*Statistical Science*
2019, Vol. 34, No. 4, 635–656
https://doi.org/10.1214/19-STS718
© Institute of Mathematical Statistics, 2019

# Gaussianization Machines for Non-Gaussian Function Estimation Models

**T. Tony Cai**

*Abstract.* A wide range of nonparametric function estimation models have been studied individually in the literature. Among them the homoscedastic nonparametric Gaussian regression is arguably the best known and understood. Inspired by the asymptotic equivalence theory, Brown, Cai and Zhou (*Ann. Statist.* **36** (2008) 2055–2084; *Ann. Statist.* **38** (2010) 2005–2046) and Brown et al. (*Probab. Theory Related Fields* **146** (2010) 401–433) developed a unified approach to turn a collection of non-Gaussian function estimation models into a standard Gaussian regression and any good Gaussian nonparametric regression method can then be used.

These Gaussianization Machines have two key components, binning and transformation. When combined with BlockJS, a wavelet thresholding procedure for Gaussian regression, the procedures are computationally efficient with strong theoretical guarantees. Technical analysis given in Brown, Cai and Zhou (*Ann. Statist.* **36** (2008) 2055–2084; *Ann. Statist.* **38** (2010) 2005–2046) and Brown et al. (*Probab. Theory Related Fields* **146** (2010) 401–433) shows that the estimators attain the optimal rate of convergence adaptively over a large set of Besov spaces and across a collection of non-Gaussian function estimation models, including robust nonparametric regression, density estimation, and nonparametric regression in exponential families. The estimators are also spatially adaptive.

The Gaussianization Machines significantly extend the flexibility and scope of the theories and methodologies originally developed for the conventional nonparametric Gaussian regression. This article aims to provide a concise account of the Gaussianization Machines developed in Brown, Cai and Zhou (*Ann. Statist.* **36** (2008) 2055–2084; *Ann. Statist.* **38** (2010) 2005–2046), Brown et al. (*Probab. Theory Related Fields* **146** (2010) 401–433).

*Key words and phrases:* Adaptivity, asymptotic equivalence, block thresholding, density estimation, exponential family, mean matching, nonparametric function estimation, quadratic variance function, quantile coupling, robust regression, variance stabilizing transformation, wavelets.

## 1. INTRODUCTION

Motivated by a wide range of applications, many nonparametric function estimation models such as Gaussian regression, density estimation, Poisson regression, and binomial regression have been considered separately in the literature. Among these function estimation models, the standard nonparametric regression with additive homoscedastic Gaussian noise, where one observes

$$Y_i = f\left(\frac{i}{n}\right) + \xi_i, \quad i = 1, \dots, n \tag{1}$$

with $\xi_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, is perhaps the best studied and understood. For example, on the theoretical side, much

*Tony Cai is Daniel H. Silberberg Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: tcai@wharton.upenn.edu).*

This paper is dedicated to the memory of Larry Brown, a great scholar, a mentor and a friend.

work has been done on developing minimax theories and adaptation theories for global and local estimation, confidence sets, and hypothesis testing. On the methodological side, significant efforts have been made on the construction of adaptive inference procedures that automatically adjust to the unknown smoothness of the underlying function and achieve optimality simultaneously over a large set of parameter spaces. In particular, wavelet thresholding procedures have been developed and shown to achieve significant successes in terms of adaptivity, spatial adaptivity and computational efficiency in the context of the standard nonparametric Gaussian regression. See, for example, Tsybakov (2009) and Johnstone (2011).

Inspired by the asymptotic equivalence theory, in a series of papers, Brown, Cai and Zhou (2008, 2010) and Brown et al. (2010) developed a unified approach to adaptive estimation for a collection of non-Gaussian function estimation models, including nonparametric Poisson regression, density estimation, regression in exponential families with a quadratic variance function, and robust nonparametric regression with an unknown and potentially heavy-tailed error distribution. This approach turns these non-Gaussian function estimation models in a unified way into the standard nonparametric regression with additive homoscedastic Gaussian noise. Then in principle any good nonparametric Gaussian regression method can be used to solve these more complicated problems and many of the theoretical results developed for the standard Gaussian regression can be carried over as well. Since the methods "gaussianize" non-Gaussian models, we call them *Gaussianization Machines* in the present paper. These Gaussianization Machines were developed further in Cai and Zhou (2009) for robust regression with symmetric error distributions and in Cai and Zhou (2010) for nonparametric regression in general natural exponential families.

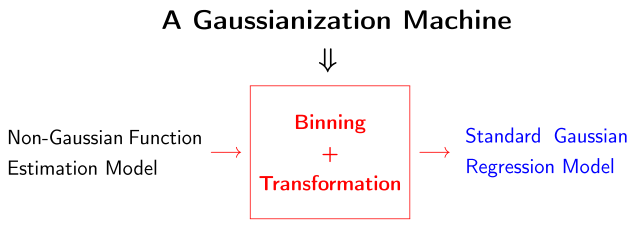
A Gaussianization Machine
⇓
Non-Gaussian Function Estimation Model
→
Binning
+
Transformation
→
Standard Gaussian Regression Model

FIG. 1. *A Gaussianization Machine converts non-Gaussian function estimation problems into the standard nonparametric Gaussian regression through binning and transformation.*

A Gaussianization Machine, which is illustrated in Figure 1, has two key components, binning and transformation. The Gaussianization Machine first groups the data into small bins and then applies a local transformation to each bin. The binning step is essentially the same for all these non-Gaussian problems considered, but the transformation step is model specific. In the case of density estimation and nonparametric regression in the natural exponential families with a quadratic variance function, the transformation is the mean-matching variance stabilizing transformation (MM-VST), and for robust nonparametric regression, it is the local median transformation. The binned and transformed data can then be treated as if they were generated from the standard nonparametric regression with additive homoscedastic Gaussian noise and any good Gaussian regression procedure such as a wavelet thresholding estimator can be applied.

The Gaussianization Machines significantly extend the flexibility and scope of the theories and methodologies originally developed for the conventional nonparametric Gaussian regression. The goal of the present paper is to give a concise account of the Gaussianization Machines for these non-Gaussian function estimation problems. The connections as well as differences among these problems will be discussed.

### 1.1 Asymptotic Equivalence

Asymptotic equivalence theory, pioneered by Lucien Le Cam with the early focus on parametric models, provides a deep understanding of the fundamental connections among seemingly different statistical models. The main goal of the asymptotic equivalence theory is to approximate complicated statistical models by simple ones so that the study of the complex model can be essentially simplified. For example, optimal procedures and theoretical results developed for the simple model can be carried over to the complex ones.

Brown and Low (1996) was the first to establish the global asymptotic equivalence result for nonparametric function estimation models. In this seminal paper, asymptotic equivalence between the homoscedastic nonparametric Gaussian regression with equispaced design and the white noise with drift model is established. Since then there has been significant efforts on establishing asymptotic equivalence among different nonparametric function estimation models. Many interesting and important results have been obtained. Global asymptotic equivalence theory has been developed in a wide range of settings, including nonparametric density estimation and Poisson process in Nussbaum (1996) and Brown et al. (2004), nonparametric regression with random design in Brown et al.

(2002), nonparametric regression with a known error distribution in Grama and Nussbaum (2002), generalized linear models in Grama and Nussbaum (2002), nonparametric autoregression in Milstein and Nussbaum (1998), diffusion models in Delattre and Hoffmann (2002) and Genon-Catalot, Laredo and Nussbaum (2002), GARCH model in Brown, Wang and Zhao (2003), nonparametric autoregression in Grama and Neumann (2006), scalar ergodic diffusions in Dalalyan and Reiß (2006) and multidimensional ergodic diffusions in Dalalyan and Reiß (2007), nonparametric regression with multivariate and random design in Reiß (2008), spectral density estimation in Golubev, Nussbaum and Zhou (2010), inference on the volatility from noisy observations in Reiß (2011), and between functional linear regression and a white noise inverse problem in Meister (2011). In addition, results on asymptotic nonequivalence have also been developed in Efromovich and Samarov (1996), Brown and Zhang (1998), and Wang (2002).

Although asymptotic equivalence theory provides deep theoretical insights into various statistical models and is intuitively appealing, it does have several drawbacks that limit its usefulness in practice. One is that the equivalence mappings typically require randomizations and so are not practical in applications. Another is that asymptotic equivalence results mostly focus on bounded loss functions so are not applicable in general in many common settings where losses such as the squared error loss are unbounded. In addition, full asymptotic equivalence in Le Cam's sense is a very stringent goal and often the failures occur only in some pathological cases which do not commonly arise in many applications of interest.

### 1.2 The Gaussianization Machines

Instead of pursuing full asymptotic equivalence in Le Cam's sense, significant efforts have been made to develop deterministic and practical algorithms using the ideas from the equivalence theory to convert a range of non-Gaussian function estimation problems into a standard homoscedastic Gaussian regression problem, which has been well understood in the literature. In this paper, we focus on a unified approach originally developed in Brown, Cai and Zhou (2008) for robust nonparametric regression, Brown et al. (2010) for nonparametric density estimation, and Brown, Cai and Zhou (2010) for nonparametric regression in exponential families with a quadratic variance function, which includes, for example, nonparametric Poisson regression, binomial regression, and Gamma regression as special cases.

The main ideas behind the Gaussianization Machines can be most easily explained using the example of nonparametric Poisson regression. In this case, one observes

$$Y_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(\lambda\left(\frac{i}{n}\right)\right), \quad i = 1, \dots, n,$$

and wishes to estimate the intensity function $\lambda(t)$. It is well known that the usual variance stabilizing transformation (VST) for Poisson distribution is the root transform. As will be discussed in Section 2, the MM-VST in this case (and for density estimation) is the root transform with a correction of $1/4$. This is a key step in the Gaussianization Machine. The Gaussianization Machine algorithm for estimating the intensity function $\lambda(t)$ can be summarized in the following four simple steps.

**Algorithm 1** A Gaussianization Machine for Poisson Regression

```
1: Binning: Divide the indices $\{1, \dots, n\}$ into $T$ equal sized groups $I_1, \dots, I_T$ with $m$ consecutive indices each. Let $Q_j = \sum_{i \in I_j} Y_i$, $j = 1, \dots, T$.
2: MM-VST: Let $Y_j^* = \sqrt{(Q_j + \frac{1}{4})/m}$, $j = 1, \dots, T$. Then treat $Y^* = \{Y_1^*, \dots, Y_T^*\}$ as a new sample from a homoscedastic Gaussian regression with equispaced design.
3: Standard Gaussian Regression: Apply your favorite Gaussian regression procedure to the binned and transformed data $Y^*$ to obtain an estimate $\widehat{\sqrt{\lambda(\cdot)}}$ of $\sqrt{\lambda(\cdot)}$.
4: Inverse Transformation: Estimate the intensity function $\lambda(\cdot)$ by $\widehat{\lambda(\cdot)} = (\widehat{\sqrt{\lambda(\cdot)}})^2$.
```

This algorithm also applies to nonparametric density estimation without essential changes. For other distributions in the natural exponential families with a quadratic variance function, the only difference is the MM-VST in Step 2 and the corresponding inverse transformation in Step 4. For example, for nonparametric binomial regression, the MM-VST in Step 2 is the arcsine transformation with a suitable correction and Step 4 is the sine transformation.

A key component in the Gaussianization Machine outlined in Algorithm 1 is the MM-VST for the natural exponential families. The advantage of the MM-VST over the classical VST is that it reduces the bias due to the transformation up to a certain level while still stabilizing the variance. Here the bias reduction (or equivalently mean-matching) is crucial. However, as shown in Brown, Cai and Zhou (2010), for the one-parameter natural exponential families, MM-VST exists only for those with a quadratic variance function. Cai and Zhou (2010) extends the scope of Brown, Cai and Zhou (2010) to the general natural exponential families where the conventional VST exists but the MM-VST may not. A new explicit procedure based on the usual VST is proposed. This approach significantly reduces the bias of the inverse transformation and as a consequence it enables a general Gaussianization Machine to be applicable to a wider class of exponential families. The drawback of this approach is that the inverse transformation may not have a closed form. See Section 4 for more discussions.

In robust nonparametric regression, one observes a signal with i.i.d. additive noise as in (1), but the distribution of the errors $\xi_i$ is unknown and potentially heavy tailed. Here we assume $\text{median}(\xi_i) = 0$ and the mean of $\xi_i$ may not exist. The goal is to recover the median function $f$. In such a setting, a direct application of a standard Gaussian regression method to the data $\{Y_1, \dots, Y_n\}$ could fail badly. A similar Gaussianization Machine is built in Brown, Cai and Zhou (2008) to turn this problem into a standard Gaussian regression problem. The main difference is that Step 2 in Algorithm 1 is replaced by a local median transformation where $Y_j^* = \text{median}(Y_i : i \in I_j)$ with $I_j$ being the index set of the observations in the $j$th bin. For general error distributions, Step 4 in Algorithm 1 is replaced by a simple bias correction step. The Gaussianization Machine algorithm for robust nonparametric regression also has four simple steps.

