Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions

Jerrod Ankenman [a], William Leeb [b,*]

[a] *Susquehanna International Group, Bala Cynwyd, PA 19004, United States*
[b] *PACM, Princeton University, Princeton, NJ 08544, United States*

ARTICLE INFO

Article history:
Received 13 January 2016
Received in revised form 6 December 2016
Accepted 23 January 2017
Available online xxxx
Communicated by Charles K. Chui

MSC:
62G08
42C40
26B35
05C05

Keywords:
Hölder
Mixed Hölder
Tensor product
Tree metric
Wavelet
Haar system
Wavelet shrinkage
Besov space
Calderón–Zygmund decomposition

ABSTRACT

This paper concerns two related problems in the analysis of data matrices whose rows and columns are equipped with tree metrics. First is the problem of recovering a matrix that has been corrupted by additive noise. Under the assumption that the clean matrix exhibits a specific regularity condition, known as the mixed Hölder condition, we adapt the well-known Donoho–Johnstone wavelet shrinkage methods from classical nonparametric statistics to obtain estimators that are within a logarithmic factor of the minimax error rate with respect to mean squared error loss.

The second part of this paper develops a theory of Besov spaces on products of tree geometries. We show that matrices with small Besov norm can be written as a sum of a mixed Hölder matrix and a matrix with small support. Such decompositions are known as Calderón–Zygmund decompositions and are of general interest in harmonic analysis. The decompositions we establish impose fewer conditions on the function with small support than previous decompositions of this type while maintaining the same guarantees on the mixed Hölder matrix. As such, they are applicable to a greater variety of matrices and should find use in many data organization problems. As part of our analysis, we provide characterizations of the underlying Besov spaces using wavelets and other multiscale difference operators that are analogous to those from the classical Euclidean theory.

1. Introduction

This paper is concerned with matrix decompositions of the following form: if $f(x, y)$ is a matrix, by which we mean a function on the product of two discrete sets X and Y, we seek to write $f = g + b$, where g is a "good" matrix satisfying a certain regularity condition known as the mixed Hölder condition that we describe in Section 2, and b is a "bad" matrix that is nevertheless under control in some way.

* Corresponding author.
E-mail addresses: mathofpoker@gmail.com (J. Ankenman), wleeb@math.princeton.edu (W. Leeb).

http://dx.doi.org/10.1016/j.acha.2017.01.003
1063-5203/

Such decompositions are encountered throughout analysis and its applications, such as in signal and image processing [1].

In Sections 1.1–1.6, we briefly introduce the high-level ideas used throughout this paper. In Section 1.7, we discuss the contributions of this paper.

1.1. Wavelets and multiresolution analysis

We give a brief summary of some relevant facts from wavelet theory. Of particular concern to us will be the notion of a *multiresolution analysis* of $L^2(\mathbb{R})$ [2–4]. One starts with a function $\phi(x)$, and considers all its dyadic dilates and integer translates, given by

$$\phi_{j,k}(x) = 2^{-j/2}\phi(2^{-j}x - k) \tag{1}$$

We define V_j as the linear span of the functions $\phi_{j,k}$ over all integers k. Under suitable conditions on ϕ, these spaces will be nested; that is, $V_j \subsetneq V_{j-1}$, or in other words, we can write $\phi(x)$ as a linear combination of the functions $\phi(2x - k)$; and their union is all of $L^2(\mathbb{R})$. In this case the system of subspaces V_j forms what is called a "multiresolution analysis" of $L^2(\mathbb{R})$, as each subspace captures activity at a certain dyadic scale, or resolution.

Wavelet analysis arises by looking at the orthogonal complement of V_j in V_{j-1}, which we denote by W_j. Given a multiresolution analysis as just described, one can construct a function $\psi(x)$ whose integer translates span W_0, and consequently where the functions $\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k)$ span W_j. The function ϕ is known as the "father wavelet", or "scaling function" and the function ψ is known as the "mother wavelet".

Perhaps the simplest example of such a system is the Haar system. Here, the father wavelet ϕ is the indicator function of the interval $[0, 1]$, and the mother wavelet is the function $\chi_{[0,1/2]} - \chi_{[1/2,1]}$. The space V_j is the span of indicator functions of dyadic intervals $[2^{-j}k, 2^{-j}(k+1)]$ for all integers k. It is very simple to generalize this particular multiresolution analysis to the setting of partition trees on abstract sets [5], as we will describe in more detail later.

1.2. The classical Besov spaces

Given a metric space (X, d), a natural way of measuring the variation of a function f defined on X is its Lipschitz norm, defined by

$$\sup_{x \neq y} \frac{f(x) - f(y)}{d(x, y)}. \tag{2}$$

If f is a differentiable function on $\mathbb{R}$, the Lipschitz norm (2) is equal to $\|f'\|_\infty$, the supremum of f's derivative. Expression (2), however, is defined for non-differentiable functions and makes sense in the abstract setting of any metric space.

A generalization of the Lipschitz norm is the Hölder norm, which replaces the metric $d(x, y)$ by $d(x, y)^\alpha$ for some parameter $\alpha > 0$. For functions on $\mathbb{R}$, this space is only non-trivial when $0 < \alpha \leq 1$. The space of Hölder functions when α is strictly less than 1 has nicer algebraic properties than the Lipschitz space; in particular, the Hölder norm of a function can be characterized by the size of its wavelet coefficients. If we take a sufficiently nice wavelet basis $\{\psi_{j,k}\}$ of $\mathbb{R}^n$ (where $j \in \mathbb{Z}$ indexes the dyadic scale 2^{-j} and $k \in \mathbb{Z}$ the location), then the expression

$$\sup_{j,k} 2^{j(\alpha+1/2)} |\langle f, \psi_{j,k} \rangle| \tag{3}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

is equivalent in size to (2), which is to say that the ratio of the two quantities is bounded above and below by finite constants not depending on f. The wavelet coefficients $\langle f, \psi_{j,k} \rangle$ can be thought of as measuring f's variation across scales. The corresponding formula for $\alpha = 1$ yields not the Lipschitz space but the Zygmund space [2].

There is a generalization of the Hölder spaces, known as the Besov spaces, that replace the L^∞ norms used to define the Hölder spaces by L^p norms. To see how this works, rewrite the Hölder norm of f as:

$$\sup_{t>0} t^{-\alpha} w_{t,\infty}(f) \tag{4}$$

where

$$w_{t,\infty}(f) = \sup_{|h| \leq t} \|f - f(\cdot - h)\|_\infty \tag{5}$$

is the L^∞ modulus of continuity of f. If we replace the L^∞ norm defining $w_{t,\infty}(f)$ with an L^p norm, and the supremum in t in (4) with an $L^s(dt/t)$ norm, we obtain the Besov norm:

$$\left(\int_0^\infty t^{-\alpha s} w_{t,p}(f)^s \frac{dt}{t} \right)^{1/s} \tag{6}$$

where

$$w_{t,p}(f) = \sup_{|h| \leq t} \|f - f(\cdot - h)\|_p \tag{7}$$

is the p-modulus of continuity of f. Like the Hölder norm, there are simple characterizations of this Besov norm using wavelets and other multiscale operators: for example, the quantity

$$\left(\sum_j 2^{j(\alpha + 1/2 - 1/p)s} \Big[\sum_k |\langle f, \psi_{j,k} \rangle|^p \Big]^{s/p} \right)^{1/s} \tag{8}$$

is equivalent to (6), in the sense that the ratio of the two is bounded above and below by constants independent of f.

Because the definition of the modulus of continuity $w_{t,p}$ makes use of the translation structure on Euclidean space, it is unclear how one might define this quantity on an abstract metric/measure space, such as the tree metrics we consider later in this paper. It follows from results in Chapter 7 of [6] that $w_{t,p}(f)$ is equivalent in size to the quantity

$$\omega_{t,p}(f) = \left(\int_{\mathbb{R}} \frac{1}{|B(x,t)|} \int_{B(x,t)} |f(x) - f(y)|^p dy dx \right)^{1/p}. \tag{9}$$

As the definition of $\omega_{t,p}(f)$ only makes use of the metric and measure on $\mathbb{R}$, it provides a way of extending the modulus of continuity to abstract metric/measure spaces.

Returning to Euclidean space, we note that the Besov spaces are also defined when $\alpha \geq 1$. For non-integer α, one simply replaces α by $\alpha - \lfloor \alpha \rfloor$ and f with its $\lfloor \alpha \rfloor^{th}$ derivative. Finally, we note that for technical reasons, in this paper we will restrict attention to the case when $p = s$; such two-parameter Besov spaces (the parameters being α and p) may be referred to in the literature as Slobodeckij spaces, Aronszajn spaces, Gagliardo spaces, Sobolev–Slobodeckij spaces, or other variations thereof. The first chapter of [7] contains a history of these spaces and other approaches to their construction.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

1.3. Mixed Besov spaces

The function spaces we discussed in Section 1.2 are defined with respect to a single metric space. Many domains arising in applications, however, are not modeled well by one metric space, but rather the product of several metric spaces. Such datasets arise naturally in a variety of applications. For example, in the theory of transposable arrays [8,9], both the rows and columns of a dataset are studied. Similarly, methods of co-clustering [10,11] search for a clustering of both the row and column sets of a data matrix, and so fits into the same framework.

For simplicity, we restrict attention to the product of two spaces, (X, d_X) and (Y, d_Y). The notion of regularity we consider for a function f defined on $X \times Y$ is the mixed Hölder condition, which requires f to have bounded mixed difference quotients; that is,

$$\sup_{x \neq x', y \neq y'} \frac{f(x,y) - f(x,y') - f(x',y) + f(x',y')}{d_X(x,x')^\alpha d_Y(y,y')^\alpha} \tag{10}$$

must be finite. We give a more formal definition of the mixed Hölder space in Section 2.

The expression (10) is unnatural in many applications in the Euclidean setting, depending as it does on the choice of x- and y-axes. In fact, it is not rotationally invariant. For simplicity, fix $\alpha = 1$ and observe that the expression (10) is equivalent in size to $\|\partial^2_{x,y} f\|_\infty$ for smooth functions on $\mathbb{R}^2$. Consider a function of the form $f(x,y) = g(x)h(y)$, for smooth g and h. Then

$$\partial^2_{xy} f(x,y) = g'(x)h'(y) \tag{11}$$

whereas

$$\partial^2_{xy} f\left(\frac{x+y}{\sqrt{2}}, \frac{x-y}{\sqrt{2}}\right) = \frac{1}{2} g''\left(\frac{x+y}{\sqrt{2}}\right) h\left(\frac{x-y}{\sqrt{2}}\right) - \frac{1}{2} g\left(\frac{x-y}{\sqrt{2}}\right) h''\left(\frac{x+y}{\sqrt{2}}\right) \tag{12}$$

and by constructing functions g and h with large zeroth and second derivatives but small first derivatives, one sees that the size of the quantity (10) depends on the coordinate system, a limitation in physical settings that exhibit rotational invariance and hence where the axes are essentially arbitrary.

By contrast, in many data-analysis problems the axes are not arbitrary, but rather intrinsic to the problem itself; consider, for example, the word/document axes of a word-document database [12], or the time/frequency axes of a spectrogram [13,14]. For such problems it is reasonable to look at norms, like the mixed Hölder norm, that depend on the choice of axes; indeed, such norms make the most sense in this context.

We can define mixed Besov spaces in much the same way as for a single space; we replace the L^∞ norms implicit in the definition of the mixed Hölder norm by an $L^p(dx)$ norm in space and an $L^s(dt/t)$ norm across scales. This gives the norm:

$$\left(\int_0^\infty \int_0^\infty w_{t,t',p}(f)^s \frac{dt}{t} \frac{dt'}{t'} \right)^{1/s} \tag{13}$$

where the modulus of continuity $w_{t,t',p}$ is given by

$$w_{t,t',p}(f) = \sup_{|h| \leq t, |h'| \leq t'} \left(\int_{\mathbb{R}} \int_{\mathbb{R}} |f(x,y) - f(x-h,y) - f(x,y-h') + f(x-h,y-h')|^p dx dy \right)^{1/p} \tag{14}$$

Such spaces have been well-studied and given characterizations similar to their one-dimensional versions. If f is defined on, say, $[0,1] \times [0,1]$ and has bounded mixed difference quotients, many favorable properties

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

follow. For instance, such functions can be reconstructed to precision ϵ using only $\mathcal{O}((1/\epsilon)\log(1/\epsilon))$ samples; these samples form what is known as a "sparse grid" [15,16]. This compares favorably with the $\Omega(1/\epsilon^2)$ points that would be needed if f were only known to be Lipschitz. Similarly, only $\mathcal{O}((1/\epsilon)\log(1/\epsilon))$ coefficients from a suitable wavelet basis are needed to reconstruct such a function f to precision ϵ [17]. Of particular relevance to the present work, statistical estimators of such an f from noisy samples achieve higher minimax rates than estimators of functions that are merely assumed to be Lipschitz [13,14]; this terminology will be explained in Section 1.4.

1.4. Statistical estimation and wavelet shrinkage

Statistical estimation is concerned with recovering a function f from noisy samples of the form f + noise. To have any hope of recovering the signal, we must impose some assumptions on it. Parametric models assume that the signal lies in a family defined by a finite number of parameters – for instance, a low-degree polynomial; the task is to estimate the parameters of this model. Nonparametric models impose weaker assumption on the function f, assuming only that it lies in, for example, a Besov ball (or a relative thereof, such as a Triebel or Sobolev ball).

For any estimation problem, a standard way of measuring the quality of an estimator is the minimax criterion. Here, we look at the estimator's expected loss; that is, if the true function without noise is f and our estimator is $\hat{f}$, we pay a price $L(f,\hat{f})$. A standard loss is the squared Euclidean distance $\|f-\hat{f}\|_2^2$ between the true function and our estimate.

The observed signal, and hence any estimator and its loss, are random quantities. We therefore consider the expected loss, or *risk*, of an estimator, which depends on the unknown true function f:

$$R(f,\hat{f}) = \mathbb{E}_f L(f,\hat{f}) \tag{15}$$

If the true function f lies in a family $\mathcal{F}$, then the worst performance of an estimator $\hat{f}$ is its maximum risk over $\mathcal{F}$, or $\sup_{f\in\mathcal{F}} R(f,\hat{f})$. The minimax criterion for estimation seeks an estimator $\hat{f}^*$ that minimizes this maximum risk; that is,

$$\hat{f}^* = \operatorname*{argmin}_{\hat{f}} \sup_{f\in\mathcal{F}} R(f,\hat{f}) \tag{16}$$

In this paper, we will be concerned only with how the quantity $\min_{\hat{f}} \sup_{f\in\mathcal{F}} R(f,\hat{f})$ depends on the problem size (the number of observations) and the problem parameters (the parameters defining the space $\mathcal{F}$). Consequently, rather than seek estimators that achieve the exact minimax risk, we relax this problem to find estimators that differ from the minimax risk by at most a multiplicative constant C that does not depend on the number of samples or the relevant parameters. We will say that such an estimator "achieves the minimax rate" of the problem.

When the parameters of the Besov space are completely specified, explicit estimators can be derived that achieve the minimax rate; these estimators are also linear functions of the observed data, and amount to applying a low-pass filter to get rid of the noise. However, in many cases it is not reasonable to assume that the parameters of the Besov space – which, after all, is often only a heuristic model – are specified. It is therefore desirable to have estimators that perform well for a range of Besov spaces – that is, for a range of parameters.

The wavelet shrinkage estimators of Donoho and Johnstone succeed at this task [18–22]. These are defined by shrinking the wavelet coefficients of the observations towards zero. The shrinkage operator is a non-linear function of one variable, given by

$$\eta_t(x) = \operatorname{sgn}(x)(|x|-t)_+. \tag{17}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

Donoho and Johnstone prove that the wavelet shrinkage estimator is nearly minimax for large ranges of Besov spaces simultaneously. More precisely, the error rate it achieves equals the minimax rate times a term $\log(n)^r$, where n is the number of samples and $r < 1$ is a parameter depending on the space $\mathcal{F}$.

1.5. Calderón–Zygmund decompositions

A Calderón–Zygmund decomposition breaks a function f into a sum of two functions $g + b$, where g is well-behaved (for instance, is not too oscillatory) and b, while it may be highly irregular, has small support. In the classical Calderón–Zygmund decomposition, f is assumed to be in L^1, and g then has small L^2 norm, while b is highly oscillatory but is supported on a small set and has mean zero. See, for instance, [23,24,4]. This decomposition is a critical ingredient in proving classical results such as the Markinciewicz Interpolation Theorem [23,24,4,1]. It is also highly useful in the study of certain operators [1]. For instance, it is employed in the proof that the important class of Calderón-Zygmund operators map L^1 to weak L^1 [25].

Another class of Calderón–Zygmund decompositions impose conditions not on f itself but rather its gradient – for example, ∇f might be assumed to lie in L^p. The function g is then constructed to satisfy a stronger condition on its gradient, such as $|\nabla g| \in L^\infty$. See, for instance, [26,27] for results along these lines.

A theorem of this kind was shown in [28] in the context of tree metrics, which are the kind of metrics we consider in this paper and which we will discuss in Section 1.6. The function f is assumed to lie in a particular mixed Besov space, and the good function g can be taken to be mixed Hölder. The decomposition of f into $g + b$ depends on a wavelet expansion of f; certain wavelet coefficients are grouped to form g, and the remaining ones form b.

1.6. Tree metrics

The spaces discussed in Sections 1.2 and 1.3 were defined for functions on Euclidean space. The present paper is concerned, however, with an analogous theory for a different kind of metric space, where the Euclidean metric is replaced by an abstract tree metric. Tree metrics are defined by breaking the space into a collection of folders, and placing a weight on each folder that defines its diameter. In this paper, the weights will equal the folder's volume.

Because of their simple structure, tree metrics appear throughout pure and applied mathematics. There are numerous applications in computer science, where certain metric tasks are very simple to perform for tree metrics; by approximating an arbitrary metric by a family of tree metrics, one obtains fast, approximate solutions [29–31].

In [28,5] it is shown that the same wavelet characterizations of the Hölder and mixed Hölder spaces of functions on Euclidean space can be adapted to the Hölder and mixed Hölder spaces on trees and products of trees. The multiple partitions induced by the tree parallel the multiresolution analyses discussed in Section 1.1, and the classical wavelets are replaced by a family of Haar wavelets defined with respect to the partition trees; we will discuss their definition in more detail in Section 2.

In [28], it is shown that many of the properties of mixed Hölder functions on $[0, 1] \times [0, 1]$ also hold for mixed Hölder functions on the product of tree metric spaces. For instance, it is possible to compress a mixed Hölder function to high precision using a small number of its tensor Haar coefficients, and one can define the same notion of sparse grid for reconstruction as in the Euclidean case.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

1.7. The contributions of this paper

The present work further develops the theory of harmonic analysis on partition trees and their products. As stated earlier, we focus on the task of writing a function $f(x, y)$ in the form $f = g + b$, where g has a small mixed Hölder norm and b is controlled in one of two ways.

The first such decomposition tries to make b look as unstructured as possible – that is, we want b to look like noise. The methods we develop in Sections 3 and 4 emerge from the statistical estimation theory discussed in Section 1.4. More precisely, in Section 3 we give a formal statement of the estimation problem in the presence of Gaussian noise, and in Section 3.1, specifically Theorem 2, we establish its minimax rate. In Section 4, we then show that the wavelet shrinkage estimators of Donoho and Johnstone, applied to the tensor Haar basis on products of trees, come within a logarithmic factor of achieving this rate, which is to be expected from the classical theory. In Section 4.3, we also show that the denoised function g will be smooth with high probability, and in Section 4.4 we provide approximation guarantees when multiple estimators are averaged over many trees whose metrics approximate another (non-tree) metric.

The second model for the bad function b, developed in Section 5, forces b to have small support; in other words, we establish a family of Calderón–Zygmund decompositions as discussed in Section 1.5. As a precursor to these decompositions, in Section 5.1 we develop analogues of the classical Besov spaces for tree metrics. In particular, we define a natural modulus of continuity for functions on tree metric spaces, define the analogues of the Besov norm (13) (when $p = s$), and give simple characterizations of the Besov norm of a function using its coefficients in various expansions.

From these simple characterizations of the Besov norms naturally emerges a broad class of Calderón–Zygmund decompositions. Theorem 6 generalizes the Haar-based decomposition from [28] to the Besov spaces introduced in Section 5.1. Theorem 7 presents an entirely new decomposition that does not depend on the function f's Haar series, but rather on a minimal-cost expansion in an overdetermined set of simple functions. This decomposition imposes fewer constraints on the function b, and is able to recover the natural decomposition for a much broader range of functions. We illustrate the advantages of this decomposition over the Haar-based decomposition on numerical examples.

