Nonparametric Regression in Exponential Families

Lawrence D. Brown[1], T. Tony Cai[2] and Harrison H. Zhou[3]
University of Pennsylvania and Yale University

Abstract

Most results in nonparametric regression theory are developed only for the case of additive noise. In such a setting many smoothing techniques including wavelet thresholding methods have been developed and shown to be highly adaptive. In this paper we consider nonparametric regression in exponential families which include, for example, Poisson regression, binomial regression, and gamma regression. We propose a unified approach of using a mean-matching variance stabilizing transformation to turn the relatively complicated problem of nonparametric regression in exponential families into a standard homoscedastic Gaussian regression problem. Then in principle any good nonparametric Gaussian regression procedure can be applied to the transformed data. In this paper we use a wavelet block thresholding rule to construct the final estimator of the regression function. The procedure is easily implementable. Both theoretical and numerical properties of the estimator are investigated. The estimator is shown to enjoy a high degree of adaptivity and spatial adaptivity. It simultaneously attains the optimal rates of convergence under integrated squared error over a wide range of Besov spaces and achieves adaptive local minimax rate for estimating functions at a point. The estimator also performs well numerically.

Keywords: Adaptivity; Asymptotic equivalence; Exponential family; James-Stein estimator; Nonparametric Gaussian regression; Quadratic variance function; Quantile coupling; Wavelets.

AMS 2000 Subject Classification: Primary 62G08, Secondary 62G20.

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.
[2]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.
The research of Tony Cai was supported in part by NSF Grants DMS-0072578 and DMS-0306576.
[3]Department of Statistics, Yale University, New Haven, CT 06511. The research of Harrison Zhou was supported in part by NSF Career Award DMS-0645676.

1 Introduction

Theory and methodology for nonparametric regression is now well developed for the case of additive noise particularly additive homoscedastic Gaussian noise. In such a setting many smoothing techniques including wavelet thresholding methods have been developed and shown to be adaptive and enjoy other desirable properties over a wide range of function spaces. However, in many applications the noise is not additive and the conventional methods are not readily applicable. For example, such is the case when the data are counts or proportions.

In this paper we consider nonparametric regression in exponential families. These include, for example, Poisson regression, binomial regression, and gamma regression. We present a unified treatment of these regression problems by using a mean-matching variance stabilizing transformation (VST) approach. The mean-matching VST turns relatively complicated problem of regression in exponential families into a standard homoscedastic Gaussian regression problem and then any good nonparametric Gaussian regression procedure can be applied.

Variance stabilizing transformations and closely related normalizing transformations have been widely used in many parametric statistical inference problems. See Hoyle (1973), Efron (1982) and Bar-Lev and Enis (1990). In the more standard parametric problems, the goal of VST is often to optimally stabilize the variance. That is, one desires the variance of the transformed variable to be as close to a constant as possible. For example, Anscombe (1948) introduced VSTs for binomial, Poisson and negative binomial distributions that provide the greatest asymptotic control over the variance of the resulting transformed variables. In the context of nonparametric function estimation, Anscombe's variance stabilizing transformation has also been briefly discussed in Donoho (1993) for density estimation. However, for our purposes it is much more essential to have optimal asymptotic control over the bias of the transformed variables. A mean-matching VST minimizes the bias of the transformed data while also stabilizes the variance.

Our procedure begins by grouping the data into many small size bins, and by then applying the mean-matching VST to the binned data. In principle any good Gaussian regression procedure could be applied to the transformed data to construct the final estimator of the regression function. In this paper we employ a wavelet block thresholding procedure. Wavelet thresholding methods have achieved considerable success in nonparametric regression in terms of spatial adaptivity and asymptotic optimality. In particular, block thresholding rules have been shown to possess impressive properties. The estimators make simultaneous decisions to retain or to discard all the coefficients within a block and increase estimation accuracy by utilizing information about neighboring coefficients. In the context of nonparametric regression local block thresholding has been studied, for example,

in Hall, Kerkyacharian, and Picard (1998), Cai (1999, 2002) and Cai and Silverman (2001). A block thresholding procedure first divides the empirical coefficients at each resolution level into non-overlapping blocks and then simultaneously estimate all the coefficients within a block. Motivated by the analysis of block thresholding rules for nonparametric regression in Cai (1999), we shall use a blockwise James-Stein rule with the block size $\log n$.

Both theoretical and numerical properties of our estimator are investigated. It is shown that the estimator enjoys excellent asymptotic adaptivity and spatial adaptivity. The procedure simultaneously attains the optimal rate of convergence under the integrated squared error over a wide range of the Besov classes. The estimator also automatically adapts to the local smoothness of the underlying function; it attains the local adaptive minimax rate for estimating functions at a point. A key step in the technical argument is the use of the quantile coupling inequality of Komlós, Major and Tusnády (1975) to approximate the binned and transformed data by independent normal variables. The procedure is easy to implement, at the computational cost of $O(n)$. In addition to enjoy the desirable theoretical properties, the procedure also performs well numerically.

We should note that nonparametric regression in exponential families has been considered in the literature. Among individual exponential families, the Poisson case is perhaps the most studied. 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) and Fryźlewicz and Nason (2001). In the case of Bernoulli regression, Antoniadis and Leblanc (2001) introduced a wavelet procedure based on diagonal linear shrinkers. Unified treatments for nonparametric regression in exponential families have also been proposed. Antoniadis and Sapatinas (2001) introduced a wavelet shrinkage and modulation method and showed that the estimator attains the optimal rate over the classical Sobolev spaces. Kolaczyk and Nowak (2005) proposed a recursive partition and complexity-penalized likelihood method. The estimator was shown to be within a logarithmic factor of the minimax rate under squared Hellinger loss over Besov spaces.

The paper is organized as follows. Section 2 discusses the mean-matching variance stabilizing transformation for natural exponential families. In Section 3, We first introduce the general approach of using the mean-matching VST to convert nonparametric regression in exponential families into a nonparametric Gaussian regression problem, and then present in detail a specific estimation procedure based on the mean-matching VST and wavelet block thresholding. Theoretical properties of the procedure are treated in Section 4. Section 5 investigates the numerical performance of the estimator. We also illustrate our estimation procedure in the analysis of two real data sets: a gamma-ray burst data set and a packet loss data set. Technical proofs are given in Section 6.

2 Mean-matching variance stabilizing transformation

We begin by considering mean-matching variance stabilizing transformations (VST) for natural exponential families. As mentioned in the introduction, VST has been widely used in many contexts and the conventional goal of VST is to optimally stabilize the variance. See, for example, Anscombe (1948) and Hoyle (1973). For our purpose of nonparametric regression in exponential families, we shall first develop a new class of VSTs, called mean-matching VSTs, which asymptotic minimizes the bias of the transformed variables while at the same time stabilizes the variance.

Let $X_1, X_2, ..., X_m$ be a random sample from a distribution in a natural exponential families with the probability density/mass function

$$q(x|\eta) = e^{\eta x - \psi(\eta)} h(x).$$

Here η is called the natural parameter. The mean and variance are respectively

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

A special subclass of interest is the one with a quadratic variance function (QVF),

$$\sigma^2 \equiv V(\mu) = a_0 + a_1\mu + a_2\mu^2. \tag{1}$$

In this case we shall write $X_i \sim NQ(\mu)$. The NEF-QVF families consist of six distributions, three continuous: normal, gamma, and NEF-GHS distributions and three discrete: binomial, negative binomial, and Poisson. See, e.g., Morris (1982) and Brown (1986).

Set $X = \sum_{i=1}^m X_i$. According to the Central Limit Theorem,

$$\sqrt{m}(X/m - \mu(\eta)) \xrightarrow{L} N(0, V(\mu(\eta))), \quad \text{as } m \to \infty.$$

A variance stabilizing transformation (VST) is a function $G : \mathbb{R} \to \mathbb{R}$ such that

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

The standard delta method then yields

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

It is known that the variance stabilizing properties can often be further improved by using a transformation of the form

$$H_m(X) = G(\frac{X + a}{m + b}) \tag{3}$$

with suitable choice of constants a and b. See, e.g., Anscombe (1948). In this paper we shall use the VST as a tool for nonparametric regression in exponential families. For this

purpose, 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 $E\{H_m(X)\}$ optimally matches $G(\mu(\eta))$.

To derive the optimal choice of a and b, we need the following expansions for the mean and variance of the transformed variable $H_m(X)$.

Lemma 1 *Let Θ_0 be a compact subset of the natural parameter space Θ. Assume that $\eta \in \Theta_0$ and the variance $\sigma^2(\eta)$ is positive on Θ_0. Then for constants a and b*

$$E\{H_m(X)\} - G(\mu(\eta)) = \frac{1}{\sigma(\eta)}(a - b\mu(\eta) - \frac{\mu''(\eta)}{4\mu'(\eta)}) \cdot m^{-1} + O(m^{-2}) \tag{4}$$

and

$$Var\{H_m(X)\} = \frac{1}{m} + O(m^{-2}). \tag{5}$$

Moreover, there exist constants a and b such that

$$E\{G(\frac{X+a}{m+b})\} - G(\mu(\eta)) = O(m^{-2}) \tag{6}$$

if and only if the exponential family has a quadratic variance function.

The proof of Lemma 1 is given in Section 6. The last part of Lemma 1 can be easily explained as follows. Equation (4) implies that Equation (6) holds if and only if

$$a - b\mu(\eta) - \frac{\mu''(\eta)}{4\mu'(\eta)} = 0$$

i.e., $\mu''(\eta) = 4a\mu'(\eta) - 4b\mu(\eta)\mu'(\eta)$. Solving this differential equation yields

$$\sigma^2(\eta) = \mu'(\eta) = a_0 + 4a\mu(\eta) - 2b\mu^2(\eta) \tag{7}$$

for some constant a_0. Hence the solution of the differential equation is exactly the subclass of natural exponential family with a quadratic variance function (QVF).

It follows from Equation (7) that among the VSTs of the form (3) for the exponential family with a quadratic variance function

$$\sigma^2 = a_0 + a_1\mu + a_2\mu^2$$

the best constants a and b for mean-matching are

$$a = \frac{1}{4}a_1 \quad \text{and} \quad b = -\frac{1}{2}a_2. \tag{8}$$

We shall call the VST (3) with the constants a and b given in (8) the mean-matching VST. The following are the specific expressions of the mean-matching VST H_m for the five distributions (other than normal) in the NEF-QVF families.

