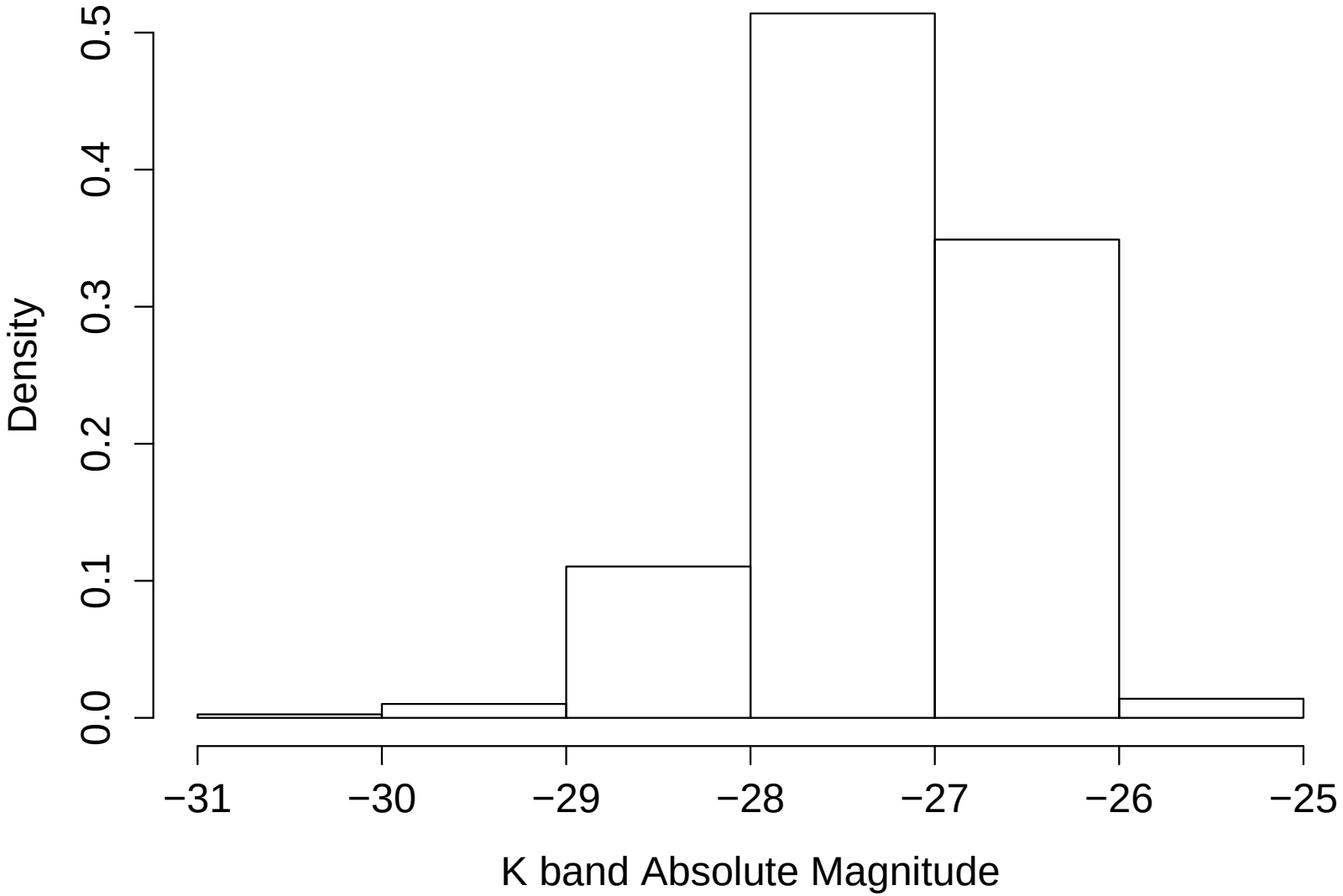
Histogram of the absolute magnitudes. (Using `hist()` in R.)

Histograms



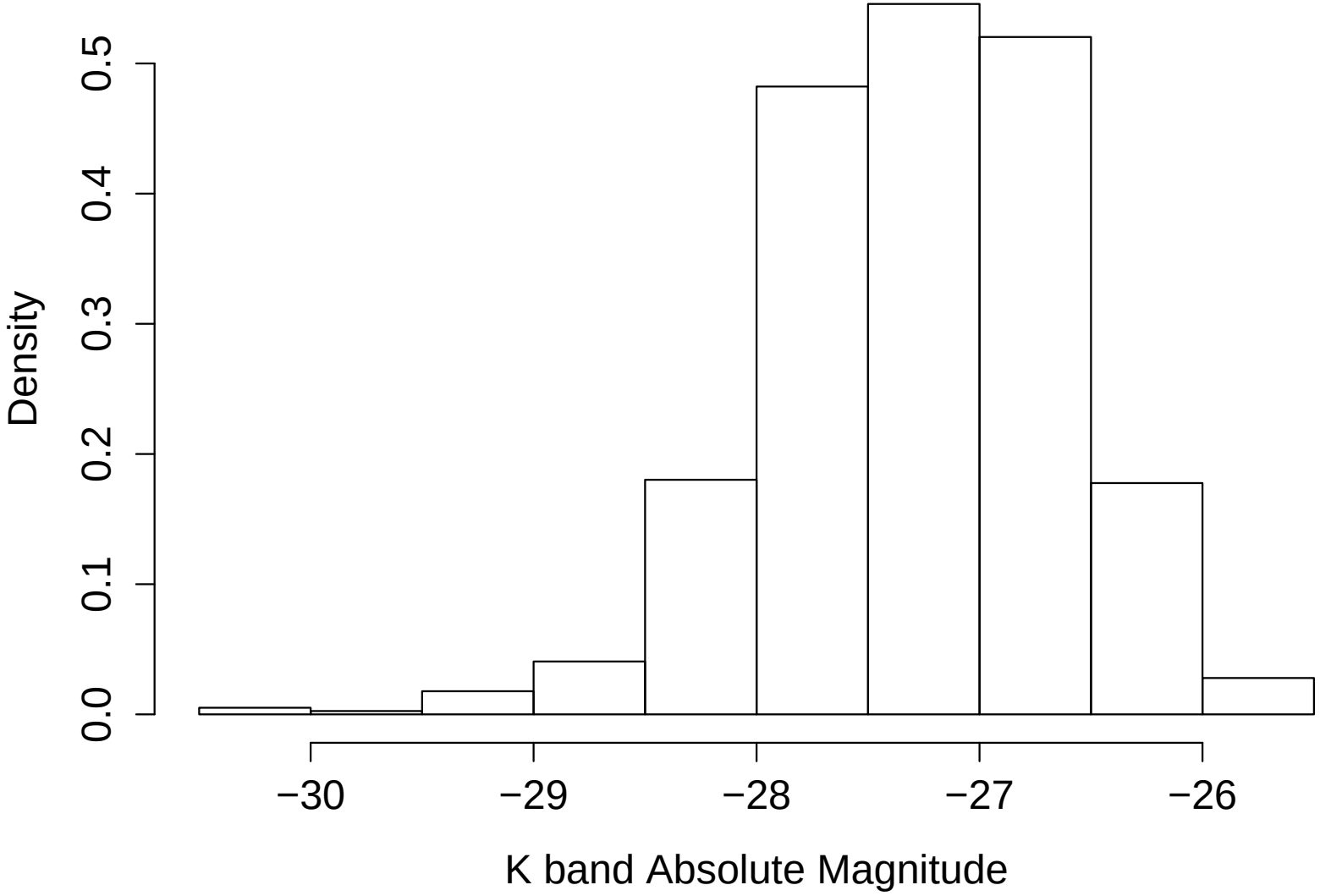
Seems like too many bins, i.e. h is too small.

Histograms



Seems like too few bins, i.e. h is too large.

Nonparametric Estimator



Seems better.

Errors in Density Estimators

Error at one value x:

$$(f(x) - \widehat{f}(x))^2$$

Error accumulated over all x:

$$\text{ISE} = \int (f(x) - \widehat{f}(x))^2 dx$$

Mean Integrated Squared Error (MISE):