**Algorithm 2** A Gaussianization Machine Algorithm for Robust Regression

```
1: Binning: Divide the indices $\{1, \dots, n\}$ into $T$ equal sized groups $I_1, \dots, I_T$ with $m$ consecutive indices each.
2: Median Transformation: Let $Y_j^* = \text{median}(Y_i : i \in I_j)$, $j = 1, \dots, T$. Then $Y_j^* \dot\sim N(g(\frac{j}{T}), \sigma^2)$ with $g(\cdot) = f(\cdot) + b_m$ and $\sigma^2 = \frac{1}{4mh^2(0)}$, where $h$ is the density function of $\xi_i$ and $b_m = \mathbb{E}\{\text{median}(\xi_1, \dots, \xi_m)\}$.
3: Standard Gaussian Regression: Apply your favorite nonparametric regression procedure to the binned and transformed data $\{Y_1^*, \dots, Y_T^*\}$ to obtain an estimate $\hat{g}$.
4: Bias Correction: The median function $f$ is then estimated by $\hat{f}(t) = \hat{g}(t) - \hat{b}_m$ where $\hat{b}_m$ is an estimator of the bias $b_m$.
```

See Section 5 for a detailed discussion, including the construction of an estimate $\hat{b}_m$ for the bias $b_m$.

Cai and Zhou (2009) further extends this approach to robust nonparametric regression with symmetric error distributions. When the error distribution is known to be symmetric, the algorithm can be simplified as $b_m = 0$ so Step 4 can be eliminated. Furthermore, much refined theoretical results can be established. See Section 5 for more discussions.

### 1.3 Technical Insights

A key question is: Why do the Gaussianization Machines work? That is, why the binned and transformed data can be treated as if they were observations from a homoscedastic Gaussian regression model? The critical technical engine underpinning the Gaussianization Machines is the quantile coupling inequalities. This is again most easily explained in the case of nonparametric Poisson regression.

The Gaussianization Machine for nonparametric Poisson regression consists of two steps: Binning and MM-VST. In the binning steps, each $Q_j$ in the first step of Algorithm 1 is the sum of independent Poisson observations in the $j$th bin. So $Q_j$ is itself a Poisson variable with mean $\lambda_{Q_j} = \sum_{i \in I_j} \lambda(\frac{i}{n})$, which grows with the bin size $m$ under some mild assumption on the intensity function $\lambda(\cdot)$. It then follows from the Central Limit Theorem (CLT) that $Q_j$ is asymptotically normal. However, the CLT is not sufficient for the technical analysis here. What is needed is a tight bound for the quantile coupling between a Poisson variable and a normal variable. This is provided by a coupling inequality given in Brown et al. (2010). The mean-matching property of the MM-VST together with the coupling inequality for Poisson variables show that under mild conditions, the binned and transformed data $Y_j^*$ can be viewed as

$$Y_j^* \approx \sqrt{\bar{\lambda}_j} + \frac{1}{2} Z \quad \text{with } Z \sim N(0, 1),$$

where $\bar{\lambda}_j = T \int_{\frac{j-1}{T}}^{\frac{j}{T}} f(x)\,dx$ is the average of $\lambda(t)$ over the $j$th subinterval, which becomes a standard nonparametric Gaussian regression problem.

The analysis above also applies to nonparametric density estimation, after a Poissonization argument is used. The quantile coupling inequality for Poisson variables can be extended to the natural exponential families with a quadratic variance function. See Section 6 for further discussions and see Brown et al. (2010), Brown, Cai and Zhou (2010) for a detailed technical analysis. The readers are also referred to Mason and Zhou (2012) for more on quantile coupling.

For robust nonparametric regression, the key step in the Gaussianization Machine is the local median transformation. The critical technical tool is a quantile coupling for the sample median, which shows precisely how well the sample median can be approximated by a normal variable. It is an analog of the coupling inequality for the sample mean given in Komlós, Major and Tusnády (1975). See Section 6 and Brown, Cai and Zhou (2008) for more discussions.

After binning and transformation in Steps 1 and 2, the data can be treated as if they were observations from the standard Gaussian regression model. Although in principle any good nonparametric Gaussian regression method can be used in Step 3, BlockJS, a wavelet block thresholding procedure proposed in Cai (1999), was use in Brown, Cai and Zhou (2008, 2010) and Brown et al. (2010) for illustration. We shall also use the same wavelet procedure in our discussion in this paper.

When combined with the BlockJS procedure for Gaussian regression, the Gaussianization Machine algorithms are computationally efficient with strong asymptotic properties. Theoretical analysis given in Brown, Cai and Zhou (2008, 2010) and Brown et al. (2010) shows that the procedures adaptive attain the optimal rate of convergence over a wide collection of the Besov spaces, without prior knowledge of the smoothness of the underlying functions across the collection of these non-Gaussian function estimation models, including robust nonparametric regression, density estimation, and nonparametric regression in exponential families. The estimator also automatically adapts to the local smoothness of the underlying function, and attains the local adaptive minimax rate for estimating functions at a point. Essentially, the optimality results for the standard nonparametric Gaussian regression carry over to this collection of non-Gaussian problems.

### 1.4 Organization and Notation

The rest paper is organized as follows. Section 2 presents a detailed description of the Gaussianization Machine for nonparametric Poisson regression. This case provides the essential insights into the general principles behind the approach. We also introduce the BlockJS procedure for nonparametric Gaussian regression. Section 3 considers nonparametric density estimation. The procedure for Poisson regression can be applied to density estimation directly without any essential changes. Section 4 generalizes the Gaussianization Machine for Poisson regression to treat nonparametric regression in the natural exponential families. Section 5 considers the Gaussianization Machine for robust nonparametric regression. Key technical insights and theoretical properties are discussed in Section 6. The paper is concluded with a discussion in Section 7.

## 2. NONPARAMETRIC POISSON REGRESSION

In this section, we consider nonparametric Poisson regression where one observes

$$X_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(\lambda\left(\frac{i}{n}\right)\right), \quad i = 1, \dots, n,$$

and wishes to estimate the intensity function $\lambda(t)$, which is assumed to be a smooth function.

Nonparametric Poisson regression is of significant interest in its own right. Regression with count data arises in a range of applications, see, for example, Ver Hoef and Boveng (2007), Winkelmann (2003), Berk and MacDonald (2008), Kroll (2019). The Poisson regression model is one of the most natural approaches to count data regression. Besbeas, De Feis and Sapatinas (2004) provided a review of the literature on the nonparametric Poisson regression and carried out an extensive numerical comparison of several estimation procedures including Donoho (1993), Kolaczyk (1999a, 1999b), Fryzlewicz and Nason (2004). As will be seen later, nonparametric Poisson regression can also be viewed as a prototypical model for nonparametric density estimation as well as a special case of nonparametric regression in natural exponential families with a quadratic variance function.

For Poisson regression, the noise is non-additive, non-Gaussian, and heteroscedastic. Applying a standard Gaussian regression method directly to the data in general does not yield desirable results. One strategy is to turn this problem into a standard Gaussian regression problem through a Gaussianization Machine, which as mentioned in the Introduction has two key components, binning and MM-VST. We begin by discussing the MM-VST and then introduce in detail the Gaussianization Machine for nonparametric Poisson regression.

### 2.1 Mean-Matching Variance Stabilizing Transformation

Variance stabilizing transformation (VST) has been used in many statistical applications. For Poisson distributions, Bartlett (1947) was the first to introduce the root transform $\sqrt{X}$ in a homoscedastic linear model where $X \sim \text{Poisson}(\lambda)$. For $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$, the sample mean $\bar{X}$ satisfies

$$\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \xrightarrow{L} N\left(0, \frac{1}{4}\right).$$

In the context of nonparametric Poisson regression, the vanilla root transform results in too much bias and it is necessary to consider a more general form of the root transform.

Lemma 1 (Brown et al., 2010). *Let $X \sim \text{Poisson}(\lambda)$ with $\lambda > 0$ and let $c \geq 0$ be a constant. Then*

$$\begin{aligned}\mathbb{E}(\sqrt{X+c}) &= \lambda^{\frac{1}{2}} + \frac{4c-1}{8} \cdot \lambda^{-\frac{1}{2}} \\ &\quad - \frac{16c^2 - 24c + 7}{128} \cdot \lambda^{-\frac{3}{2}} + O(\lambda^{-\frac{5}{2}}),\end{aligned} \tag{2}$$

$$\begin{aligned}\text{Var}(\sqrt{X+c}) &= \frac{1}{4} + \frac{3-8c}{32} \cdot \lambda^{-1} \\ &\quad + \frac{32c^2 - 52c + 17}{128} \cdot \lambda^{-2} + O(\lambda^{-3}).\end{aligned} \tag{3}$$

It is clear from Lemma 1 that the choice of $c = \frac{3}{8}$ yields optimal variance stabilization. This is Anscombe's VST proposed in Anscombe (1948). In comparison to variance stabilization, mean matching, in the sense of making the expectation of $\sqrt{X+c}$ as close to $\sqrt{\lambda}$ as possible, is more important for the nonparametric Poisson regression problem we consider here. Lemma 1 shows that the choice of $c = \frac{1}{4}$ is optimal for mean-matching while stabilizing the variance. We shall call the mapping $x \mapsto \sqrt{x + \frac{1}{4}}$ *mean-matching variance stabilizing transformation* (MM-VST).

The effect of the constant $c$ in the root transform on the mean and variance can also be easily seen empirically. In Figure 2, the left panel plots the bias $\mathbb{E}_\lambda(\sqrt{X+c}) - \sqrt{\lambda}$ as a function of $\lambda$ for $c = 0$, $c = \frac{1}{4}$ and $c = \frac{3}{8}$. It is clear that the choice of $c = \frac{1}{4}$ is the best among the three for mean-matching. With $c = \frac{1}{4}$ the bias is negligible for $\lambda$ as small as 2. The right panel shows the variance of $\sqrt{X+c}$ for $c = 0$, $c = \frac{1}{4}$ and $c = \frac{3}{8}$. In this case, $c = \frac{3}{8}$ is the best choice. The choice of $c = \frac{1}{4}$ is slightly worse than but comparable to the case with $c = \frac{3}{8}$. For both mean-matching and variance stabilization, $c = 0$ is clearly the worst choice of the three.

### 2.2 Estimation Procedure

We now return to the nonparametric Poisson regression problem where we wish to estimate the intensity function $\lambda(\cdot)$ based on the observations $X_i \overset{\text{ind.}}{\sim} \text{Poisson}(\lambda(\frac{i}{n}))$, $i = 1, \dots, n$. As mentioned earlier, the first step of the Gaussianization Machine is binning. Let $T \asymp n^{3/4}$ be some positive integer. We begin by dividing the indices $\{1, \dots, n\}$ into $T$ nonoverlapping and equal sized groups $I_1, \dots, I_T$ with $m = n/T$ consecutive indices each. Let $Q_j$ be the sum of observations in the $j$th bin $I_j$,

$$Q_j = \sum_{i \in I_j} Y_i, \quad j = 1, \dots, T. \tag{4}$$

The sums $Q_j$ can be treated as observations for a Gaussian regression directly, but this in general leads to a heteroscedastic problem. Instead, we apply the MM-VST and let

$$Y_j^* = \sqrt{\left(Q_j + \frac{1}{4}\right)/m}, \quad j = 1, \dots, T. \tag{5}$$

The transformed data $Y^* = (Y_1^*, \dots, Y_T^*)$ is then treated as a new sample for a Gaussian nonparametric regression with equispaced design and in principle any good Gaussian nonparametric regression method can be applied to construct an estimate $\widehat{\sqrt{\lambda(\cdot)}}$. The intensity function $\lambda(\cdot)$ is estimated by $(\widehat{\sqrt{\lambda(\cdot)}})^2$. The four-step estimation procedure has been summarized in Algorithm 1 in Section 1.

### 2.3 BlockJS for the Standard Gaussian Regression

In this paper we use a wavelet thresholding procedure, BlockJS, proposed in Cai (1999) for the standard Gaussian regression in Step 3 of Algorithm 1 as an illustration. BlockJS makes simultaneous decisions to keep or kill all the coefficients within a block and increases estimation accuracy by utilizing information about neighboring coefficients.