- Poisson: $a = 1/4$, $b = 0$, and $H_m(X) = 2\sqrt{(X + \frac{1}{4})/m}$.
- Binomial(r, p): $a = 1/4$, $b = \frac{1}{2r}$, and $H_m(X) = 2\sqrt{r}\arcsin\left(\sqrt{\frac{X+1/4}{rm+1/2}}\right)$.
- Negative Binomial(r, p): $a = 1/4$, $b = -\frac{1}{2r}$, and

$$H_m(X) = 2\sqrt{r}\ln\left(\sqrt{\frac{X+1/4}{mr-1/2}} + \sqrt{1+\frac{X+1/4}{mr-1/2}}\right).$$

- Gamma(r, λ) (with r known): $a = 0$, $b = -\frac{1}{2r}$, and $H_m(X) = \sqrt{r}\ln(\frac{X}{rm-1/2})$.
- NEF-GHS(r, λ) (with r known): $a = 0$, $b = -\frac{1}{2r}$, and

$$H_m(X) = \sqrt{r}\ln\left(\frac{X}{rm-1/2} + \sqrt{1+\frac{X^2}{(mr-1/2)^2}}\right).$$

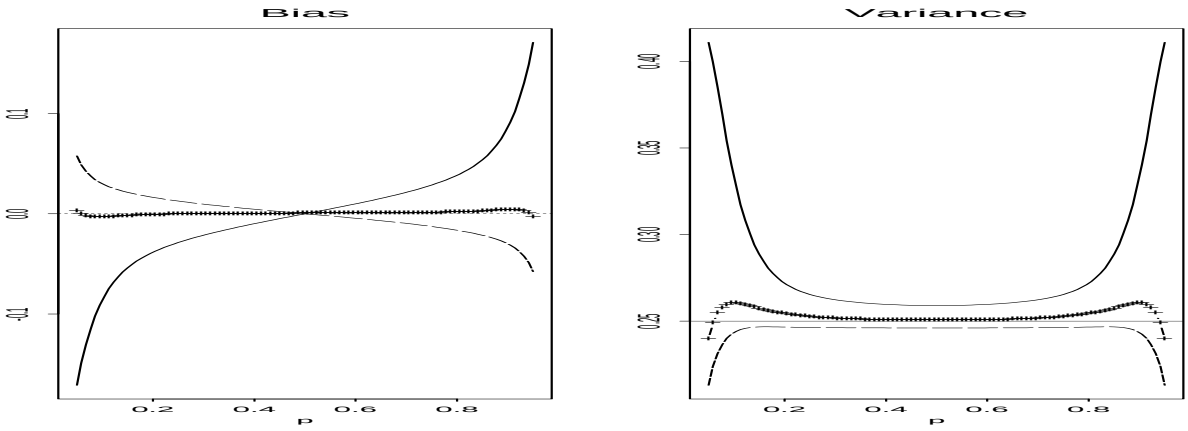

Figure 1: Comparison of the mean (left panel) and variance (right panel) of the arcsine transformations for Binomial$(30, p)$ with $c = 0$ (solid line), $c = \frac{1}{4}$ (+ line) and $c = \frac{3}{8}$ (dashed line).

Note that the mean-matching VST is different from the more conventional VST that optimally stabilizes the variance. Take the binomial distribution as an example. In this case the VST is an arcsine transformation. Let $X_1, ..., X_m \overset{iid}{\sim}$ Bernoulli(p) and then $X = \sum_{i=1}^{m} X_i \sim$ Binomial(m, p). Figure 1 compares the mean and variance of three arcsine transformations of the form

$$\arcsin\left(\sqrt{\frac{X+c}{m+2c}}\right)$$

for the binomial variable X with $m = 30$. The choice of $c = 0$ gives the usual arcsine transformation, $c = 3/8$ optimally stabilizes the variance asymptotically, and $c = 1/4$ yields the mean-matching arcsine transformation. The left panel of Figure 1 plots the bias

$$\sqrt{m}(E_p \arcsin(\sqrt{(X+c)/(m+2c)}) - \arcsin(\sqrt{p}))$$

as a function of p for $c = 0$, $c = \frac{1}{4}$ and $c = \frac{3}{8}$. It is clear from the plot that $c = \frac{1}{4}$ is the best choice among the three for matching the mean. On the other hand, the arcsine transformation with $c = 0$ yields significant bias and the transformation with $c = \frac{3}{8}$ also produces noticeably larger bias. The right panel plots the variance of $\sqrt{m}\arcsin(\sqrt{(X+c)/(m+2c)})$ for $c = 0$, $c = \frac{1}{4}$ and $c = \frac{3}{8}$. Interestingly, over a wide range of values of p near the center the arcsine transformation with $c = \frac{1}{4}$ is even slightly better than the case with $c = \frac{3}{8}$ and clearly $c = 0$ is the worst choice of the three. Figure 2 below shows similar behavior for the Poisson case.

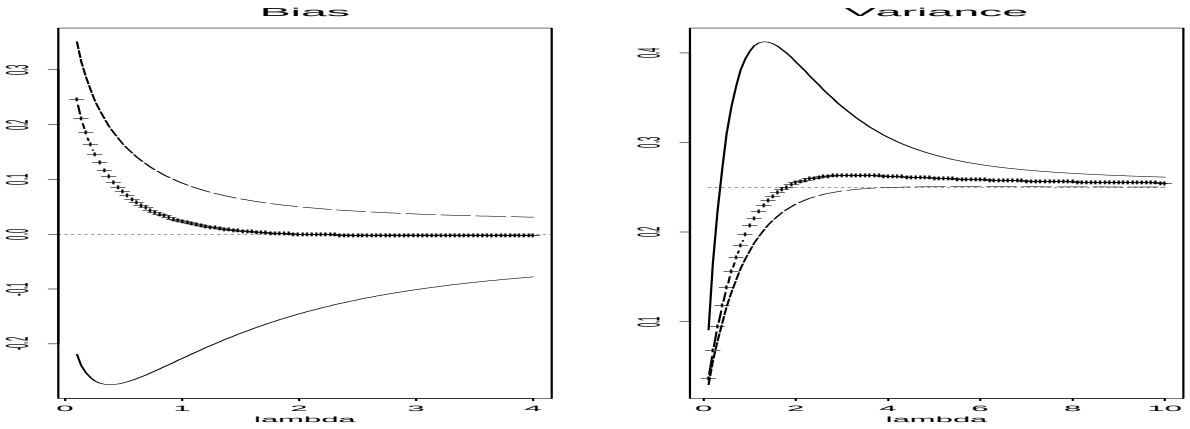

Figure 2: Comparison of the mean (left panel) and variance (right panel) of the root transformations for Poisson(λ) with $c = 0$ (solid line), $c = \frac{1}{4}$ (+ line) and $c = \frac{3}{8}$ (dashed line).

Let us now consider the Gamma distribution as an example for the continuous case. The VST in this case is a log transformation. Let $X_1, ..., X_m \overset{iid}{\sim}$ Exponential(λ). Then $X = \sum_{i=1}^{m} X_i \sim$ Gamma(m, λ). Figure 3 compares the mean and variance of two log transformations of the form

$$\ln\left(\frac{X}{m-c}\right) \tag{9}$$

for the Gamma variable X with $\lambda = 1$ and m ranging from 3 to 40. The choice of $c = 0$ gives the usual log transformation, and $c = 1/2$ yields the mean-matching log transformation. The left panel of Figure 3 plots the bias as a function of m for $c = 0$ and $c = \frac{1}{2}$. It is clear from the plot that $c = \frac{1}{2}$ is a much better choice than $c = 0$ for matching the mean. It is interesting to note that in this case there do not exist constants a and b that optimally stabilize the variance. The right panel plots the variance of $\sqrt{m}\ln(X)$, i.e., $c = 0$, as a function of m. In this case, it is obvious that the variances are the same with $c = 0$ and $c = 1/2$ for the variable in (9).

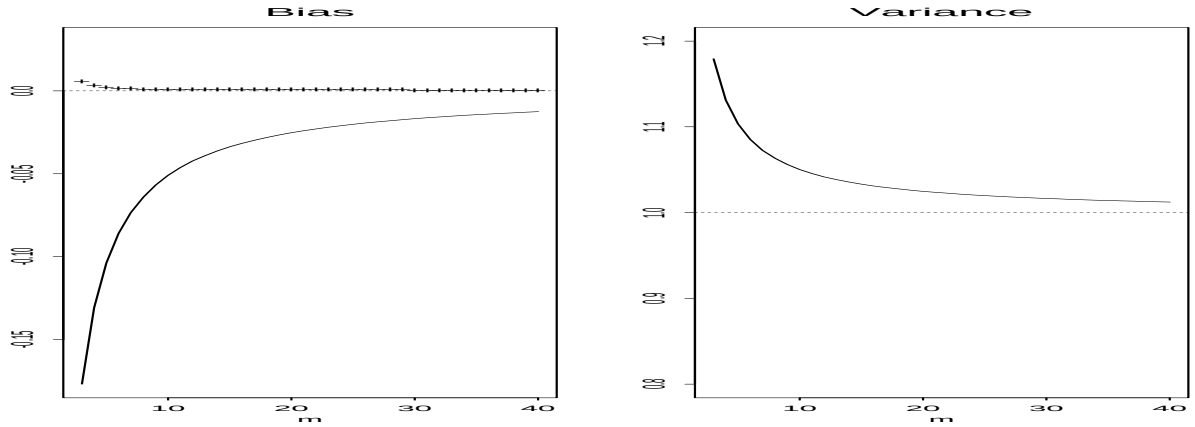

Figure 3: Comparison of the mean (left panel) and variance (right panel) of the log transformations for Gamma(m, λ) with $c = 0$ (solid line) and $c = \frac{1}{2}$ (+ line).

3 Nonparametric regression in exponential families

As mentioned earlier, our main interest in the present paper is to develop new techniques for nonparametric regression in exponential families. A major step is the mean-matching VST developed in Section 2. Suppose we observe

$$Y_i \overset{ind.}{\sim} NQ(f(t_i)), \;\; i = 1, ..., n, \; t_i = \frac{i}{n} \tag{10}$$

and wish to estimate the mean function $f(t)$. In this setting, for the five NEF-QVF families discussed in the last section the noise is not additive and non-Gaussian. Applying standard nonparametric regression methods directly to the data $\{Y_i\}$ in general do not yield desirable results. Our strategy is to use the mean-matching VST to reduce this problem to a standard Gaussian regression problem where

$$\tilde{Y}_j \sim N\left(G\left(f\left(t_j\right)\right), \frac{T}{n}\right), t_j = j/T, j = 1, 2, \ldots, T.$$

Here G is the VST defined in (2), and T will be specified later.

We begin by dividing the interval into T equi-length subintervals and let Q_i be the sum of observations on the i-th subinterval $I_i = [\frac{i-1}{T}, \frac{i}{T})$, $i = 1, 2, \ldots T$. Set $m = \frac{n}{T}$. The sums $\{Q_i\}$ can be treated as observations for a Gaussian regression directly, but this is often a heteroscedastic problem. Instead, we apply the mean-matching VST discussed in Section 2, and then treat $H_m(Q_i)$ as new regression observations. The constants a and b are chosen as in Equation (8) to match the means. We will estimate $G\left(f\left(t_i\right)\right)$ first, then take a transformation of the estimator to estimate the mean function f. After the original regression problem is turned into a Gaussian regression problem through the mean-matching VST, in principle any good nonparametric Gaussian regression method can be

applied to the transformed data $\{\tilde{Y}_j\}$ to construct an estimate of $G(f(\cdot))$. The general ideas for our approach can be summarized as follows.

1. **Binning:** Divide $\{Y_i\}$ into T equal length intervals between 0 and 1. Let $Q_1, Q_2, ..., Q_T$ be the sum of the observations in each of the intervals. A suitable choice of T will be given in Section 3.2.

