Vol. 0 (0000) 1–0

Asymptotically sufficient statistics in nonparametric regression experiments with correlated noise

Andrew V Carter

University of California, Santa Barbara
Santa Barbara, CA 93106-3110 e-mail: carter@pstat.ucsb.edu

Abstract: We find asymptotically sufficient statistics that could help simplify inference in nonparametric regression problems with correlated errors. These statistics are derived from a wavelet decomposition that is used to whiten the noise process and to effectively separate high resolution and low resolution components. The lower resolution components contain nearly all the available information about the mean function, and the higher resolution components can be used to estimate the error covariances. The strength of the correlation among the errors is related to the speed at which the variance of the higher resolution components shrinks, and this is considered an additional nuisance parameter in the model. We show that the NPR experiment with correlated noise is asymptotically equivalent to an experiment that observes the mean function in the presence of a continuous Gaussian process that is similar to a fractional Brownian motion. These results provide a theoretical motivation for some commonly proposed wavelet estimation techniques.

AMS 2000 subject classifications: Primary 62B15; secondary 62G20, 62G08.
Keywords and phrases: nonparametric regression, asymptotic equivalence, asymptotic sufficiency, fractional Brownian motion, correlated errors.

*Supported by NSF Grant DMS-08-05481

1. Introduction

A nonparametric regression (NPR) problem consists of estimating an unknown mean function that smoothly changes between observations at different design points. There are n observations Y_i of the form

$$Y_i = \mu\,(i/n) + \xi_i \qquad \text{for } i = 1, \ldots, n \tag{1}$$

where μ is the unknown smooth mean function on $[0,1]$ and the errors ξ_i are observations from a zero-mean Gaussian process. For NPR problems that have a particular long memory structure to the covariance of the error terms, we will find a continuous Gaussian experiment approximation to the problem of estimating the mean.

Brown and Low (1996) showed that the NPR experiment is asymptotically equivalent to the white-noise model where the mean function is observed in the presence of a Brownian motion process. This result paralleled work in Nussbaum (1996) in showing that asymptotic results in nonparametric function estimation problems can be simplified using approximations by the continuous white-noise experiments that Pinsker (1980) studied. The original asymptotic equivalence results for NPR experiments were extended by Brown et al. (2002) and Carter (2006, 2007) along with refinements in the approximations from Rohde (2004) and Reiss (2008).

All of these results assume that the errors ξ_i in (1) are all independent, and this assumption is critical in establishing the appropriateness of a white-noise model that also has independent increments. We want to consider the effect of correlation between the observations on these approximations. Presumably, if the correlation is weak then the effect washes out asymptotically. However, we wish to consider cases where there is sufficient long-range correlation to affect the form of the approximation. In particular, we will show that the appropriate approximation is by a continuous Gaussian process experiment that is no longer white noise but is closer to a fractional Brownian motion.

Our approach is motivated by the work in Johnstone and Silverman (1997) and Johnstone (1999). They investigated the wavelet decomposition of data of this type and used a fractional Brownian motion approximation in the limit:

$$dY(t) = \mu(t)\,dt + n^{-(\beta+1)/2} dB_K(t) \qquad t \in [0,1]. \tag{2}$$

They argued that the wavelet decomposition resulted in nearly independent coefficients which simplified the inference significantly. We will assume that the $B_K(t)$ process is decorrelated by a wavelet decomposition, and then show that this continuous model is asymptotically equivalent to the NPR experiment with the same covariance structure.

Theorem 1. *The nonparametric regression experiment $\mathcal{F}$ observes Y_i as in (1) for an unknown mean function μ from a parameter set $\mathcal{M}(M,\alpha)$ defined in Section 1.2 and a known covariance structure as described in Section 1.3. This experiment is asymptotically equivalent to the experiment $\mathcal{E}$ that observes*

$$dY(t) = \mu(t)\,dt + \sigma n^{-(\beta+1)/2} dB_K(t) \tag{3}$$

where $B_K(t)$ is a Brownian motion with covariance kernel K.

This will be proven in two steps. First, Lemma 1 proves that the first n wavelet coefficients in a decomposition of $dY(t)$ are asymptotically sufficient in $\mathcal{E}$ for estimating μ. For the second step, Lemma 2 shows that a discrete wavelet transform of the observations from $\mathcal{F}$ produces observations with nearly the same distribution as these asymptotically sufficient statistics.

Furthermore, in both experiments the lower frequency terms in the wavelet decomposition are sufficient for estimating the means, allowing the higher frequency terms to be used to give information about the variance process. This leads to Theorem 2, which proposes an experiment that allows some flexibility in the error structure.

Theorem 2. *The NPR experiment $\tilde{\mathcal{F}}$ observes the Y_i as in (1), where the covariance structure depends on the parameters β and γ and is such that the variance of the wavelet coefficients is $2^{\gamma+\beta(j+1)}$.*

The experiment $\hat{\mathcal{E}}$ observes the pair

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\beta} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \gamma \\ \beta \end{pmatrix}, \frac{\log 2}{2} \left(\mathbf{x}^{\mathsf{T}} \Lambda \mathbf{x}\right)^{-1} \right) \tag{4}$$

(where $\mathbf{x}$ and Λ are defined in Section 4) and then observes the continuous Gaussian process conditionally on $\hat{\gamma}$ and $\hat{\beta}$

$$dY(t) = \mu(t)\, dt + n^{-(1+\tilde{\beta})} 2^{\tilde{\gamma}} dB_{\hat{K}}(t)$$

where the covariance $\hat{K}$ is such that $\mathrm{Var}(B_{\hat{K}}(\psi_{jk})) = 2^{\tilde{\beta}(j+1)}$. The estimators $\tilde{\beta}$ and $\tilde{\gamma}$ are the same as $(\hat{\beta}, \hat{\gamma})$ but truncated so that $-1 \le \tilde{\beta} \le 0$, and $\tilde{\gamma} \ge -c$.

For $\mu \in \mathcal{M}(M, \alpha)$, $-1 < \beta \le 0$, and $\gamma > -c$ for some constant c, the experiments $\tilde{\mathcal{F}}$ and $\hat{\mathcal{E}}$ are asymptotically equivalent.

This theorem can be seen as an extension of Carter (2007) Theorem 1 from a case where there is a single unknown variance for all the wavelet coefficients to a case where the variance changes as a log-linear function of the resolution level (or frequency).

Wang (1996) addressed the issue of asymptotically sufficient statistics in the fractional Brownian motion process. In Section 3 of that article there is an argument that bounds the difference between minimax errors in an NPR experiment with correlated errors and an experiment that observes the mean in the presence of fractional Brownian motion error. This result extends the sort of approximation in Donoho and Johnstone (1999) to correlated errors, and is very much in the spirit of our Theorem 1 here. Our results differ from Wang (1996) in that we have made a stronger assumption on the covariance structure of the errors in order to obtain the full asymptotic equivalence of the experiments as discussed in section 1.1.

Lemma 1 is presented and proven in Section 2. Section 3 presents Lemma 2 and the proof of Theorem 1. The proof for Theorem 2 is in Section 4 with some relevant bounds in Sections 5 and 6.

1.1. Asymptotic sufficiency

Instead of focusing on single estimation techniques, we will consider approximations of the entire statistical experiment. For large sample sizes, there is often a simpler statistical experiment that can approximate the problem at hand. One benefit of finding an approximating experiment is that it may have convenient sufficient statistics even when they are not available in the original experiment.