Let $\{\phi, \psi\}$ be a pair of compactly supported father and mother wavelets with $\int \phi = 1$. Denote $\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)$ and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$. For simplicity, we work with periodized wavelet bases on

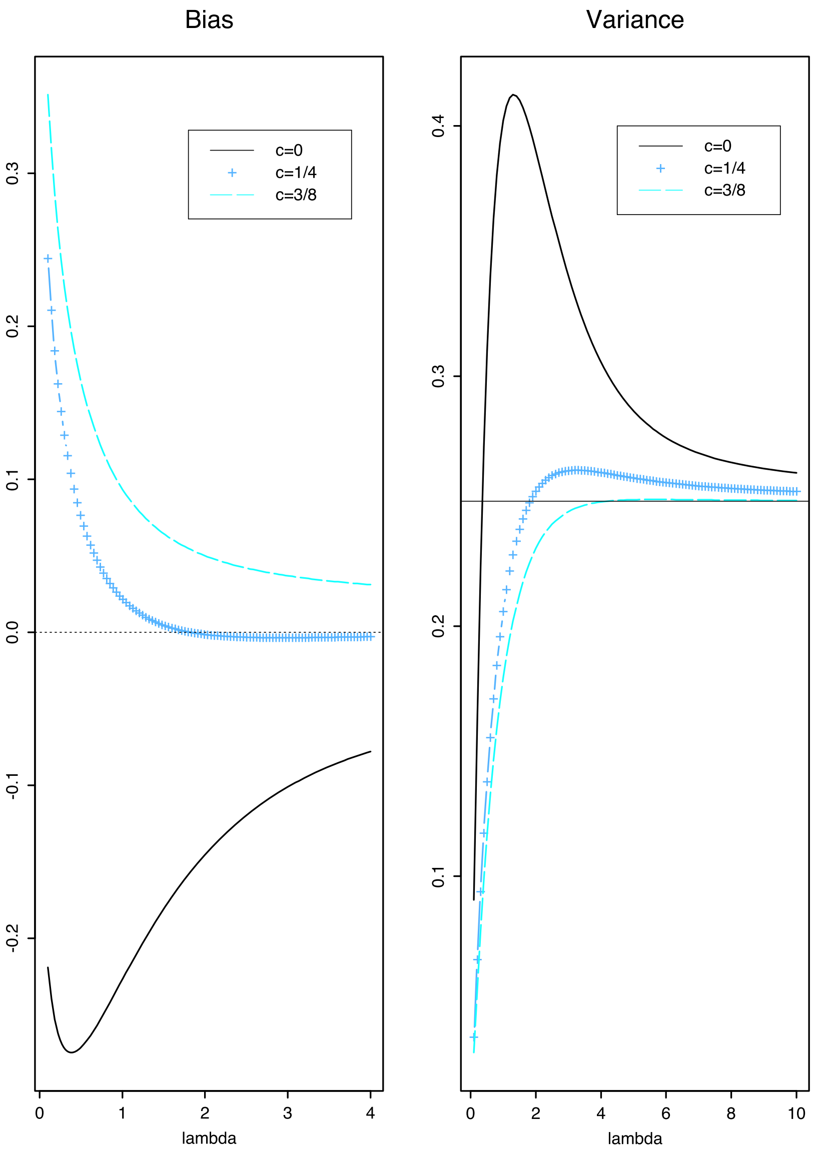

FIG. 2. *Comparison of the bias (left panel) and variance (right panel) of root transforms $x \mapsto \sqrt{x+c}$ with $c = 0, \frac{1}{4}$ and $\frac{3}{8}$. It is clear that $c = \frac{1}{4}$ is optimal for mean-matching and $c = \frac{3}{8}$ is the best for variance stabilization.*

[0, 1] and let

$$\phi^p_{j,k}(t) = \sum_{l=-\infty}^{\infty} \phi_{j,k}(t-l),$$

$$\psi^p_{j,k}(t) = \sum_{l=-\infty}^{\infty} \psi_{j,k}(t-l) \quad \text{for } t \in [0,1].$$

The collection $\{\phi^p_{j_0,k}, k = 1, \dots, 2^{j_0}; \psi^p_{j,k}, j \geq j_0 \geq 0, k = 1, \dots, 2^j\}$ is then an orthonormal basis of $L^2[0,1]$, provided the primary resolution level $j_0$ is suitably chosen. The superscript "$p$" will be suppressed from the notation for convenience. An orthonormal wavelet basis has an associated orthogonal Discrete Wavelet Transform (DWT) which transforms sampled data into the wavelet coefficients. See, for example, Daubechies (1992) for details on wavelets and DWT. A square-integrable function $f$ on [0, 1] can be expanded into a wavelet series:

$$f(t) = \sum_{k=1}^{2^{j_0}} \tilde{\theta}_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{j,k}\psi_{j,k}(t), \tag{6}$$

where $\tilde{\theta}_{j,k} = \langle f, \phi_{j,k}\rangle$, $\theta_{j,k} = \langle f, \psi_{j,k}\rangle$ are the wavelet coefficients of $f$.

BlockJS was proposed in Cai (1999) for nonparametric Gaussian regression. Let the sample $Y = (Y_1, \ldots, Y_n)^{\mathsf{T}}$ be given as in (1) with $\xi_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ and $n = 2^J$ for some integer $J$. Apply DWT to the data $Y$ and let $U = n^{-\frac{1}{2}} WY$ be the empirical wavelet coefficients, where $W$ is the discrete wavelet transformation matrix. Then $U$ can be written as

$$\begin{aligned} U = (&\tilde{u}_{j_0,1}, \ldots, \tilde{u}_{j_0,2^{j_0}}, u_{j_0,1}, \ldots, \\ &u_{j_0,2^{j_0}}, \ldots, u_{J-1,1}, \ldots, u_{J-1,2^{J-1}})^{\mathsf{T}}. \end{aligned} \tag{7}$$

Here $\tilde{u}_{j_0,k}$ are the gross structure terms at the lowest resolution level, and $u_{j,k}$ for $k = 1, \ldots, 2^j$ and $j_0 \le j \le J - 1$ are the empirical wavelet coefficients. Note that $u_{j,k}$ can be viewed as

$$\begin{aligned} u_{j,k} = \theta_{j,k} & \\ &+ z_{j,k}, \quad j_0 \le j \le J-1, k = 1, \ldots, 2^j, \end{aligned} \tag{8}$$

where $\theta_{j,k}$ are approximately the wavelet coefficients of the regression function $f$ and $z_{j,k} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2/n)$ are the noise. Divide each resolution level $j_0 \le j \le J - 1$ into nonoverlapping blocks of length $L = [\log n]$ (or $L = 2^{\lfloor \log_2(\log n) \rfloor} \approx \log n$). Let $B_j^i = \{(j,k) : (i-1)L + 1 \le k \le iL\}$ denote the $i$th block at level $j$ and let $S_{j,i}^2 \equiv \sum_{(j,k) \in B_j^i} y_{j,k}^2$. For $(j,k) \in B_j^i$, $\theta_{j,k}$ is estimated by a James–Stein type shrinkage estimate

$$\hat{\theta}_{j,k} = \begin{cases} \left(1 - \dfrac{\lambda_* L \sigma^2}{n S_{j,i}^2}\right)_+ y_{j,k} & \text{for } (j,k) \in B_j^i, \\ & \quad j_0 \le j < J, \\ 0 & \text{for } j \ge J, \end{cases} \tag{9}$$

where $\lambda_* = 4.50524$ is a constant chosen according to an oracle inequality and a minimax criterion to satisfy $\lambda_* - \log \lambda_* - 3 = 0$. The estimate of the regression function $f$ is given by

$$\begin{aligned} \hat{f}(t|Y) = &\sum_{k=1}^{2^{j_0}} \tilde{u}_{j_0,k} \phi_{j_0,k}(t) \\ &+ \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{j,k} \psi_{j,k}(t). \end{aligned} \tag{10}$$

Here we write the estimate of $f$ as $\hat{f}(t|Y)$ to emphasize the dependence of the estimator $\hat{f}$ on the input sample $Y$. BlockJS is easily implementable and enjoys a high degree of adaptivity and spatial adaptivity. See Cai (1999) for details.

### 2.4 A Wavelet Procedure for Poisson Regression

We now return to nonparametric Poisson regression. Set $J = J_n = \lfloor \log_2 n^{3/4} \rfloor$ and let $T = 2^J \asymp n^{3/4}$. Applying BlockJS to the binned and root transformed data $Y^* = (Y_1^*, \ldots, Y_T^*)^{\mathsf{T}}$ to obtain the estimate of the function $\sqrt{\lambda}(\cdot)$,

$$\begin{aligned} \widehat{\sqrt{\lambda}}(t|Y^*) = &\sum_{k=1}^{2^{j_0}} \hat{\tilde{\theta}}_{j_0,k} \phi_{j_0,k}(t) \\ &+ \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{j,k} \psi_{j,k}(t). \end{aligned} \tag{11}$$

The final estimator of the intensity function $\lambda$ is given by the square of $\widehat{\sqrt{\lambda}}$:

$$\begin{aligned} \hat{\lambda}(t) = &\left(\sum_{k=1}^{2^{j_0}} \hat{\tilde{\theta}}_{j_0,k} \phi_{j_0,k}(t)\right. \\ &\left. + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{j,k} \psi_{j,k}(t)\right)^2. \end{aligned} \tag{12}$$

The estimator $\hat{\lambda}(t)$ given in (12) performs well numerically. Figure 3 illustrates the estimation procedure using BlockJS. The estimator also enjoys a high degree of adaptivity and spatial adaptivity with near optimal asymptotic performance over a large collection of Besov spaces. There are two key technical tools in the theoretical analysis. One is a coupling inequality to approximate the binned and root transformed data by independent normal variables and another is a risk bound for block thresholding in the case where the noise is not necessarily Gaussian. See more discussions in Section 6.

## 3. NONPARAMETRIC DENSITY ESTIMATION

Nonparametric density estimation, which aims to recover the underlying density function based on an i.i.d. sample $\{X_1, \ldots, X_n\}$, is one of the most fundamental problems in data analysis and has been extensively studied in statistics. See, for example, Silverman (1986).

Nonparametric density estimation is traditionally considered separately from regression. Using the same Gaussianization Machine, density estimation can be converted into a standard homoscedastic Gaussian regression problem. A key step in understanding the procedure is Poissonization, which turns nonparametric density estimation into nonparametric Poisson regression. Then the Gaussianization Machine described in

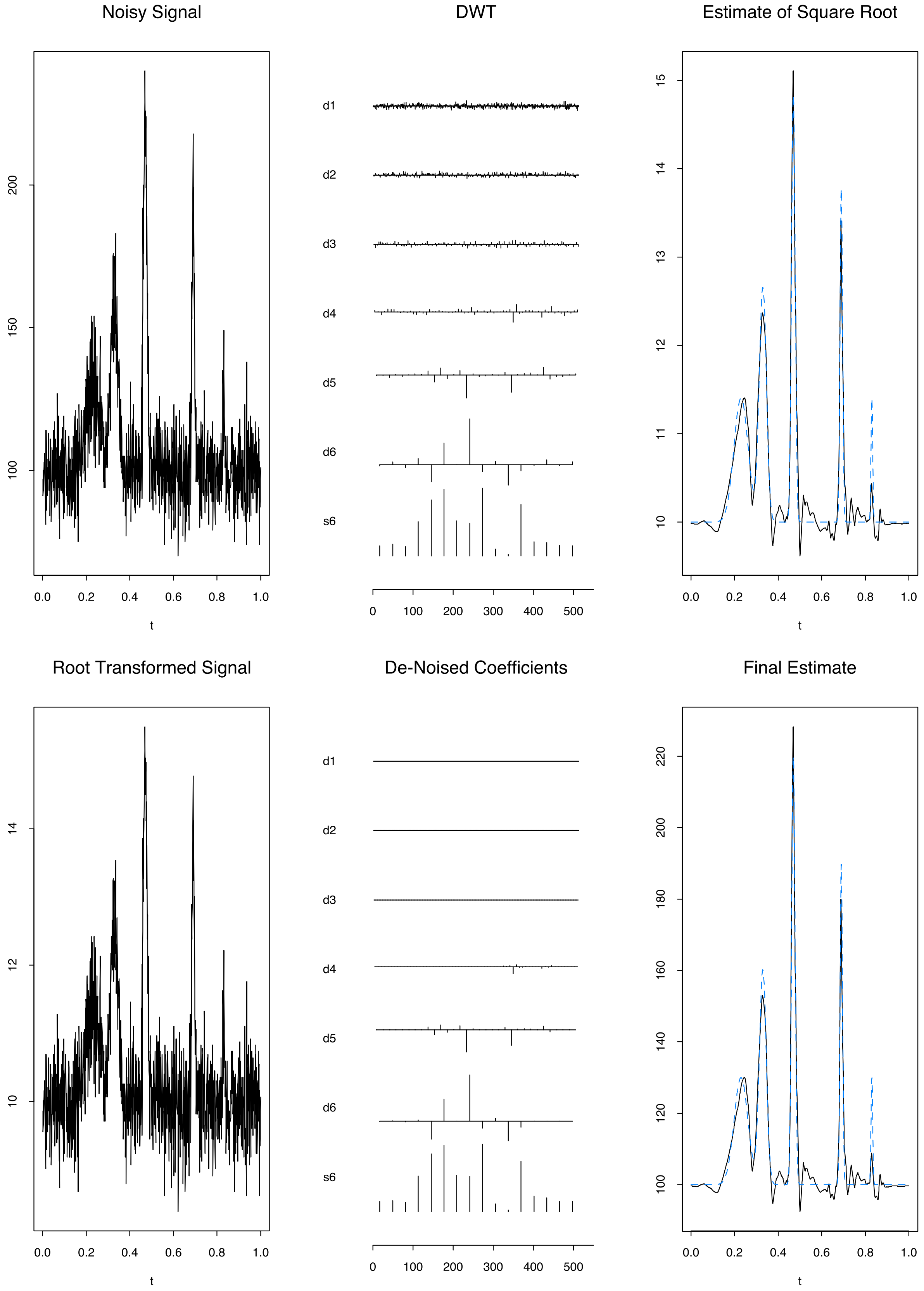

FIG. 3. *The Gaussianization Machine for nonparametric Poisson regression as implemented via BlockJS.*

Section 2 applies. In essence, density estimation can be solved in the exactly same way as nonparametric Poisson regression is solved via binning and MM-VST as described in the last section.

