8-2015

Nonparametric Regression In Natural Exponential Families: A Simulation Study

YIJUN CHEN
Clemson University, yijunc@g.clemson.edu

NONPARAMETRIC REGRESSION IN NATURAL EXPONENTIAL FAMILIES:A SIMULATION STUDY

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Mathematical Sciences

by
Yijun Chen
August 2015

Accepted by:
Dr. Jun Luo, Committee Chair
Dr. Patrick Gerard
Dr. William Bridges

Abstract

Nonparametric regression has been particularly well developed. Base on the asymptotic equivalence theory, there are some procedures that can turn more complicated nonparametric estimation problems into a standard nonparametric regression, especially in natural exponential families. This procedure is described in detail with a wavelet thresholding estimator for Gaussian nonparametric regression and simulation study shed light on the behavior of this method under different sample sizes and parameterizations of exponential distribution. The resulting estimators have a high degree of adaptivity in [6].

Table of Contents

List of Tables

Chapter 1

Introduction

The idea we use for the procedure of nonparametric regression is asymptotic equivalence theory, by which we can turn some more complicated nonparametric estimation problems into a standard Gaussian nonparametric regression problem. Meanwhile, this method is appealing because it can also eliminate the transformation bias.

In statistical decision theory, the asymptotic equivalence theory has an important position, it describes limiting behavior and the functions eventually becoming essentially equal.

In recent years, there are a lot of efforts to construct practical procedures to turn complicated nonparametric estimation problems into a standard regression base on the asymptotic equivalence theory. Which turn the research object to a relatively simple and already well studied model. For example, in [1] and [5],the methods base on binning and taking the median have been well developed for nonparametric regression with general additive noise. Especially in [2], it introduced the nonparametric regression in exponential families with a quadratic variance function, such as binomial regression, gamma regression and Poisson regression. Also in [4], the procedure is base on a root-unroot transformation for density estimation.

The bias reduction is a crucial problem. Some methods have been proposed and well studied , such as the mean matching variance stabilizing transformation [2] reduces the bias from the transformation while still stabling the variance.

The procedure we are going to talk begins by [2] grouping the data into bins with size of order $(log n)^{1+v}$ for some $v > 0$, where n is the sample size, and we test v to choose some appropriate value and explore the influence of sample size and the value of v to the regression problem. After that, the Gaussian regression procedure can applied to the new group of data from transformation. We construct the inverse VST of the estimator obtained in the Gaussian regression to get the final estimator of the original regression function.

In this report, we applied the new procedure for simulation in exponential distribution and Poisson distribution, use the *log* transformation as the variance stabilized transformation to get the new group of data of exponential distribution, and square root function as VST for Poisson distribution. We get the performance of regression base on the different sample sizes and the different binning sizes. A much smaller sample size and a larger v value will have less mean square error for the predicted value and the true value.

The report is organized as follows. Chapter 2 and Chapter 3 introduced of using the VST to convert nonparametric regression in exponential families into a Gaussian nonparametric regression problem. Chapter 4 introduced the wavelet procedure. Simulation study design and result are shown in Chapter 5 and Chapter 6. Theoretical proof for the related properties are presented in Appendix A.

There are still some more future work for this report. First, the VST we used in this report is the *log* transformation for exponential distribution random error data and square root function for Poisson distribution random error data. We tested for different group pf sample sizes and v value, we may have some technique to find the best combination which could have the least mean square error. Second, the data set we used in this report is randomly generated , the future work may apply this new procedure to the large real data set.

Chapter 2

Related Work

The noise is not additive and non-Gaussian for nonparametric regression problem in natural exponential families. Thus we can not get desirable results by applying standard nonparametric regression methods to the data directly. In order to solve this problem, we need to use a variance stabilizing transformation(VST) to turn this problem to a standard Gaussian regression problem.