Our approximations will therefore be of *experiments* rather than particular distributions. A statistical experiment $\mathcal{P}$ that observes data X consists of a set of probability distributions $\{\mathbb{P}_\theta\}$ indexed by the parameter set $\theta \in \Theta$. We wish to compare the information about θ in $\mathcal{P}$ to another experiment $\mathcal{Q}$ that observes data Y from among the set of distributions $\{\mathbb{Q}_\theta\}$ that are indexed by the same parameter θ. Implicitly, we are concerned with two sequences of experiments $\mathcal{P}_n$ and $\mathcal{Q}_n$ where n roughly denotes the increasing sample size, but generally, we will leave off the subscript n. It will always be understood that the distributions depend on the "sample size."

The NPR experiment will be approximated using Le Cam's notion of asymptotically equivalent experiments (Le Cam, 1964, 1986) and asymptotically sufficient statistics (Le Cam, 1974). Asymptotically equivalent experiments have corresponding inference procedures (such as estimators or tests) in each experiment that perform nearly as well. Specifically, if there is an estimator $\tau(X)$ in $\mathcal{P}$ with risk $\mathbb{P}_\theta L(\tau(X))$ then, for bounded loss functions, there is an estimator $\sigma(Y)$ such that

$$\sup_\theta \mathbb{P}_\theta L(\tau(X)) - \mathbb{Q}_\theta L(\sigma(Y)) \to 0$$

as $n \to \infty$. These asymptotic equivalence results are stronger than what the equivalence of minimax rates that is derived under a similar model in for example Wang (1996). Our results imply a correspondence over a range of bounded loss functions. Thus, the equivalence holds for a global L_2 error as well as local error measurements or other distances.

Asymptotic sufficiency is a stronger notion, where if $T(X)$ is a sufficient statistic for inference about θ in $\mathcal{P}$, then $T(Y)$ is asymptotically sufficient for $\mathcal{Q}$ when the total-variation distance between $\mathbb{P}_\theta$ and $\mathbb{Q}_\theta$ is negligible. These asymptotically sufficient statistics generate experiments that are all asymptotically equivalent. In particular, $\mathcal{P}$ and $\mathcal{Q}$ are asymptotically equivalent, and they are also asymptotically equivalent to the experiments generated by the distributions of $T(X)$ and $T(Y)$. As a result, an estimator in $\mathcal{P}$ should generally be of the form $\tau(T(X))$ and there is a corresponding estimator $\tau(T(Y))$ that performs nearly as well in the $\mathcal{Q}$ experiment. There is a basic transitive property to the asymptotic equivalence that implies if $\mathcal{P}$ is asymptotically equivalent to $\mathcal{Q}$, and $\mathcal{Q}$ is asymptotically equivalent to $\mathcal{R}$, then $\mathcal{P}$ is asymptotically equivalent to $\mathcal{R}$.

Le Cam's asymptotic equivalence is characterized using the total-variation distance $\delta(\mathbb{P}_\theta, \mathbb{Q}_\theta)$ between the distributions. We will abuse this notation a bit by writing $\delta(\mathcal{P}, \mathcal{Q}) = \sup_\theta \delta(\mathbb{P}_\theta, \mathbb{Q}_\theta)$. It will often often be more convenient to use the Kullback–Leibler divergence ($\mathbf{D}(\mathbb{P}, \mathbb{Q}) = \mathbb{P} \log [d\mathbb{P}/d\mathbb{Q}]$) to bound the

total variation distance

$$\delta(\mathbb{P}_\theta, \mathbb{Q}_\theta) \leq 2\mathbf{D}(\mathbb{P}_\theta, \mathbb{Q}_\theta)^{1/2}$$

(Kullback, 1967). The divergence is convenient for product experiments because $\mathbf{D}(\prod \mathbb{P}_i, \prod_i \mathbb{Q}_i) = \sum_i \mathbf{D}(\mathbb{P}_i, \mathbb{Q}_i)$.

1.2. Wavelet basis

We will use orthonormal wavelet bases to characterize the function space and to simplify the covariance structure of the errors.

Assuming we are considering periodic functions on the interval $[0, 1]$, we can construct periodic wavelet bases as in Daubechies (1992) Chapter 9.3. We start with a space $\mathcal{V}_j$ which consists of functions of the form

$$f_j(t) = \sum_{k=1}^{2^j} a_{jk}\phi_{jk}(t), \qquad t \in [0, 1]$$

where ϕ_{jk} is an orthonormal set of periodic functions generated via

$$\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k).$$

We will work with a ϕ function having finite support $[0, N]$, and at the boundaries of the interval the $\phi_{jk}(t)$ are given the proper periodic extensions. This space generates wavelet functions ψ_{jk} that span the difference between the $\mathcal{V}_j$ and $\mathcal{V}_{j-1}$, and can be written $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$ with the proper periodic adjustment at the boundary (for example, $\psi_{j,2^j}(\epsilon) = 2^{j/2}\psi\left(2^j\epsilon\right)$ for a small ϵ). This periodic adjustment has a small effect at the high resolution levels, but is a larger factor for small values of j. In particular, the scaling function at level 0 is $\phi_0(t) = \sum_{k=0}^{N} \phi(k + t) = 1$.

The mean functions $\mu(t)$ will be assumed to be constructed from this wavelet basis:

$$\mu(t) = \theta_0\phi_0(t) + \sum_{j\geq 0}\sum_{k=1}^{2^j} \theta_{jk}\psi_{jk}(t).$$

We will restrict the mean functions to those that belong to a Hölder(α) class of functions. Specifically, the class of periodic mean functions $\mu(t)$ is $\mathcal{M}(M, \alpha)$

$$\sup_{\mathcal{M}(M,\alpha)} \sup_{t,s} \frac{|\mu(t) - \mu(s)|}{|t - s|^\alpha} \leq M$$

for some $1/2 < \alpha < 1$ and $M > 0$. This smoothness condition on the functions bounds the rate of growth of the higher frequency terms in the orthonormal expansion. Originally from Meyer (1990), in Daubechies (1992) p. 299 equation (9.2.1) gives the bound

$$\max_k |\theta_{jk}| \leq M_2 2^{-j(\alpha+1/2)}$$

for a constant M_2 related to M. This bounds implies the useful bound on the squared error in the tail of the basis expansion

$$\sum_{j>j^*} 2^j \sum_{k=1}^{2^j} \theta_{jk}^2 \leq M_2^2 2^{-\varepsilon j^*} \left(1 - 2^{-\varepsilon}\right)^{-1} \tag{5}$$

where $\varepsilon = 2\alpha - 1$.

1.3. Error structure

These results rely on a specific structure to the covariance matrix of the errors in the NPR experiment. As in Johnstone (1999), the fractional Brownian motion is the motivating example for our continuous Gaussian model. However, this model does not necessarily provide the independent coefficients that would simplify the inference. Instead, an error structure that has roughly some of the properties of the fractional Brownian motion will be considered.

Traditionally, the asymptotics of the NPR experiment have assumed independent noise. This white-noise model is especially convenient because all of the eigenvalues of the covariance operator are equal. Thus, any orthonormal basis generates a set of independent standard normal coefficients. With a more general covariance function, the eigenvalues are different and only particular decompositions lead to independent coefficients. Thus there is much less flexibility in the choice of basis, and this basis determines some of the structure of the covariance.

