The Annals of Applied Statistics
2015, Vol. 9, No. 2, 1076–1101
DOI: 10.1214/15-AOAS829

WAVELET-DOMAIN REGRESSION AND PREDICTIVE INFERENCE IN PSYCHIATRIC NEUROIMAGING

BY PHILIP T. REISS[1,*,†], LAN HUO[2,*], YIHONG ZHAO[*], CLARE KELLY[2,*] AND R. TODD OGDEN[3,‡]

New York University, Nathan S. Kline Institute for Psychiatric Research*† *and Columbia University*‡

An increasingly important goal of psychiatry is the use of brain imaging data to develop predictive models. Here we present two contributions to statistical methodology for this purpose. First, we propose and compare a set of wavelet-domain procedures for fitting generalized linear models with scalar responses and image predictors: sparse variants of principal component regression and of partial least squares, and the elastic net. Second, we consider assessing the contribution of image predictors over and above available scalar predictors, in particular, via permutation tests and an extension of the idea of confounding to the case of functional or image predictors. Using the proposed methods, we assess whether maps of a spontaneous brain activity measure, derived from functional magnetic resonance imaging, can meaningfully predict presence or absence of attention deficit/hyperactivity disorder (ADHD). Our results shed light on the role of confounding in the surprising outcome of the recent ADHD-200 Global Competition, which challenged researchers to develop algorithms for automated image-based diagnosis of the disorder.

1. Introduction. A major goal of current psychiatric neuroimaging research is to predict clinical outcomes on the basis of quantitative image data. Many studies have focused on "predicting" current disease states from brain images [e.g., Craddock et al. (2009), Sun et al. (2009)]. While seemingly less difficult than accurate prediction of *future* outcomes, the goal of clinically useful imaging-based diagnosis has proved highly challenging [Honorio et al. (2012), Kapur, Phillips and Insel (2012)].

This paper addresses two important limitations of standard methods for using brain images to predict psychiatric outcomes:

Received August 2013; revised February 2015.

[1]Supported in part by NSF Grant DMS-09-07017 and National Institutes of Health Grants 5R01EB009744-03 and 1R01MH095836-01A1.

[2]Supported in part by National Institutes of Health Grant 1R01MH095836-01A1.

[3]Supported in part by National Institutes of Health Grant 5R01EB009744-03.

Key words and phrases. ADHD-200, elastic net, functional confounding, functional magnetic resonance imaging, functional regression, sparse principal component regression, sparse partial least squares.

(i) Ordinarily, the voxels (volume units) of the brain are treated as interchangeable predictors or "features." Accuracy might be improved by properly exploiting the spatial arrangement of the brain.

(ii) In some cases brain images may prove successful for diagnostic classification, but only because the images are related to one or more scalar covariates that drive the association. This is a nonstandard form of confounding, and there seems to be no existing methodology for detecting it. In other words, little is known about how to assess whether image data offers "added value" for prediction, beyond what is available from nonimage data—which will typically be much simpler to acquire.

To address limitation (i), we approach the general problem as one of regressing scalar responses on *image predictors*, which are viewed, as in Reiss (2006) and Reiss and Ogden (2010), as a challenging special case of functional predictors [Ramsay and Silverman (2005)]. The responses $y_1, \ldots, y_n$ are assumed to be generated independently by the model

$$y_i \sim EF(\mu_i, \phi), \tag{1}$$

$$g(\mu_i) = \mathbf{t}_i^T \boldsymbol{\delta} + \int_{\mathcal{S}} x_i(\mathbf{s})\beta(\mathbf{s})\, d\mathbf{s}. \tag{2}$$

Here $EF(\mu_i, \phi)$ denotes an exponential family distribution with mean μ_i and scale parameter ϕ, along with a link function g; $\mathbf{t}_i$ is an m-dimensional vector of (scalar) covariates, of which the first is the constant 1; $x_i : \mathcal{S} \longrightarrow \mathbb{R}$ is a functional predictor with domain $\mathcal{S} \subset \mathbb{R}^2$ or $\subset \mathbb{R}^3$; and the corresponding effect, the *coefficient function* or *coefficient image* $\beta : \mathcal{S} \longrightarrow \mathbb{R}$, is the parameter of interest. The simplest special case is the linear model

$$y_i = \mathbf{t}_i^T \boldsymbol{\delta} + \int_{\mathcal{S}} x_i(\mathbf{s})\beta(\mathbf{s})\, d\mathbf{s} + \varepsilon_i, \tag{3}$$

where the ε_i are independent and identically distributed errors with mean 0 and variance $\sigma^2 (= \phi)$. When $\mathbf{t}_i \equiv 1$ (i.e., no scalar covariates), model (3) is the extension, from one-dimensional to multidimensional predictors, of the functional linear model that has been studied by Marx and Eilers (1999), Cardot, Ferraty and Sarda (1999), Müller and Stadtmüller (2005), Ramsay and Silverman (2005), Hall and Horowitz (2007), Reiss and Ogden (2007), Goldsmith et al. (2011) and many others.