Appendix A contains basic results in the metric geometry of partition trees and estimation of mixed Hölder functions on tree metrics. In some cases, these results have not been stated previously in the literature, while in other cases they are sharper versions of estimates that have appeared in [28]. These estimates can also be used to improve results from [28] that are not used here, such as the sparse grid construction.

2. Preliminaries

Throughout this paper, X and Y will denote two spaces with n_X and n_Y points, respectively, and we let $n = n_X \cdot n_Y$ denote the number of points in $X \times Y$. We will think of X and Y as, respectively, the set of rows and columns of a matrix. We will equip X and Y with finite measures, and assume for simplicity (and without loss of generality) that the total mass of each is 1. We denote by $|S|$ the measure of a set, and $\#S$ the number of elements it contains.

As a matter of notation, we will often use the letter C to denote an arbitrary constant. When stating results we will sometimes write expressions like $C = C(a, b, c, d)$ to specify that C depends on the parameters a, b, c, d. In long strings of inequalities, the value of C may change from line to line.

2.1. Tree metrics, Hölder functions and mixed Hölder functions

We assume throughout this paper that both X and Y are equipped with partition trees, denoted by $\mathcal{T}_X$ and $\mathcal{T}_Y$. A partition tree on X (or Y) is a collection of subsets of X, called "folders", that include X itself

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

and all singletons $\{x\}$, and with the property that for any two folders I and I', either $I \subset I'$, $I' \subset I$, or I and I' are disjoint.

Associated to any partition tree on X (or Y) is a corresponding volume-based tree metric. If $I_{x,x'}$ denotes the smallest folder in $\mathcal{T}_X$ containing both x and x', then the tree distance $d_X(x,x')$ equals

$$d_X(x,x') = \begin{cases} |I_{x,x'}|, \text{ if } x \neq x' \\ 0, \text{ if } x = x' \end{cases} \tag{18}$$

Given a folder I, we will refer to the smallest folder containing but not equal to I as I's "parent". If I' is I's parent, we will also say that I is a "child" of I'.

As in [32,28,5], the critical assumption we will impose on the partition tree is that it is balanced, in the sense that there are constants B_L and B_U such that

$$0 < B_L \leq \frac{|child|}{|parent|} \leq B_U < 1. \tag{19}$$

We will denote by $\mathcal{T}_Y$ a partition tree on Y, and assume the same balance condition (19) on Y's folders. We will denote the tree distance on Y by $d_Y(y,y')$.

For any folder $I \in \mathcal{T}_X$, define

$$(m_I f)(y) = \frac{1}{|I|} \int_I f(x,y)dx \tag{20}$$

and define $(m_J f)(x)$ similarly for $J \in \mathcal{T}_Y$.

Given any function f on $X \times Y$ and any $\alpha > 0$, we say that f has mixed Hölder(α) norm $L = L(f,\alpha)$ if the maximum of the terms

$$\sup_{x \neq x', y \neq y'} \frac{f(x,y) - f(x,y') - f(x',y) + f(x',y')}{d_X(x,x')^\alpha d_Y(y,y')^\alpha} \tag{21}$$

$$\sup_{x \neq x'} \frac{(m_Y f)(x) - (m_Y f)(x')}{d_X(x,x')^\alpha} \tag{22}$$

$$\sup_{y \neq y'} \frac{(m_X f)(y) - (m_X f)(y')}{d_Y(y,y')^\alpha} \tag{23}$$

is bounded above by L. We will refer to the quantity (21) alone as the *mixed variation* of f, and denote it $M(f,\alpha)$.

2.2. Haar systems on trees and tensor products

Given a set X with tree $\mathcal{T}_X$, we can build functions defined on X that mimic the classical Haar system. These functions are described in [32,28,5], and we will only review the basic properties here. Each Haar function ϕ is supported on a single non-singleton folder $I \in \mathcal{T}_X$, and is constant on the children of I. Furthermore, ϕ has mean zero, is normalized to have L^2 norm 1, and is orthogonal to every other Haar function. Consequently, the collection of all Haar functions and the constant function 1 on the space form an orthonormal basis for the space of all functions on X.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

Given Haar systems $\{\phi\}$ on X and $\{\psi\}$ on Y, we can consider their tensor product; this is the collection of all functions $\Phi(x,y)$ of the form $\phi(x)\psi(y)$, $\phi(x)$, and $\psi(y)$. These functions form an orthonormal basis for the collection of all functions with mean zero on $X \times Y$.

Given a Haar function $\phi(x)$ on X, we will use $I(\phi)$ to denote the smallest folder containing the support of ϕ; similarly, $J(\psi)$ will denote the smallest folder containing the support of a Haar function ψ on Y. Given a tensor Haar function $\Phi(x,y) = \phi(x)\psi(y)$ on $X \times Y$, we will denote by $R(\Phi) = I(\phi) \times J(\psi)$ the smallest rectangle containing the support of Φ.

The following result, which gives a characterization of the mixed Hölder(α) norm of a function based on its tensor Haar coefficients $\langle f, \Phi \rangle$, is key:

Theorem 1. *For any $\alpha > 0$, there is a constant $C = C(B_L, B_U, \alpha) > 1$ such that for any function f on $X \times Y$,*

$$\frac{1}{C} L(f,\alpha) \leq \sup_{\Phi} \frac{|\langle f, \Phi \rangle|}{|R(\Phi)|^{\alpha+1/2}} \leq L(f,\alpha) \tag{24}$$

where the supremum is over all Haar functions $\Phi(x,y)$.

The proof is essentially contained in [28].

3. Matrix denoising

In this section, we assume that X and Y are equipped with normalized counting measure; in other words, the measure of any single point $x \in X$ is n_X^{-1}, and the measure of any single point $y \in Y$ is n_Y^{-1}. Recall that given any set S, we will always denote by $\#S$ the number of elements of S. If $S \subset X$, we will write its measure as $|S|$. To illustrate the notation, observe that for any $S \subset X$, since $|S|$ denotes the measure of a set and $\#S$ denotes the number of elements, $|S| = (\#S)/n_X$; similarly, if $S \subset X \times Y$, $|S| = (\#S)/(n_X n_Y) = (\#S)/n$.

We suppose that we have a function f defined on $X \times Y$, and that f has mixed Hölder(α) norm not exceeding L for some $\alpha > 0$ and $L > 0$. Let $\mathcal{F} = \mathcal{F}(\alpha, L)$ denote the collection of all such functions. We do not observe f directly, however, but rather

$$T(x,y) = f(x,y) + \eta(x,y) \tag{25}$$

where $\eta(x,y)$ is a random variable. We will assume that the $\eta(x,y)$'s are independent, have mean zero, and have maximum variance $\nu < \infty$.

Our goal is to estimate f from the noisy samples T, taking as our primary measure of the loss of an estimator $\hat{f}_n$ the mean squared loss, or $\|\hat{f}_n - f\|_2^2$. As explained in Section 1.4, we are concerned with the minimax rate $\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\hat{f}_n - f\|_2^2$, where the infimum is over all estimators.

In Section 3.1, we will give an explicit estimator $\hat{f}_n$ for the function f, and derive bounds on the expected error of this estimator in terms of the parameters ν, α and L and the problem size, n. We will show in Section 3.1.3 that this estimator achieves the minimax rate in terms of its dependence on L, ν and n when the noise η is Gaussian.

The estimator $\hat{f}_n$ we define in Section 3.1 depends on specification of L and α, which might be unnatural in practice, and the bounds we obtain are only for the L and α specified. In Section 4, we adapt the wavelet shrinkage estimators of Donoho and Johnstone [19] to define estimators that are nearly minimax for all $\alpha > 0$ simultaneously when the noise is Gaussian.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

3.1. Optimal matrix denoising

We now consider the problem of recovering a mixed Hölder function f that has been corrupted by noise. We assume that we are given two spaces X and Y with, respectively, n_X and n_Y points, and partition trees $\mathcal{T}_X$ and $\mathcal{T}_Y$ whose folders satisfy (19).

We can think of X as the rows of a matrix and Y as the columns. A matrix itself will be described by a function f on $X \times Y$. We suppose that f has mixed Hölder(α) norm not exceeding L, which by Theorem 1 implies that the size each Haar coefficients $\langle f, \Phi \rangle$ is bounded by $L|R(\Phi)|^{\alpha+1/2}$.

Before turning to the wavelet shrinkage estimator, however, we will first analyze how well we can do if we are willing to specify the parameters L and α, in the regime $n \to \infty$. We will write down an explicit estimator and estimate its expected error. In Section 3.1.3, specifically Theorem 2, we will prove that this estimator's mean squared error is optimal in its dependence on $n = n_X n_Y$, L and ν for all n sufficiently large.

Define $\epsilon > 0$ by

$$\epsilon = \min \left\{ 1, \left(\frac{\nu}{nL^2} \right)^{1/(2\alpha+1)} \right\}. \tag{26}$$

Suppose for now that $\epsilon < 1$. We will consider the case where $\epsilon = 1$ separately. Define the estimator

$$\hat{f}_n(x,y) = \sum_{\Phi : |R(\Phi)| \geq \epsilon} \langle T, \Phi \rangle \Phi(x,y) \tag{27}$$

where Φ runs over all Haar functions and the constant function 1. We will bound the expected squared error of $\hat{f}_n$, both pointwise and in L^2.

3.1.1. Mean squared error of $\hat{f}_n$

Define the deterministic function

$$g(x,y) = \sum_{\Phi : |R(\Phi)| \geq \epsilon} \langle f, \Phi \rangle \Phi(x,y). \tag{28}$$

We have $\|\hat{f}_n - f\|_2^2 = \|\hat{f}_n - g\|_2^2 + \|g - f\|_2^2 + 2\langle \hat{f}_n - g, g - f \rangle$. Observe that $\mathbb{E}_f T(x,y) = f(x,y)$; consequently, by linearity we have $\mathbb{E}_f \hat{f}_n = g$ and in particular $\mathbb{E}_f \langle \hat{f}_n - g, g - f \rangle = \langle \mathbb{E}_f \hat{f}_n - g, g - f \rangle = 0$. Therefore, we establish the bias-variance decomposition of the expected error

$$\mathbb{E}_f \|\hat{f}_n - f\|_2^2 = \mathbb{E}_f \|\hat{f}_n - g\|_2^2 + \|g - f\|_2^2. \tag{29}$$

Now, by Corollary 3 we can control the bias term:

$$\|g - f\|_2^2 \leq C L^2 \epsilon^{2\alpha} (\log_{B_U^{-1}}(1/\epsilon) + 1), \tag{30}$$

where $C = C(B_L, B_U, \alpha)$ is a constant. We will now derive an upper bound for $\mathbb{E}_f \|\hat{f}_n - g\|_2^2$. Observe that

$$\mathbb{E}_f \|\hat{f}_n - g\|_2^2 = \sum_{\Phi : |R(\Phi)| \geq \epsilon} \mathbb{E}_f \langle f - T, \Phi \rangle^2 = \sum_{\Phi : |R(\Phi)| \geq \epsilon} \mathbb{E}_f \langle \eta, \Phi \rangle^2. \tag{31}$$

To compute an upper bound on $\mathbb{E}_f \langle \eta, \Phi \rangle^2$, observe that $\mathbb{E}_f \langle \eta, \Phi \rangle = 0$; consequently,

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\begin{aligned}\mathbb{E}_f\langle\eta,\Phi\rangle^2 &= \mathrm{Var}(\langle\eta,\Phi\rangle) = \mathrm{Var}\left\{\frac{1}{n}\sum_{x,y}\eta(x,y)\Phi(x,y)\right\}\\ &= \frac{1}{n^2}\sum_{x,y}\mathrm{Var}(\eta(x,y))\Phi(x,y)^2 \le \frac{\nu}{n}\|\Phi\|_2^2 = \frac{\nu}{n}.\end{aligned} \tag{32}$$

Combining (31), (32) and Corollary 4 gives

$$\begin{aligned}\mathbb{E}_f\|\hat{f}_n - g\|_2^2 &= \sum_{\Phi:|R(\Phi)|\ge\epsilon}\mathbb{E}_f\langle\eta,\Phi\rangle^2 \le \frac{1}{B_L}\frac{\nu}{n}\#\{R : |R(\Phi)| \ge \epsilon\}\\ &\le \frac{1}{B_L(1-B_U)}\frac{\nu}{n}\left(\frac{\log_{B_U^{-1}}(1/\epsilon)+1}{\epsilon}\right).\end{aligned} \tag{33}$$

Since $\epsilon = (\nu/(nL^2))^{1/(2\alpha+1)}$, combining (29), (30), and (33) gives

$$\sup_{f\in\mathcal{F}}\mathbb{E}_f\|\hat{f}_n - f\|_2^2 \le C\nu^{2\alpha/(2\alpha+1)}L^{2/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}\log_{B_U^{-1}}(L^2n/\nu) \tag{34}$$

where $C = C(B_L, B_U, \alpha)$ is a constant. We show in Section 3.1.3 that, when n is sufficiently big, no other estimator of f can outperform this one in terms of its dependence on L, ν and n.

Finally, note that if $\epsilon = 1$, we can combine (29), (30), and (33) to obtain the estimate $\sup_{f\in\mathcal{F}}\mathbb{E}_f\|\hat{f}_n - f\|_2^2 \le C\nu/n$.

3.1.2. Pointwise squared error of $\hat{f}_n$

We will now derive an upper bound on the pointwise squared error of the estimator $\hat{f}_n$. That is, for any point $(x_0, y_0) \in X \times Y$, we will bound $\mathbb{E}_f\{(f(x_0,y_0) - \hat{f}_n(x_0,y_0))^2\}$. Note that we only need to estimate the size of wavelet coefficients for wavelets containing (x_0, y_0); in particular, we only require that f be mixed Hölder(α) at (x_0, y_0).

Suppose first that $\epsilon < 1$. Take the function g defined by (28). Observe that, since $\mathbb{E}_f\hat{f}_n(x_0,y_0) = g(x_0,y_0)$, we can write

$$\begin{aligned}&\mathbb{E}_f\{(f(x_0,y_0) - \hat{f}_n(x_0,y_0))^2\}\\ &\quad= \mathbb{E}_f(\hat{f}_n(x_0,y_0) - g(x_0,y_0))^2 + \{(f(x_0,y_0) - g(x_0,y_0))^2\}.\end{aligned} \tag{35}$$

From Corollary 6, we have the bound

$$(f(x_0,y_0) - g(x_0,y_0))^2 \le CL^2\epsilon^{2\alpha}(\log_{B_U^{-1}}(1/\epsilon)+1)^2. \tag{36}$$

As for the second term on the right side of (35), we have (with the sum over all wavelets Φ with $(x_0,y_0) \in R(\Phi)$ and $|R(\Phi)| \ge \epsilon$)

$$\begin{aligned}\mathbb{E}_f(\hat{f}_n(x_0,y_0) - g(x_0,y_0))^2 &= \mathbb{E}_f\left\{\sum_\Phi\langle\eta,\Phi\rangle\Phi(x_0,y_0)\right\}^2\\ &\le C\left\{\sum_\Phi\sqrt{\mathbb{E}_f\langle\eta,\Phi\rangle^2|R(\Phi)|^{-1}}\right\}^2\\ &\le C\frac{\nu}{n}\left\{\sum_\Phi|R(\Phi)|^{-1/2}\right\}^2 \le C\frac{\nu}{n}\frac{(\log_{B_U^{-1}}(1/\epsilon)+1)^2}{\epsilon}\end{aligned} \tag{37}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

where we have used Corollary 5. Substituting the value $\epsilon = (\nu/(nL^2))^{1/(2\alpha+1)}$ and combining (35), (36) and (37) gives

$$\mathbb{E}_f\{(\hat{f}_n(x_0,y_0) - f(x_0,y_0))^2\} \leq C\nu^{2\alpha/(2\alpha+1)}L^{2/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}\log^2_{B_U^{-1}}(L^2n/\nu), \tag{38}$$

where $C = C(B_L, B_U, \alpha)$ is a constant. Note that the pointwise bound is worse by an extra factor of $\log_{B_U^{-1}}(L^2n/\nu)$ when compared to the mean squared estimate (34).

Finally, observe that if $\epsilon = 1$, (35), (36) and (37) give $\sup_{f\in\mathcal{F}}\mathbb{E}_f\{(\hat{f}_n(x_0,y_0) - f(x_0,y_0))^2\} \leq C\nu/n$.

3.1.3. The minimax lower bound

In this section, we will show that under the Gaussian noise model with variance $\nu = \sigma^2$ the mean squared error of the estimator $\hat{f}_n$ from Section 3.1 cannot be improved in its dependence on L, ν and n, as given by the bound (34) whenever n is sufficiently big, the parameters $L, \sigma^2, B_L, B_U, C_L, C_U$ are assumed fixed, and n_X and n_Y are of comparable size: that is, we assume that there are positive constants C_L and C_U such that

$$C_L \leq \frac{n_X}{n_Y} \leq C_U. \tag{39}$$

Theorem 2. *Let X and Y be finite sets with, respectively, n_X and n_Y points satisfying (39) and partition trees $\mathcal{T}_X$ and $\mathcal{T}_Y$ satisfying (19). Equip X and Y with normalized counting measure.*

Fix $\alpha > 0$, and let $\mathcal{F} = \mathcal{F}(\alpha, L)$ denote the set of functions with mixed Hölder(α) norm not exceeding L. Suppose we observe a single draw from each of a collection of independent random variables $T(x,y) \sim N(f(x,y), \sigma^2)$, $(x,y) \in X \times Y$, where σ^2 is a known variance and $f \in \mathcal{F}$ is an unknown function. Then there is an $N = N(B_L, B_U, \alpha, L, \sigma^2, C_L, C_U)$ constant $C = C(B_L, B_U, \alpha, C_L, C_U) > 0$ such that for all $n \geq N$,

$$\inf_{\hat{f}_n}\sup_{f\in\mathcal{F}}\mathbb{E}_f\|\hat{f}_n - f\|_2^2 \geq C\sigma^{4\alpha/(2\alpha+1)}L^{2/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}\log_{B_U^{-1}}(L^2n/\sigma^2). \tag{40}$$

The infimum is over all estimators of f.

The proof we give is adapted from proofs of classical minimax lower bounds (see, for instance, [33,34]), and rests on the following lemma from Chapter 2 of [34].

Lemma 1. *Suppose $f_0, \ldots, f_M \in \mathcal{F}$ satisfy*

1. *$\|f_j - f_{j'}\|_2 \geq 2s_n > 0$ for all $j \neq j'$;*
2. *The likelihood ratios of the distribution under f_0 and the distribution under f_j are of the form*

$$\Lambda(f_0, f_j) = \frac{d\mathbb{P}_{f_0}}{d\mathbb{P}_{f_j}} = \exp\{\Delta_j - \lambda_j \ln M\}$$

where $0 < \lambda_j < \lambda < 1$ and the random variables Δ_j are positive with probability bounded away from 0 under the model f_j, i.e. $\mathbb{P}_{f_j}(\Delta_j \geq 0) \geq p > 0$. Here, Δ_j is a random variable that may depend on n, and λ_j is a non-random number that may depend on n. The non-random constants p and λ are independent of n and j.

Then for any estimator $\hat{f}_n$,

$$\max_{1\leq j\leq M}\mathbb{P}_{f_j}(\|\hat{f}_n - f_j\|_2 \geq s_n) \geq p/2.$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

The strategy of the proof is as follows. We will construct a collection of functions $f_0, f_1, \ldots, f_M \in \mathcal{F}$ satisfying the conditions of Lemma 1 when n is sufficiently large with

$$s_n^2 = C\sigma^{4\alpha/(2\alpha+1)} L^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)} \log_{B_U^{-1}}(L^2 n/\sigma^2) \tag{41}$$

where $C = C(B_L, B_U, \alpha, C_L, C_U) > 0$ is a constant, and where $p = 1/2$. Chebyshev's inequality tells us that

$$\begin{aligned} \sup_{f\in\mathcal{F}} \mathbb{E}_f(s_n^{-2}\|\hat{f}_n - f\|_2^2) &\geq \sup_{f\in\mathcal{F}} \mathbb{P}_f(\|\hat{f}_n - f\|_2 \geq s_n) \\ &\geq \max_{1\leq j\leq M} \mathbb{P}_{f_j}(\|\hat{f}_n - f_j\|_2 \geq s_n) \end{aligned} \tag{42}$$

Consequently,

$$\begin{aligned} \sup_{f\in\mathcal{F}} \mathbb{E}_f(\|\hat{f}_n - f\|_2^2) &\geq s_n^2 \max_{1\leq j\leq M} \mathbb{P}_{f_j}(\|\hat{f}_n - f_j\|_2 \geq s_n) \\ &\geq \frac{p}{2} C\sigma^{4\alpha/(2\alpha+1)} L^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)} \log_{B_U^{-1}}(L^2 n/\sigma^2) \end{aligned} \tag{43}$$

which is the desired result.