2.1 Binning

In [6] Divided $\{Y_i\}$ into T equal length intervals between 0 and 1. Let $Q_1, Q_2, \cdots, Q_T$ be the sum of the observations in each of the intervals. The choice of T satisfying $T = n/log^{1+v}n$. We begin by dividing the interval into T qui-length subintervals with $m = n/T$ observations in each subintervals. Let Q_j be the sum of observations on the j^{th} subinterval $I_j = (\frac{j-1}{T}, \frac{j}{T}]$, $j = 1, 2, \cdots, T$

$$Q_j = \sum_{i=(j-1)m+1}^{jm} Y_i \tag{2.1}$$

The sum $\{Q_j\}$can be treated as a group of new observations for a Gaussian regression directly, but this will lead to a heteroscedastic problem. Instead, we apply the VST and then treat $G(\frac{Q_j}{m})$ as new observations in a homoscedastic Gaussian regression problem. Let

$$Y_j^* = G(\frac{Q_j}{m}), j = 1, \cdots, T. \tag{2.2}$$

The transformed data $Y^* = (Y_1^*, \cdots, Y_T^*)$ is then treated as the new qui-spaced sample for a Gaussian nonparametric regression problem.

2.2 VST

The VST for natural exponential families has been widely used . For example, in [4] for an extensive review. Here we focused on exponential distribution and the VST for exponential distribution is *log* function, and the situations likely to occur with certain kinds of reaction times, waiting times, and financial data.

In [6] the VST for natural exponential families has beed widely used in many contexts. Note that the probability density/mass function of a distribution in a natural one-parameter exponential families can be written as

$$q(x|\eta) = e^{\eta x - \psi(\eta) h(x)}, \tag{2.3}$$

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

$$\mu(\eta) = \psi^{'}(\eta), and \sigma^2(\eta) = \psi^{''}(\eta). \tag{2.4}$$

We shall denote the distribution by $NEF(\mu)$. Let $X_1, \cdots, X_m \sim NEF(\mu)$ be a random sample and set $X = \sum_{i=1}^m X_i$. The Central Limit Theorem yields that

$$\sqrt{\{}(X/m) - (\mu(\eta))\} \longrightarrow N(0, V(\mu(\eta))), as m \to \infty \tag{2.5}$$

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

$$G'(\mu) = V^{-\frac{1}{2}}(\mu). \tag{2.6}$$

The standard delta method then yields

$$\sqrt(m)\{G(X/m) - G(\mu(\eta))\} \rightarrow N(0, 1) \tag{2.7}$$

Since the natural exponential can be mean parameterized, we define

$$H_m(\mu) = \mathbb{E}G(X/m),$$

where H_m depends on m. For notational simplicity, we shall drop the subscript m hereafter.

2.3 Nonparametric Regression

The general nonparametric regression model is written as

$$y_i = f(x_i') + \epsilon_i \tag{2.8}$$

with the function f left unspecified. Where $x_i^{'}$ is a vector of predictors for the i^{th} of n observations; the errors ϵ_i are assumed to be normally and independently distributed with mean 0 and variance σ^2.

In [6] , now consider nonparametric regression in natural exponential families. Suppose we observe

$$Y_i \sim NEF(f(t_i)), i = 1, \cdots, n, t_i = \frac{i}{n} \tag{2.9}$$

and wish to estimate the mean function $f(t)$. As mentioned earlier, applying standard nonparametric regression methods directly to the data $\{Y_i\}$ in general do not yield desirable results. We shall turn this problem to a standard Gaussian regression problem based on a sample $\{\widetilde{Y_j} : j = 1, \cdots, T\}$

where

$$\widetilde{Y_j} \sim N(H(f(t_j)), m^{-1}), t_j = j/T, j = 1, 2, \cdots, T. \tag{2.10}$$

Here T is defined as before, T is the number of bins, and m is the number of observations in each bin.

Chapter 3

Effects of binning and VST

The transformed data $\{Y_j^*\}$ can be treated as if they were data from a homoscedastic Gaussian nonparametric regression problem after binning and VST. We need to introduced some background to explain why binning and VST can work here, quantile coupling provides an important technical tool to shed insights on the procedure in [6].

The following result, which is a direct consequence of the quantile coupling inequality developed in [9], shows that the binned and transformed data can be well approximated by independent normal variables. See also in [14].

