Non-parametric Density Estimation

Tobias Kuhlmann, Rui Zhang

UMass Amherst

December 12, 2018

Outline

Research aim

Compare asymptotic MISE behaviour for kernel and logspline density estimators as $n \to \infty$ in a Monte Carlo experiment.

Data

For our study, we use simulated univariate data

$$\{y_i\}_{i=1}^{n}, i \in \{1, ..., n\},$$

where $y_i \sim iid$ and a known smooth density f(x), i.e., $y_i \sim_{iid} f(x)$, where $x \in R$.

Kernel density estimation

We use a univariate kernel density function following Wand and Jones (1995). A density function can be estimated by

$$\hat{f}(x; h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\},$$

where K is a kernel function satisfying $\int K(x)dx = 1$ and h is the bandwidth.

Asymptotic MISE approximations

$$MISE\{\hat{f}(\cdot;h)\} = E\int\{\hat{f}(x;h) - f(x)\}^2 dx$$

h_{MISE} is the minimiser of $MISE\{\hat{f}(\cdot;h)\}$ then

$$h_{MISE} \sim \left[\frac{R(K)}{\mu(K)^2 R(f'')n}\right]^{\frac{1}{5}} = C_1 n^{-\frac{1}{5}}$$

$$\inf \mathrm{MISE}_{h>0}\{\hat{f}(;h)\} \sim \frac{5}{4}\{\mu_2(K)^2 R(K)^4 R(f'')\}^{\frac{1}{5}} n^{-\frac{4}{5}} = C_2 n^{-\frac{4}{5}}$$

These expressions give the rate of convergence of the MISE-optimal bandwidth and the minimum MISE to zero as $n \to \infty$. The best obtainable rate of convergence of the MISE of the kernel estimator is of $O(n^{-4/5})$.

Asymptotic MISE approximations

Asymptotic MISE approximations can also be used to make comparisons of the kernel estimator to the histogram.
Let b be the binwidth of the histogram $\hat{f}_H(\cdot; b)$:

$$b_{MISE} \sim \{6/R(f')\}^{\frac{1}{3}} n^{-\frac{1}{3}}$$

$$\inf \text{MISE}_{b>0}\{\hat{f}(\cdot; b)\} \sim \frac{1}{4}\{36R(f')\}^{\frac{1}{3}} n^{-\frac{2}{3}}$$

$$\text{MISE} = C_3 n^{-\frac{2}{3}}$$

$$\log(MISE) = -\frac{2}{3}\log(C_3 n)$$

Thus, the MISE of the histogram is asymptotically inferior to the kernel density estimator since its convergence rate is $O(n^{-2/3})$ compared to the kernel estimator's $O(n^{-4/5})$ rate.

Logspline density estimation

Let B be a set of basis functions. β be a collection of feasible column vectors . A column vector β is said to be feasible if $\int_L^U exp(\beta_1 B_1(x) + \cdots + \beta_J B_J(x))dx < \infty$.
Given $\beta \in B$, set

$$f(x;\beta) = exp(\beta_1 B_1(x) + \cdots + \beta_J B_J(x) - C(\beta)), L < x < U$$

where

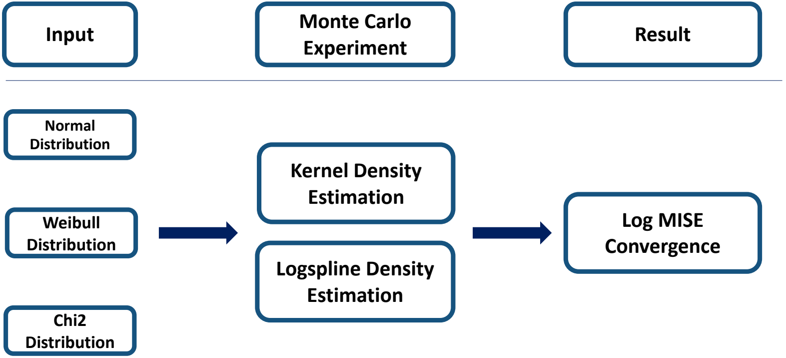
$$C(\beta) = log(\int_L^U exp(\beta_1 B_1(x) + \cdots + \beta_J B_J(x))dx).$$

Then $f(y;\beta)$ is a positive density function on (L,U), and $\int_R f(x;\beta)dx = 1$.