We now turn to the construction of the functions $f_0, f_1, \ldots, f_M$. Let $\epsilon = (\sigma^2/(nL^2))^{1/(2\alpha+1)}$. Let $\mathcal{R}$ denote the collection of rectangles R with area $\epsilon B_L \leq |R| \leq \epsilon/B_L$, and let N_ϵ denote the number of such rectangles.

For each rectangle $R \in \mathcal{R}$, pick any wavelet $\Phi_R(x, y)$ supported on R. Define

$$f_R(x, y) = \delta L \epsilon^{\alpha+1/2} \Phi_R(x, y) \tag{44}$$

where $\delta > 0$ is a small constant to be chosen later. Since $\|\Phi_R\|_2 = 1$, $\|f_R\|_2^2 = \delta^2 L^2 \epsilon^{2\alpha+1}$. The functions f_R are the basic building blocks of the functions f_j, as we now describe.

The Varshamov–Gilbert bound states that there at least $M \equiv \lfloor 2^{N_\epsilon/8} \rfloor$ binary vectors of length N_ϵ whose pairwise Hamming distance ρ_H exceeds $N_\epsilon/16$; see [34] for a proof. For each such vector $\omega_j = (\omega_{j,1}, \ldots, \omega_{j,N_\epsilon})$, define the function

$$f_j(x, y) = \sum_{i=1}^{N_\epsilon} \omega_{j,i} f_{R_i}(x, y) \tag{45}$$

where $R_1, \ldots, R_{N_\epsilon}$ is some fixed but otherwise arbitrary ordering of the rectangles in $\mathcal{R}$. Observe that the magnitude of the wavelet coefficients $\langle f_j, \Phi_R \rangle$ are either 0 (if $R \notin \mathcal{R}$) or of size

$$\delta L \epsilon^{\alpha+1/2} \leq \delta L |R|^{\alpha+1/2} / B_L^{\alpha+1/2}. \tag{46}$$

By Theorem 1, by picking $\delta = \delta(B_L, B_U, \alpha)$ sufficiently small, we can guarantee that f_j has mixed Hölder(α) norm not exceeding L.

Since the functions f_{R_i} are pairwise orthogonal, the lower bound on N_ϵ provided by Lemma 8 yields, for all n sufficiently large,

$$\begin{aligned} \|f_j - f_{j'}\|_2^2 &= \sum_{i=1}^{N_\epsilon} (\omega_{j,i} - \omega_{j',i})^2 \|f_{R_i}\|_2^2 = \delta^2 L^2 \epsilon^{2\alpha+1} \rho_H(\omega_j, \omega_{j'}) \\ &\geq \frac{N_\epsilon}{16} \delta^2 L^2 \epsilon^{2\alpha+1} \geq C\delta^2 L^2 \epsilon^{2\alpha} \log(1/\epsilon) \\ &= C\sigma^{4\alpha/(2\alpha+1)} L^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)} \log_{B_U^{-1}}(L^2 n/\sigma^2) \end{aligned} \tag{47}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

where in the last line we have substituted $\epsilon = (\sigma^2/(nL^2))^{1/(2\alpha+1)}$ and where $C = C(B_L, C_L, C_U, \alpha)$ is a constant.

Now, let $f_0 \equiv 0$. Then the likelihood ratio $\Lambda(f_0, f_j)$ can be explicitly computed:

$$\begin{aligned}\Lambda(f_0, f_j) &= \frac{\prod_{(x,y)} \exp\{-(2\sigma^2)^{-1} T(x,y)^2\}}{\prod_{(x,y)} \exp\{-(2\sigma^2)^{-1}(T(x,y) - f_j(x,y))^2\}} \\ &= \prod_{(x,y)} \exp\{-(2\sigma^2)^{-1}(2T(x,y)f_j(x,y) - f_j(x,y)^2)\} \\ &= \exp\left\{\frac{1}{2\sigma^2}\sum_{(x,y)} f_j(x,y)^2 - \frac{1}{\sigma^2}\sum_{(x,y)} T(x,y)f_j(x,y)\right\} \\ &= \exp\left\{\frac{1}{\sigma^2}\sum_{(x,y)}(f_j(x,y) - T(x,y))f_j(x,y) - \frac{1}{2\sigma^2}n\|f_j\|_2^2\right\}\end{aligned} \tag{48}$$

Let $\Delta_j = \frac{1}{\sigma^2}\sum_{(x,y)}(f_j(x,y) - T(x,y))f_j(x,y)$ and $\lambda_j \ln M = \frac{1}{2\sigma^2}n\|f_j\|_2^2$. Since $T(x,y) - f_j(x,y) \sim N(0,\sigma^2)$ under the model f_j, $\mathbb{P}_{f_j}(\Delta_j \geq 0) = 1/2 \equiv p$. To apply Lemma 1, it remains to show that there is a $\lambda \in (0,1)$ such that $\lambda_j < \lambda$. We have

$$\begin{aligned}\lambda_j \ln M &= \frac{1}{2\sigma^2}n\|f_j\|_2^2 = \frac{1}{2\sigma^2}n\sum_{j=1}^{N_\epsilon}\delta^2 L^2\epsilon^{2\alpha+1} \\ &= \frac{1}{2\sigma^2}nN_\epsilon\delta^2L^2\epsilon^{2\alpha+1} = \frac{1}{2\sigma^2}nN_\epsilon\delta^2L^2\frac{\sigma^2}{nL^2} = \frac{N_\epsilon\delta^2}{2}\end{aligned} \tag{49}$$

where we have used that $\epsilon = (\sigma^2/(nL^2))^{1/(2\alpha+1)}$. Since $\log_2(M+1) \geq N_\epsilon/8$, we can choose δ small enough so that $\lambda \equiv (\delta^2 N_\epsilon)/\ln M < 1$. This completes the proof.

4. Simultaneous adaptation to all Hölder classes

We adapt the work of Donoho and Johnstone [19] to develop an estimator that is nearly optimal over all smoothness classes (that is, over all $\alpha > 0$) in the presence of Gaussian noise. As in [19], this follows by developing an estimate of each wavelet coefficient of the function. Since the wavelet transform is an orthogonal transformation, each wavelet coefficient of the observed, noisy function is normally distributed around the true value, and we can reduce the argument we give to that of estimating a single normally distributed random variable.

Suppose $Y \sim N(\theta, 1)$. For any $\delta \in (0,1)$, $p \in (0,2)$, and $\tau = \sqrt{2\log(1/\delta)}$, define $m_\delta^p(\theta)$ by

$$m_\delta^p(\theta) = \tau^p \min\{(\theta/\tau)^2, 1\}. \tag{50}$$

Donoho and Johnstone consider the problem of estimating θ using with loss function $(\hat\theta - \theta)^2/(\delta + m_\delta^p(\theta))$, where $\hat\theta$ is the estimator. Since this loss function penalizes mistakes made at small values of θ more than those made at large values of θ, it pays to "shrink" the standard estimate Y of θ towards 0. Formally, for any $\tau > 0$ we define the function

$$\eta_\tau(x) = (|x| - \tau)_+ \operatorname{sgn}(x) \tag{51}$$

which moves x closer to 0 by $\min\{\tau, |x|\}$. We then estimate θ by $\eta_\tau(Y)$ with $\tau = \sqrt{2\log(1/\delta)}$.

Let $M_p^*(\delta)$ denote the maximum risk of this estimator; that is,

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$M_p^*(\delta) = \sup_\theta \frac{\mathbb{E}_\theta[(\theta - \hat{\theta})^2]}{\delta + m_\delta^p(\theta)}. \tag{52}$$

We then have the following lemma from [19]:

Lemma 2. *There is a $C > 0$ so that for all δ sufficiently small,*

$$M_p^*(\delta) \le C \log(1/\delta)^{1-p/2}. \tag{53}$$

We will make use of the following trivial corollary to Lemma 2:

Corollary 1. *Suppose $Y \sim N(\theta, \sigma^2)$, where σ^2 is assumed to be known. Define the estimator of θ to be $\hat{\theta} = \eta_{\sigma\tau}(Y)$ where $\tau = \sqrt{2\log(1/\delta)}$ and $\delta \in (0,1)$. Then there is a constant $C > 0$ such that for all δ sufficiently small,*

$$\sup_\theta \frac{\mathbb{E}_\theta[(\hat{\theta} - \theta)^2]}{\delta + m_\delta^p(\theta/\sigma)} \le C\sigma^2 \log(1/\delta)^{1-p/2}. \tag{54}$$

We now define the estimator we will use. Following the notation in [19], let $r = 2\alpha/(2\alpha+1)$ and let $p = 2(1-r)$. Expand f in a two-dimensional Haar series:

$$f(x,y) = \sum_\Phi \langle f, \Phi\rangle \Phi(x,y) \equiv \sum_\Phi A_\Phi \Phi(x,y) \tag{55}$$

Since the Haar transform is orthogonal, each observed Haar coefficient $\langle T, \Phi\rangle$ is normally distributed around the true coefficient $\langle f, \Phi\rangle$, with variance $\epsilon^2 \equiv \sigma^2/n$. We will define the estimator $\hat{f}_n^*(x,y)$ by letting its wavelet coefficients be shrinkage estimators of f's wavelet coefficients; that is, let $\hat{A}_\Phi = \eta_{\epsilon\tau}(\langle T, \Phi\rangle)$ for non-constant tensor Haar functions Φ, and $\hat{A}_{\mathbf{1}} = \langle T, \mathbf{1}\rangle$ and define $\hat{f}_n^*(x,y)$ by

$$\hat{f}_n^*(x,y) = \sum_\Phi \hat{A}_\Phi \Phi(x,y) \tag{56}$$

where $\tau = \sqrt{2\log(n)}$. We will bound the mean squared error and the pointwise squared error of $\hat{f}_n^*$.

4.1. Mean squared error of $\hat{f}_n^*$

Suppose that the function f has mixed Hölder(α) norm L, for some $\alpha > 0$. Since the squared error from estimating $\langle f, \mathbf{1}\rangle$ is of size $\mathcal{O}(1/n)$, we can ignore its contribution for, as we shall see, this is a smaller order of magnitude than that contributed by estimating the other Haar coefficients. By Corollary 1 with $\delta = 1/n$ we have

$$\begin{aligned} \mathbb{E}_f \|f - \hat{f}_n^*\|_2^2 &= \sum_\Phi \mathbb{E}_f\{(A_\Phi - \hat{A}_\Phi)^2\} \le C\log(n)^r \sum_\Phi (\epsilon^2/n + \epsilon^2 m_{1/n}^p(A_\Phi/\epsilon)) \\ &= C\log(n)^r \sum_\Phi \epsilon^2/n + C\log(n)^r \epsilon^2 \sum_\Phi m_{1/n}^p(A_\Phi/\epsilon). \end{aligned} \tag{57}$$

We will estimate each of the two sums on the right side of (57). For the first, since any rectangle can support no more than a constant (in fact, $B_L^{-2} - 1$) number of Haar functions, and the number of rectangles is no more than $\mathcal{O}(n)$, we have:

$$\log(n)^r \sum_\Phi \epsilon^2/n \le C\log(n)^r \epsilon^2 = C\log(n)^r \frac{\sigma^2}{n}. \tag{58}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

As $n \to \infty$, this term will become negligible.

We turn to the second sum from (57). Because f has mixed Hölder(α) norm L, $|A_\Phi| \leq L|R(\Phi)|^{\alpha+1/2}$. If we let $\xi_\Phi = L|R(\Phi)|^{\alpha+1/2}/\epsilon$, we can write

$$\begin{aligned} \epsilon^2 m^p_{1/n}(A_\Phi/\epsilon) &\leq \epsilon^2 m^p_{1/n}(\xi_\Phi) \\ &= L^{2/(2\alpha+1)}|R(\Phi)|\epsilon^{2r}\xi_\Phi^{2(r-1)} m^p_{1/n}(\xi_\Phi) \\ &= L^{2/(2\alpha+1)}|R(\Phi)|\epsilon^{2r} \min\left\{\left(\frac{\xi_\Phi}{\tau_n}\right)^{2r}, \left(\frac{\tau_n}{\xi_\Phi}\right)^{2(1-r)}\right\} \end{aligned} \tag{59}$$

where $\tau_n = \sqrt{2\log(n)}$. The last equality is easily verified from the definition of $m^p_{1/n}(\xi_\Phi)$.

Let $S > 0$ satisfy $S^{\alpha+1/2} = \epsilon\tau_n/L$. Then we observe that $\xi_\Phi/\tau_n = (|R(\Phi)|/S)^{\alpha+1/2}$, and substituting this into (59) yields

$$\epsilon^2 m^p_{1/n}(A_\Phi/\epsilon) \leq L^{2/(2\alpha+1)}|R(\Phi)|\epsilon^{2r} \min\left\{\left(\frac{|R(\Phi)|}{S}\right)^{2\alpha}, \frac{S}{|R(\Phi)|}\right\}. \tag{60}$$

Summing over all wavelets Φ yields (with the constant C changing meaning from line to line):

$$\begin{aligned} \epsilon^2 \sum_\Phi m^p_{1/n}(A_\Phi/\epsilon) &\leq L^{2/(2\alpha+1)}\epsilon^{2r} \sum_\Phi |R(\Phi)| \min\left\{\left(\frac{|R(\Phi)|}{S}\right)^{2\alpha}, \frac{S}{|R(\Phi)|}\right\} \\ &\leq CL^{2/(2\alpha+1)}\epsilon^{2r} \sum_R |R| \min\left\{\left(\frac{|R|}{S}\right)^{2\alpha}, \frac{S}{|R|}\right\} \\ &= CL^{2/(2\alpha+1)}\epsilon^{2r}\left\{\sum_{R:|R|\geq S} S + \sum_{R:|R|<S} \frac{|R|^{2\alpha+1}}{S^{2\alpha}}\right\} \\ &\leq CL^{2/(2\alpha+1)}\epsilon^{2r}\left\{\log_{B_U^{-1}}(1/S)+1\right\} \leq CL^{2/(2\alpha+1)}\epsilon^{2r}\log_{B_U^{-1}}(nL/\sigma^2), \end{aligned} \tag{61}$$

where we have used Corollary 4 and Corollary 5. Combining (57), (58) and (61) yields the estimate:

$$\begin{aligned} \mathbb{E}_f\|f - \hat{f}_n^*\|_2^2 &\leq C\log(n)^r\left(\frac{\sigma^2}{n} + L^{2/(2\alpha+1)}\epsilon^{2r}\log_{B_U^{-1}}(nL/\sigma^2)\right) \\ &\leq C\log^{2\alpha/(2\alpha+1)}(n)(\sigma^2 L^{1/\alpha}/n)^{2\alpha/(2\alpha+1)}\log_{B_U^{-1}}(nL/\sigma^2)(1+o(1)) \end{aligned} \tag{62}$$

Comparing (62) to (34), we see that the bound on the shrinkage estimator is only a factor of $C\log(n)^{2\alpha/(2\alpha+1)}$ worse than that on the estimator (27) from Section 3.1, with $C = C(B_L, B_U, \alpha)$, even though the shrinkage estimator is not defined using the parameters L and α.

4.2. Pointwise squared error of $\hat{f}_n^*$

We bound the expected squared error of $\hat{f}_n^*$ at an arbitrary point $(x_0, y_0) \in X \times Y$. As we observed when estimating the mean squared error, the squared error from estimating $\langle f, \mathbf{1}\rangle$ is of size $\mathcal{O}(1/n)$, and hence we can ignore its contribution because, as we will see, this is a smaller order of magnitude than that contributed by estimating the other Haar coefficients. We have

$$f(x_0, y_0) - \hat{f}_n^*(x_0, y_0) = \sum_\Phi (A_\Phi - \hat{A}_\Phi)\Phi(x_0, y_0) \tag{63}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

the sum being over all Φ whose support contains (x_0, y_0). We have:

$$\begin{aligned}\mathbb{E}(f(x_0,y_0)-\hat{f}_n^*(x_0,y_0))^2 &\le \Big(\sum_\Phi \sqrt{\mathbb{E}\{(A_\Phi-\hat{A}_\Phi)^2\}|\Phi(x_0,y_0)|^2}\Big)^2\\ &\le C\Big(\sum_\Phi \sqrt{\mathbb{E}\{(A_\Phi-\hat{A}_\Phi)^2\}|R(\Phi)|^{-1}}\Big)^2\end{aligned} \tag{64}$$

where we have used the bound $\|\Phi\|_\infty = \mathcal{O}(|R(\Phi)|^{-1/2})$.

Observe that by Corollary 1 with $\delta = 1/n$ we get

$$\mathbb{E}\{(A_\Phi-\hat{A}_\Phi)^2\} \le C\log(n)^{1-p/2}\{\epsilon^2/n + \epsilon^2 m^p_{1/n}(A_\Phi/\epsilon)\}. \tag{65}$$

Substituting (65) into (64) and using $(a+b)^{1/2} \le a^{1/2} + b^{1/2}$ gives

$$\begin{aligned}\mathbb{E}(f(x_0,y_0)-\hat{f}_n^*(x_0,y_0))^2 &\le C\Big(\sum_\Phi \sqrt{|R(\Phi)|^{-1}\log(n)^r\{\epsilon^2/n+\epsilon^2 m^p_{1/n}(A_\Phi/\epsilon)\}}\Big)^2\\ &\le C\Big(\sum_\Phi \sqrt{|R(\Phi)|^{-1}\log(n)^r\epsilon^2/n}\\ &\qquad + C\sum_\Phi \sqrt{|R(\Phi)|^{-1}\log(n)^r\epsilon^2 m^p_{1/n}(A_\Phi/\epsilon)}\Big)^2.\end{aligned} \tag{66}$$

We estimate each summand on the right side of (66) separately. For the first sum, we have

$$\begin{aligned}\sum_\Phi \sqrt{|R(\Phi)|^{-1}\log(n)^r\epsilon^2/n} &= \log(n)^{r/2}\epsilon n^{-1/2}\sum_\Phi |R(\Phi)|^{-1/2}\\ &\le C\log(n)^{r/2}\epsilon n^{-1/2}n^{1/2}\log_{B_U^{-1}}(n)\\ &= C\log(n)^{r/2}\epsilon\log_{B_U^{-1}}(n) = C\frac{\sigma}{n^{1/2}}\log(n)^{r/2}\log_{B_U^{-1}}(n)\end{aligned} \tag{67}$$

where we have used Corollary 5. This term will become negligible as $n \to \infty$.

As for the second sum on the right side of (66), by inequality (60) we have

$$\begin{aligned}&\sum_\Phi \sqrt{|R(\Phi)|^{-1}\log(n)^r\epsilon^2 m^p_{1/n}(A_\Phi/\epsilon)}\\ &\qquad\le C\log(n)^{r/2}L^{1/(2\alpha+1)}\epsilon^r\sum_R \min\left\{\left(\frac{|R|}{S}\right)^\alpha, \left(\frac{S}{|R|}\right)^{1/2}\right\}\end{aligned} \tag{68}$$

where the sum is over all rectangles containing (x_0, y_0), and $S^{\alpha+1/2} = \epsilon\tau_n$.

We can write the sum over R as:

$$\begin{aligned}\sum_R \min\left\{\left(\frac{|R|}{S}\right)^\alpha, \left(\frac{S}{|R|}\right)^{1/2}\right\} &= \sum_{|R|\le S}\left(\frac{|R|}{S}\right)^\alpha + \sum_{|R|>S}\left(\frac{S}{|R|}\right)^{1/2}\\ &\le C(\log_{B_U^{-1}}(1/S)+1) \le C\log_{B_U^{-1}}(nL/\sigma)\end{aligned} \tag{69}$$

where we have used Corollary 5.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

From (68) and (69), we get

$$\sum_{\Phi} \sqrt{|R(\Phi)|^{-1} \log(n)^r \epsilon^2 m_{1/n}^p(A_\Phi/\epsilon)} \leq C \log(n)^{r/2} \epsilon^r \log_{B_U^{-1}}(nL/\sigma). \tag{70}$$

Consequently, combining (66), (67) and (70) with $\epsilon = \sigma^2/n$ we get

$$\begin{aligned} &\mathbb{E}(f(x_0, y_0) - \hat{f}_n^*(x_0, y_0))^2 \\ &\leq C \log(n)^{2\alpha/(2\alpha+1)} (\sigma^2 L^{1/\alpha}/n)^{2\alpha/(2\alpha+1)} \log^2_{B_U^{-1}}(nL/\sigma^2)(1 + o(1)) \end{aligned} \tag{71}$$

This is only a factor of $C \log(n)^{2\alpha/(2\alpha+1)}$ worse than (38), with $C = C(B_L, B_U, \alpha)$, even though the shrinkage estimator is not defined using the parameters L and α.