$$\text{MISE} = \mathbb{E}\left(\int (f(x) - \widehat{f}(x))^2 dx \right)$$

The Bias–Variance Tradeoff

Can write

$$\text{MISE} = \int b^2(x)dx + \int v(x)dx$$

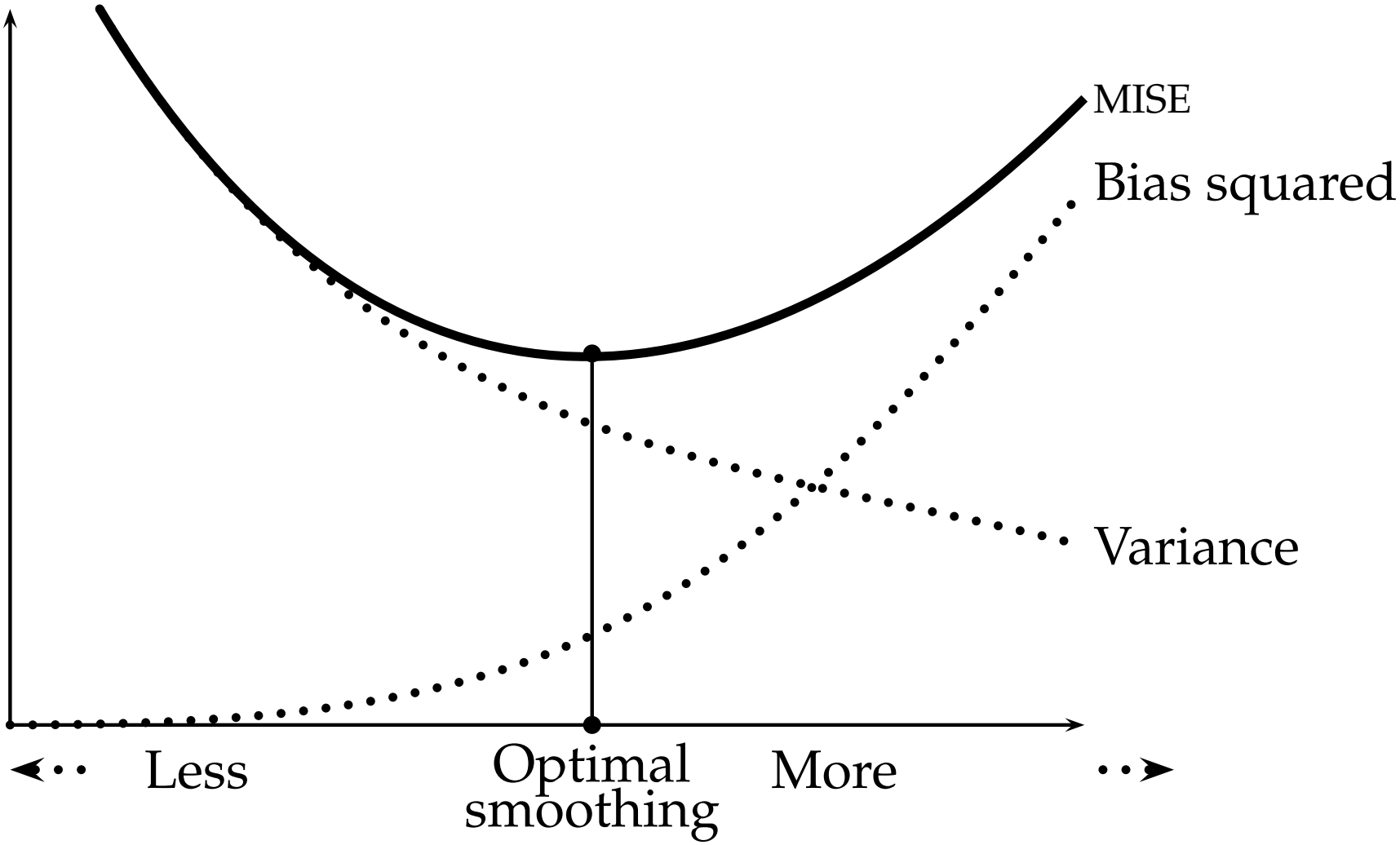
where the bias at x is

$$b(x) = \mathbb{E}(\widehat{f}(x)) - f(x)$$

and the variance at x is

$$v(x) = \mathbb{V}(\widehat{f}(x))$$

The Bias–Variance Tradeoff



Histograms

For histograms,

$$\text{MISE} = \mathbb{E}\left(\int (f(x) - \widehat{f}(x))^2 dx\right) \approx \frac{h^2}{12}\int (f'(u))^2 du + \frac{1}{nh}$$

The value h^* that minimizes this is

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int (f'(u))^2 du}\right)^{1/3}.$$

and then

$$\text{MISE} \sim \frac{C_2}{n^{2/3}}$$

Kernel Density Estimation

We can improve histograms by introducing smoothing.

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Here, $K(\cdot)$ is the kernel function, itself a smooth density.

Kernels

A **kernel** is any smooth function K such that $K(x) \geq 0$ and

$$\int K(x)\,dx = 1, \quad \int xK(x)dx = 0 \quad \text{and} \quad \sigma_K^2 \equiv \int x^2 K(x)dx > 0.$$

Some commonly used kernels are the following:

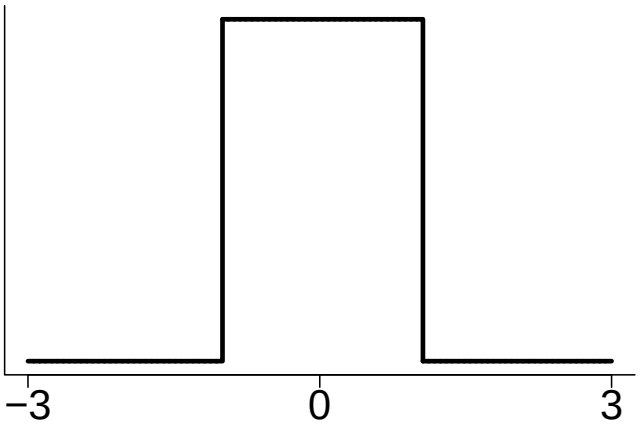
$$\textbf{the boxcar kernel}: \qquad K(x) = \frac{1}{2} I(|x| < 1),$$

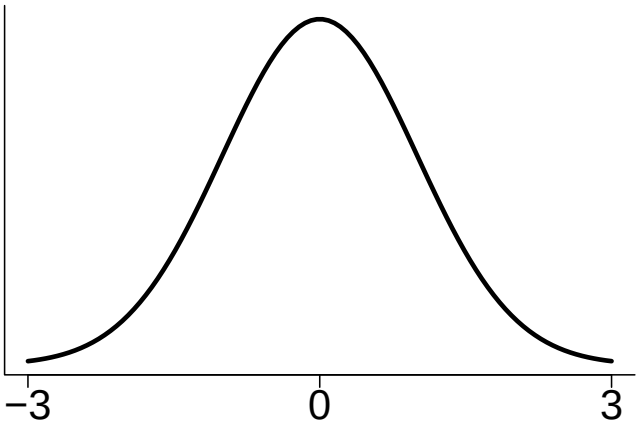
$$\textbf{the Gaussian kernel}: \qquad K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

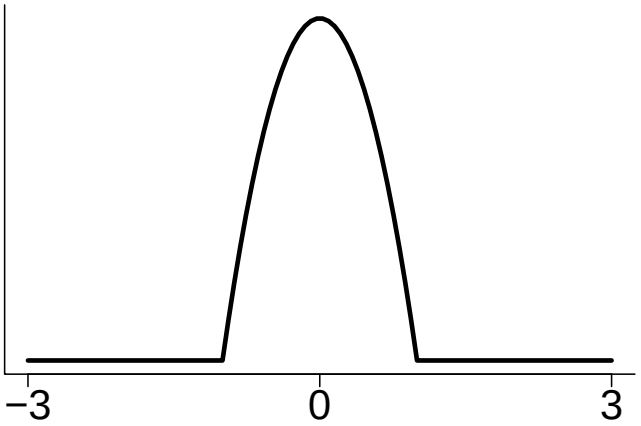
$$\textbf{the Epanechnikov kernel}: \qquad K(x) = \frac{3}{4}(1 - x^2) I(|x| < 1)$$

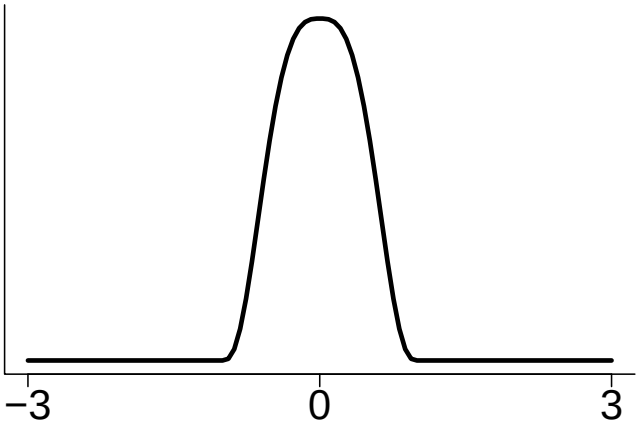
$$\textbf{the tricube kernel}: \qquad K(x) = \frac{70}{81}(1 - |x|^3)^3 I(|x| < 1).$$