Following Johnstone (1999); Johnstone and Silverman (1997); Zhang and Waiter (1994); Wang (1996) and Cavalier (2004) among others, we will assume a covariance structure that is whitened by a wavelet decomposition. When there is a long-range positive correlation between the observations, the wavelet decomposition tends to decorrelate the error process because the wavelet functions act like band-pass filters.

We will assume that there exists an orthonormal basis ϕ_0 and ψ_{jk} for $j \geq 0$ and $k = 0, \ldots, 2^j - 1$ such that the decomposition of the error process generates independent normal coefficients. In other words, the error process is a zero-mean Gaussian process that is roughly

$$dB_K(t) = \xi_0 \phi_0(t) + \sum_{j \geq 0} \sum_k \xi_{jk} \psi_{jk}(t)$$

in the distributional sense where the ξ_{kj} are independent normals. The $\mathrm{Var}(\xi_{jk})$ will be assumed to depend on j and not k as a sort of stationarity condition. In particular, we will assume that $\mathrm{Var}(\xi_0) = \sigma^2$ and then $\mathrm{Var}(\xi_{jk}) = \sigma^2 2^{\beta(j+1)}$ for some β in the interval $(-1, 0]$. If $\beta = 0$ then this is the white-noise process.

This is a convenient form for the error, but not completely unrealistic. Wavelet decompositions nearly whiten the fractional Brownian motion process. Wornell (1990) argued that long-memory processes can be constructed via a wavelet basis with variances at resolution level j shrinking like $2^{-\gamma j}$ for $0 < \gamma < 2$. McCoy

and Walden (1996) showed that the discrete wavelet transform nearly decorrelates the noise in fractionally differenced white-noise processes. Alternatively, Wang (1996) used a wavelet–vaguelette decomposition (Donoho, 1995) to find a decomposition of the fractional Brownian motion that results in independent coefficients for a nearly orthonormal basis.

Section 7 demonstrates some properties of the specific Gaussian process generated by using the Haar basis as the wavelet basis. These properties are consistent with the sort of behavior that we want in the covariances of our observations. The correlation between observations decreases like $d^{-(\beta+1)}$ for $\beta < 0$ where d measures the distance between the locations of the coefficients.

A well established method for estimating the parameter β in these long-range dependent models is to fit a linear function to the log of an estimate of the variances of the coefficients at each resolution level. This idea goes back to at least Abry and Veitch (1998) and is now a standard approach that has been improved upon in subsequent work, see Veitch and Abry (1999); Stoev et al. (2002) among others. This motivates the asymptotic sufficient statistics in Theorem 2 which are least squares estimates from the fitted line.

The assumptions in Theorem 1 on the covariance structure of the errors is strong and could limit the applicability of the result. However, if we allow the variances at different scales to have a range of linear relationship, we could then have a sufficiently rich class of error models. Theorem 2 allows for this somewhat larger class of models, and it seems likely that the changing magnitude of the variances over different resolutions level will have a greater effect on the distribution of the observed errors than the underlying basis.

2. Approximate sufficiency in the Gaussian sequence experiment

The first step in the proof of Theorem 1 is to establish that a truncated wavelet decomposition is asymptotically sufficient for the continuous Gaussian experiment.

Lemma 1. *The experiment $\mathcal{E}$ is a Gaussian process experiment that observes*

$$dY(t) = \mu(t)\, dt + \sigma n^{-(\beta+1)/2} dB_K(t) \tag{6}$$

where B_K is a zero-mean continuous Gaussian process with covariance K. There are asymptotically sufficient statistics for estimating $\mu \in \mathcal{M}(M, \alpha)$

$$y_0,\ y_{jk} \sim \mathcal{N}\left(\theta_{jk}, \sigma^2 n^{-\beta-1} 2^{\beta j}\right)$$

for $0 \le j \le j^*$ as long as

$$j^* > \frac{\log_2 n}{2\alpha}.$$

In the Gaussian sequence experiment $\mathcal{E}$ where only the mean $\mu(t)$ is to be estimated, the likelihood is

$$\frac{d\mathbb{P}_\mu}{d\mathbb{P}_0}(\mathbf{y}) = \exp\left[\frac{2y_0\theta_0 - \theta_0^2}{2\sigma^2} + \sigma^{-2} n^{(1+\beta)} \sum_{j\ge 0} \sigma^{-2} 2^{-\beta(j+1)} \sum_k \left(\theta_{jk} y_{jk} - \frac{1}{2}\theta_{kj}^2\right)\right]$$

where $\mathbb{P}_\mu$ is the distribution of $Y(t)$ and $\mathbb{P}_0$ is the distribution of the version with mean 0 which would just be $\sigma n^{-(\beta+1)/2} B_K(t)$.

We want to approximate this experiment $\mathcal{E}$ by a similar experiment $\mathcal{F}$ where the mean is projected onto the first j^* resolution levels, i.e. μ is replaced by $\bar{\mu}$

$$\bar{\mu}(t) = \theta_0 \phi_0(t) + \sum_{j=0}^{j^*} \sum_k \theta_{jk} \psi_{jk}(t). \tag{7}$$

The likelihood becomes

$$\frac{d\mathbb{Q}_\mu}{d\mathbb{Q}_0}(\mathbf{y}) = \exp\left[\frac{2y_0\theta_0 - \theta_0^2}{2\sigma^2} + \sigma^{-2} n^{(1+\beta)} \sum_{j=0}^{j^*} 2^{-\beta j} \sum_k \left(\theta_{jk} y_{jk} - \frac{1}{2}\theta_{kj}^2 \right) \right]. \tag{8}$$

Therefore, this experiment $\mathcal{F}$ has sufficient statistics y_0 and y_{jk} for $0 \le j \le j^*$. These observations are approximately sufficient in the $\mathcal{E}$ experiment if the distance between the distributions in the two experiments is small.

By (15), the distance between these two sets of experiments is

$$\delta(\mathbb{P}, \mathbb{Q}) \le \sigma^{-1} n^{(\beta+1)/2} \left(\sum_{j>J} 2^{-\beta j} \sum_k \theta_{jk}^2 \right)^{1/2}.$$

For the parameter space $\mathcal{M}(M, \alpha)$, (16) bounds the distance between the two experiments as

$$\begin{aligned} \delta(\mathbb{P}, \mathbb{Q}) &\le \sigma^{-1} n^{(\beta+1)/2} M 2^{-\varepsilon J} \\ &\le M \sigma^{-1} 2^{(J-j^*)(1+\beta)/2 - \varepsilon j^*} \end{aligned}$$

which is negligible as $n \to \infty$ when the dimension of the sufficient statistics increases like

$$j^* > \left(\frac{1+\beta}{\beta + 2\alpha} \right) \log_2 n$$

for $-1 < \beta < 0$. The worst case is when $\beta = 0$, and thus we have proved Lemma 1.

3. Approximating the NPR experiment

Theorem 1 can now be proven by approximating the sufficient statistics from Lemma 1 using the observations from the NPR experiment $\mathcal{F}$.

We suppose that we have n observations from the NPR experiment as in (1) where the ξ_i are Gaussian random variables with a specified covariance function. Specifically, let $\mathbf{W}$ be the $n \times n$ matrix that performs the discrete wavelet transform, and $\mathbf{W}^\mathsf{T}$ is its inverse. The vector of random wavelet coefficients from Lemma 1, $\mathbf{y} = (y_0, y_{00}, y_{01}, \ldots, y_{J-1,2^{J-1}-1})^\mathsf{T}$ where $J = \log_2 n$, can be transformed via the discrete wavelet transform to create $\tilde{\mathbf{y}}_J = \mathbf{W}^\mathsf{T}\mathbf{y}$.