In the density experiment, for a given sample size $n$ one generates an i.i.d. sample $\{X_1, \dots, X_n\}$ from some distribution $F$ with the density function $f$. For a fixed $n$, in the Poissonized density experiment one first draws a Poisson random variable $N \sim \text{Poisson}(n)$ and then generates an i.i.d. sample $\{X_1, \dots, X_N\}$ from the same distribution. It is shown in Low and Zhou (2007) that these two experiments are asymptotically equivalent under regularity conditions. Poissonization shows that the problem of estimating a density with the fixed sample size is not essentially different from estimating the density where the sample size is a Poisson random variable. Given an i.i.d. sample with a fixed sample size $n$, the bin counts jointly have a multinomial distribution and so the counts for different bins are dependent. Poissonization allows one to treat the bin counts as if they were independent Poisson variables. Poissonization has been studied, for example, in Le Cam (1974), Low and Zhou (2007).

We now return to density estimation, where one observes an i.i.d. sample $\{X_1, \dots, X_n\}$ from some distribution $F$ with the density function $f$. For simplicity, we assume the unknown density $f$ is supported on a finite interval, say $[0, 1]$. Given a random sample $\{X_1, \dots, X_n\}$, we first create a histogram with $T$ equal sized bins from 0 to 1. Let $Q_j$ be the number of observations in the $j$th subinterval. Then $\{Q_1, \dots, Q_T\}$ jointly have a multinomial distribution. Note that if the sample size is Poissonized, then the counts $\{Q_1, \dots, Q_T\}$ are independent Poisson random variables with

$$Q_j \sim \text{Poisson}(np_j) \quad \text{where } p_j = T\int_{\frac{j-1}{T}}^{\frac{j}{T}} f(x)\,dx.$$

The same mean-matching root transform can be applied to the $Q_j$'s. Let

$$Y_j^* = \sqrt{Q_j + \frac{1}{4}}, \quad j = 1, \dots, T. \tag{13}$$

Then the binned and transformed data $Y^* = (Y_1^*, \dots, Y_T^*)$ can be treated as a new equispaced sample for a nonparametric Gaussian regression problem. The final density estimator can be obtained by normalizing the square of $\widehat{\sqrt{f}}$. In summary, the Gaussianization Machine leads to the following algorithm for density estimation.

**Algorithm 3** A Gaussianization Machine for Density Estimation

```
1: Binning: Create a histogram with $T$ equal sized bins from 0 to 1. Let $Q_1, \dots, Q_T$ be the number of observations in the bins.
2: Mean-matching Root Transform: Let $Y_j^* = \sqrt{(Q_j + \frac{1}{4})/n}$, $j = 1, \dots, T$, and treat $Y^* = \{Y_1^*, \dots, Y_T^*\}$ as the new equispaced sample for a homoscedastic Gaussian regression problem.
3: Standard Gaussian Regression: Apply your favorite Gaussian regression procedure to the binned and transformed data $Y^*$ to obtain an estimate $\widehat{\sqrt{f}}$ of $\sqrt{f}$.
4: Unroot & Normalization: Let $\tilde{f} = (\widehat{\sqrt{f}})^2$ and estimate the density function $f$ by
       $\hat{f}(t) = \tilde{f}(t) / \int_0^1 \tilde{f}(t)\,dt.$
```

In Step 4, $\hat{f}$ may not integrate to 1 and so the normalization is needed. For Gaussian regression in Step 3, we again use the BlockJS procedure as illustration. Figure 4 shows the steps of the algorithm for density estimation.

The density estimator $\hat{f}$ enjoys essentially the same properties as the estimator of the intensity function in nonparametric Poisson regression discussed in the last section. Given the results for the Poisson regression, the key step in the technical argument for density estimation is Poissonization. See Brown et al. (2010) for a detailed technical analysis.

## 4. NONPARAMETRIC REGRESSION IN NATURAL EXPONENTIAL FAMILIES

Poisson distribution is a member of the natural exponential families. The Gaussianization Machine for nonparametric Poisson regression discussed in Section 2 can be generalized for nonparametric regression in natural exponential families with a quadratic variance function (NEF-QVF), which includes, for example, nonparametric binomial regression and nonparametric exponential regression. These regression problems have been studied separately in the literature. See Brown, Cai and Zhou (2010) for further references and discussions.

As in the Poisson case, for nonparametric regression in the NEF-QVF, the noise is non-additive, non-Gaussian, and heteroscedastic, and applying standard nonparametric regression methods directly to the data

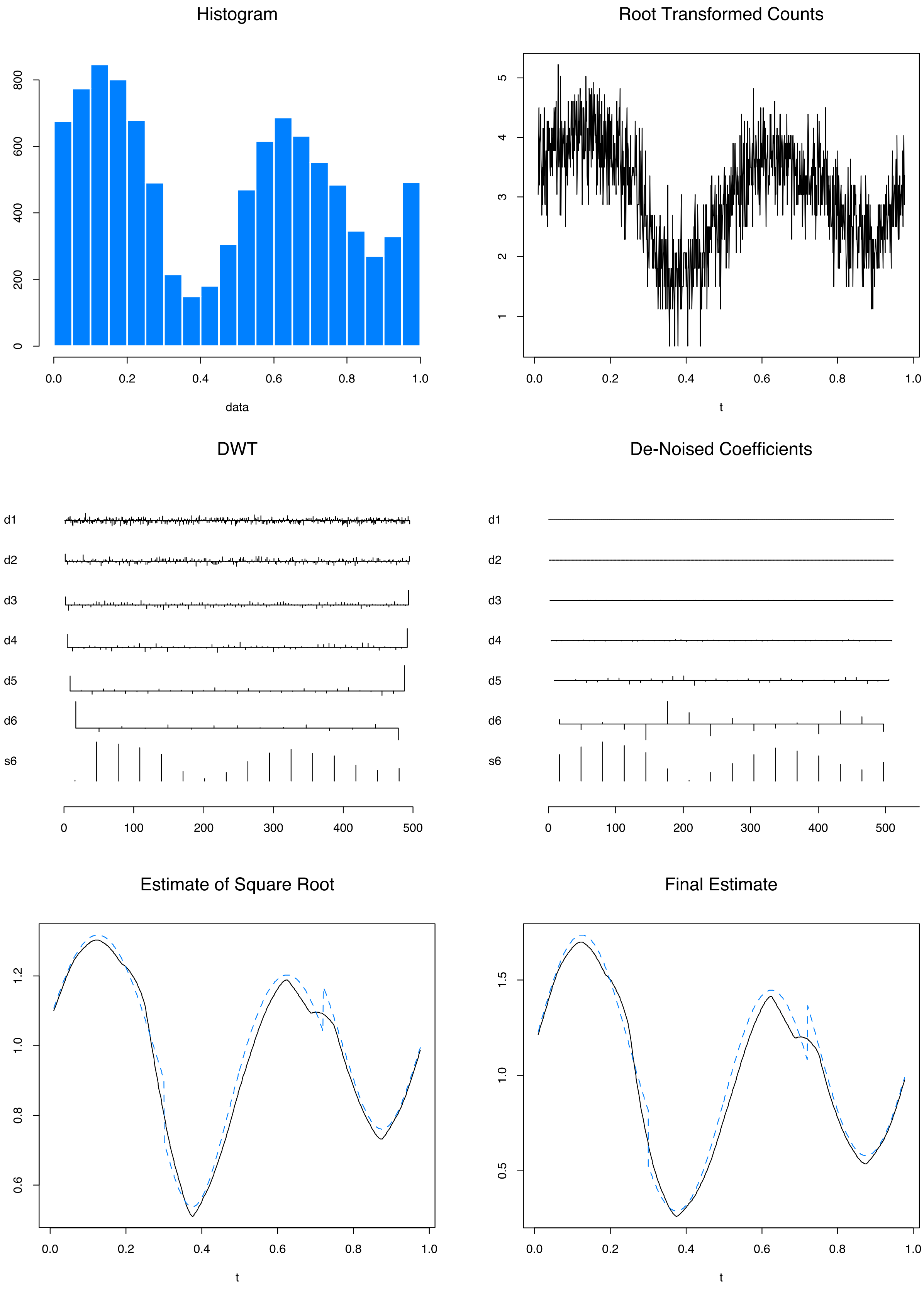

FIG. 4. *The Gaussianization Machine for nonparametric density estimation as implemented via BlockJS.*

in general do not yield desirable results. The Gaussianization Machine designed for the Poisson regression and density estimation can be generalized to turn this problem into a homoscedastic Gaussian regression problem.

We begin by discussing the MM-VST for the natural exponential families (NEF). Conventional VST has been widely used in statistics. See Hoyle (1973) for an extensive review of the literature. The probability density/mass function of a distribution in a one-parameter NEF can be written as $q(x|\eta) = e^{\eta x - \psi(\eta)} h(x)$, where $\eta$ is the natural parameter. The mean and variance are given by

$$\mu(\eta) = \psi'(\eta) \quad \text{and} \quad \sigma^2(\eta) = \psi''(\eta).$$

Such a distribution can be parametrized by its mean and we denote the distribution by NEF($\mu$). A special subclass of NEF is the one with a quadratic variance function (QVF),

$$\sigma^2 \equiv V(\mu) = a_0 + a_1\mu + a_2\mu^2. \tag{14}$$

We write in this case $X_i \sim \mathrm{NQ}(\mu)$. The NEF-QVF consists of three discrete distributions: binomial, negative binomial, and Poisson, and three continuous distributions: normal, gamma, and NEF-GHS. See Morris (1982) and Brown (1986).

Let $X_1, \dots, X_m \overset{\text{i.i.d.}}{\sim} \mathrm{NEF}(\mu)$ be a random sample and set $X = \sum_{i=1}^m X_i$. It follows from the CLT that $\sqrt{m}(\frac{X}{m} - \mu) \xrightarrow{L} N(0, V(\mu))$, as $m \to \infty$. A VST is a function $G : \mathbb{R} \to \mathbb{R}$ such that $G'(\mu) = V^{-\frac{1}{2}}(\mu)$. The standard delta method yields

$$\sqrt{m}\left(G\left(\frac{X}{m}\right) - G(\mu)\right) \xrightarrow{L} N(0, 1).$$

As in the Poisson case, the mean-matching and variance stabilizing properties can often be further improved by using a more general transformation of the form

$$H_m(X) = G\left(\frac{X + a}{m + b}\right) \tag{15}$$

with a suitable choice of constants $a$ and $b$. For our purpose, same as in Poisson regression, it is more important to optimally match the means than to optimally stabilize the variance. That is, we wish to choose the constants $a$ and $b$ such that $\mathbb{E}\{H_m(X)\} - G(\mu)$ is minimized. The following expansions for the mean and variance of the transformed variable $H_m(X)$ is useful for finding the optimal choice of $a$ and $b$.