Lemma 1. In [6] let $X_i \sim NEF(\mu)$ with variance V for $i = 1, \cdots, m$ and let $X = \sum_{i=1}^m X_i$. There exists a standard normal random variable $Z \sim N(0,1)$ and positive constants $c_i, i = 1, 2, 3, 4, 5$, not depending on m such that whenever the event $A = \{|X - m\mu| \leq c_1 m\}$ occurs,

$$|X - m\mu - \sqrt{mV}Z| > c_2 Z^2 + c_3 \tag{3.1}$$

$$\mathbb{P}(|X - m\mu - \sqrt{mV}Z| > a) \leq c_4 exp(-c_5 a), \tag{3.2}$$

uniformly over μ in a compact set in the interior of the natural parameter space.

Hence, for large m, X can be treated as a normal random variable with mean $m\mu$ and variance mV. Let $Y = G(X/m)$, and Z be a standard normal variable satisfying (2). Then Y can be written as

$$Y = H(\mu) + m^{-\frac{1}{2}} Z + \xi, \tag{3.3}$$

where

$$\xi = G(\frac{X}{m}) - H(\mu) - m^{-\frac{1}{2}} Z, \tag{3.4}$$

is a zero mean and "stochastically small" random variable. The following result is proved in section 6.1.

Lemma 2. [6] Let $X_i \sim NEF(\mu)$ with variance V for $i = 1, \cdots, m$ and let $X = \sum_{i=1}^{m} X_i$. Let Z be the standard normal variable given as in Lemma 1 and let ξ be given in (8). Then for any integer $k \geq 1$ there exist positive constants $c_k > 0$ such that for all $a > 0$,

$$P(m|\xi| > a) \leq c_1 exp(-c_2 a) + c_3 exp(-c_4 m), \tag{3.5}$$

So far the discussion has focused on the effects of the VST for i.i.d. observations. However, observations in each bin are independent not identically distributed since the mean function f is smooth. Let $X_i \sim NEF(\mu_i), i = 1. \cdots, m$, be independent. Here the means f are "close" but they are not equal. Let μ be a value close to the $\mu_i' s$. The analysis given in [6] shows that X_i in each bin can in fact be coupled with i.i.d random variables $X_{i,c}$ with $X_{i,c} \sim^{iid} NEF(\mu_c^*)$, for some $\mu_c^* > 0$. See Lemma 4 in proof section for a precise statement. Lemma 1,2 and 4 together yield the following result which shows how far away are the transformed data $\{Y_j^*\}$from the ideal Gaussian model.

Theorem 1. [6] Let $\{Y_j^*\} = G(\frac{Q_j}{m})$ be given as in (2.2). Assume that f is continuous, and for all $x \in [0, 1]$, $f(x) \in [\varepsilon, v]$, a compact set in the interior of the mean parameter space of the natural exponential family. Then $\{Y_j^*\}$ can be written as

$$Y_j^* = H(f(\frac{j_*}{T})) + m^{-\frac{1}{2}} Z_j + \xi_j, \qquad j = 1, 2, \cdots, T, \tag{3.6}$$

where $jm + 1 \leq j_*(j + 1)m$, $Z_j \sim^{i.i.d.} N(0, 1)$, and ξ_j are independent and "stochastically small" random variables satisfying that for any integer $k > 0$ and any constant $a > 0$

$$P(m|\xi_j| > a) \leq c_1 exp(-c_2 a) + c_3 exp(-c_4 m), \tag{3.7}$$

where $c_k > 0$.

Theorem 1 provides explicit bounds for both the deterministic and stochastic errors.

Chapter 4

A Wavelet procedure for generalized regression

[13]Wavelets are special basis functions with two appealing features. First, it can be computed quickly. Second, the resulting estimators are spatially adaptive, this means we can accommodate local features in the data.