2. **VST:** Let $Y_j^* = H_m(Q_i)$, $i = 1, \cdots, T$, and treat $Y^* = (Y_1^*, Y_2^*, \ldots, Y_T^*)$ as the new equi-spaced sample for a nonparametric Gaussian regression problem.

3. **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 VST:** Estimate the mean function f by $\widehat{f} = G^{-1}\left(\widehat{G(f)}\right)$. We define $G^{-1}(a) = 0$ when $a < 0$ in the case of Negative Binomial and NEF-GHS distributions.

3.1 Wavelet thresholding

As mentioned earlier, through the mean-matching VST the original problem of regression in exponential families is turned into a standard homoscedastic Gaussian nonparametric regression problem and then a Gaussian regression procedure can be applied. In the present paper we use wavelet block thresholding to construct the final estimator of the regression function. Before we can give a detailed description of our procedure, we need a brief review of basic notation and definitions.

Let $\{\phi,\ \psi\}$ be a pair of father and mother wavelets. The functions ϕ and ψ are assumed to be compactly supported and $\int \phi = 1$. Dilation and translation of ϕ and ψ generates an orthonormal wavelet basis. For simplicity in exposition, in the present paper we work with periodized wavelet bases on $[0, 1]$. Let

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

where $\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)$ and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$. The collection $\{\phi_{j_0,k}^p,\ k = 1, \ldots, 2^{j_0};\ \psi_{j,k}^p,\ j \geq j_0 \geq 0, k = 1, ..., 2^j\}$ is then an orthonormal basis of $L^2[0, 1]$, provided the primary resolution level j_0 is large enough to ensure that the support of the scaling functions and wavelets at level j_0 is not the whole of $[0, 1]$. 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 Daubechies (1992) and Strang (1992) for further details

about the wavelets and discrete wavelet transform. A square-integrable function f on $[0,1]$ can be expanded into a wavelet series:

$$f(t) = \sum_{k=1}^{2^{j_0}} \tilde{\theta}_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{j,k}\psi_{j,k}(t) \tag{11}$$

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.

3.2 Wavelet procedure for generalized regression

Suppose we observe $Y_1, \ldots, Y_n$ as in (10). Set $J = J_n = \lfloor \log_2 n^{3/4} \rfloor$. Let $T = 2^J$ and $m = n/T$. Divide the interval $(0,1]$ into T equal-length subintervals, $(\frac{j-1}{T}, \frac{j}{T}]$ for $j = 1, 2, .., T$. Set

$$Q_j = \sum_{(j-1)\frac{n}{m}+1 \le i \le j\frac{n}{m}} Y_i. \tag{12}$$

Let

$$Y_j^* = \sqrt{m}G(\frac{Q_j + a}{m + b}), \quad i = 1, \cdots, T, \tag{13}$$

and treat $Y^* = (Y_1^*, \ldots, Y_T^*)$ as the new equi-spaced sample for a nonparametric Gaussian regression problem.

Apply the discrete wavelet transform to the binned and transformed data Y^*, and let $U = T^{-\frac{1}{2}}WY^*$ be the empirical wavelet coefficients, where W is the discrete wavelet transformation matrix. Write

$$U = (\widetilde{y}_{j_0,1}, \cdots, \widetilde{y}_{j_0,2^{j_0}}, y_{j_0,1}, \cdots, y_{j_0,2^{j_0}}, \cdots, y_{J-1,1}, \cdots, y_{J-1,2^{J-1}})'. \tag{14}$$

Here $\widetilde{y}_{j_0,k}$ are the gross structure terms at the lowest resolution level, and $y_{j,k}$ ($j = j_0, \cdots, J-1, k = 1, \cdots, 2^j$) are empirical wavelet coefficients at level j which represent fine structure at scale 2^j. The empirical wavelet coefficients can then be written as

$$y_{j,k} = \theta_{j,k} + \epsilon_{j,k} + \frac{1}{\sqrt{n}}z_{j,k} + \xi_{j,k}, \tag{15}$$

where $\theta_{j,k}$ are the true wavelet coefficients of $G(f)$, $\epsilon_{j,k}$ are "small" deterministic approximation errors, $z_{j,k}$ are i.i.d. $N(0,1)$, and $\xi_{j,k}$ are some "small" stochastic errors. The theoretical calculations given in Section 6 will show that both the approximation errors $\epsilon_{j,k}$ and the stochastic errors $\xi_{j,k}$ are negligible. For example, Proposition 2 in Section 6 shows that the tail probability $P(|\xi_j| > a)$ decays faster than any polynomial of n and is thus negligible relative to the Gaussian noise $\frac{1}{\sqrt{n}}z_{j,k}$. If these negligible errors are ignored then we have

$$y_{j,k} \approx \theta_{j,k} + \frac{1}{\sqrt{n}}z_{j,k}, \tag{16}$$

which is the idealized Gaussian sequence model with noise level $\sigma = 1/\sqrt{n}$.

We shall apply the BlockJS procedure (Cai, 1999) to the empirical coefficients $y_{j,k}$ as if they are observed as in (16). More specifically, at each resolution level j, the empirical wavelet coefficients $y_{j,k}$ are grouped into nonoverlapping blocks of length L. As in the sequence estimation setting let $B_j^i = \{(j,k) : (i-1)L + 1 \le k \le iL\}$ and let $S_{j,i}^2 \equiv \sum_{(j,k)\in B_j^i} y_{j,k}^2$. A modified James-Stein shrinkage rule is then applied to each block B_j^i, i.e.,

$$\hat{\theta}_{j,k} = \left(1 - \frac{\lambda_* L}{n S_{j,i}^2}\right)_+ y_{j,k} \quad \text{for } (j,k) \in B_j^i, \tag{17}$$

where $\lambda_* = 4.50524$ is the solution to the equation $\lambda_* - \log \lambda_* = 3$. For the gross structure terms at the lowest resolution level j_0, we set $\hat{\tilde{\theta}}_{j_0,k} = \tilde{y}_{j_0,k}$. The estimate of $G(f(\cdot))$ at the equally spaced sample points $\{\frac{i}{T} : i = 1, \cdots, T\}$ is then obtained by applying the inverse discrete wavelet transform (IDWT) to the denoised wavelet coefficients. That is, $\{G(f(\frac{i}{T})) : i = 1, \cdots, T\}$ is estimated by $\widehat{G(f)} = \{\widehat{G(f(\frac{i}{T}))} : i = 1, \cdots, T\}$ with $\widehat{G(f)} = T^{\frac{1}{2}} W^{-1} \cdot \hat{\theta}$. The estimate of the whole function $G(f)$ is given by

$$\widehat{G(f(t))} = \sum_{k=1}^{2^{j_0}} \hat{\tilde{\theta}}_{j_0,k} \phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{j,k} \psi_{j,k}(t).$$

The mean function f is estimated by

$$\widehat{f}(t) = G^{-1}(\widehat{G(f(t))}). \tag{18}$$

Figure 4 shows the steps of the procedure for an example in the case of nonparametric Gamma regression.

Remark 1 The choice of block size $L = \log n$ is important for achieving simultaneously the optimal global and local adaptivity. Theorems 1 and 2 in Section 4 do not hold simultaneously if, for example, $L = (\log n)^\rho$ for $\rho \neq 1$. The thresholding constant λ_* is chosen based on a block thresholding oracle risk inequality for $L = \log n$ in a similar but more delicate way than the threshold $\sqrt{2 \log n}$ is chosen in term-by-term thresholding. See Cai (1999) for further details.

4 Theoretical properties

We shall now investigate the asymptotic properties of the procedure proposed in Section 3. Numerical results will be given in Section 5.

We study the theoretical properties of our procedure over the Besov spaces that are by now standard for the analysis of wavelet regression methods. Besov spaces are a very rich

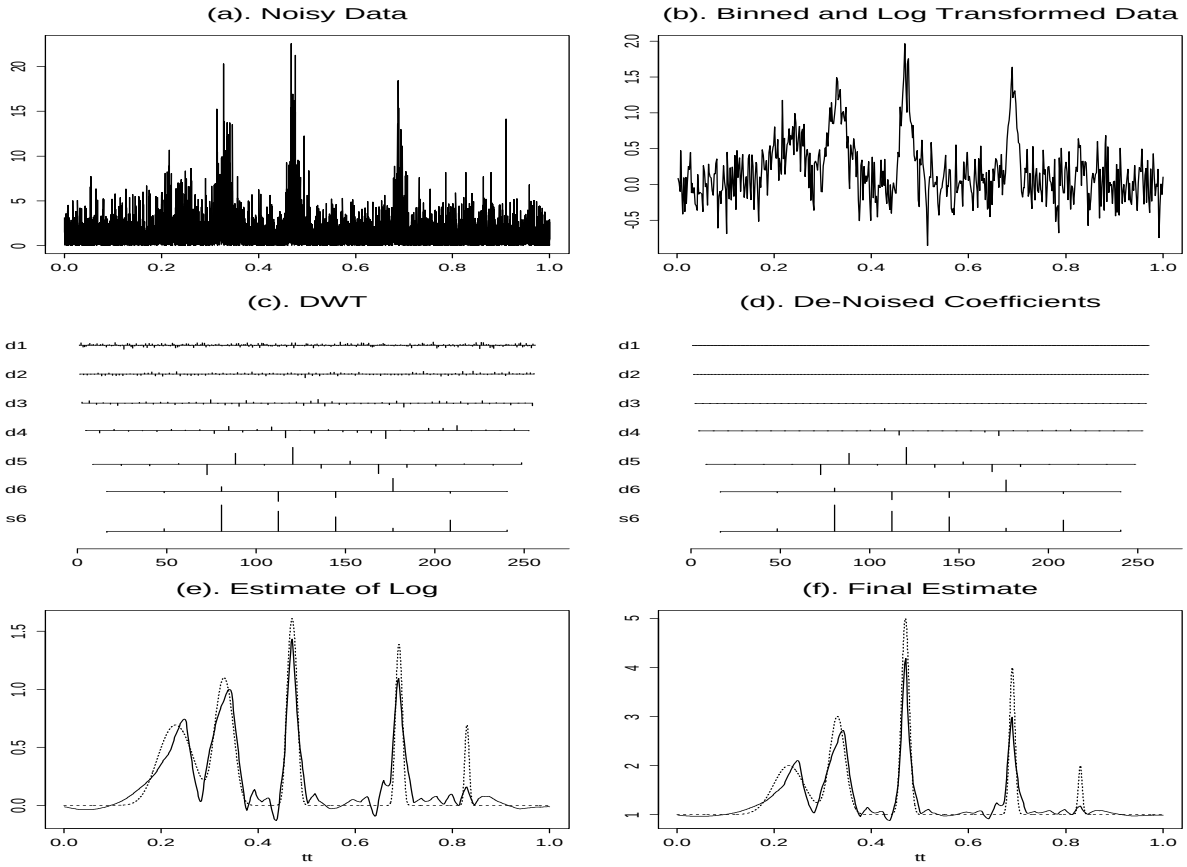

Figure 4: An example of nonparametric Gamma regression using the mean-matching VST and wavelet block thresholding.