For the case of one-dimensional functional predictors, a popular way to take spatial information into account is to restrict $\beta(\cdot)$ to the span of a spline basis [e.g., Marx and Eilers (1999)]. Spline methods for two-dimensional predictors have been studied by Marx and Eilers (2005) and Guillas and Lai (2010), and by Reiss and Ogden (2010), whose work was motivated by neuroimaging applications.

Some more recent work has considered neuroimaging applications with two- and three-dimensional predictors [Goldsmith, Huang and Crainiceanu (2014), Huang et al. (2013), Zhou, Li and Zhu (2013)]. In this paper, we propose a set

of new approaches based on a wavelet representation of the coefficient image. The idea of transforming the images to the wavelet domain has previously appeared in the brain mapping literature, where it is customary to fit separate models at each voxel, with the image-derived quantity regressed on demographic or clinical variables of interest [e.g., Ruttimann et al. (1998), Van De Ville et al. (2007)]. But for our objective of using entire images in a *single* model to predict a scalar response, working in the wavelet domain has been mentioned as a natural idea [Grosenick et al. (2013)] but rarely if ever pursued, at least until the very recent work of Wang et al. (2014). Unlike spline bases, wavelet bases are designed for sparse representation and yield estimates that adapt to the features of the coefficient image.

Limitation (ii) was highlighted by the results of the recent ADHD-200 Global Competition for automated diagnosis of attention deficit/hyperactivity disorder [ADHD-200 Consortium (2012)]. Teams were provided with functional magnetic resonance images from ADHD subjects and controls on which to train diagnostic algorithms, and then applied these algorithms to predict diagnosis in a separate set of images. A team of biostatisticians from Johns Hopkins University, whose methods are described by Eloyan et al. (2012), achieved the highest score for correct imaging-based classification and were declared the winners. But a team from the University of Alberta, which discarded the images and used just four scalar predictors [age, sex, handedness and IQ; see Brown et al. (2012)], attained a slightly higher classification score [see Caffo et al. (2012) for related discussion].

To address limitation (ii), we test the effect of image predictors via a permutation-based approach originally proposed in the neuroimaging literature [Golland and Fischl (2003)], which we extend to allow for scalar covariates. We also consider how to extend the traditional notion of confounding to settings with both scalar and image predictors. These ideas are illustrated using our wavelet methods, but are not specific to them; rather, they are applicable with other approaches to functional or high-dimensional regression.

Our contributions can be summarized as follows: (i) We propose novel wavelet-domain methodology for regression with image predictors. While Wang et al. (2014) and Zhao, Chen and Ogden (2015) have studied the wavelet-domain lasso for image predictors, we also propose and compare several other methods, and consider the generalized linear case and the role of scalar covariates. (ii) We extend predictive performance-based hypothesis testing [Golland and Fischl (2003)] to the case where scalar confounders are present, providing a new way to assess the usefulness of image-based prediction.

In Section 2 we introduce wavelet bases, and motivate and outline a general template for scalar-on-image regression in the wavelet domain. Section 3 describes three specific algorithms, which are evaluated in simulations in Section 4. Section 5 considers hypothesis testing and confounding with image predictors. In Section 6 the proposed methods are applied to a portion of the ADHD-200 data set, and the results point to a possible role of confounding in the competition's surprising result. Section 7 offers a concluding discussion.

2. Wavelets and their use in regression on images.

2.1. *A brief introduction to wavelet basis representations.* Wavelet bases are a popular way to obtain a sparse representation for functional data, in particular, when the degree of smoothness exhibits local variation [see Nason (2008), Ogden (1997), Vidakovic (1999) for statistically-oriented treatments]. A wavelet basis for $L^2(\mathbb{R})$ is constructed from a scaling function (or "father wavelet") ϕ and a wavelet function ("mother wavelet") ψ [see Figure 1(a)–(b)], with the following properties:

- For each $j \in \mathbb{Z}$, the shifted and dilated functions $\{\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k) : k \in \mathbb{Z}\}$ form an orthonormal basis for V_j, where $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$ is a nested sequence of subspaces whose union is a dense subspace of $L^2(\mathbb{R})$.
- For each $j \in \mathbb{Z}$, $\{\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) : k \in \mathbb{Z}\}$ form an orthonormal basis for a "detail space" W_j satisfying $V_{j+1} = V_j \oplus W_j$.

Hence, for any integer $j_0 \geq 0$, $V_{j_0} \oplus W_{j_0} \oplus W_{j_0+1} \oplus \cdots$ is a dense subspace of $L^2(\mathbb{R})$.