4.3. The mixed Hölder norm of $\hat{f}_n^*$

We have shown that the shrinkage $\hat{f}_n^*$ estimator is expected to be reasonably close to the true function f, measured in L^2 or pointwise. However, this property alone does not guarantee that $\hat{f}_n^*$ itself has small mixed Hölder norm. It follows from a result of Donoho [20] that $\hat{f}_n^*$ will have mixed Hölder(α) within a constant factor of f's with high probability (converging to 1 as $n \to \infty$). Indeed, from Theorem 4.1 in [20] we have:

Theorem 3. *The probability*

$$\Pr\{|\hat{A}_\Phi| \leq |A_\Phi| \text{ for all } \Phi\} \tag{72}$$

that all the wavelet coefficients of the shrinkage estimator do not exceed those of the true function f converges to 1 as $n \to \infty$.

Using the characterization of the mixed Hölder(α) norm $L(f, \alpha)$ via the magnitude of the wavelet coefficients, the next result is immediate:

Theorem 4. *For any $\alpha > 0$, there is a constant $C = C(B_L, B_U, \alpha)$ such that*

$$\Pr\{L(\hat{f}_n^*, \alpha) \leq C L(f, \alpha)\} \tag{73}$$

converges to 1 as $n \to \infty$.

In other words, the estimator $\hat{f}_n^*$ is almost as smooth as the true function f with high probability.

We illustrate the performance of the shrinkage estimator $\hat{f}_n^*$ numerically. We note that in order for the theory to apply, we must know the variance σ^2. Absent this knowledge, Donoho and Johnstone [18,20,21] propose using the robust estimator of σ defined as the median absolute deviation of the finest scale wavelet coefficients, divided by .6745; see also [35–37] for justification of this estimator. We adapt this idea to our setting and take our estimator $\hat{\sigma}$ to be the median absolute deviation of the Haar coefficients at the product of the finest scales of each estimator, divided by .6745.

We note that there is a tendency for the shrinkage estimator to oversmooth the data; for instance, it is apparent from Fig. 1 that some of the small-scale Haar coefficients in the original function have been washed away by the shrinkage estimator. This phenomenon is apparent in the classical application of wavelet shrinkage estimators to signal denoising [20]. In the classical setting, it can be remedied by changing the amount of shrinkage based on the wavelet's scale; the resulting estimator is known as the SUREShrink estimator [18]. However, with the SUREShrink estimator one loses the probabilistic guarantees that the resulting estimator is as smooth as f.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

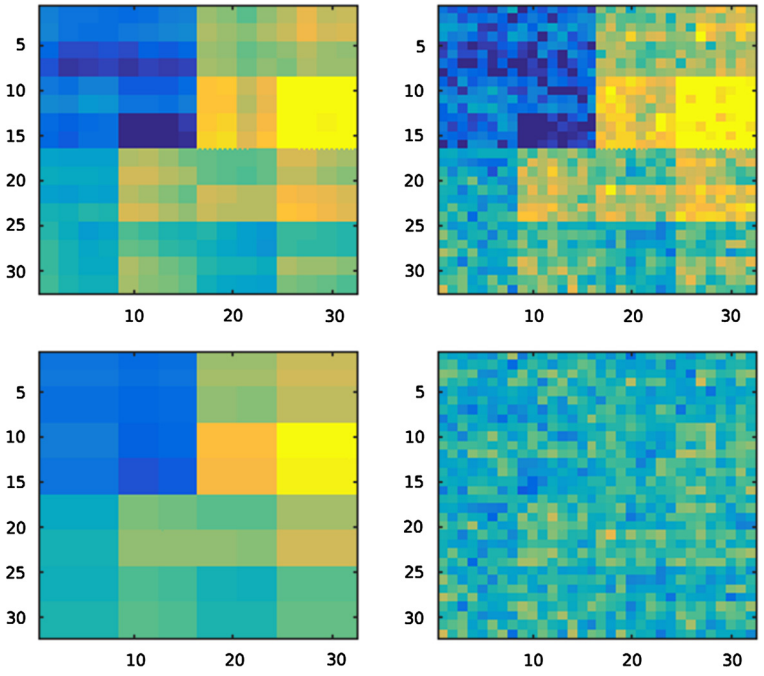

Fig. 1. Right to left, top to bottom: the original matrix; the matrix plus i.i.d. Gaussian noise; the denoised matrix; the residual.

4.4. Averaging over trees

In many problems in data analysis, tree metrics are only used to approximate some other, continuous metric (or quasi-metric). The tree will most likely separate certain points at a high level that are actually close together in the true geometry. In many constructions of trees used in machine learning (see, e.g. [38–40], among others), the trees depend on certain parameters. By varying these parameters, we obtain a family of trees and corresponding tree metrics. Although any one tree will introduce artificial breaks between data points, combining the output from all the trees will help wash away these artifacts.

On this subject, we mention that there is a vast literature in theoretical computer science on approximating arbitrary metrics by averaging tree metrics; seminal papers on this subject include [29–31]. A conceptually similar idea appears in the paper [41], in which Coifman and Donoho confront the problem that classical wavelets are attached to the dyadic grid, which creates artifacts when the signal being processed does not align with the grid. Their proposed solution involves combining the outputs from a number of shifted dyadic grids.

In the context of the present work, we consider the following scenario. Suppose that there are a family of pairs of tree metrics on X and Y, and that f has mixed Hölder norm not exceeding L for all pairs. This assumption will hold if, for example, there are intrinsic metrics on X and Y with respect to which f has mixed Hölder norm L, and each tree metric on X or Y dominates the intrinsic metric on that space.

We will denote by Π a distribution over pairs of trees; so if G is some random variable that depends on the tree pairs, $\mathbb{E}_\Pi[G]$ denotes its expected value. Let $\hat{f}^*_{\mathcal{T},n}$ denote the shrinkage estimator based on the pair of trees $\mathcal{T} = (\mathcal{T}_X, \mathcal{T}_Y)$, and let $\hat{f}^*_{\Pi,n} = \mathbb{E}_\Pi \hat{f}^*_{\mathcal{T},n}$ denote the expectation of these estimators over the family of random trees equipped with distribution Π.

For any function f, it follows easily from Jensen's inequality that:

$$\mathbb{E}_f \|f - \hat{f}^*_{\Pi,n}\|_2^2 \leq \mathbb{E}_\Pi \mathbb{E}_f \|f - \hat{f}^*_{\mathcal{T},n}\|_2^2 \tag{74}$$

Taking supremums then yields:

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

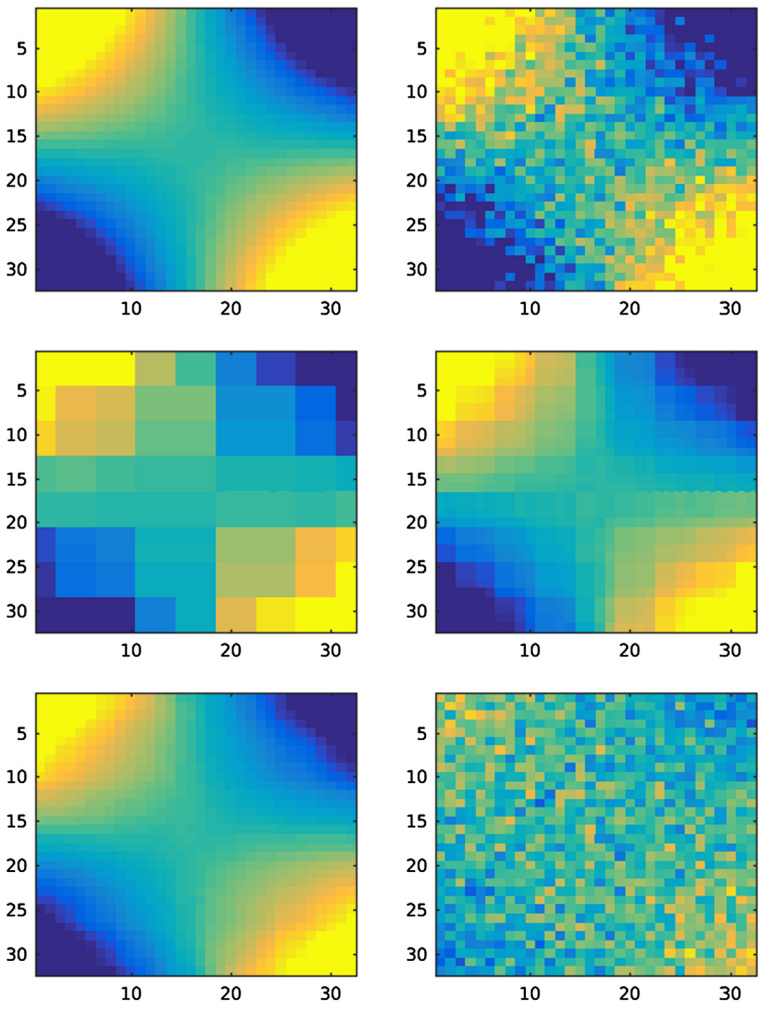

Fig. 2. Right to left, top to bottom: the clean function f, the function f with i.i.d. Gaussian noise added, the estimator with a single tree pair, the average of 5 estimators using random shifts of the trees, the average of 50 estimators using random shifts of the trees, the final residual.

$$\sup_f \mathbb{E}_f \|f - \hat{f}^*_{\Pi,n}\|_2^2 \leq \mathbb{E}_\Pi \sup_f \mathbb{E}_f \|f - \hat{f}^*_{\mathcal{T},n}\|_2^2$$
$$\leq C \log^{2\alpha/(2\alpha+1)}(n)(\sigma^2 L^{1/\alpha}/n)^{-2\alpha/(2\alpha+1)} \log_{B_U^{-1}}(nL/\sigma^2)(1+o(1)) \tag{75}$$

This implies that as long as f has the same mixed Hölder norm L for all trees being averaged, then the maximum risk for the average estimator cannot be worse than the expected risk for each individual estimator. Consequently, we do not expect to do worse by averaging over multiple trees; in fact, we should do much better as averaging will wash away many of the artificial discontinuities any estimator based on a single tree will be faced with.

To illustrate this observation, we took the smooth function $f(x,y) = (x-.5)(y-.5)$ on $[0,1] \times [0,1]$, sampled at an 32-by-32 equispaced grid of points, and built a family of randomly shifted (on the circle) dyadic trees. Since each tree metric dominates the Euclidean metric, if the function f has mixed Hölder(α) norm L with respect to the Euclidean metric then it will have mixed Hölder(α) norm not exceeding L with respect to all the tree metrics, and so we may apply our analysis.

In Fig. 2, we show the original function with and without noise, and the denoised versions based on 1, 5 and 50 random tree pairs, as well as the final residual (from the average of 50 estimators). In Fig. 3,

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

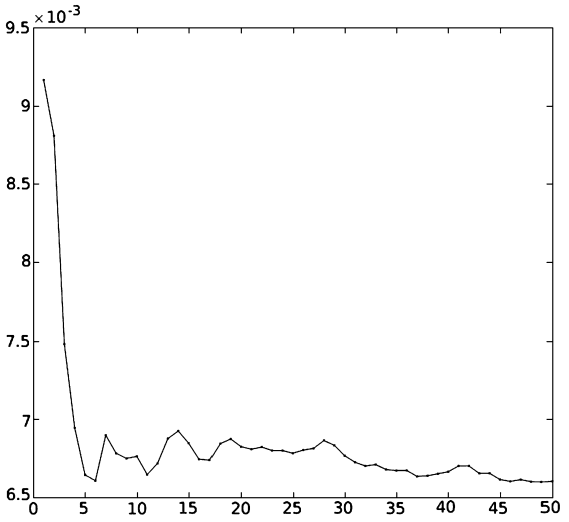

Fig. 3. The root mean square error of the average estimator, as a function of the number of tree pairs ranging from 1 to 50.

we plot the root mean squared error of the estimator obtained by averaging the individual estimators over increasing numbers of trees. We observe a dramatic improvement in the error rate.

5. Besov spaces and Calderón–Zygmund type decompositions

Sections 3 and 4 were concerned with recovering a mixed Hölder matrix that has been corrupted by noise. When the noise is Gaussian, an adaptation of the wavelet shrinkage estimator of Donoho and Johnstone was adapted to produce a matrix that is expected to be close to the noise-free matrix, and have mixed Hölder constant almost as small with high probability. In other words, if $T(x, y)$ was the matrix of observations, we wrote $T = \hat{f}_n^* + \delta$, where $\hat{f}_n^*$ is as almost as smooth as the unknown f with high probability, and close to f in L^2; and the residuals δ are approximately Gaussian.

The generative model we impose on the data is not always realistic. In particular, in many problems the partition trees on the rows and columns of the matrix are not known ahead of time; only the raw data matrix is given, and the part of the task is to *discover* the geometry by rearranging the rows and columns of the matrix. The trees built after such a process are likely to violate the assumption that the noise is independent. The analysis we have given in Section 4 of the wavelet shrinkage operator does not apply in this scenario.

We therefore turn to another way of extracting a mixed Hölder matrix from an arbitrary data matrix that does not impose a possibly unrealistic statistical model on the data when the partition trees are not known in advance. Instead of the model from Sections 3 and 4, we instead impose a weaker regularity condition on the data matrix – namely, that it lies in a Besov ball, which will be defined shortly. As discussed in the introduction, the Besov norms provide a different way of measuring a function's regularity by looking at its variation at different scales. In Section 5.1, we will introduce multiple Besov norms analogous to those encountered in the Euclidean setting and prove their equivalence.

In [28], the norm

$$\left(\sum_{\Phi} |\langle f, \Phi \rangle|^p \right)^{1/p} \tag{76}$$

is introduced, where $0 < p < 2$. It is shown in [28] that one can write a function f as a sum of a "good" function g, which is smooth, and a "bad" function b, which has small support, where the guarantees on g's smoothness and b's support size improve as this norm on f shrinks.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

In this section, we extend this result to a broader family of Besov norms, and give characterizations of these spaces without using the Haar system. These alternate characterizations will yield additional Calderón–Zygmund decompositions of f that are of potentially greater applicability than the one from [28].

In Appendix B.1 we will define the Besov norms and prove their equivalence for a single space X equipped with a partition tree. In Section 5.1, we will generalize these results to the case of the product $X \times Y$ of two spaces. In Section 5.2, we will prove additional Calderón-Zygmund decompositions based on these Besov norms.

5.1. Product Besov spaces

We now turn to defining equivalent Besov-type norms on the product $X \times Y$ of two spaces with tree metrics d_X and d_Y, respectively. We first define the mixed modulus of continuity in the product of folders $R = I \times J \in \mathcal{T}_X \times \mathcal{T}_Y$:

$$\omega_{R,p}(f) = \left(\frac{1}{|R|} \int_R \int_R |f(x,y) - f(x,y') - f(x',y) + f(x',y')|^p dx dy dx' dy' \right)^{1/p}. \tag{77}$$

We then define the Besov norm, as in the case of a single space, by

$$\|f\|_{\alpha,p} = \left(\sum_{\substack{R=I\times J: \\ I\neq X, J\neq Y}} |R|^{-\alpha p} \omega_{R,p}(f)^p \right)^{1/p} \tag{78}$$

We will show that whenever $p \geq 1$ and $\alpha > 0$, this norm is equivalent to five other norms that measure the mixed variation of f on rectangles. We will also show that these five other norms are also equivalent to each other when $p > 0$, even though they are not equivalent to $\|f\|_{\alpha,p}$ for $0 < p < 1$.

Let $p > 0$. We define the local p-variation of a function f on a rectangle $R = I \times J$ to be

$$v_{R,p}(f) = \left(\int_R |f(x,y) - m_{X,I} f(y) - m_{Y,J} f(x) + m_R(f)|^p dx dy \right)^{1/p} \tag{79}$$

and for $\alpha > 0, p > 0$, we define the norm:

$$\|f\|^{(0)}_{\alpha,p} = \left(\sum_R |R|^{-\alpha p} v_{R,p}(f)^p \right)^{1/p}. \tag{80}$$

We also define the wavelet norm

$$\|f\|^{(1)}_{\alpha,p} = \left(\sum_\Phi |R(\Phi)|^{(-\alpha-1/2+1/p)p} |\langle f, \Phi\rangle|^p \right)^{1/p} \tag{81}$$

where the sum is over all tensor products of Haar functions on X and Y, that do not include the constant functions.

Another equivalent norm is

$$\|f\|^{(2)}_{\alpha,p} = \left(\sum_R |R|^{-\alpha p} \|\Delta_R f\|_p^p \right)^{1/p} \tag{82}$$

where the difference operators Δ_R, $R = I \times J$, are defined by

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\Delta_R f(x,y) = \Delta_{X,I}\Delta_{Y,J} f(x,y) \tag{83}$$

where

$$\Delta_{X,I} f(x,y) = \sum_{\tilde{I}\in \mathrm{sub}(I)} m_{X,\tilde{I}}(f)\chi_{\tilde{I}}(x,y) - m_{X,I}(f)\chi_I(x,y) \tag{84}$$

and

$$m_{X,I} f(y) = \frac{1}{|I|}\int_I f(x,y)dx \tag{85}$$

and the corresponding operators for Y are defined similarly.

Another equivalent norm is

$$\|f\|^{(3)}_{\alpha,p} = \left(\sum_R |R|^{-(\alpha+1/p)p}|\delta_R(f)|^p\right)^{1/p} \tag{86}$$

where for any folders $I \in \mathcal{T}_X$ and $J \in \mathcal{T}_Y$, we define

$$\delta_{I\times J}(f) = \frac{1}{|I||J|}\int_{I\times J} f - \frac{1}{|I'||J|}\int_{I'\times J} f - \frac{1}{|I||J'|}\int_{I\times J'} f + \frac{1}{|I'||J'|}\int_{I'\times J'} f \tag{87}$$

where I' and J' denote the parent folders of I and J, respectively. It will also be convenient to define the one-dimensional differences

$$\delta_{X,I}(f) = \frac{1}{|I|}\int_{I\times Y} f - \frac{1}{|I'|}\int_{I'\times Y} f \tag{88}$$

and

$$\delta_{Y,J}(f) = \frac{1}{|J|}\int_{X\times J} f - \frac{1}{|J'|}\int_{X\times J'} f. \tag{89}$$

Note that we can expand f as

$$f = \sum_{I\neq X, J\neq Y} \delta_{I\times J}(f)\chi_{I\times J}(x,y) + \sum_{I\neq X}\delta_{X,I}(f)\chi_I(x) + \sum_{J\neq Y}\delta_{Y,J}(f)\chi_J(y) + \int_{X\times Y} f. \tag{90}$$

It will be convenient to define, for $I \in \mathcal{T}_X$ and $J \in \mathcal{T}_Y$

$$\hat{\delta}_I f(y) = \frac{1}{|I|}\int_I f(x,y)dx - \frac{1}{|I'|}\int_{I'} f(x,y)dx \tag{91}$$

and

$$\hat{\delta}_J f(x) = \frac{1}{|J|}\int_J f(x,y)dy - \frac{1}{|J'|}\int_{J'} f(x,y)dy \tag{92}$$

where I' denotes the parent of I and J' the parent of J. Note that $\delta_{I\times J} f = \hat{\delta}_I\hat{\delta}_J f = \delta_I(\hat{\delta}_J f) = \delta_J(\hat{\delta}_I f)$, where δ_I and δ_J are defined by

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\delta_I(f) = \frac{1}{|I|}\int_I f - \frac{1}{|I'|}\int_{I'} f \tag{93}$$

and similarly for δ_J.

Finally, we also define the norm

$$\|f\|^{(4)}_{\alpha,p} = \inf\left\{\Big(\sum_{\substack{R=I\times J:\\ I\neq X, J\neq Y}} |a_R|^p |R|^{(-\alpha+1/p)p}\Big)^{1/p} : f = \sum_R a_R\chi_R\right\} \tag{94}$$

In other words, for every expansion of f as a linear combination of indicator functions on rectangles, we look at the weighted p-norm of the expansion coefficients (excluding those rectangles of the form $X \times J$ or $I \times Y$), where the weights are powers of the rectangle's area. The norm of f is then the minimum such p-norm.

