# Robust And Generalized Nonparametric Regression

T. Tony Cai
Department of Statistics
The Wharton School
University of Pennsylvania

*http://stat.wharton.upenn.edu/~tcai*

Joint Work with Larry Brown and Harrison Zhou

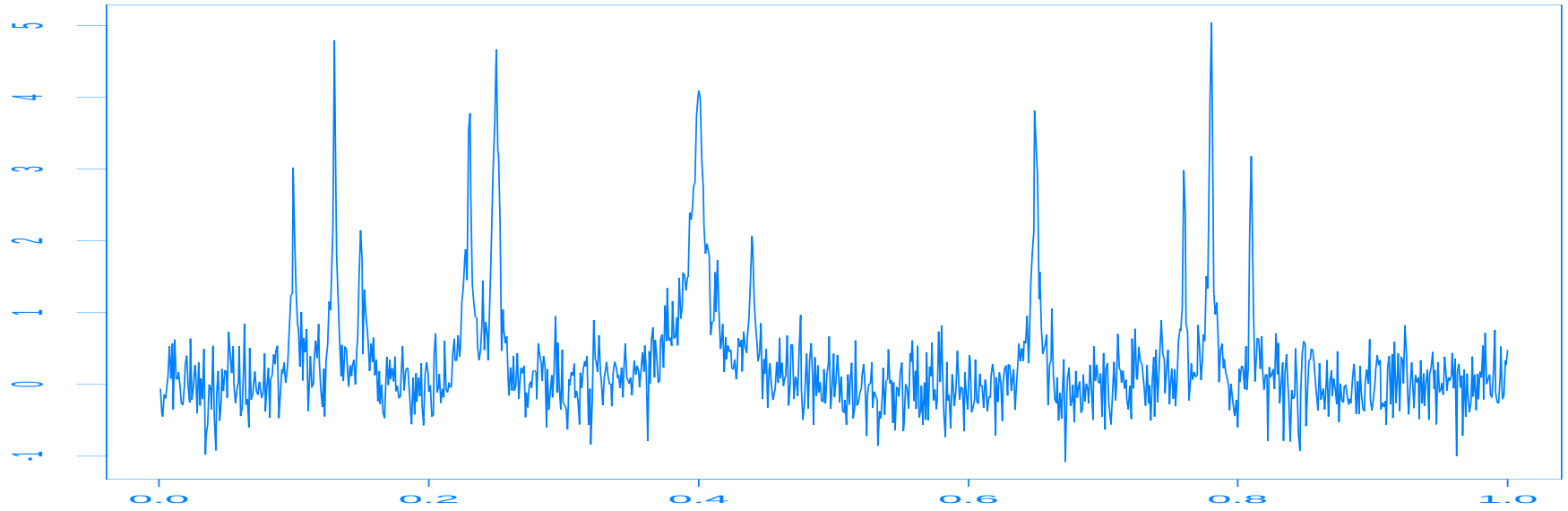
# Outline

- Introduction
- Robust Nonparametric Regression
  - Median Coupling
- Nonparametric Regression in Exponential Families
  - Mean-Matching VST for Exponential Families
- Discussion:
  - Density Estimation

## "A Gaussianization Machine"

# Gaussian Nonparametric Regression

$$y_i = f(t_i) + \sigma z_i, \quad z_i \overset{iid}{\sim} N(0,1), \quad i = 1, \cdots, n.$$




- **Method:** Kernel, local polynomial, spline, wavelet thresholding, ...
- **Theory:** Minimax theory, (global/local) adaptation theory, ...

# What About Non-Gaussian Noise?

**"Nothing is Gaussian"**

# Robust Nonparametric Regression

Observe

$$Y_i = f(t_i) + \xi_i, \quad i = 1, ..., n$$

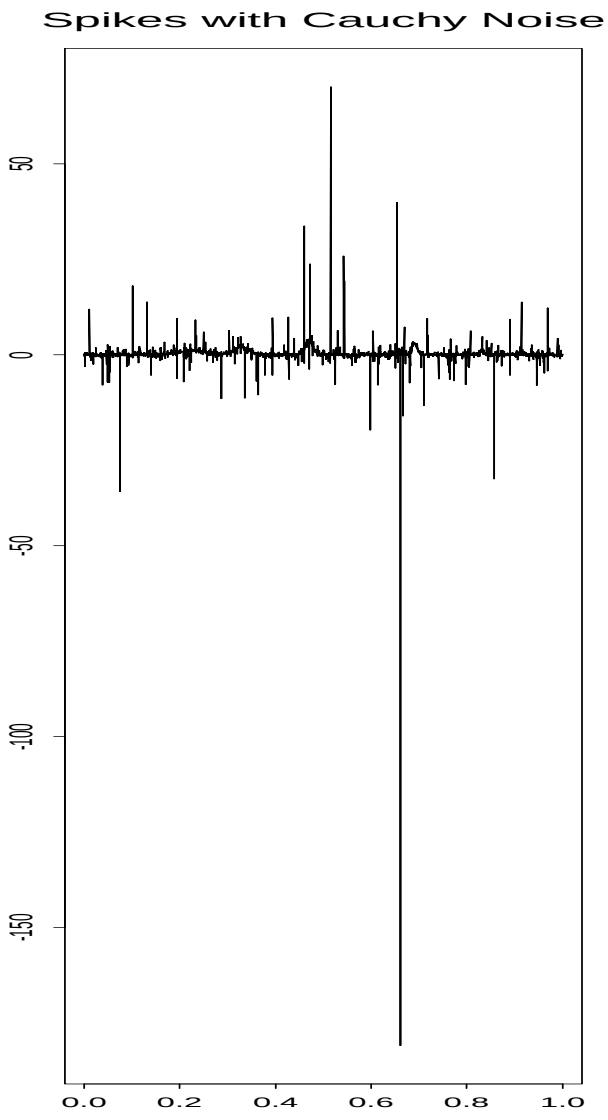
where $\xi_i$ are iid with an unknown distribution and median$(\xi_i) = 0$.

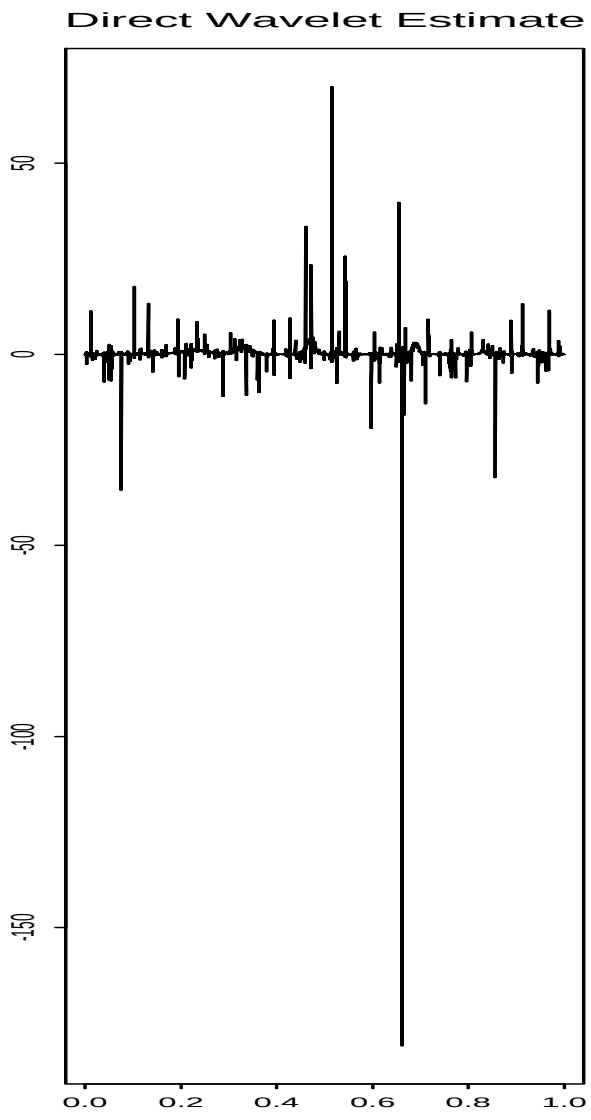
Standard methods such as kernel smoothing and wavelet thresholding would fail when the noise is heavy-tailed.

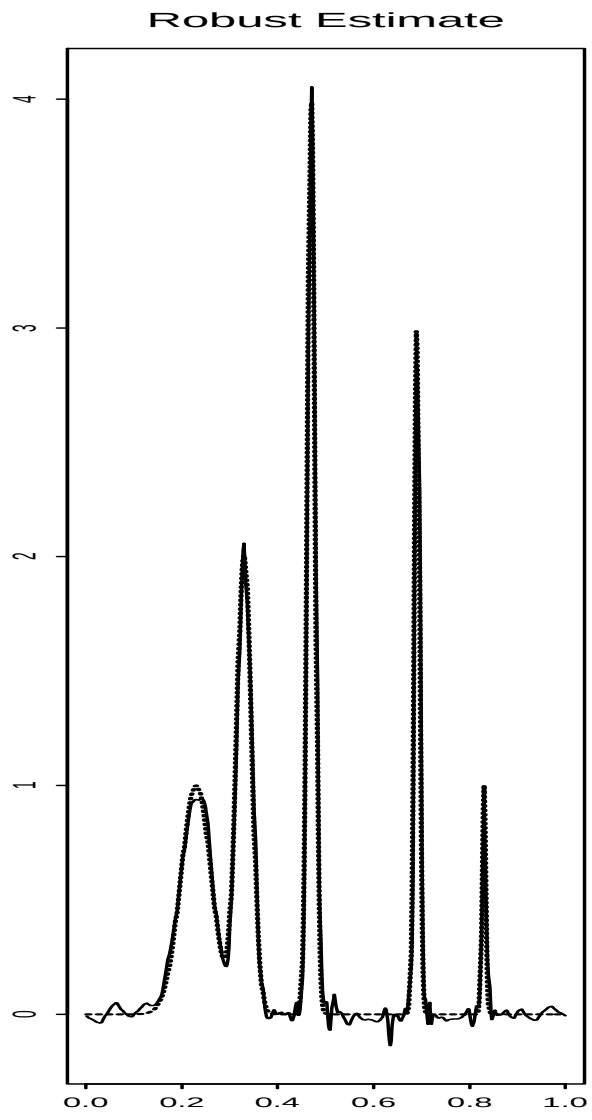
**Example:** In Cauchy regression where $\xi_i$ has a Cauchy distribution, typical realizations of $\xi_i$ contain a few extremely large observations of order $n$ since

$$P(\max\{\xi_i\} \geq n) = \left(\frac{1}{\pi}\arctan(n) + \frac{1}{2}\right)^n \to \exp\left(-\frac{1}{\pi}\right).$$

In contrast, in Gaussian regression $\max\{\xi_i\} \asymp \sqrt{2\log n}$.