Kernels



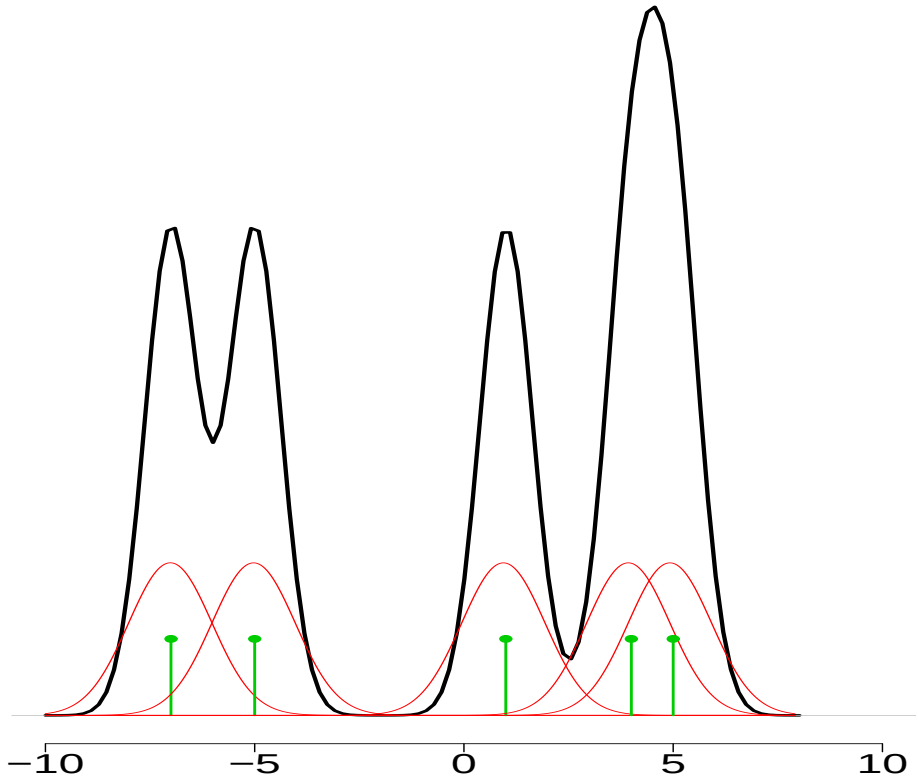






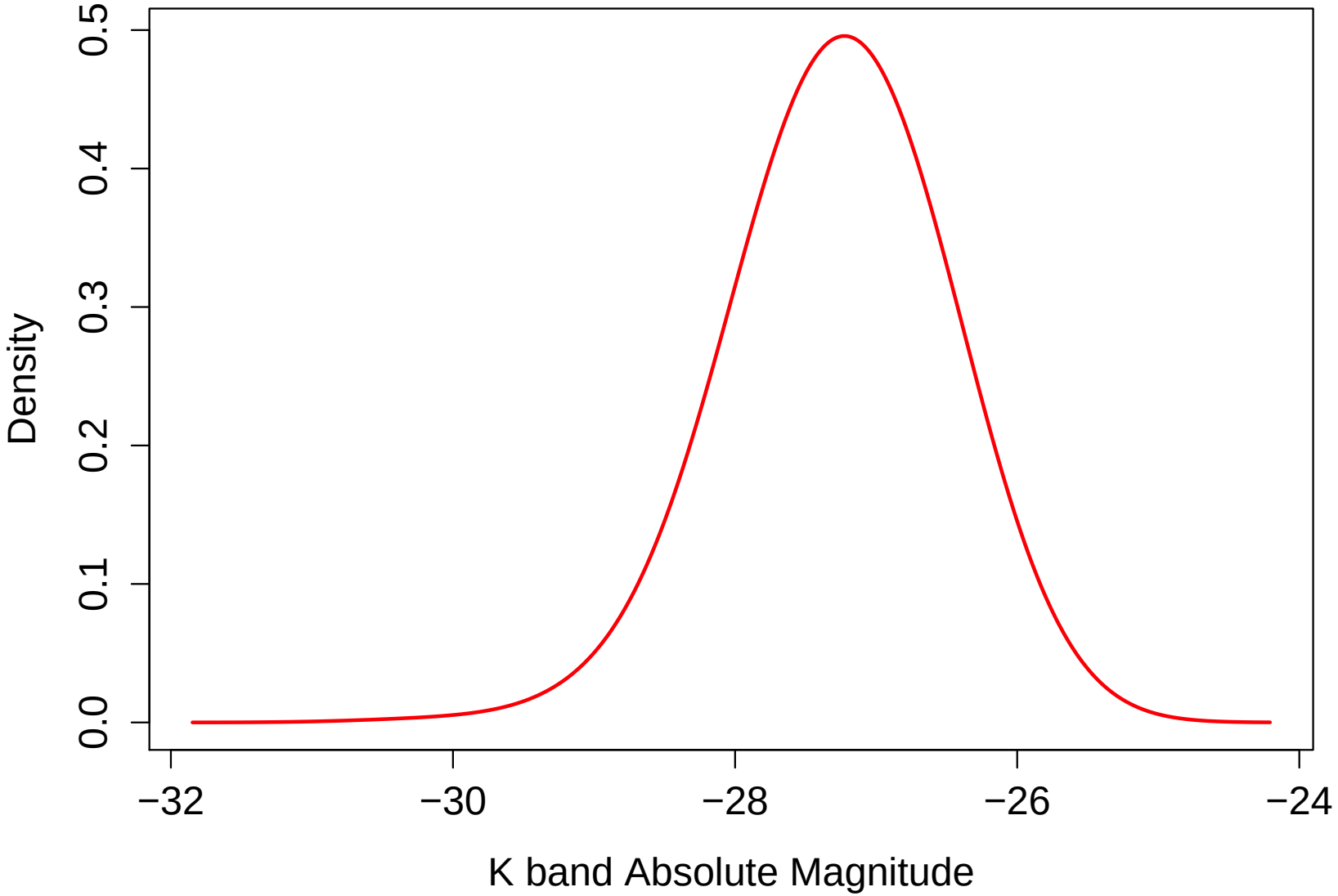
Kernel Density Estimation

Places a smoothed-out lump of mass of size $1/n$ over each data point



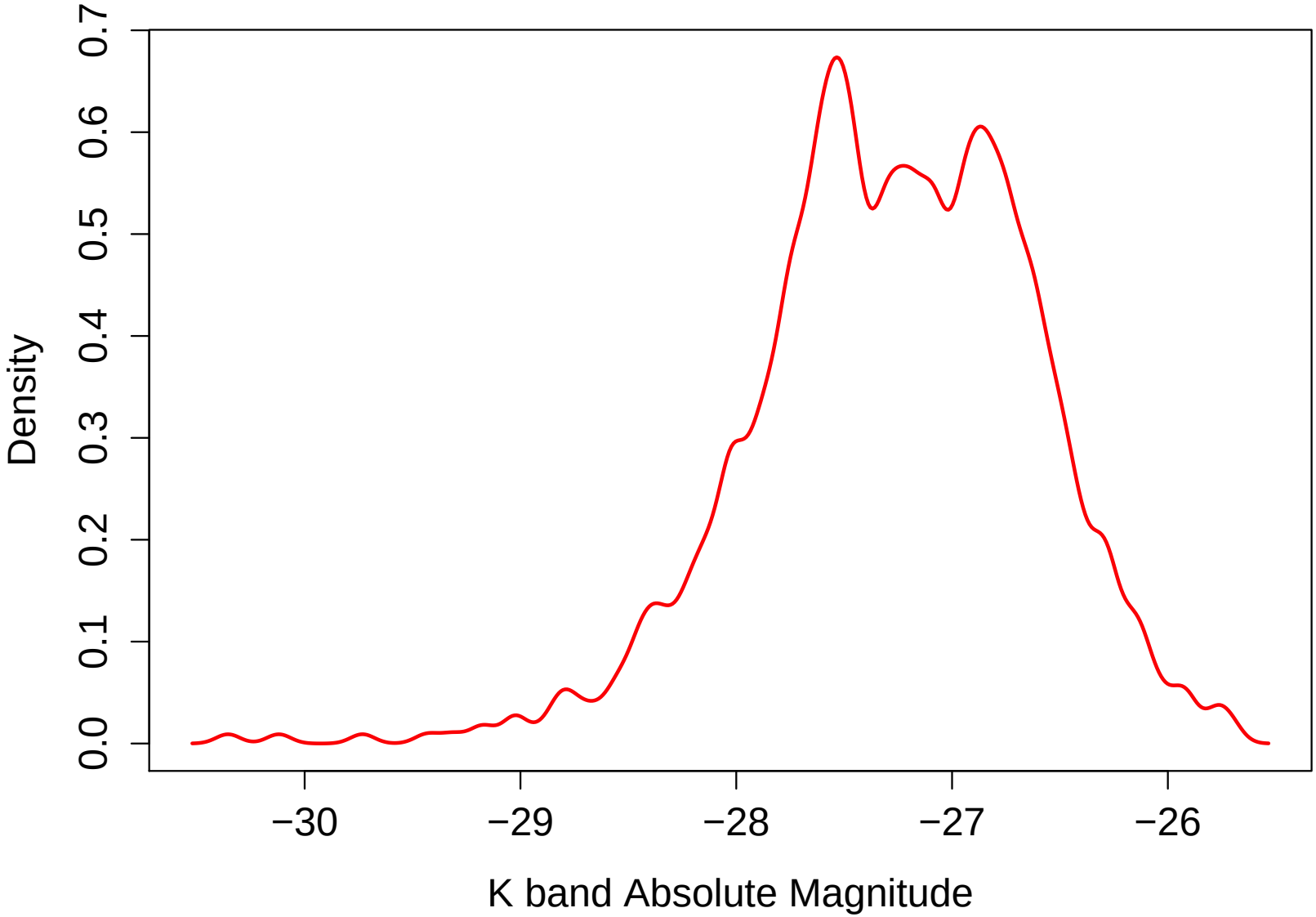
The shape of the "lumps" is controlled by $K(\cdot)$; their width is controlled by h.

Kernel Density Estimation



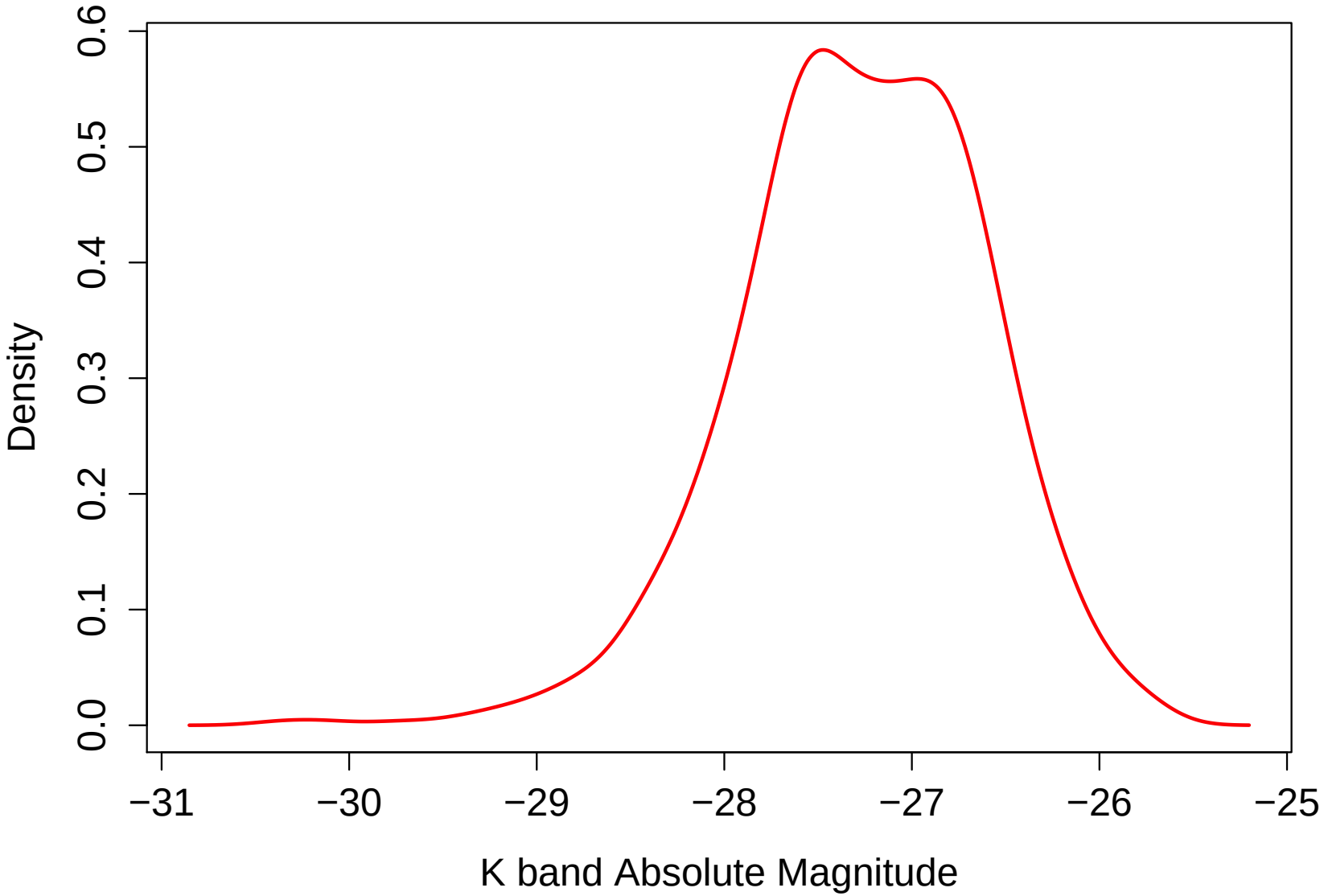
Kernel density estimator with h chosen too large.