Note that the norms $\|f\|_{\alpha,p}$ and $\|f\|^{(i)}_{\alpha,p}, 0 \le i \le 4$ do not change if we add to f any function that is constant in one of the variables. Strictly speaking, we should use the term "semi-norm", though we will continue to use the term "norm" instead.

We say two norms are "equivalent" if the ratio of those two norms of a function f are bounded above and below by constants that depend only on the intrinsic parameters of the space, namely B_L, B_U, α and p. We prove the following result in Appendix B:

Theorem 5. *The norms $\|f\|_{\alpha,p}$ and $\|f\|^{(i)}_{\alpha,p}$, $0 \le i \le 4$, are equivalent for all $p \ge 1$ and all $\alpha > 0$. The norms $\|f\|^{(i)}_{\alpha,p}$, $0 \le i \le 4$ are equivalent for all $p > 0$ and all $\alpha > 0$.*

5.2. Calderón–Zygmund decompositions

We now turn to the main result of this section, namely the use of the Besov norms in writing Calderón–Zygmund decompositions of f. In particular, we show that a matrix f can be decomposed into a sum of two matrices, one with a prescribed a mixed Hölder norm and the other with small support; the quality of this decomposition can be controlled by the sizes of the Besov norms from Section 5.1.

We will define two explicit decompositions, one based on the norm $\|f\|^{(1)}_{\alpha,p}$ and the other on the norm $\|f\|^{(4)}_{\alpha,p}$. For the first, we expand f in a Haar series:

$$f = \sum_R \langle f, \Phi\rangle \Phi. \tag{95}$$

We define the function S by

$$S(x,y) = \sum_R |R|^{(-\alpha-1/2+1/p)p} \frac{|\langle f, \Phi\rangle|^p}{|R|}\chi_R(x,y) \tag{96}$$

and, for a parameter $\lambda > 0$, we also define the set E_λ by

$$E_\lambda = \{(x,y) : S(x,y) \ge \lambda\}. \tag{97}$$

We then define the bad function b_λ by

$$b_\lambda(x,y) = \sum_{R(\Phi)\subset E_\lambda} \langle f, \Phi\rangle \chi_R(x,y), \tag{98}$$

and the good function g_λ by

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$g_\lambda = f - b_\lambda = \sum_{R: R \cap E_\lambda^c \neq \emptyset} \langle f, \Phi \rangle \chi_R. \tag{99}$$

We then have the following theorem:

Theorem 6. *Suppose $p > 0$ and $\alpha > 0$. Let f be a function on $X \times Y$. Then for any $\lambda > 0$, with g_λ and b_λ defined by (98) and (99), we have $f = g_\lambda + b_\lambda$, and the following properties hold:*

1. *g_λ has mixed variation $M(g, \alpha) \leq C\lambda^{1/p}$, for a constant $C = C(B_L, B_U, \alpha)$*
2. *The support of b_λ is contained in the set E_λ, and*

$$|E_\lambda| \leq \frac{(\|f\|^{(1)}_{\alpha,p})^p}{\lambda} \tag{100}$$

3. *$E_{\lambda_2} \subset E_{\lambda_1}$ whenever $\lambda_1 < \lambda_2$.*
4. *b_λ has mean zero and zero marginals (that is, $m_X b_\lambda = m_Y b_\lambda = 0$).*

When $\alpha = 1/p - 1/2$ and $0 < p < 2$, this is exactly the decomposition described in [28]. We next consider a different decomposition, which imposes fewer restrictions on the bad function b_λ; in particular, it is not required to have zero marginals. We illustrate on selected examples that the new decomposition is far more natural in many settings.

Suppose we expand f as a linear combination of indicator functions of rectangles:

$$f = \sum_R a_R \chi_R. \tag{101}$$

Define $D = D(f, \{a_R\}_R, \alpha, p)$ by

$$D = \left(\sum_{\substack{R = I \times J: \\ I \neq X, J \neq Y}} |R|^{(-\alpha + 1/p)p} |a_R|^p \right)^{1/p}. \tag{102}$$

Note that with appropriate choices of coefficients a_R, we can have $D = \|f\|^{(3)}_{\alpha,p}$, or $D = \|f\|^{(4)}_{\alpha,p}$.

We define the function S by

$$S(x, y) = \sum_{\substack{R = I \times J: \\ I \neq X, J \neq Y}} |R|^{(-\alpha + 1/p)p} \frac{|a_R|^p}{|R|} \chi_R(x, y) \tag{103}$$

and, for a parameter $\lambda > 0$, we also define the set E_λ by

$$E_\lambda = \{(x, y) : S(x, y) \geq \lambda\}. \tag{104}$$

We then define the bad function b_λ by

$$b_\lambda(x, y) = \sum_{R \subset E_\lambda} a_R \chi_R(x, y), \tag{105}$$

and the good function g_λ by

$$g_\lambda = f - b_\lambda = \sum_{R: R \cap E_\lambda^c \neq \emptyset} a_R \chi_R. \tag{106}$$

We then have the following theorem:

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

Theorem 7. *Suppose $p > 0$ and $\alpha > 0$. Let f be a function on $X \times Y$. Then for any $\lambda > 0$, with g_λ and b_λ defined by (105) and (106), we have $f = g_\lambda + b_\lambda$, and the following properties hold:*

1. *g_λ has mixed variation $M(g, \alpha) \leq C\lambda^{1/p}$, for a constant $C = C(B_L, B_U, \alpha, p)$*
2. *The support of b_λ is contained in the set E_λ, and*

$$|E_\lambda| \leq \frac{D^p}{\lambda} \tag{107}$$

3. *$E_{\lambda_2} \subset E_{\lambda_1}$ whenever $\lambda_1 < \lambda_2$.*

Unlike Theorem 6, there is no requirement that b_λ will have zero marginals, and in general, it will not. However, this turns out to be an advantage in many applications. Many quite reasonable models of noise or outliers do not satisfy such a stringent assumption. We will illustrate this by example following the proof.

The proof of Theorem 7 mimics the proof of the result from [28]; the proof of Theorem 6 is similar and will be omitted.

Proof of Theorem 7. First, observe that the function S satisfies:

$$\int_{X \times Y} S(x, y) dx dy = \sum_R |a_R|^p |R|^{(-\alpha+1/p)p} = D^p. \tag{108}$$

Consequently, since $E_\lambda \equiv \{(x, y) : S(x, y) \geq \lambda\}$ it follows from Chebyshev's inequality

$$|E_\lambda| \leq \frac{D^p}{\lambda}. \tag{109}$$

It is also obvious that if $\lambda_1 < \lambda_2$, then $E_{\lambda_2} \subset E_{\lambda_1}$.

Finally, for any R not contained in E_λ, there is some $(x, y) \in R$ with $S(x, y) < \lambda$; consequently,

$$|R|^{(-\alpha+1/p)p} \frac{|a_R|^p}{|R|} \leq S(x, y) < \lambda \tag{110}$$

and so $|a_R| \leq \lambda^{1/p} |R|^\alpha$. From Proposition 1, it follows that $M(g_\lambda, \alpha) \leq C\lambda^{1/p}$, as desired. □

Finding the decomposition $f = \sum_R a_R \chi_R$ with minimum value of $\sum_R |R|^{(-\alpha+1/p)p} |a_R|^p$ is a convex optimization problem when $p \geq 1$. The case $p = 1$ is particularly well-suited to the Calderón–Zygmund decomposition, as expansions that minimize l_1 norms are generally sparse [42,43] and so encourage the sparsity of the function b_λ.

We tested two different decompositions on functions of the form $f = g + b$, where g is a function with small mixed Hölder(1) norm and b is a "bad" function. For the first choice of b, we took a function b consisting of randomly placed bottom-level (on both trees) tensor Haar functions, with large coefficients. The resulting function f is shown in Fig. 4, along with the individual functions g and b. Note that the color scalings are different so that the non-zero values of b can be shown more clearly.

We compared a Haar-based decomposition described in Theorem 6 with a minimization-based decomposition from Theorem 7, where the coefficients are obtained by solving the minimization problem, with parameters $p = 1$ and $\alpha = 1$. This minimization problem can be formulated as a linear program; to compute the solution numerically, we used the CVX optimization package [44,45].

In general, the coefficients $\{a_R\}_R$ that solve the minimization problem are not uniquely defined, and hence there is a not a unique definition of the functions g_λ and b_λ using the minimization-based decomposition.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

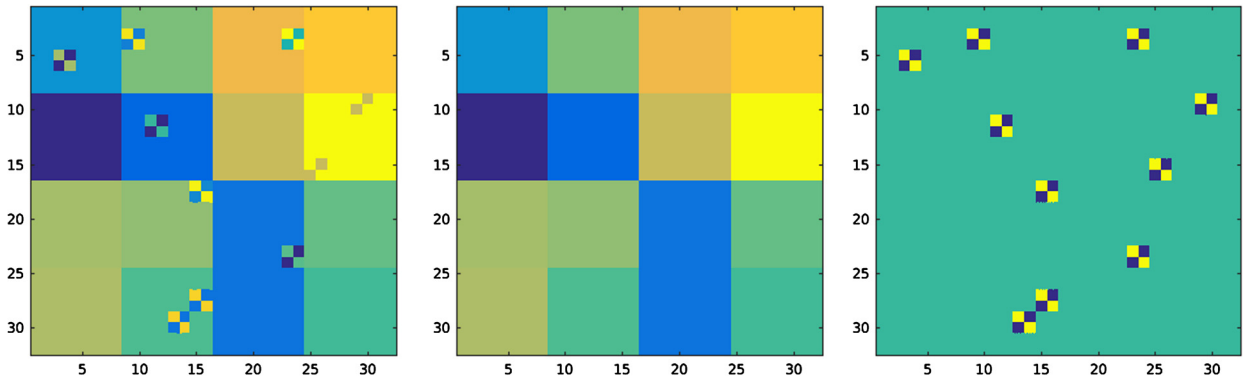

Fig. 4. The function $f = g + b$.

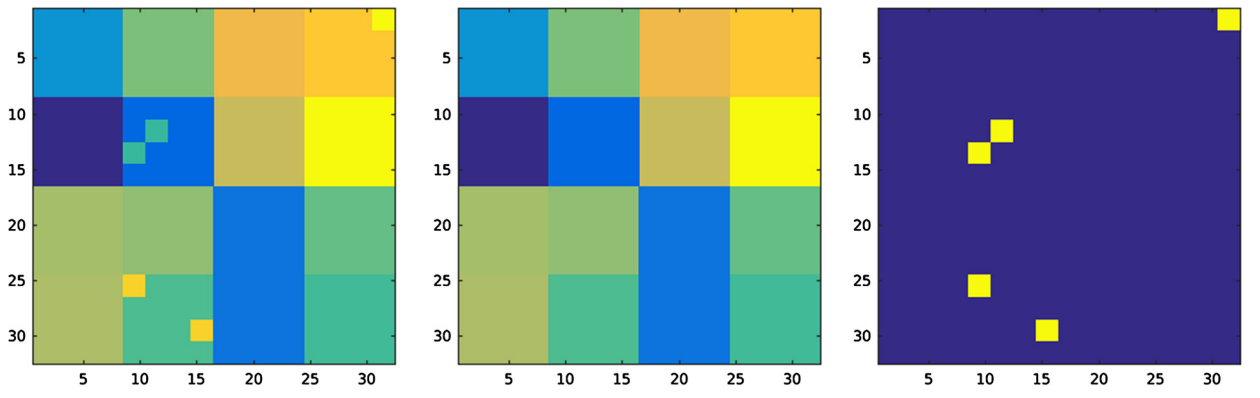

Fig. 5. The functions $f = g + b$ (on the left), g (in the middle) and b (on the right). The colors for b have been rescaled for display purposes.

However, the solutions we obtained all yielded functions b_λ whose support coincided with the spikes. In this sense, both decompositions are successful in that they identify the support of the bad function b.

However, the two decompositions are not equally successful when we run the same experiment with a different bad function b, consisting of rectangular bumps instead of Haar functions. The resulting function f is shown in Fig. 5, along with the individual functions g and b; the color scalings are different so that the non-zero values of b can be shown more clearly. In Fig. 6, we show the Haar-based decomposition for the choice of λ with the smallest error between the original function g and the function g_λ. Since the bad function b in this case cannot be written as a sum of Haar functions with small support, the Haar-based decomposition fails to cleanly split g and b for any value of λ; that is, it does not recover the correct decomposition.

By contrast, for a large range of values of λ the minimization approach recovers a bad function b_λ whose support coincides with the function b. We show the functions g_λ and b_λ from this decomposition in Fig. 6 as well.

6. Conclusion

This paper has considered two different ways of writing a data matrix with tree metrics on the rows and columns as a sum of a mixed Hölder matrix and a well-modeled residual. In Sections 3 and 4, we adapted the wavelet shrinkage estimators of Donoho and Johnstone to remove Gaussian noise from a mixed Hölder matrix. In the classical case, these results depend largely on the equivalence of the Hölder norm of a function and the magnitude of its wavelet coefficients; because this equivalence also applies to the mixed Hölder norm on tree metrics and the size of the tensor Haar coefficients, the same analysis goes through.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

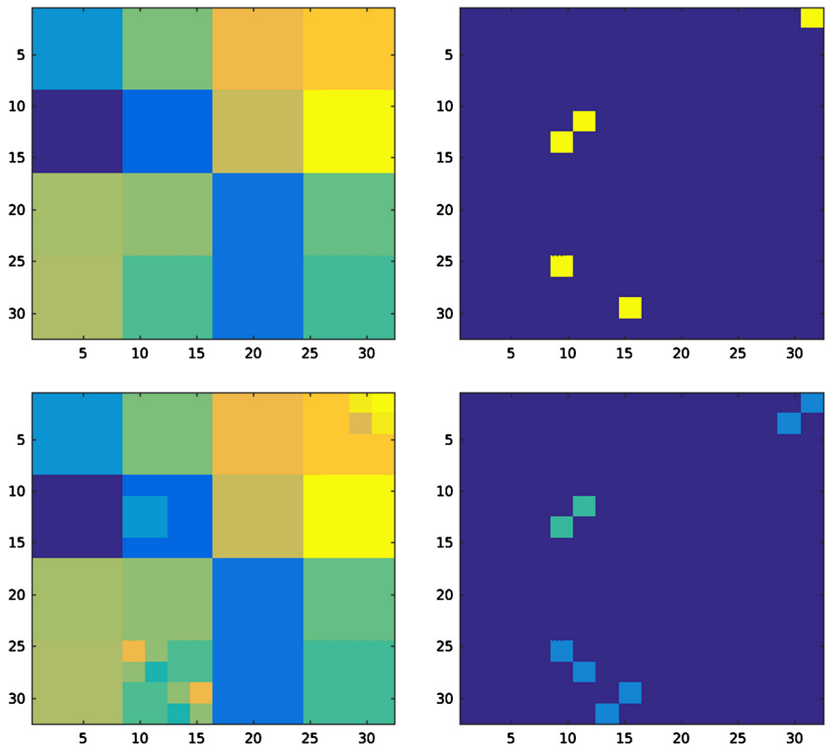

Fig. 6. The first row shows the Calderón–Zygmund decomposition using l_1 minimization, while the second row shows the Haar-based decomposition.

The primary technical challenge in this theory results from the non-homogeneity of the trees (there is not a strong notion of the level of the tree, since the decay rates of folders can differ drastically down different branches).

In Section 5, we developed a theory of Besov spaces on trees and the products of trees. In particular, we defined a natural modulus of continuity and a corresponding multiscale Besov norm, and showed that this norm is equivalent to other norms that measure the variation of a function across different scales. Extending the results from one dimension to multiple dimensions was straightforward in all cases; we apply the one-dimensional result to the function's variation in each variable separately. The main application of this theory is that a function can be decomposed into a sum of a mixed Hölder function and a function with small support; the Besov norm controls the quality of this decomposition.

To apply these decompositions to a real data matrix, the user must have row and column trees so that the matrix in question satisfies the necessary conditions. In many data analysis problems that we encounter in practice, however, such trees may not be known a priori, and it is incumbent on the practitioner to construct these trees from scratch.

In [32,28], heuristic methods for achieving a good matrix organization are discussed. The basic technique for matrix organization described there is to iterate between organizing the rows and columns of the matrix, using the organization of one to refine the organization of the other. Much recent work has been done following this same framework; detailed experimental results are contained in Jerrod Ankenman's doctoral dissertation [46], which can be explored interactively online [47]. The algorithms used in this work are based on a metric between vectors related to the Earth Mover's Distance (EMD) between probability measures [48]; the computation of EMD on trees is studied in the papers [49] and [50].

Finally, we note that it is straightforward to extend the theoretical results on the product of two spaces to the product of $d \geq 3$ spaces. In applications where data is indexed by three or more coordinate axes this extension may be useful; for instance, see the recent work [51], in which an experimental database with three

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

axes is studied by building trees on each axis. The tree-building algorithms are based on multi-dimensional versions of EMD and related metrics, described in [50].

Acknowledgments

The authors thank Ronald Coifman for his guidance, support, and numerous insights. Additionally, William Leeb thanks Justin Solomon for his assistance with CVX. William Leeb acknowledges support from the Simons Collaboration on Algorithms and Geometry.

Appendix A. Estimates for trees and function approximation

In this subsection, we will prove some basic estimates we will need throughout the paper. Some of these are improved versions of estimates found in [28].

Lemma 3. *For every* $0 < \epsilon \leq 1$

$$\sum_{I:|I|\geq\epsilon} |I| \leq \log_{B_U^{-1}}(1/\epsilon) + 1.$$

Proof. Let $\mathcal{P}_0 = \{X\}$, and for $n \geq 1$ let $\mathcal{P}_n$ contain all the folders in $\mathcal{T}_X$ that are children of the folders in $\mathcal{P}_{n-1}$. Then by condition (19), for every $I \in \mathcal{P}_n$ $|I| \leq B_U^n$. Consequently, if $n > \log_{B_U^{-1}}(1/\epsilon)$, then every folder in $\mathcal{P}_n$ is of size smaller than ϵ. So

$$\sum_{I:|I|\geq\epsilon} |I| \leq \sum_{n=0}^{\lfloor \log_{B_U^{-1}}(1/\epsilon)\rfloor} \sum_{I\in\mathcal{P}_n} |I| \leq \lfloor \log_{B_U^{-1}}(1/\epsilon)\rfloor + 1$$

which gives the desired result. □

Lemma 4. *For every* $0 < \epsilon \leq 1$ *and* $\beta > 0$

$$\sum_{I:|I|\leq\epsilon} |I|^{\beta+1} \leq \frac{1}{1 - B_U^\beta}\epsilon^\beta.$$

Proof. Let $\mathcal{P}_0$ be the set of all folders I such that $|I| \leq \epsilon$ but whose parent has size strictly bigger than ϵ; if $\epsilon = 1$, take $\mathcal{P}_0 = \{X\}$. It is easy to see that the folders in $\mathcal{P}_0$ are disjoint. For every $n \geq 1$, recursively define $\mathcal{P}_n$ to be the collection of all children of the folders in the partition $\mathcal{P}_{n-1}$; note that the folders in $\mathcal{P}_n$ are also disjoint. Then

$$\sum_{I:|I|\leq\epsilon} |I|^{\beta+1} = \sum_{n\geq 0} \sum_{I\in\mathcal{P}_n} |I|^{\beta+1}.$$

We will estimate $\sum_{I\in\mathcal{P}_n} |I|^{\beta+1}$ for each $n \geq 0$. Observe that for any folder $I \in \mathcal{P}_n$, $|I| \leq B_U^n\epsilon$. Since the folders in $\mathcal{P}_n$ are disjoint, the sum of their measures is no greater than 1; therefore

$$\sum_{I\in\mathcal{P}_n} |I|^{\beta+1} \leq (B_U^n\epsilon)^\beta \sum_{I\in\mathcal{P}_n} |I| \leq (B_U^n\epsilon)^\beta.$$

Consequently,

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\sum_{I:|I|\le\epsilon} |I|^{\beta+1} \le \epsilon^\beta \sum_{n\ge 0} B_U^{n\beta} = \frac{1}{1-B_U^\beta}\epsilon^\beta$$

as desired. □

Corollary 2. *For any $0<\epsilon<1$, we have*

$$\sum_{R:|R|\le\epsilon} |R|^{\beta+1} \le \epsilon^\beta \left(\frac{1}{1-B_U^\beta}\log_{B_U^{-1}}(1/\epsilon) + \frac{1}{(1-B_U^\beta)^2} + \frac{1}{1-B_U^\beta}\right).$$

(The sum is over rectangles $R=I\times J$ of area not exceeding ϵ, where $I\in\mathcal{T}_X$ and $J\in\mathcal{T}_Y$.)