class of function spaces and contain as special cases many traditional smoothness spaces such as Hölder and Sobolev Spaces. Roughly speaking, the Besov space $B^{\alpha}_{p,q}$ contains functions having α bounded derivatives in L^p norm, the third parameter q gives a finer gradation of smoothness. Full details of Besov spaces are given, for example, in Triebel (1983) and DeVore and Popov (1988). For a given r-regular mother wavelet ψ with $r > \alpha$ and a fixed primary resolution level j_0, the Besov sequence norm $\|\cdot\|_{b^{\alpha}_{p,q}}$ of the wavelet coefficients of a function f is then defined by

$$\|f\|_{b^{\alpha}_{p,q}} = \|\underline{\xi}_{j_0}\|_p + \left(\sum_{j=j_0}^{\infty}(2^{js}\|\underline{\theta}_j\|_p)^q\right)^{\frac{1}{q}} \tag{19}$$

where $\underline{\xi}_{j_0}$ is the vector of the father wavelet coefficients at the primary resolution level j_0, $\underline{\theta}_j$ is the vector of the wavelet coefficients at level j, and $s = \alpha + \frac{1}{2} - \frac{1}{p} > 0$. Note that the Besov function norm of index (α, p, q) of a function f is equivalent to the sequence norm

(19) of the wavelet coefficients of the function. See Meyer (1992). We define

$$B_{p,q}^{\alpha}(M) = \left\{ f; \|f\|_{b_{p,q}^{\alpha}} \leq M \right\}. \tag{20}$$

and

$$F_{p,q}^{\alpha}(M,\epsilon) = \{f : f \in B_{p,q}^{\alpha}(M), f(x) \geq \epsilon \text{ for all } x \in [0,1]\}. \tag{21}$$

Note that when f is bounded below from 0 and above from a constant, the condition $f \in B_{p,q}^{\alpha}(M)$ is equivalent to that there exists $M' > 0$ such that $G(f) \in B_{p,q}^{\alpha}(M')$. See Runst (1986). The following theorem shows that our estimator achieves optimal global adaptation under integrated squared error for a wide range of Besov balls.

Theorem 1 *Suppose the wavelet ψ is r-regular. Let $X_i \sim NQ(f(t_i)),\ i = 1, ..., n,\ t_i = \frac{i}{n}$. Let $T = cn^{\frac{3}{4}}$. Then the estimator $\hat{f}$ defined in* (18) *satisfies*

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

For functions of spatial inhomogeneity, the local smoothness of the functions varies significantly from point to point and global risk given in Theorem 1 cannot wholly reflect the performance of estimators at a point. We use the local risk measure

$$R(\widehat{f}(t_0),\ f(t_0)) = E(\widehat{f}(t_0) - f(t_0))^2 \tag{22}$$

for spatial adaptivity.

The local smoothness of a function can be measured by its local Hölder smoothness index. For a fixed point $t_0 \in (0,1)$ and $0 < \alpha \leq 1$, define the local Hölder class $\Lambda^{\alpha}(M, t_0, \delta)$ as follows:

$$\Lambda^{\alpha}(M, t_0, \delta) \ = \ \{f : |f(t) - f(t_0)| \leq M\,|t - t_0|^{\alpha},\ \text{ for } t \in (t_0 - \delta,\ t_0 + \delta)\}.$$

If $\alpha > 1$, then

$$\Lambda^{\alpha}(M, t_0, \delta) \ = \ \{f : |f^{(\lfloor \alpha \rfloor)}(t) - f^{(\lfloor \alpha \rfloor)}(t_0)| \leq M\,|t - t_0|^{\alpha'} \text{ for } t \in (t_0 - \delta,\ t_0 + \delta)\}$$

where $\lfloor \alpha \rfloor$ is the largest integer less than α and $\alpha' = \alpha - \lfloor \alpha \rfloor$.

In Gaussian nonparametric regression setting, it is a well known fact that for estimation at a point, one must pay a price for adaptation. The optimal rate of convergence for estimating $f(t_0)$ over function class $\Lambda^{\alpha}(M, t_0, \delta)$ with α completely known is $n^{-2\alpha/(1+2\alpha)}$. Lepski (1990) and Brown and Low (1996) showed that one has to pay a price for adaptation of at least a logarithmic factor. It is shown that the local adaptive minimax rate over the Hölder class $\Lambda^{\alpha}(M, t_0, \delta)$ is $(\log n/n)^{2\alpha/(1+2\alpha)}$.

The following theorem shows that our estimator achieves optimal local adaptation with the minimal cost.

Theorem 2 *Suppose the wavelet ψ is r-regular with $r \geq \alpha > 1/6$. Let $t_0 \in (0,1)$ be fixed. Let $X_i \sim NQ(f(t_i))$, $i = 1, ..., n$, $t_i = \frac{i}{n}$. Let $T = cn^{\frac{3}{4}}$. Then the estimator $\hat{f}$ defined in* (18) *satisfies*

$$\sup_{f \in \Lambda^\alpha(M, t_0, \delta)} E(\widehat{f}(t_0) - f(t_0))^2 \leq C \cdot (\frac{\log n}{n})^{\frac{2\alpha}{1+2\alpha}}. \tag{23}$$

Theorem 2 shows that the estimator automatically attains the local adaptive minimax rate for estimating functions at a point, without prior knowledge of the smoothness of the underlying functions.

5 Numerical study

In this section we study the numerical performance of our estimator. The procedure introduced in Section 3 is easily implementable. We shall first consider simulation results and then apply our procedure in the analysis of two real data sets.

5.1 Simulation results

As discussed the Section 2, there are several different versions of the VST in the literature and we have emphasized the importance of using the mean-matching VST for theoretical reasons. We shall now consider the effect of the choice of the VST on the numerical performance of the resulting estimator. To save space we only consider the Poisson and Bernoulli cases. We shall compare the numerical performance of the mean-matching VST with those of classical transformations by Bartlett (1936) and Anscombe (1948) using simulations. The transformation formulae are given as follows. (In the following tables and figures, we shall use MM for mean-matching.)

	MM	Bartlett	Anscombe
Poi(λ)	$\sqrt{X + 1/4}$	$\sqrt{X}$	$\sqrt{X + 3/8}$
Bin(m, p)	$\sin^{-1}\sqrt{\frac{X+1/4}{m+1/2}}$	$\sin^{-1}\sqrt{\frac{X}{m}}$	$\sin^{-1}\sqrt{\frac{X+3/8}{m+3/4}}$

.

Four standard test functions, Doppler, Bumps, Blocks and HeaviSine, representing different level of spatial variability are used for the comparison of the three VSTs. See Donoho and Johnstone (1994) for the formulae of the four test functions. These test functions are suitably normalized so that they are positive and taking values between 0 and 1 (in the binomial case). Sample sizes vary from a few hundred to a few hundred thousand. We use Daubechies' compactly supported wavelet *Symmlet* 8 for wavelet transformation. As is the case in general, it is possible to obtain better estimates with different wavelets for different signals. But for uniformity, we use the same wavelet for all cases. Although our

asymptotic theory only gives a justification for the choice of the bin size of order $n^{1/4}$ due to technical reasons, our extensive numerical studies have shown that the procedure works well when the number of counts in each bin is between 5 and 10 for the Poisson case, and similarly for the Bernoulli case the average number of successes and failures in each bin is between 5 and 10. We follow this guideline in our simulation study. Table 1 reports the average squared errors over 100 replications. A graphical presentation is given in Figure 5.

Bernoulli	MM	Bartlett	Anscombe		MM	Bartlett	Anscombe
Doppler				Bumps			
1280	12.117	11.197	12.673	1280	7.756	8.631	7.896
5120	3.767	3.593	4.110	5120	7.455	7.733	7.768
20480	1.282	1.556	1.417	20480	3.073	3.476	3.450
81920	0.447	0.772	0.540	81920	1.203	1.953	1.485
327680	0.116	0.528	0.169	327680	0.331	1.312	0.535
Blocks				HeaviSine			
1280	18.451	17.171	18.875	1280	2.129	2.966	2.083
5120	7.582	6.911	7.996	5120	0.842	1.422	0.860
20480	3.288	3.072	3.545	20480	0.549	0.992	0.603
81920	1.580	1.587	1.737	81920	0.285	0.681	0.339
327680	0.594	0.781	0.681	327680	0.138	0.532	0.195
Poisson	MM	Bartlett	Anscombe		MM	Bartlett	Anscombe
Doppler				Bumps			
640	8.101	8.282	8.205	640	107.860	103.696	109.023
2560	3.066	3.352	3.160	2560	70.034	68.616	70.495
10240	1.069	1.426	1.146	10240	24.427	24.268	24.653
40960	0.415	0.743	0.502	40960	9.427	9.469	9.620
163840	0.108	0.461	0.190	163840	3.004	3.098	3.204
Blocks				HeaviSine			
640	12.219	12.250	12.320	640	2.831	3.552	2.851
2560	5.687	6.209	5.724	2560	0.849	1.468	0.884
10240	2.955	3.363	3.005	10240	0.425	0.852	0.501
40960	1.424	1.773	1.495	40960	0.213	0.560	0.298
163840	0.508	0.890	0.573	163840	0.118	0.455	0.206

Table 1: Mean squared error (MSE) from 100 replications. The MSE is in units of 10^{-3} for Bernoulli case and 10^{-2} for Poisson case.

