2002

Root-Unroot Methods for Nonparametric Density Estimation and Poisson Random-Effects Models

Lwawrence D. Brown
University of Pennsylvania

Ren Zhang

Linda H. Zhao
University of Pennsylvania

Root-Unroot Methods for Nonparametric Density Estimation and Poisson Random-Effects Models

Disciplines

Statistics and Probability

Root-Unroot Methods for Nonparametric Density Estimation And Poisson random-effects models[1]

Lawrence D. Brown
University of Pennsylvania

Statistics Department
Philadelphia, PA 19104-6302
(lbrown@wharton.upenn.edu)

Joint work with:
Ren Zhang
Linda Zhao

1. For departmental Statistics seminar 1/16/2002. Revision and expansion of invited talk delivered at the Joint Statistical Meetings in Atlanta, GA, August, 2001

Outline of Presentation

0. Preamble (motivation).

1. Description of the "root-unroot" method for density estimation. (Really, a "methodette.)

2. Motivation via Poissonization and variance stabilization.

3. Properties of the root-unroot step.

4. Some simulation pictures.

5. Comments about signal-to-noise ratio.

6. An empirical example from a telephone call-in service center.

---

7. A two-way random-effects analysis of call arrival rates.

Motivation

- Nonparametric regression and nonparametric density estimation are siblings. (Maybe even non-identical twins.)

- But I like better working with the regression formulation. (I feel a better sense of rapport and understanding!)

- So my goal is to convert a density problem into a regression one; so I can work with it as such.

Description of the Methodette

1. **DATA**: $\{X_1, \ldots, X_n\}$ i.i.d. from a density f on [0, 1].

2. **BIN**: Create K *(equal width)* bins. Let

   $\mathbf{N_k}$ = # of $\{X_i\}$ in the kth bin, $\mathbf{Bin_k}$
   $\mathbf{T_k}$ = Center of $\mathbf{Bin_k}$.

3. **"ROOT":** Calculate

$$\mathbf{Y_k} = \sqrt{\frac{\mathbf{K}}{\mathbf{n}}} \times \sqrt{\mathbf{N_k} + \mathbf{1/4}}\,.$$

4. **ESTIMATE**: Apply your favorite nonparametric regression estimator to $\{T_k, Y_k\}$. This produces an estimate:

$$\hat{\mathbf{h}}\mathbf{(t)} \text{ of } \mathbf{h(t)} \approx \sqrt{\mathbf{f(t)}}\,.$$

Since $\mathbf{E(Y_k)} \approx \sqrt{\mathbf{f(t_k)}} = \mathbf{h(t_k)}$.

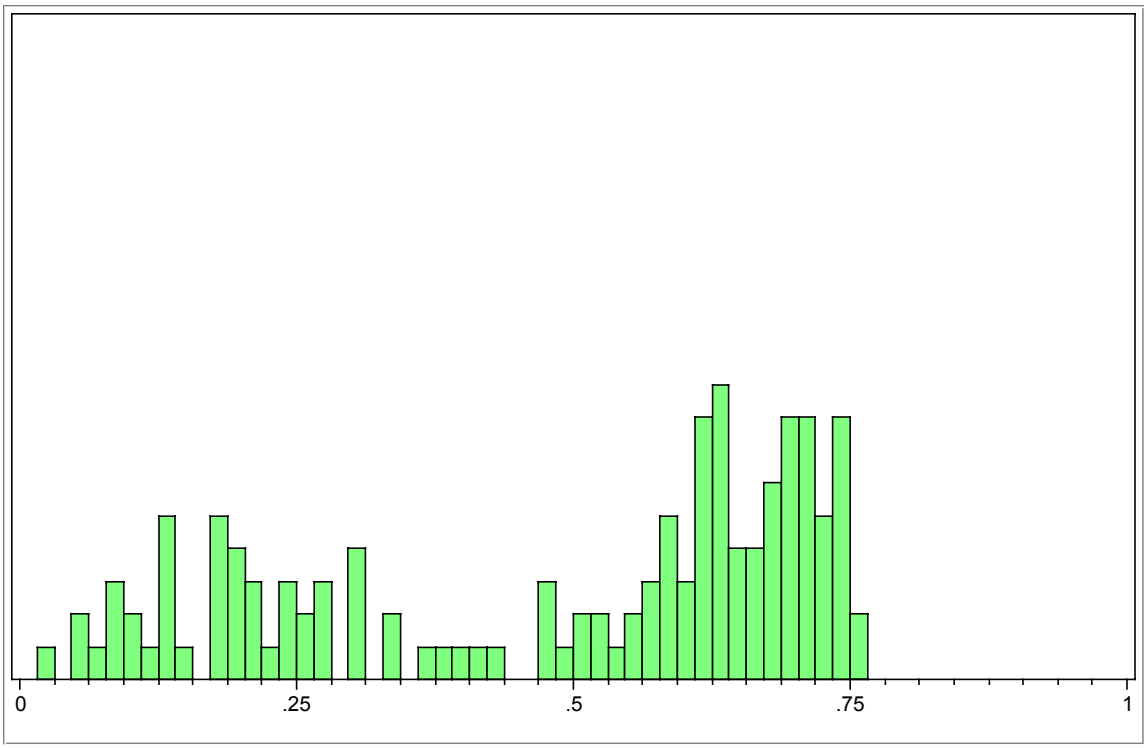
If helpful, capitalize on $\mathrm{Var}(Y_k) \approx K/4n$ for all k.

5. **"UN-ROOT"**: Calculate $\left[\hat{\mathbf{h}}(\mathbf{t})\right]^2$ and re-normalize to be a density. For equal width bins this gives

$$\hat{\mathbf{f}}(\mathbf{t_k}) = \frac{(\hat{\mathbf{h}}(\mathbf{t_k})_+)^2}{\sum_{\mathbf{j=1}}^{\mathbf{K}} (\hat{\mathbf{h}}(\mathbf{t_j})_+)^2} \times \mathbf{K}.$$

Data Example

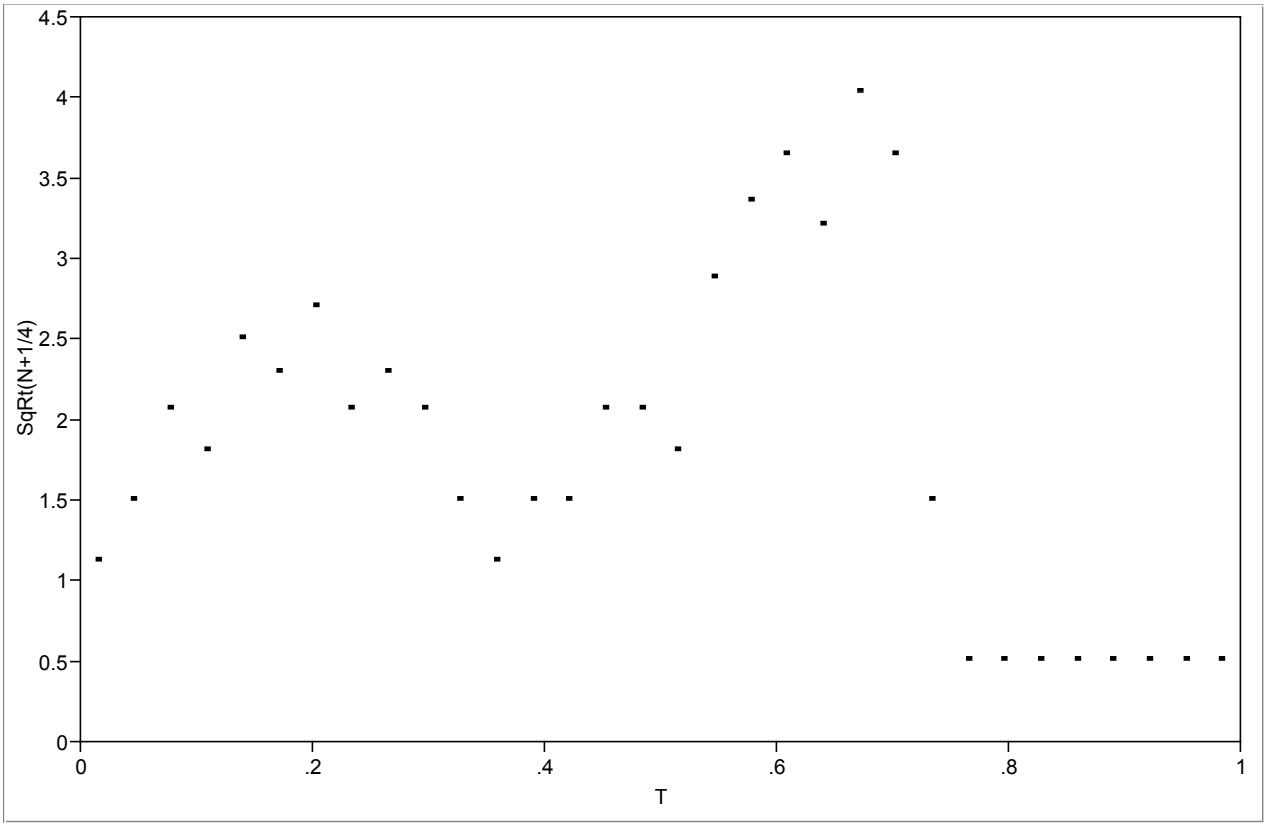
1. DATA:



Histogram of Data: n = 128

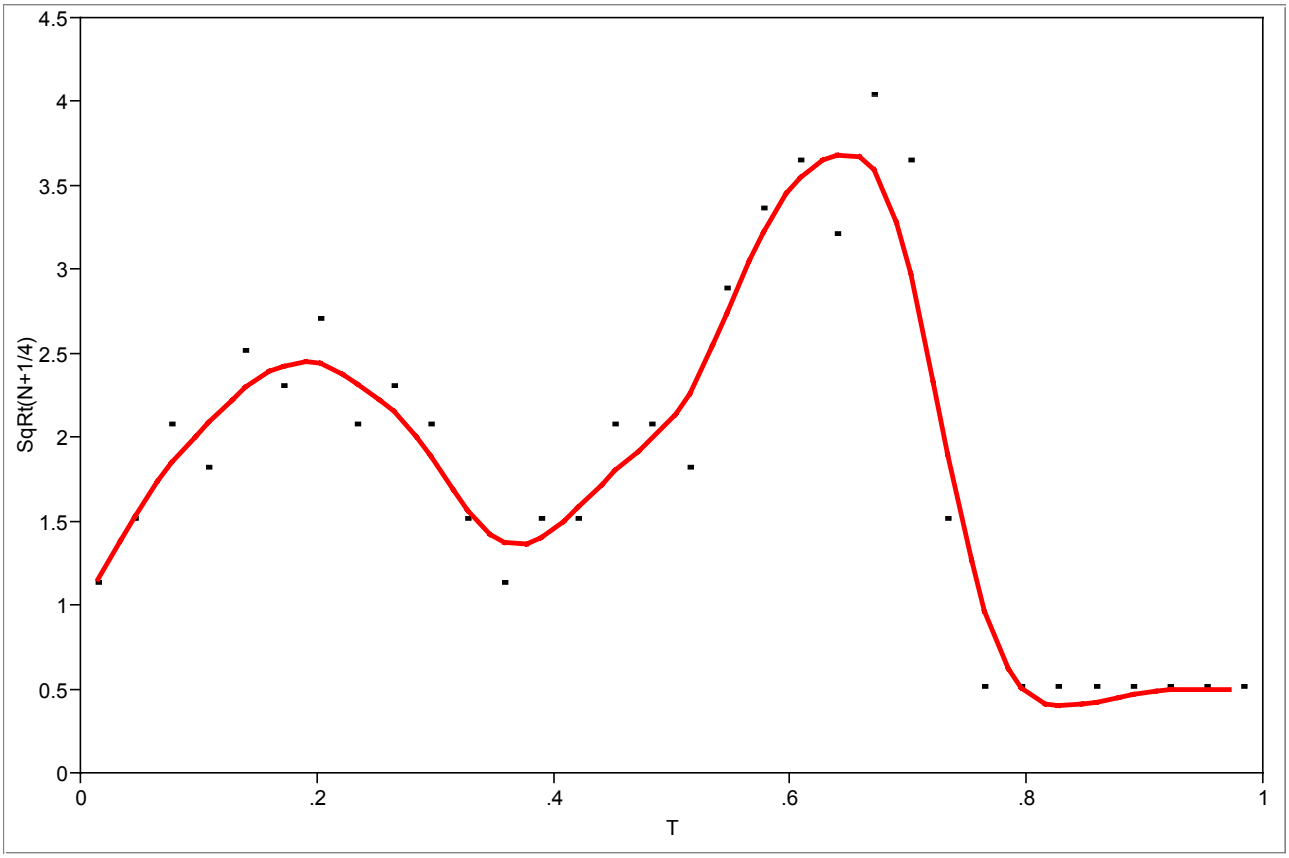
- It's accidental that n is a power of 2; this is irrelevant to the methodette.
- This data is a random sample from the "two-humps" density in Wahba (1983).