Spikes with Cauchy Noise
50
0
-50
-100
-150
0.0
0.2
0.4
0.6
0.8
1.0


Direct Wavelet Estimate
50
0
-50
-100
-150
0.0
0.2
0.4
0.6
0.8
1.0


Robust Estimate
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0

# Transformation

Transformation is one of the most commonly used tools in statistics. The goal of transformation is to convert a complicated-looking problem into a more familiar/simpler problem.

- **Normalizing transformation**
- **Symmetrizing transformation**
- **Linearizing transformation**
- **Variance-stabilizing transformation**

# Robust Regression

1. **Binning & Taking Median**: Divide the indices $\{1, ..., n\}$ into $T$ equilength bins $\{I_j : j = 1, ..., T\}$ of size $m$ and let

$$Y_j^* = \text{median}(Y_i : \ i \in I_j).$$

Then $Y_j^*$ can be treated as if it were a normal random variable with mean $g(\frac{j}{T}) = f(\frac{j}{T}) + b_m$ and variance $\sigma^2 = 1/(4mh^2(0))$, where $h$ is the density function of $\epsilon_i$ and

$$b_m = E\{\text{median}(\xi_1, \ldots, \xi_m)\}. \tag{1}$$

Both the variance $\sigma^2$ and the bias $b_m$ can be estimated easily.

2. **Gaussian Regression:** Applying your favorite Gaussian regression procedure to $\{Y_j^*, \ j = 1, ..., T\}$ to yield an estimator $\hat{g}$. The regression function $f$ is then estimated by $\hat{f} = \hat{g} - \hat{b}_m$ where $\hat{b}_m$ is an estimator of the bias $b_m$.

# The BlockJS Procedure

To illustrate the ideas, we shall use the **BlockJS** procedure in the second step.

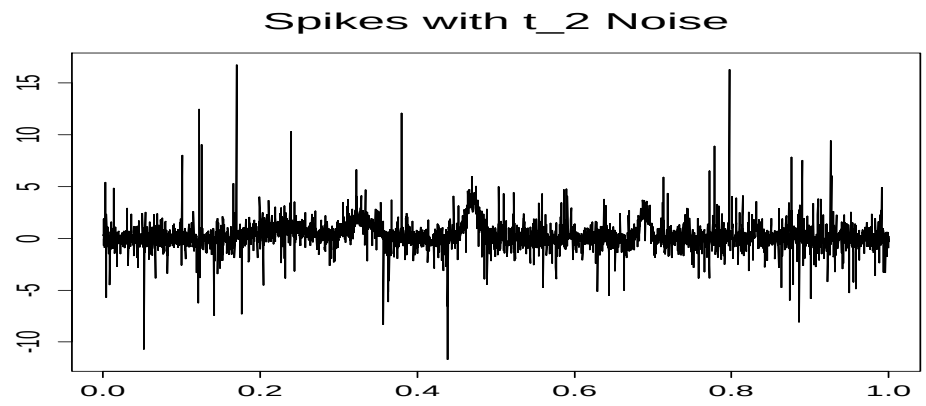
1. Transform the noisy data via the DWT.
2. At each resolution level, estimate wavelet coefficients via block thresholding

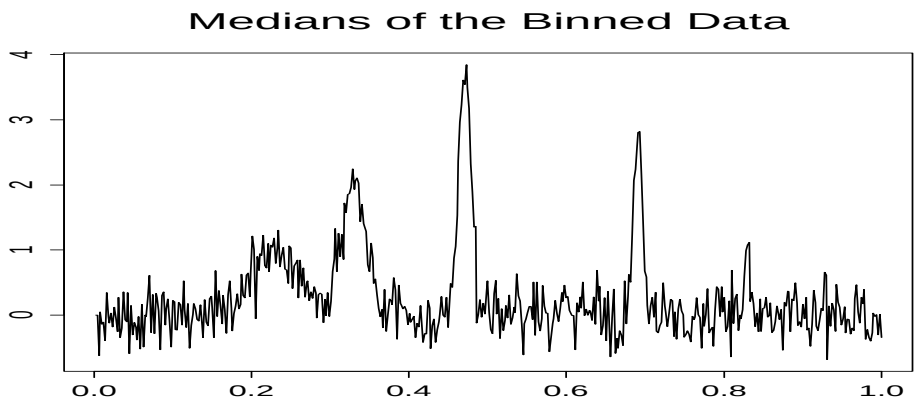
$$\hat{\theta}_{B_j} = (1 - \frac{\gamma L_* \sigma^2}{S_j^2})_+ \tilde{y}_{B_j},$$

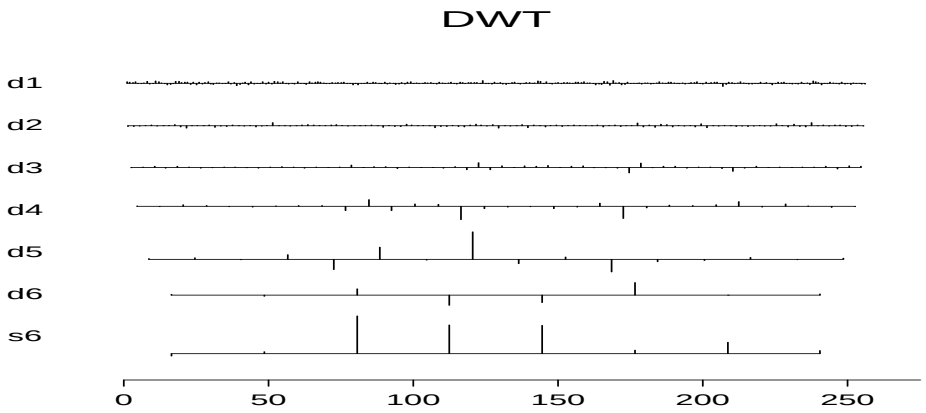
   where $L_* = \log n$ and $\gamma = 4.5053$ (root of $\gamma - \log \gamma - 3 = 0$).
3. Estimate the function $f$ via the IDWT.

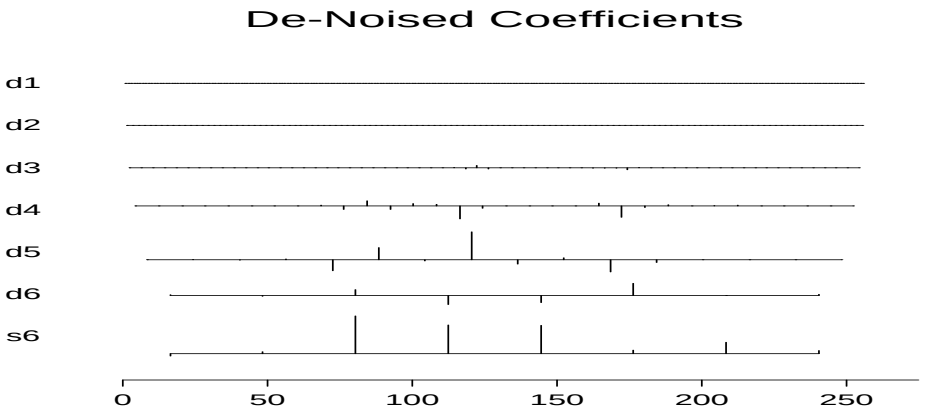
**Properties:**

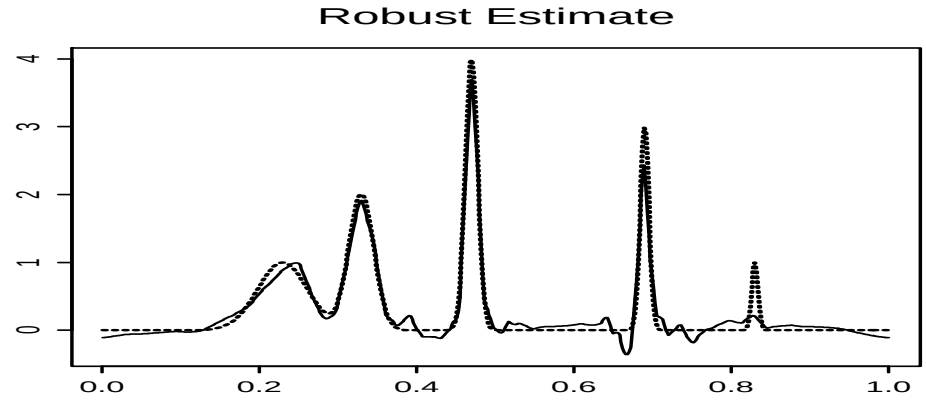
- **Globally adaptive**
- **Spatially adaptive**