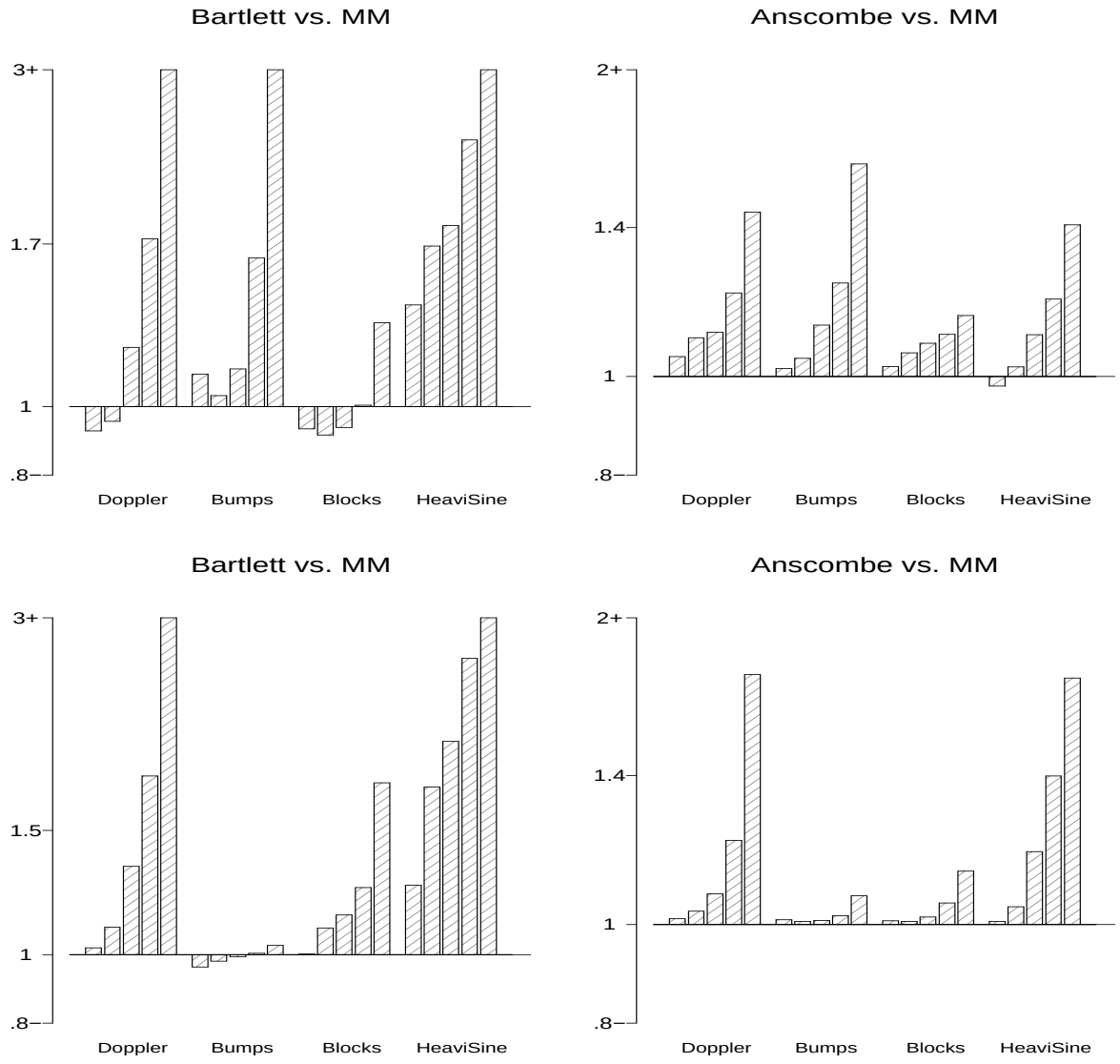

Figure 5: Left panels: The vertical bars represent the ratios of the MSE of the estimator using the Bartlett VST to the corresponding MSE of our estimator using the mean-matching VST. Right Panels: The bars represent the ratios of the MSE of the estimator using the Anscombe VST to the corresponding MSE of the estimator using the mean-matching VST. The higher the bar the better the relative performance of our estimator. The bars are plotted on a log scale and the original ratios are truncated at the value 3 for the Bartlett VST and at 2 for the Anscombe VST. For each signal the bars are ordered from left to right in the order of increasing sample size. The top row is for the Bernoulli case and the bottom row for the Poisson case.

Table 1 compares the performance of three nonparametric function estimators constructed from three VSTs and wavelet block thresholding for Bernoulli and Poisson regressions. The three VSTs are the mean-matching, Bartlett and Anscombe transformations given above. The results show the mean-matching VST outperforms the classical transformations for nonparametric estimation in most cases. The improvement becomes more significant as the sample size increases.

In the Poisson regression, the mean-matching VST outperforms the Bartlett VST in 17 out of 20 cases and uniformly outperforms the Anscombe VST in all 20 cases. The case of Bernoulli regression is similar: the mean-matching VST is better than the Bartlett VST in 15 out of 20 cases and better than the Anscombe VST in 19 out of 20 cases. Although the mean-matching VST does not uniformly dominate either the Bartlett VST or the Anscombe VST, the improvement of the mean-matching VST over the other two VSTs is significant as the sample size increases for all four test functions. The simulation results show that mean-matching VST yields good numerical results in comparison to other VSTs. These numerical findings is consistent with the theoretical results given in Section 4 which show that the estimator constructed from the mean-matching VST enjoys desirable adaptivity properties.

We have so far considered the effect of the choice of VST on the performance of the estimator. We now discuss the Poisson case in more detail and compare the numerical performance of our procedure with other estimators proposed in the literature. As mentioned in the introduction, Besbeas, De Feis and Sapatinas, T. (2004) carried out an extensive simulation studies comparing several nonparametric Poisson regression estimators including the estimator given in Donoho (1993). The estimator in Donoho (1993) was constructed by first applying the Anscombe (1948) VST to the binned data and by then using a wavelet procedure with a global threshold such as VisuShrink (Donoho and Johnstone (1994)) to the transformed data as if the data were actually Gaussian. The simulation study carried out in Besbeas, De Feis and Sapatinas (2004) showed Donoho's method is comparable with other Poisson regression procedures. For reasons of space, we thus only compare our procedure with Donoho's estimator. Figure 6 plots the ratios of the MSE of Donoho's estimator to the corresponding MSE of our estimator. The results show that our estimator outperforms Donoho's estimator in all but one case and in many cases our estimator has the MSE less than one half and sometimes even one third of that of Donoho's estimator. This combined with those simulation results given in Besbeas, De Feis and Sapatinas (2004) show that our procedure performs well numerically in comparison with other nonparametric Poisson regression estimators.

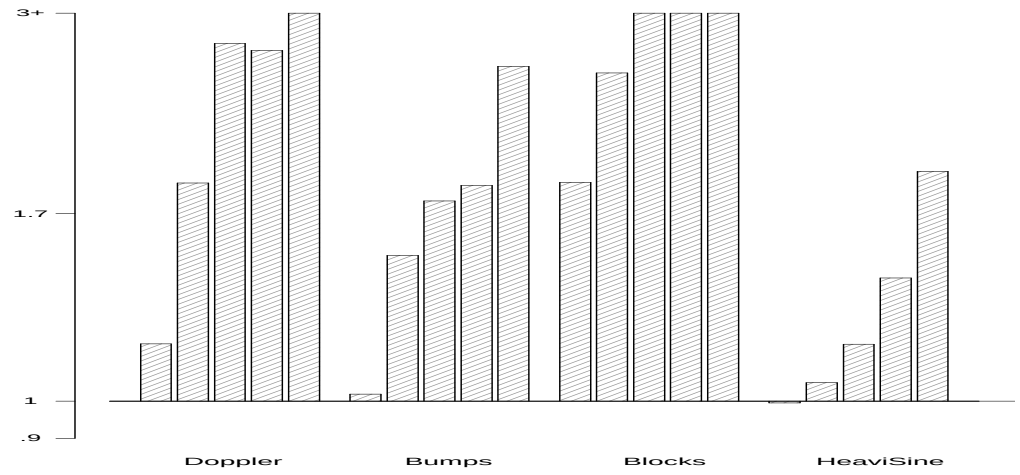

Figure 6: The vertical bars represent the ratios of the MSE of Donoho's estimator to the corresponding MSE of our estimator. The higher the bar the better the relative performance of our estimator. The bars are plotted on a log scale and the original ratios are truncated at the value 3. For each signal the bars are ordered from left to right in the order of increasing sample size.

5.2 Real data applications

We now demonstrate our estimation method in the analysis of two real data sets, a gamma-ray burst data set (GRBs) and a packet loss data set. These two data sets have been discussed in Kolaczyk and Nowak (2005).

Cosmic gamma-ray bursts were first discovered in the late 1960s. In 1991, NASA launched the Compton Gamma Ray Observatory and its Burst and Transient Source Explorer (BATSE) instrument, a sensitive gamma-ray detector. Much burst data has been collected since then, followed by extensive studies and many important scientific discoveries during the past few decades, however the source of GRBs remains unknown (Kaneko, 2005). For more details see the NASA website http://www.batse.msfc.nasa.gov/batse/. GRBs seem to be connected to massive stars and become powerful probes of the star formation history of the universe. However not many redshifts are known and there is still much work to be done to determine the mechanisms that produce these enigmatic events. Statistical methods for temporal studies are necessary to characterize their properties and hence to identify the physical properties of the emission mechanism. One of the difficulties in analyzing the time profiles of GRBs is the transient nature of GRBs which means that the usual assumptions for Fourier transform techniques do not hold (Quilligan et al. (2001)). We may model the time series data by an inhomogeneous Poisson process, and

apply our wavelet procedure. The data set we use is called BATSE 551 with the sample size 7808. In Figure 7, the top panel is the histogram of the data with 1024 bins such that the number of observations in each bin would be between 5 and 10. In fact we have on average 7.6 observations. The middle panel is the estimate of the intensity function using our procedure. If we double the width of each bin, i.e., the total number of bins is now 512, the new estimator in the bottom panel is noticeably different from previous one since it does not capture the fine structure from time 200 to 300. The study of the number of pulses in GRBs and their time structure is important to provide evidence for rotation powered systems with intense magnetic fields and the added complexity of a jet.

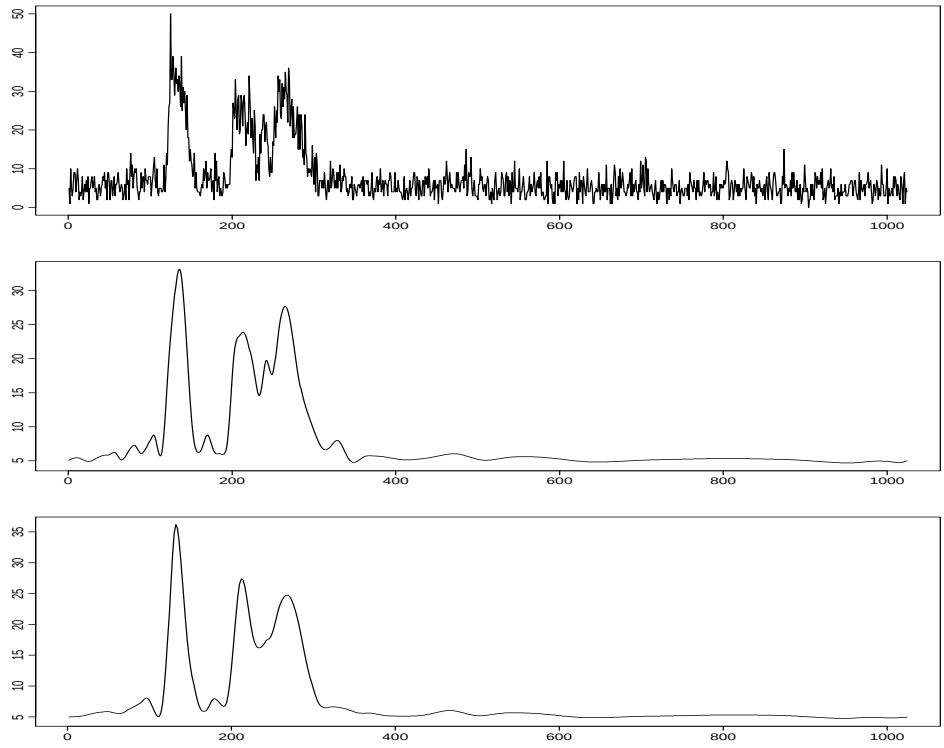

Figure 7: Gamma-ray burst. The top panel is the histogram of BATSE 551 with 1024 bins. The middle panel is our estimator based on 1024 bins, and the bottom panel is the estimator with 512 bins.

Packet loss describes an error condition in internet traffic in which data packets appear to be transmitted correctly at one end of a connection, but never arrive at the other. So, if 10 packets were sent out, but only 8 made it through, then there would be 20% overall packet loss. The following data were originally collected and analyzed by Yajnik et al. (1999). The objective is to understand packet loss by modeling. It measures the reliability of a connection and is of fundamental importance in network applications such as audio/video

conferencing and Internet telephony. Understanding the loss seen by such applications is important in their design and performance analysis. The measurements are of loss as seen by packet probes sent at regular time intervals. The packets were transmitted from the University of Massachusetts at Amherst to the Swedish Institute of Computer Science. The records note whether each packet arrived or was lost. It is a Bernoulli time series, and can be naturally modeled as Binomial after binning the data. The following figure gives the histogram and our corresponding estimator. The average sum of failures in each bin is about 10. The estimator in Kolaczyk and Nowak (2005) is comparable to ours. But our procedure is more easily implemented.