The expected value of this transformed vector is

$$\mathbb{E}\tilde{y}_{Ji} = \tilde{\mu}_{Ji} = \int 2^{J/2}\phi(2^J t - i)\mu(t)\, dt.$$

For a $\mu(t)$ function that is nearly constant around i/n, $\tilde{\mu}_{Ji} \approx 2^{-J/2}\mu(i/n)$ so we can approximate $\tilde{\mathbf{y}}_J$ by $\frac{1}{\sqrt{n}}(Y_1, Y_2, \ldots, Y_N)^{\mathsf{T}}$.

In the original NPR experiment, the variances $\mathrm{Var}(Y_i) = \mathrm{Var}(\xi_i) = C\sigma^2$ for a constant C that depends on β and the basis we'll be using. The covariance matrix for these Y_i will be assumed to be $\Sigma = \mathbf{WDW}^{\mathsf{T}}$ where $\mathbf{D}$ is the diagonal matrix of $\mathrm{Var}(y_{jk}) = \sigma_n^2 2^{\beta(j+1)}$. The variance of the $\tilde{y}_{Jk}$ should be the same as that of $Y_i n^{-1/2}$, and in the model described, $\mathrm{Var}(\tilde{y}_{Jk}) \propto \sigma_n^2 n^{\beta}$. Therefore, σ_n^2 should be set to $\sigma^2 n^{-1-\beta}$.

Lemma 2. *If the mean function $\mu(t)$ is in $\mathcal{M}(M,\alpha)$ then the total variation distance between the distribution $\mathbb{P}_\mu$ of the vector $\frac{1}{\sqrt{n}}(Y_1, Y_2, \ldots, Y_N)^{\mathsf{T}}$ and the distribution $\tilde{\mathbb{P}}_\mu$ of the vector $\tilde{y}_J$ is*

$$\sup_{\mu\in\mathcal{M}(M,\alpha)} \delta\left(\mathbb{P}_\mu, \tilde{\mathbb{P}}_\mu\right) \le \left(\frac{MN^{\alpha+1}}{\alpha+1}\right) n^{1/2-\alpha+\beta/2}.$$

This lemma essentially goes back to the original work of Mallat (1989) and parallels some of what was done in Brown and Low (1996), Johnstone (1999), Rohde (2004), and Carter (2006).

The NPR observations are such that the covariance matrix of ξ_i is also Σ, and therefore the total variation distance between the distributions is bounded in (14) by

$$\delta(\mathbb{P}, \tilde{\mathbb{P}}) \le \frac{1}{\sqrt{2\pi}}\Delta^{1/2}$$

with

$$\Delta = \left(\mu(i/n) - \tilde{\mu}_{Ji}\right)^{\mathsf{T}} \Sigma^{-1} \left(\mu(i/n) - \tilde{\mu}_{Ji}\right).$$

A standard calculation bounds the difference between the means when $\phi(t) < M$ with support on $[0, N]$ and the $\mu(t)$ are Hölder (α) for $\alpha < 1$

$$\left|\mu\left(\frac{i}{n}\right) - \tilde{\mu}_{Ji}\right| \le n^{-1/2-\alpha}\frac{M}{\alpha+1}N^{\alpha+1}.$$

The covariance matrix is a positive definite matrix such that $\Sigma^{-1} = \mathbf{MD}^{-1}\mathbf{M}^{\mathsf{T}}$. The first column of the wavelet transform matrix is $\sqrt{n}\mathbf{1}$ where $\mathbf{1}$ is the vector of 1's. Therefore,

$$\Delta \le \left(\frac{MN^{\alpha+1}}{\alpha+1}\right) n^{-1-2\alpha} n n^{1+\beta} = \left(\frac{MN^{\alpha+1}}{\alpha+1}\right) n^{1-2\alpha+\beta}$$

which is negligible for large n and $\alpha > 1/2$.

3.1. *Proof of Theorem 1*

The theorem follows from the fact that the observations $y_0, \{y_{jk}\}$ for $j = 0, \ldots, J-1$ are asymptotically sufficient for the continuous process in (2). Then a linear function of these sufficient statistics $\tilde{\mathbf{y}} = \mathbf{W}^{\mathsf{T}}\mathbf{y}$ is still approximately sufficient. Thus, the experiment that seeks to draw inference about μ from the observations $\tilde{y}_{Ji}$ is asymptotically equivalent to the experiment that observes (2) by Lemma 1.

Furthermore, by Lemma 2, the original NPR experiment that has the same covariance structure as $\tilde{y}_{Ji}$ is asymptotically equivalent to that experiment and thus, by transitivity, to the experiment that observes the process (2) as well. This proves Theorem 1.

3.2. *Remarks on the covariance structure*

This result is restrictive in that it requires a specific known covariance structure. We are working under the assumption that the covariance matrix has eigenfunctions that correspond to a wavelet basis. This does not generally lead to a typical covariance structure. It does not even necessarily lead to a stationary Gaussian process; see the Haar basis example below.

The difficulty is that the requirement for having asymptotically equivalent experiments is quite strict, and the total variation distance between the processes with even small differences in the structure of the covariance is not negligible. For two multivariate Guassian distributions with the same means but where one covariance matrix is Σ and the other is D, a diagonal matrix with the same diagonal elements as Σ, the Kullback–Leibler divergence between the distributions is $\log|\Sigma| - \log|D|$.

If the correlation between the highest level coefficients $\mathrm{Corr}\,(\xi_{j^*,k},\ \xi_{j^*,k+1}) = \rho$ then the contribution to the difference of the log determinants is on the order of $\rho 2^{j^*}$. The dimension of the problem is growing while the correlations are generally not going to 0 significantly quickly. For instance, in a typical wavelet basis decomposition of the true fractional Brownian motion $\mathrm{Corr}\,(\xi_{j^*,k},\ \xi_{j^*,k+1}) = c_\beta$ where c_β is a constant that depends on β but not j^* or n.

Thus, the difference $\log|\Sigma| - \log|D|$ will not go to 0 as the sample size increases. Therefore, for the sort of long-range correlation structures that we are considering here, the eigenfunctions of the kernel K need to be known or else the experiments will not be asymptotically equivalent.

4. Estimating the covariance of the increments.

The key limitation of Theorem 1 is that it supposes that the covariance structure of the errors is known to the experimenter. To make the approximation more useful, it would help if the covariance structure was more flexible. A strategy similar to that used by Carter (2007) can be used to estimate the variances of the coefficients.

In Carter (2007), I showed that a model with a variance that changes slowly over time can still be approximated by the Gaussian process as long as all of the observations are independent. Our result here is that for correlated observations, if the variance is a linear function of the frequency then a similar technique can be used to establish a set of asymptotically sufficient statistics.

Flexibility with regard to the covariance structure is added by allowing the magnitude of the $\mathrm{Var}(y_{jk})$ to depend on the resolution level j. The variances will be described by two parameters γ and β, which characterize the size of the error and the speed that it shrinks at higher resolution levels. These nuisance parameters can be estimated using part of the data, and then the inference can be carried out conditionally on the estimates.