Given appropriate boundary handling, such as modifying the scaling and wavelet functions to be periodic, one can likewise construct orthonormal wavelet bases for $L^2[0, 1]$, of the form

$$\underbrace{\{\phi_{j_0,0}, \ldots, \phi_{j_0,2^{j_0}-1}\}}_{\in V_{j_0}} \cup \underbrace{\{\psi_{j_0,0}, \ldots, \psi_{j_0,2^{j_0}-1}\}}_{\in W_{j_0}} \cup \underbrace{\{\psi_{j_0+1,0}, \ldots, \psi_{j_0+1,2^{j_0+1}-1}\}}_{\in W_{j_0+1}} \cup \cdots$$

—that is, 2^{j_0} scaling functions (corresponding to the large-scale features of the data), 2^{j_0} wavelet functions at level j_0, 2^{j_0+1} wavelet functions at level $j_0 + 1$ and so on, with higher wavelet levels capturing finer-scale details. This multiscale structure is what makes wavelet bases so useful for sparse representation of functions with varying degrees of smoothness.

The wavelet decomposition level j_0 acts as a tuning parameter. A small j_0 implies that a small number (2^{j_0}) of scaling functions are used to construct the macro

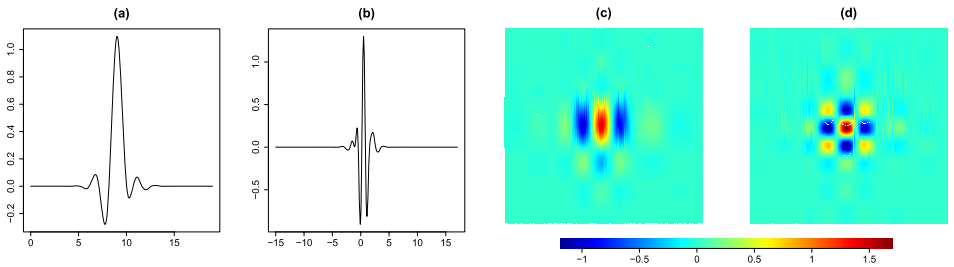

FIG. 1. (a) *Scaling function* ϕ *and* (b) *wavelet function* ψ *for* 1*D Daubechies* (1988) *"least-asymmetric" wavelets with* 10 *vanishing moments.* 2*D basis functions are formed from tensor products such as* (c) $(x, y) \mapsto \psi(x)\phi(y)$ *and* (d) $(x, y) \mapsto \psi(x)\psi(y)$.

features of the function, with most of the basis elements dedicated to providing detail at a variety of scales. A large j_0 allows for many more scaling functions, each at higher resolution, and thus fewer basis elements corresponding to detail.

In practice, a function $f \in L^2[0,1]$ is observed at finitely many points, ordinarily taken to be the $N = 2^J$ (for some positive integer J) equally spaced points $0, \frac{1}{N}, \ldots, \frac{N-1}{N}$. (When the function is observed at a number of points that is not a power of 2, one can insert zeroes before and after to attain the next highest power of 2.) The observed values can then be interpolated by the N-dimensional truncated basis

$$\{\phi_{j_0,0}, \ldots, \phi_{j_0,2^{j_0}-1}\} \cup \{\psi_{j_0,0}, \ldots, \psi_{j_0,2^{j_0}-1}\} \cup \cdots \cup \{\psi_{J-1,0}, \ldots, \psi_{J-1,2^{J-1}-1}\}.$$

The discrete wavelet transform (DWT), implemented by the $O(N)$ pyramid algorithm of Mallat (1989), expands f with respect to this basis. Given a judicious choice of ϕ and ψ, signals of varying smoothness can be well represented with a small number of coefficients. Throughout this paper we use the compactly supported Daubechies (1988) "least-asymmetric" wavelets with 10 vanishing moments, displayed in Figure 1.

Wavelet bases for two dimensions can be constructed by taking tensor products of the ϕ and ψ functions. The two-dimensional scaling function is $\phi(x)\phi(y)$ and there are three types of 2D wavelets: $\phi(x)\psi(y)$, $\psi(x)\phi(y)$ and $\psi(x)\psi(y)$, roughly corresponding to "horizontal," "vertical" and "diagonal" detail, respectively [see Figure 1(c)–(d)]. These functions are dilated and translated just as their 1D counterparts are. Wavelet bases for 3D are constructed similarly. Morris et al. (2011) discuss alternative wavelet transforms for images that are not constructed as tensor products.