Proof. We can write

$$\begin{aligned}
\sum_{R:|R|\le\epsilon} |R|^{\beta+1} &= \sum_{\substack{I\in\mathcal{T}_X:\\ |I|\le\epsilon}} |I|^{\beta+1} \sum_{J\in\mathcal{T}_Y} |J|^{\beta+1} + \sum_{\substack{I\in\mathcal{T}_X:\\ |I|>\epsilon}} |I|^{\beta+1} \sum_{\substack{J\in\mathcal{T}_Y:\\ |J|\le\epsilon/|I|}} |J|^{\beta+1} \\
&\le \frac{1}{(1-B_U^\beta)^2}\epsilon^\beta + \frac{1}{1-B_U^\beta} \sum_{I:|I|>\epsilon} |I|^{\beta+1}\epsilon^\beta |I|^{-\beta} \\
&\le \frac{1}{(1-B_U^\beta)^2}\epsilon^\beta + \frac{1}{1-B_U^\beta}\epsilon^\beta \sum_{I:|I|>\epsilon} |I| \\
&\le \frac{1}{(1-B_U^\beta)^2}\epsilon^\beta + \frac{1}{1-B_U^\beta}\epsilon^\beta(\log_{B_U^{-1}}(1/\epsilon)+1). \qquad \square
\end{aligned}$$

We deduce a sharper version of an approximation theorem found in [28].

Corollary 3. *Let f have mixed Hölder(α) constant L. Let*

$$g(x,y) = \sum_{\Phi:|R(\Phi)|\ge\epsilon} \langle f,\Phi\rangle\Phi(x,y).$$

Then

$$\|f-g\|_2^2 \le \frac{L^2}{B_L^2}\epsilon^{2\alpha}\left(\frac{1}{1-B_U^{2\alpha}}\log_{B_U^{-1}}(1/\epsilon) + \frac{1}{(1-B_U^{2\alpha})^2} + \frac{1}{1-B_U^{2\alpha}}\right).$$

Proof. The proof is exactly the same as in [28], except we use the tighter estimate for $\sum_{R:|R|\le\epsilon}|R|^{2\alpha+1}$. □

Lemma 5. *For every $0<\epsilon\le 1$, the number of folders $I\in\mathcal{T}_X$ of area greater than or equal to ϵ is no more than*

$$\frac{1}{1-B_U}\frac{1}{\epsilon}.$$

Proof. Let $\mathcal{S}_0$ denote the set of all folders of size greater than or equal to ϵ and with the additional property that they are either singletons or all of their children are of size strictly less than ϵ. Then the folders in $\mathcal{S}_0$ are all disjoint. For $n\ge 1$, inductively define the set $\mathcal{S}_n$, $n\ge 1$, to be the collection of folders I such that I is a parent of a folder in $\mathcal{S}_{n-1}$ and I does not contain any other parent of a folder in $\mathcal{S}_{n-1}$. Again, the folders in $\mathcal{S}_n$ are all disjoint. Furthermore, each folder in $\mathcal{S}_n$ has size greater than or equal to $B_U^{-n}\epsilon$, and consequently,

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\#\mathcal{S}_n \frac{\epsilon}{B_U^n} \leq \sum_{I \in \mathcal{S}_n} |I| \leq |X| = 1$$

and so $\#\mathcal{S}_n \leq B_U^n/\epsilon$.

We will show that any folder of size greater than or equal to ϵ must lie in some $\mathcal{S}_n$. Assuming this for the moment, it implies that the total number of such folders can be upper bounded by

$$\sum_{n \geq 0} \#\mathcal{S}_n \leq \sum_{n \geq 0} \frac{B_U^n}{\epsilon} = \frac{1}{\epsilon} \frac{1}{1 - B_U},$$

which is the desired result.

We now show that any folder of size greater than or equal to ϵ lies in some $\mathcal{S}_n$. Given the collection $\mathcal{S}_n$, call I a k^{th} *generation ancestor* of $\mathcal{S}_n$ if I contains some folder in $\mathcal{S}_n$, and if k is the maximum number of folders sitting between I and some folder in $\mathcal{S}_n$, not including this folder or I.

We will prove the following claim by induction on k: for any n, all k^{th} generation ancestors of $\mathcal{S}_n$ lie in $\mathcal{S}_m$ for some $m \geq n$. To establish the base case $k = 0$, suppose that I is a 0^{th} generation ancestor of $\mathcal{S}_n$. Then I is the parent of some folder $J \in \mathcal{S}_n$. Suppose that I contained the parent $\hat{J}'$ of some other folder $\hat{J} \in \mathcal{S}_n$. Then we have the chain of inclusions $\hat{J} \subsetneq \hat{J}' \subsetneq I$, with $\hat{J} \in \mathcal{S}_n$; but this violates the condition that the maximal such chain ending with I can have length 0. By this contradiction, I does not contain the parent of any folder in $\mathcal{S}_n$, and hence is an element of $\mathcal{S}_{n+1}$.

Now suppose the claim is true for some $k \geq 0$. Take any n and any $(k+1)^{st}$ generation ancestor I of $\mathcal{S}_n$. Then there is a chain of folders $J_0 \subsetneq J_1 \subsetneq \cdots \subsetneq J_{k+1} \subsetneq I$ where $J_0 \in \mathcal{S}_n$; and no longer such chain exists. In particular, J_l is the parent folder of J_{l-1}.

Now, J_1 is the parent of $J_0 \in \mathcal{S}_n$; so the only way for J_1 to not be in $\mathcal{S}_{n+1}$ would be if it contained the parent of some other folder in $\mathcal{S}_n$; i.e. if there were folders F_0, F_1 where $F_0 \in \mathcal{S}_n$ and $F_0 \subsetneq F_1 \subsetneq J_1$. But this is impossible, since we could then form the chain $F_0 \subsetneq F_1 \subsetneq J_1 \subsetneq \cdots \subsetneq J_{k+1} \subsetneq I$ which has length $k+2$. Consequently, J_1 must lie in $\mathcal{S}_{n+1}$. But then the chain $J_1 \subsetneq \cdots \subsetneq J_{k+1} \subsetneq I$ shows that I is a k^{th} generation ancestor of $\mathcal{S}_{n+1}$, and by the induction hypothesis $I \in \mathcal{S}_m$ for some $m \geq n+1$.

In particular, we have shown that all the ancestors of $\mathcal{S}_0$ must lie in some $\mathcal{S}_n$. We conclude by observing that any folder of size greater than or equal to ϵ is an ancestor of $\mathcal{S}_n$. To see this, take any such folder I and let I_0 be its largest child; let I_1 be the largest child of I_0; and proceed in this manner. Since the folder sizes decay, eventually some I_l will either be a singleton of size exceeding ϵ, which is an element of $\mathcal{S}_0$; or some I_l will be of size less than ϵ. Take the first such I_l; so $|I_{l-1}| \geq \epsilon$. Since I_l is the largest child of I_{l-1}, all the children of I_{l-1} have size less than ϵ, and hence $I_{l-1} \in \mathcal{S}_0$. This completes the proof. □

Corollary 4. *For every $0 < \epsilon \leq 1$, the number of rectangles $R = I \times J$ of area greater than or equal to ϵ is bounded above by*

$$\frac{1}{1 - B_U} \frac{1}{\epsilon} (\log_{B_U^{-1}}(1/\epsilon) + 1).$$

Proof. The number of rectangles of area greater than or equal to ϵ is

$$\sum_{R:|R| \geq \epsilon} 1 = \sum_{\substack{J \in \mathcal{T}_Y: \\ |J| \geq \epsilon}} \sum_{\substack{I \in \mathcal{T}_X: \\ |I| \geq \epsilon/|J|}} 1 \leq \frac{1}{1 - B_U} \sum_{\substack{J \in \mathcal{T}_Y: \\ |J| \geq \epsilon}} \frac{|J|}{\epsilon} \leq \frac{1}{1 - B_U} \frac{1}{\epsilon} (\log_{B_U^{-1}}(1/\epsilon) + 1)$$

which is the desired result. □

Lemma 6. *Suppose $x \in X$, $\beta > 0$ and $0 < \epsilon \leq 1$. Then*

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\sum_{I\ni x:|I|\le\epsilon} |I|^\beta \le \frac{1}{1-B_U^\beta}\epsilon^\beta \tag{A.1}$$

and

$$\sum_{I\ni x:|I|\ge\epsilon} |I|^{-\beta} \le \frac{1}{1-B_U^\beta}\epsilon^{-\beta}. \tag{A.2}$$

Proof. For the first inequality, order the folders in the sum as $I_0 \supsetneq I_1 \supsetneq \dots$. Then $|I_j| \le B_U^j|I_0| \le B_U^j\epsilon$, and consequently

$$\sum_{I\in\mathcal{S}_{x,\epsilon}} |I|^\beta \le \epsilon^\beta \sum_{j\ge 0} B_U^{j\beta} = \epsilon^\beta \frac{1}{1-B_U^\beta}. \tag{A.3}$$

The proof of the second inequality is similar. □

Corollary 5. *Suppose* $(x,y) \in X\times Y$, $\beta > 0$ *and* $0 < \epsilon \le 1$. *Let* $\mathcal{S}_{x,y,\epsilon}$ *denote the set of all rectangles* $R = I\times J \in \mathcal{T}_X \times \mathcal{T}_Y$ *containing* (x,y) *and of size not exceeding* ϵ. *Then*

$$\sum_{R\ni(x,y):|R|\le\epsilon} |R|^\beta \le \epsilon^\beta\left(\frac{1}{1-B_U^\beta}\log_{B_U^{-1}}(1/\epsilon) + \frac{1}{(1-B_U^\beta)^2} + \frac{1}{1-B_U^\beta}\right) \tag{A.4}$$

and

$$\sum_{R\ni(x,y):|R|\ge\epsilon} |R|^{-\beta} \le \frac{1}{1-B_U^\beta}\epsilon^{-\beta}(\log_{B_U^{-1}}(1/\epsilon)+1). \tag{A.5}$$

Proof. To prove the first inequality, we have

$$\begin{aligned}
\sum_{R\ni(x,y):|R|\le\epsilon} |R|^\beta &= \sum_{I\ni x} |I|^\beta \sum_{\substack{J\ni y:\\ |J|\le\min\{\epsilon/|I|,1\}}} |J|^\beta \\
&= \sum_{\substack{I\ni x:\\ |I|\ge\epsilon}} |I|^\beta \sum_{\substack{J\ni y:\\ |J|\le|I|/\epsilon}} |J|^\beta + \sum_{\substack{I\ni x:\\ |I|<\epsilon}} |I|^\beta \sum_{J\ni y} |J|^\beta \\
&\le \sum_{\substack{I\ni x:\\ |I|\ge\epsilon}} |I|^\beta \sum_{\substack{J\ni y:\\ |J|\le|I|/\epsilon}} |J|^\beta + \frac{1}{(1-B_U^\beta)^2}\epsilon^\beta \\
&\le \frac{1}{1-B_U^\beta} \sum_{\substack{I\ni x:\\ |I|\ge\epsilon}} |I|^\beta \frac{\epsilon^\beta}{|I|^\beta} + \frac{1}{(1-B_U^\beta)^2}\epsilon^\beta \\
&\le \frac{\epsilon^\beta}{1-B_U^\beta}(\log_{B_U^{-1}}(1/\epsilon)+1) + \frac{1}{(1-B_U^\beta)^2}\epsilon^\beta
\end{aligned}$$

as desired.

For the second inequality, we have

$$\begin{aligned}
\sum_{R\in\mathcal{S}_{x,y,\epsilon}} |R|^{-\beta} &= \sum_{\substack{I\ni x:\\ |I|\ge\epsilon}} |I|^{-\beta} \sum_{\substack{J\ni y:\\ |J|\ge\epsilon/|I|}} |J|^{-\beta} \\
&\le \sum_{\substack{I\ni x:\\ |I|\ge\epsilon}} |I|^{-\beta}\frac{\epsilon^{-\beta}}{1-B_U^\beta}|I|^\beta \le \frac{1}{1-B_U^\beta}\epsilon^{-\beta}(\log_{B_U^{-1}}(1/\epsilon)+1). \qquad \square
\end{aligned}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

We can deduce a stronger version of a result from [28].

Corollary 6. *Let f have mixed Hölder(α) constant L. Let*

$$g(x,y) = \sum_{\Phi:|R(\Phi)|\geq\epsilon} \langle f,\Phi\rangle\Phi(x,y) \tag{A.6}$$

and let $C(B_L,B_U) = \sup_\Phi \|\Phi\||R(\Phi)|^{1/2}$. Then

$$\|f-g\|_\infty \leq \frac{C(B_L,B_U)}{B_L^2} L\epsilon^\alpha \left(\frac{1}{1-B_U^\alpha}\log_{B_U^{-1}}(1/\epsilon) + \frac{1}{(1-B_U^\alpha)^2} + \frac{1}{1-B_U^\alpha}\right).$$

Proof. The proof is exactly the same as in [28], except we use the tighter estimate for $\sum_{R:|R|\leq\epsilon}|R|^\alpha$. □

Corollary 7. *For any $\beta>0$ and for any rectangle $R\in\mathcal{T}_X\times\mathcal{T}_Y$,*

$$\sum_{\tilde{R}\supseteq R} |\tilde{R}|^{-\beta} \leq \left(\frac{1}{1-B_U^\beta}\right)^2 |R|^{-\beta}. \tag{A.7}$$

Proof. First, note that for any folder $|I|$, taking $\epsilon=|I|$ Lemma 6 gives $\sum_{J\supseteq I}|J|^{-\beta} \leq \frac{1}{1-B_U^\beta}|I|^{-\beta}$. The result follows easily. □

The following result will be useful in Section 5.

Proposition 1. *Suppose $f=\sum_R a_R\chi_R$. Suppose that for every rectangle $R=I\times J$ with $I\neq X$ and $J\neq Y$ we have $|a_R|\leq A_f|R|^\alpha$. Then the mixed variation $M(f,\alpha)$ of f does not exceed $C\cdot A_f$, where $C=C(B_L,B_U,\alpha)$ is a constant.*

We will deduce this from the one-dimensional version:

Lemma 7. *Suppose $f=\sum_I a_I\chi_I$. Suppose that for every folder $I\neq X$, $|a_I|\leq A_f|I|^\alpha$. Then $f(x)-f(y)\leq CA_f d_X(x,y)^\alpha$ for all $x\neq y$, where $C=C(B_L,B_U,\alpha)$ is a constant.*

Proof. The proof is nearly identical to the corresponding result for the Haar system in [5]. □

Proof of Proposition 1. For any fixed folder $I\neq X$, the function $\sum_{J\in\mathcal{T}_Y} a_{I\times J}\chi_J(y)$ has coefficients $|a_{I\times J}|\leq A_f|I|^\alpha|J|^\alpha$ when $J\neq Y$; so by Lemma 7,

$$\left|\sum_{J\in\mathcal{T}_Y} a_{I\times J}(\chi_J(y)-\chi_J(y'))\right| \leq CA_f|I|^\alpha d_Y(y,y')^\alpha \tag{A.8}$$

Now, for any fixed $y,y'\in Y$, we write the difference

$$f(x,y)-f(x,y') = \sum_{I\in\mathcal{T}_X}\chi_I(x)\left\{\sum_{J\in\mathcal{T}_Y} a_{I\times J}(\chi_J(y)-\chi_J(y'))\right\}. \tag{A.9}$$

Viewing this as a function of x, by (A.8) and Lemma 7 we see

$$f(x,y)-f(x,y')-f(x',y)-f(x',y') \leq CA_f d_Y(y,y')^\alpha d_X(x,x')^\alpha \tag{A.10}$$

which is the desired result. □

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

For the next result, we will assume that the number of points in X and Y are comparable; more precisely, that there are positive constants C_L and C_U such that

$$C_L \leq \frac{n_X}{n_Y} \leq C_U \tag{A.11}$$

Lemma 8. *Suppose (A.11) holds, and suppose that X and Y are equipped with normalized counting measure; that is, every singleton in X has measure n_X^{-1}, and every singleton in Y has measure n_Y^{-1}. Let $b > 0$, $\beta > 0$, and suppose that $\epsilon = b \cdot n^{-1/(\beta+1)}$. Let N_ϵ denote the number of rectangles R such that*

$$\epsilon B_L \leq |R| \leq \epsilon / B_L. \tag{A.12}$$

Then there is an $N = N(B_L, C_L, C_U, b, \beta)$ and a constant $C = C(B_L, C_L, C_U, \beta)$ such that for all $n > N$

$$N_\epsilon \geq C \frac{1}{\epsilon} \log_{B_L^{-1}}(n). \tag{A.13}$$

Proof. For any $\delta > 0$, let $\mathcal{T}_X(\delta)$ denote the set of all folders $I \in \mathcal{T}_X$ of size $|I| \leq \delta$, but whose parent I' has size $|I'| > \delta$. As long as $\delta \geq 1/n_X$, the folders in $\mathcal{T}_X(\delta)$ form a partition of X (they are disjoint and cover X). Consequently,

$$1 = |X| = \sum_{I \in \mathcal{T}_X(\delta)} |I| \leq \delta(\#\mathcal{T}_X(\delta))$$

and so there are at least $1/\delta$ folders in $\mathcal{T}_X(\delta)$. Note too that any folder I in $\mathcal{T}_X(\delta)$ satisfies $|I| \geq B_L|I'| \geq B_L\delta$, where I' denotes I's parent. Define $\mathcal{T}_Y(\delta)$ similarly.

We will count rectangles of the form $R = I \times J$ where $I \in \mathcal{T}_X(\sqrt{\epsilon} B_L^{-l-1})$ and $J \in \mathcal{T}_Y(\sqrt{\epsilon} B_L^l)$. To ensure $n_X^{-1} \leq |I| \leq 1$ and $n_Y^{-1} \leq |J| \leq 1$, it suffices to take values of l satisfying

$$l \geq \frac{-\beta}{2(\beta+1)} \log_{B_L^{-1}}(n) - \frac{1}{2} \log_{B_L^{-1}}(C_L) - \frac{1}{2} \log_{B_L^{-1}}(b) \tag{A.14}$$

$$l \leq \frac{1}{2(\beta+1)} \log_{B_L^{-1}}(n) - \frac{1}{2} \log_{B_L^{-1}}(b) - 1 \tag{A.15}$$

$$l \leq \frac{\beta}{2(\beta+1)} \log_{B_L^{-1}}(n) - \frac{1}{2} \log_{B_L^{-1}}(C_U) + \frac{1}{2} \log_{B_L^{-1}}(b) - 1 \tag{A.16}$$

$$l \geq \frac{-1}{2(\beta+1)} \log_{B_L^{-1}}(n) + \frac{1}{2} \log_{B_L^{-1}}(b). \tag{A.17}$$

When n is sufficiently large, there are at least $C \log_{B_L^{-1}}(n)$ such values of l, where $C = C(B_L, C_L, C_U, \beta)$. Furthermore, for each l, there are at least $\epsilon^{-1/2} B_L^{l+1}$ folders in $\mathcal{T}_X(\sqrt{\epsilon} B_L^{-l-1})$, and at least $\epsilon^{-1/2} B_L^{-l}$ folders in $\mathcal{T}_Y(\sqrt{\epsilon} B_L^l)$; consequently, there are at least B_L/ϵ such rectangles R for each l, and so in total there are at least $C\epsilon^{-1} \log_{B_L^{-1}}(n)$ rectangles.

Observe that for any such R, $|R| \leq \sqrt{\epsilon} B_L^{-l-1} \sqrt{\epsilon} B_L^l = \epsilon / B_L$ and $|R| \geq B_L^2 \sqrt{\epsilon} B_L^{-l-1} \sqrt{\epsilon} B_L^l = \epsilon B_L$, so these rectangles satisfy condition (A.12). This proves the desired lower bound on N_ϵ. □

Appendix B. Besov spaces and the proof of Theorem 5

The purpose of this section is to prove Theorem 5, on the equivalence of the product Besov norms. However, before doing so it is convenient to state and prove the analogous result for Besov norms on a single space. The proof for product spaces will be easier to derive using the one-dimensional result.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

B.1. Besov spaces on a single space

We define the analogue of the classical modulus of continuity as follows. For $p > 0$ and any folder $I \in \mathcal{T}_X$, let

$$\omega_{I,p}(f) = \left(\frac{1}{|I|} \int_I \int_I |f(x) - f(y)|^p dx dy \right)^{1/p} \tag{B.1}$$

when $p < \infty$, and

$$\omega_{I,\infty}(f) = \sup_{x,y \in I} |f(x) - f(y)|. \tag{B.2}$$

$\omega_{I,p}(f)$ measures the p-variation of f on the folder I. Let $\alpha > 0$ and $p > 0$. We define the following Besov-type norm for functions on X.

$$\|f\|_{\alpha,p} = \left(\sum_{I \neq X} |I|^{-\alpha p} \omega_{I,p}(f)^p \right)^{1/p}. \tag{B.3}$$

In other words, we measure the variation of f on each folder I, as measured by $\omega_{I,p}(f)$, and then take a weighted L^p-norm of these variations. When $p = \infty$, the p-norm of the summands is replaced by the supremum; note that $\|f\|_{\alpha,\infty}$ is simply the Hölder(α) norm of f.