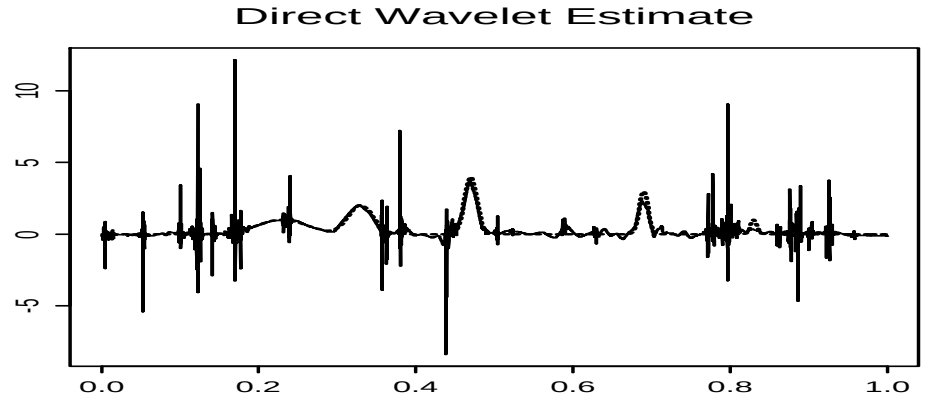

Spikes with t_2 Noise
15
10
5
0
-5
-10
0.0
0.2
0.4
0.6
0.8
1.0


Medians of the Binned Data
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0


DWT
d1
d2
d3
d4
d5
d6
s6
0
50
100
150
200
250


De-Noised Coefficients
d1
d2
d3
d4
d5
d6
s6
0
50
100
150
200
250


Robust Estimate
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0


Direct Wavelet Estimate
10
5
0
-5
0.0
0.2
0.4
0.6
0.8
1.0

# More Details

- **The Bias** $b_m$ is the systematic bias due to the expectation of the median of the noise $\xi_i$ in each bin. It can be estimated as follows.
  - Divide each bin $I_j$ into two sub-bins with the first bin of the size $\lfloor \frac{m}{2} \rfloor$. Let $X_j^*$ be the median of observations in the first sub-bin and set

$$\hat{b}_m = \frac{1}{T} \sum_j \left( X_j^* - X_j \right). \tag{2}$$

- **The empirical wavelet coefficients** can be written as

$$y_{j,k} = \theta_{j,k} + \epsilon_{j,k} + \frac{1}{2h\left(0\right)\sqrt{n}} z_{j,k} + \xi_{j,k}, \tag{3}$$

where $\theta_{j,k}$ are the true coefficients of $g = f + b_m$, $\epsilon_{j,k}$ are "small" deterministic approximation errors, $z_{j,k}$ are i.i.d. $N(0,1)$, and $\xi_{j,k}$ are some "small" stochastic errors.

# Adaptivity of the Procedure

**Theorem 1** *Let* $m = Cn^{\frac{3}{4}}$. *Then*

$$\sup_{h\in\mathcal{H}} \sup_{f\in B^{\alpha}_{p,q}(M)} E\|\widehat{f}_n - f\|_2^2 \le \begin{cases} Cn^{-\frac{2\alpha}{1+2\alpha}} & p \ge 2,\ \frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0 \\ Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}} & 1 \le p < 2,\ \frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0. \end{cases}$$

**Theorem 2** *Let* $m = Cn^{\frac{3}{4}}$ *and* $\alpha > 1/6$. *Then*

$$\sup_{h\in\mathcal{H}} \sup_{f\in\Lambda^{\alpha}(M,t_0,\delta)} E(\widehat{f}_n(t_0) - f(t_0))^2 \le C \cdot (\frac{\log n}{n})^{\frac{2\alpha}{1+2\alpha}}. \qquad (4)$$

## Main Technical Tool: Median Coupling

- **Median Coupling Inequality** Let $Z \sim N(0,1)$ and let $X_1, \ldots, X_n$ be i.i.d. with density $h$ such that $\int_{-\infty}^{0} h(x) = \frac{1}{2}$, $h(0) > 0$, and $h(x)$ is Lipschitz at $x = 0$. Then for every $n$ there is a mapping $\widetilde{X}_{med}(Z) : \mathbb{R} \mapsto \mathbb{R}$ such that $\mathcal{L}\left(\widetilde{X}_{med}(Z)\right) = \mathcal{L}(X_{med})$ and

$$\left|\sqrt{4n}h(0)\,\widetilde{X}_{med} - Z\right| \leq \frac{C}{\sqrt{n}}\left(1 + |Z|^2\right), \text{ when } |Z| \leq \varepsilon\sqrt{n} \tag{5}$$

## Remark

If the error distribution is known/assumed to be **symmetric**, much stronger results can be obtained. (In this case, $b_m \equiv 0$.)

- **Asymptotic Equivalence:** The medians $\{Y_j^*\}$ can be shown to be asymptotically equivalent to a Gaussian experiment.
- **Smaller Bin Size:** For robust estimation of the regression function, the bin size can be taken to be logarithmic.
- **Estimating Functionals:** Similar results can be derived for estimating functionals such as linear or quadratic functionals.

# Generalized Nonparametric Regression

## Other Nonparametric Function Estimation Models

- *White Noise Model:*

$$dY(t) = f(t)dt + \epsilon dW(t)$$

  where $W$ is a standard Brownian motion.

- *Density Estimation:* $X_1,\ X_2,\ \cdots, X_n \overset{i.i.d.}{\sim} f.$

- *Poisson Regression:* $X_i \sim \text{Poisson}(\lambda(t_i)),\ \ i = 1, ..., n.$

- *Binomial Regression:* $X_i \sim \text{Binomial}(r, p(t_i)),\ \ i = 1, ..., n.$

- ...

## Asymptotic Equivalence Theory

- Regression $\Longleftrightarrow$ White Noise

  Brown & Low (1996), Brown, Cai, Low, & Zhang (2002).

- Density Estimation $\Longleftrightarrow$ White Noise

  Nussbaum (1996), Brown, Carter, Low, & Zhang (2004).

- Density Estimation $\Longleftrightarrow$ Poisson Regression (Poissonization)

  Low and Zhou (2007).

# Natural Exponential Families

Let $X_1, X_2, ..., X_n$ be a random sample from the distribution in the natural exponential family (NEF) with the pdf/pmf

$$f(x|\xi) = e^{\xi x - \psi(\xi)} h(x).$$

The mean and variance are $\mu(\xi) = \psi'(\xi)$, and $\sigma^2(\xi) = \psi''(\xi)$ respectively.

In the subclass with a quadratic variance function (QVF),

$$\sigma^2 \equiv V(\mu) = v_0 + v_1\mu + v_2\mu^2. \tag{6}$$

We shall write

$$X_1, X_2, ..., X_n \sim NQ(\mu).$$

NEF-QVF families consist of six important distributions, three continuous: **Normal, Gamma**, and **NEF-GHS** distributions; three discrete: **Binomial, Negative Binomial**, and **Poisson** (see, e.g., Morris (1982) and Brown (1986)).

# Variance Stabilizing Transformation (VST)

Set $X = \sum_{i=1}^{n} X_i$. Then

$$\sqrt{n}(X/n - \mu(\xi)) \xrightarrow{L} N(0, V(\mu(\xi))).$$

The VST is a function $G : I\!R \to I\!R$ such that

$$G'(\mu) = V^{-\frac{1}{2}}(\mu). \tag{7}$$

The delta method yields

$$\sqrt{n}\{G(X/n) - G(\mu(\xi))\} \xrightarrow{L} N(0, 1).$$

The variance stabilizing properties can be improved by using

$$H(X) = G(\frac{X + a}{n + b})$$

with suitable choice of $a$ and $b$. See, e.g., Anscombe (1948).

# Mean-Matching VST

For us, mean matching is more important than variance stabilizing.

**Lemma 1**

$$E\{G(\frac{X+a}{n+b})\} - G(\mu(\xi)) = \frac{1}{\sigma(\xi)}(a - b\mu(\xi) - \frac{\mu''(\xi)}{4\mu'(\xi)}) \cdot n^{-1} + O(n^{-2}). \quad (8)$$

*There exist constants $a$ and $b$ such that*

$$E\{G(\frac{X+a}{n+b})\} - G(\mu(\xi)) = O(n^{-2}) \quad (9)$$

*if and only if the NEF has a QVF.*

In the NEF-QVF with

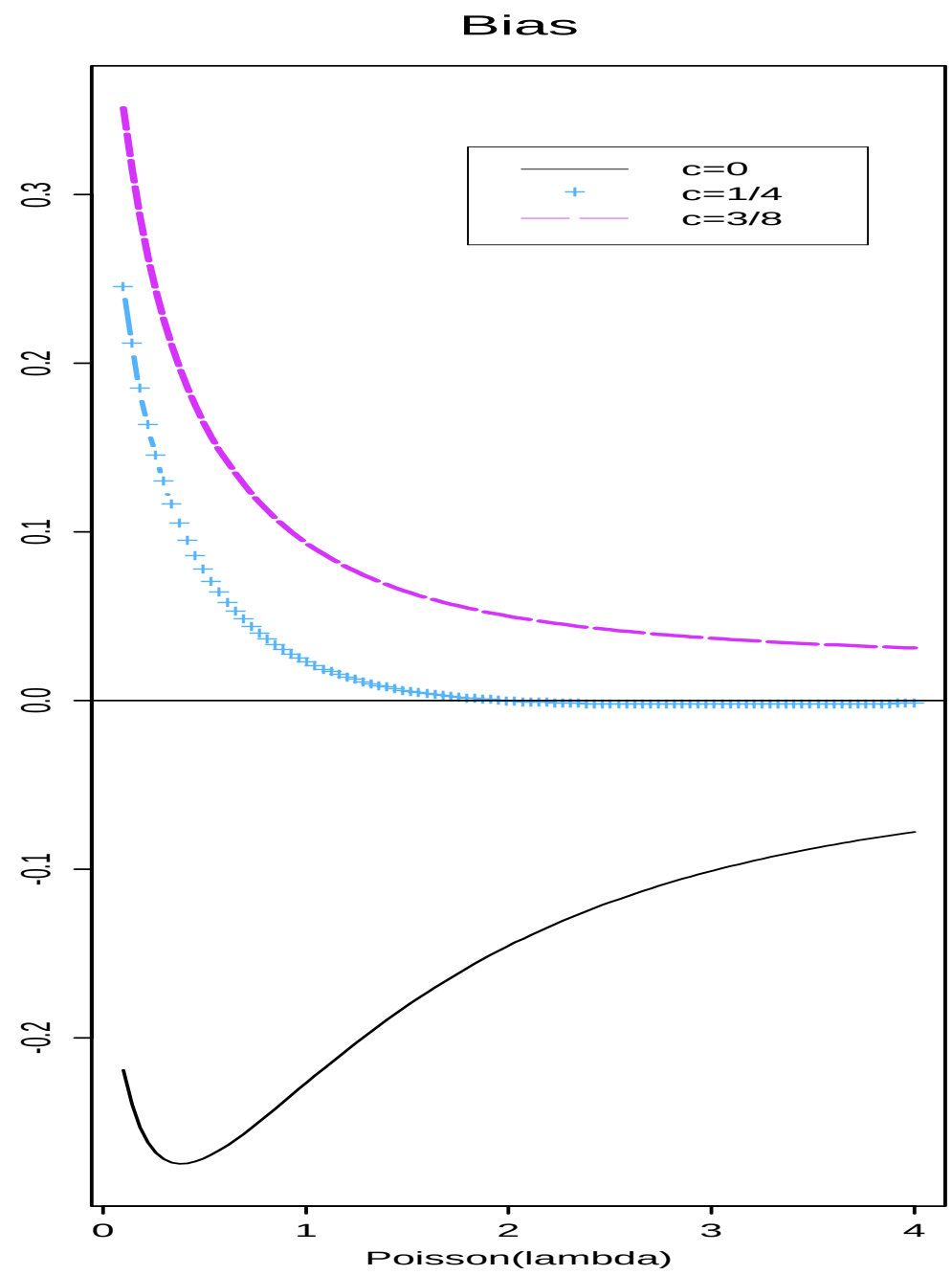
$$V(\mu) = v_0 + v_1\mu + v_2\mu^2 \quad (10)$$