2.2. *A meta-algorithm for scalar-on-image regression.* Henceforth, the functional predictor $x_i(\cdot)$ of (2), (3) will be replaced by the ith discretized image observation $\mathbf{x}_i = (x_1, \ldots, x_N)^T \equiv [x_i(s_1), \ldots, x_i(s_N)]^T$, where $s_1, \ldots, s_N \in \mathcal{S}$ are distinct spatial locations at which the function x_i is measured. Often, in practice, each image is given as a matrix or 3D array; $\mathbf{x}_i$ is then obtained by converting this into a vector. From now until Section 3.5 we focus on the linear model (3), which can now be written in matrix form as

$$\mathbf{y} = \mathbf{T}\boldsymbol{\delta} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{4}$$

Here $\mathbf{y} = (y_1, \ldots, y_n)^T$; $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$; $\mathbf{T}$ is the $n \times m$ matrix with ith row $\mathbf{t}_i^T$; $\mathbf{X}$ is the $n \times N$ matrix with ith row $\mathbf{x}_i^T$; and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^T$ is a similarly discretized version of the coefficient image β. More precisely, for $j = 1, \ldots, N$, $\beta_j = w_j \beta(s_j)$, where the w_j's are quadrature weights such that $\mathbf{x}_i^T \boldsymbol{\beta}$ is a good approximation to the integral in (3); but for image data, $s_1, \ldots, s_N$ typically form an equally spaced grid, so these weights are taken as constant and hence ignored. With these definitions, (3) is just the ith of the n equations that make up the vector equation (4).

To simplify the notation, we shall use a single subscript and denote the wavelet basis functions for a given j_0 as $\{\psi_1, \psi_2, \ldots, \psi_N\}$. The wavelet representation of the ith observed image x_i is $x_i(s) = \sum_{k=1}^{N} \tilde{x}_{ik}\psi_k(s)$, in which the *wavelet coefficients* are given by $\tilde{x}_{ik} = \langle x_i, \psi_k \rangle$. The coefficient vector $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{iN})^T$ can be written as $\tilde{\mathbf{x}}_i = \mathcal{W}\mathbf{x}_i$, where $\mathcal{W}$ is an $N \times N$ orthonormal matrix (which is not formed explicitly when $\tilde{\mathbf{x}}_i$ is computed by the DWT). Similarly the discretized coefficient function $\boldsymbol{\beta}$ can be represented in terms of its wavelet coefficients as $\tilde{\boldsymbol{\beta}} = \mathcal{W}\boldsymbol{\beta}$, leading to the wavelet-domain version of model (4):

$$\begin{aligned} \mathbf{y} &= \mathbf{T}\boldsymbol{\delta} + \mathbf{X}\mathcal{W}^T\mathcal{W}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{T}\boldsymbol{\delta} + \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}, \end{aligned} \tag{5}$$

where $\tilde{\mathbf{X}}$ is the $n \times N$ matrix with ith row $\tilde{\mathbf{x}}_i^T$.

The key point is that the wavelet-domain form (5) is better suited than the original form (4) for applying sparse techniques for high-dimensional regression—both because wavelet bases are designed for sparse representation of images [Mallat (2009)] and because the DWT approximately decorrelates or "whitens" data [Vidakovic (1999)]. We can thus formulate a "meta-algorithm" for scalar-on-image regression in the wavelet domain:

1. Apply the DWT to the image predictors to transform model (4) into model (5).
2. Use some high-dimensional regression methodology to derive a sparse estimate $\hat{\tilde{\boldsymbol{\beta}}}$.
3. Apply the inverse DWT to $\hat{\tilde{\boldsymbol{\beta}}}$ to obtain a coefficient image estimate $\hat{\boldsymbol{\beta}}$ for the original model (4).

Different choices for step 2 lead to specific algorithms, as described in the next section.

The above general scheme can be extended to multiple image predictors [cf. Zhu, Vannucci and Cox (2010)]. We note that this meta-algorithm has been applied before for 1D functional predictors [Brown, Fearn and Vannucci (2001), Malloy et al. (2010), Wang, Ray and Mallick (2007), Zhao, Ogden and Reiss (2012)] and more for image predictors [Wang et al. (2014), Zhao, Chen and Ogden (2015)]. Past work on wavelet-domain classification, as opposed to regression [e.g., Berlinet, Biau and Rouvière (2008), Chang, Chen and Ogden (2014), Zhu, Brown and Morris (2012)], may bear comparison to our proposed methods. Morris et al. (2011) develop wavelet-domain functional mixed models with images as *responses*.

3. Three wavelet-domain algorithms.

3.1. *Sparse wavelet-domain principal component regression.* The functional linear model (3) is often fitted by assuming the coefficient function has a truncated functional principal component, or Karhunen–Loève, representation $\beta(s) =$

$\sum_{j=1}^{m} c_j \rho_j(s)$, where m is a positive integer and $\rho_1, \rho_2, \ldots, \rho_m$ are the first m eigenfunctions of the covariance operator associated with the predictor functions x_i [e.g., Cai and Hall (2006), Cardot, Ferraty and Sarda (1999), Müller and Stadtmüller (2005)]. The eigenfunctions $\rho_1, \rho_2, \ldots, \rho_m$ can be estimated by viewing the functional predictors as (highly) multivariate data, and applying ordinary principal component analysis to the predictor matrix $\mathbf{X}$.