Specifically, the experiment $\mathcal{E}_n$ observes independent components $y_0 \sim \mathcal{N}\left(\theta_0, n^{-(1+\beta)}2^{\gamma}\right)$ and

$$y_{jk} \sim \mathcal{N}\left(\theta_{jk},\ \ n^{-(1+\beta)}2^{\gamma+(j+1)\beta}\right) \qquad \text{for } 0 \le j \le J$$

where the $n^{-(1+\beta)}$ factor is included to match up with the scaling functions at the Jth resolution level. These observations form a new experiment with a parameter set that includes (μ, γ, β) where $\mu(t) \in \mathcal{M}(M, \varepsilon)$, $-1 < \beta < 0$ and γ is bounded below by a constant $-c$.

This experiment $\mathcal{E}_n$ with the parametric variances is no longer an approximately sufficient statistic for the experiment that observes all of the θ_{jk}. That experiment has too much information about the variance. If we observed the entire sequence at all resolution levels j then γ and β could be estimated exactly. We need to adopt another approximating experiment as in Carter (2007). Many of the bounds in this section follow arguments from that paper.

4.1. Proof of Theorem 2

The theorem can be proven by applying Lemma 2 and then a version of Lemma 1 that uses only a small proportion of the low frequency wavelet coefficients. The rest of the coefficients can be used to fix the parameters in the covariance of the observations.

The first step is to decompose the nonparametric regression into a set of wavelet coefficients. The n NPR observations Y_i can be transformed by dividing by $\sqrt{n}$ and then performing the discrete wavelet transformation as in Lemma 2. The result is that a sequence of n wavelet coefficients y_0 and y_{jk} for $j = 0, \ldots, J-1$ is equivalent to the original NPR observations with a total-variation distance between the distributions of

$$\delta\left(\mathbb{P}_{\mu,\gamma,\beta}, \tilde{\mathbb{P}}_{\mu\gamma\beta}\right) \le C 2^{-\gamma} n^{1-2\alpha+\beta}.$$

The supremum of this bound over all $\gamma > -c$ and $\beta < 0$ is

$$\delta\left(\mathcal{P}, \tilde{\mathcal{P}}\right) \le C 2^{c} n^{1-2\alpha}$$

which will be negative for $\alpha > 1/2$.

The key strategy is to break the observations from this wavelet composition into pieces starting at level j^*, where observations on $j \leq j^*$ are assumed to be informative about the means, and the higher resolution levels are used to estimate the covariance structure.

For each resolution level with $j > j^*$, we generate the approximately sufficient statistics $V_j = \sum_k y_{jk}^2$. Along with the y_{jk} for $j \leq j^*$, the collection of V_j are exactly sufficient if the means are $\theta_{jk} = 0$ for $j > j^*$, because if there is no information about the means in the higher frequency terms, then we have a piece of the experiment that is like a normal scale family. This new experiment $\mathcal{E}_v$ is asymptotically equivalent to our $\mathcal{E}_n$.

The error in approximating $\mathcal{E}_n$ by $\mathcal{E}_v$, where the means of the higher frequencies coefficients are 0, is bounded by (15)

$$\delta(\mathcal{E}_v, \mathcal{E}_n) \leq \left(\sum_{j=j^*+1}^{J-1} n^{1+\beta} 2^{-(\gamma+(j+1)\beta)} \sum_k \theta_{jk}^2 \right)^{1/2}.$$

For θ_{jk} in $\mathcal{M}(M, \varepsilon)$ space, (16) bounds the distance as bound

$$\delta\left(\mathcal{E}_v, \mathcal{E}_n\right) \leq M 2^{-\gamma/2} 2^{(J-j^*-1)(1+\beta)/2-\beta/2-\varepsilon j^*} \leq M 2^{c/2+1/2} 2^{J/2-(1/2+\varepsilon)j^*} \quad (9)$$

which is negligible when $j^* > J/(1+2\varepsilon)$.

This $\mathcal{E}_v$ experiment has sufficient statistics y_0, y_{jk} for $j \leq j^*$, and

$$V_j \sim \Gamma\left(2^j, \;\; n^{-(1+\beta)} 2^{\gamma+\beta(j+1)-j}\right) \text{ for } j = j^*+1, \ldots, J-1.$$

Furthermore, there are approximately sufficient statistics in this experiment $(y_{jk}, \hat{\gamma}, \hat{\beta})$ where $\hat{\gamma}$ and $\hat{\beta}$ are the weighted least-squares estimates of γ and β from the data $\log V_j$. These are exactly sufficient statistics in the experiment $\mathcal{E}_\ell$ that observes the y_0 and y_{jk} for the lower resolution levels $j \leq j^*$ as before, in addition to the observations 2^{W_j} for $j^* < j < J$ where

$$W_j \sim \mathcal{N}\left(-(1+\beta)J + \gamma + \beta(j+1), 2^{-j+1}(\log 2)\right).$$

The distance between $\mathcal{E}_\ell$ and $\mathcal{E}_v$ depends on the distance between the distribution of the log of the Gamma variables and the normal approximation to this distribution. The calculation in section 10.1 of Carter (2007) gives a bound on the Kullback–Leibler divergence of $\mathbf{D}\left(\mathbb{Q}_j, \breve{\mathbb{Q}}_j\right) \leq 2^{-j}$ where $\mathbb{Q}_j$ is the distribution of V_j, and $\breve{\mathbb{Q}}_j$ is the distribution of 2^{W_j}. Therefore, the total error between the two experiments is

$$\delta(\mathcal{E}_\ell, \mathcal{E}_v) \leq \left(\sum_{j=j^*+1}^{J-1} 2^{-j} \right)^{1/2} \leq 2^{-j^*/2}.$$

Therefore, the observations in $\mathcal{E}_\ell$ are asymptotically sufficient for $\mathcal{E}_v$ and thus also $\mathcal{E}_n$ (as long as $j^* \to \infty$ with n).

In the experiment $\mathcal{E}_\ell$, the sufficient statistics for estimating γ and β are the weighted least-squares estimators $\hat{\gamma}$ and $\hat{\beta}$

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\beta} \end{pmatrix} = \left(\mathbf{x}^{\mathsf{T}}\Lambda\mathbf{x}\right)^{-1}\mathbf{x}^{\mathsf{T}}\Lambda\left(\mathbf{W}+J\right)$$

where Λ is the diagonal matrix with 2^j for $j = j^*+1,\ldots,J-1$ along its diagonal, $\mathbf{x}$ is the design matrix with rows $(1, j-J+1)$, and $\mathbf{W}+J$ is the column of observations $W_j - J$. The vector of estimators is normal with mean $(\gamma \;\; \beta)^{\mathsf{T}}$ and covariance $\frac{1}{2}(\log 2)\left(\mathbf{x}^{\mathsf{T}}\Lambda\mathbf{x}\right)^{-1}$.