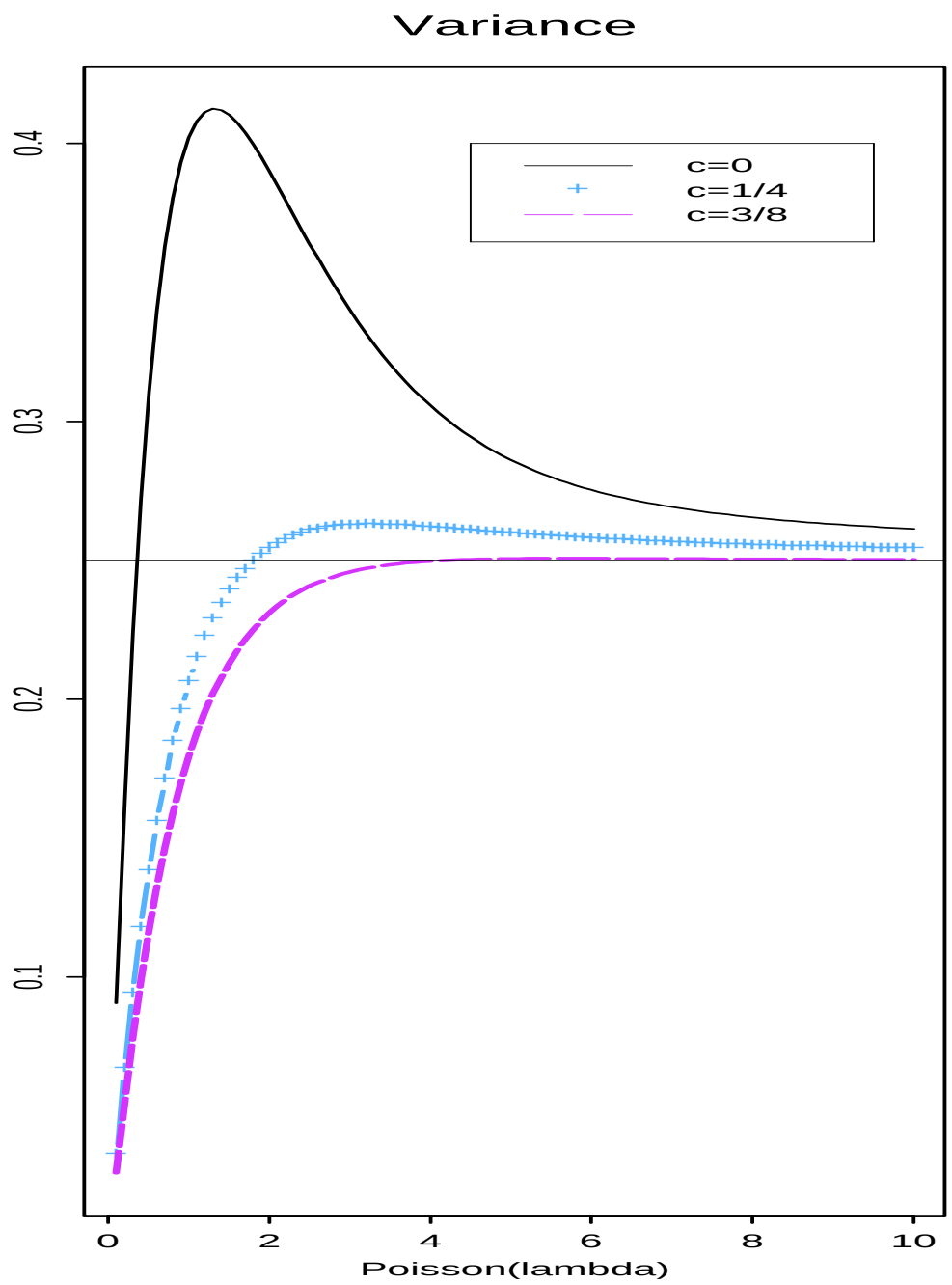
The optimal constants for the **mean-matching VST** are

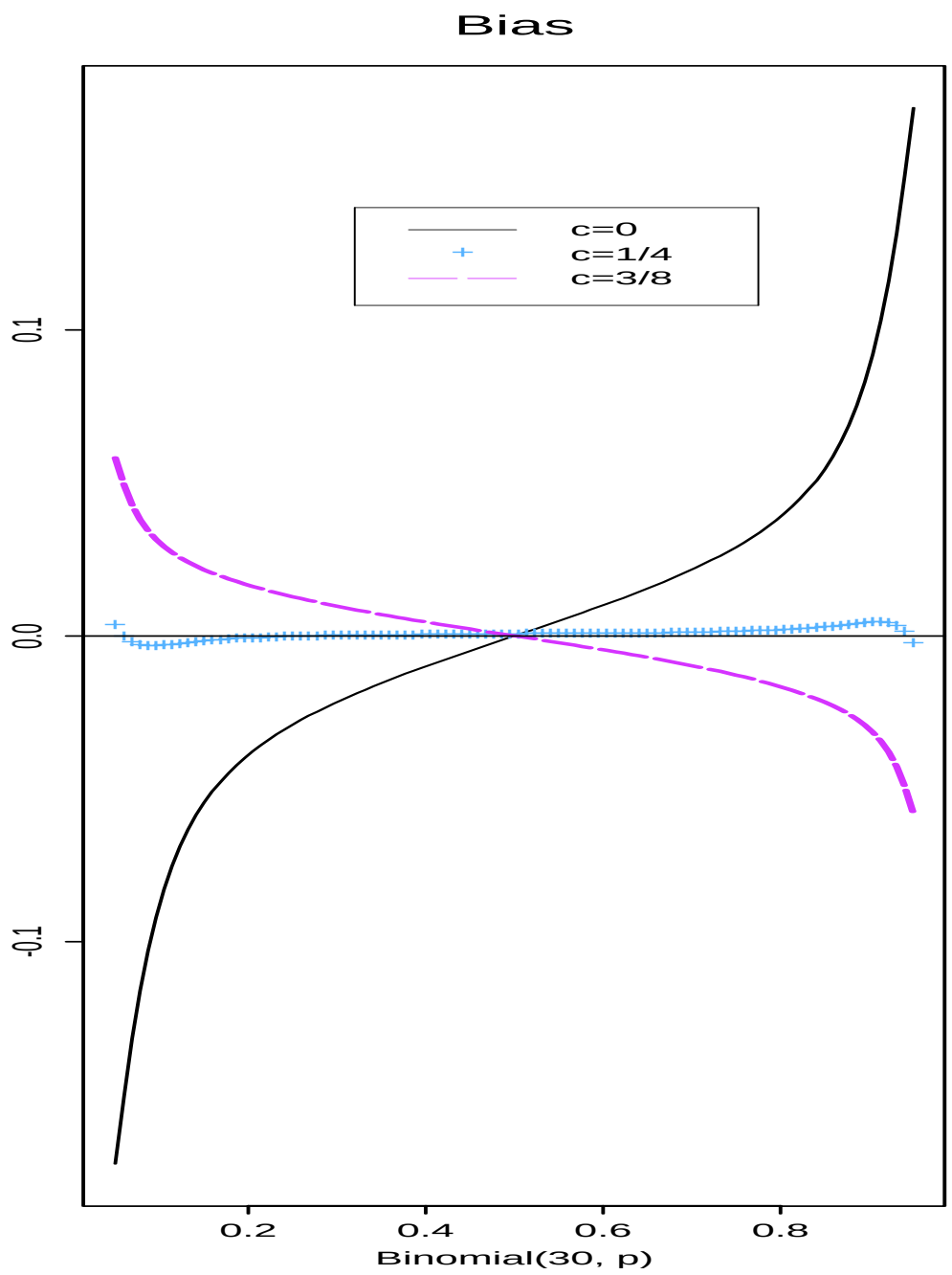
$$a = \frac{1}{4}v_1 \text{ and } b = -\frac{1}{2}v_2. \quad (11)$$