LEMMA 2 (Brown, Cai and Zhou, 2010). *Let $\Theta$ be a compact set in the interior of the natural parameter space. Then for $\eta \in \Theta$ and for constants $a$ and $b$*

$$\begin{aligned}\mathbb{E}\{H_m(X)\} - G(\mu(\eta)) &\\ = \frac{1}{\sigma(\eta)}\left(a - b\mu(\eta) - \frac{\mu''(\eta)}{4\mu'(\eta)}\right) &\\ \cdot m^{-1} + O(m^{-2})&\end{aligned} \tag{16}$$

*and*

$$\mathrm{Var}\{H_m(X)\} = \frac{1}{m} + O(m^{-2}). \tag{17}$$

*Moreover, there exist constants $a$ and $b$ such that*

$$\mathbb{E}\left\{G\left(\frac{X + a}{m + b}\right)\right\} - G(\mu) = O(m^{-2}) \tag{18}$$

*for all $\eta \in \Theta$ with a positive Lebesgue measure if and only if the NEF has a quadratic variance function.*

As shown in Brown, Cai and Zhou (2010), among the VSTs of the form (15) for the NEF-QVF with $\sigma^2 = a_0 + a_1\mu + a_2\mu^2$, the best constants $a$ and $b$ for mean-matching are

$$a = \frac{1}{4}a_1 \quad \text{and} \quad b = -\frac{1}{2}a_2. \tag{19}$$

The VST (15) with the constants $a$ and $b$ given in (19) is called the MM-VST. For the five distributions (other than normal) in the NEF-QVF families, the specific expressions for the MM-VST $H_m$ are as follows:

- Poisson: $a = \frac{1}{4}, b = 0$, and $H_m(X) = 2\sqrt{(X + \frac{1}{4})/m}$.
- Binomial$(r, p)$: $a = \frac{1}{4}$, $b = \frac{1}{2r}$, and $H_m(X) = 2 \cdot \sqrt{r}\arcsin(\sqrt{\frac{X+1/4}{rm+1/2}})$.
- Negative Binomial$(r, p)$: $a = \frac{1}{4}$, $b = -\frac{1}{2r}$, and $H_m(X) = 2\sqrt{r}\ln(\sqrt{\frac{X+1/4}{mr-1/2}} + \sqrt{1 + \frac{X+1/4}{mr-1/2}})$.
- Gamma$(r, \lambda)$ (with $r$ known): $a = 0$, $b = -\frac{1}{2r}$, and $H_m(X) = \sqrt{r}\ln(\frac{X}{rm-1/2})$.
- NEF-GHS$(r, \lambda)$ (with $r$ known): $a = 0$, $b = -\frac{1}{2r}$, and $H_m(X) = \sqrt{r}\ln(\frac{X}{rm-1/2} + \sqrt{1 + \frac{X^2}{(mr-1/2)^2}})$.

Note that the MM-VST is different from the one that optimally stabilizes the variance. Take, for example, the binomial distribution with $r = 1$. In this case the VST that optimally stabilizes the variance is $\arcsin(\sqrt{(X + c)/(m + 2c)})$ with $c = 3/8$. Figure 5 compares the bias and variance for $c = 0, 1/4$ and $3/8$ in the binomial case with $m = 30$. The plots show similar behavior to that in the Poisson case as seen in Figure 2.

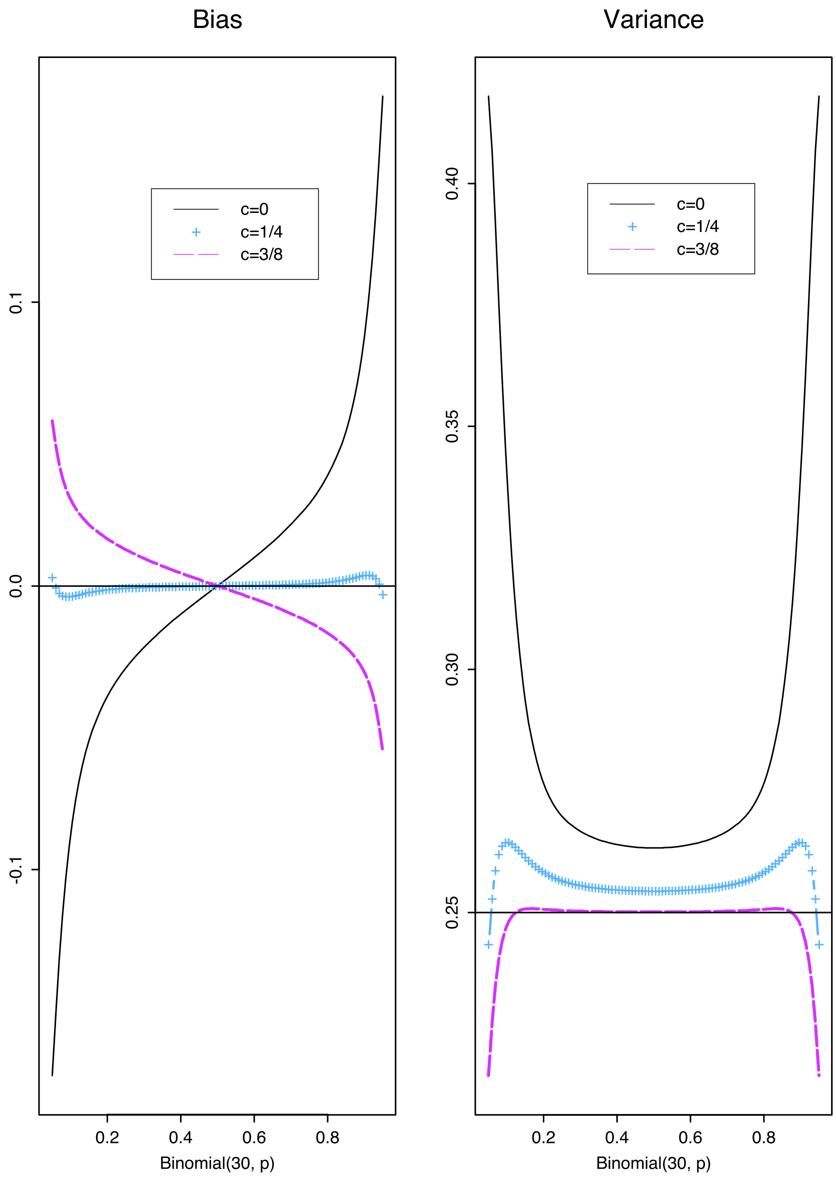

FIG. 5. *Comparison of the bias (left panel) and variance (right panel) of the VSTs for* Binomial(30, $p$) *with* $c = 0$ *(solid line),* $c = \frac{1}{4}$ *(+ line) and* $c = \frac{3}{8}$ *(dashed line). It is clear that* $c = \frac{1}{4}$ *is optimal for mean-matching and* $c = \frac{3}{8}$ *is the best for variance stabilizing.*

It is also interesting to also consider a continuous case and we use the exponential distribution as an example. Let $X_1, \dots, X_m \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$. Then $X = \sum_{i=1}^m X_i \sim \text{Gamma}(m, \lambda)$. The VST in this case is the log transformation. Figure 6, which is from Cai and Zhou (2010), compares the mean and variance of two log transformations of the form $\ln(\frac{X}{m-c})$ with $c = 1/2$, which is the MM-VST, and with $c = 0$, which is the usual log transformation. It is clear from the left panel of Figure 6 that the bias with $c = \frac{1}{2}$ is much smaller than corresponding bias with $c = 0$. For the variance, it is obvious that it does not depend on the value of $c$. In fact, in this case there do not exist constants $a$ and $b$ that optimally stabilize the variance.

Let us return to nonparametric regression in the NEF-QVF. Suppose we observe

$$Y_i \overset{\text{ind.}}{\sim} \text{NQ}\left(f\left(\frac{i}{n}\right)\right), \quad i = 1, \dots, n, \tag{20}$$

and wish to estimate the mean function $f(t)$. With the MM-VST in place, nonparametric regression in the NEF-QVF can be implemented via the Gaussianization Machine in exactly the same way as in the algorithm

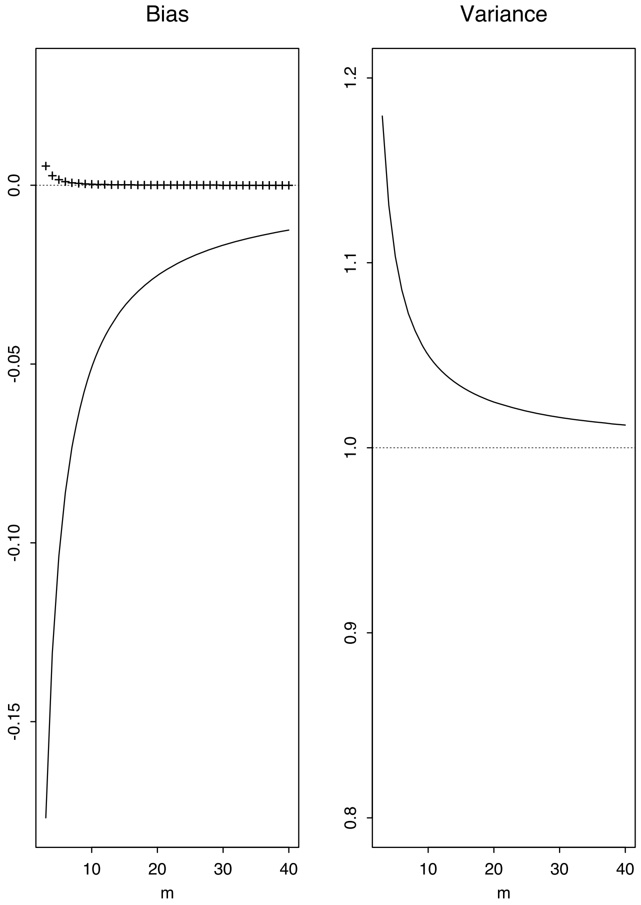

FIG. 6. *Comparison of the mean* (*left panel*) *and variance* (*right panel*) *of the log transformations for* Gamma$(m, \lambda)$ *with* $c = 0$ (*solid line*) *and* $c = \frac{1}{2}$ (+ *line*).

for Poisson regression.The procedure can be summarized as follows.

**Algorithm 4** A Gaussianization Machine for Nonparametric Regression in the NEF-QVF

```
1: Binning: Divide the indices $\{1, \dots, n\}$ into $T \asymp n^{3/4}$ equal sized groups $I_1, \dots, I_T$ with $m$ consecutive indices each. Let $Q_j = \sum_{i \in I_j} Y_i$, $j = 1, \dots, T$.
2: MM-VST: Let $Y_j^* = H_m(Q_j)$, $j = 1, \dots, T$. Then treat $Y^* = \{Y_1^*, \dots, Y_T^*\}$ as the new equispaced sample for a homoscedastic Gaussian regression problem.
3: Standard Gaussian Regression: Apply your favorite nonparametric regression procedure to the binned and transformed data $Y^*$ to obtain an estimate $\widehat{G(f)}$ of $G(f)$.
4: Inverse Transformation: Estimate the mean function $f$ by $\hat{f} = G^{-1}(\widehat{G(f)})$.
```

In Step 4, if due to randomness $\widehat{G(f)}$ is not in the domain of $G^{-1}$, then the usual correction is applied as follows. The domain of $G^{-1}$ is an interval, say between $a$ and $b$, one sets $G^{-1}(\widehat{G(f)}) = G^{-1}(a)$ if $\widehat{G(f)} < a$ and sets $G^{-1}(\widehat{G(f)}) = G^{-1}(b)$ if $\widehat{G(f)} > b$.

Figure 7 illustrates the steps of the Gaussianization Machine algorithm for nonparametric binomial regression where BlockJS is used for Gaussian regression in Step 3.

### 4.1 Extensions to General One-Parameter NEF

Cai and Zhou (2010) further extends the Gaussianization Machine built in Brown, Cai and Zhou (2010) to treat nonparametric regression in general natural exponential families. When the variance is not a quadratic function of the mean, the VST still exists, although the MM-VST may not exist. Let $X_1, \dots, X_m \overset{\text{i.i.d.}}{\sim} \text{NEF}(\mu)$ and set $\bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$. Let $G$ be the usual VST and define

$$H_m(\mu) = \mathbb{E}G(\bar{X}). \tag{21}$$

As mentioned earlier, The MM-VST only exists in the NEF-QVF. Cai and Zhou (2010) uses the usual VST $G(\cdot)$ instead. To control the transformation bias, a different inverse transformation is used. To be more specific, suppose one observes

$$Y_i \overset{\text{ind.}}{\sim} \text{NEF}\left(f\left(\frac{i}{n}\right)\right), \quad i = 1, \dots, n, \tag{22}$$

and wishes to estimate the mean function $f(t)$. Again, group the observations into $T$ bins with $m = n/T$ observations in each bin. Let $Q_j$ be the sum of observations in the $j$th bin, that is, $Q_j = \sum_{i=(j-1)m+1}^{jm} Y_i$. Apply the VST to obtain $Y_j^* = G(\frac{Q_j}{m})$ and treat $Y^* = (Y_1^*, \dots, Y_T^*)^{\mathsf{T}}$ as a sample for a standard Gaussian regression problem with the regression function being $H_m(f(\cdot))$. Once an estimator $\widehat{H_m(f(\cdot))}$ of $H_m(f(\cdot))$ is obtained, the mean function $f$ is estimated by $\hat{f} = H_m^{-1}(\widehat{H_m(f)})$. Algorithmically, the key difference is the use of the usual VST as the transformation in the Gaussianization Machine and the use of $H_m^{-1}$ as the inverse transformation. See Cai and Zhou (2010) for more details.