2 & 3. BIN AND ROOT:



Scatterplot of SqRt(N_k+1/4) versus T_k

4. ESTIMATE:



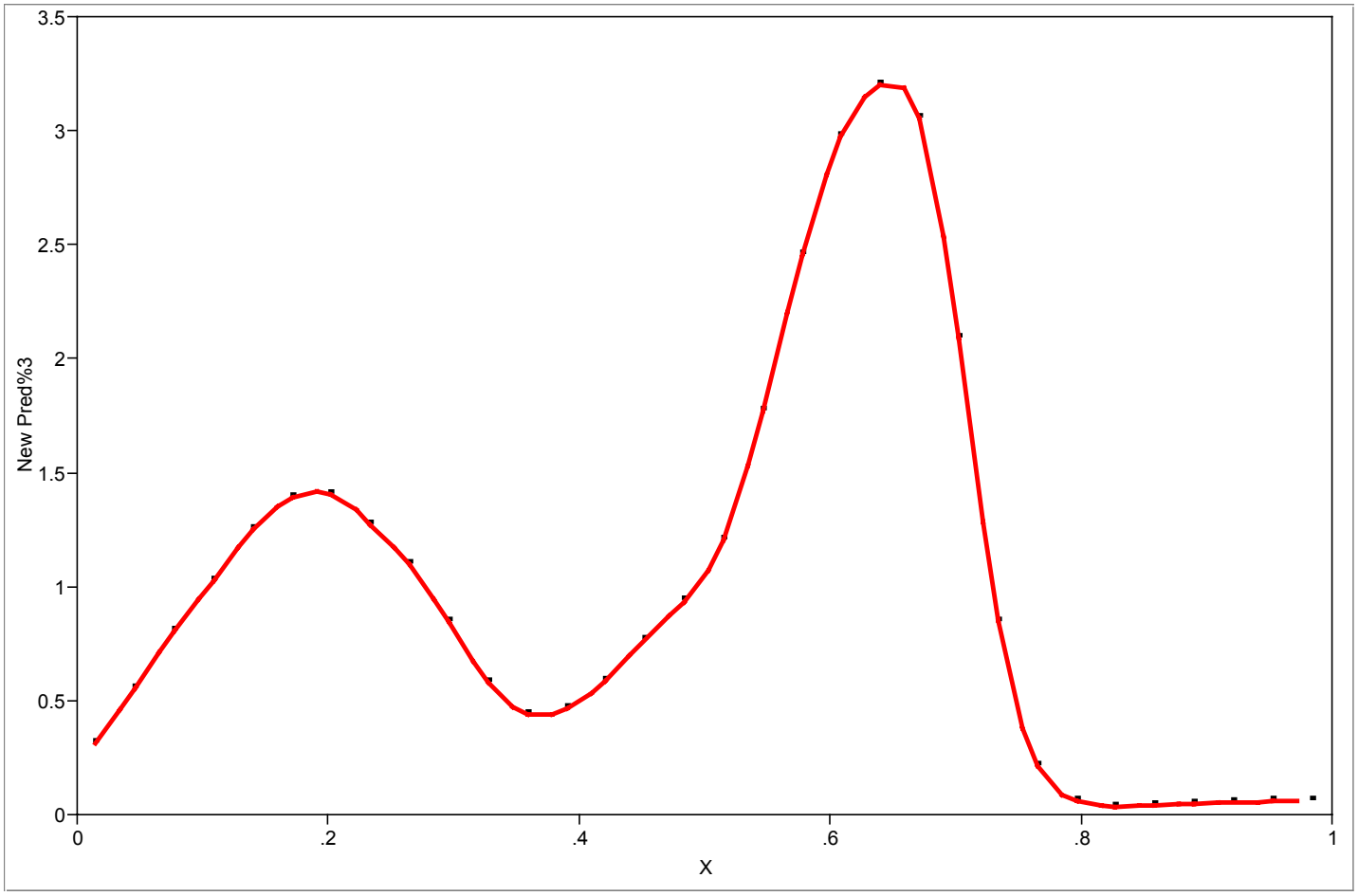
Smoothing Spline Fit

Smoothing Spline Fit, lambda=0.00005

R-Square	0.94
Sum of Squares Error	2.01

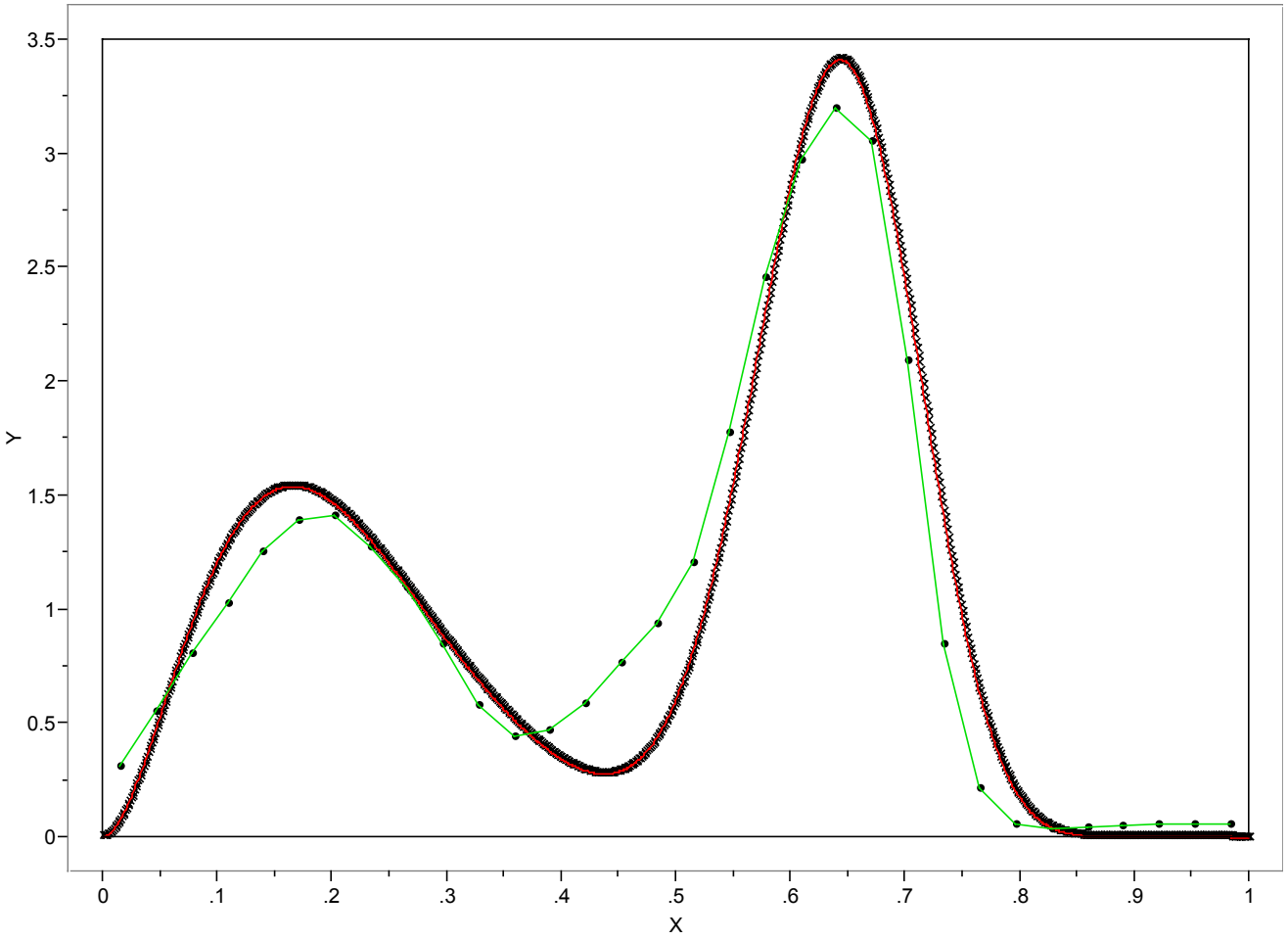
This "fit" has about 8 df for error ($8 \approx \frac{2.01}{1/4}$); and consequently about 24 df for the model.

6. UN-ROOT:



Estimate of f(x)

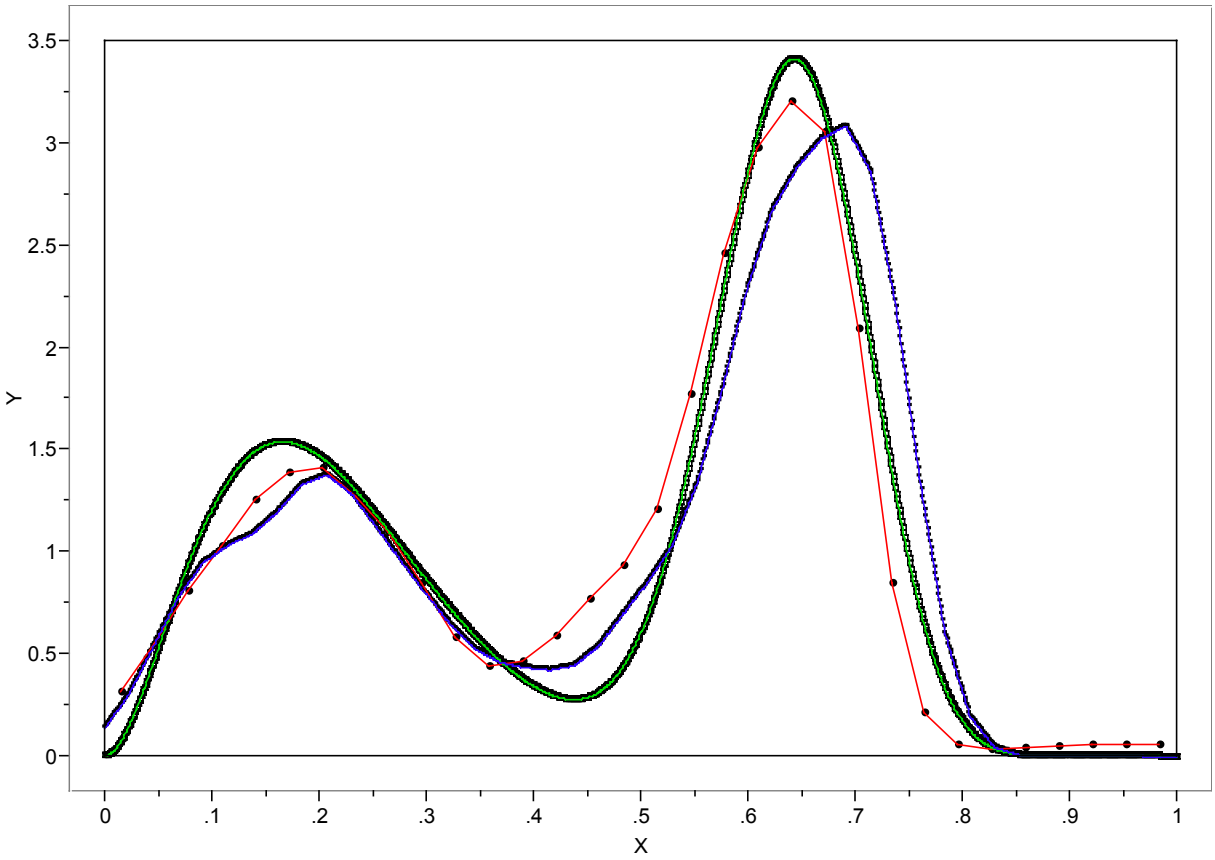
How Well Did We Do? (1)



Plot of Estimate and True Density

True density = thick curve
Estimate = narrow curve

How Well Did We Do? (2)



Comparison of Estimators

Thin curve: Root-unroot estimator

Medium curve: Kernel density estimator

Heavy curve: True density

- Both estimators capture the qualitative two-humps feature.
- Neither estimator provides a tremendously good fit. (More on this later.)
- Fair and revealing Monte Carlo comparisons are difficult to construct; the comparisons depend much more on the appropriateness of the respective density and regression estimators than on the integrity of direct density estimation versus the root-unroot regression paradigm. More, later.

Motivation

Poissonization:

- Suppose n = N, a Poisson(λ) random variable. Then

$$N_k \sim \text{Poisson}(\boldsymbol{\lambda_k})$$

with $\boldsymbol{\lambda_k = \lambda \bar{f}(t_k)}$ and $\boldsymbol{\bar{f}(t_k)}$ = Average of f over Bin_k.

These N_k are **independent**.

- We then wish to make inference about $\boldsymbol{\bar{f}(t_k) = \frac{\lambda_k}{\lambda}}$.

- $\sum N_k$ is ancillary to $\boldsymbol{\lambda_k/\lambda}$.

A natural way to proceed is to estimate $\boldsymbol{\lambda_k}$ *(nonparametrically)* by $\boldsymbol{\hat{\lambda}_k}$,

and then estimate $\boldsymbol{\bar{f}(t_k)}$ by $\boldsymbol{\frac{\hat{\lambda}_k}{\sum \hat{\lambda}_j}}$.