Here and in Section 3.2, we assume that $\mathbf{X}$ has mean-centered columns, that is, $\mathbf{1}^T\mathbf{X} = \mathbf{0}$. The approach of the previous paragraph then amounts to assuming $\boldsymbol{\beta} = \mathbf{V}_m\boldsymbol{\gamma}$ for some $\boldsymbol{\gamma} \in \mathbb{R}^m$, where $\mathbf{UDV}^T$ is the singular value decomposition of $\mathbf{X}$, and $\mathbf{V}_m$ comprises the leading m columns of $\mathbf{V}$. Hence, estimation reduces to choosing $\boldsymbol{\delta}, \boldsymbol{\gamma}$ to minimize the principal component regression [PCR; Massy (1965)] criterion

$$\|\mathbf{y} - \mathbf{T}\boldsymbol{\delta} - \mathbf{X}\mathbf{V}_m\boldsymbol{\gamma}\|^2. \tag{6}$$

(This is a slightly nonstandard PCR criterion, in that principal component reduction is applied only to $\mathbf{X}$ but not to $\mathbf{T}$. A similar remark applies to the other criteria introduced below.)

As shown by Reiss and Ogden (2007), PCR can be implemented more effectively by exploiting the functional character of the data. In the one-dimensional functional predictor case, this has usually meant forming smooth estimates of the eigenfunctions—as in the FPCR$_C$ method of Reiss and Ogden (2007), which expands the eigenfunctions with respect to a B-spline basis [cf. Cardot, Ferraty and Sarda (2003)]. But for image predictors, local adaptivity—the ability to capture sharp features in some areas vs. a high degree of smoothness elsewhere—becomes particularly important. This motivates using a wavelet basis, rather than a spline basis, to represent the eigenfunctions, or, in other words, developing a wavelet-domain version of PCR as an instance of the meta-algorithm of Section 2.2.

A *non*sparse wavelet-domain PCR estimate would minimize

$$\|\mathbf{y} - \mathbf{T}\boldsymbol{\delta} - \tilde{\mathbf{X}}\tilde{\mathbf{V}}_m\boldsymbol{\gamma}\|^2, \tag{7}$$

which is analogous to (6) but based on the SVD of $\tilde{\mathbf{X}}$ rather than of $\mathbf{X}$. However, the advantage of working in the wavelet domain is to obtain a sparse coefficient estimate by replacing the PC weights $\tilde{\mathbf{V}}_m$ with weights from a sparse version of PCA. Several penalty-based methods have been proposed for sparse PCA [e.g., Shen and Huang (2008), Witten, Tibshirani and Hastie (2009), Zou, Hastie and Tibshirani (2006)], but we opted for the approach of Johnstone and Lu (2009), which is simpler than the penalized methods and, unlike them, was developed with a view toward sparse wavelet representations of signals. Johnstone and Lu (2009) propose to select the features or coordinates with highest variance, and apply PCA only to these. The resulting sparse PCR criterion is

$$\|\mathbf{y} - \mathbf{T}\boldsymbol{\delta} - \tilde{\mathbf{X}}^*\tilde{\mathbf{V}}_m^*\boldsymbol{\gamma}\|^2; \tag{8}$$

here $\tilde{\mathbf{X}}^*$ consists of the c columns of $\tilde{\mathbf{X}}$ having highest variance, and $\tilde{\mathbf{V}}^*_m$ consists of the leading m columns of $\tilde{\mathbf{V}}^*$, where $\tilde{\mathbf{U}}^*\tilde{\mathbf{D}}^*\tilde{\mathbf{V}}^{*T}$ is the SVD of $\tilde{\mathbf{X}}^*$. The minimizer $(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}})$ of (8) can be obtained by simple least squares. The vector of wavelet coefficient estimates is then $\hat{\tilde{\boldsymbol{\beta}}} = \tilde{\mathbf{V}}^*_m\hat{\boldsymbol{\gamma}}$, and the coefficient image estimate $\hat{\boldsymbol{\beta}} = \mathcal{W}^T\hat{\tilde{\boldsymbol{\beta}}}$ is derived by the inverse DWT.

3.2. *Sparse wavelet-domain partial least squares.* Whereas PCR reduces dimension by regressing on the leading PCs of the predictors, partial least squares [PLS; Wold (1966)] works by regressing on a set of components that are relevant to predicting the responses. A (nonsparse) wavelet-domain PLS estimate [cf. Nadler and Coifman (2005)] is derived by minimizing

$$\|\mathbf{y} - \mathbf{T}\boldsymbol{\delta} - \tilde{\mathbf{X}}\tilde{\mathbf{R}}_m\boldsymbol{\gamma}\|^2 \tag{9}$$

[cf. (7)], where the columns of $\tilde{\mathbf{R}}_m$ are defined iteratively as follows [Stone and Brooks (1990)]:

- $\tilde{\mathbf{r}}_1 = \arg\min_{\|r\|=1} \operatorname{Cov}(\mathbf{y}, \tilde{\mathbf{X}}\mathbf{r})$;
- for $j = 2, \ldots, c$,

$$\tilde{\mathbf{r}}_j = \arg \min_{\|r\|=1,\ r^T\tilde{X}^T\tilde{X}r_m=0\ \forall m=1,\ldots,j-1} \operatorname{Cov}(\mathbf{y}, \tilde{\mathbf{X}}\mathbf{r}).$$