The advantage of using $H_m^{-1}$ as the inverse transformation is that it significantly reduces the bias of the inverse transformation. This enables one to use much smaller bin size than what is required in Brown, Cai and Zhou (2010). As a consequence, the procedure can still perform well when the regression function is less

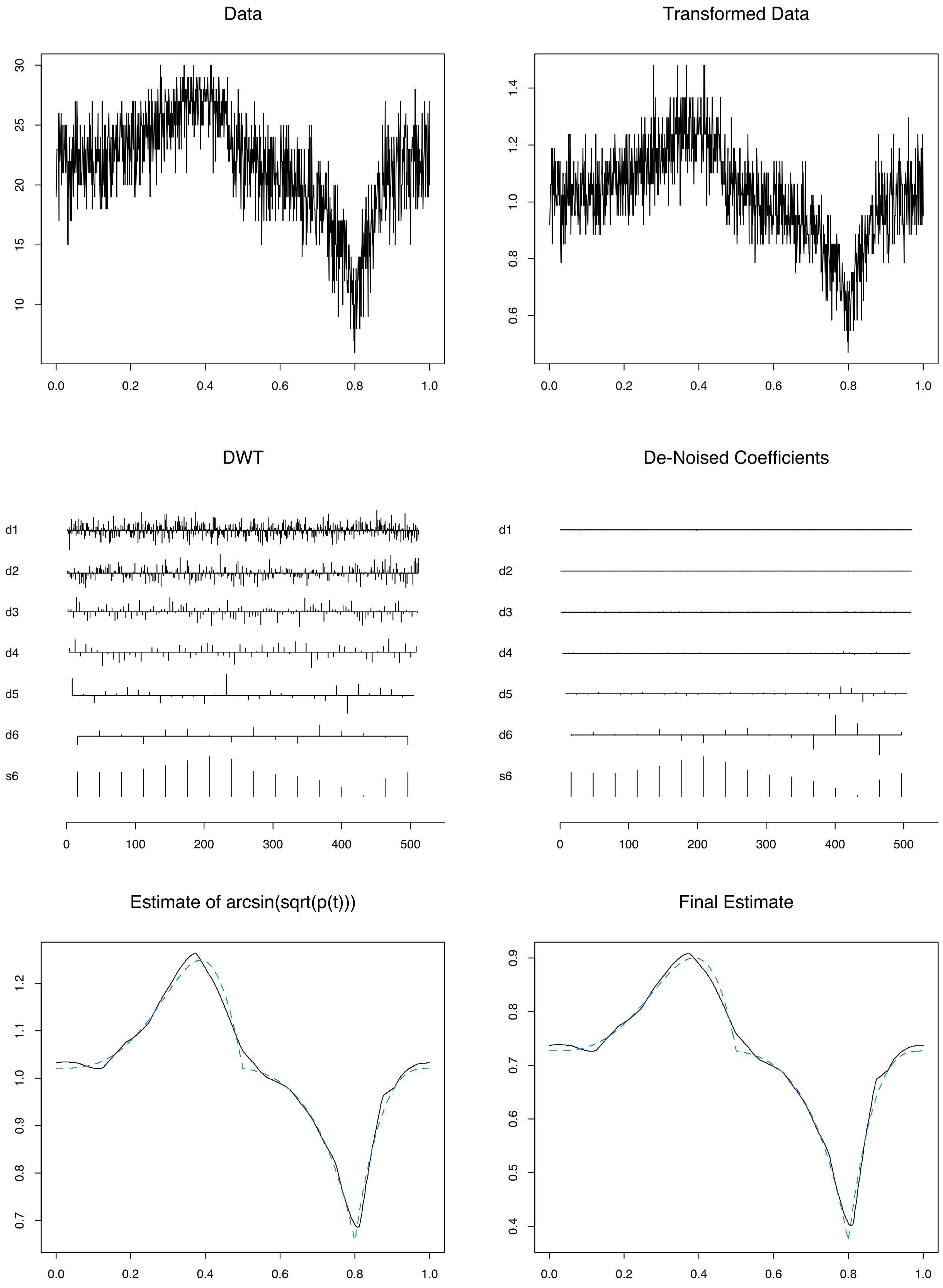

FIG. 7. *The Gaussianization Machine for nonparametric binomial regression as implemented via BlockJS.*

smooth and the method is applicable to a wider class of exponential families. The drawback of the approach is that the inverse transformation $H_m^{-1}$ typically does not have a closed form.

## 5. ROBUST NONPARAMETRIC REGRESSION

In this section, we consider robust nonparametric regression where the error distribution is unknown and possibly heavy tailed. Specifically, one observes $\{Y_1, \dots, Y_n\}$ with

$$Y_i = f\left(\frac{i}{n}\right) + \xi_i, \quad i = 1, \dots, n, \tag{23}$$

where the errors $\xi_i$ are independent and identically distributed with $\text{median}(\xi_i) = 0$ and an unknown density $h$. For some heavy-tailed distributions such as Cauchy distribution the mean does not even exist. So here we do not assume the existence of the mean and aim to estimate the median function $f(\cdot)$.

As in the case of many non-Gaussian regression problems discussed earlier, applying standard Gaussian regression methods directly to the data $\{Y_1, \dots, Y_n\}$ does not lead to good results in general. This can be seen easily by a comparison of extreme values. In Gaussian regression with $\xi_i \overset{\text{i.i.d.}}{\sim} N(0,1)$, $\xi_{\max} \equiv \max\{|\xi_1|, \dots, |\xi_n|\} = \sqrt{2\log n}(1+o(1))$ and $\mathbb{P}(\xi_{\max} \le \sqrt{2\log n}) \to 1$. This fact is the basis for the choice of the threshold level $\sqrt{2\log n}$ in the standard wavelet thresholding procedures. In contrast, when the noise $\xi_i$ has the standard Cauchy distribution, typical realizations of $\xi_i$ contain order $\frac{n}{\sqrt{\log n}}$ observations with magnitude larger than $\sqrt{2\log n}$ since

$$\begin{aligned}\mathbb{P}(|\xi_i| \ge \sqrt{2\log n}) &= 1 - \frac{2}{\pi}\arctan(\sqrt{2\log n})\\ &= \frac{\sqrt{2}}{\pi\sqrt{\log n}}(1+o(1)).\end{aligned}$$

Indeed, with probability close to 47% the extreme value $\xi_{\max} \equiv \max\{|\xi_1|, \dots, |\xi_n|\}$ can be larger than $n$ since $\mathbb{P}(\xi_{\max} \ge n) = 1 - (\frac{2}{\pi}\arctan(n))^n = 1 - \exp(-\frac{2}{\pi})(1+o(1))$. Clearly the conventional wavelet thresholding procedures designed for Gaussian noise would fail if they are applied directly to the sample $\{Y_1, \dots, Y_n\}$ when the noise is in fact heavy tailed. Figure 8, which is from Cai and Zhou (2009), illustrates the failure of such a naive approach (middle panel) and the success of the proposed robust estimate constructed by applying BlockJS to the data processed by a Gaussianization Machine (right panel). The difference is quite striking.

Similar to nonparametric regression in the NEF, robust regression can be turned into a standard Gaussian regression problem through a Gaussianization Machine and then in principle any procedure for Gaussian nonparametric regression can be used. For robust regression, the Gaussianization Machine has two components, binning and taking local median. More specifically, we group the observations $\{Y_1, \dots, Y_n\}$ into $T \asymp n^{3/4}$ bins of size $m$, and then take the median $Y_j^*$ of the observations in the $j$th bin for $j = 1, \dots, T$. Theoretical analysis shows that, for a wide range of error distributions, the medians $\{Y_1^*, \dots, Y_T^*\}$ can be viewed as if they were generated from the standard Gaussian nonparametric regression model where

$$\begin{aligned}Y_j^* &= f\left(\frac{j}{T}\right) + b_m\\ &\quad + \frac{1}{2h(0)\sqrt{m}}z_j, \quad z_j \overset{\text{i.i.d.}}{\sim} N(0,1), j = 1, \dots, T,\end{aligned} \tag{24}$$

with $b_m = \mathbb{E}\{\text{median}(\xi_1, \dots, \xi_m)\}$ being an unknown constant. Any good Gaussian regression methods can then be applied to the transformed data $\{Y_1^*, \dots, Y_T^*\}$ to obtain an estimator $\hat{g}(t)$ for $g(t) = f(t) + b_m$.

In order to construct the final estimator for $f$, an additional step of estimating $b_m$ is needed. Although the median of individual $\xi_i$ is 0, the expectation of the sample median of $\xi_1, \dots, \xi_m$ is nonzero in general. The quantity $b_m$ can be viewed as the systematic bias due to the expectation of the sample median of the noise $\xi_i$ in each bin. We estimate $b_m$ as follows. Divide each bin $I_j$ into two sub-bins with the first bin of the size $\lfloor \frac{m}{2} \rfloor$. Let $\tilde{Y}_j^*$ be the median of observations in the first sub-bin. We set

$$\hat{b}_m = \frac{1}{T}\sum_{j=1}^{T}(\tilde{Y}_j^* - Y_j^*). \tag{25}$$

The estimate $\hat{b}_m$ can be viewed as the bias correction. The final estimator of $f$ is given by

$$\hat{f}(t) = \hat{g}(t) - \hat{b}_m.$$

The four-step Gaussianization Machine for robust nonparametric regression has been summarized in Algorithm 2 in Section 1.

As in the previous sections, we use BlockJS for the Gaussian regression to obtain an estimator $\hat{g}$ for illustration. In this case, we set $T = 2^J$ with $J = \lfloor \log_2 n^{3/4} \rfloor$ and $m = n/T$. Define

$$\hat{f}(t) = \hat{g}(t) - \hat{b}_m, \tag{26}$$

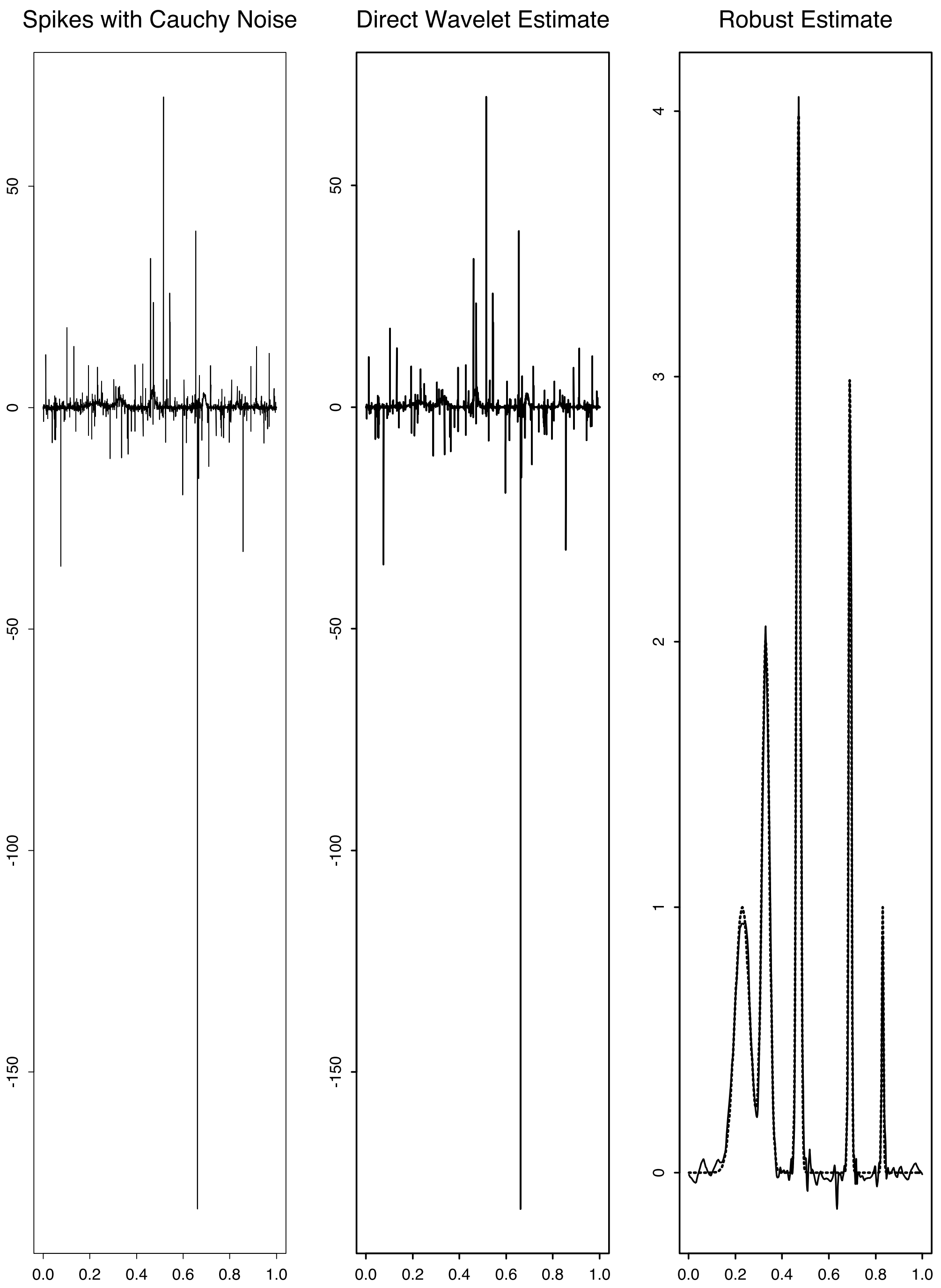

FIG. 8. *Left panel*: *Spikes signal with Cauchy noise*; *Middle panel*: *An estimate obtained by applying directly a wavelet procedure to the original noisy signal*; *Right panel*: *A robust estimate by applying BlockJS to the gaussianized data.*

where $\hat{g}(t)$ is the BlockJS estimator based on the binned and transformed data $\{Y_1^*, \dots, Y_T^*\}$.