We start with the simplest wavelet, the Haar wavelet. The Haar father wavelet or Harr calling function is defined by

$$\phi(x) = \begin{cases} 1 \text{ if } 0 \leq x < 1 \\ 0 \text{ otherwise .} \end{cases} \tag{4.1}$$

The mother Haar wavelet is defined by

$$\psi(x) = \begin{cases} -1 \text{ if } 0 \leq x < \frac{1}{2} \\ 1 \text{ if } \frac{1}{2} \leq x < 1 \end{cases} \tag{4.2}$$

Let $\phi_{jk}(x)$=$2^{\frac{j}{2}}\psi(2^j x - k)$.

Father ϕ

Level 1 ψ

Level 2 ψ_{10} ψ_{11}

Level 3 ψ_{20} ψ_{21} ψ_{22} ψ_{23}

Level 4 ψ_{30} ψ_{31} ψ_{32} ψ_{33} ψ_{34} ψ_{35} ψ_{36} ψ_{37}.

The set of functions $\{\{\phi\}, \{\psi_{00}\}, \{\psi_{10}, \psi_{11}\}, \{\psi_{20}, \psi_{21}, \psi_{22}, \psi_{23}\}, \cdots\}$ is an orthonormal basis for $L_2(0,1)$. So, as $J \to \infty$,

$$f(x) \approx \alpha\phi(x) + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x) \tag{4.3}$$

where $\alpha = \int_0^1 f(x)\phi(x)dx$, $\beta_{jk} = \int_0^1 f(x)\psi_{jk}(x)dx$.

Many functions f hace sparse expansions in a wavelet basis. Smooth functions are sparse. (Smooth + jumps) is also sparse. So, we expect that for many functions we can write $f(x) = \alpha\phi(x) + \sum_j \sum_j \beta_{jk}\psi_{jk}(x)$, where most $\beta_{jk} \approx 0$. The idea is to estimate the $\beta_{jk}^{'}s$ then set all $\hat{\beta}_{jk} = 0$ except or a few large coefficients.Father ϕ and mother ψ. Smooth wavelets cannot be written in closed form but they can be computed quickly. Still have:

$$f(x) = \sum_k \alpha_{0k}\phi_{0k}(x) + \sum_{j=0}^{\infty} \sum_k \beta_{jk}\psi_{jk}(x), \tag{4.4}$$

where $\alpha_{0k} = \int f(x)\phi_{0k}(x)dx$ and $\beta_{jk} = \int f(x)\psi_{jk}(x)dx$.

Let $Y_i = f(x_i) + \sigma\epsilon_i$, where $x_i = i/n$.(Adjustments are needed for non-equally spaced data), the procedures are:

- Form the preliminary estimate:

$$\tilde{\beta}_{jk} = \frac{1}{n}\sum_i Y_i\psi_{jk}(x_i) \approx \int f(x)\psi_{jk}(x)dx = \beta_{jk}, \tag{4.5}$$

- Shrink: $\hat{\beta}_{jk} \longleftarrow \text{shrink}(\tilde{\beta}_{jk})$

- Reconstruct function:

$$\hat{f}(x) = \sum_k \hat{\alpha}_{0k}\phi_{0k}(x) + \sum_{j=0}^{\infty}\sum_k \hat{\beta}_{jk}\psi_{jk}(x), \tag{4.6}$$

In practice, the preliminary estimates are computed using the discrete wavelet transform(DWT). Two types of shrinkage are used: hard thresholding and soft thresholding.

The hard threshold estimator is

$$\hat{\beta}_{jk} \begin{cases} 0 \text{ if } |\tilde{\beta}_{jk}| < \lambda \\ \tilde{\beta}_{jk} \text{ if } |\tilde{\beta}_{jk}| \geq \lambda \end{cases} \tag{4.7}$$

The soft threshold estimator is

$$\hat{\beta}_{jk} = sign(\tilde{\beta}_{jk})(|\tilde{\beta}_{jk}| - \lambda)_+, \tag{4.8}$$

We still need to choose the threshold λ. There are several methods for choosing λ. The simplest rule is the universal threshold defined by

$$\lambda = \hat{\sigma}\sqrt{\frac{2\log n}{n}}. \tag{4.9}$$

[11]The following list describes some of the different methods that can be used to choose the list. Such as exact minimax method, SURE , cross-validation, false-discovery rate, Bayesian methods, and Ogden's methods.