Once again, however, the point of working in the wavelet domain is to obtain a sparse estimate. To define sparse wavelet-domain PLS, as with PCR, we could have used penalization to derive sparse PLS components [Chun and Keleş (2010)], but we instead opted to build on the aforementioned approach of Johnstone and Lu (2009) to sparse PCA. A natural PLS analogue of that approach is to select those features $\tilde{x}_j$ whose covariance with y has the greatest magnitude. This results in the sparse PLS criterion

$$\|\mathbf{y} - \mathbf{T}\boldsymbol{\delta} - \tilde{\mathbf{X}}^\dagger\tilde{\mathbf{R}}^\dagger_m\boldsymbol{\gamma}\|^2; \tag{10}$$

here $\tilde{\mathbf{X}}^\dagger$ consists of the c columns of $\tilde{\mathbf{X}}$ having highest covariance with $\mathbf{y}$, and the columns of $\tilde{\mathbf{R}}^\dagger_m$ are defined analogously to those of $\tilde{\mathbf{R}}_m$ in (9). As for PCR, the least-squares minimizer $(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}})$ of (10) leads directly to estimates of the wavelet coefficients $\tilde{\boldsymbol{\beta}}$ and of the resulting coefficient image $\boldsymbol{\beta}$.

Our PLS algorithm is a wavelet-domain counterpart of the spline-based functional PLS procedure denoted by FPLS$_C$ in Reiss and Ogden (2007). We note that Preda and Saporta (2005) and Delaigle and Hall (2012b) have proposed more explicitly functional formulations of PLS, based on covariance operators on function spaces.

3.3. *Wavelet-domain elastic net.* Since wavelet bases are well suited for sparse representation of functions, recent work has considered combining them with sparsity-inducing penalties, both for semiparametric regression [Wand and Ormerod (2011)] and for regression with functional or image predictors [Wang et al. (2014), Zhao, Chen and Ogden (2015), Zhao, Ogden and Reiss (2012)]. The latter papers focused on ℓ_1 penalization, also known as the lasso [Tibshirani (1996)], in the wavelet domain. Alternatives to the lasso include the SCAD penalty [Fan and Li (2001)] and the adaptive lasso [Zou (2006)]. Here we consider the elastic net (EN) estimator for wavelet-domain model (5), which minimizes

$$\|\mathbf{y} - \mathbf{T}\boldsymbol{\delta} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}\|^2 + \lambda\big[\alpha\|\tilde{\boldsymbol{\beta}}\|_1 + (1-\alpha)\|\tilde{\boldsymbol{\beta}}\|_2^2\big] \tag{11}$$

over $(\boldsymbol{\delta}, \tilde{\boldsymbol{\beta}})$, for a regularization parameter $\lambda > 0$ and a mixing parameter $\alpha \in [0, 1]$ which controls the relative strength of the ℓ_1 and ℓ_2 penalties on the coefficients [Zou and Hastie (2005)].

In the original nomenclature of Zou and Hastie (2005), the minimizer of (11) is the "naïve" EN, whereas EN is a rescaled version. Since we shall make use of the generalized linear extension of EN as implemented by Friedman, Hastie and Tibshirani (2010), we follow these authors in omitting the rescaling step. When $\alpha > 0$, the ℓ_1 penalty shrinks small coefficients to zero, leading to a sparse wavelet representation. The wavelet-domain lasso is obtained when $\alpha = 1$. As explained by Zou and Hastie (2005), given a group of important features that are highly correlated, the lasso tends to select just one, whereas EN selects the entire group, which is often preferable—even in the wavelet domain, notwithstanding the "whitening" property of the discrete wavelet transform.

3.4. *Summary*: *Alternative routes to sparsity.* All three of the above methods seek to represent the coefficient image $\beta(\cdot)$ sparsely, as a linear combination of a subset of the wavelet basis functions, but they deploy very different strategies to choose that subset. The ℓ_1 penalty in the elastic net criterion (11) has the effect of shrinking small coefficients to zero. This can be interpreted as imposing a prior that favors a sparse estimate. The PCR criterion (8) eliminates basis elements *before* performing regression, based on an implicit assumption that those basis elements with low variance in the data have little to contribute to the coefficient image. This assumption is broadly consistent, on the one hand, with the assumption of Johnstone and Lu (2009) that such basis elements are merely capturing noise; and, on the other hand, with the underlying assumption of PCR, namely, that the highest-variance principal components are most relevant in regression [see Cook (2007) for some relevant discussion]. The PLS criterion (10) likewise lets the data determine which basis elements to include; but here, instead of considering only the wavelet-transformed image data $\tilde{\mathbf{X}}$ as in PCR, we define relevant components by iteratively maximizing covariance with the responses $\mathbf{y}$.