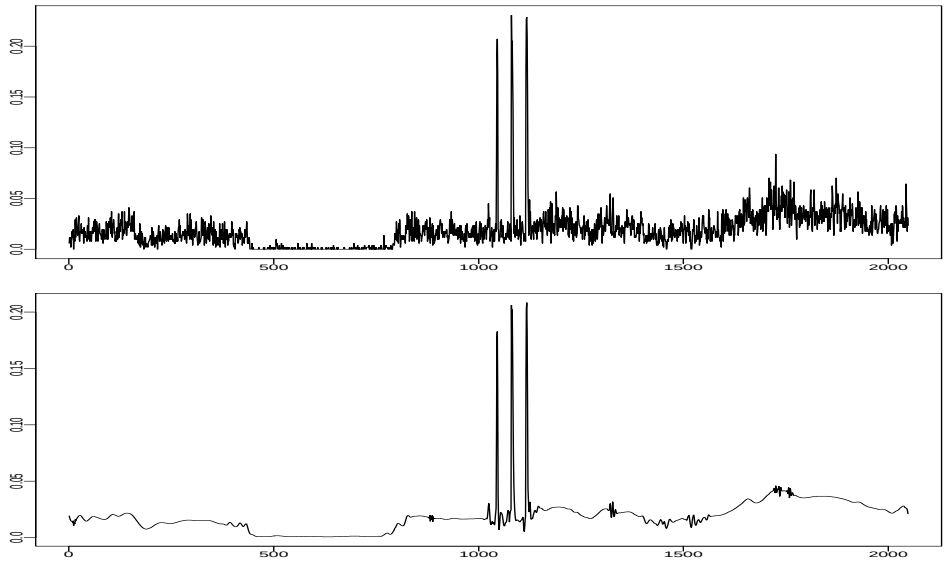

Figure 8: Packet loss data. The first plot is the histogram with 2048 bins. The second plot is our estimator based on the binned data.

6 Proofs

Proof of Lemma 1 . We only prove the first part of the lemma. By Taylor expansion we write

$$G\left(\frac{X+a}{m+b}\right) - G\left(\mu(\eta)\right) = T_1 + T_2 + T_3 + T_4$$

where

$$\begin{aligned} T_1 &= G'\left(\mu(\eta)\right)\left(\frac{X+a}{m+b} - \mu(\eta)\right),\ T_2 = \frac{1}{2}G''\left(\mu(\eta)\right)\left(\frac{X+a}{m+b} - \mu(\eta)\right)^2 \\ T_3 &= \frac{1}{6}G'''\left(\mu(\eta)\right)\left(\frac{X+a}{m+b} - \mu(\eta)\right)^3, T_4 = \frac{1}{24}G^{(4)}\left(b\left(\eta\right)\right)\left(\frac{X+a}{m+b} - \mu(\eta)\right)^4 \end{aligned}$$

By definition, $G'(\mu(\eta)) = I(\eta)^{-1/2}$ with $I(\eta) = \mu'(\eta)$, then

$$G''(\mu(\eta))\,\mu'(\eta) = -\frac{1}{2} I(\eta)^{-3/2} I'(\eta)$$

i.e.,

$$G''(\mu(\eta)) = -\frac{1}{2} I(\eta)^{-5/2} I'(\eta)$$

then

$$\begin{aligned} ET_1 &= I(\eta)^{-1/2} \frac{a - \mu(\eta) b}{m + b} \\ ET_2 &= -\frac{1}{4} I(\eta)^{-5/2} I'(\eta) \left[\left(\frac{a - \mu(\eta) b}{m + b} \right)^2 + \frac{m I(\eta)}{(m + b)^2} \right]. \end{aligned}$$

Note that $G'(\mu(\eta))$ is uniformly bounded on Θ_0 by the assumption in the lemma, then we have

$$\begin{aligned} E(T_1 + T_2) &= \frac{m}{(m + b)^2} I(\eta)^{-1/2} \left[a - \mu(\eta) b - \frac{1}{4} (\log I(\eta))' \right] + O\left(\frac{1}{m^2} \right) \\ &= \frac{m}{(m + b)^2 I(\eta)^{1/2}} \left(a - \mu(\eta) b - \frac{\mu''(\eta)}{4\mu'(\eta)} \right) + O\left(\frac{1}{m^2} \right) \\ &= \frac{1}{m I(\eta)^{1/2}} \left(a - \mu(\eta) b - \frac{\mu''(\eta)}{4\mu'(\eta)} \right) + O\left(\frac{1}{m^2} \right). \end{aligned}$$

Similarly it can be shown

$$|ET_3| = O\left(\frac{1}{m^2} \right),\ |ET_4| = O\left(\frac{1}{m^2} \right),$$

and so equation (4) is established. ∎

6.1 Proof of the Main Results

The proof of Theorem 1 contains two main steps: coupling and bounding the risk of block thresholding estimators.

6.1.1 Coupling and preparatory results

We shall use the quantile coupling inequality of Komlós, Major and Tusnády (1975) to approximate the binned and transformed data by independent normal variables. The following lemma is a direct consequence of the results given in Komlós, Major and Tusnády (1975) and Zhou (2006).

Lemma 2 *Let $X_i \overset{iid}{\sim} NQ(\mu)$ with variance V for $i = 1, ..., m$ and let $X = \sum_{i=1}^{m} X_i$. Under the assumption in Lemma 1, there exists a standard normal random variable $Z \sim N(0,1)$ and constants $c_1, c_2, c_3 > 0$ not depending on n such that whenever the event $A = \{|X - m\mu| \leq c_1 m\}$ occurs,*

$$|X - m\mu - \sqrt{mV}Z| < c_2 Z^2 + c_3. \tag{24}$$

We shall develop tight bounds for both the deterministic approximation errors $\epsilon_{j,k}$ and the stochastic errors $\xi_{j,k}$ in the decomposition of the empirical wavelet coefficients given in (15). Let $Y = \sqrt{m}H_m(X) = \sqrt{m}G(\frac{X+a}{m+b})$, $\epsilon = \sqrt{m}EY - \sqrt{m}G(\mu)$ and Z be a standard normal variable satisfying (24). Let

$$\xi = \sqrt{m}G(\frac{X+a}{m+b}) - \sqrt{m}G(\mu) - \epsilon - Z \tag{25}$$

and write

$$Y = \sqrt{m}G(\mu) + \epsilon + Z + \xi$$

It follows from Lemma 1 that when m is large, ϵ is "small", $|\epsilon| \leq cm^{-\frac{3}{2}}(1 + o(1))$ for some constant $c > 0$. We shall show, using Lemma 2, that the random variable ξ is "stochastically small".

Lemma 3 *Let $X_i \sim NQ(\mu)$ with variance V , $X = \sum_{i=1}^{m} X_i$ and let Z be the standard normal variable given as in Lemma 2. Let ξ be given as in (25). Then for any integer $i \geq 1$ there exists a constant $C_i > 0$ such that for all $\lambda \geq 1$ and all $a > 0$,*

$$E|\xi|^i \leq C_i m^{-\frac{i}{2}} \quad \text{and} \quad P(|\xi| > a) \leq C_i (a^2 m)^{-\frac{i}{2}}. \tag{26}$$

Proof: By Taylor expansion we write

$$G\left(\frac{X+a}{m+b}\right) - G(\mu) = G'(\mu)\left(\frac{X+a}{m+b} - \mu\right) + \frac{1}{2}G''(\mu^*)\left(\frac{X+a}{m+b} - \mu\right)^2.$$

Then write

$$\xi = \xi_1 + \xi_2 + \xi_3$$

where

$$\begin{aligned}
\xi_1 &= \sqrt{m}G'(\mu)\left(\frac{X+a}{m+b} - \frac{X}{m}\right) - \epsilon = G'(\mu)\frac{am - bX}{(m+b)\sqrt{m}} - \epsilon \\
\xi_2 &= \sqrt{m}G'(\mu)\left(\frac{X}{m} - \mu - \sqrt{\frac{V}{m}}Z\right) = \frac{G'(\mu)}{\sqrt{m}}\left(X - m\mu - \sqrt{mV}Z\right) \\
\xi_3 &= \frac{1}{2}\sqrt{m}G''(\mu^*)\left(\frac{X+a}{m+b} - \mu\right)^2 = \frac{1}{2}\sqrt{m}G''(\mu^*)\left(\frac{X-m\mu}{m+b} + \frac{a-b\mu}{m+b}\right)^2
\end{aligned}$$

It is very easy to see $E|\xi_1|^i \leq C_i m^{-\frac{i}{2}}$ and $E|\xi_3|^i \leq C_i m^{-\frac{i}{2}}$. An application of Lemma 2 implies $E|\xi_2|^i \leq C_i m^{-\frac{i}{2}}$.

The second bound in (26) is a direct consequence of the first one and Markov inequality.

∎

Remark 2 *Variance stabilizing transformation considered in Section 2 is for i.i.d. observations. In the function estimation procedure, observations in each bin are independent but not identically distributed. However, observations in each bin can be treated as i.i.d. random variables through coupling. Let $X_i \sim NQ(\mu_i)$, $i = 1, ..., m$, be independent. Here the means μ_i are "close" but not equal. Let μ' be a value close of the μ_i's. The following argument shows that X_i can be coupled with i.i.d. random variables X_i' where $X_i' \overset{iid}{\sim} NQ(\mu')$. Let F_μ denote the distribution function of $NQ(\mu)$. We couple X_i and X_i' as follows*

$$X_i = F_{\mu_i}^{-1}(U_i) \text{ and } X_i' = F_{\mu'}^{-1}(U_i),\ U_i \overset{iid}{\sim} Uniform[0,1].$$

Let $R = \frac{1}{m}\sum_{i=1}^m (X_i - X_i')$. Assume that $|\mu_i - \mu'| \leq CT^{-d}$ for all i. Then $E|X_i - X_i'|^l \leq CT^{-d}$ for all positive integer l. Since $X_i - X_i'$ are independent, it can be shown that

$$E\left(\sum_{i=1}^m |X_i - X_i'|\right)^l \leq C_l\left[mT^{-d} + \left(mT^{-d}\right)^l\right]$$

by a straightforward expansion of the product $\left(\sum_{i=1}^m |X_i - X_i'|\right)^l$. This implies

$$E|R|^l \leq C_l\left(T^{-dl} + T^{-d}/m^{l-1}\right). \tag{27}$$

Let

$$D = \sqrt{m}\left[G\left(\frac{\sum_{i=1}^m X_i + a}{m+b}\right) - G\left(\frac{\sum_{i=1}^m X_i' + a}{m+b}\right)\right]$$

and $\epsilon' = E(D)$, and $\xi' = D - \epsilon'$. The inequality (27) gives

$$\epsilon' \leq C\sqrt{m}T^{-d} \text{ and } E|\xi'|^l \leq C\left[\left(\sqrt{m}T^{-d}\right)^l + mT^{-d}/m^{l/2}\right].$$

¿From the assumption in Theorems 1 or 2 it is easy to check $\frac{T}{n}(\epsilon')^2 \leq CT^{-2d} = o\left(n^{-2\alpha/(2\alpha+1)}\right)$ and$\sqrt{m}T^{-d}$ converges to 0 as a power of n. Therefore the contribution of ϵ' and ξ' to the final risk bounds in Theorems 1 and 2 is negligible as that of ϵ_i and ξ_i in the proposition below.