Therefore, we can compare this experiment $\mathcal{E}_\ell$ to an experiment $\hat{\mathcal{E}}$ that observes the same $(\hat{\gamma} \;\; \hat{\beta})^{\mathsf{T}}$, but the y_0 and y_{jk} for $j \le j^*$ are replaced by Gaussian random variables $\hat{y}_{jk}$ with variances (conditional on $(\hat{\gamma},\hat{\beta})$) that are $\mathrm{Var}(\hat{y}_{jk}) = 2^{\hat{\gamma}+(j-J+1)\hat{\beta}-J}$. The error in this approximation depends on the distance between the two sets of independent normal experiments with different variances. Letting $\mathbb{P}_{jk}$ be the distribution of y_{jk} and $\hat{\mathbb{P}}_{jk}$ the distribution of $\hat{y}_{jk}$, the bound (22) in Section 6 gives

$$\mathbf{D}\left(\mathbb{Q}_{jk},\hat{\mathbb{Q}}_{jk}\right) \le \frac{(\log n)^2}{n-2^{j^*}} + O\left(\frac{(\log n)^4}{(n-2^{j^*+1})^2}\right),$$

and

$$\mathbf{D}\left(\mathbb{Q}_0,\hat{\mathbb{Q}}_0\right) \le \frac{(\log n)^2}{n-2^{j^*}} + O\left(\frac{(\log n)^4}{(n-2^{j^*+1})^2}\right).$$

There are 2^{j^*+1} independent normals y_{jk} for $j \le j^*$ so that the total divergence is

$$\delta(\mathcal{E}_\ell,\hat{\mathcal{E}})^2 \le \frac{2(\log n)^2}{2^{J-j^*}-1} + O\left(\frac{2^{j^*}(\log n)^4}{(n-2^{j^*+1})^2}\right). \tag{10}$$

Therefore, the experiments $\mathcal{E}_\ell$ and $\hat{\mathcal{E}}$ are asymptotically equivalent for

$$j^* = J - 2\log_2 J - \eta_n \tag{11}$$

for some $\eta_n \to \infty$.

We can improve this approximation by replacing the estimators $\hat{\beta}$ and $\hat{\gamma}$ in $\hat{\mathcal{E}}$ by using

$$\tilde{\beta} = \begin{cases} -1 & \hat{\beta} \le -1 \\ \hat{\beta} & -1 < \hat{\beta} < 0 \\ 0 & \hat{\beta} \ge 0 \end{cases}$$

and $\tilde{\gamma} = \hat{\gamma} \vee c$ to match up with the bounds on the parameter space. The new version of this experiment therefore observes $(\hat{\gamma},\hat{\beta})$ and the normal coordinates $y_{jk} \sim \mathcal{N}\left(\theta_{jk},\, n^{1+\tilde{\beta}}2^{\tilde{\gamma}}2^{\tilde{\beta}(j+1)}\right)$. for $0 \le j \le j^*$. The error between $\mathcal{E}_\ell$ and this new version of $\hat{\mathcal{E}}$ is smaller because $\left|\gamma-\tilde{\gamma}+(\beta-\tilde{\beta})(j-J)\right| \le \left|\gamma-\hat{\gamma}+(\beta-\hat{\beta})(j-J)\right|$, which makes the bound in (19) uniformly smaller.

Finally, we create a continuous Gaussian version of the $\hat{\mathcal{E}}$ experiment. This approximation $\tilde{\mathcal{E}}$ observes all the $\tilde{y}_{kj}$ for $k \geq 0$ with means θ_{jk} and variances $n^{-(1+\tilde{\beta})}2^{\tilde{\gamma}+\tilde{\beta}(j+1)}$. The $\hat{\mathcal{E}}$ are actually sufficient statistics for an experiment that observes $(\hat{\gamma}, \hat{\beta})$ and y_{jk} for $0 \leq j \leq j^*$ and for $j > j^*$

$$y_{jk} \sim \mathcal{N}\left(0, n^{-(1+\tilde{\beta})}2^{\tilde{\gamma}+\tilde{\beta}(j+1)}\right).$$

The difference between the experiments $\hat{\mathcal{E}}$ and $\tilde{\mathcal{E}}$ conditional on $\tilde{\gamma}$ and $\tilde{\beta}$ is as in Section 2 and (16) less than $M2^{-\tilde{\gamma}}2^{(J-j^*)(1+\tilde{\beta})/2-\tilde{\beta}/2-\varepsilon j^*}$. The expectation of this bound when averaged over the possible values of $(\tilde{\gamma}, \tilde{\beta})$ is a bound on the unconditional error. Furthermore, this expectation is less than the minimum over possible values of $(\tilde{\gamma}, \tilde{\beta})$ (this is the real advantage that comes from going from $(\hat{\gamma}, \hat{\beta})$ to $(\tilde{\gamma}, \tilde{\beta})$). Thus,

$$\delta\left(\hat{\mathcal{E}}, \tilde{\mathcal{E}}\right) \leq 2^{c-1}2^{(J-j^*)/2-\varepsilon j^*}. \tag{12}$$

As before, this is asymptotically negligible when $j^* > J/(1+2\varepsilon)$.

All that is left to do is choose the level j^* at which to split the data so that the requirements from (9) and (12) that $j^* > J/(1+2\varepsilon)$ and from (11) that $j^* = J - 2\log_2 J - \eta_n$ are all fulfilled. We could choose $\eta_n = \frac{\varepsilon J}{1+2\varepsilon} - \log_2 J$ so that

$$j^* = \frac{J}{1+2\varepsilon} + \frac{\varepsilon J}{1+2\varepsilon} - \log_2 J$$

which is greater than $J/(1+2\varepsilon)$ for $\varepsilon > 0$. This choice of j^* plugged into the bound in (9) gives us

$$\delta(\mathcal{E}_v, \mathcal{E}_n) \leq N2^{(c+1)/2}J^{\frac{1}{2}+\varepsilon}2^{-\varepsilon J/2} \to 0$$

as $J \to \infty$ for $\varepsilon > 0$. At the same time, the bound in (10) becomes

$$\delta(\mathcal{E}_\ell, \hat{\mathcal{E}})^2 \leq \frac{2(\log 2)^2 J}{2^{J\varepsilon/(1+2\varepsilon)} - J^{-1}} \to 0.$$

Thus, Theorem 2 is established.

5. Bounding the total variation distance.

We need a bound on the distance between two multivariate normal distributions with different means in order to bound the error in many of our approximations.

For shifted Gaussian processes, the total-variation distance between the distributions is

$$\delta(\mathbb{P}_{\mu_1}, \mathbb{P}_{\mu_2}) = 1 - 2\Phi(-\Delta^{1/2}/2) \tag{13}$$

where $\Delta = (\mu_1 - \mu_2)^{\mathsf{T}}\Sigma^{-1}(\mu_1 - \mu_2)$. The expression in (13) for the total variation distance is concave for positive Δ, so a simple expansion gives

$$\delta(\mathbb{P}_{\mu_1}, \mathbb{P}_{\mu_2}) \leq \frac{1}{\sqrt{2\pi}}\Delta^{1/2}. \tag{14}$$

For the Gaussian process with correlated components, we will assume that the variance of each wavelet coefficient is of the form $\mathrm{Var}\,(\psi_{jk}) = \sigma^2 n^{-(1+\beta)} 2^{\beta(j+1)}$ where the variance is calibrated so that $\mathrm{Var}(Y_i) = \sigma^2 = n\,\mathrm{Var}(B_K(\phi_{J,\ell}))$. A bound on the error in the projection onto the span of ψ_{jk} for $j > j^*$ comes from (14) which depends on

$$\begin{aligned}\Delta &= \frac{n^{1+\beta}}{\sigma^2} \sum_{j>j^*} 2^{-\beta(j+1)} \sum_k \theta_{jk}^2 \qquad (15)\\ &\le \frac{n^{1+\beta}}{\sigma^2} 2^{-(1+\beta)j^*-\beta} \sum_{j>j^*} 2^j \sum_j \theta_{jk}^2 \\ &\le M_2^2 \sigma^{-2} 2^{(J-j^*)(1+\beta)-\varepsilon j^*} \left(2^\beta - 2^{\varepsilon+\beta}\right)^{-1} \qquad (16)\end{aligned}$$

where the upper bound in (16) follows from $n = 2^J$, the definition of $\mathcal{M}(M,\alpha)$ and the bound in (5), and $-1 < \beta < 0$. This error is negligible as $J \to \infty$ whenever

$$j^* > \frac{J}{2\alpha}. \qquad (17)$$

6. Bounds from the estimated variances.

In order to expand our asymptotically sufficient statistics out into a continuous Gaussian experiment, we need a bound on the total-variation distance between $\mathcal{E}_n$ which, for $0 \le j \le j^*$, observes a sequence of normals with variances $n^{-(1+\beta)} 2^{\gamma+\beta(j+1)}$ and $\mathcal{E}_g$, which observes a similar set of normals with variances $n^{-(1+\beta)} 2^{\hat\gamma+\hat\beta(j+1)}$.