Kernel Density Estimation



Kernel density estimator with h chosen too small.

Kernel Density Estimation



Kernel density estimator with h chosen "optimally."

Kernel Density Estimation

$$\text{MISE} \approx \frac{1}{4} c_1^2 h^4 \int (f''(x))^2 dx + \frac{\int K^2(x) dx}{nh}$$

Optimal bandwidth is

$$h_* = \left(\frac{c_2}{c_1^2 A(f) n} \right)^{1/5}$$

where $c_1 = \int x^2 K(x) dx$, $c_2 = \int K(x)^2 dx$ and $A(f) = \int (f''(x))^2 dx$.

Then,

$$\text{MISE} \sim \frac{C_3}{n^{4/5}}.$$

The Bias–Variance Tradeoff

For many smoothers:

$$\text{MISE} \approx c_1 h^4 + \frac{c_2}{nh}$$

which is minimized at

$$h = O\left(\frac{1}{n^{1/5}}\right)$$

Hence,

$$\text{MISE} = O\left(\frac{1}{n^{4/5}}\right)$$

whereas, for parametric problems

$$\text{MISE} = O\left(\frac{1}{n}\right)$$

Comparisons

So, for parametric form, if it's correct,

$$\text{MISE} \sim \frac{C_1}{n}$$

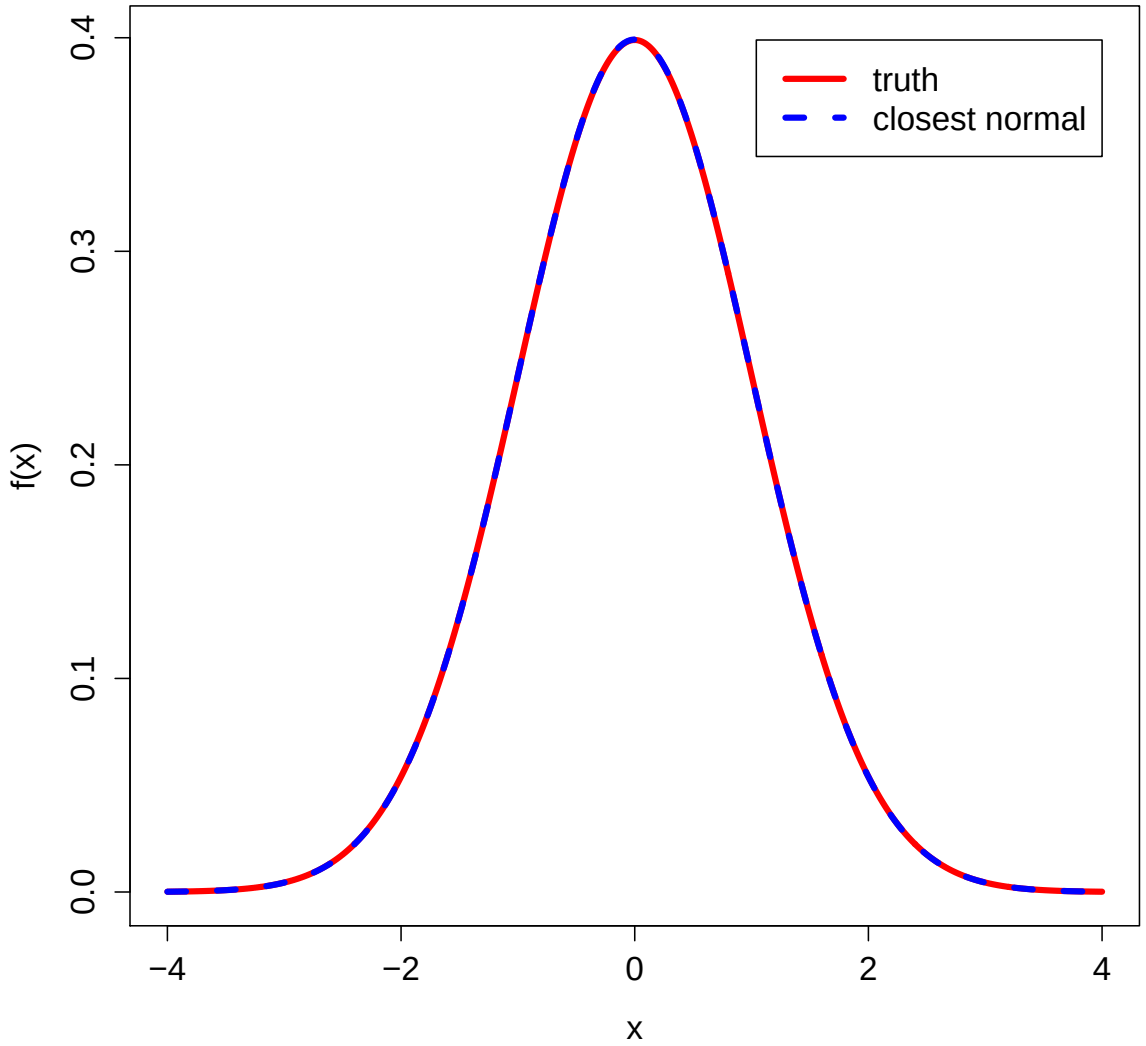
For histograms,

$$\text{MISE} \sim \frac{C_2}{n^{2/3}}$$

For estimators based on smoothing,

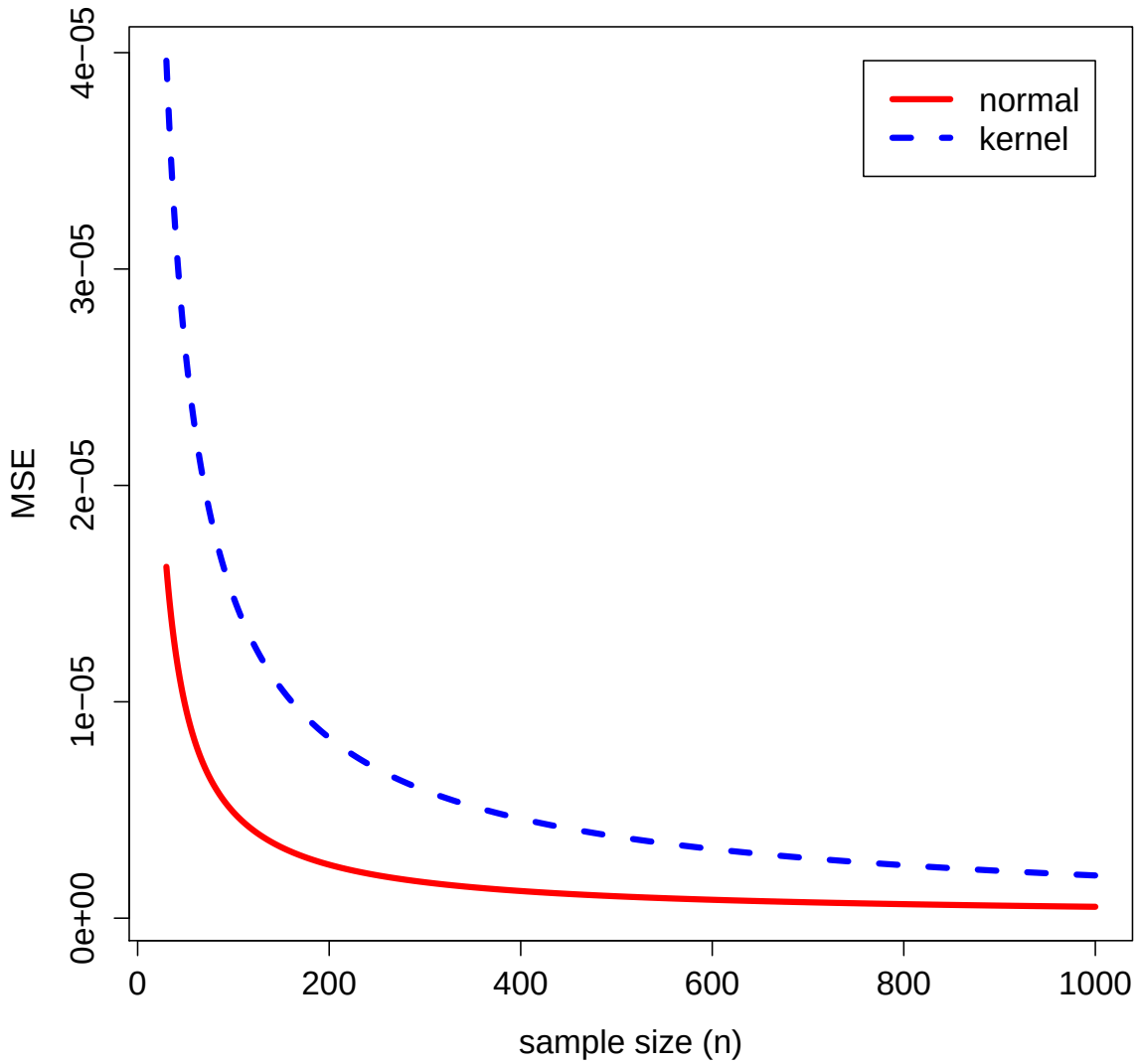
$$\text{MISE} \sim \frac{C_3}{n^{4/5}}$$