The following Figure 9 is from Cai and Zhou (2009). It illustrates the main steps of the Gaussianization machine as implemented by the BlockJS procedure. It also shows a comparison of the robust estimate given in (26) by applying BlockJS to the gaussianized data (bottom left panel) and a direct application of BlockJS to the original data (bottom right panel). Here the noise has $t$ distribution with 2 degrees of freedom.

### 5.1 The Case of Symmetric Error Distributions

Cai and Zhou (2009) considered robust nonparametric regression where the error distribution is assumed

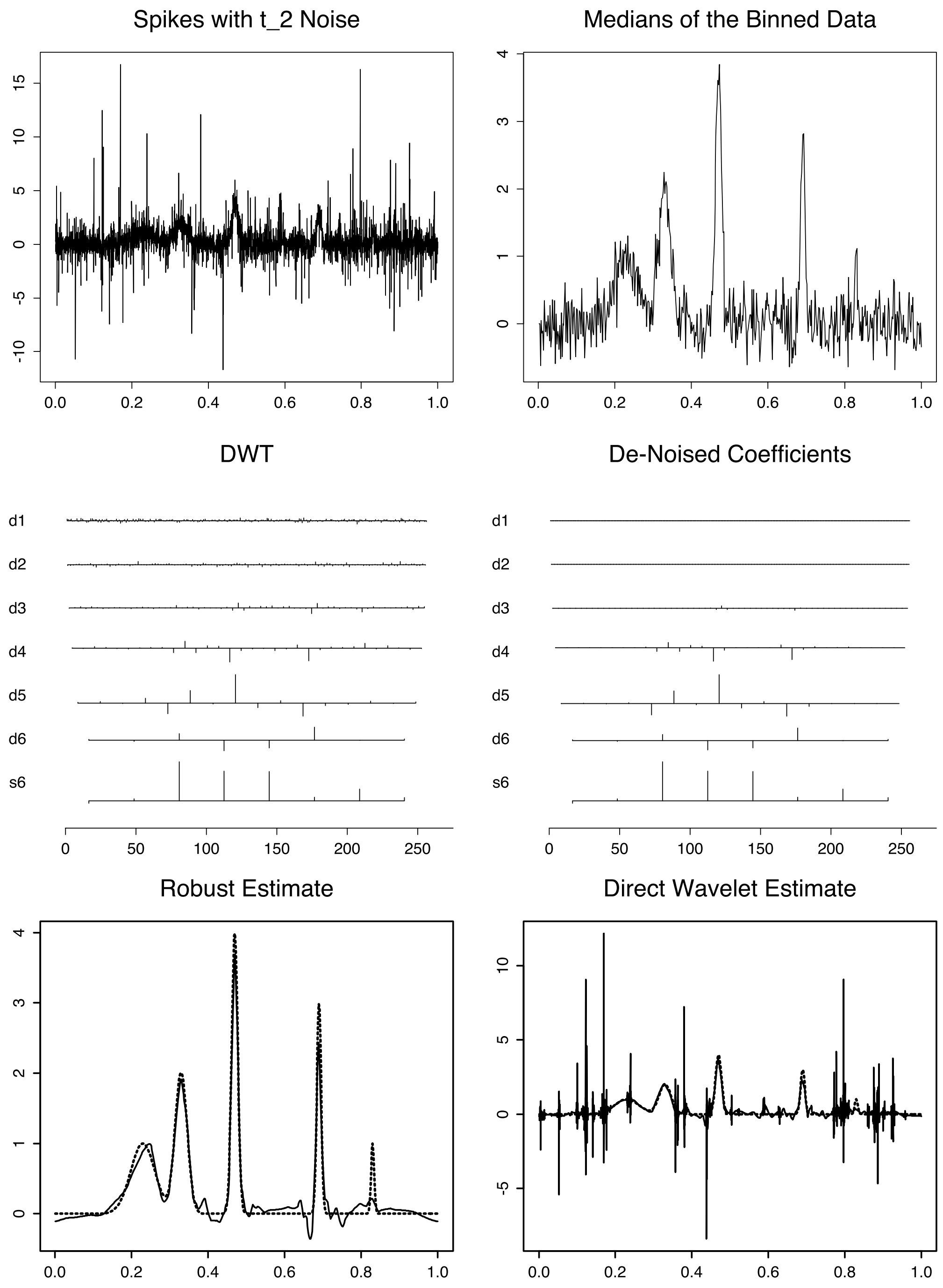

FIG. 9. *The binning and local median algorithm for robust nonparametric regression as implemented via BlockJS. The noise distribution is t distribution with* 2 *degrees of freedom. Bottom right*: *BlockJS applied directly to the original data.*

to be symmetric with median 0. In this case, the procedure can be simplified. Note that $b_m \equiv 0$ when the error distribution is symmetric and so the bias correction step can be eliminated. More importantly, the bin size can be chosen to be logarithmic in $n$, much smaller than what is required in Brown, Cai and Zhou (2008) and much stronger results can be obtained. In the case of symmetric error distribution, Cai and Zhou (2009) establishes the asymptotic equivalent for a large class of unbounded losses between the experiment of observing the local medians $\{Y_1^*, \dots, Y_T^*\}$ and a standard nonparametric Gaussian regression experiment.

The results on asymptotic equivalence have direct implications for estimation, confidence sets, and hypothesis testing. Cai and Zhou (2009) constructed easily implementable, robust, and adaptive procedures for estimation of the regression function and estimation of a quadratic functional. Other problems such as estimation and confidence intervals for linear functionals can be handled in a similar way.

## 6. TECHNICAL INSIGHTS AND THEORETICAL PROPERTIES

The Gaussianization Machines discussed in the previous sections for various non-Gaussian problems share the same features: binning and transformation. As shown earlier, the binned and transformed data $\{Y_1^*, \dots, Y_T^*\}$ can be treated as if they were generated from a homoscedastic nonparametric Gaussian regression model. As mentioned briefly in Section 1.3, the fundamental technical tool for understanding why the Gaussianization Machines work is the quantile coupling inequalities. We now explain the technical insights in more detail and discuss the theoretical properties of the Gaussianization Machine algorithms.

### 6.1 Technical Insights

We begin with the nonparametric Poisson regression. The quantile coupling inequality given in Brown et al. (2010), which is a direct consequence of the results developed in Komlós, Major and Tusnády (1975), provides a tight bound for the quantile coupling between a Poisson variable and a normal variable.

LEMMA 3 (Brown et al., 2010). *Let $\lambda > 0$ and let $X \sim \text{Poisson}(\lambda)$. There exist a standard normal random variable $Z \sim N(0, 1)$ and constants $c_1, c_2 > 0$ not depending on $\lambda$ such that whenever the event $A = \{|X - \lambda| \le c_1 \lambda\}$ occurs,*

$$|X - \lambda - \sqrt{\lambda} Z| < c_2(Z^2 + 1). \tag{27}$$

This result can be used to show precisely how well the transformed data is approximated by independent normal variables. To understand the effect of the Gaussianization Machine, let us consider $X \sim \text{Poisson}(\lambda)$ and denote $Y = \sqrt{X + \frac{1}{4}}$ and $\epsilon = \mathbb{E}Y - \sqrt{\lambda}$. Let $Z$ be a standard normal variable satisfying (27). Then $Y$ can be expressed as

$$Y = \sqrt{\lambda} + \epsilon + \frac{1}{2} Z + \delta,$$

where

$$\begin{aligned}\delta = {} & \frac{X - \lambda}{\sqrt{X + \frac{1}{4}} + \sqrt{\lambda + \frac{1}{4}}} - \frac{1}{2} Z \\ & - \mathbb{E}\Bigl(\frac{X - \lambda}{\sqrt{X + \frac{1}{4}} + \sqrt{\lambda + \frac{1}{4}}}\Bigr).\end{aligned} \tag{28}$$

The approximation result on the MM-VST given in Lemma 1 in Section 2 shows that $\epsilon = \mathbb{E}Y - \sqrt{\lambda}$ is "small" when $\lambda$ is large, which is the case after binning. (It also shows the importance of the correction factor $1/4$.) Lemma 3 implies that the random variable $\delta$ is "stochastically small". Hence for practical purposes, the binned and transformed data $Y_j^*$ can be viewed as independent homoscedastic normal variables where $Y_j^* \mathrel{\dot\sim} N(\sqrt{\lambda_j}, \frac{1}{4})$, and the Poisson regression problem is thus"gaussianized".

The main ideas for Poisson regression can be extended easily to the nonparametric regression in the NEF-QVF. The transformation is the MM-VST $H_m$ defined in (15). Let $X_1, \dots, X_m \overset{\text{i.i.d.}}{\sim} \text{NQ}(\mu)$ with variance $V$. Let $X = \sum_{i=1}^m X_i$, $Y = H_m(X) = G(\frac{X+a}{m+b})$, and $\epsilon = \mathbb{E}Y - G(\mu)$. The following coupling inequality shows that $X$ can be treated as a normal random variable with mean $m\mu$ and variance $mV$ when $m$ is large.

LEMMA 4 (Brown, Cai and Zhou, 2010). *There exist a standard normal random variable $Z$ and constants $c_1, c_2 > 0$ not depending on $m$ such that whenever the event $A = \{|X - m\mu| \le c_1 m\}$ occurs,*

$$|X - m\mu - \sqrt{mV} Z| < c_2(Z^2 + 1). \tag{29}$$

Let $Y = H_m(X) = G(\frac{X+a}{m+b})$, $\epsilon = \mathbb{E}Y - G(\mu)$ and $Z \sim N(0, 1)$ satisfying (29). Then $Y$ can be written as

$$Y = G(\mu) + \epsilon + m^{-\frac{1}{2}} Z + \delta, \tag{30}$$

where $\epsilon$ is a deterministic approximation error and $\delta$ is a stochastic error. The mean-matching property of $H_m$ guarantees that the deterministic approximation error $\epsilon$ is small and the quantile coupling inequality ensures the stochastic error $\delta$ to be small. For nonparametric regression in the NEF-QVF, the observations in the same bin have different means in general and the bias increases with $m$. On the other hand, the stochastic error $\delta$ decreases as $m$ increases. The choice of $m \approx n^{1/4}$ (or equivalently $T \asymp n^{3/4}$) balances these two different kinds of errors. See Brown, Cai and Zhou (2010) for a detailed technical analysis.

We now turn to the robust nonparametric regression. In this case, the transformation in the Gaussianization Machine is the local median transformation. Let $\xi_1, \dots, \xi_m$ be i.i.d. with density function $h$ such that $\int_{-\infty}^0 h(x)\,dx = \frac{1}{2}$, $h(0) > 0$, and $h(x)$ is Lipschitz at $x = 0$. Let $\xi_{\text{med}} = \text{median}(\xi_1, \dots, \xi_m)$. It follows from the CLT that $\xi_{\text{med}}$ is approximately normal when $m$ is large. A precise quantification of the approximation error is need. The following quantile coupling inequality for the sample median is a key tool for understanding the effects of the Gaussianization Machine.