3.5. *Extension to the generalized linear case.* The above three wavelet-domain algorithms can be straightforwardly extended from linear to generalized linear models (GLMs) of the form

$$g\big[E(\mathbf{y})\big] = \mathbf{T}\boldsymbol{\delta} + \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}, \tag{12}$$

for a link function g, generalizing (5). For PCR, one simply fits a GLM, as opposed to a linear model, to the sparse PCs. For the elastic net, the `glmnet` algorithm of Friedman, Hastie and Tibshirani (2010) is available for the generalized linear case.

PLS is sometimes performed in an iteratively reweighted manner for GLMs [Marx (1996)], but in high-dimensional settings, such algorithms may require nontrivial modification [e.g., Ding and Gentleman (2005)] to avoid convergence problems. Here we view PLS as a generic approach to constructing relevant components, which may be employed beyond the linear regression setting [e.g., Delaigle and Hall (2012a), Nguyen and Rocke (2002)]. Thus, we construct PLS components exactly as we would for a linear model, but then use these components to fit a GLM.

3.6. *Tuning parameter selection.* For wavelet-domain PCR and PLS, three tuning parameters must be selected: the resolution-level parameter j_0; the number c of wavelet coefficients to retain [i.e., the number of columns of $\tilde{\mathbf{X}}^*$ in (8) or of $\tilde{\mathbf{X}}^\dagger$ in (10)]; and the number m of PCs or PLS components. We generally fix $j_0 = 4$, since we have found that resolution level to be generally either optimal or near-optimal as measured by cross-validation (CV). For wavelet-domain elastic net, one must choose j_0 and the two penalty parameters α and λ in (11), but we again prefer to fix $j_0 = 4$.

These tuning parameters are chosen by repeated K-fold CV. In the rth of R repetitions we divide the data points $(y_i, \mathbf{t}_i, \mathbf{x}_i)$ $(i = 1, \ldots, n)$ into K equal-sized validation sets indexed by $I_{r,1}, \ldots, I_{r,K}$. We can then choose the tuning parameters to minimize the CV score

$$\frac{1}{RK}\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{i\in I_{r,k}} L(y_i; \hat{\boldsymbol{\delta}}_{-r,k}, \hat{\tilde{\boldsymbol{\beta}}}_{-r,k}), \tag{13}$$

where $\hat{\boldsymbol{\delta}}_{-r,k}, \hat{\tilde{\boldsymbol{\beta}}}_{-r,k}$ are the estimates that result when model (12) is fitted (by PCR, PLS or EN) with the observations indexed by $I_{r,k}$ excluded, and L is an appropriate loss function. For linear regression the standard loss function is the squared error $L(\mathbf{y}_i; \boldsymbol{\delta}, \tilde{\boldsymbol{\beta}}) = (y_i - \mathbf{t}_i^T\boldsymbol{\delta} - \tilde{\mathbf{x}}_i^T\tilde{\boldsymbol{\beta}})^2$. For the generalized linear case, following Zhu and Hastie (2004), we use the deviance $D(y_i; \boldsymbol{\delta}, \tilde{\boldsymbol{\beta}})$ as the loss function. Specifically for logistic regression, unusually large summands can dominate criterion (13). Therefore, similarly to Chi and Scott (2014), we instead choose the tuning parameters by a robust CV score that takes the median rather than the mean

over each set of K validation sets:

$$\frac{1}{R}\sum_{r=1}^{R}\underset{k\in\{1,\ldots,K\}}{\text{median}}\sum_{i\in I_{r,k}} D(y_i;\hat{\boldsymbol{\delta}}_{-r,k},\hat{\hat{\boldsymbol{\beta}}}_{-r,k}). \tag{14}$$

4. Comparative simulation study. To test the performance of our methods with realistic image predictors, we created a data set based on the positron emission tomography (PET) data previously studied by Reiss and Ogden (2010). That data set included axial slices from 33 amyloid beta maps, from which we extracted a square region of 64×64 voxels. To generate a larger sample of $n = 500$ images, we applied a procedure similar to that of Goldsmith, Huang and Crainiceanu (2014):

1. We estimated the (vectorized) principal components (eigenimages)

$$\hat{\boldsymbol{\rho}}_1,\ldots,\hat{\boldsymbol{\rho}}_{32}\in\mathbb{R}^{64^2},$$

with corresponding eigenvalues $\lambda_1,\ldots,\lambda_{32}$.

2. For $i = 1,\ldots,500$, we generated the ith simulated predictor image as $\mathbf{x}_i = \sum_{j=1}^{32} c_{ij}\hat{\boldsymbol{\rho}}_j$, with the c_{ij}'s simulated independently from the $N(0,\lambda_j)$ distribution.

In step 1 above we used the sparse PCA method of Johnstone and Lu (2009), including the 492 wavelet coefficients having the highest variance. This number of wavelet coefficients was sufficient to capture 99.5% of the "excess" variance, in the sense of Section 4.2 of Johnstone and Lu (2009).