For two normal distributions with the same means μ, and variances σ_1^2 and σ_2^2 respectively, the Kullback–Leibler divergence is

$$\mathbf{D}(N_1, N_2) = \frac{1}{2}\left[\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)\right]. \qquad (18)$$

Thus, for $\mathbb{Q}_{kj}$ (the distribution of the y_{kj}) and $\hat{\mathbb{Q}}_{kj}$ (the distribution of the $\hat y_{kj}$), the divergence between the conditional distributions given $\hat\gamma$ and $\hat\beta$ is

$$\begin{aligned}\mathbf{D}\left(\mathbb{Q}_{kj}, \hat{\mathbb{Q}}_{kj} \mid \hat\gamma, \hat\beta\right) = \frac{1}{2} 2^{\gamma-\hat\gamma+(j+1-J)\left(\beta-\hat\beta\right)} + \\ - \frac{1}{2} - \frac{1}{2}\left[\gamma - \hat\gamma + (j+1-J)\left(\beta-\hat\beta\right)\right]\log 2, \qquad (19)\end{aligned}$$

and

$$\mathbf{D}\left(\mathbb{Q}_0, \hat{\mathbb{Q}}_0 \mid \hat\gamma, \hat\beta\right) = \frac{1}{2} 2^{\gamma-\hat\gamma} - \frac{1}{2} - \frac{(\gamma-\hat\gamma)\log 2}{2}. \qquad (20)$$

This divergence between conditional distributions can be used to bound the joint divergence

$$\mathbf{D}\left(\mathbb{Q}_{kj}, \hat{\mathbb{Q}}_{kj}\right) = \mathbb{E}\mathbf{D}\left(\mathbb{Q}_{kj}, \hat{\mathbb{Q}}_{kj} \mid \hat\gamma, \hat\beta\right)$$

where the expectation is taken over the estimators $\hat{\gamma}$ and $\hat{\beta}$.

To bound the expected value of the divergence in (19), we need the distribution of the estimators

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\beta} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \gamma \\ \beta \end{pmatrix}, \frac{\log 2}{2} \left(\mathbf{x}^{\mathsf{T}} \Lambda \mathbf{x}\right)^{-1} \right).$$

This implies that

$$\mathbb{E} \exp\left[\log 2 \left(\gamma - \hat{\gamma} + (j+1-J)\left(\beta - \hat{\beta}\right)\right)\right] = \exp\left[\frac{(\log 2)^2}{2} \operatorname{Var}\left(\hat{\gamma} + (j+1-J)\hat{\beta}\right)\right],$$

and therefore

$$\mathbb{E}\mathbf{D}\left(\mathbb{Q}_{kj}, \hat{\mathbb{Q}}_{kj} \mid \hat{\gamma}, \hat{\beta}\right) = \exp\left[\frac{(\log 2)^2}{2} \operatorname{Var}\left(\hat{\gamma} + (j+1-J)\hat{\beta}\right)\right] - \frac{1}{2}.$$

Via elementary linear algebra calculations we get that for

$$\zeta = \left(\sum_{i=j^*+1}^{J-1} 2^i\right)\left(\sum_{i=j^*+1}^{J-1} (J-i-1)^2 2^i\right) + \left(\sum_{i=j^*+1}^{J-1} (J-i-1)2^i\right)^2,$$

$$\begin{aligned}
\operatorname{Var}\left(\hat{\gamma}\right) &= \frac{\log 2}{2}\left(\sum_{i=j^*+1}^{J-1} (J-i-1)^2 2^i\right)\zeta^{-1} \\
\operatorname{Var}\left(\hat{\beta}\right) &= \frac{\log 2}{2}\left(\sum_{i=j^*+1}^{J-1} 2^i\right)\zeta^{-1} \\
\operatorname{Cov}\left(\hat{\gamma}, \hat{\beta}\right) &= \frac{\log 2}{2}\left(\sum_{i=j^*+1}^{J-1} (J-i-1)2^i\right)\zeta^{-1}.
\end{aligned}$$

As a result,

$$\begin{aligned}
&\operatorname{Var}\left(\hat{\gamma} + (j+1-J)\hat{\beta}\right) = \frac{\log 2}{2} \times \\
&\left[\sum_i (J-i-1)^2 2^i + (J-j-1)^2 \sum_i 2^i - 2(J-j-1)\sum_i (J-i-1)2^i\right]\zeta^{-1}.
\end{aligned} \tag{21}$$

To simplify this expression, let X be a random variable with probability mass function proportional to 2^i for $i = j^*+1, \ldots, J-1$. Thus, the ζ is equal to $\left(\sum 2^i\right)^2 \left[\mathbb{E}(J-X-1)^2 - [\mathbb{E}(J-X-1)]^2\right] = \left(\sum 2^i\right)^2 \operatorname{Var}(X)$. Similarly, the main factor in (21) is equal to

$$\left(\sum 2^i\right)\left[\mathbb{E}(J-X-1)^2 + (J-j-1)^2 - 2(J-j-1)\mathbb{E}(J-X-1)\right] = \left(\sum 2^i\right)\mathbb{E}(X-j)^2.$$

A simple bound of $0 < X - j < J$ leads to $\mathbb{E}(X-j)^2 < J^2$. Furthermore, the variance of X is decreasing in j^*, and thus it is greater than $2/9$ when $j^* < J-2$. Therefore,

$$\operatorname{Var}\left(\hat{\gamma} + (j+1-J)\hat{\beta}\right) \leq \frac{9\log 2}{4}\frac{J^2}{\sum 2^i} \leq \frac{J^2}{2^{J-1}-2^{j^*}}$$

because $9\log 2/4 < 2$.