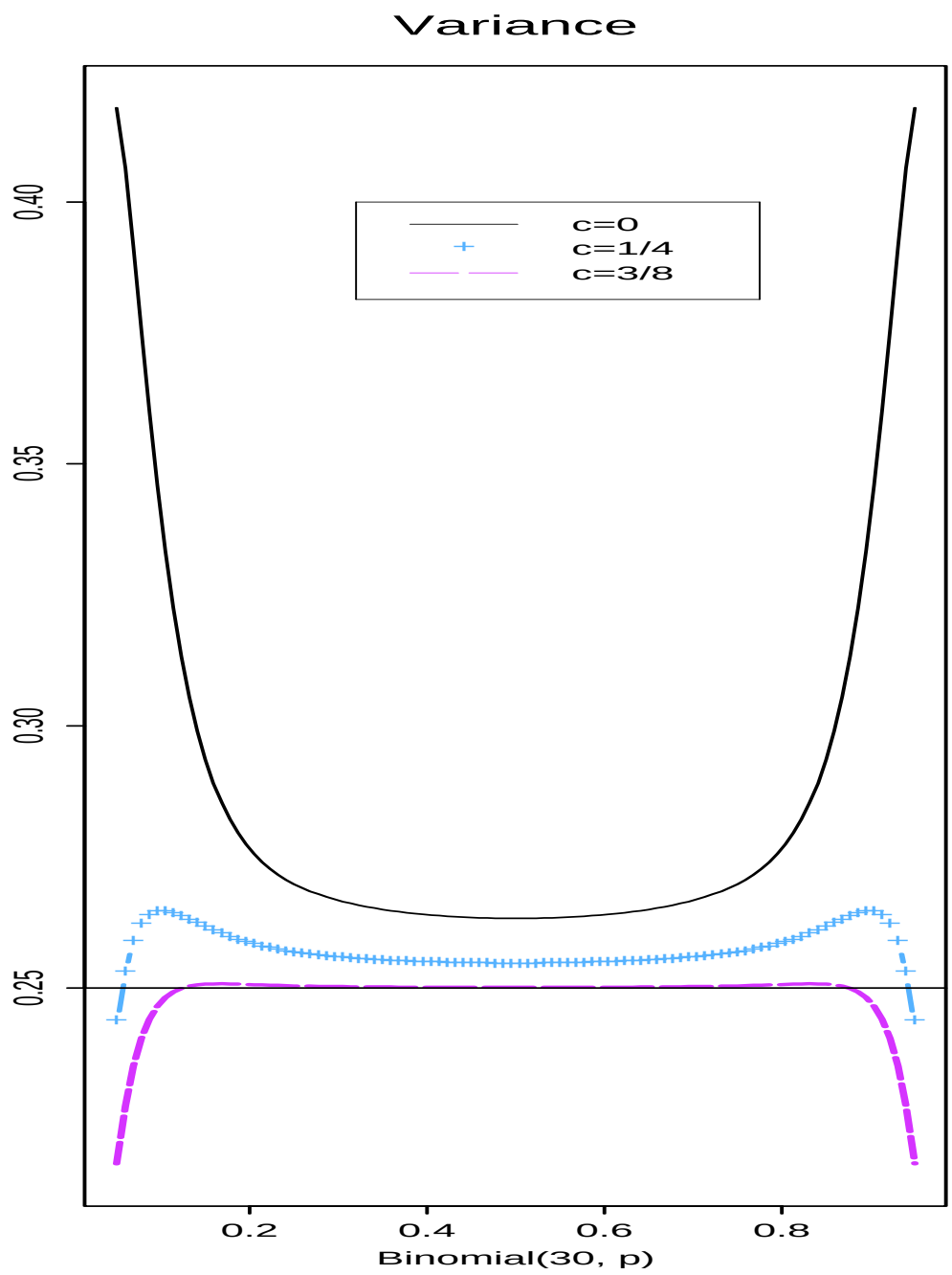
# Mean-Matching VST

- Poisson: $H(X) = \sqrt{(X + \frac{1}{4})/n}$.
- Binomial: $H(X) = \arcsin\left(\sqrt{\frac{X+1/4}{n+1/2}}\right)$.
- Negative Binomial: $H(X) = \ln\left(\sqrt{\frac{X+1/4}{n-1/2}} + \sqrt{1 + \frac{X+1/4}{n-1/2}}\right)$.
- Gamma$(r, \lambda)$ (with $r$ known): $H(X) = \ln(\frac{X}{n-(1/2r)})$.
- NEF-GHS$(r, \lambda)$ (with $r$ known): $H(X) = \ln\left(\frac{X}{n-(1/2r)} + \sqrt{1 + \frac{X^2}{(n-(1/2r))^2}}\right)$.


Bias
c=0
c=1/4
c=3/8
0.3
0.2
0.1
0.0
-0.1
-0.2
0
1
2
3
4
Poisson(lambda)


Variance
c=0
c=1/4
c=3/8
0.4
0.3
0.2
0.1
0
2
4
6
8
10
Poisson(lambda)


Bias
c=0
c=1/4
c=3/8
0.1
0.0
-0.1
0.2
0.4
0.6
0.8
Binomial(30, p)


Variance
c=0
c=1/4
c=3/8
0.40
0.35
0.30
0.25
0.2
0.4
0.6
0.8
Binomial(30, p)

## Nonparametric Regression in Exp. Families

Observe

$$Y_i \overset{ind.}{\sim} NQ(\mu(t_i)), \quad i = 1, ..., n, \; t_i = \frac{i}{n}$$

and wish to estimate the mean function $\mu(t)$.

Examples:

- Poisson Regression: $Y_i \overset{ind.}{\sim} \text{Poisson}(\lambda(t_i))$.
- Binomial Regression: $Y_i \overset{ind.}{\sim} \text{Binomial}(r, \; p(t_i))$.
- Exponential Regression: $Y_i \overset{ind.}{\sim} \text{Exponential}(\lambda(t_i))$.

# The Procedure

1. **Binning**: Group the observations in $T$ equilength bins. Let $m \equiv n/T$ and set

$$N_j = \sum_{(j-1)m+1 \le i \le jm} Y_i.$$

2. **MM-VST**: Let

$$Y_j^* = H(N_j).$$

When $m$ is large, then

$$\boxed{\mathbf{Y_j^* \dot{\sim} N(G(\mu(\frac{j}{T})),\ \frac{1}{m})}}$$

or equivalent

$$Y_j^* \approx G(\mu(\frac{j}{T})) + \frac{1}{\sqrt{m}} Z_j, \quad Z_j \overset{iid}{\sim} N(0,1), \quad j = 1, \cdots, T.$$

## The Procedure (Cont.)

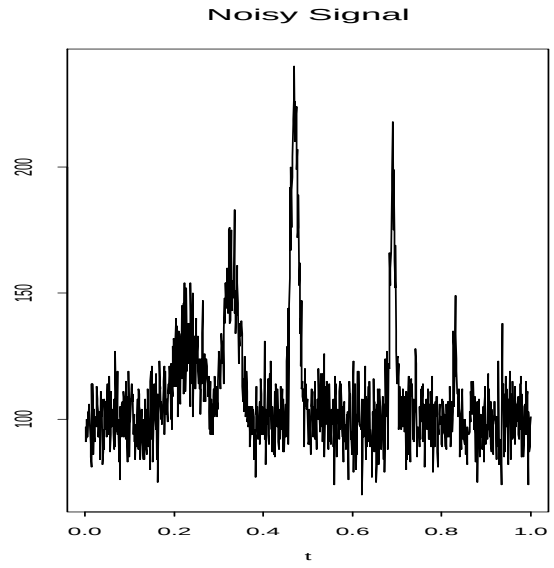
1. **Gaussian Regression:** Apply your favorite Gaussian regression procedure to the binned and transformed data $Y^*$ to obtain an estimator $\widehat{G(\mu(\cdot))}$ of $G(\mu(\cdot))$.
2. **Inverse VST:** Estimate $\mu(t)$ by

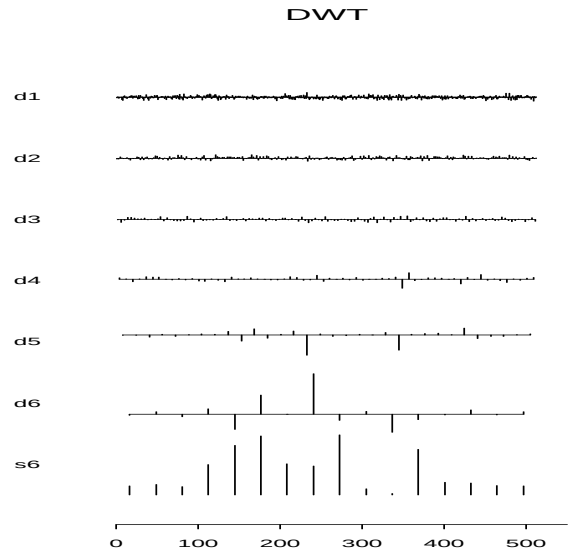
$$\hat{\mu}(t) = G^{-1}(\widehat{G(\mu(t))}).$$

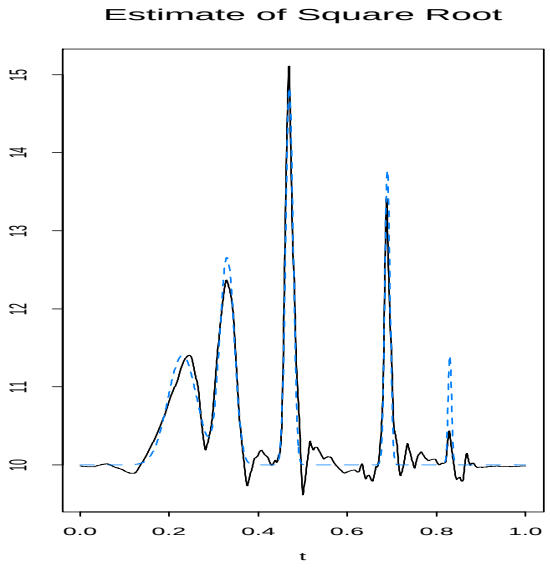
## Nonparametric Regression in Exp. Families