Chapter 5

Simulation Study Design

The above method for nonparametric regression developed by [6]. To do so, we choose the exponential distribution with parameter 1, and then tested the value v to determine the length for each bin and the number T. We also need to make sure the length is between 0 and 1.

However , there are some tips we need consider when we decide the length T. First, the we need to make sure that there are enough bins, so the length has to be close to 0 as possible. After transforming the data, a wavelet regression procedure then applied to the new group of data.

We want to explore the influence of binning and sample size, so we chose three group pf sample sizes, 50,100,500, and three group of v value, 3,3.1,3.2, to compare the mean square error.

First we generate the data base on the nonparametric model $y = f(x_i) + \epsilon$, define a simple function $f = e^x$ as the smooth mean function, use exponential distribution with parameter 1 as random error function first, then use Poisson distribution with parameter 1. Since the data are added with error, we assume the error are iid distribution. Then we need to calculate the set of value of length. After this, we start binning data, and get the summation of each variables in subintervals. Then we apply the VST to binned data, here we use log function for exponential distribution and square root function for Poisson distribution, in order to get a good performance, we use $log(s + 10)$ where s is the binned data, same as Poisson distribution. After we binned and transformed the data, we need to apply the wavelet procedure to the new group pf data. Another important thing is the wavelet regression function in r require the data length is power of 2, so we need first pads with zeros of the data in order to apply wavelet procedure.

In [10] for the wavelet procedure, we can perform the interpolation of the old data values to the new regularly spaced grid, also we can apply wd function directly to the data set. The next step is to threshold the coefficients. And after this thresholding we obtain the estimate. The estimate is not extremely bad but could no doubt be improved by a judicious choice of threshold method and parameters.The results are listed in conclusion part.

Chapter 6

Simulation Results, Conclusions and Discussion

6.1 Simulation Results

The results for three different sample sizes are listed below. The difference between v value have been shown in three sample sizes. For each combination of sample size and v value, we calculate the mean square error five times, then find the average value of them to compare and analyze the result. First we will give the result of exponential distribution random error , followed with comparison of different group. Then we give the result of Poisson distribution random error along with the comparison table.

6.2 Conclusions and future work

For the exponential distribution random error data, when the v remain the same, it is clear that larger sample sizes can increase mean square error, possible because there are more points of both the predicted and original value, so the number of term in MSE is greater than smaller sample size . Meanwhile, larger sample sizes could result larger interval length.

Sample Size	*v*	*Interval*	***MSE***	***Avg MSE***
50	3	0.2135	0.5729 0.2529 0.4448 1.5418 0.7746	0.7174
50	3.1	0.1863	0.3218 0.2107 0.4702 0.7566 0.2722	0.4063
50	3.2	0.1625	0.1423 0.1054 0.1769 0.9382 0.1377	0.3001
100	3	0.2223	1.0443 0.6990 1.9724 0.9688 0.5133	1.0396
100	3.1	0.1908	1.7805 0.6594 1.2239 0.6482 0.9767	0.9767
100	3.2	0.1638	0.5923 0.5839 0.6692 0.422 0.5573	0.5651
500	3	0.3352	1.3374 1.7787 1.8583 0.8376 3.0475	1.7719
500	3.1	0.2792	1.5275 0.5796 2.1313 0.3475 1.6001	1.2372
500	3.2	0.2326	1.2180 2.0142 0.3152 0.8519 1.6132	1.2025

Table 6.1: Exponential distribution

Sample Size	*v*	*Interval*	***Avg MSE***
50	3	0.2135	0.7174
50	3.1	0.1863	0.4063
50	3.2	0.1625	0.3001

Table 6.2: `n=50`

Sample Size	*v*	*Interval*	***Avg MSE***
100	3	0.2223	1.0396
100	3.1	0.1908	0.9767
100	3.2	0.1638	0.5651