Lemmas 1, 2 and 3 together yield the following result.

Proposition 1 *Let $Y_i = G(\frac{Q_j + a}{m+b})$ be given as in (13). Then Y_i can be written as*

$$Y_i = \sqrt{m}G(f(i/T)) + \epsilon_i + Z_i + \xi_i, \quad i = 1, 2, \ldots, T, \tag{28}$$

where $Z_i \overset{i.i.d.}{\sim} N(0,1)$, ϵ_i *are constants satisfying* $|\epsilon_i| \le cm^{-\frac{3}{2}}$ *and consequently for some constant* $C > 0$

$$\frac{1}{n}\sum_{i=1}^{T} \epsilon_i^2 \le C \cdot m^{-4}, \tag{29}$$

and ξ_i *are independent and "stochastically small" random variables satisfying*

$$E|\xi_i|^l \le C_l m^{-\frac{l}{2}} \quad \textit{and} \quad P(|\xi_i| > a) \le C_l (a^2 m)^{-\frac{l}{2}} \tag{30}$$

where $l > 0$, $a > 0$ *and* $C_l > 0$ *is a constant depending on* l *only.*

We need the following moment bounds for an orthogonal transform of independent variables.

Lemma 4 *Let* $X_1, \ldots, X_n$ *be independent variables with* $E(X_i) = 0$ *for* $i = 1, \ldots, n$. *Suppose that* $E|X_i|^k < M_k$ *for all* i *and all* $k > 0$ *with* $M_k > 0$ *some constant not depending on* n. *Let* $Y = WX$ *be an orthogonal transform of* $X = (X_1, ..., X_n)'$. *Then there exist constants* M_k' *not depending on* n *such that* $E|Y_i|^k < M_k'$ *for all* $i = 1, \ldots, n$ *and all* $k > 0$.

From (28) in Proposition 1 we can write $\frac{1}{\sqrt{n}} Y_i = \frac{G(f(i/T))}{\sqrt{T}} + \frac{\epsilon_i}{\sqrt{n}} + \frac{Z_i}{\sqrt{n}} + \frac{\xi_i}{\sqrt{n}}$. Let $(u_{j,k}) = n^{-\frac{1}{2}} W \cdot Y$ be the discrete wavelet transform of the binned and transformed data. Then one may write

$$u_{j,k} = \theta'_{j,k} + \epsilon_{j,k} + \frac{1}{\sqrt{n}} z_{j,k} + \xi_{j,k} \tag{31}$$

where θ'_{jk} are the discrete wavelet transform of $(\frac{\sqrt{p_i}}{\sqrt{T}})$ which are approximately equal to the true wavelet coefficients of $G(f)$, $z_{j,k}$ are the transform of the Z_i's and so are i.i.d. $N(0,1)$ and $\epsilon_{j,k}$ and $\xi_{j,k}$ are respectively the transforms of $(\frac{\epsilon_i}{\sqrt{n}})$ and $(\frac{\xi_i}{\sqrt{n}})$. Then it follows from Proposition 1 that

$$\sum_j \sum_k \epsilon_{j,k}^2 = \frac{1}{n} \sum_i \epsilon_i^2 \le Cm^{-4}. \tag{32}$$

It now follows from Lemma 4 and Proposition 1 that for all $i > 0$ and $a > 0$

$$E|\xi_{j,k}|^i \le C_i'(mn)^{-\frac{i}{2}} \quad \text{and} \quad P(|\xi_{j,k}| > a) \le C_i'(a^2 mn)^{-\frac{i}{2}}. \tag{33}$$

6.2 Risk bound for a single block

We need the following auxiliary results for block thresholding estimators without the normality assumption (see Brown, Cai, Zhang, Zhao and Zhou (2007) for details).

Lemma 5 *Suppose* $y_i = \theta_i + z_i, \quad i = 1, ..., L$, *where* θ_i *are constants and* z_i *are random variables. Let* $S^2 = \sum_{i=1}^{L} y_i^2$ *and let* $\hat{\theta}_i = (1 - \frac{\lambda L}{S^2})_+ y_i$. *Then*

$$E\|\hat{\theta} - \theta\|_2^2 \le \|\theta\|_2^2 \wedge 4\lambda L + 4E\left[\|z\|_2^2 I(\|z\|_2^2 > \lambda L)\right]. \tag{34}$$

Lemma 6 *Let $X \sim \chi^2_L$ and $\lambda > 1$. Then*

$$P(X \geq \lambda L) \leq e^{-\frac{L}{2}(\lambda - \log \lambda - 1)} \quad \textit{and} \quad EXI(X \geq \lambda L) \leq \lambda L e^{-\frac{L}{2}(\lambda - \log \lambda - 1)}. \tag{35}$$

Proposition 2 *Let the empirical wavelet coefficients $u_{j,k} = \theta'_{j,k} + \epsilon_{j,k} + \frac{1}{2\sqrt{n}} z_{j,k} + \xi_{j,k}$ be given as in (31) and let the block thresholding estimator $\hat{\theta}_{j,k}$ be defined as in (17). Then for some constant $C > 0$*

$$E \sum_{(j,k) \in B^i_j} (\hat{\theta}_{j,k} - \theta'_{j,k})^2 \leq \min \left\{ 4 \sum_{(j,k) \in B^i_j} (\theta'_{j,k})^2,\ 8\lambda_* L n^{-1} \right\} + 6 \sum_{(j,k) \in B^i_j} \epsilon^2_{j,k} + CLn^{-2}. \tag{36}$$

Lemma 7 *Let $T = 2^J$ and $d = \min(\alpha - \frac{1}{p}, 1)$. Set $\bar{g}_J(x) = \sum_{k=1}^T \frac{1}{\sqrt{T}} G\left(f\left(k/n\right)\right) \phi_{J,k}(x)$. Then for some constant $C > 0$*

$$\sup_{g \in F^\alpha_{p,q}(M,\epsilon)} \| \bar{g}_J - G\left(f\right) \|_2^2 \leq CT^{-2d}. \tag{37}$$

6.3 Proof of Theorem 1

Let Y and $\hat{\theta}$ be given as in (10) and (17) respectively. Then,

$$\begin{aligned} E\| \widehat{G\left(f\right)} - G\left(f\right) \|_2^2 &= \sum_k E(\hat{\tilde{\theta}}_{j_0,k} - \tilde{\theta}_{j,k})^2 + \sum_{j=j_0}^{J-1} \sum_k E(\hat{\theta}_{j,k} - \theta_{j,k})^2 + \sum_{j=J}^{\infty} \sum_k \theta^2_{j,k} \\ &\equiv S_1 + S_2 + S_3 \end{aligned} \tag{38}$$

It is easy to see that the first term S_1 and the third term S_3 are small.

$$S_1 = 2^{j_0} n^{-1} \epsilon^2 = o(n^{-2\alpha/(1+2\alpha)}) \tag{39}$$

Note that for $x \in I\!R^m$ and $0 < p_1 \leq p_2 \leq \infty$,

$$\|x\|_{p_2} \leq \|x\|_{p_1} \leq m^{\frac{1}{p_1} - \frac{1}{p_2}} \|x\|_{p_2} \tag{40}$$

Since $f \in B^\alpha_{p,q}(M)$, so $2^{js}(\sum_{k=1}^{2^j} |\theta_{jk}|^p)^{1/p} \leq M$. Now (40) yields that

$$S_3 = \sum_{j=J}^{\infty} \sum_k \theta^2_{j,k} \leq C 2^{-2J(\alpha \wedge (\alpha + \frac{1}{2} - \frac{1}{p}))}. \tag{41}$$

Proposition 2, Lemma 7 and Equation (32) yield that

$$\begin{aligned} S_2 &\leq 2\sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{j,k}-\theta'_{j,k})^2 + 2\sum_{j=j_0}^{J-1}\sum_k(\theta'_{j,k}-\theta_{j,k})^2 \\ &\leq \sum_{j=j_0}^{J-1}\sum_{i=1}^{2^j/L}\min\left\{8\sum_{(j,k)\in B_j^i}\theta_{j,k}^2,\ 8\lambda_* Ln^{-1}\right\} + 6\sum_{j=j_0}^{J-1}\sum_k\epsilon_{j,k}^2 + Cn^{-1} + 10\sum_{j=j_0}^{J-1}\sum_k(\theta'_{j,k}-\theta_{j,k})^2 \\ &\leq \sum_{j=j_0}^{J-1}\sum_{i=1}^{2^j/L}\min\left\{8\sum_{(j,k)\in B_j^i}\theta_{j,k}^2,\ 8\lambda_* Ln^{-1}\right\} + Cm^{-4} + Cn^{-1} + CT^{-2d} \end{aligned} \tag{42}$$

we now divide into two cases. First consider the case $p \geq 2$. Let $J_1 = [\frac{1}{1+2\alpha}\log_2 n]$. So, $2^{J_1} \approx n^{1/(1+2\alpha)}$. Then (42) and (40) yield

$$S_2 \leq 8\lambda_* \sum_{j=j_0}^{J_1-1}\sum_{i=1}^{2^j/L} Ln^{-1} + 8\sum_{j=J_1}^{J-1}\sum_k\theta_{j,k}^2 + Cn^{-1} + CT^{-2d} \leq Cn^{-2\alpha/(1+2\alpha)} \tag{43}$$

By combining (43) with (39) and (41), we have $E\|\hat{\theta}-\theta\|_2^2 \leq Cn^{-2\alpha/(1+2\alpha)}$, for $p \geq 2$.

Now let us consider the case $p < 2$. First we state the following lemma without proof.