Variance Stabilization:

●An asymptotically unbiased,

second-order mean-stabilizing,
first-order variance stabilizing transformation
for the independent counts, N_k, is

$$\sqrt{N_k + 1/4}\,.$$

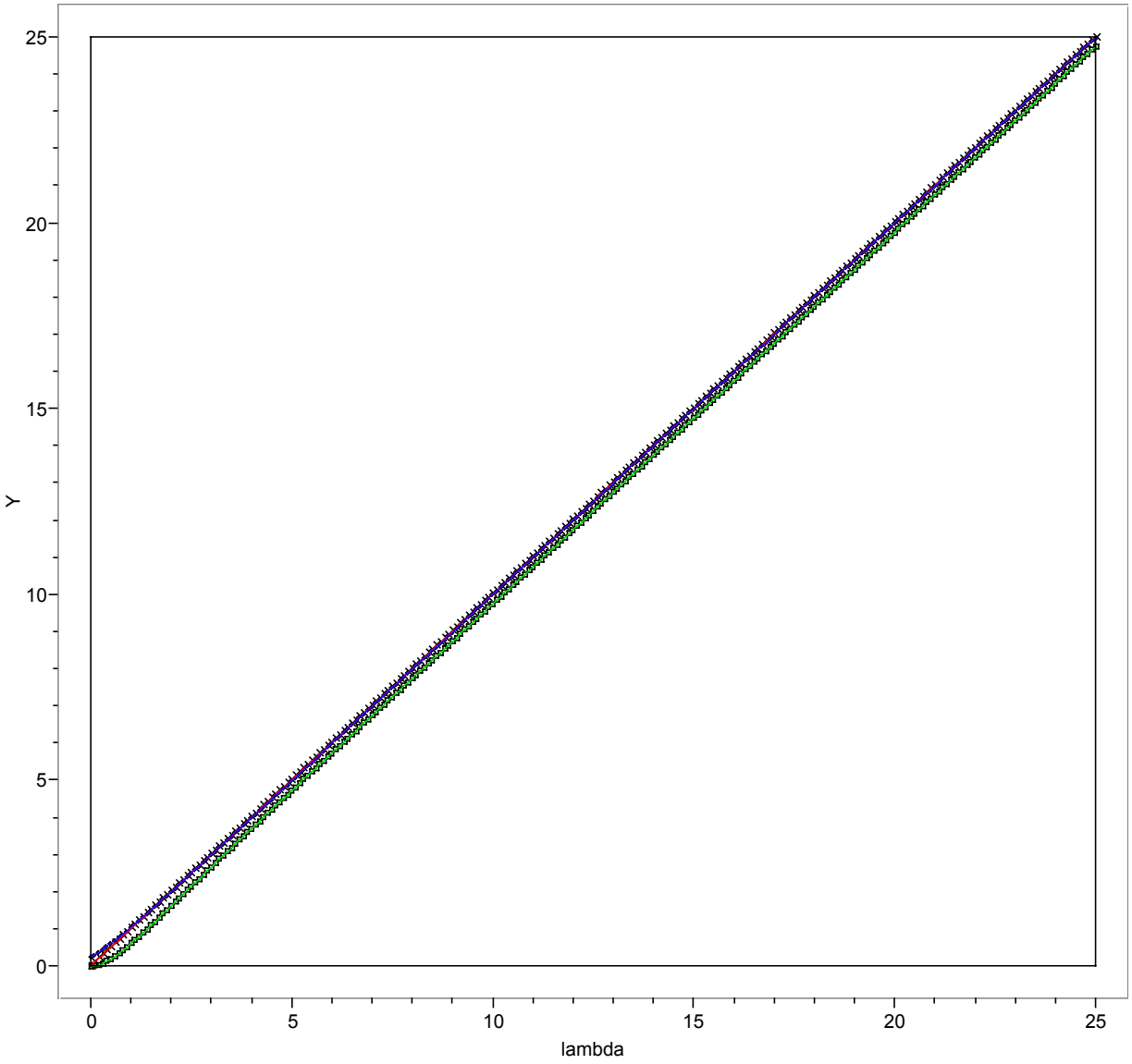
●These variables are approximately normal,

with mean = $\sqrt{\lambda_k}$, and variance = 1/4,
as well as independent.

- Variance stabilization and Poissonization each have extensive statistical pedigrees.
- Density estimation schemes closely related to the above have also been often proposed. See Fan & Gijbels (1996), and Vidakovic (1999) for two recent versions, each of which is somewhat different from the above
- Nussbaum (1996), Nussbaum and Klemela (1999), Carter (2001) and Brown, Low and Zhang(2001) all contain asymptotic equivalence theorems formally justifying (sometimes complicated) versions of the above scheme, under suitable regularity conditions.

Properties of the Root-Unroot Step

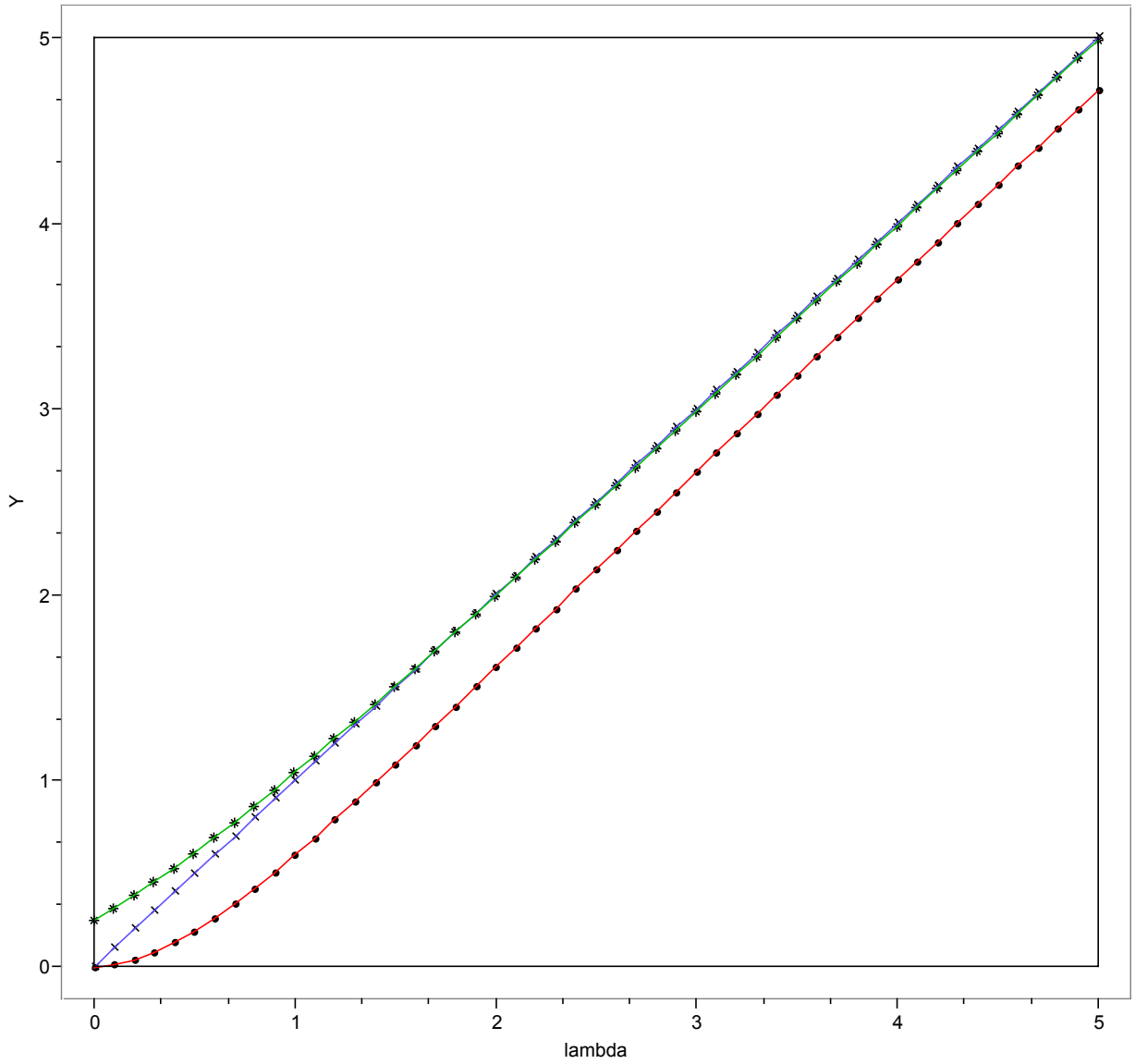
- Stabilizes the mean:



Mean of $\left\{\mathbf{E}\left(\sqrt{\mathbf{Poiss}_{\lambda}+\mathbf{c}}\right)\right\}^2$ for c = 0 and c = 1/4

Also shown is the line Y = λ.

Detail from Previous Plot

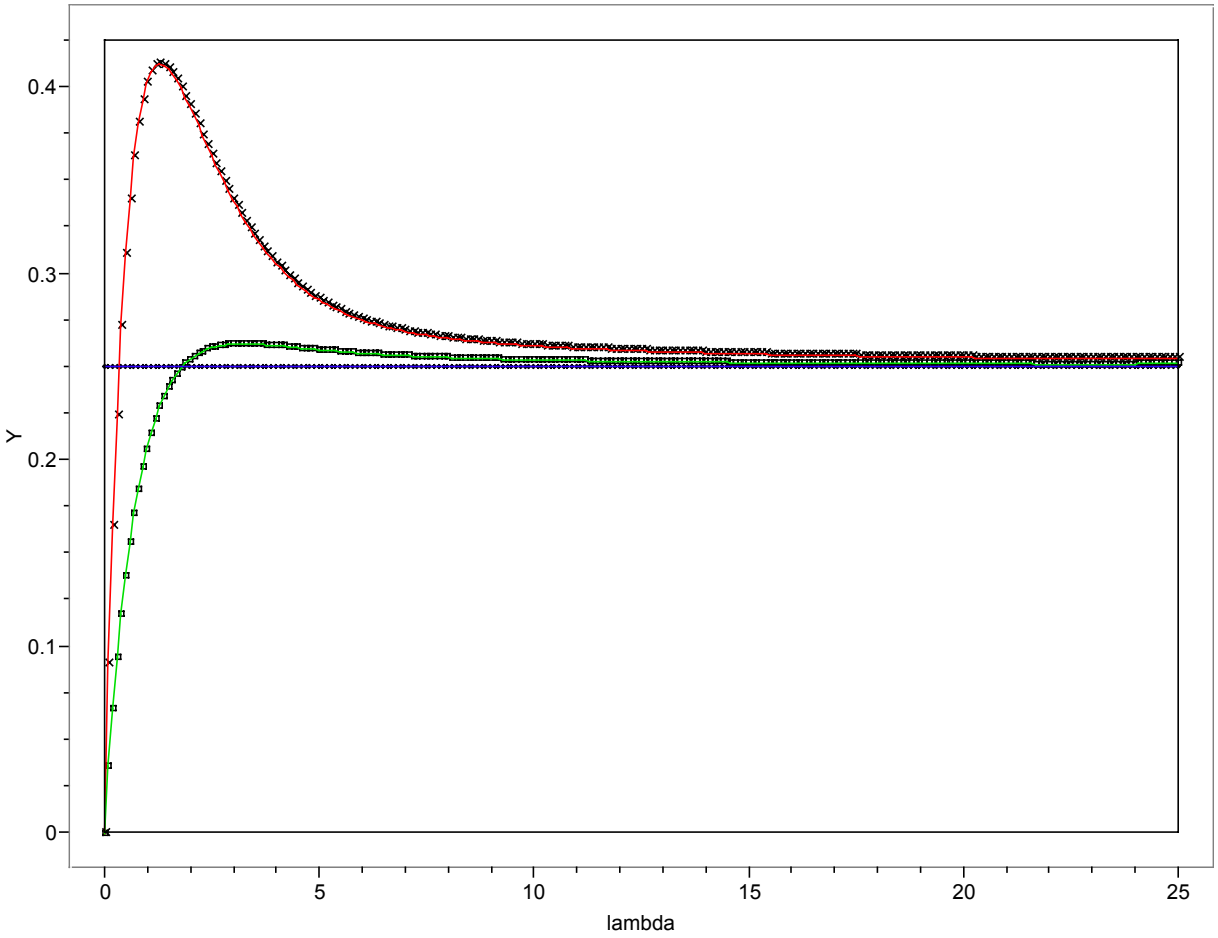


$$\mu_\lambda^2 = \left\{ \mathbf{E}\left(\sqrt{\mathbf{Poiss}_\lambda + \mathbf{c}} \right) \right\}^2 \text{ for } c = 0 \text{ and } c = 1/4$$

Also shown is the line $Y = \lambda$.

(Lowest curve is for c = 0. Highest curve is for c = 1/4.)

- **Stabilizes the variance:**



$\sigma^2_\lambda = \mathbf{Var}(\sqrt{\mathbf{Poiss}_\lambda + \mathbf{c}})$ for c = 0 and c =1/4

The nominal (and limiting) value is 1/4.

Upper curve is for c = 0;
Other curve is for c = 1/4;

<u>CONCLUSION</u>: It is suitable to use the root-unroot method with values of n/K as small as 3 or 4. As n increases it seems suitable to let n/K increase <u>slowly</u>.

Notes:

- Apparently, Bartlett (1936) was the first to propose the transformation $\mathbf{Y} = \sqrt{\mathbf{Poiss}}$ as the variance stabilizing transformation. He then proposed using it in a homoscedastic linear model.