Comparisons



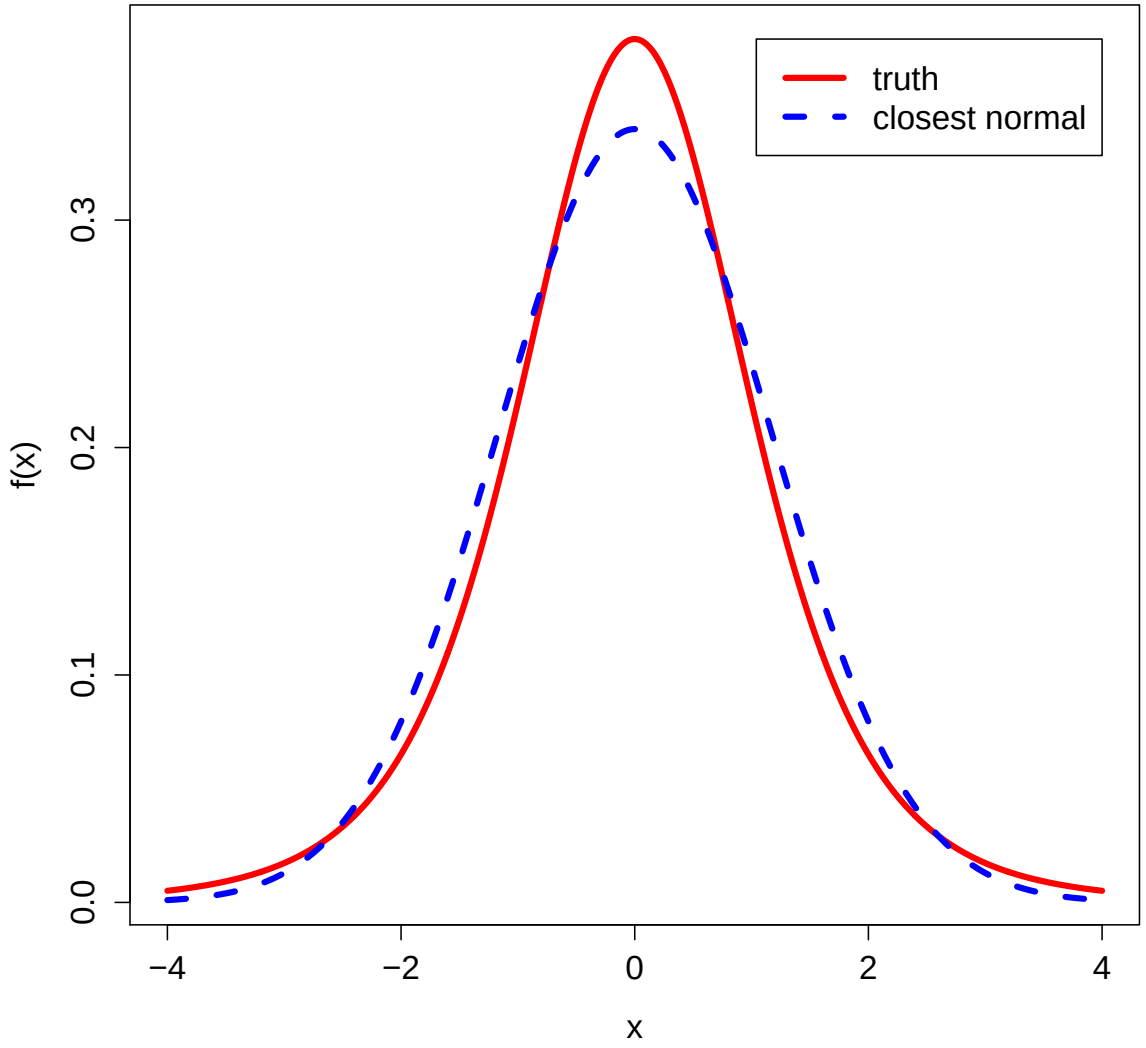
Truth is standard normal.

Comparisons



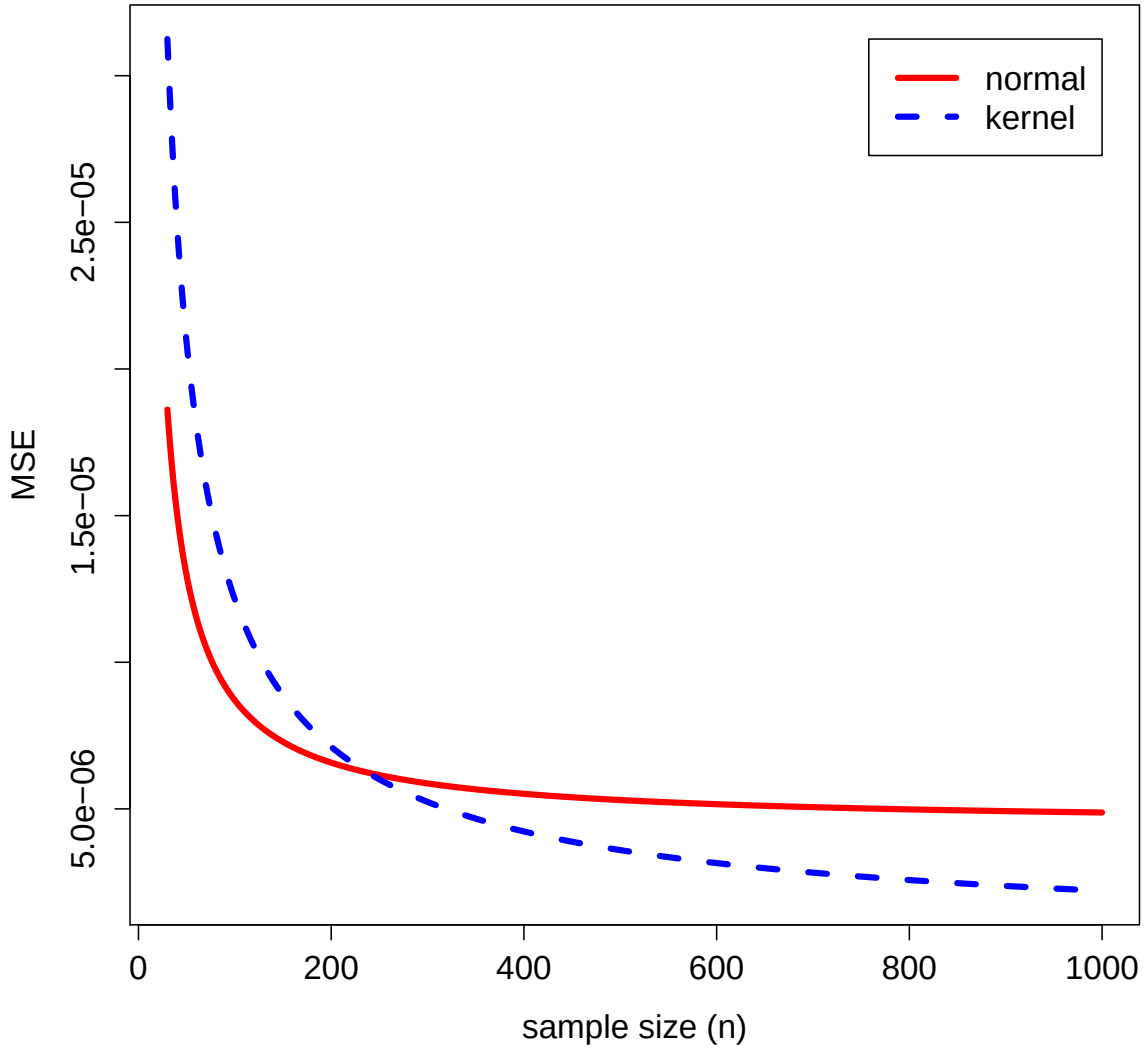
Assuming normality has its advantages.

Comparisons



Truth is t-distribution with 5 degrees of freedom.

Comparisons



Normal assumption costs, even for moderate sample size.

Cross-Validation for Selecting h

Objective, data-driven choice of h is possible:

$$\begin{aligned}
\text{MISE} &= \mathbb{E}\left(\int (f(x) - \widehat{f}(x))^2 dx\right) \\
&= \mathbb{E}\left(\int \widehat{f}(x)^2 dx\right) - 2\mathbb{E}\left(\int \widehat{f}(x) f(x) dx\right) + \int f(x)^2 dx \\
&= J(h) + \int f(x)^2 dx
\end{aligned}$$

Unbiased Estimator of $J(h)$:

$$\widehat{J}(h) = \int \left(\widehat{f}_n(x)\right)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i)$$

where $\widehat{f}_{(-i)}$ is the density estimator obtained after removing the i^{th} observation.

Kernel Density Estimation in R

The function `density()` does kernel density estimation

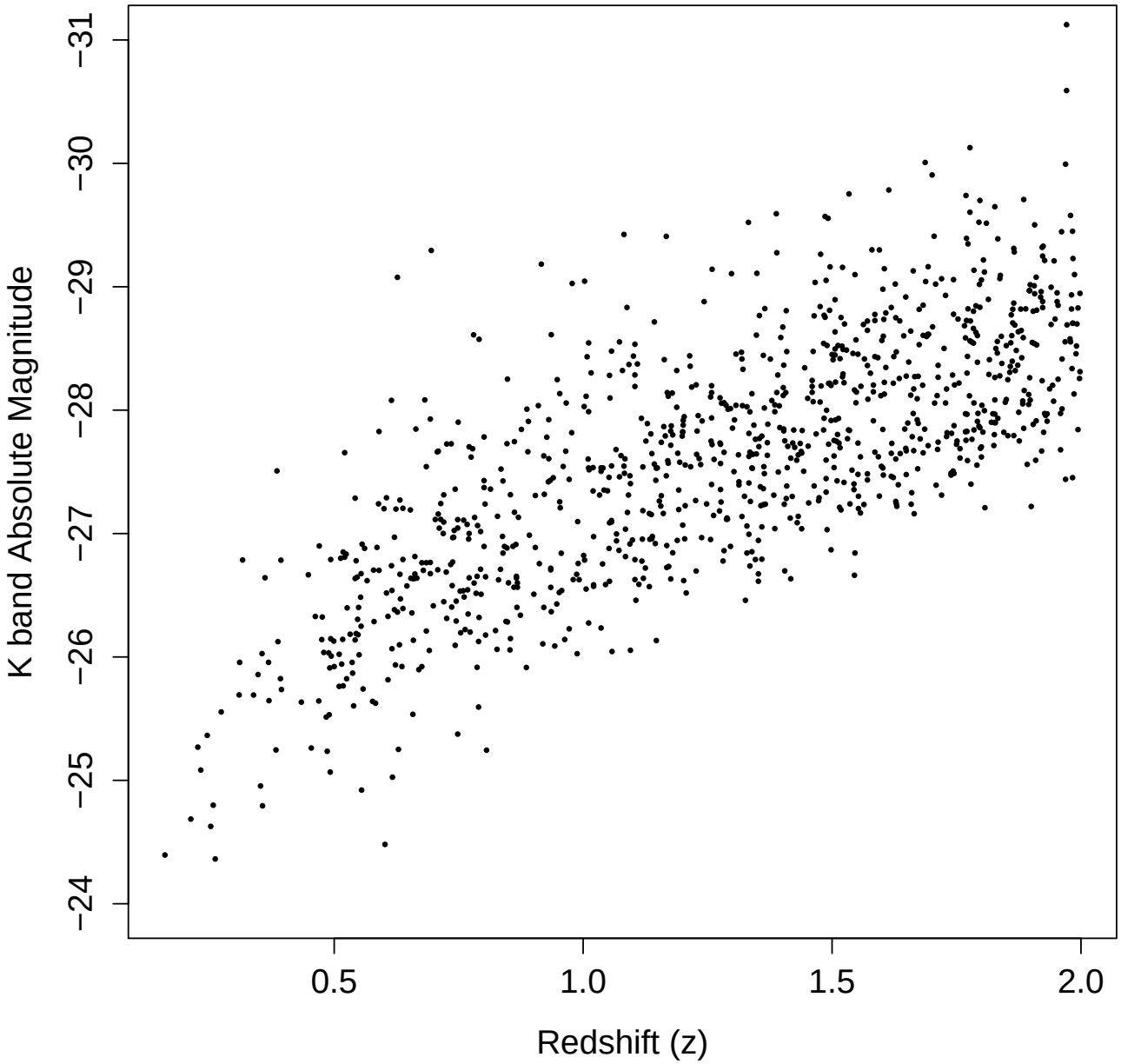
For data stored in `x`, using cross-validation to choose the smoothing parameter (aka "bandwidth") with function `bw.ucv()`:

```
> optsmooth = bw.ucv(x)
> densityest = density(x, bw = optsmooth)
```