Logspline's advantages

- As one of the penalized approaches, logspline uses a maximum likelihood approach.
- Adds knots in those parts of the density where they are most needed.
- Has a natural way to estimate densities with bounded support.
- Avoids spurious bumps and gives smooth estimates in the tail of the distribution.
- Can estimate the density even when some observations are censored.

Simulation study



1

Density estimation in R

Kernel density estimation
Matt Wand (2013). KernSmooth: Functions for kernel smoothing for Wand & Jones (1995)

```
# Univariate kernel density estimator from KernSmooth package (Wand (1995))
h <- dpik(y) # select optimal bandwidth
fit <- bkde(x=y, bandwidth=h, gridsize = 401) # kde
```

Logspline density estimation
Charles Kooperberg (2005). Logspline: Logspline Density estimation routines

```
# Logspline density estimator
fit <- logspline(y) # fit logspline
dens <- dlogspline(q=x, fit=fit) # get density values
```

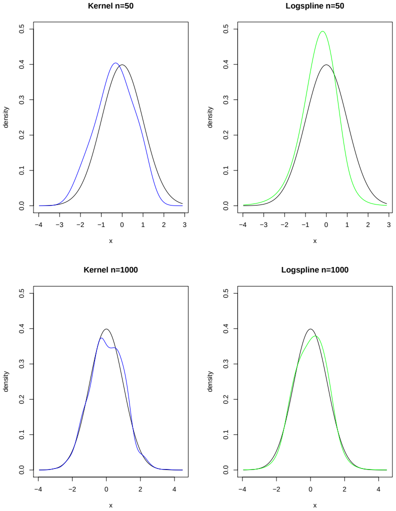
Monte Carlo experiment

Same approach as in September 12th class:

- For 20 sample sizes from 100 to 100000
  - For 10 different random samples
    - Kernel density estimation
    - Logspline density estimation

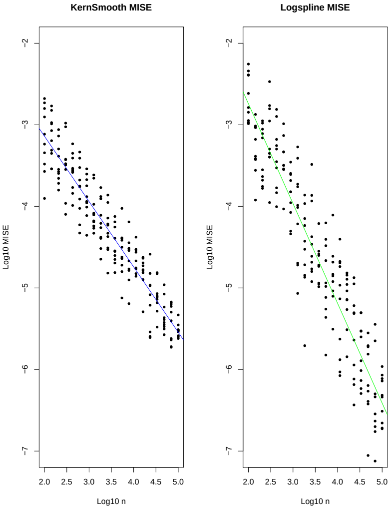
Normal distribution

N(0,1) with density estimations



Normal distribution

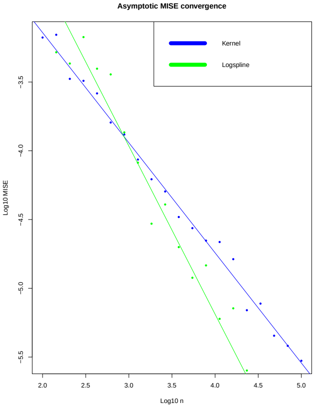
MISE Convergence rates



Normal distribution

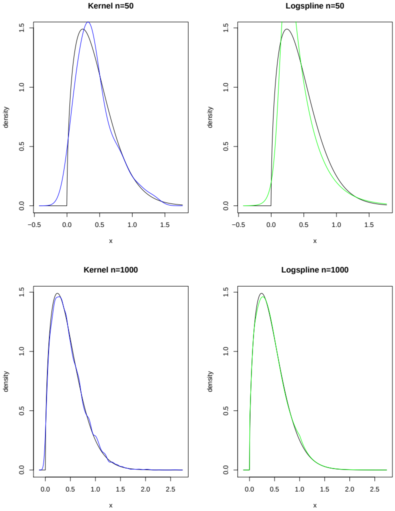
Table 1: Log MISE convergence regression results

Type	Slope estimate	95% CI
Kernel	-0.80	(-0.84,-0.76)
Logspline	-1.22	(-1.29,-1.15)