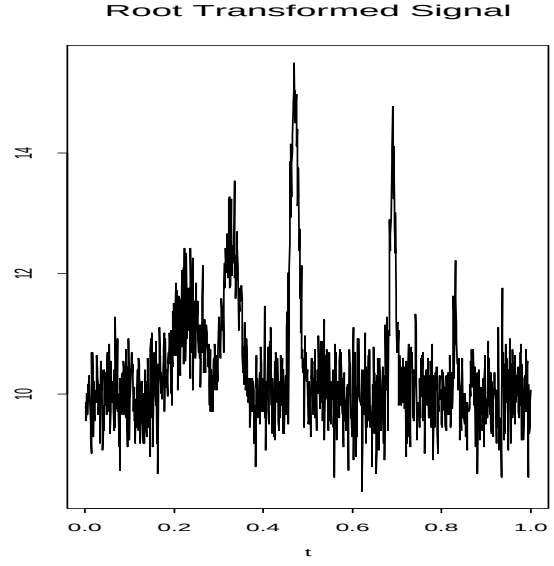
1. **Binning & Mean Matching VST**
2. **BlockJS**
3. **Inverse VST**

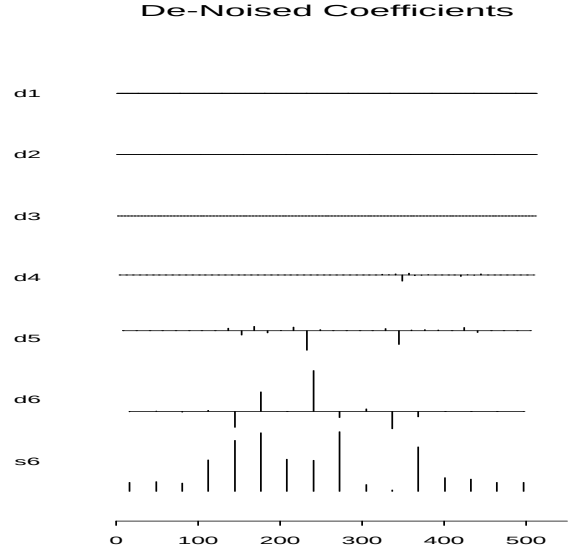
# Example: Poisson Regression

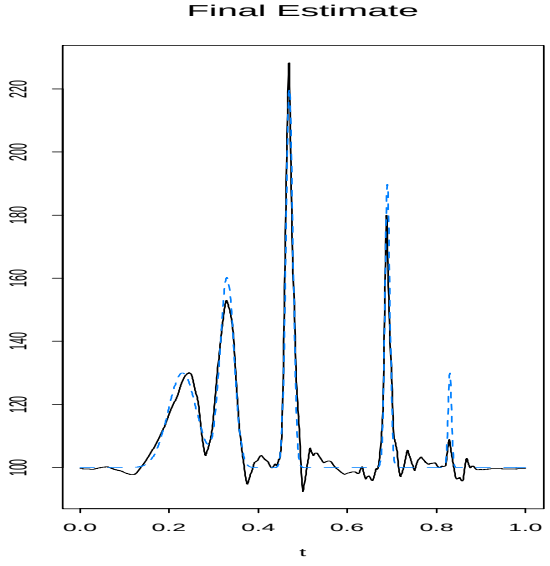


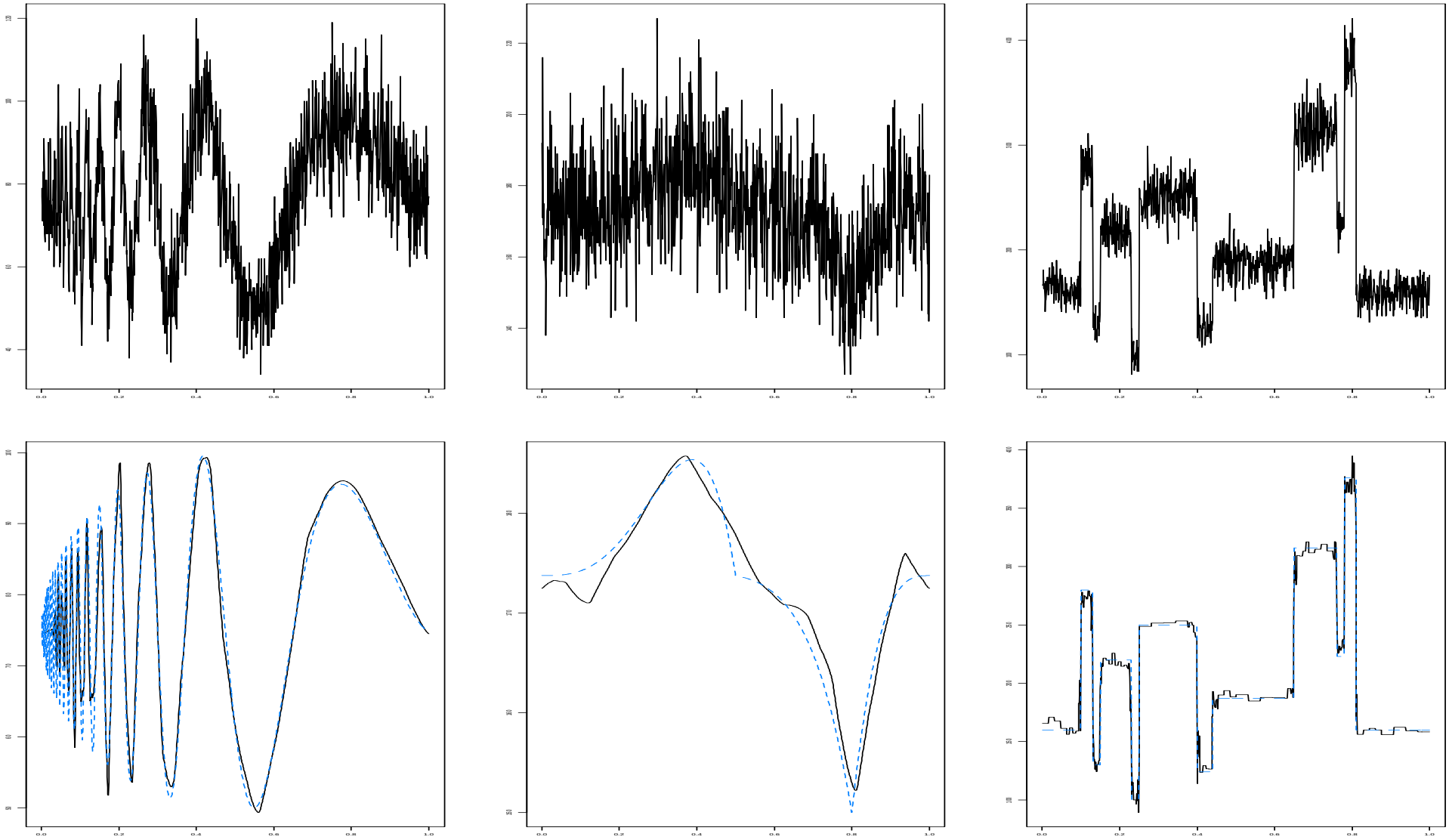



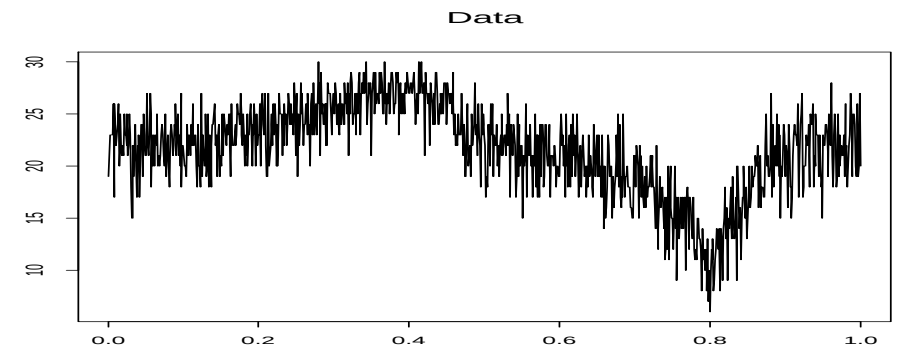



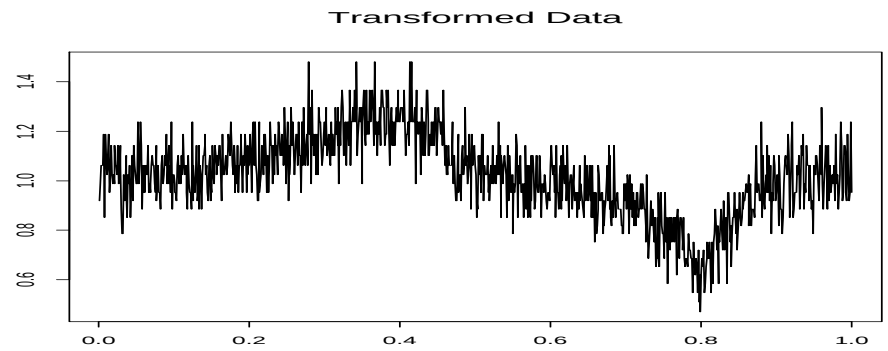



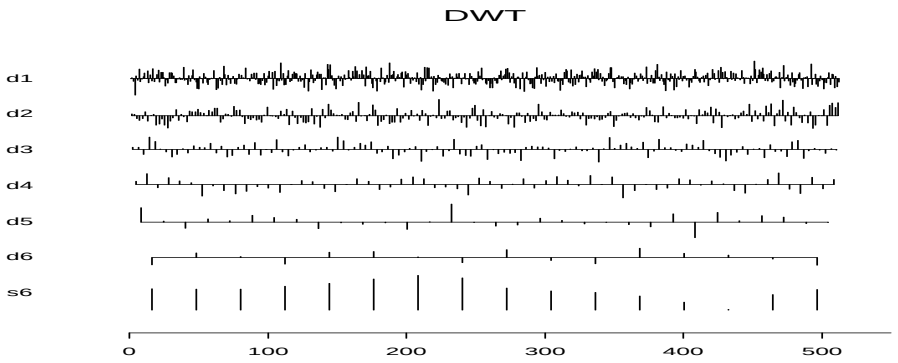
# Example: Poisson Regression

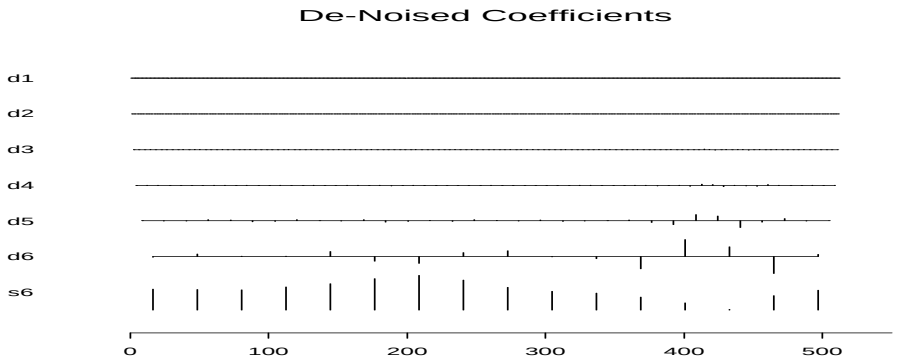


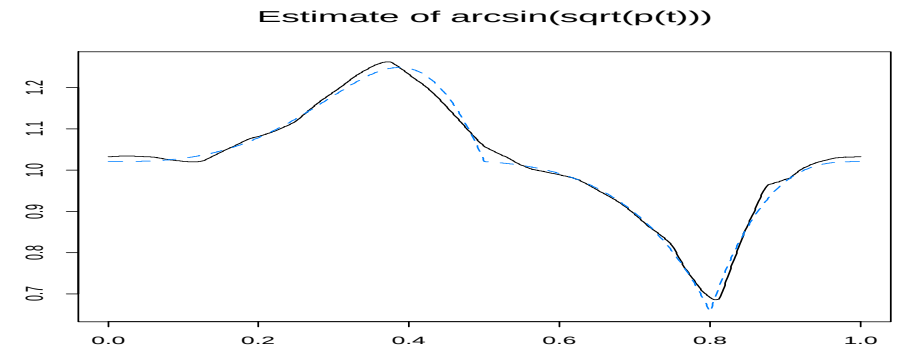