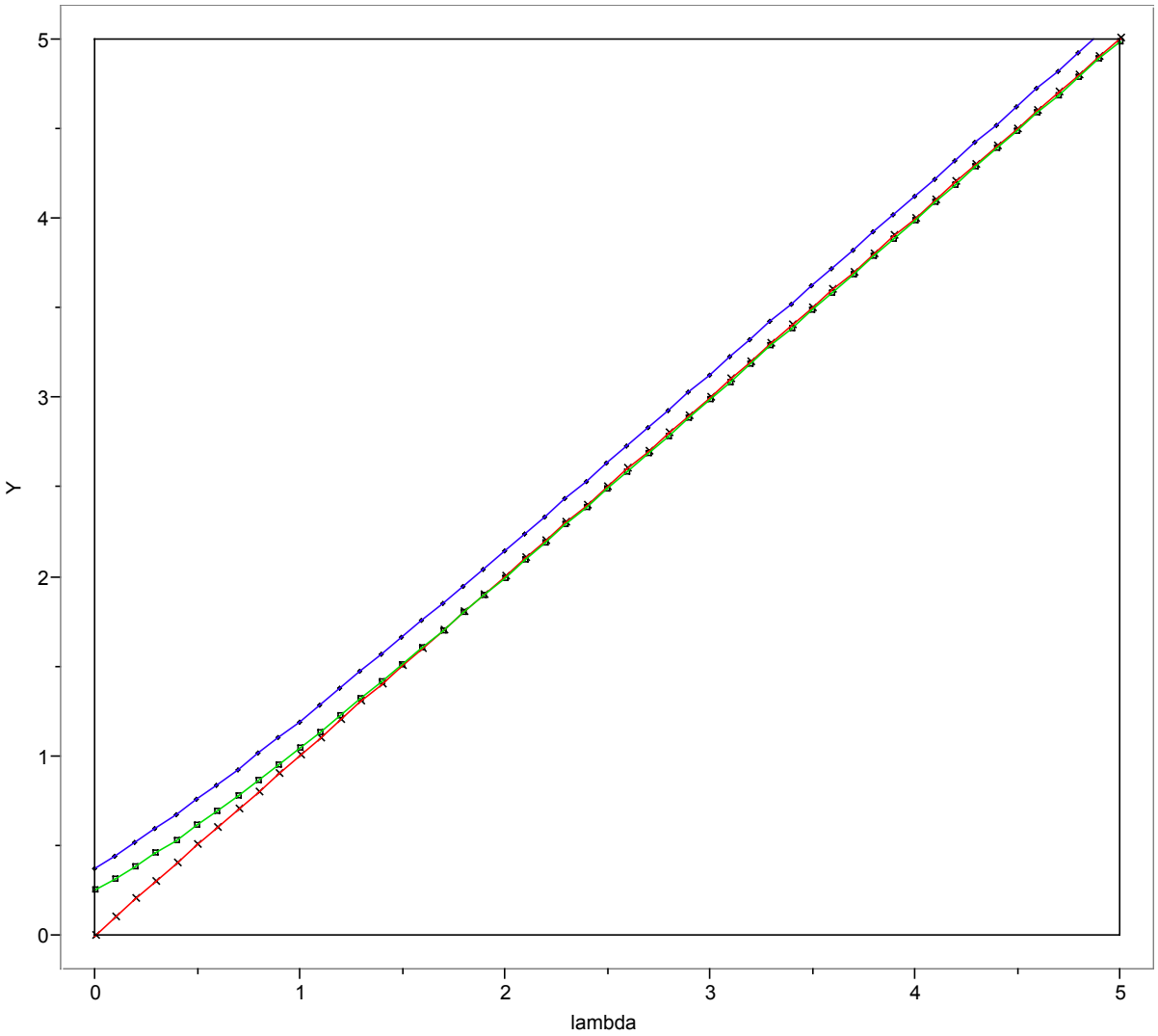
- Anscombe (1948) proposed improving the variance stabilizing properties by instead using $\mathbf{Y} = \sqrt{\mathbf{Poiss} + \mathbf{3/8}}$. (He credits this result to A. H. L. Johnson.)

- He apparently thought that this transformation was also optimal with respect to its mean. (See his equation (2.12).) In fact this transformation is slightly more biased than our $\mathbf{Y} = \sqrt{\mathbf{Poiss} + 1/4}$, as is shown by asymptotic analysis or by the plot below.

- I do not know who should be credited as the first to propose $\mathbf{Y} = \sqrt{\mathbf{Poiss} + 1/4}$ as the asymptotically unbiased, nearly variance-stabilizing transformation.

- On the grounds that bias is more important here than variance, we prefer to use $\mathbf{Y} = \sqrt{\mathbf{Poiss} + 1/4}$. This transformation has such nice properties that it doesn't seem worthwhile to try to make further improvements at the expense of complicating the formula.

Plots for Anscombe's Transformation

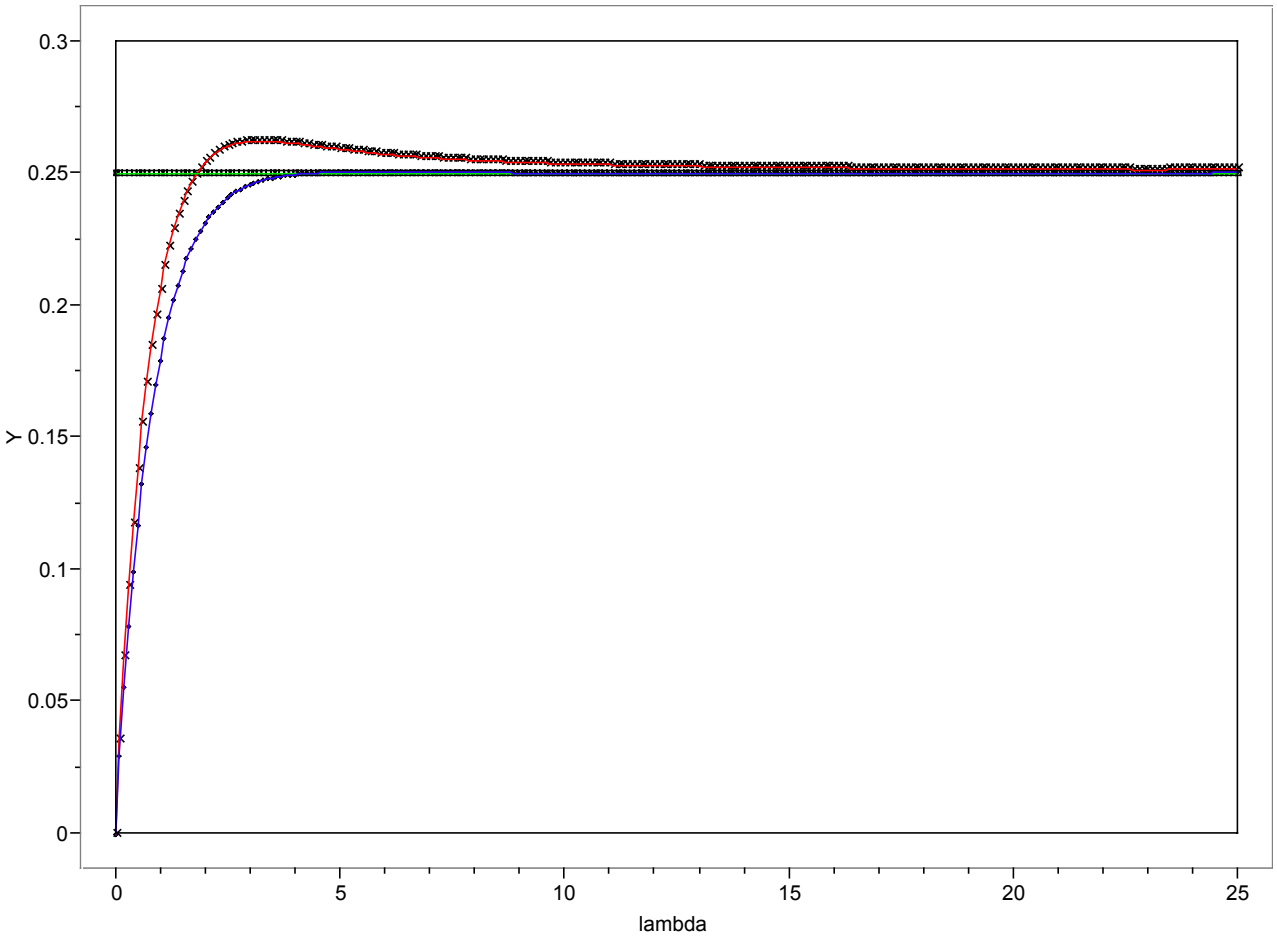


$$\mu_{\lambda}^{2} = \left\{ \mathbf{E}\left(\sqrt{\mathbf{Poiss}_{\lambda} + \mathbf{c}} \right) \right\}^{2}$$ for c = 1/4 and c = 3/8

Also shown is the line $Y = \lambda$.

Lower curve is for c = 1/4. Higher curve is for c = 3/8.

Revealed in the plot is the limiting bias of 1/8 for Anscombe's $\left\{ \mathbf{E}\left(\sqrt{\mathbf{Poiss}_{\lambda} + 3/8} \right) \right\}^{2}$.

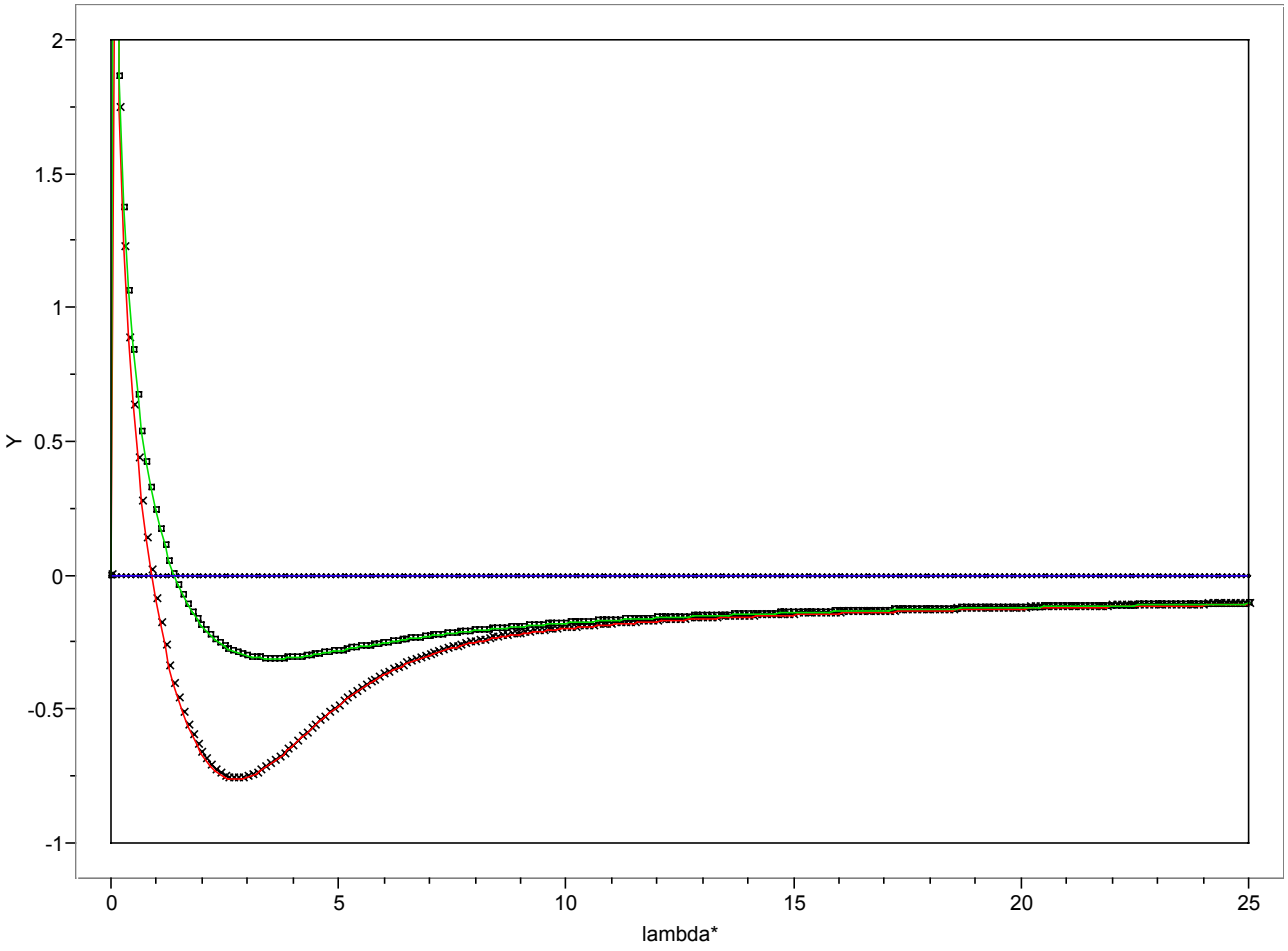


$\sigma^2_\lambda = \mathbf{Var}(\sqrt{\mathbf{Poiss}_\lambda + \mathbf{c}})$ for c = 1/4 and c = 3/8

The nominal (and limiting) value is 1/4.