Whenever $p \geq 1$ and $\alpha > 0$, we will show that the norm $\|f\|_{\alpha,p}$ is equivalent to five other norms, each of which measures the variation of the averages of f across different folders in a different way. We will also show that these five other norms are all equivalent to each other for all $p > 0$ and $\alpha > 0$.

Let f be a function on X and $\alpha > 0$. Then for any folder $I \in \mathcal{T}_X$, define the mean p-variation of f on I by

$$v_{I,p}(f) = \left(\int_I |f(x) - m_I(f)|^p dx \right)^{1/p} \tag{B.4}$$

when $p < \infty$, and

$$v_{I,\infty}(p) = \sup_{x \in I} |f(x) - m_I(f)|. \tag{B.5}$$

In other words, $v_{I,p}(f)$ measures how far f differs from its average on I, where the difference is measured in L^p.

Let $\alpha > 0$ and $p > 0$. We define:

$$\|f\|^{(0)}_{\alpha,p} = \left(\sum_{I \neq X} |I|^{-\alpha p} v_{I,p}(f)^p \right)^{1/p}. \tag{B.6}$$

In other words, we measure the mean variation of f on each folder I, as measured by $v_{I,p}(f)$, and then take a weighted L^p-norm of these variations.

We can also use the Haar coefficients to measure the function's variation, defining

$$\|f\|^{(1)}_{\alpha,p} = \left(\sum_{\phi} |I(\phi)|^{(-\alpha - 1/2 + 1/p)p} |\langle f, \phi \rangle|^p \right)^{1/p} \tag{B.7}$$

Another equivalent norm is

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\|f\|_{\alpha,p}^{(2)} = \left(\sum_{I \neq X} |I|^{-\alpha p} \|\Delta_I f\|_p^p \right)^{1/p} \tag{B.8}$$

where the difference operators Δ_I are defined by

$$\Delta_I f(x) = \sum_{\tilde{I} \in \mathrm{sub}(I)} m_{\tilde{I}}(f) \chi_{\tilde{I}}(x) - m_I(f) \chi_I(x) \tag{B.9}$$

Yet another equivalent norm is

$$\|f\|_{\alpha,p}^{(3)} = \left(\sum_{I \neq X} |I|^{(-\alpha+1/p)p} |\delta_I(f)|^p \right)^{1/p} \tag{B.10}$$

where the difference operators δ_I are defined by

$$\delta_I(f) = \frac{1}{|I|} \int_I f - \frac{1}{|I'|} \int_{I'} f \tag{B.11}$$

and where I' denotes the parent folder of the folder I.

Finally, we also define the norm

$$\|f\|_{\alpha,p}^{(4)} = \inf \left\{ \left(\sum_{I \neq X} |a_I|^p |I|^{(-\alpha+1/p)p} \right)^{1/p} : f = \sum_I a_I \chi_I \right\} \tag{B.12}$$

In other words, for every expansion of f as a linear combination of indicator functions on folders, we look at the weighted p-norm of the expansion coefficients (excluding the topmost folder $I = X$), where the weights are powers of the folder's measure. The norm of f is then the minimum such p-norm.

All the norms we have defined are equivalent in size, for appropriate parameter ranges. More precisely, we have the following theorem:

Theorem 8. *The norms $\|f\|_{\alpha,p}$ and $\|f\|_{\alpha,p}^{(i)}$, $0 \leq i \leq 4$, are equivalent for all $p \geq 1$ and all $\alpha > 0$. The norms $\|f\|_{\alpha,p}^{(i)}$, $0 \leq i \leq 4$ are equivalent for all $p > 0$ and all $\alpha > 0$.*

The remainder of this section is devoted to proving Theorem 8. In Section B.2, we will then give the proof of the two-dimensional version, Theorem 5.

Lemma 9. *For $p > 0$, there is a constant $C = C(B_L, B_U, p)$ such that $\omega_{I,p}(f) \leq C v_{I,p}(f)$ for all functions f on X.*

Proof. For $p < \infty$, we write

$$|f(x) - f(y)|^p \leq C(p)(|f(x) - m_I(f)|^p + |f(y) - m_I(f)|^p). \tag{B.13}$$

Integrating each side in both x and y over I yields and dividing by $|I|$ yields the result. The case $p = \infty$ is even simpler. □

Corollary 8. *For $p > 0$ and $\alpha > 0$, there is a constant $C = C(B_L, B_U, p)$ such that $\|f\|_{\alpha,p} \leq C\|f\|_{\alpha,p}^{(0)}$ for all functions f on X.*

Lemma 10. *For $p \geq 1$, there is a constant $C = C(B_L, B_U, p)$ such that $v_{I,p}(f) \leq C\omega_{I,p}(f)$ for all functions f on X.*

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

Proof. From Jensen's inequality, we have

$$\begin{aligned}\int_I |f(x) - m_I(f)|^p dx &= \int_I \left| f(x) - \frac{1}{|I|}\int_I f(y)dy \right|^p dx \\ &= \int_I \left| \frac{1}{|I|}\int_I f(x) - f(y)dy \right|^p dx \\ &\leq \frac{1}{|I|}\int_I \int_I |f(x) - f(y)|^p dydx \end{aligned} \tag{B.14}$$

which immediately yields the desired result. □

Corollary 9. *For $p \geq 1$ and $\alpha > 0$, there is a constant $C = C(B_L, B_U, p)$ such that $\|f\|^{(0)}_{\alpha,p} \leq C\|f\|_{\alpha,p}$ for all functions f on X.*

Lemma 10 and Corollary 9 are not true when $0 < p < 1$. It is easy to build a counterexample; take, for instance, the function f defined by

$$f(x) = \begin{cases} |I|^{-1}, \text{ if } x \in I \\ (1 - |I|)^{-1}, \text{ if } x \notin I \end{cases} \tag{B.15}$$

where I is some folder in $\mathcal{T}_X$. It is easy to check that the ratio $v_{X,p}(f)^p/\omega(X,p)(f)^p$ is proportional to $|I|^{p-1} + (1 - |I|)^{p-1}$, and so is unbounded as $|I|$ goes to zero whenever $0 < p < 1$.

Lemma 11. *For any function f on X, any folder $I \in \mathcal{T}_X$ and any $p > 0$,*

$$\|\Delta_I f\|_p^p = \sum_{\tilde{I} \in sub(I)} |\delta_{\tilde{I}}(f)|^p |\tilde{I}| \tag{B.16}$$

Proof. We have

$$\Delta_I f = \sum_{\tilde{I} \in \text{sub}(I)} m_I(f)\chi_{\tilde{I}} - \sum_{\tilde{I} \in \text{sub}(I)} m_{\tilde{I}}(f)\chi_{\tilde{I}} = \sum_{\tilde{I} \in \text{sub}(I)} \delta_{\tilde{I}}(f)\chi_{\tilde{I}} \tag{B.17}$$

and the result follows. □

The following corollary is immediate:

Corollary 10. *The norms $\|f\|^{(2)}_{\alpha,p}$ and $\|f\|^{(3)}_{\alpha,p}$ are equivalent.*

We turn to the other norms. The following lemma will be convenient.

Lemma 12. *Suppose $\{a_I\}_{I \subset I^*}$ is any collection of numbers, indexed by the folders in $\mathcal{T}_X$ contained in I^*; suppose too that $p > 0$ and $0 < s < 1/p$. Then there is a constant $C = C(p, s)$ such that*

$$\left(\sum_{I \subset I^*} |a_I||I| \right)^p \leq C|I^*|^{p(1-s)} \sum_{I \subset I^*} |I|^{sp}|a_I|^p \tag{B.18}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

Proof. If $0 < p \leq 1$, the result is immediate, since

$$\Big(\sum_{I \subset I^*} |a_I||I| \Big)^p \leq \sum_{I \subset I^*} (|a_I||I|)^p. \tag{B.19}$$

Suppose $p > 1$. Let $\mathcal{P}_l$ denote the set of subfolders of I^* such that $2^{-(l+1)} < |I|/|I^*| \leq 2^{-l}$, and let $1/p + 1/q = 1$. Then

$$\sum_{I \subset I^*} |a_I||I| = \sum_{l \geq 0} \sum_{I \in \mathcal{P}_l} |a_I||I| \leq \sum_{l \geq 0} \Big(\sum_{I \in \mathcal{P}_l} |a_I|^p \Big)^{1/p} \Big(\sum_{I \in \mathcal{P}_l} |I|^q \Big)^{1/q}. \tag{B.20}$$

Now

$$\sum_{I \in \mathcal{P}_l} |I|^q \leq |I^*|^q \sum_{I \in \mathcal{P}_l} 2^{-lq} \leq 2|I^*|^q 2^{-l(q-1)} \tag{B.21}$$

since there are no more than 2^{l+1} folders in $\mathcal{P}_l$. Consequently,

$$\begin{aligned}
\sum_{I \subset I^*} |a_I||I| &\leq \sum_{l \geq 0} \Big(\sum_{I \in \mathcal{P}_l} |a_I|^p \Big)^{1/p} \Big(\sum_{I \in \mathcal{P}_l} |I|^q \Big)^{1/q} \\
&\leq C \sum_{l \geq 0} \Big(\sum_{I \in \mathcal{P}_l} |a_I|^p \Big)^{1/p} \Big(|I^*|^q 2^{-l(q-1)} \Big)^{1/q} \\
&= C|I^*| \sum_{l \geq 0} 2^{-l(1-1/q)} \Big(\sum_{I \in \mathcal{P}_l} |a_I|^p \Big)^{1/p} \\
&\leq C|I^*| \Big(\sum_{l \geq 0} 2^{-l(q-1)} 2^{lsq} \Big)^{1/q} \Big(\sum_{l \geq 0} \sum_{I \in \mathcal{P}_l} 2^{-lsp} |a_I|^p \Big)^{1/p} \\
&\leq C|I^*|^{1-s} \Big(\sum_{l \geq 0} \sum_{I \in \mathcal{P}_l} |I|^{sp} |a_I|^p \Big)^{1/p} \\
&= C|I^*|^{1-s} \Big(\sum_{I \subset I^*} |I|^{sp} |a_I|^p \Big)^{1/p}
\end{aligned} \tag{B.22}$$

Note that the geometric series converges since $s < 1/p$. This completes the proof. □

We now show the equivalence of $\|f\|^{(3)}_{\alpha,p}$ and $\|f\|^{(4)}_{\alpha,p}$.

Proposition 2. *The norms $\|f\|^{(3)}_{\alpha,p}$ and $\|f\|^{(4)}_{\alpha,p}$ are equivalent.*

Proof. Since we can write

$$f - m_X f = \sum_{I \neq X} \delta_I(f) \chi_I \tag{B.23}$$

the inequality $\|f\|^{(4)}_{\alpha,p} \leq \|f\|^{(3)}_{\alpha,p}$ is immediate. For the other direction, take any expansion of f as a linear combination of indicator functions, $f(x) = \sum_I a_I \chi_I(x)$. Take any $I \neq X$. We will estimate the size of δ_I. First, observe that

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\delta_I = \sum_{\tilde{I}} a_{\tilde{I}} \Big(\frac{|I \cap \tilde{I}|}{|I|} - \frac{|I' \cap \tilde{I}|}{|I'|} \Big). \tag{B.24}$$

The only summands that are non-zero are those with $\tilde{I} \subsetneq I'$. Suppose $\tilde{I} \subsetneq I'$; then a crude estimate yields

$$\Big| \frac{|I \cap \tilde{I}|}{|I|} - \frac{|I' \cap \tilde{I}|}{|I'|} \Big| \leq 2 \frac{|\tilde{I}|}{|I|}. \tag{B.25}$$

We get the estimate

$$|\delta_I| \leq 2 \sum_{\tilde{I} \subsetneq I'} |a_{\tilde{I}}| \frac{|\tilde{I}|}{|I|}. \tag{B.26}$$

Consequently, if $\max\{0, 1/p - \alpha\} < s < 1/p$ we have, using Lemma 12

$$\begin{aligned} \sum_{I \neq X} |I|^{(-\alpha+1/p)p} |\delta_I|^p &\leq C \sum_{I \neq X} |I|^{-(\alpha+1/p)p} \Big(\sum_{\tilde{I} \subsetneq I'} |a_{\tilde{I}}| \frac{|\tilde{I}|}{|I|} \Big)^p \\ &\leq C \sum_{I \neq X} |I|^{(-\alpha+1/p-s)p} \sum_{\tilde{I} \subsetneq I'} |a_{\tilde{I}}|^p |\tilde{I}|^{sp} \\ &\leq C \sum_{\tilde{I} \neq X} |a_{\tilde{I}}|^p |\tilde{I}|^{sp} \sum_{I \supset \tilde{I}} |I|^{(-\alpha+1/p-s)p} \\ &\leq C \sum_{\tilde{I} \neq X} |a_{\tilde{I}}|^p |\tilde{I}|^{(-\alpha+1/p)p} \end{aligned} \tag{B.27}$$

which is the desired result. □

Next we bring the wavelet norm into the picture.

Lemma 13. *There is a constant $C = C(B_L, B_U, p)$ such that*

$$\|\Delta_I f\|_p^p \leq C \sum_{\phi : I(\phi) = I} |\langle f, \phi \rangle|^p |I|^{1-p/2} \tag{B.28}$$

for all functions f on X

Proof. One easily checks that

$$\Delta_I f = \sum_{\phi : I(\phi) = I} \langle f, \phi \rangle \phi. \tag{B.29}$$

From the estimate $\|\phi\|_\infty \leq C |I(\phi)|^{-1/2}$, it follows that $\|\phi\|_p \leq C |I(\phi)|^{1/p - 1/2}$, from which we get

$$\|\Delta_I f\|_p^p \leq C \sum_{\phi : I(\phi) = I} |\langle f, \phi \rangle|^p |I|^{1-p/2}. \qquad \square \tag{B.30}$$

Corollary 11. *There is a constant $C = C(B_L, B_U, p)$ such that $\|f\|_{\alpha,p}^{(2)} \leq C \|f\|_{\alpha,p}^{(1)}$ for all functions f on X.*

Proposition 3. *The norms $\|f\|_{\alpha,p}^{(0)}$ and $\|f\|_{\alpha,p}^{(2)}$ are equivalent.*

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

Proof. First, we write

$$\begin{aligned}(\Delta_I f)(x) &= \sum_{\tilde{I}\in \mathrm{sub}(I)} m_{\tilde{I}}(f)\chi_{\tilde{I}}(x) - m_I(f)\chi_I(x) \\ &= \sum_{\tilde{I}\in \mathrm{sub}(I)} (m_{\tilde{I}}(f) - f(x))\chi_{\tilde{I}}(x) - (m_I(f) - f(x))\chi_I(x)\end{aligned} \tag{B.31}$$

Therefore,

$$\|\Delta_I\|_p^p \le C\Big\{ \sum_{\tilde{I}\in \mathrm{sub}(I)} v_{\tilde{I},p}(f)^p + v_{I,p}(f)^p \Big\} \tag{B.32}$$

from which $\|f\|_{\alpha,p}^{(2)} \le C\|f\|_{\alpha,p}$ easily follows.

For the other direction, we write $(f - m_I(f))\chi_I$ as a telescopic sum:

$$(f - m_I(f))\chi_I = \sum_{\hat{I}\subset I} \Delta_{\hat{I}} f \tag{B.33}$$

When $0 < p \le 1$, this yields the estimate

$$v_{I,p}(f)^p \le \sum_{\hat{I}\subset I} \|\Delta_{\hat{I}} f\|_p^p \tag{B.34}$$

from which it follows that

$$\begin{aligned}\sum_{I\neq X} |I|^{-\alpha p} v_{I,p}^p &\le \sum_{I\neq X} |I|^{-\alpha p} \sum_{\hat{I}\subset I} \|\Delta_{\hat{I}} f\|_p^p \\ &= \sum_{\hat{I}\neq X} \|\Delta_{\hat{I}} f\|_p^p \sum_{I\supset \hat{I}} |I|^{-\alpha p} \le C \sum_{\hat{I}\neq X} |\hat{I}|^{-\alpha p} \|\Delta_{\hat{I}} f\|_p^p\end{aligned} \tag{B.35}$$

where $C = C(B_U, \alpha, p)$; so $\|f\|_{\alpha,p} \le C\|f\|_{\alpha,p}^{(2)}$. When $p > 1$, a little more work is required. For $l \ge 0$, let $\mathcal{P}_l$ denote the folders sitting l levels below I (so $\mathcal{P}_0 = \{I\}$, $\mathcal{P}_1 = \mathrm{sub}(I)$, etc.). (B.33) can then be written as:

$$(f - m_I(f))\chi_I = \sum_{l\ge 0} \sum_{\hat{I}\in\mathcal{P}_l} \Delta_{\hat{I}} f \tag{B.36}$$

Let $0 < s < \alpha$, and $1/p + 1/q = 1$. We have

$$\begin{aligned}|(f - m_I(f))\chi_I|^p &\le \Big(\sum_{l\ge 0} \sum_{\hat{I}\in\mathcal{P}_l} |\Delta_{\hat{I}} f| \Big)^p \le \Big(\sum_{l\ge 0} B_U^{-lsq} \Big)^{p/q} \sum_{l\ge 0} B_U^{lsp} \Big(\sum_{\hat{I}\in\mathcal{P}_l} |\Delta_{\hat{I}} f| \Big)^p \\ &= C \sum_{l\ge 0} B_U^{lsp} \sum_{\hat{I}\in\mathcal{P}_l} |\Delta_{\hat{I}} f|^p \le C|I|^{sp} \sum_{\hat{I}\subset I} |\hat{I}|^{-sp} |\Delta_{\hat{I}} f|^p\end{aligned} \tag{B.37}$$

where we have used the fact that the supports of the functions $\Delta_{\hat{I}} f$ are disjoint, so $\sum_{\hat{I}\in\mathcal{P}_l} |\Delta_{\hat{I}} f|^p = \sum_{\hat{I}\in\mathcal{P}_l} |\Delta_{\hat{I}} f|^p$. Taking the integral of each side yields:

$$v_{I,p}(f)^p \le C|I|^{sp} \sum_{\hat{I}\subset I} |\hat{I}|^{-sp} \|\Delta_{\hat{I}} f\|_p^p \tag{B.38}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

Consequently, we have

$$\begin{aligned}\sum_{I\neq X}|I|^{-\alpha p}v_{I,p}(f)^p &\leq C\sum_{I\neq X}|I|^{(-\alpha+s)p}\sum_{\hat{I}\subset I}|\hat{I}|^{-sp}\|\Delta_{\hat{I}}f\|_p^p\\ &= C\sum_{\hat{I}\neq X}|\hat{I}|^{-sp}\|\Delta_{\hat{I}}f\|_p^p\sum_{I\supset\hat{I}}|I|^{(-\alpha+s)p}\\ &\leq C\sum_{\hat{I}\neq X}|\hat{I}|^{-\alpha p}\|\Delta_{\hat{I}}f\|_p^p\end{aligned} \tag{B.39}$$

i.e. $\|f\|_{\alpha,p}\leq C\|f\|_{\alpha,p}^{(2)}$. □

We now complete the full proof of equivalence of all the norms:

Proposition 4. *There is a constant $C=C(B_L,B_U,\alpha,p)$ such that $\|f\|_{\alpha,p}^{(1)}\leq C\|f\|_{\alpha,p}^{(4)}$ for all functions f on X.*

Proof. We introduce some notation. For any Haar function $\phi(x)$, if I is a child folder of $I(\phi)$ let a_I^ϕ denote the unique value that ϕ takes on I. We can therefore write $\phi=\sum_{I\in\mathrm{sub}(I(\phi))}a_I^\phi\chi_I$. We have the estimate $|a_I^\phi|\leq C|I(\phi)|^{-1/2}$.