# Example: Binomial Regression

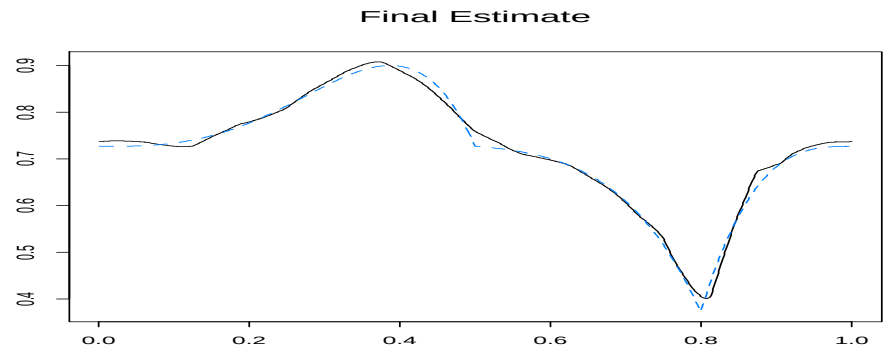


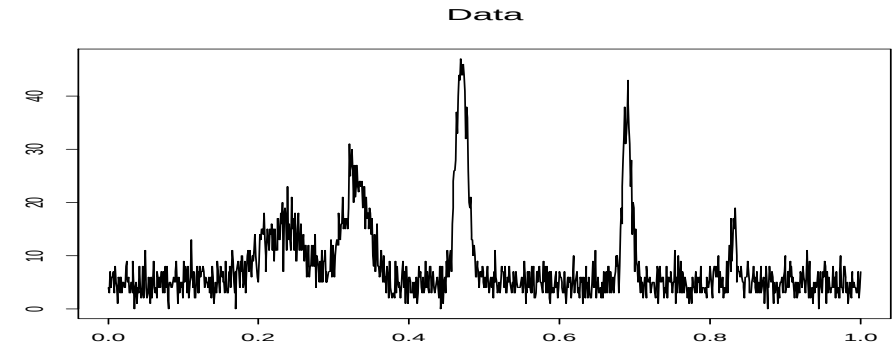



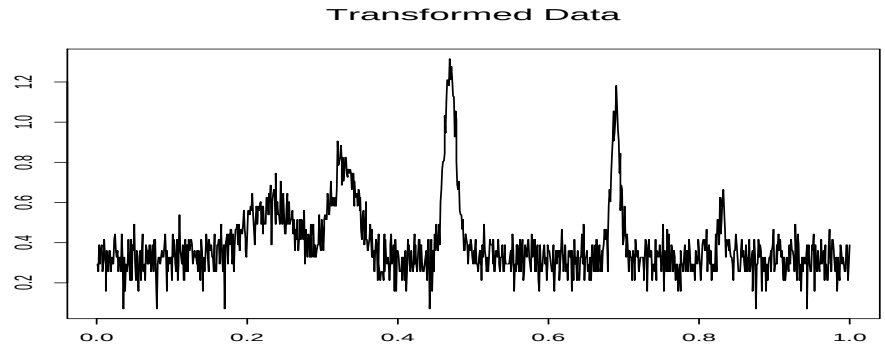



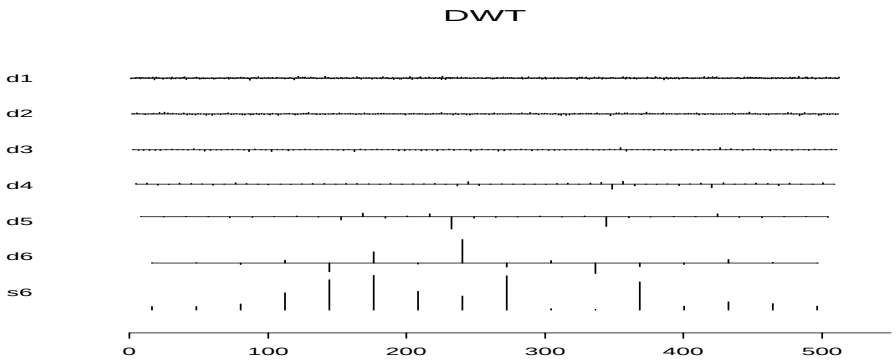



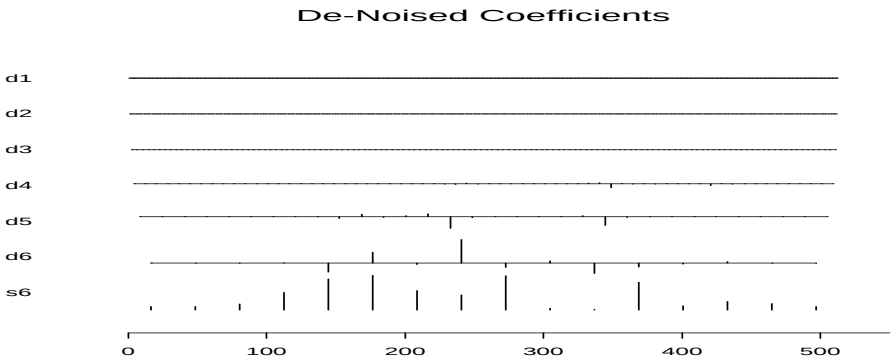



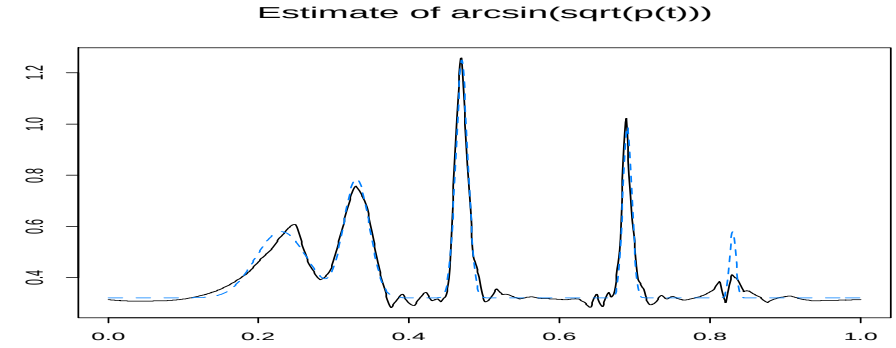
# Example: Binomial Regression

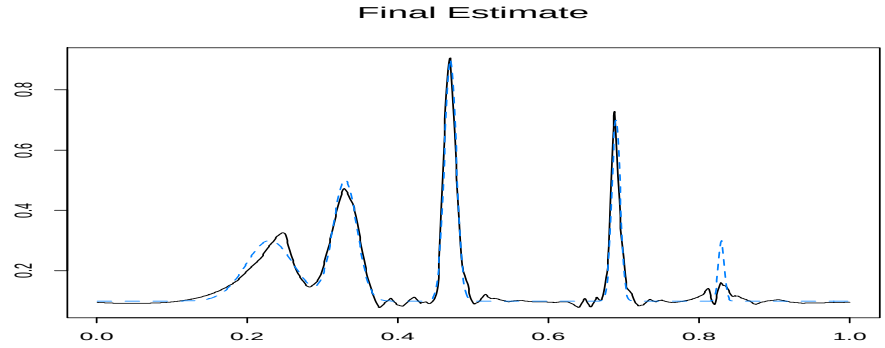


# Global & Local Adaptivity

Define the function class

$$F^{\alpha}_{p,q}(M,\epsilon) = \{f : f \in B^{\alpha}_{p,q}(M), f(x) \geq \epsilon \text{ for all } x \in [0,1]\}. \tag{12}$$