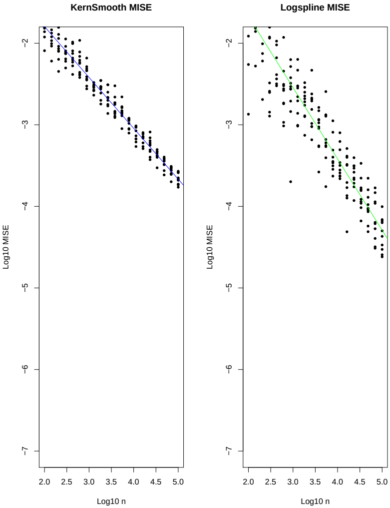
Weibull distribution

Weibull(0, 1.5, 0.5) with density estimations



Weibull distribution

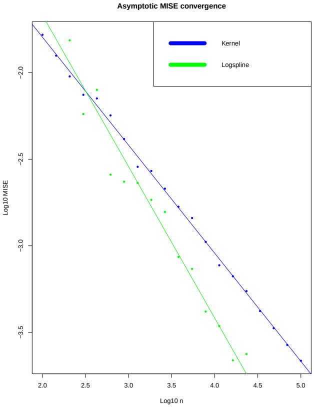
MISE Convergence rates



Weibull distribution

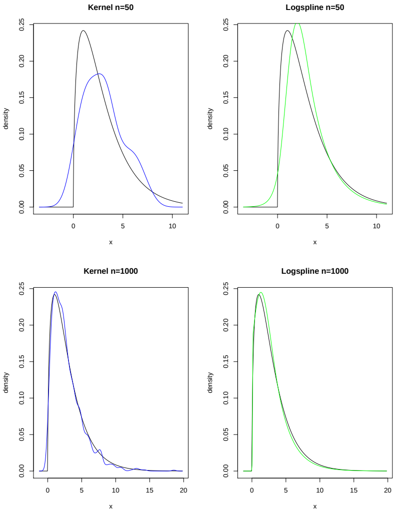
Table 2: Log MISE convergence regression results

Type	Slope estimate	95% CI
Kernel	-0.62	(-0.64,-0.61)
Logspline	-0.87	(-0.92,-0.82)



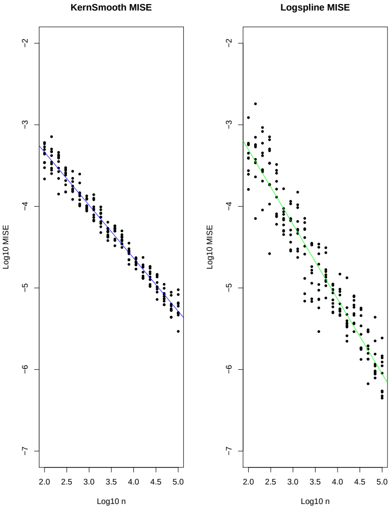
Chi squared distribution

Chisquared(3) with density estimations



Chi squared distribution

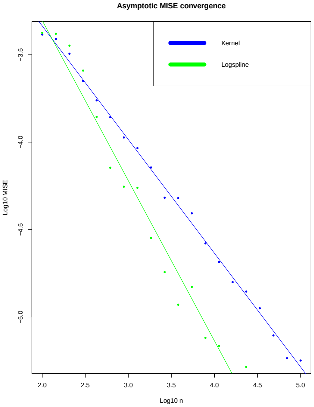
MISE Convergence rates



Chi squared distribution

Table 3: Log MISE convergence regression results

Type	Slope estimate	95% CI
Kernel	-0.65	(-0.66,-0.63)
Logspline	-0.91	(-0.96,-0.87)



Conclusions

- Low sample density estimates are highly inaccurate.
- Logspline MISE converges faster to zero than kernel density estimation in all three experiments.
- Be careful with bounds, ensuring density is smooth.

Open Questions

- Theoretical derivation of asymptotic logspline MISE.
- Does logspline log mise asymptotic behavior linear?
- Why are convergence rates different?
- Think about and try distribution bounds.

References

- Stone, Hansen, Kooperberg, and Truong, Polynomial Splines and their Tensor Products in Extended Linear Modeling, Annals of Statistics, Volume 25,Issue 4(Aug., 1997), 1371-1425.
- MP. Wand and M.C.Jones, Kernel Smoothing, Chapman& Hall, 1995.