Lemma 8 *Let $0 < p < 1$ and $S = \{x \in \mathbb{R}^k : \sum_{i=1}^k x_i^p \leq B,\ x_i \geq 0,\ i = 1, \cdots, k\}$. Then $\sup_{x\in S}\sum_{i=1}^k (x_i \wedge A) \leq B \cdot A^{1-p}$ for all $A > 0$.*

Let J_2 be an integer satisfying $2^{J_2} \asymp n^{1/(1+2\alpha)}(\log n)^{(2-p)/p(1+2\alpha)}$. Note that

$$\sum_{i=1}^{2^j/L}\left(\sum_{(j,k)\in B_j^i}\theta_{j,k}^2\right)^{\frac{p}{2}} \leq \sum_{k=1}^{2^j}(\theta_{j,k}^2)^{\frac{p}{2}} \leq M2^{-jsp}.$$

It then follows from Lemma 8 that

$$\sum_{j=J_2}^{J-1}\sum_{i=1}^{2^j/L}\min\left\{8\sum_{(j,k)\in B_j^i}\theta_{j,k}^2,\ 8\lambda_* Ln^{-1}\right\} \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2-p}{p(1+2\alpha)}}. \tag{44}$$

On the other hand,

$$\sum_{j=j_0}^{J_2-1}\sum_{i=1}^{2^j/L}\min\left\{8\sum_{(j,k)\in B_j^i}\theta_{j,k}^2,\ 8\lambda_* Ln^{-1}\right\} \leq \sum_{j=j_0}^{J_2-1}\sum_b 8\lambda_* Ln^{-1} \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2-p}{p(1+2\alpha)}}. \tag{45}$$

Putting (39), (41), (44) and (45) together yields $E\|\hat{\theta}-\theta\|_2^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2-p}{p(1+2\alpha)}}$. ∎

Remark 3 *To make the other terms negligible (or at least not dominant) for all α, we need to have $m^{-4} = O(n^{-\frac{2\alpha}{1+2\alpha}})$ and $T^{-2((\alpha-\frac{1}{p})\wedge 1)} = O(n^{-\frac{2\alpha}{1+2\alpha}})$. This condition puts constraints on both m and α (and p). We choose $m = n^{\frac{1}{4}}$ and so $T = n^{\frac{3}{4}}$. Then we need $\frac{3}{2}(\alpha-\frac{1}{p}) > \frac{2\alpha}{1+2\alpha}$ or equivalently $\frac{2\alpha^2-\alpha/3}{1+2\alpha} > \frac{1}{p}$. This last condition is purely due to approximation error over Besov spaces. The other condition, $m \geq n^{\frac{1}{4}}$, is needed for bounding the stochastic error.*

Asymptotic optimality under L_2 Loss

Write

$$\begin{aligned} E\|\widehat{f} - f\|_2^2 &= E\|G^{-1}[\widehat{G(f)}] - G^{-1}[G(f)]\|_2^2 = E\|(G^{-1})'(g)\,[\widehat{G(f)} - G(f)]\|_2^2 \\ &= E\int V\left(G^{-1}(g)\right)[\widehat{G(f)} - G(f)]^2 dt \end{aligned}$$

where g is function in between $\widehat{G(f)}$ and $G(f)$.

It then suffices to show that there exists a constant C such that

$$\sup_{f\in F^{\alpha}_{p,q}(M,\epsilon)} P\left\{\left\|V\left(G^{-1}(g)\right)\right\|_{\infty} > C\right\} \leq C_l n^{-l}$$

Note that G^{-1} is an increasing and nonnegative function, and V is a quadratic variance function (see equation 1). It is then enough to show

$$\sup_{f\in F^{\alpha}_{p,q}(M,\epsilon)} P\left\{\left\|\widehat{G(f)}\right\|_{\infty} > C\right\} \leq C_l n^{-l}$$

for any $l > 1$.

Recall that we can write the discrete wavelet transform of the binned data as

$$u_{j,k} = \theta'_{j,k} + \epsilon_{j,k} + \frac{1}{2\sqrt{n}} z_{j,k} + \xi_{j,k}$$

where θ'_{jk} are the discrete wavelet transform of $(\frac{G(f(i/T))}{\sqrt{T}})$ which are approximately equal to the true wavelet coefficients θ_{jk} of $G(f)$. Note that $\left|\theta'_{jk} - \theta_{jk}\right| = O\left(2^{-j(d+1/2)}\right)$, for $d = \min(\alpha - 1/p, 1)$. Note also that a Besov Ball $B^{\alpha}_{p,q}(M)$ can be embedded in $B^{d}_{\infty,\infty}(M_1)$ for some $M_1 > 0$. (See, e.g., Meyer (1992)). From the equation above, we have

$$\sum_{k=1}^{2^{j_0}} \widetilde{\theta}_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \theta'_{j,k}\psi_{j,k}(t) \in B^{d}_{\infty,\infty}(M_2)$$

for some $M_2 > 0$. Applying the Block thresholding approach, we have

$$\begin{aligned} \hat{\theta}_{jk} &= (1 - \frac{\lambda L\sigma^2}{S^2_{(j,i)}})_+\theta'_{j,k} + (1 - \frac{\lambda L\sigma^2}{S^2_{(j,i)}})_+\epsilon_{j,k} + (1 - \frac{\lambda L\sigma^2}{S^2_{(j,i)}})_+\left(\frac{1}{2\sqrt{n}} z_{j,k} + \xi_{j,k}\right) \\ &= \hat{\theta}_{1,jk} + \hat{\theta}_{2,jk} + \hat{\theta}_{3,jk} \text{ , for } (j,k) \in B^i_j,\ j_0 \leq j < J. \end{aligned}$$

Note that $\left|\hat{\theta}_{1,jk}\right| \leq \left|\theta'_{j,k}\right|$ and so $\widehat{g}_1 = \sum_{k=1}^{2^{j_0}} \widetilde{\theta}'_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \hat{\theta}_{1,j,k}\psi_{j,k} \in B^d_{\infty,\infty}(M_2)$. This implies $\widehat{g}_1$ is uniformly bounded. Note that $T^{\frac{1}{2}}\left(\sum_{j,k}\left(\epsilon^2_{j,k}\right)\right)^{1/2} = T^{\frac{1}{2}} \cdot O\left(m^{-2}\right) = o(1)$, so $W^{-1} \cdot T^{\frac{1}{2}}\left(\hat{\theta}_{2,jk}\right)$ is a uniformly bounded vector. For $0 < \beta < 1/6$ and a constant $a > 0$ we have

$$\begin{aligned} P\left(\left|\hat{\theta}_{3,jk}\right| > a2^{-j(\beta+1/2)}\right) &\leq P\left(\left|\hat{\theta}_{3,jk}\right| > aT^{-(\beta+1/2)}\right) \\ &\leq P\left(\left|\frac{1}{2\sqrt{n}}z_{j,k}\right| > \frac{1}{2}aT^{-(\beta+1/2)}\right) + P\left(|\xi_{j,k}| > \frac{1}{2}aT^{-(\beta+1/2)}\right) \\ &\leq A_l n^{-l} \end{aligned}$$

for any $l > 1$ by Mill's ratio inequality and equation (30). Let $A = \underset{j,k}{\cup}\left\{\left|\hat{\theta}_{3,jk}\right| > a2^{-j(\beta+1/2)}\right\}$. Then $P(A) = C_l n^{-l}$. On the event A^c we have

$$\widehat{g}_3(t) = \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \hat{\theta}_{3,jk}\psi_{j,k}(t) \in B^{\beta}_{\infty,\infty}(M_3), \text{ for some } M_3 > 0$$

which is uniformly bounded. Combining these results we know that for C sufficiently large,

$$\sup_{f \in F^{\alpha}_{p,q}(M,\epsilon)} P\left\{\left\|\widehat{G(f)}\right\|_{\infty} > C\right\} \leq \sup_{f \in F^{\alpha}_{p,q}(M,\epsilon)} P(A) = C_l n^{-l}. \quad \blacksquare \tag{46}$$

$\blacksquare$

Acknowledgments: The authors would like to thank Eric Kolaczyk for providing the BATSE data and Packet loss data.

References

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

[2] Antonidis, A and Leblanc, F. (2000). Nonparametric wavelet regression for binary response. *Statistics* **34**, 183-213.

[3] Antoniadis, A. & Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika* **88**, 805-820.

[4] Bar-Lev, S. K. and Enis, P. (1990). On the construction of classes of variance stabilizing transformations. *Statist. Probab. Lett.* **10**, 95-100.

[5] Bartlett, M. S. (1936). The square root transformation in analysis of variance. *J. Roy. Statist. Soc. Suppl.* **3**, 68-78.

[6] Besbeas, P., De Feis, I. and Sapatinas, T. (2004). A Comparative Simulation Study of Wavelet Shrinkage Estimators for Poisson Counts. *Internat. Statist. Rev.* **72**, 209-237.

[7] Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, Inst. of Math. Statist., Hayward, California.

[8] Brown, L. D., Cai, T. T., Zhang, R., Zhao, L. H. and Zhou, H. H. (2006). The Root-unroot algorithm for density estimation as implemented via wavelet block thresholding. Manuscript.

[9] Brown L. D. and Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24**, 2524-2535.

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

[11] Cai, T. (2002). On block thresholding in wavelet regression: Adaptivity, block Size, and threshold level. *Statistica Sinica* **12**, 1241-1273.

[12] Cai, T. and Silverman, B.W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya Ser. B* **63**, 127-148.

[13] Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

[14] DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Amer. Math. Soc.* **305**, 397-414.

[15] Donoho, D.L. (1993). Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In *Different perspectives on Wavelets* (I. Daubechies Ed.), *Proc. Symp. Appl. Math.* **47**, 173-205.

[16] Donoho, D.L. & Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

[17] Efron, B. (1982). Transformation theory: How normal is a family of a distributions? *Ann. Statist.* **10**, 323-339.

[18] Fryźlewicz, P. and Nason, G.P. (2001). Poisson intensity estimation using wavelets and the Fisz transformation. Technical Report 01/10, Department of Mathematics, University of Bristol, United Kingdom.

[19] Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26**, 922-942.

[20] Hoyle, M. H. (1973). Transformations - an introduction and bibliography. *International Statistical Review* **41**, 203-223.

[21] Kaneko, Y. (2004). Spectral studies of Gamma-Ray burst prompt emission. Ph. D. Thesis. University of Alabama in Huntsville.

[22] Kolaczyk, E. D. (1999). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statist. Sinica* **9**, 119-135.

[23] Kolaczyk, E.D. (1999). Bayesian multiscale models for Poisson processes. *J. Amer. Statist. Assoc.* **94**, 920-933.

[24] Kolaczyk, E. D. and Nowak, R.D. (2005). Multiscale generalized linear models for nonparametric function estimation. *Biometrika* **92**, 119-133.

[25] 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.

[26] Lepski, O. V. (1990). On a problem of adaptive estimation in white gaussian noise. *Theor. Probab. Appl.* **35**, 454-466.

[27] Meyer, Y. (1992). *Wavelets and Operators*. Cambridge University Press, Cambridge.

[28] Morris, C. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10**, 65–80.

[29] Quilligan, F., McBreen, B., Hanlon, L. ,McBreen, S. , Hurley, K. J. and Watson, D. (2001). Temporal properties of gamma-ray bursts as signatures of jets from the central engine. *Astronomy & Astrophysics* **385**, 377.

[30] Runst, T. (1986). Mapping properties of non-linear operators in spaces of Triebel-Lizorkin and Besov type, *Anal. Math.* **12**, 313-346.

[31] Strang, G. (1992). Wavelet and dilation equations: a brief introduction. *SIAM Review* **31**, 614-627.

[32] Triebel, H. (1992). *Theory of Function Spaces II*. Birkhäuser Verlag, Basel.

[33] Yajnik, M., Moon, S. Kurose, J. and Towsley D. (1999). Measurement and Modelling of the Temporal Dependence in Packet Loss. In *Proc. 18^{th} Annual Conference IEEE Computer and Communications Societies* (INFOCOM), New York, NY, 345-353.

[34] Zhou, H. H. (2006). A note on quantile coupling inequalities and their applications. Technical report. Available from www.stat.yale.edu/˜hz68 .