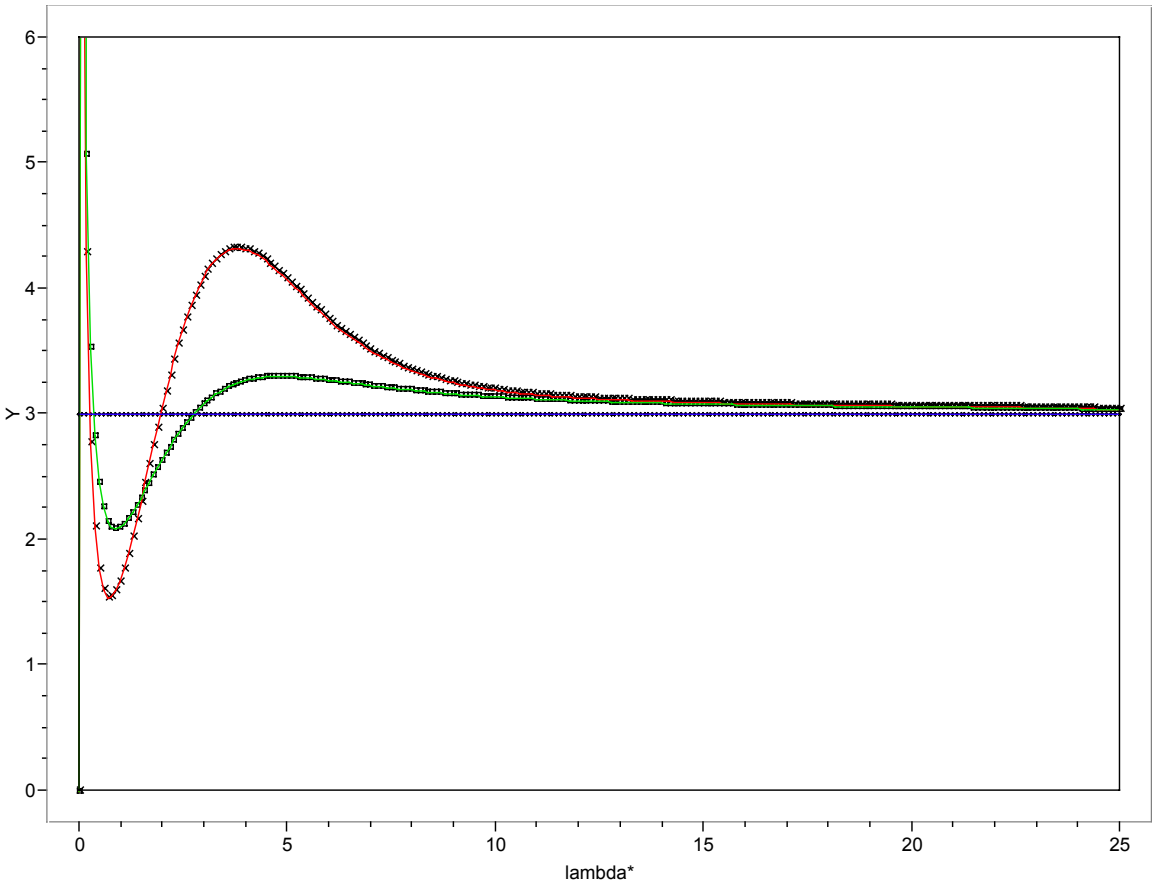
Upper curve is for c = 1/4;
Other curve is for c = 3/8;

- Our transformation does pretty well with higher moments too:
  - ◊ Skewness



$$\text{Skewness} = \frac{\mathbf{E}((\mathbf{Poiss}_{\lambda} - \boldsymbol{\mu}_{\lambda})^{3})}{\boldsymbol{\sigma}_{\lambda}^{3}}$$

Lower curve is for c = 0;
Other curve is for c = ¼;
The nominal (and limiting) value is 0.



$$\text{Kurtosis} = \frac{\mathbf{E}((\mathbf{Poiss}_{\lambda} - \boldsymbol{\mu}_{\lambda})^4)}{\boldsymbol{\sigma}_{\lambda}^4}$$

The more variable curve is for $c = 0$;

Other curve is for $c = 1/4$;

The nominal (and limiting) value is 3.

1. A “digression” because this is about one Poisson observation, but the principal goal here is density estimation and related inference. .

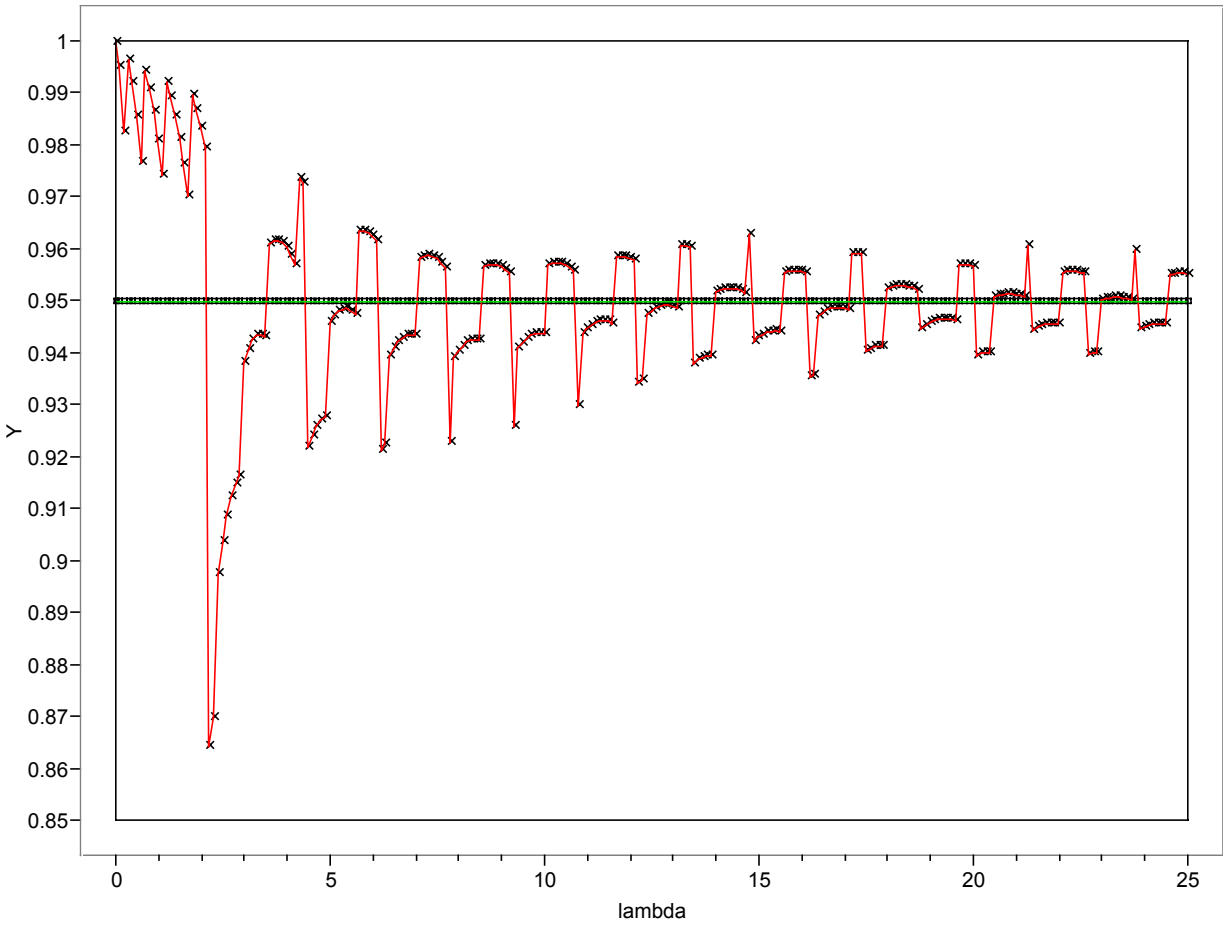
Digression[1]: Confidence Intervals For single Poisson variables

The transformation $\sqrt{N + 1/4}$ produces good individual confidence intervals. Here is a plot of the coverage of the nominal 95% interval

$$\left(\sqrt{N+\frac{1}{4}}-1.96\times\frac{1}{2}\right)_+^2,\ \left(\sqrt{N+\frac{1}{4}}+1.96\times\frac{1}{2}\right)^2$$

for λ, where N~Poiss(λ).

1. A “digression” because this is about one Poisson observation, but the principal goal here is density estimation and related inference. .



Y = Coverage of 95% root-unroot interval

as a function of λ

A naïve 95% confidence interval could be formed by exploiting the fact that N is the MLE of Var(N) = λ. This interval would be

$$\left(N - 1.96\sqrt{N}\right)_+ ,\ N + 1.96\sqrt{N}\,.$$

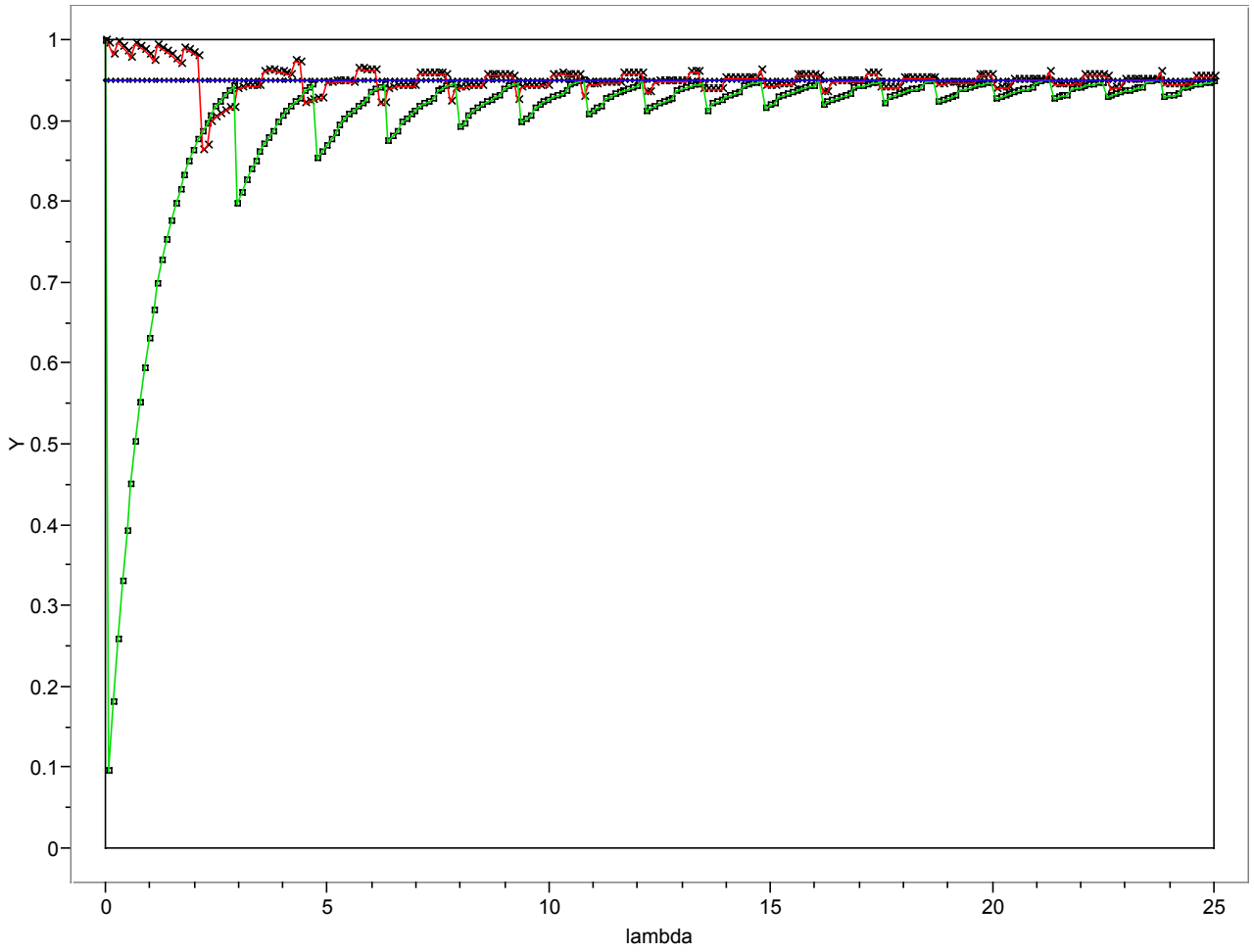
Note that the 95% root-unroot interval is

$$N - 1.96\sqrt{N} + \left(\frac{1.96}{2}\right)^2 ,\ N + 1.96\sqrt{N} + \left(\frac{1.96}{2}\right)^2$$

so long as $\sqrt{N + \frac{1}{4}} - 1.96 \times \frac{1}{2} \geq 0.$

Thus, for N$\geq$4 the two intervals have the same length, but the root-unroot interval is (always) shifted to the right.

Here is a comparison of the coverage of the two intervals:



Y = Coverage of
95% root-unroot interval and of
95% conventional Wald interval
as a function of λ

End of “Digression”

An Asymptotic Comparison

●*Fair and revealing Monte Carlo comparisons are difficult to construct; the comparisons depend much more on the appropriateness of the respective density and regression estimators than on the integrity of direct density estimation versus the root-unroot regression paradigm.*

Here is an asymptotic comparison to give an idea of the situation.

Consider a standard kernel estimator, with kernel W.

●For the ordinary density problem this is

$$\tilde{f}(t) = \frac{1}{n}\sum_i \frac{1}{d} W\left(\frac{X_i - t}{d}\right).$$

● For the root-unroot method this uses the formula

$$\hat{h}(t) = \frac{1}{K}\sum_k \frac{1}{d} W\left(\frac{t_k - t}{d}\right) Y_k,$$

and then un-roots according to $\hat{f}(t) = (\hat{h}(t))^2$.

(Asymptotically $\int \left(\hat{h}(t)\right)^2 dt \cong 1$, very nearly, so we not not carry out the renormaliztion step.)