Table 6.3: `n=100`

Sample Size	*v*	*Interval*	***Avg MSE***
500	3	0.3352	1.7719
500	3.1	0.2792	1.2372
500	3.2	0.2326	1.2025

Table 6.4: `n=500`

v	***Sample Size***	*Interval*	***Avg MSE***
3	50	0.2135	0.7174
3	100	0.2223	1.0396
3	500	0.3352	1.7719

Table 6.5: `v=3`

v	***Sample Size***	*Interval*	***Avg MSE***
3.1	50	0.1863	0.4063
3.1	100	0.1908	0.9767
3.1	500	0.2792	1.2372

Table 6.6: v=3.1

v	***Sample Size***	*Interval*	***Avg MSE***
3.2	50	0.1625	0.3001
3.2	100	0.1638	0.5651
3.2	500	0.2326	1.2025

Table 6.7: v=3.2

Sample Size	*v*	*Interval*	***MSE***	***Avg MSE***
50	3	0.2135	1.5816 1.7895 1.5211 0.9742 2.3175	1.6368
50	3.1	0.1863	0.5594 2.0433 2.2403 1.6387 1.6491	1.6262
50	3.2	0.1625	0.4153 2.0963 0.5242 0.2798 0.6360	0.7903
100	3	0.2223	1.7023 2.4350 2.4728 1.3189 0.7753	1.7609
100	3.1	0.1908	2.0818 1.0402 2.4050 0.8007 1.8478	1.6351
100	3.2	0.1638	1.0951 0.5225 1.2749 0.4384 0.9415	0.8545
500	3	0.3352	3.0987 3.8024 2.7733 3.6794 3.6620	3.4032
500	3.1	0.2792	3.8911 2.4412 1.9825 3.1100 3.5684	2.9986
500	3.2	0.2326	3.5383 2.5069 1.4557 2.2087 1.7535	2.2926

Table 6.8: Poisson distribution

Sample Size	*v*	*Interval*	***Avg MSE***
50	3	0.2135	1.6368
50	3.1	0.1863	1.6262
50	3.2	0.1625	0.7903

Table 6.9: n=50

Sample Size	*v*	*Interval*	***Avg MSE***
100	3	0.2223	1.7609
100	3.1	0.1908	1.6351
100	3.2	0.1638	0.8545

Table 6.10: n=100

Sample Size	*v*	*Interval*	***Avg MSE***
500	3	0.3352	3.4032
500	3.1	0.2792	2.9986
500	3.2	0.2326	2.2926

Table 6.11: `n=500`

v	***Sample Size***	*Interval*	***Avg MSE***
3	50	0.2135	1.6368
3	100	0.2223	1.7609
3	500	0.3352	3.4032

Table 6.12: `v=3`

v	***Sample Size***	*Interval*	***Avg MSE***
3.1	50	0.1863	1.6262
3.1	100	0.1908	1.6351
3.1	500	0.2792	2.9986

Table 6.13: `v=3.1`

v	***Sample Size***	*Interval*	***Avg MSE***
3.2	50	0.1625	0.7903
3.2	100	0.1638	0.8545
3.2	500	0.2326	2.2926

Table 6.14: `v=3.2`

When the sample size remains the same, it is clear that larger v can reduce the mean square error. Combined with the above, the larger v value and smaller sample size always along with a smaller mean square error. So the smallest mean square error in all the results corresponding to the largest v value , and the smallest sample size.

For the Poisson distribution random error data, the conclusions are the same.

The tables will give a straightforward comparison. It is clear that the subtable of $v = 3.2$ and the smallest sample size 50 perform well. There are still some more future work for this report. First, the technique of how to choose the value of v and the appropriate sample size. The value used in this report was tested by hand in order to make sure the length of the subinterval lies between 0 and 1. However, may be in the future, we could set the ideal range of mean square error that we what, then return back to get the satisfied v value which will make the binning and transformation procedure work.

Moreover, the data set we used in this report is generated by r, maybe in the future when we explore enough technique about how to choose v , control and sample size and control the ideal mean square error, we could apply this new procedure to the large real data set.