Now plot the result:

```
> plot(densityest$x, densityest$y,
   xlab="Absolute Magnitude", ylab="Density",type="l")
```

Multivariate Density Estimation



1,000 quasars from Peth, et al. (2011) catalog.

Multivariate Density Estimation

The kernel density estimator extends naturally to higher dimensions:

If x is d-dimensional,

$$\widehat{f}(x) = \frac{1}{n|\mathbf{H}|^{1/2}} \sum_{i=1}^{n} K\left(\mathbf{H}^{-1}(x - X_i)\right).$$

This puts a d-dimensional "bump" at each data point.

The matrix $\mathbf{H}$ controls the size and shape of the bump.

Multivariate Density Estimation

In R, use package `ks`.

`Hscv()` finds the bandwidth matrix using cross-validation

Warning: Finding the bandwidth can be time consuming. Instead:

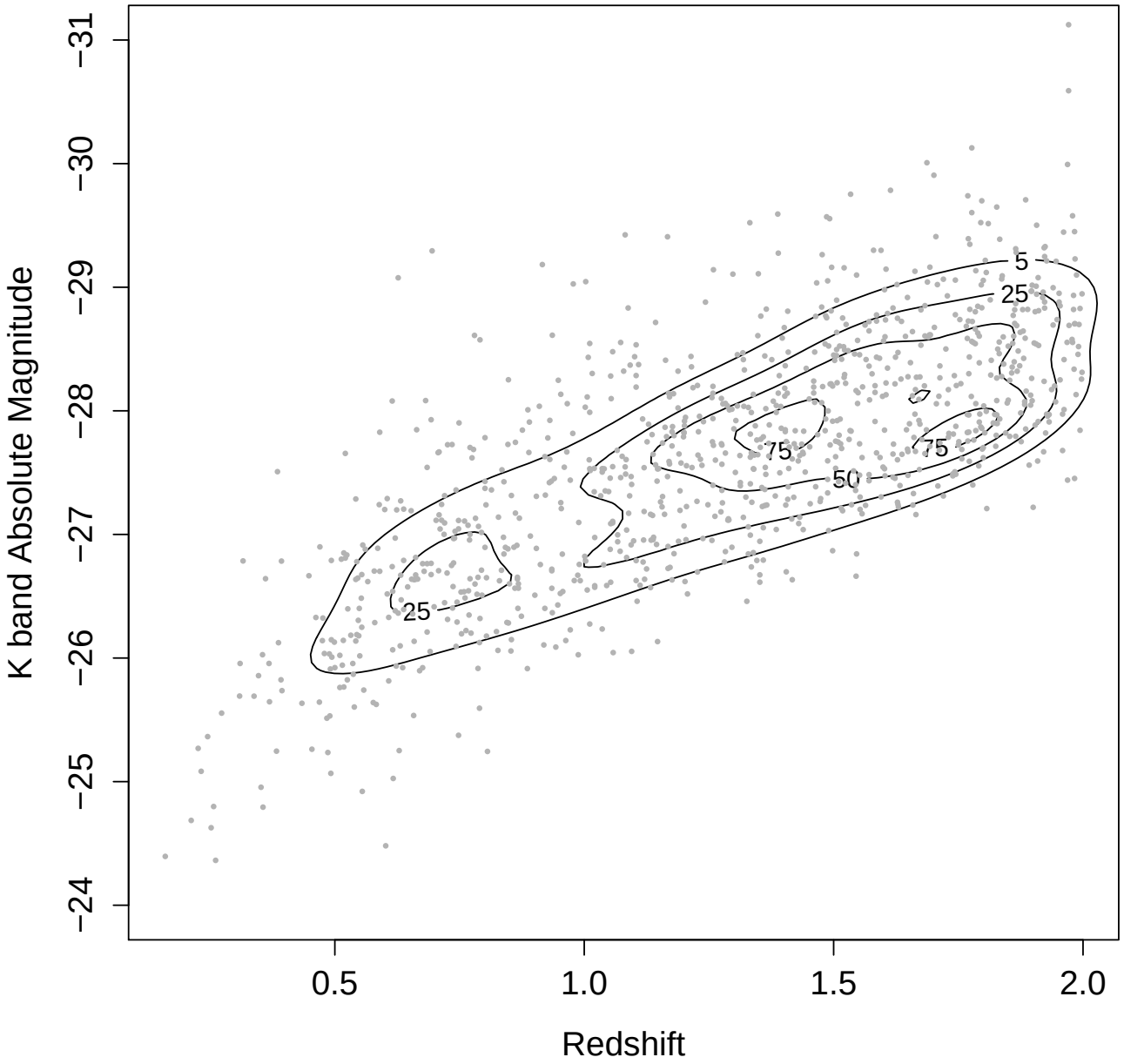
- Use a subsample to find $\mathbf{H}$
- Use the option `binned = T`
- Use `Hscv.diag()`, which forces $\mathbf{H}$ to be diagonal

The function `kde()` finds the kernel density estimate.

Multivariate Density Estimation

```
library(ks)

Hopt = Hscv(cbind(Redshift,KbandAbsMag),binned=T)

kdeout = kde(cbind(Redshift,KbandAbsMag),Hopt)

plot(kdeout, xlab="Redshift",
   ylab="K band Absolute Magnitude",
   ylim=c(-24,-31),cont=c(5,25,50,75))

points(Redshift, KbandAbsMag, pch=16, cex=0.5,
   col=rgb(0.7,0.7,0.7))
```

Multivariate Density Estimation



1,000 quasars from Peth, et al. (2011) catalog.

Curse of Dimensionality

For kernel density estimator in d dimensions,

$$\text{MISE} \sim \frac{C_4}{n^{4/(4+d)}}$$

So, to achieve same MISE in d dimensions as you would have with 150 observations in one dimension, need $\exp(4+d)$ observations.

d	1	3	5	7	9
$\exp(4+d)$	150	1,100	8,100	60,000	450,000

Curse of Dimensionality

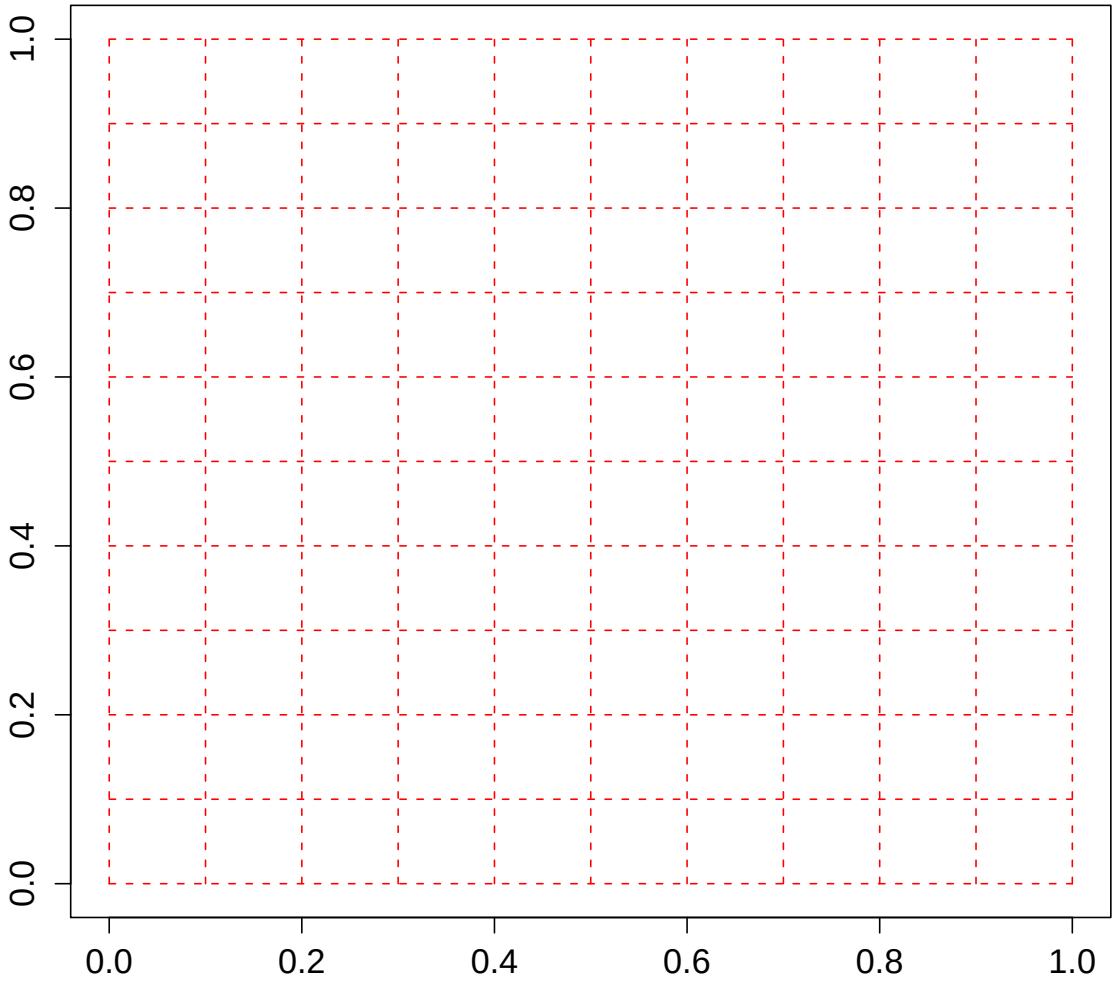
What to do?