●Take W = Unif(-1/2, 1/2) for numerical simplicity, and

$$d = C / n^{1/5}.$$

●Define the IMSE Risk for an estimator $\breve{f}$ as

$$R = \int (\breve{f}(t) - f(t))^2 \, dt \approx \frac{1}{K}\sum (\breve{f}(t_k) - f(t_k))^2 .$$

Note that R can be decomposed as

R = $\int$Bias^2 + $\int$Var.

●Then, for the density estimator

$$n^{4/5}R_{dens} = \frac{C^4}{24^2}\int (f'')^2 + \frac{1}{C} \equiv C^4\Psi^2_{dens} + \frac{1}{C},$$

and for the Root-unroot estimator

$$n^{4/5}R_{r-u} \sim \frac{C^4}{24^2}\int (f'' - \frac{(f')^2}{2f})^2 + \frac{1}{C} \equiv C^4\Psi^2_{R-U} + \frac{1}{C},$$

since

$$\left(\hat{h}^2 - h^2\right)^2 \sim_{(prob)} \left(\hat{h} - h\right)^2 4h^2 \quad \text{so that}$$

$$\int bias^2_{\hat{h}^2} \sim \int 4h^2 \left(bias_{\hat{h}}\right)^2 =$$

$$\int 4h^2 \times (h'')^2 = \int (f'' - \frac{(f')^2}{2f})^2.$$

●Clearly, it is possible that $R_{dens} < R_{r\text{-}u}$ (if f is linear); and also that $R_{dens} > R_{r\text{-}u}$ (if $h = \sqrt{f}$ is linear).

•Here is a table showing the results for some simple situations:

Density = .95**f**+.05 : **f**=	Ψ^2_{dens}	$\Psi^2_{R\text{-}U}$	R^{**}_{dens}	$R^{**}_{R\text{-}U}$
$2\sin^2\pi x$	1.23	1.06	1.72	1.67
$30x^2(1-x)^2$	1.13	0.96	1.69	1.64
$6x(1-x)$	0.23	3.60	1.23	2.13

Table of $\int \text{Bias}^2$ term, Ψ^2,

for 3 simple density functions.

The table also shows the Oracle Risk, R**, where

$$R^{**} = \inf_C \{n^{4/5} R\} = \left(\frac{5}{4^{4/5}}\right) \times \Psi^{2/5}.$$

Conclusion: The root-unroot estimator is very slightly superior to the original density estimator when f is "smooth" in the region where it's small. It is considerably inferior when f is linear in that region.

●The previous considerations suggest that it might be more sensible to model the problem in terms of $h = \sqrt{f}$; and to judge the quality of an estimator according to its squared Hellinger risk –

$$H \cdot R = \int \left(\sqrt{\breve{f}(t)} - \sqrt{f(t)} \right)^2 dt \approx \frac{1}{K} \sum \left(\sqrt{\breve{f}(t_k)} - \sqrt{f(t_k)} \right)^2$$

.

(The above general conclusion remains valid.)

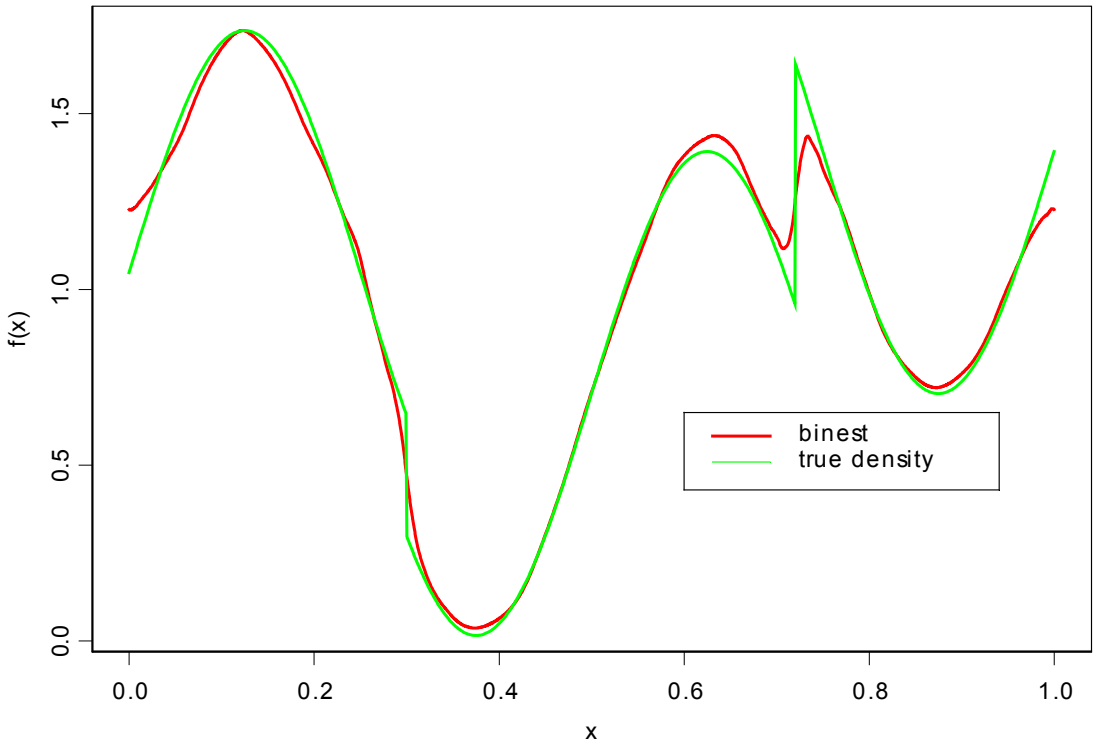
Simulations

●The Root-unroot scheme enables reduction of the problem to one effectively having $K = 2^m$ equally spaced, ***homoscedastic*** and ***independent*** observations. This is **exactly** suitable for wavelet analyses.

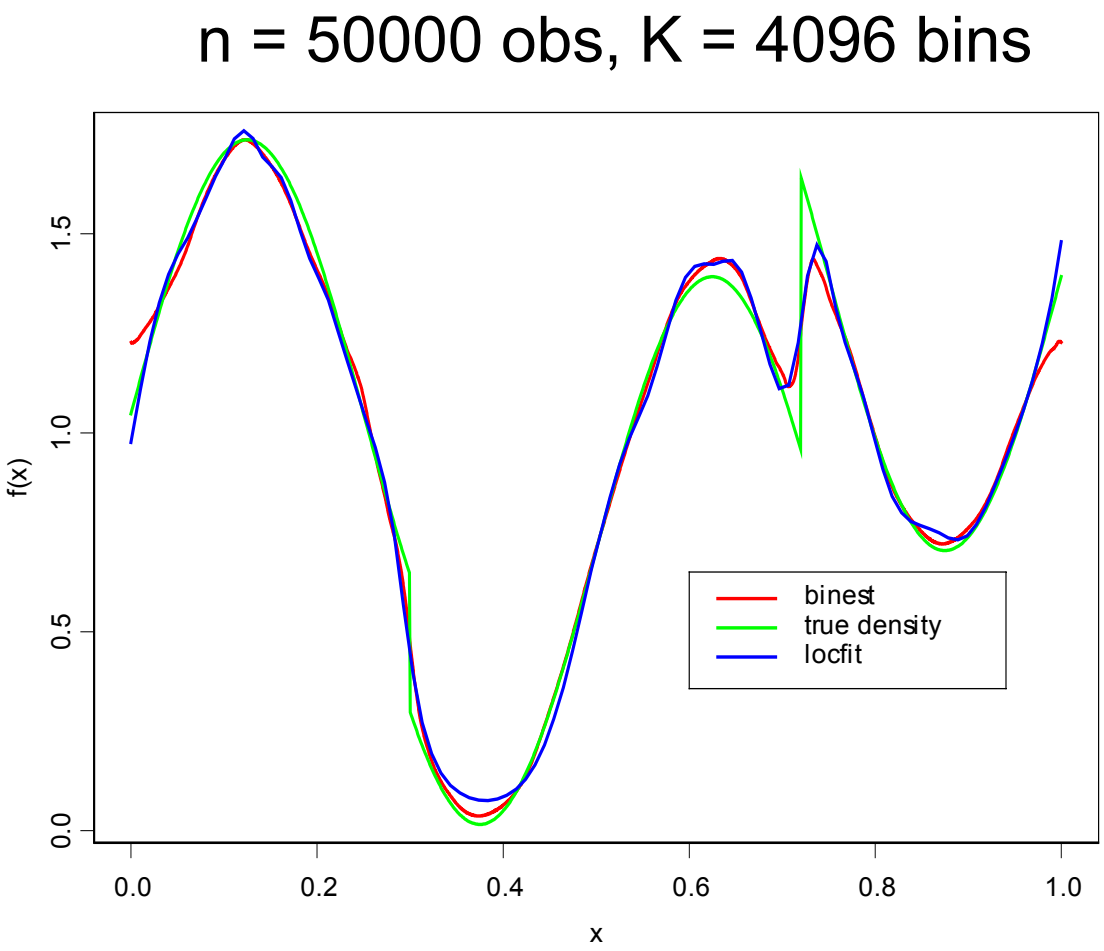
●Here are plots showing the result of using the root-unroot scheme along with the Block J-S wavelet procedure in Cai and Silverman(1998+). The functions exhibited here are "hard-to-fit" forms adapted from Donoho and Johnstone (1995). ("Heavisine" and "Blocks")

n = 50000 obs, K = 4096 bins,

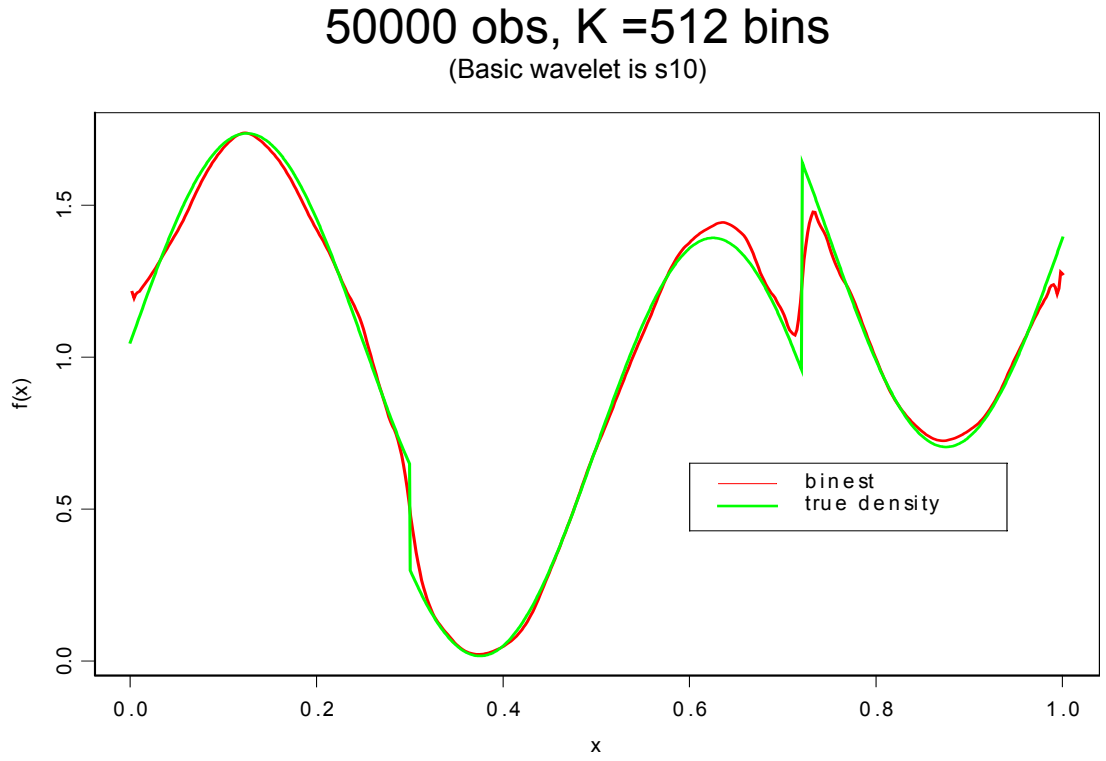
(Basic wavelet is s10.)