Take any expansion of f as a linear combination of indicator functions, $f=\sum_I a_I\chi_I$. Since the Haar functions are all mean zero, we have

$$\langle f,\phi\rangle=\sum_{\tilde{I}\subsetneq I(\phi)}\sum_{I\in\mathrm{sub}(I(\phi))}a_{\tilde{I}}a_I^\phi|\tilde{I}\cap I|\leq C|I(\phi)|^{-1/2}\sum_{\tilde{I}\subsetneq I(\Phi)}|a_{\tilde{I}}||\tilde{I}|. \tag{B.40}$$

Using this estimate, we apply Lemma 12 and get, for $\max\{0,1/p-\alpha\}<s<1/p$,

$$\begin{aligned}\sum_\phi|I(\phi)|^{(-\alpha-1/2+1/p)p}|\langle f,\phi\rangle|^p &\leq C\sum_\phi|I(\phi)|^{(-\alpha-1+1/p)p}\Big(\sum_{I\subsetneq I(\phi)}|a_I||I|\Big)^p\\ &\leq C\sum_{\tilde{I}}|\tilde{I}|^{(-\alpha+1/p-s)p}\sum_{I\subset\tilde{I}}|a_I|^p|I|^{sp}\\ &= C\sum_I|a_I|^p|I|^{sp}\sum_{\tilde{I}\supset I}|\tilde{I}|^{(-\alpha+1/p-s)p}\\ &\leq C\sum_I|a_I|^p|I|^{(-\alpha+1/p)p}\end{aligned} \tag{B.41}$$

or in other words, $\|f\|_{\alpha,p}^{(1)}\leq C\|f\|_{\alpha,p}^{(4)}$, as desired. □

Combining Corollary 8, Corollary 9, Corollary 10, Corollary 11, Proposition 2, Proposition 3 and Proposition 4, we have completed the proof of Theorem 8.

B.2. Proof of Theorem 5

We now turn to the proof of Theorem 5 on the equivalence of the two-dimensional Besov norms.

Lemma 14. *For $p>0$, there is a constant $C=C(B_L,B_U,p)$ such that $\omega_{R,p}(f)\leq Cv_{R,p}(f)$ for all functions f on $X\times Y$ and rectangles $R=I\times J$.*

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

Proof. We apply Lemma 9, the corresponding result for one space, twice. For fixed x, consider $f(x,y) - m_{X,I}f(y)$ as a function of y. By Lemma 9, we have

$$\frac{1}{|J|}\int_J\int_J |f(x,y) - m_{X,I}f(y) - f(x,y') + m_{X,I}f(y')|^p dy dy'$$

$$\leq C\int_J |f(x,y) - m_{X,I}f(y) - m_{Y,J}f(x) + m_R f|^p dy. \tag{B.42}$$

Now we fix y and y' and consider $f(x,y) - f(x,y')$ as a function of x. Applying Lemma 9 again to this function yields

$$\frac{1}{|I|}\int_I\int_I |f(x,y) - f(x,y') - f(x',y) + f(x',y')|^p dx dx'$$

$$\leq C\int_I |f(x,y) - f(x,y') - m_{X,I}f(y) + m_{X,I}f(y')|^p dx \tag{B.43}$$

Putting (B.42) and (B.43) together gives the result. □

Corollary 12. *For $p > 0$ and $\alpha > 0$, there is a constant $C = C(B_L, B_U, p)$ such that $\|f\|_{\alpha,p} \leq C\|f\|^{(0)}_{\alpha,p}$ for all functions f on X.*

Lemma 15. *For $p \geq 1$, there is a constant $C = C(B_L, B_U, p)$ such that $v_{I,p}(f) \leq C\omega_{I,p}(f)$ for all functions f on X.*

Proof. As in the proof of Lemma 14, applying Lemma 10 to the functions $x \mapsto f(x,y) - f(x,y')$ and $y \mapsto f(x,y) - m_{X,I}f(y)$ gives the result. □

Corollary 13. *For $p \geq 1$ and $\alpha > 0$, there is a constant $C = C(B_L, B_U, p)$ such that $\|f\|^{(0)}_{\alpha,p} \leq C\|f\|_{\alpha,p}$ for all functions f on X.*

Lemma 16. *For any function f on $X \times Y$, rectangle $R = I \times J$, $I \in \mathcal{T}_X$ and $J \in \mathcal{T}_Y$, and any $p > 0$,*

$$\|\Delta_R\|_p^p = \sum_{\tilde{I}\in sub(I)}\sum_{\tilde{J}\in sub(J)} |\delta_{\tilde{I}\times\tilde{J}}(f)|^p |\tilde{I}|^p |\tilde{J}|^p. \tag{B.44}$$

Proof. This follows by repeated application of the corresponding one-dimensional result, Lemma 11. By Lemma 11, we have

$$\int |\Delta_{X,I}\Delta_{Y,J}f(x,y)|^p dx = \sum_{\tilde{I}\in \mathrm{sub}(I)} |\hat{\delta}_{\tilde{I}}\Delta_{Y,J}f(y)|^p |\tilde{I}|. \tag{B.45}$$

Integrating out in y and applying Lemma 11 again then yields

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\begin{aligned}\int\int |\Delta_{X,I}\Delta_{Y,J} f(x,y)|^p dxdy &= \int \sum_{\tilde{I}\in \mathrm{sub}(I)} |\hat{\delta}_{\tilde{I}}\Delta_{Y,J} f(y)|^p |\tilde{I}| dy \\ &= \sum_{\tilde{I}\in \mathrm{sub}(I)} \int |\Delta_{Y,J}\hat{\delta}_{\tilde{I}} f(y)|^p |\tilde{I}| dy \\ &= \sum_{\tilde{I}\in \mathrm{sub}(I)} \sum_{\tilde{J}\in \mathrm{sub}(J)} |\hat{\delta}_{\tilde{J}}\hat{\delta}_{\tilde{I}} f|^p |\tilde{I}||\tilde{J}| \\ &= \sum_{\tilde{I}\in \mathrm{sub}(I)} \sum_{\tilde{J}\in \mathrm{sub}(J)} |\delta_{\tilde{I}\times\tilde{J}} f|^p |\tilde{I}||\tilde{J}| \end{aligned} \tag{B.46}$$

which is the desired result. □

Corollary 14. *The norms $\|f\|^{(2)}_{\alpha,p}$ and $\|f\|^{(3)}_{\alpha,p}$ are equivalent.*

Proposition 5. *The norms $\|f\|^{(3)}_{\alpha,p}$ and $\|f\|^{(4)}_{\alpha,p}$ are equivalent.*

Proof. Since we can write

$$f = \sum_{I\neq X, J\neq Y} \delta_{I\times J}(f)\chi_{I\times J}(x,y) + \sum_{I\neq X} \delta_I(f)\chi_I(x) + \sum_{J\neq Y} \delta_J(f)\chi_J(y) + \int_{X\times Y} f, \tag{B.47}$$

the inequality $\|f\|^{(4)}_{\alpha,p} \leq \|f\|^{(3)}_{\alpha,p}$ is immediate.

For the other direction, take any expansion of f as a linear combination of indicator functions of rectangles: $f = \sum_R a_R\chi_R$. Fix any folder I, and define the function

$$g_I(y) = \sum_{J\in\mathcal{T}_Y} a_{I\times J}\chi_J(y). \tag{B.48}$$

By Proposition 2, we have the inequality

$$\sum_{J\neq Y} |\delta_J(g)|^p |J|^{(-\alpha+1/p)p} \leq C \sum_{J\neq Y} |a_{I\times J}|^p |J|^{(-\alpha+1/p)p} \tag{B.49}$$

Now, we have

$$\hat{\delta}_J f(x) = \sum_{I\in\mathcal{T}_X} \delta_J(g_I)\chi_I(x) \tag{B.50}$$

and so another application of Proposition 2 yields

$$\begin{aligned}\sum_{I\neq X} |\delta_{I\times J}(f)|^p |I|^{(-\alpha+1/p)p} &= \sum_{I\neq X} |\delta_I(\hat{\delta}_J f)|^p |I|^{(-\alpha+1/p)p} \\ &\leq C \sum_{I\neq X} |\delta_J(g_I)|^p |I|^{(-\alpha+1/p)p} \end{aligned} \tag{B.51}$$

Combining (B.49) and (B.51) yields the desired result. □

Lemma 17. *There is a constant $C = C(B_L, B_U, p)$ such that*

$$\|\Delta_R f\|_p^p \leq C \sum_{\Phi: R(\Phi)=R} |\langle f,\Phi\rangle|^p |R|^{1-p/2} \tag{B.52}$$

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

for all functions f on $X \times Y$. Note that the sum is over all tensor Haar functions $\phi(x)\psi(y)$ (non-constant) so that $I(\phi) \times J(\psi) = R$.

Proof. We prove this by repeated application of the corresponding one-dimensional result, Lemma 13. Suppose $R = I \times J$. We fix an arbitrary y and consider $\Delta_{Y,J} f(x, y)$ as a function of x. Lemma 13 then gives

$$\int |\Delta_{X,I}\Delta_{Y,J} f(x,y)|^p dx \leq C \sum_{\phi: I(\phi)=I} |\langle \Delta_{Y,J} f(\cdot, y), \phi\rangle|^p |I|^{1-p/2}. \tag{B.53}$$

Now fix any ϕ with $I(\phi) = I$ and consider the function $y \mapsto \langle f(\cdot, y), \phi\rangle$. From Lemma 13 applied to this function, we observe:

$$\begin{aligned}\int |\langle \Delta_{Y,J} f(\cdot, y), \phi\rangle|^p dy &= \int |\Delta_{Y,J} \langle f(\cdot, y), \phi\rangle|^p dy \\ &\leq C \sum_{\psi: J(\psi)=J} |\langle \langle f, \phi\rangle, \psi\rangle|^p |J|^{1-p/2} \\ &= C \sum_{\psi: J(\psi)=J} |\langle f, \phi\psi\rangle|^p |J|^{1-p/2}. \end{aligned} \tag{B.54}$$

Combining (B.53) and (B.54) yields the desired result. □

Corollary 15. *There is a constant $C = C(B_L, B_U, p)$ such that $\|f\|^{(2)}_{\alpha,p} \leq C\|f\|^{(1)}_{\alpha,p}$ for all functions f on $X \times Y$.*

Proposition 6. *There is a constant $C = C(B_L, B_U, \alpha, p)$ such that $\|f\|^{(1)}_{\alpha,p} \leq C\|f\|^{(2)}_{\alpha,p}$ for all functions f on $X \times Y$.*

Proof. The proof is nearly identical to that of Proposition 7 below; we leave the details to the reader. □

We complete the proof that all norms are equivalent by showing $\|f\|^{(0)}_{\alpha,p} \simeq \|f\|^{(1)}_{\alpha,p}$.

Proposition 7. *The norms $\|f\|^{(0)}_{\alpha,p}$ and $\|f\|^{(1)}_{\alpha,p}$ are equivalent.*

Proof. As usual, we reduce to the corresponding statement for the one-dimensional norms. Fix a Haar function ϕ on X let $g_\phi = \langle f(\cdot, y), \phi\rangle$. Observe that

$$\begin{aligned} v_{J,p}(g_\phi)^p &= \int_J |\langle f(\cdot, y), \phi\rangle - m_J\langle f, \phi\rangle|^p dy \\ &= \int_J |\langle f(\cdot, y) - m_{Y,J}(f), \phi\rangle|^p dy. \end{aligned} \tag{B.55}$$

Then Theorem 8 implies

$$\begin{aligned} \sum_\psi |J(\psi)|^{(-\alpha-1/2+1/p)p} |\langle g_\phi, \psi\rangle|^p &\simeq \sum_{J \neq Y} |J|^{-\alpha p} |v_{p,J}(g_\phi)|^p \\ &= \sum_{J \neq Y} |J|^{-\alpha p} \int_J |\langle f(\cdot, y) - m_{Y,J}(f), \phi\rangle|^p dy. \end{aligned} \tag{B.56}$$

Now, again by Theorem 8 we have

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

$$\sum_{\phi} |I(\phi)|^{(-\alpha-1/2+1/p)p} |\langle f(\cdot,y) - m_{Y,J}(f), \phi\rangle|^p$$

$$\simeq \sum_{I\neq X} |I|^{-\alpha p} \int_I |f(x,y) - m_{Y,J} f(x) - m_{X,I} f(y) + m_{I\times J}(f)|^p dx. \tag{B.57}$$

Putting (B.56) and (B.57) together, we get:

$$\begin{aligned}
\|f\|^{(1)}_{\alpha,p} &= \sum_{\phi}\sum_{\psi} (|I(\phi)||J(\psi)|)^{(-\alpha-1/2+1/p)p} |\langle f, \phi\psi\rangle|^p \\
&= \sum_{\phi} |I(\phi)|^{(-\alpha+1/2)p} \sum_{\psi} |J(\psi)|^{(-\alpha-1/2+1/p)p} |\langle g_\phi, \psi\rangle|^p \\
&\simeq \sum_{\phi} |I(\phi)|^{(-\alpha-1/2+1/p)p} \sum_{J\neq Y} |J|^{-\alpha p} \int_J |\langle f(\cdot,y) - m_{Y,J}(f), \phi\rangle|^p dy \\
&= C \sum_{J\neq Y} |J|^{-\alpha p} \int_J \sum_{\phi} |I(\phi)|^{(-\alpha-1/2+1/p)p} |\langle f(\cdot,y) - m_{Y,J}(f), \phi\rangle|^p dy \\
&\simeq \sum_{J\neq Y} |J|^{-\alpha p} \int_J \sum_{I\neq X} |I|^{-\alpha p} \int_I |f(x,y) - m_{Y,J} f(x) - m_{X,I} f(y) + m_{I\times J}(f)|^p dx dy \\
&= C\|f\|^{(0)}_{\alpha,p}
\end{aligned} \tag{B.58}$$

which is the desired result. □

This completes the proof of Theorem 5.

References

[1] Y. Meyer, Oscillating Patterns in Image Processing and Nonlinear Evolution Equations, Univ. Lecture Ser., vol. 22, American Mathematical Society, 2001.
[2] Y. Meyer, Wavelets and Operators, Cambridge University Press, 1992.
[3] S. Mallat, A Wavelet Tour of Signal Processing, 2nd edition, Academic Press, 1999.
[4] P. Wojtaszcyk, A Mathematical Introduction to Wavelets, Cambridge University Press, 1997.
[5] M. Gavish, B. Nadler, R.R. Coifman, Multiscale wavelets on trees, graphs and high dimensional data: theory and applications to semi supervised learning, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 367–374.
[6] P.P. Petrushev, V.A. Popov, Rational Approximation of Real Functions, Cambridge University Press, 1987.
[7] H. Triebel, Theory of Function Spaces II, Birkhäuser Verlag, 1992.
[8] G.I. Allen, R. Tibshirani, Transposable regularized covariance models with an application to missing data imputation, Ann. Appl. Stat. 4 (2) (2010) 764–790.
[9] G.I. Allen, R. Tibshirani, Inference with transposable data: modelling the effects of row and column correlations, J. R. Stat. Soc. Ser. B. Stat. Methodol. 74 (4) (2012) 721–743.
[10] J.A. Hartigan, Direct clustering of a data matrix, J. Amer. Statist. Assoc. 67 (337) (1972) 123–129.
[11] I.S. Dhillon, S. Mallela, D.S. Modha, Information-theoretic co-clustering, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 89–98.
[12] I.S. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001, pp. 269–274.
[13] M.H. Neumann, Multivariate wavelet thresholding in anisotropic function spaces, Statist. Sinica 10 (2) (2000) 399–432.
[14] M.H. Neumann, R. von Sachs, Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra, Ann. Statist. 25 (1) (1997) 38–76.
[15] S.A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Dokl. Akad. Nauk SSSR 4 (1963) 240–243.
[16] T. Gerstner, M. Griebel, Numerical integration using sparse grids, Numer. Algorithms 18 (1998) 209–232.
[17] J.-A. Strömberg, Computation with wavelets in higher dimensions, Doc. Math. 3 (1998) 523–532.
[18] D.L. Donoho, I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, J. Amer. Statist. Assoc. 90 (432) (1995) 1200–1224.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003

[19] D.L. Donoho, I.M. Johnstone, Neo-classical minimax problems, thresholding and adaptive function estimation, Bernoulli 2 (1) (1996) 39–62.
[20] D.L. Donoho, De-noising by soft-thresholding, IEEE Trans. Inform. Theory 41 (3) (1995) 613–627.
[21] D.L. Donoho, I.M. Johnstone, Wavelet shrinkage: asymptotia?, J. R. Stat. Soc. Ser. B. Stat. Methodol. 57 (2) (1995) 301–369.
[22] D.L. Donoho, I.M. Johnstone, Minimax estimation via wavelet shrinkage, Ann. Statist. 26 (3) (1998) 879–921.
[23] E.M. Stein, Singular Integrals and Differentiability Properties of Functions, Princeton University Press, 1970.
[24] D.G. Deng, Y. Han, Harmonic Analysis on Spaces of Homogeneous Type, Springer, 2009.
[25] Y. Meyer, R. Coifman, Wavelets: Calderón–Zygmund and Multilinear Operators, Cambridge University Press, 1997.
[26] P. Auscher, T. Coulhon, Riesz transform on manifolds and Poincaré inequalities, Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) 4 (5) (2005) 531–555.
[27] P. Auscher, On L^p estimates for square roots of second order elliptic operators on $\mathbb{R}^n$, Publ. Math. 48 (2004) 159–186.
[28] M. Gavish, R.R. Coifman, Sampling, denoising and compression of matrices by coherent matrix organization, Appl. Comput. Harmon. Anal. 33 (3) (2012) 354–369.
[29] Y. Bartal, Probabilistic approximation of metric spaces and its algorithmic applications, in: 37th Annual Symposium on Foundations of Computer Science, IEEE, 1996, pp. 184–193.
[30] Y. Bartal, On approximating arbitrary metrics by tree metrics, in: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, ACM Press, 1998, pp. 161–168.
[31] J. Fakcharoenphol, S. Rao, K. Talwar, A tight bound on approximating arbitrary metrics by tree metrics, in: Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing, ACM, 2003, pp. 448–455.
[32] M. Gavish, R. Coifman, Harmonic analysis of digital databases, in: J. Cohen, A.I. Zayed (Eds.), Wavelets and Multiscale Analysis, Birkhäuser, 2011, pp. 161–197.
[33] A. Antoniadis, F. Leblanc, Nonparametric wavelet regression for binary response, Statistics 34 (3) (2000) 183–213.
[34] A.P. Korostelev, A.B. Tsybakov, Minimax Theory of Image Reconstruction, Springer, 1993.
[35] F.R. Hampel, The influence curve and its role in robust estimation, J. Amer. Statist. Assoc. 69 (346) (1974) 383–393.
[36] F.R. Hampel, E.M. Ronchetti, P. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, Vol. 114, John Wiley & Sons, 2011.
[37] P. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, J. Amer. Statist. Assoc. 88 (424) (1993) 1273–1283.
[38] R. Kannan, S. Vempala, A. Vetta, On clusterings: good, bad and spectral, J. ACM 51 (2004) 497–515.
[39] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 888–905.
[40] B. Nadler, M. Galun, Fundamental limitations of spectral clustering, in: Advances in Neural Information Processing Systems, 2006, pp. 1017–1024.
[41] R.R. Coifman, D.L. Donoho, Translation-invariant de-noising, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, Springer, 1995, pp. 125–150.
[42] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1) (1998) 33–61.
[43] E.J. Candès, M.B. Wakin, S.P. Boyd, Enhancing sparsity by reweighted l_1 minimization, J. Fourier Anal. Appl. 14 (2008) 877–905.
[44] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, http://cvxr.com/cvx, March 2014.
[45] M. Grant, S. Boyd, Graph implementations for nonsmooth convex programs, in: V. Blondel, S. Boyd, H. Kimura (Eds.), Recent Advances in Learning and Control, in: Lecture Notes in Control and Inform. Sci., Springer-Verlag Limited, 2008, pp. 95–110.
[46] J.I. Ankenman, Geometry and Analysis of Dual Networks on Questionnaires, Ph.D. thesis, Yale University, May 2014.
[47] [link]. https://github.com/hgfalling/pyquest.
[48] C. Villani, Topics in Optimal Transportation, American Mathematical Society, 2003.
[49] M.S. Charikar, Similarity estimation techniques from rounding algorithms, in: Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing, 2002, pp. 380–388.
[50] W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008, in press.
[51] G. Mishne, R. Talmon, R. Meir, J. Schiller, U. Dubin, R.R. Coifman, Hierarchical coupled geometry analysis for neuronal structure and activity pattern discovery, arXiv:1511.02086.

Please cite this article in press as: J. Ankenman, W. Leeb, Mixed Hölder matrix discovery via wavelet shrinkage and Calderón–Zygmund decompositions, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.01.003