Thus,

$$\begin{aligned}
\mathbf{D}\left(\mathbb{Q}_{jk}, \hat{\mathbb{Q}}_{jk}\right) &= \frac{1}{2}\exp\left[\frac{(\log 2)^2}{2}\operatorname{Var}\left(\hat{\gamma} + (j+1-J)\hat{\beta}\right)\right] - \frac{1}{2} \\
&\leq \frac{1}{2}\exp\left[\frac{(\log 2)^2 J^2}{2^J - 2^{j^*+1}}\right] - \frac{1}{2} \\
&\leq \frac{(\log 2)^2 J^2}{2^J - 2^{j^*+1}} + \frac{CJ^4}{(2^J - 2^{j^*})^2} \\
&= \frac{(\log n)^2}{n - 2^{j^*+1}} + O\left(\frac{(\log n)^4}{(n-2^{j^*+1})^2}\right). \qquad (22)
\end{aligned}$$

Analogously, the expected divergence between $\mathbb{Q}_0$ and $\hat{\mathbb{Q}}_{jk}$ is bounded by

$$\begin{aligned}
\mathbb{E}\mathbf{D}\left(\mathbb{Q}_0, \hat{\mathbb{Q}}_0 \mid \hat{\gamma}, \hat{\beta}\right) &= \frac{1}{2}\exp\left[-\frac{(\log 2)^2}{2}\operatorname{Var}(\hat{\gamma})\right] - \frac{1}{2} \\
&= \frac{1}{2}\exp\left[-\frac{(\log 2)^2}{2}\left(\frac{(\log 2)\mathbb{E}(J-X-1)^2}{2\left(\sum 2^i\right)\operatorname{Var}(X)}\right)\right] - \frac{1}{2} \\
&\leq \frac{(\log 2)^2 J^2}{2^J - 2^{j^*+1}} + \frac{CJ^4}{(2^J - 2^{j^*})^2}.
\end{aligned}$$

If we add up these errors over the 2^{j^*} observations in the experiment, we get that the error in the approximation is less than $C(\log n)^2/(n2^{-j^*}-1)$, which is negligible for j^* sufficiently small.

7. Haar basis covariance

The Haar basis is a simple enough wavelet basis that we can make some explicit calculations of the properties of the error distribution. We will show that the resulting errors ξ_i will have variances of approximately n^β as we expected, and the correlation between ξ_i and ξ_j will decrease at about a rate of $|i-j|^{-(1+\beta)}$.

The scaling functions for the Haar basis are constant on 2^j dyadic intervals at the resolution level j. The assumption is that we have a single scaling function coefficient with $\operatorname{Var}(y_0) = 1$, and then every wavelet coefficient y_{jk} is independent and has variance $2^{\beta(j+1)}$. Then the covariances can be calculated from the synthesis formula for the Haar basis.

The formula for synthesizing the scaling function coefficients $\tilde{y}_{Jk}$ from the wavelet decomposition is

$$\tilde{y}_{Jk} = 2^{-J/2} y_0 + \sum_{j=0}^{J-1} \zeta_{j,J,k} 2^{(j-J)/2} y_{jk^*}$$

where k^* is the index such that ψ_{jk^*} has support that includes the support of ϕ_{Jk}. The $\zeta_{j,J,k}$ is either 1 or -1 depending on whether ϕ_{Jk} sits in the positive or negative half of the ψ_{jk^*} function.

Using the covariance structure described above, the variance of $\tilde{y}_{Jk}$ is

$$\begin{aligned}
\text{Var}\left(\tilde{y}_{Jk}\right) &= 2^{-J} + \sum_{j=0}^{J-1} 2^{(j-J)+\beta(j+1)} \\
&= 2^{-J} + 2^{\beta-J}\left[\frac{2^{(1+\beta)J}-1}{2^{1+\beta}-1}\right] \\
&= 2^{\beta J}\left[\frac{1}{2-2^{-\beta}}\right] - 2^{-J}\left[\frac{2^{-\beta}-1}{2-2^{-\beta}}\right]
\end{aligned}$$

for $-1 < \beta < 0$. For $\beta = 0$, the variance of each scaling function coefficient is 1 as in white noise. For $\beta = -1$, direct calculation leads to a variance of $2^{-J}(1 + J/2)$.

To find the covariance between two variables $\tilde{y}_{Jk_1}$ and $\tilde{y}_{Jk_2}$, we need j^* which is the highest resolution level such that the support of $\psi_{j^*k^*}$ includes the support of both scaling functions ϕ_{Jk_1} and ϕ_{Jk_2}. The covariance is thus

$$\begin{aligned}
\text{Cov}\left(\tilde{y}_{Jk_1}, \tilde{y}_{Jk_2}\right) &= 2^{-J} + \sum_{j=0}^{j^*-1} 2^{(j-J)+\beta(j+1)} - 2^{j^*-J+\beta(j^*+1)} \\
&= 2^{-J} - 2^{j^*-J+\beta(j^*+1)} + 2^{\beta-J}\left[\frac{2^{(1+\beta)j^*}-1}{2^{1+\beta}-1}\right] \\
&= 2^{j^*-J+\beta(j^*+1)}\left[\frac{2-2^{1+\beta}}{2^{1+\beta}-1}\right] - 2^{-J}\left[\frac{2^{-\beta}-1}{2-2^{-\beta}}\right] \\
&= 2^{\beta J}\left(2^{-(1+\beta)(J-j^*)}\right)\left[\frac{1-2^{\beta}}{1-2^{-\beta-1}}\right] - 2^{-J}\left[\frac{2^{-\beta}-1}{2-2^{-\beta}}\right].
\end{aligned}$$

For large J the correlation is on the order of $d^{-(1+\beta)}$ where $d = 2^{J-j^*}$ is a proxy for the distance between the observations. For $\beta = 0$, all of these covariances are 0. For $\beta = -1$, the correlation is $j^*/(J+2)$.

References

Abry, P. and Veitch, D. (1998). Wavelet analysis of long-range-dependent traffic. *IEEE Trans. Inform. Theory* **44** 2–15.

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

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

Carter, A. (2006). A continuous Gaussian approximation to a nonparametric regression in two dimensions. *Bernoulli* **12** 143–156.

Carter, A. (2007). Asymptotic approximation of nonparametric regression experiments with unknown variances. *Ann. Statist.* **35** 1644–1673.

Cavalier, L. (2004). Estimation in a problem of fractional integration. *Inverse Problems* **20** 1445–1454.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

Donoho, D. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2** 101–126.

Donoho, D. and Johnstone, I. (1999). Asymptotic minimaxity of wavelet estimators with sampled data. *Statistica Sinica* **9** 1–32.

Johnstone, I. M. (1999). Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sinica* **9** 51–83.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B* **59** 319–351.

Kullback, S. (1967). A lower bound for discrimination in terms of variation. *IEEE Trans. Information Theory* **13** 126–127.

Le Cam, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35** 1419–1455.

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

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer–Verlag, New York.

Mallat, S. G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Math. Soc.* **315** 69–87.

McCoy, E. J. and Walden, A. T. (1996). Wavelet analysis and synthesis of stationary long-memory processes. *Journal of Computational and Graphical Statistics* **5** 26–56.

Meyer, Y. (1990). *Ondelettes et opérateurs. I.* Actualités Mathématiques. [Current Mathematical Topics]. Hermann, Paris. Ondelettes. [Wavelets].

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

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission* **16** 120–133.

Reiss, M. (2008). Asymptotic equivalence for nonparametric regression with multivariate and random design. *Annals of Statistics* **36** 1957–1982.

Rohde, A. (2004). On the asymptotic equivalence and rate of convergence of nonparametric regression and Gaussian white noise. *Statistics & Decisions* **22** 235–243.

Stoev, S., Pipiras, V. and Taqqu, M. S. (2002). Estimation of the self-similarity parameter in linear fractional stable motion. *Signal Process.* **82**

1873–1901.

Veitch, D. and Abry, P. (1999). A wavelet-based joint estimator of the parameters of long-range dependence. *IEEE Trans. Inform. Theory* **45** 878–897.

Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Ann. Statist.* **24** 466–484.

Wornell, G. W. (1990). A Karhunen–Loève-like expansion for $1/f$ processes via wavelets. *IEEE Transactions on information theory* **36** 859–861.

Zhang, J. and Waiter, G. (1994). A wavelet-based KL-like expansion for wide-sense stationary random processes. *Signal Processing, IEEE Transactions on* **42** 1737–1745.