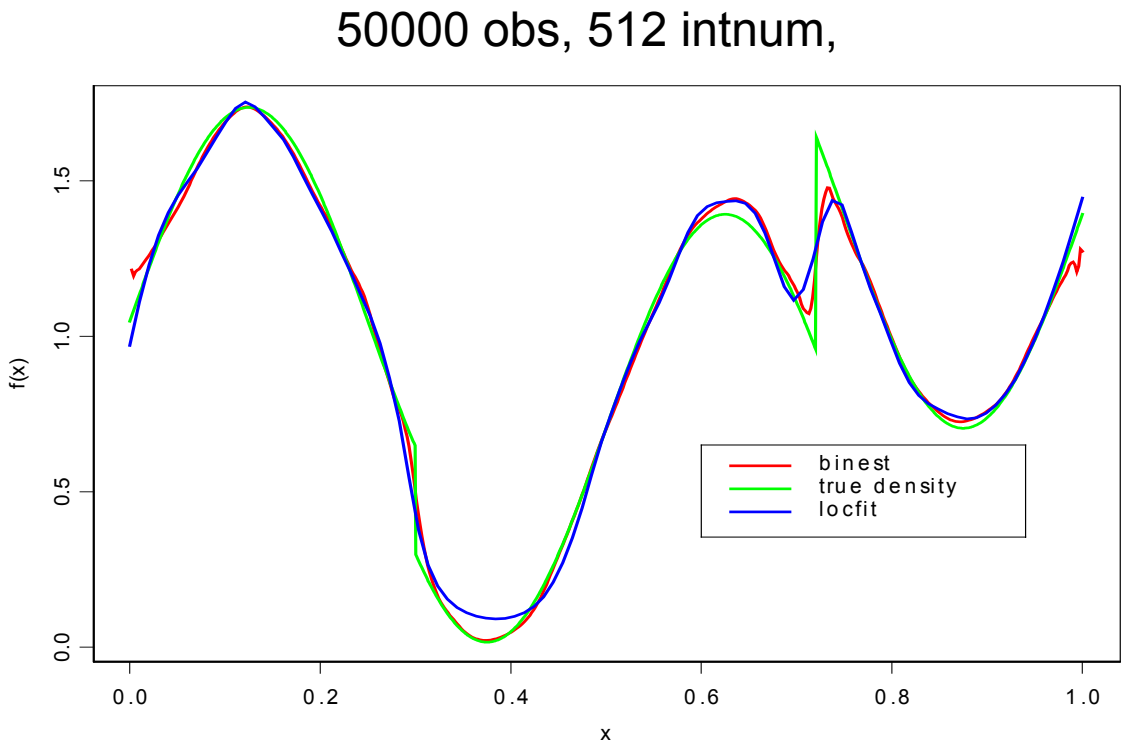


Here is the same plot, also showing a local linear fit estimator.



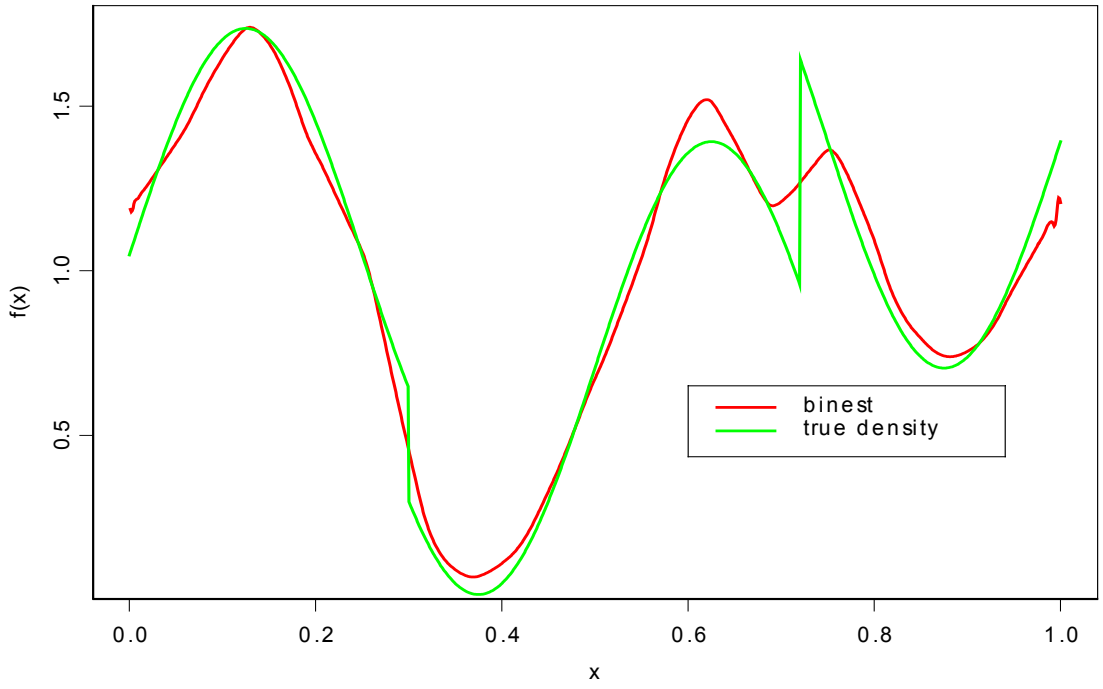
More Heavisine plots



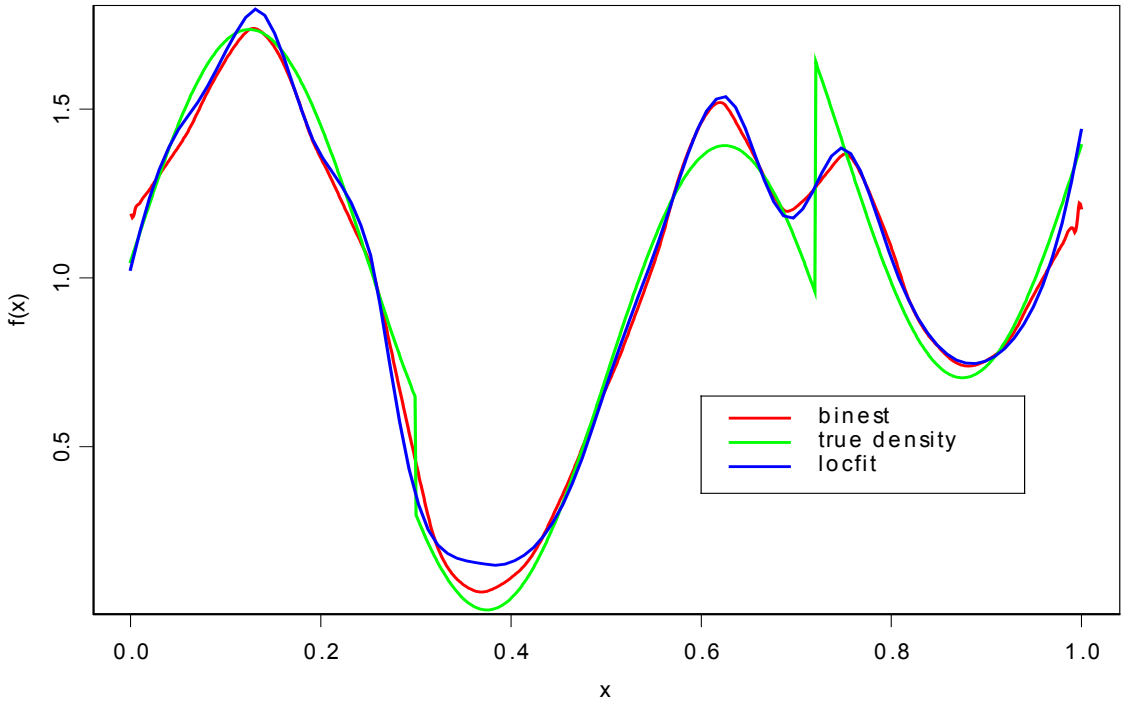


Here are Heavisine plots with fewer observations. Note that the quality of fit is much poorer.

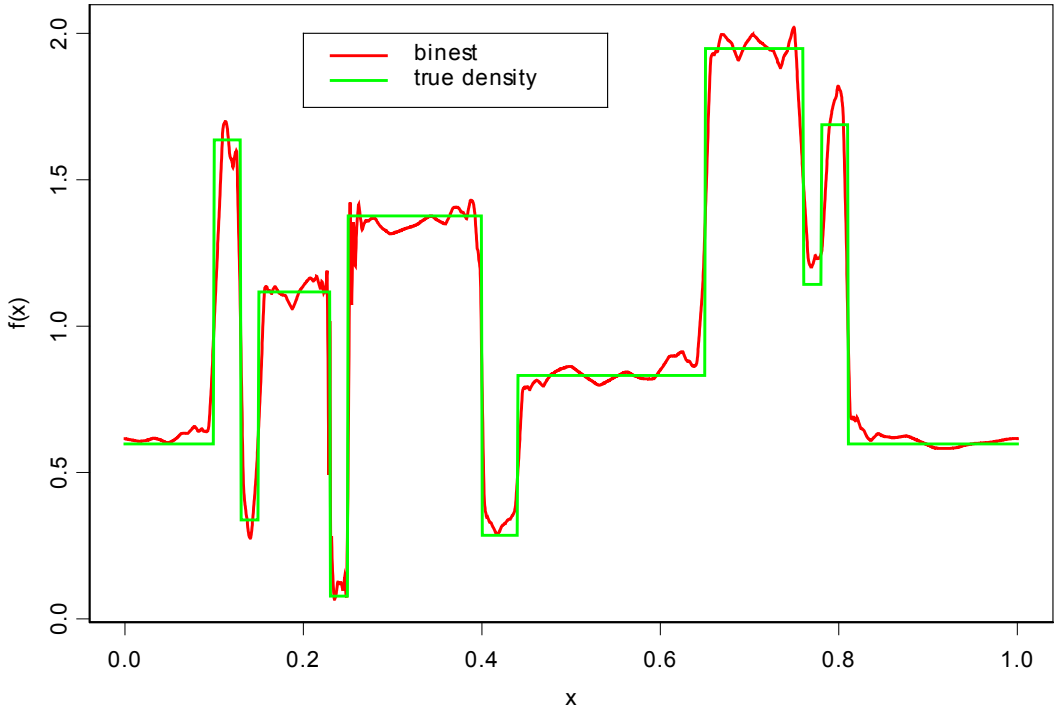
6000 obs, 1024 intnum,



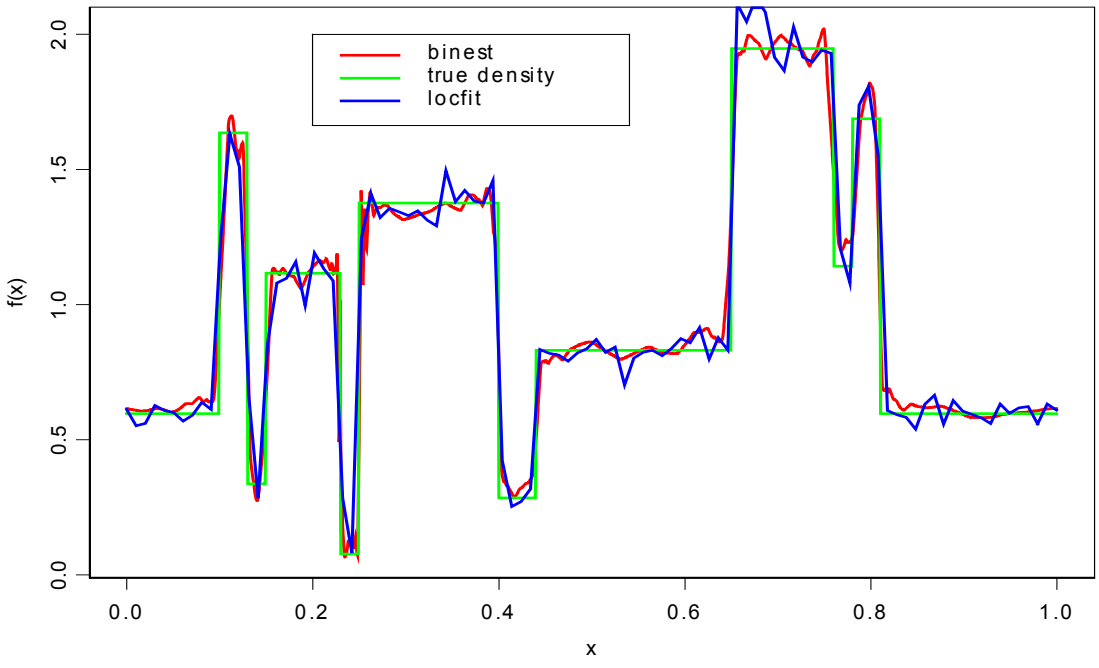
6000 obs, 1024 intnum



BLOCKS: n = 50000 obs, K = 4096 (s8)

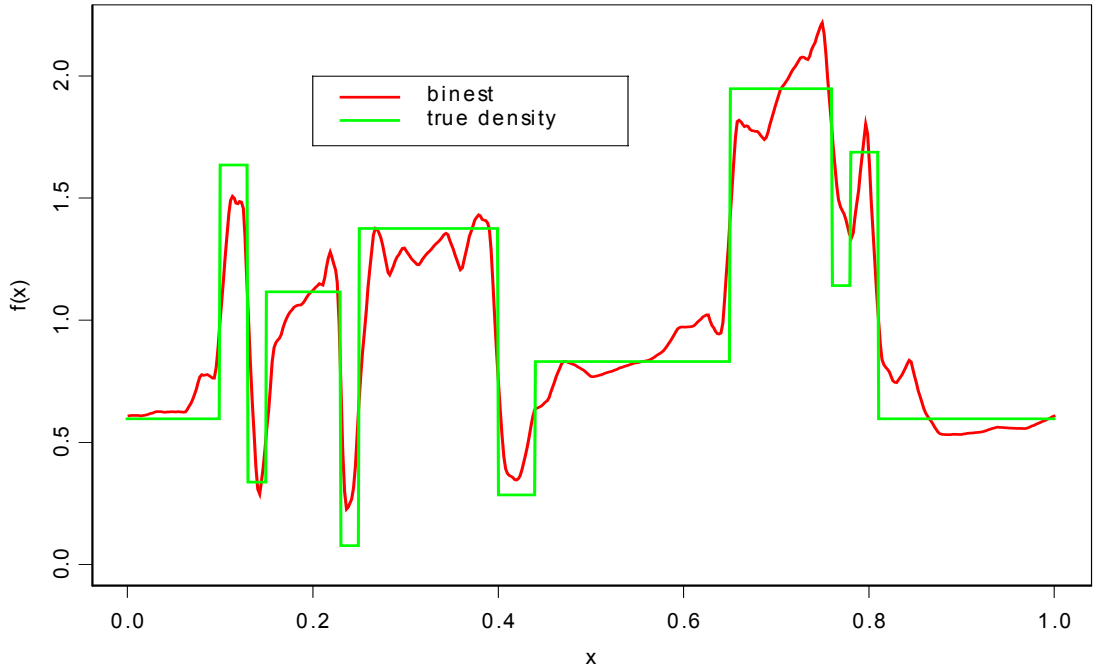


n = 50000 obs, K = 4096 (s8)