Appendices

Appendix A Proofs

Lemma 1, Lemma 2, and Lemma 3 are used to prove Theorem 3.1 . Thus they are stated first and proven first. They are followed by eh proof of Theorem 3.1. All results proven by [6].

Proof of Lemma 2. Write

$$G(\frac{X}{m}) - G_m(\mu) - \frac{1}{\sqrt{m}}Z \tag{1}$$

$$[G(\frac{X}{m}) - G(\mu + \sqrt{\frac{V}{m}})Z] + [G(\frac{X}{m}) - G(\mu) - \frac{1}{\sqrt{m}}Z] + [G(\mu) - G_m(\mu)] \tag{2}$$

Taylor expansion yields

$$G(\frac{X}{m}) - G_m(\mu) - \frac{1}{\sqrt{m}}Z = G'(\mu_1^*)(\frac{X}{m} - \mu - \sqrt{\frac{V}{m}}Z) + G''(\mu_2^*)\frac{V}{m}Z^2 + G(\mu) - G_m(\mu) \tag{3}$$

From Lemma 1 we have

$$P(m|\frac{X}{m} - \mu - \sqrt{\frac{V}{m}}Z| > a) \leq c_1 \exp(-c_2 a), \tag{4}$$

Since Z is standard normal, Z^2 has an exponential tail

$$P(m|\frac{V}{m}Z^2| > a) \leq c_3 \exp(-c_4 a). \tag{5}$$

It is easy to see that

$$|G(\mu) - G_m(\mu)| = O(\frac{1}{m}). \tag{6}$$

Lemma 3 *Let*$X_i \sim NEF(\mu_i), i = 1, \cdots, m,$ *be independent with* $\mu_i \in [\epsilon, v]$ *a compact subset in the interior of the mean parameter space of the natural exponential family. Assume that* $|min_i \mu_i - max_i \mu_i| \leq C\delta$. *Then there are i.i.d random variables* $X_{i,c}$ *where* $X_{i,c} \sim NEF(\mu_c^*)$

with $\mu_c^* \in [min_i \mu_i, max_i \mu_i]$ *such that* $ED = 0$ *and*

$$P(\{X_i \neq X_{i,c}\}) \leq C\delta, \tag{7}$$

and for *any* $fixed$ *integer* $k \geq 1$ *there exists a constant* $C_k > 0$ *such that* for *all* $a > 0$,

$$P(|D| > a) \leq c_1 exp(-c_2 a^2 m) + c_3 exp(-c_4 m). \tag{8}$$

Proof of Lemma 3 There is a classical coupling identity for the Total variation distance. Let P and Q be distribution of two random variables X and Y on the same sample space respectively, then there is a random variable Y_c with distribution Q such that $P(X \neq Y_c) = |P - Q|_{TV}$. See, for example, page 256 in [**?**, 10] The proof for the first identity follows from that identity and the inequalty that $|NEF(\mu_i - NEF(\mu_c^*))|_{TV} \leq C|mnu_i - \mu_c^*|$ for some $C > 0$ which only depends on the family of the distribution of X_i and $[\epsilon, v]$.

Using Taylor expansion we can rewrite D as $D = G'(\zeta)\frac{\sum_{i=1}^m X_i - X_{i,c}}{m}$ for some ζ between $\frac{\sum_{i=1}^m X_i}{m}$ and $\frac{\sum_{i=1}^m X_{i,c}}{m}$. Write

$$\frac{\sum_{i=1}^m X_i - X_{i,c}}{m} = \frac{\sum_{i=1}^m X_i - EX_i}{m} - \frac{\sum_{i=1}^m X_{i,c} - EX_{i,c}}{m} + \frac{\sum_{i=1}^m EX_i - EX_{i,c}}{m}. \tag{9}$$

Since the distributions X_i and $X_{i,c}$ are in exponential family, we have

$$P(|\frac{\sum_{i=1}^m X_i - EX_i}{m}| \geq a) \leq c_1 exp(-c_2 a^2 m), \tag{10}$$

$$P(|\frac{\sum_{i=1}^m X_{i,c} - EX_{i,c}}{m}| \geq a) \leq c_3 exp(-c_4 a^2 m), \tag{11}$$