Assume multivariate parametric form, then MISE $\sim C_1/n$, but this is even less realistic in high dimensions.

Discard dimensions.

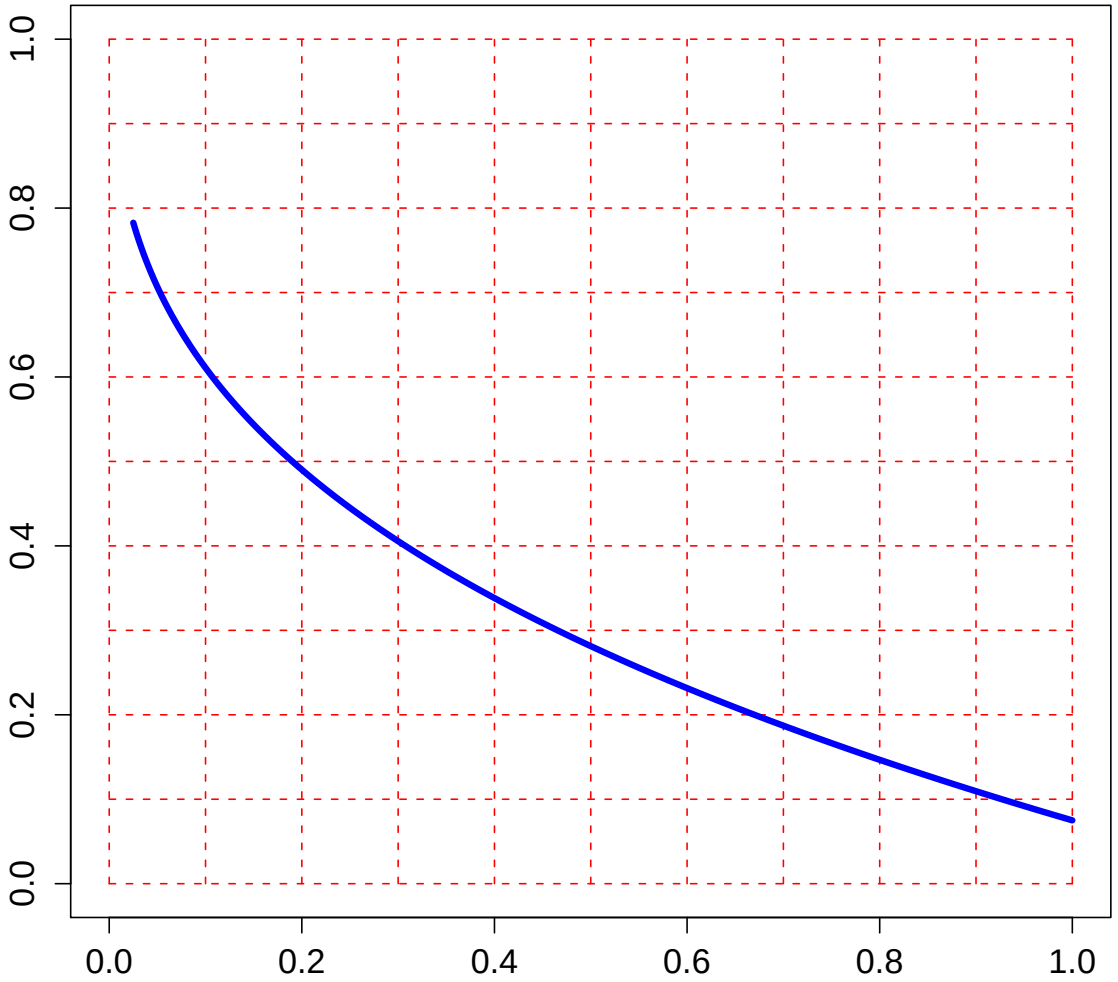
Reparametrize to low-dimensional space in a way which preserves main variations in physical system.

Curse of Dimensionality



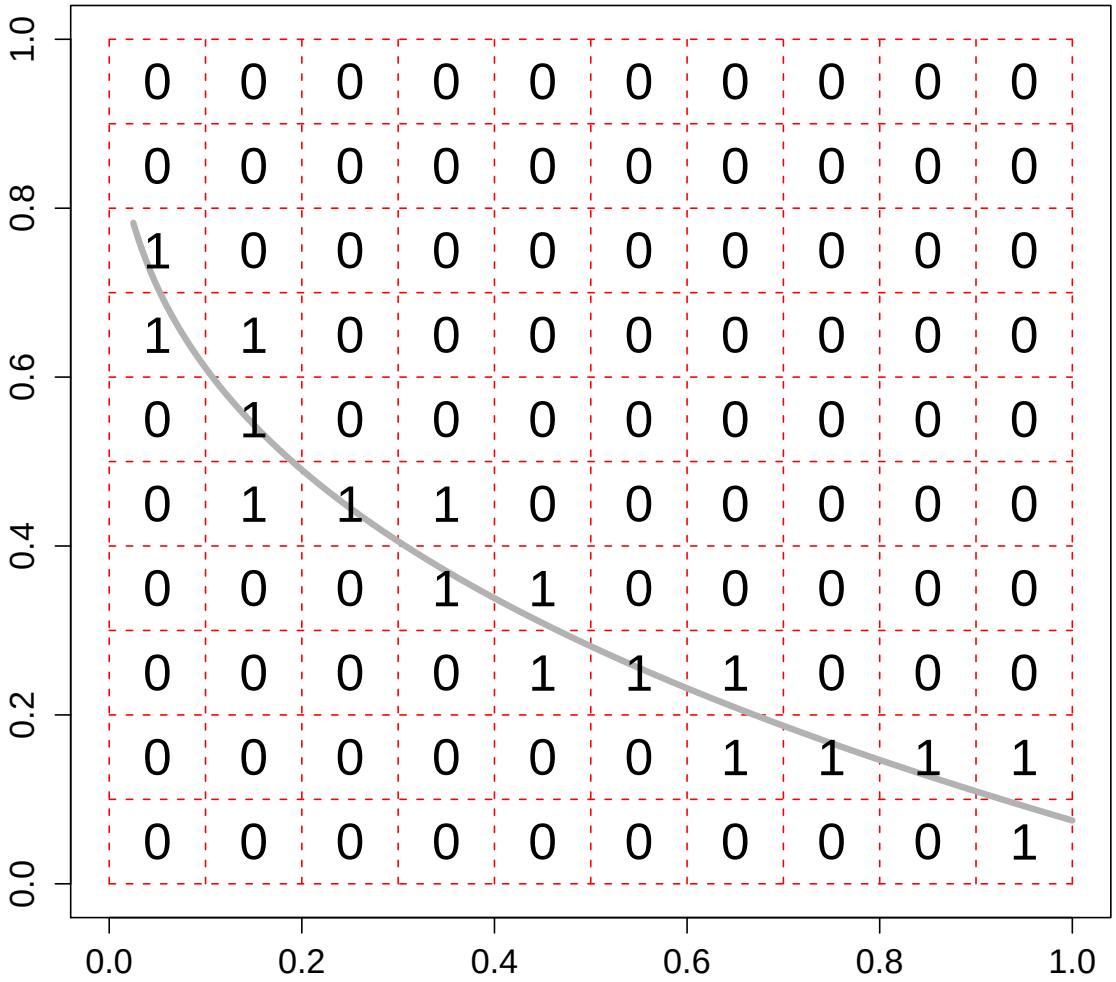
Imagine an m by m grid.

Curse of Dimensionality



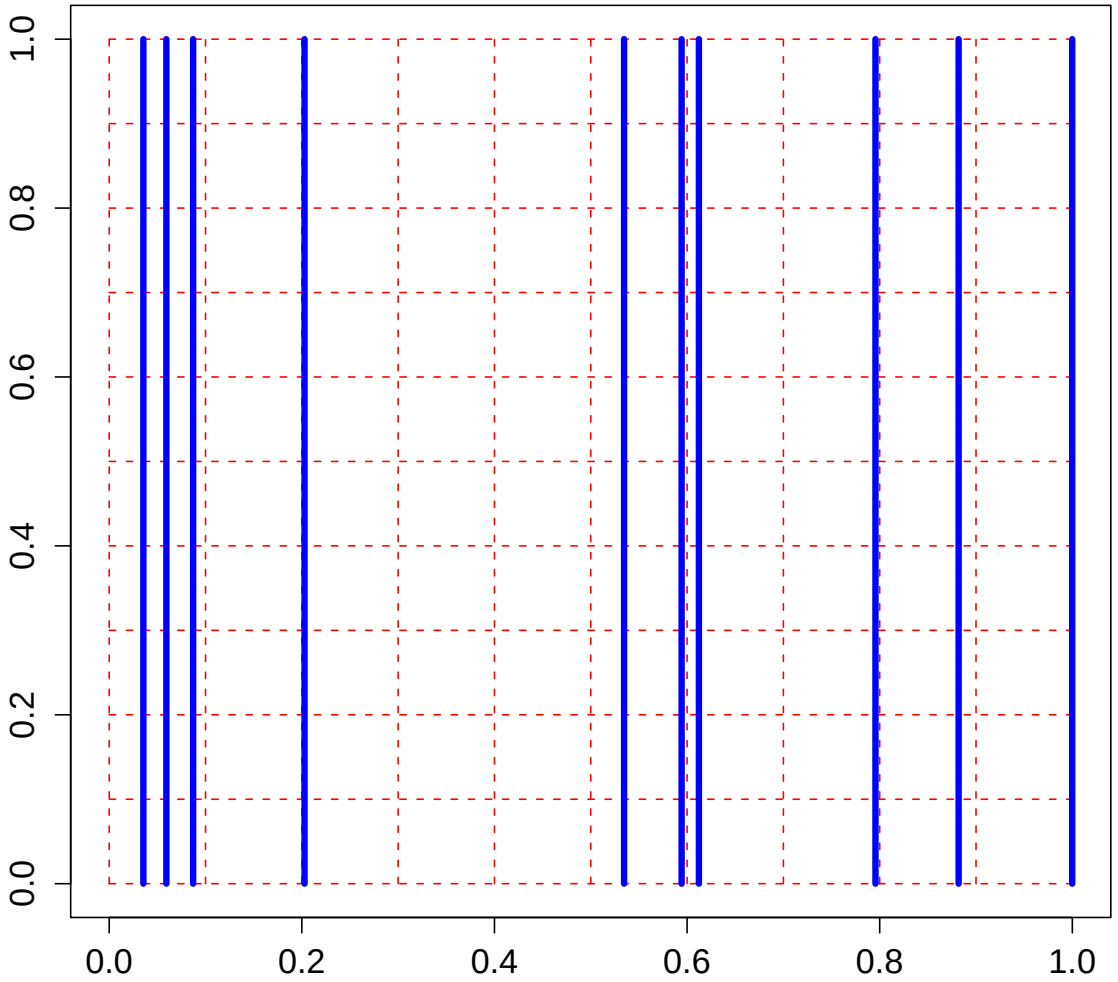
With a path.