LEMMA 5 (Median Coupling Inequality, Brown, Cai and Zhou (2008)). *For every odd integer $m \geq 3$, there exist a standard normal random variable $Z$ and constants $c_1, c_2 > 0$ not depending on $m$ such that*

$$\left|\xi_{\text{med}} - \frac{1}{2\sqrt{mh(0)}}Z\right| \leq \frac{c_2}{mh(0)}(|Z|^2 + 1) \quad \textit{when } |Z| \leq c_1\sqrt{m}. \tag{31}$$

A similar result holds when the bin size $m$ is an even integer. For robust nonparametric regression, one observes $\{Y_1, \ldots, Y_n\}$ as in (23). The Gaussianization Machine first groups the data into $T$ bins with $m$ observations each. Take the first bin as an example. In this bin, the observations are $Y_i = f(\frac{i}{n}) + \xi_i$, $i = 1, \ldots, m$. Note that the values of $f(\frac{i}{n})$ are not equal. Let $Z \sim N(0, 1)$ satisfy (31). The sample median $Y_1^* = \text{median}(Y_1, \ldots, Y_m)$ can be written as

$$Y_1^* = \bar{f}_1 + \epsilon_1 + b_m + \frac{1}{2\sqrt{mh(0)}}Z + \delta_1,$$

where $\bar{f}_1 = T\int_0^{\frac{1}{T}} f(x)\,dx$ is the average of $f$ in the subinterval $[0, \frac{1}{T}]$, $\epsilon_1$ is a deterministic approximation error, $b_m = \mathbb{E}(\xi_{\text{med}})$ is a constant depending only on the error density function $h$ and bin size $m$, and $\delta_1$ is the stochastic error. As in the case of nonparametric regression in the NEF-QVF, the approximation error $\epsilon_1$ increases with $m$ and the stochastic error $\delta_1$ decreases with $m$. The choice of $m \approx n^{1/4}$ balances these two different types of errors.

When the error distribution is known to be symmetric, $b_m = 0$ and the stochastic error $\delta_1$ is much smaller for a given $m$. This allows for choosing a smaller $m$ and the estimator achieves optimality over a larger collection of function spaces. See Brown, Cai and Zhou (2008) and Cai and Zhou (2009) for more details.

### 6.2 Theoretical Properties

Brown, Cai and Zhou (2008, 2010) and Brown et al. (2010) use the BlockJS procedure in Step 3 of the Gaussianization Machine algorithms and analyze the theoretical properties of the procedures for nonparametric density estimation, regression in the NEF-QVF, and robust nonparametric regression. Let us denote such as estimator by $\hat{f}_{\text{BJS}}$ in the following discussion. The performance of the estimator $\hat{f}_{\text{BJS}}$ is measured globally by the mean integrated squared error

$$R(\hat{f}_{\text{BJS}}, f) = \mathbb{E}\|\hat{f}_{\text{BJS}} - f\|_2^2, \tag{32}$$

and locally at any given point $t_0 \in (0, 1)$ by the pointwise mean squared error

$$R(\hat{f}_{\text{BJS}}(t_0), f(t_0)) = \mathbb{E}(\hat{f}_{\text{BJS}}(t_0) - f(t_0))^2. \tag{33}$$

The estimator $\hat{f}_{\text{BJS}}$ is computationally easy to implement and performs well numerically. Theoretically, the global performance of the estimator $\hat{f}_{\text{BJS}}$ is evaluated over a collection of the Besov spaces, which occur naturally in many fields of analysis. They contain a number of traditional smoothness spaces such as Hölder and Sobolev spaces as special cases. Roughly speaking, the Besov space $B^\alpha_{p,q}$ contains functions having $\alpha$ bounded derivatives in $L^p$ norm, the third parameter $q$ gives a finer gradation of smoothness. See Triebel (1983) for detailed discussion on Besov spaces.

Brown, Cai and Zhou (2008, 2010) and Brown et al. (2010) show that, under mild regularity conditions, $\hat{f}_{\text{BJS}}$ adaptively achieves the optimal rate of convergence, $n^{-2\alpha/(1+2\alpha)}$, for global estimation over a compact ball in a wide collection of the Besov spaces $B^\alpha_{p,q}$, without prior knowledge of the smoothness of the underlying functions. The estimator $\hat{f}_{\text{BJS}}$ is also spatially adaptive. For local estimation, $\hat{f}_{\text{BJS}}$ attains the pointwise adaptive minimax rate over a set of local Hölder classes, without prior knowledge of the local smoothness of the underlying functions. These results hold across the collection of all the non-Gaussian function estimation models discussed in the previous sections. In other words, the optimality results for BlockJS in the case of the standard nonparametric Gaussian regression essentially carry over to these non-Gaussian problems without change.

The analysis of the theoretical properties of $\hat{f}_{\text{BJS}}$ relies heavily on the quantile coupling inequalities and a general oracle inequality for block thresholding where the noise is not necessarily Gaussian. This risk bound is useful in turning the analysis of the estimator $\hat{f}_{\text{BJS}}$ in the non-Gaussian setting into the bias-variance tradeoff calculation which is often used in the more standard nonparametric Gaussian regression.

## 7. DISCUSSION AND CONCLUDING REMARKS

The Gaussianization Machines discussed in the present paper are practical and easily implementable. They extend the theories and methodologies developed for the standard nonparametric Gaussian regression to a much larger class of models. The focus so far has mainly been on estimation, in particular the global recovery of the regression function. It is also of significant interest to consider other estimation problems and statistical inference, including estimation of the linear and quadratic functionals as well as confidence intervals and hypothesis testing, under these non-Gaussian models. It is interesting to investigate the extend to which the Gaussianization Machines work for these estimation and inference problems.

In addition to the NEF-QVF, the MM-VST also exists for some other important families of distributions, including the gamma-Poisson family and the beta-binomial family (Brown, Cai and Zhou, 2010). The Gaussianization Machine developed for nonparametric regression in the NEF-QVF can be extended to nonparametric regression in these families as well. In addition, as discussed in the Introduction, global asymptotic equivalence theory has been established in a wide range of settings. It is of practical interest to develop similar user-friendly Gaussianization Machines for other non-Gaussian models such as nonparametric autoregression and GARCH models.

Much recent attention in statistics has been on high-dimensional statistical inference. As usual, the Gaussian models occupy a particularly important place in the high-dimensional settings. Another interest direction is to investigate when and how the main ideas behind the Gaussianization Machines discussed in the present paper can be applied in the high-dimensional settings to extend the theories and methodologies originally developed in the Gaussian case to the non-Gaussian cases.

## ACKNOWLEDGMENTS

Supported in part by NSF Grant DMS-1712735 and NIH Grants R01-GM129781 and R01-GM123056.

## REFERENCES

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35** 246–254. MR0028556

Bartlett, M. S. (1936). The square root transformation in analysis of variance. *Suppl. J. R. Stat. Soc.* **3** 68–78.

Berk, R. and MacDonald, J. M. (2008). Overdispersion and Poisson regression. *J. Quant. Criminol.* **24** 269–284.

Besbeas, P., De Feis, I. and Sapatinas, T. (2004). A comparative simulation study of wavelet shrinkage estimators for Poisson counts. *Int. Stat. Rev.* **72** 209–237.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **9**. IMS, Hayward, CA. MR0882001

Brown, L. D., Cai, T. T. and Zhou, H. H. (2008). Robust nonparametric estimation via wavelet median regression. *Ann. Statist.* **36** 2055–2084. MR2458179

Brown, L. D., Cai, T. T. and Zhou, H. H. (2010). Nonparametric regression in exponential families. *Ann. Statist.* **38** 2005–2046. MR2676882

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958

Brown, L. D., Wang, Y. and Zhao, L. H. (2003). On the statistical equivalence at suitable frequencies of GARCH and stochastic volatility models with the corresponding diffusion model. *Statist. Sinica* **13** 993–1013.

Brown, L. D. and Zhang, C.-H. (1998). Asymptotic nonequivalence of nonparametric experiments when the smoothness index is $1/2$. *Ann. Statist.* **26** 279–287. MR1611772

Brown, L. D., Cai, T. T., Low, M. G. and Zhang, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30** 688–707.

Brown, L. D., Carter, A. V., Low, M. G. and Zhang, C.-H. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.* **32** 2074–2097. MR2102503

Brown, L., Cai, T. T., Zhang, R., Zhao, L. and Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probab. Theory Related Fields* **146** 401–433. MR2574733

Cai, T. T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.* **27** 898–924. MR1724035

Cai, T. T. and Zhou, H. H. (2009). Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann. Statist.* **37** 3204–3235. MR2549558

Cai, T. T. and Zhou, H. H. (2010). Nonparametric regression in natural exponential families. In *Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat. (IMS) Collect.* **6** 199–215. IMS, Beachwood, OH. MR2798520

Dalalyan, A. and Reiss, M. (2006). Asymptotic statistical equivalence for scalar ergodic diffusions. *Probab. Theory Related Fields* **134** 248–282. MR2222384

Dalalyan, A. and Reiss, M. (2007). Asymptotic statistical equivalence for ergodic diffusions: The multidimensional case. *Probab. Theory Related Fields* **137** 25–47. MR2278451

Daubechies, I. (1992). *Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics* **61**. SIAM, Philadelphia, PA. MR1162107

Delattre, S. and Hoffmann, M. (2002). Asymptotic equivalence for a null recurrent diffusion. *Bernoulli* **8** 139–174. MR1895888

Donoho, D. L. (1993). Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In *Different Perspectives on Wavelets* (*San Antonio*, *TX*, 1993). *Proc. Sympos. Appl. Math.* **47** 173–205. Amer. Math. Soc., Providence, RI. MR1268002

Efromovich, S. and Samarov, A. (1996). Asymptotic equivalence of nonparametric regression and white noise model has its limits. *Statist. Probab. Lett.* **28** 143–145. MR1394666

Fryzlewicz, P. and Nason, G. P. (2004). A Haar–Fisz algorithm for Poisson intensity estimation. *J. Comput. Graph. Statist.* **13** 621–638. MR2087718

Genon-Catalot, V., Laredo, C. and Nussbaum, M. (2002). Asymptotic equivalence of estimating a Poisson intensity and a positive diffusion drift. *Ann. Statist.* **30** 731–753.

GOLUBEV, G. K., NUSSBAUM, M. and ZHOU, H. H. (2010). Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Ann. Statist.* **38** 181–214. MR2589320

GRAMA, I. G. and NEUMANN, M. H. (2006). Asymptotic equivalence of nonparametric autoregression and nonparametric regression. *Ann. Statist.* **34** 1701–1732. MR2283714

GRAMA, I. and NUSSBAUM, M. (2002). Asymptotic equivalence for nonparametric regression. *Math. Methods Statist.* **11** 1–36. MR1900972

HOYLE, M. H. (1973). Transformations–an introduction and a bibliography. *Int. Stat. Rev.* **41** 203–223. MR0423611

JOHNSTONE, I. M. (2011). *Gaussian Estimation: Sequence and Wavelet Models*. Unpublished manuscript.

KOLACZYK, E. D. (1999a). Bayesian multiscale models for Poisson processes. *J. Amer. Statist. Assoc.* **94** 920–933. MR1723303

KOLACZYK, E. D. (1999b). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statist. Sinica* **9** 119–135. MR1678884

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131. MR0375412

KROLL, M. (2019). Non-parametric Poisson regression from independent and weakly dependent observations by model selection. *J. Statist. Plann. Inference* **199** 249–270. MR3857826

LE CAM, L. (1974). On the information contained in additional observations. *Ann. Statist.* **2** 630–649. MR0436400

LOW, M. G. and ZHOU, H. H. (2007). A complement to Le Cam's theorem. *Ann. Statist.* **35** 1146–1165. MR2341701

MASON, D. M. and ZHOU, H. H. (2012). Quantile coupling inequalities and their applications. *Probab. Surv.* **9** 439–479. MR3007210

MEISTER, A. (2011). Asymptotic equivalence of functional linear regression and a white noise inverse problem. *Ann. Statist.* **39** 1471–1495. MR2850209

MILSTEIN, G. and NUSSBAUM, M. (1998). Diffusion approximation for nonparametric autoregression. *Probab. Theory Related Fields* **112** 535–543. MR1664703

MORRIS, C. N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10** 65–80. MR0642719

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430. MR1425959

REISS, M. (2008). Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist.* **36** 1957–1982. MR2435461

REISS, M. (2011). Asymptotic equivalence for inference on the volatility from noisy observations. *Ann. Statist.* **39** 772–802. MR2816338

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. CRC Press, London. MR0848134

TRIEBEL, H. (1983). *Theory of Function Spaces. Monographs in Mathematics* **78**. Birkhäuser, Basel. MR0781540

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. MR2724359

VER HOEF, J. M. and BOVENG, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* **88** 2766–2772.

WANG, Y. (2002). Asymptotic nonequivalence of GARCH models and diffusions. *Ann. Statist.* **30** 754–783.

WINKELMANN, R. (2003). *Econometric Analysis of Count Data*, 4th ed. Springer, Berlin. MR2148271