**Theorem 3** *Let* $X_i \sim NQ(\mu(t_i))$, $i = 1, ..., n$, $t_i = \frac{i}{n}$. *Let* $T = Cn^{\frac{3}{4}}$. *Then*

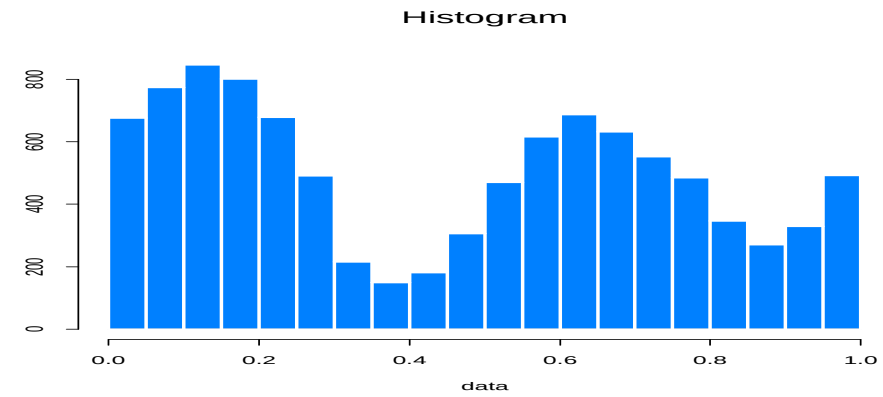
$$\sup_{\mu \in F^{\alpha}_{p,q}(M,\epsilon)} E\|\hat{\mu} - \mu\|_2^2 \leq \begin{cases} Cn^{-\frac{2\alpha}{1+2\alpha}} & p \geq 2,\ \frac{2\alpha^2 - \alpha/3}{1+2\alpha} - \frac{1}{p} > 0 \\ Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}} & 1 \leq p < 2,\ \frac{2\alpha^2 - \alpha/3}{1+2\alpha} - \frac{1}{p} > 0. \end{cases}$$

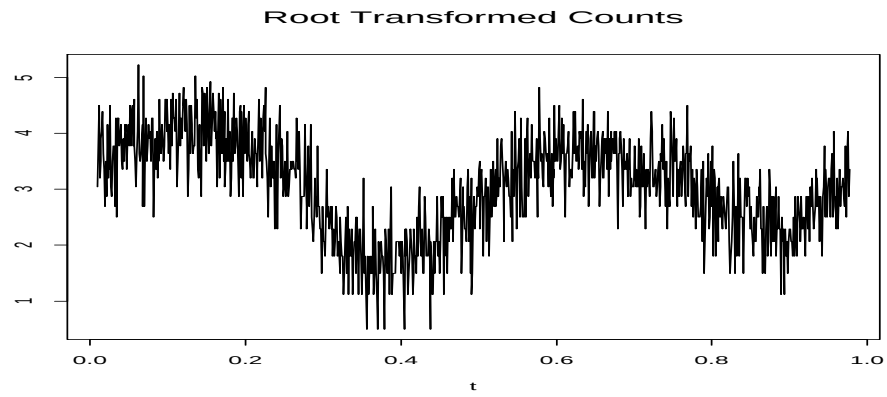
**Theorem 4** *Let* $t_0 \in (0,1)$ *be fixed and let* $\alpha > 1/6$. *Then*

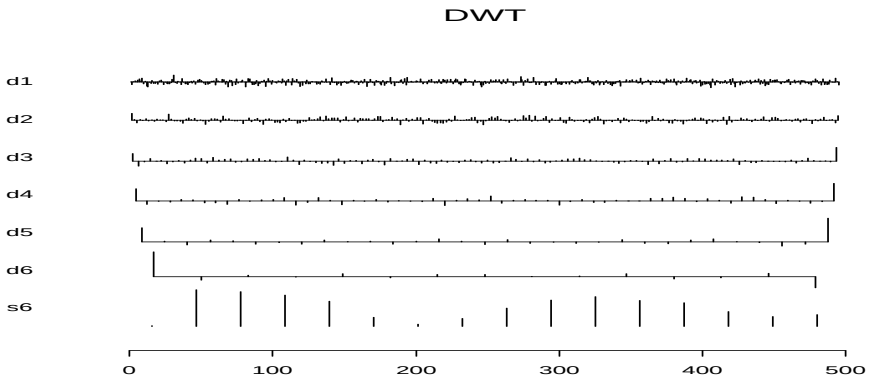
$$\sup_{\mu \in \mathrm{Lip}_{\alpha}(M;t_0)} E(\hat{\mu}(t_0) - \mu(t_0))^2 \leq C \cdot (\frac{\log n}{n})^{\frac{2\alpha}{1+2\alpha}}. \tag{13}$$

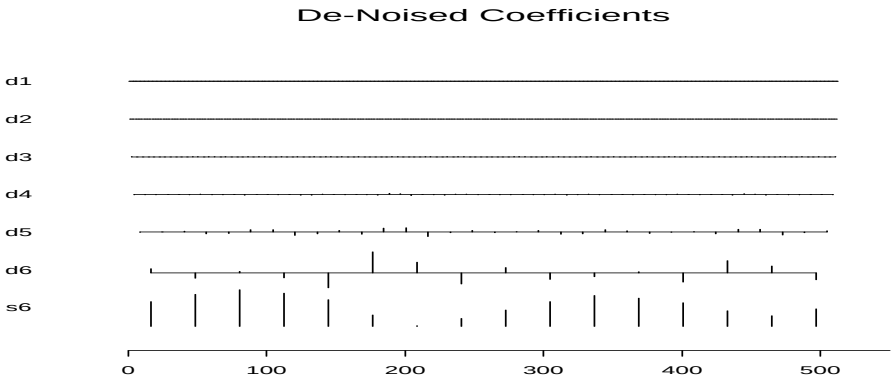
## Density Estimation

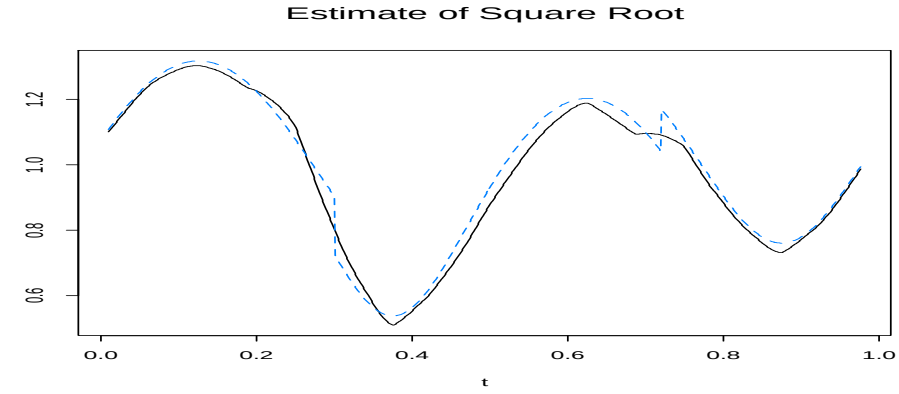
- Density estimation can be treated in a similar way. $\mathbf{X_1, ..., X_n} \overset{\mathbf{iid}}{\sim} \mathbf{f}$.
  1. **Binning**: Bin the observations into $m$ groups, $\mathbf{N_j = \#\{X_i : X_i \in G_j\}}$.
  2. **Root Transform**: $\mathbf{Y_j = \sqrt{\frac{N_j + 1/4}{m}}}$.
  3. **BlockJS**
  4. **Unroot Transform**

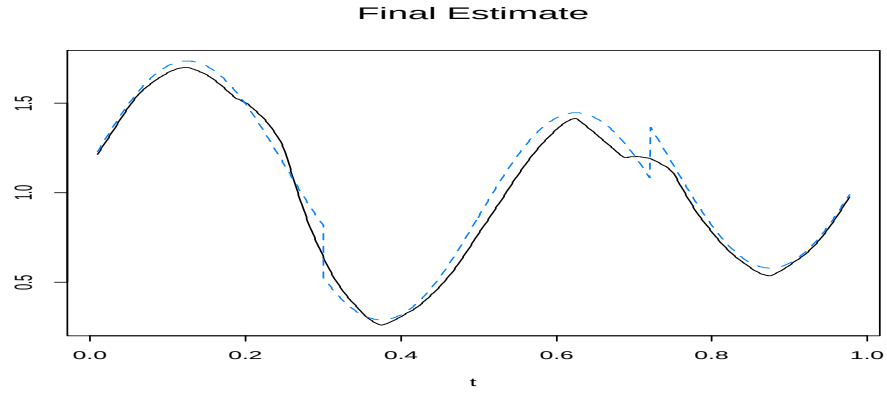

Histogram
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
data


Root Transformed Counts
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
t


DWT
d1
d2
d3
d4
d5
d6
s6
0
100
200
300
400
500


De-Noised Coefficients
d1
d2
d3
d4
d5
d6
s6
0
100
200
300
400
500


Estimate of Square Root
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
t


Final Estimate
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
t

# Summary


'Nonstandard'
Nonparametric
Regression
Transformation
Standard
Gaussian
Regression

## Papers


Brown, L. D., Cai, T. & Zhou, H. (2008). **Robust nonparametric estimation via wavelet median regression.** *Ann. Statist.* **36**, 2055-2084.

Cai, T. & Zhou, H. (2009). **Asymptotic equivalence and adaptive estimation for robust nonparametric regression.** *Ann. Statist.* **37**, in press.

Brown, L. D., Cai, T. & Zhou, H. (2009). **Nonparametric regression in exponential families.** Technical Report.


**Available at: http://stat.wharton.upenn.edu/∼tcai**