Curse of Dimensionality



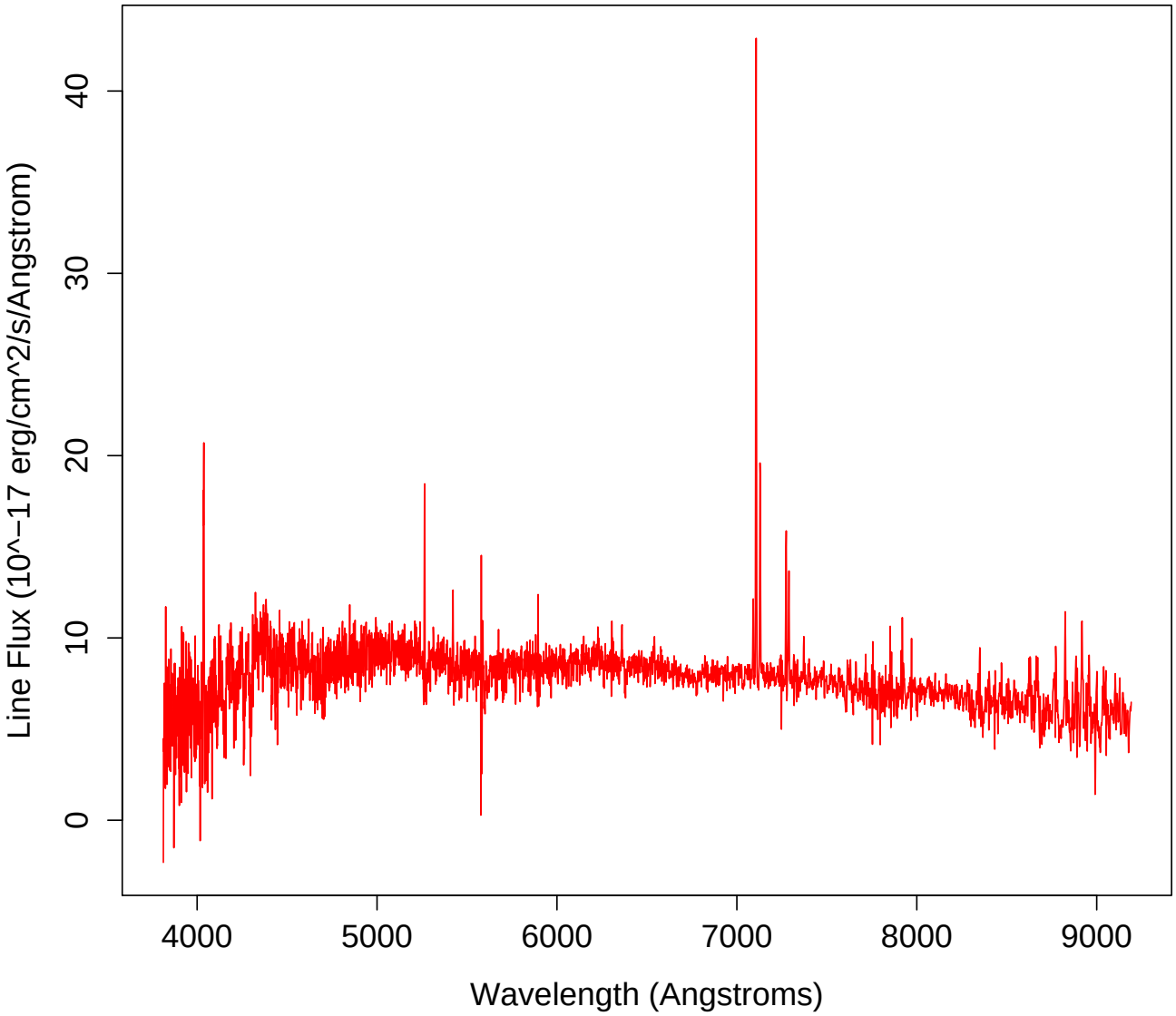
Converted to a m^2 vector.

Curse of Dimensionality



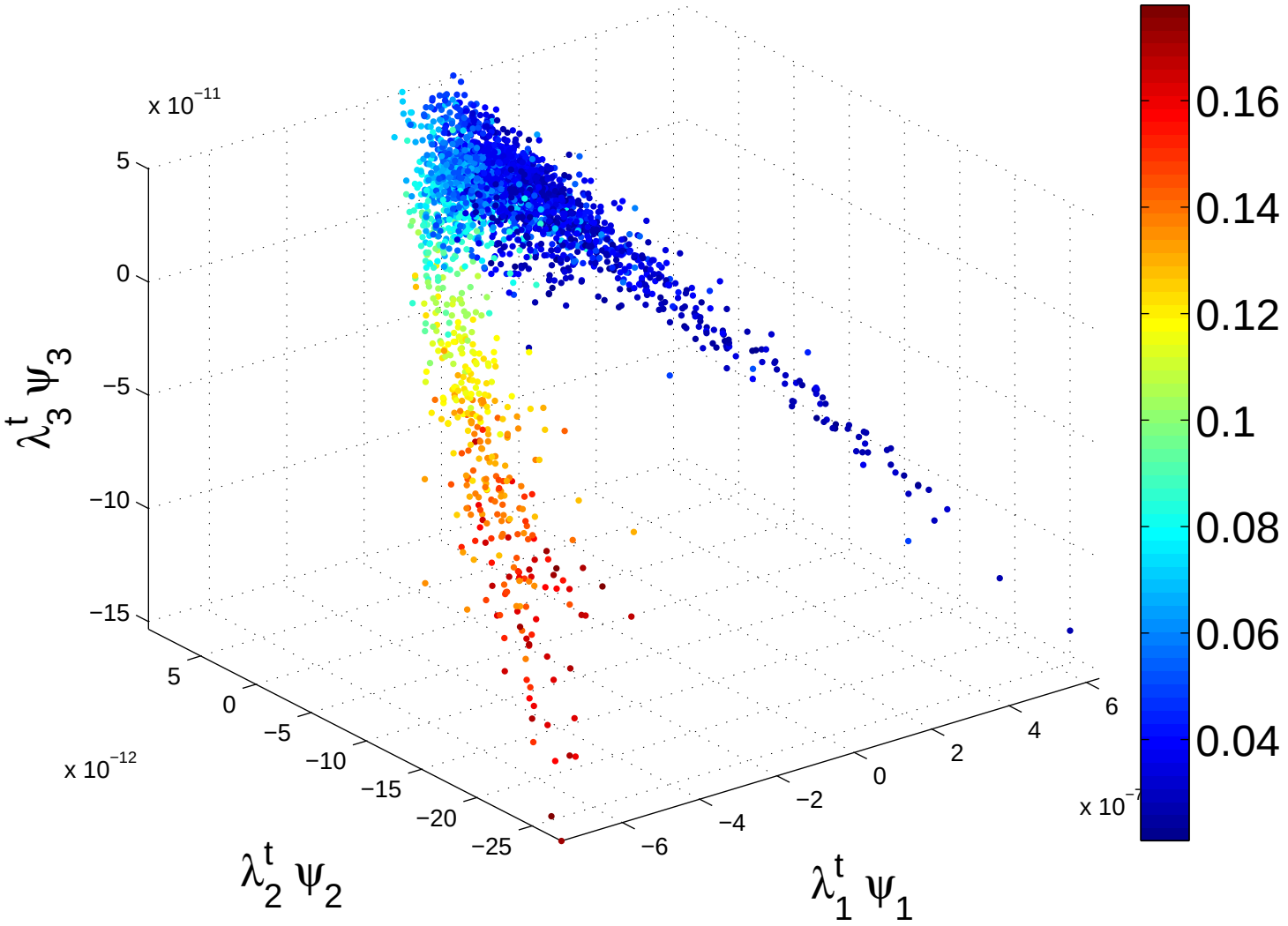
Uniform over the grid, but only one-dimensional.

High-Dimensional Data



An SDSS galaxy spectrum.

Low-Dimensional Embedding



3,846 galaxy spectra, colored by redshift (Richards, Freeman, Lee, Schafer (2009))

Conclusion

The need to move beyond the standard, parametric assumptions

"Classic" nonparametric tools

Nonparametric density estimation via kernel density estimation

MISE gains even at moderate sample sizes

Curse of Dimensionality

References

Binney and Merrifield (1998). *Galactic Astronomy*. Princeton University Press.

Blanton (2003). *ApJ*. **592**, 819-838.

Peth (2011). *AJ*. **141**. 105.

Press and Schechter (1974). *ApJ* **187**, 425-438.

Richards, Freeman, Lee, and Schafer (2009a). *ApJ*. **691** 32-42.

Schechter (1976). *ApJ* **(203)**, 297-306.

Tegler, et al. (2012). `http://arxiv.org/pdf/1203.6888.pdf` To appear in ApJ.