Note that $|\sum_{i=1}^m EX_{i,c} - EX_{i,c}m| \leq c_5\delta$. Thus we have

$$P(|\sum_{i=1}^m X_i - X_{i,c}m| > a) \leq c_6 exp(-c_7 a^2 m) \tag{12}$$

where $a > 2c_5\delta$. The equation is apparently true when $a \leq 2C\delta$. Since $X_i - X_{i,c}$ are independent, it can be shown that the second inequality in this Lemma follows immediately by observing that $G'(\zeta)$ is bounded with a probability approaching to 1 exponentially fast, since for any $\epsilon > 0$,

$$P(|\zeta| > \mu + \epsilon) \leq c_8 exp(-c_9 \varepsilon^2 m). \tag{13}$$

Proof of Theorem 1 From Lemma 3, there exist $Y^*_{j,c}$ where $X_{i,c} \sim NEF(f^*_j)$ with

$$f^*_{j,c} \in [\min_{jm+1 \leq i \leq (j+1)m} f(\frac{i}{n}), \max_{jm+1 \leq i \leq (j+1)m} f(\frac{i}{n})] \tag{14}$$

such that

$$E[Y^*_j - Y^*_{j,c}] = 0 \tag{15}$$

$$P(|Y^*_j - Y^*_{j,c}| > a) \leq c_1 exp(-c_2 a^2 m) + c_3 exp(-c_4 m) \tag{16}$$

Let $f^*_{j,c} = f(j_*/T)$, where $jm + 1 \leq j_* \leq (j+1)m$, by the intermediate value theorem. Lemmas 1 and 2 together yield

$$Y^*_j = H(f(\frac{j_*}{T})) + m^{-\frac{1}{2}} Z_j + \xi_j, j = 1, 2, \cdots, T. \tag{17}$$

$$P(|\xi_j| > a) \leq c_1 exp(-c_2 a^2 m) + c_3 exp(-c_4 m). \tag{18}$$

Theorem 1 then follows immediately by combining equations (15)-(18).

Bibliography

[1] BROWN, L.D.,CAI, T.T. and ZHOU, H.H.(2008).Robust nonparametric estimation via wavelet median regression. *Ann.Statist.* 36:2055-2084 .

[2] BROWN, L.D.,CAI, T.T. and ZHOU, H.H.(2010). Nonparametric regression in exponential families. *Ann.Statist.* To appear.

[3] BROWN, L.D.,CAI, T.T. and ZHOU, H.H.(2010). Asymptotic equivalence of nonparametric regression and white noise. *Ann.Statist.* 24:2385-2398.

[4] BROWN, L.D.,CAI, 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.

[5] CAI, T.T. and ZHOU, H.H.(2009). Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann.Statist.* 37:3024-3235.

[6] CAI, T.T. and ZHOU, H.H.(2010). Nonparametric regression in natural exponential families. *IMS Collections.* 6:199-215

[7] FOX,J.(2002). A note on nonparametric regression. Technical report

[8] HOYLE, M.H.(1973). Transformations-an introduction and a bibliography. *Internat. Statist. Rev.* 41:203-223

[9] KOMLOS,J., MAJOR,P. and TUSNADY,G. (1975). An approximation of partial sums of independent RV'-s, and the sample DF.I. *Z.Washrsch. verw. Gebiete* 111-113

[10] NASON,G.P.(2011). *Wavelet methods in statistics with R.* Springer, New York.

[11] NASON,G.P. Choice of the threshold parameter in wavelet function estimation. A Technical report

[12] POLLARD,D.(2002) *A user's guide to measure theoretic probability. Cambridge Series in Statistical and Probabilistic Mathematics* 8: Cambridge Univ. Press, Cambridge.

[13] SCHAFER,C., WASSERMAN,L. A tutorial on nonparametric inference with R

[14] ZHOU,H.(2006). A note on quantile coupling inequalities and their applications. Technical report.