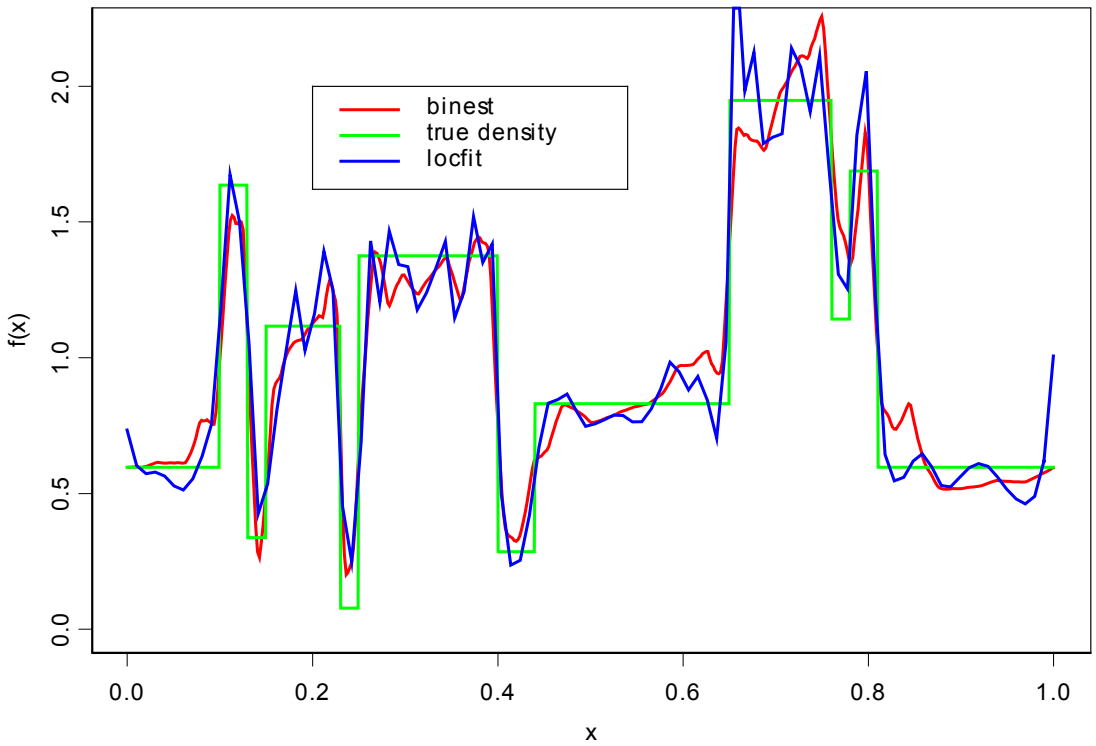


Note: Use of smaller K, K = 512, barely affects the plots.

BLOCKS: n = 6000 obs, K = 1024 (s8)



n = 6000 obs, K = 1024 (s8)



Signal to noise ratio

Question: Why are the plots for **n = 6000** so poor? Why aren't the plots for

n = 50000

nearly perfect? (That's the situation with the nonparametric regression data analyzed in Donoho, et. al. (1994, 1995, +) for these curves.)

Answer: Because density estimation is intrinsically much harder than the nonparametric regression situations of those articles.

Detail: Here is a scale-free (in Y) **D**ifficulty of **E**stimation index:

$$\mathbf{DE} = \frac{\textit{Average local S.D. of the error}}{\textit{Average intensity of the signal}}\Delta.$$

Where Δ denotes a scale free (in Y) measure of the local complexity of the signal (see below). For a homoscedastic regression problem with signal g and q observations each having S.D. = σ this is

$$\mathbf{DE} = \frac{\sigma / \sqrt{q}}{\left(\int (g(x) - \bar{g})^2\right)^{1/2}} \Delta = \frac{1}{SNR \times \sqrt{q}} \Delta \ ,$$

where SNR denotes the conventional Signal to Noise ratio,

$$SNR = \frac{\left(\int (g(x) - \bar{g})^2\right)^{1/2}}{\sigma}.$$

A suitable form for Δ is

$$\Delta^4 = \frac{\left(\int (f'')^2\right)^{1/2}}{\left(\int (f - \bar{f})^2\right)^{1/2}}$$

The root-ed density problem is a nonparametric regression with signal h = $\sqrt{f}$, and there are K observed values of Y_k, each having $\sigma = \frac{\sqrt{K}}{2\sqrt{n}}$.

We get

$$\mathbf{DE}_h = \frac{\frac{\sqrt{K}}{2\sqrt{n}} \Big/ \sqrt{K}}{\left(\int (h(x)-\bar{h})^2\right)^{1/2}} \Delta_h = \frac{1}{\left(\int (h(x)-\bar{h})^2\right)^{1/2} \times 2\sqrt{n}} \Delta_h$$

(Note that K cancels out of this expression, as it should so long as n/K is not very small.) This is the DE for the root-ed problem.

Unrooting makes the problem twice as hard since $(h+\varepsilon)^2 \approx h^2 + \mathbf{2}\varepsilon = f + \mathbf{2}\varepsilon$;

The complexity factor also changes when comparing the rooted and unrooted problems, but the change is small; thus

$$\frac{\Delta_h}{\Delta_f} = 1 \pm (a\ little)$$

Hence we should ascribe to the density problem the difficulty

$$\mathbf{DE}_{\text{orig prob}} = \frac{1}{\left(\int (h(x)-\bar{h})^2\right)^{1/2} \times \mathbf{1}\sqrt{n}} \Delta \ .$$

Here is a table showing $\mathbf{DE}_{orig}$ for our density problem as compared to the **DE**'s for the regression setup of Donoho, et. al. in which SNR = 7 and n = 4096.

$1000 \times$DE for

Function	density	regression
Heavisine	4.98Δ	$.223\Delta$
Blocks	6.13Δ	$.223\Delta$

$1000 \times$DE's for n = 4096.

(DE_{reg} is for SNR=7.)

According to this measure, the density problem is from 20 to 30 times harder than the regression one at the same sample size.

THUS

It needs from 400 to 900 as many observations to provide a comparable quality of estimate.

Empirical Example
From a Bank Call-in Center in Israel

The data are records of all telephone calls made during 1999 to a telephone assistance and service center operated by a bank in Israel.

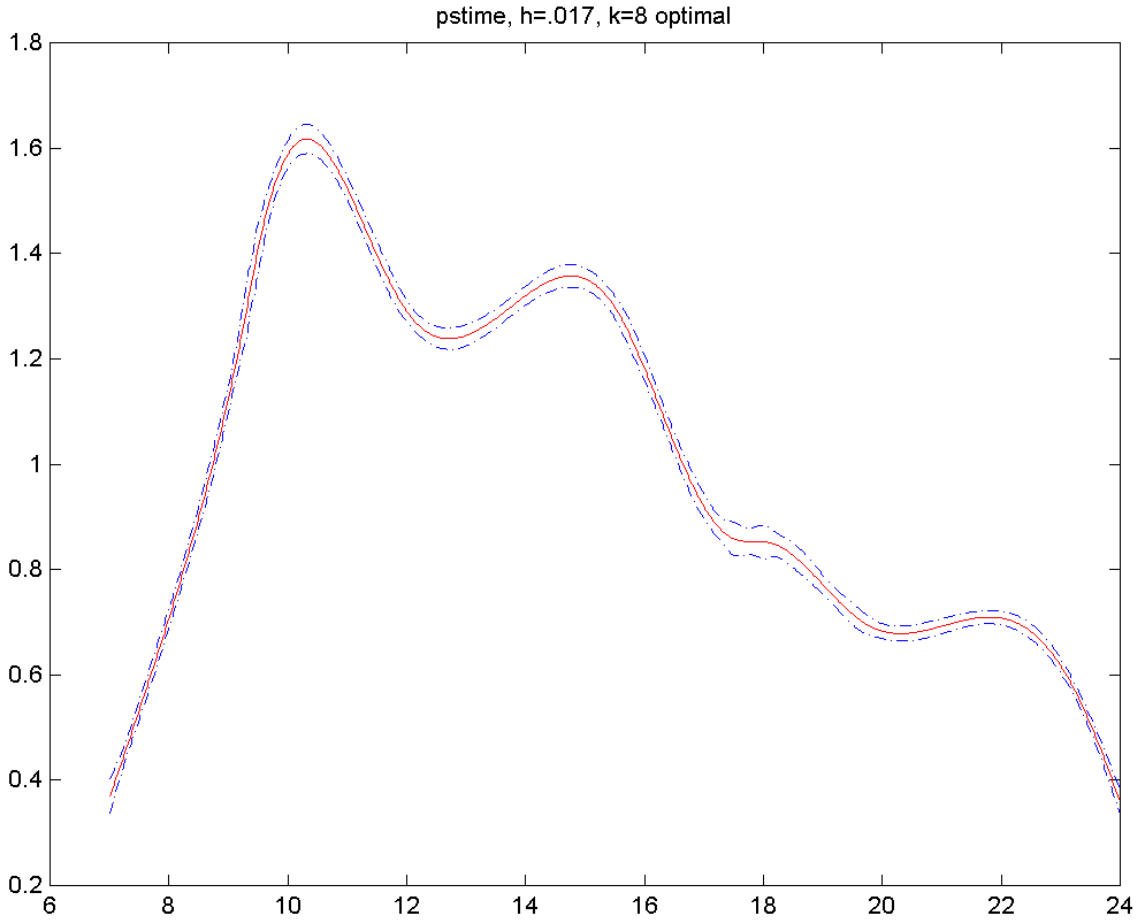
Today I'll present just an analysis of the time of day that these calls arrive at the center. We look only at telephone business hours - 7am to midnight – on regular weekdays (Sun. through Thurs. in Israel), excluding holidays.

The first plot is for REGULAR customers. (n =258,500). The second is for INTERNET customers (n = 14,320), who call a special number seeking assistance with on-line banking.

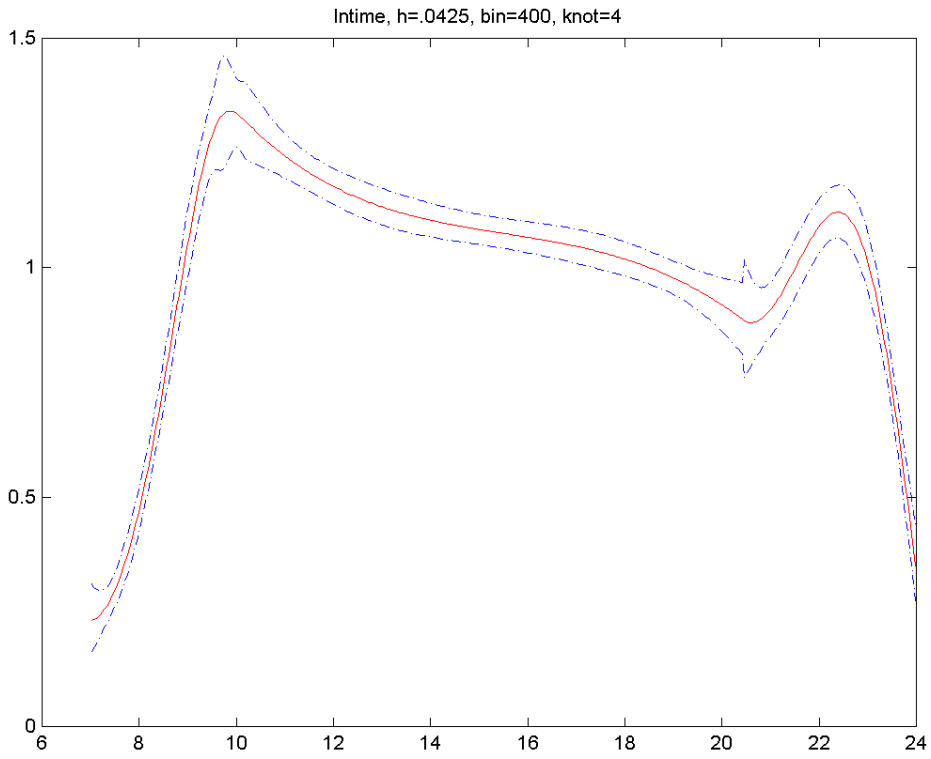
(Differences for the days of the week and the seasons of the year did exist, but were relatively minor. These are ignored in the current analysis, but will be considered in the analysis in the last section of the talk.)

An advantage of the root-unroot methodology is that it reduces the problem to a homoscedastic regression in which better-understood tools lead to 95% confidence (variance) interval bands in addition to estimates.

The following plots were produced via a root-unroot scheme, using the free-knots nonparametric regression spline methodology of Zhao (1999, rev. 2001).



Arrival time of REGULAR calls



Arrival time of INTERNET calls

Note that the arrival patterns of regular and internet customers are significantly different. (The internet calls arrive more uniformly across the day, but with a statistically significant, noticeable late night local-mode at 10-11pm.)