We used two different true coefficient images $\boldsymbol{\beta}\in\mathbb{R}^{64^2}$, which are shown in Figure 2. The first image $\boldsymbol{\beta}^{(1)}$ is similar to that used by Goldsmith, Huang and Crainiceanu (2014). Taking its domain to be $[1,64]^2$, this coefficient image is given by $\beta^{(1)} = g_1 - g_2$, where g_1, g_2 are the densities of the bivariate normal distributions

$$N\left[\begin{pmatrix}30\\20\end{pmatrix},10\mathbf{I}_2\right] \quad\text{and}\quad N\left[\begin{pmatrix}20\\55\end{pmatrix},10\mathbf{I}_2\right],$$

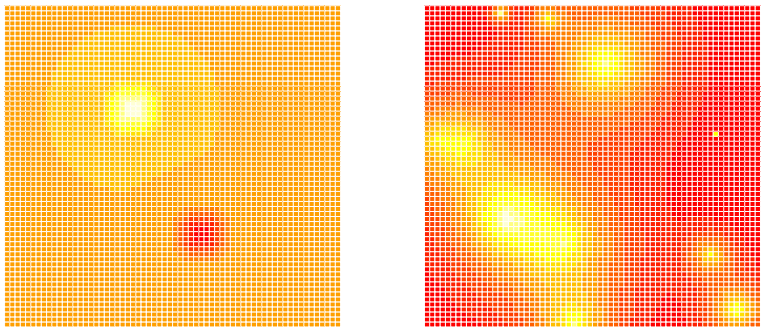

FIG. 2. *Coefficient images* $\boldsymbol{\beta}^{(1)}$ *(left) and* $\boldsymbol{\beta}^{(2)}$ *(right) used in the simulation study.*

respectively. The second image $\boldsymbol{\beta}^{(2)}$ is a two-dimensional analogue of the "bumps" function used by Donoho and Johnstone (1994), and many subsequent authors, to illustrate the properties of wavelets.

We then simulated continuous or binary outcomes $y_1, \ldots, y_n$ with specified approximate values of the coefficient of determination R^2, in the sense detailed in Supplementary Appendix A.1 [Reiss et al. (2015)]. We generated 100 sets of $n = 500$ continuous outcomes and 100 sets of 500 binary outcomes, for each of the R^2 values 0.1, 0.5.

We compared the performance of the three wavelet-domain methods described in Section 3 with three analogous "voxel-domain" methods, that is, sparse PCR, sparse PLS and elastic net without transformation to the wavelet domain. The wavelet- and voxel-domain methods are denoted by WPCR, WPLS and WNet and by VPCR, VPLS and VNet, respectively. We also included the B-spline-based functional PCR method ("FPCR$_R$," or simply FPCR) of Reiss and Ogden (2007, 2010). Tuning parameter selection was as described in Supplementary Appendix A.1 [Reiss et al. (2015)].

Performance was evaluated in terms of estimation error and prediction error. Estimation error is defined by the scaled mean squared error (MSE) $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 / \|\boldsymbol{\beta}\|^2$, where $\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}$ are the true and estimated coefficient images. Prediction error is defined using a separate set of outcomes $y_1^*, \ldots, y_n^*$, generated from the same conditional distribution as $y_1, \ldots, y_n$. We use the scaled mean squared prediction error $\frac{1}{n\sigma^2} \sum_{i=1}^n (y_i^* - \hat{y}_i)^2$ as our criterion for linear regression and the mean of the deviances of $y_1^*, \ldots, y_n^*$ for logistic regression.

Figure 3 presents boxplots of the results. In general, all seven methods differ only slightly in prediction error. Much greater differences are seen for estimation error. Compared with the corresponding voxel-domain methods, the estimation MSE for wavelet methods is either roughly equal or clearly lower on average, and the variability of the MSE is often much lower. The wavelet methods also markedly outperform B-spline-based FPCR. Somewhat contrary to expectation, the superior performance of wavelet methods is not clearly more pronounced for $\boldsymbol{\beta}^{(2)}$ than for $\boldsymbol{\beta}^{(1)}$.

While the wavelet-domain methods do not clearly attain lower estimation error than voxel-domain methods for logistic regression with $R^2 = 0.5$, they do appear superior for the $R^2 = 0.1$ setting (which seems more realistic) and for linear regression. Moreover, qualitatively, wavelet-domain modeling helps to capture the main features of the coefficient image. Figure 4 displays an example of the training-set estimates derived by wavelet-domain lasso versus ordinary lasso. The wavelet-domain estimates are clearly more similar to each other and to the true coefficient image than are the ordinary lasso estimates.

The wavelet-domain EN appears to have a slight edge overall compared with PCR and PLS. For this reason, and because wavelet EN (or at least its special case, the lasso) are now somewhat established in the literature [Wang et al. (2014),

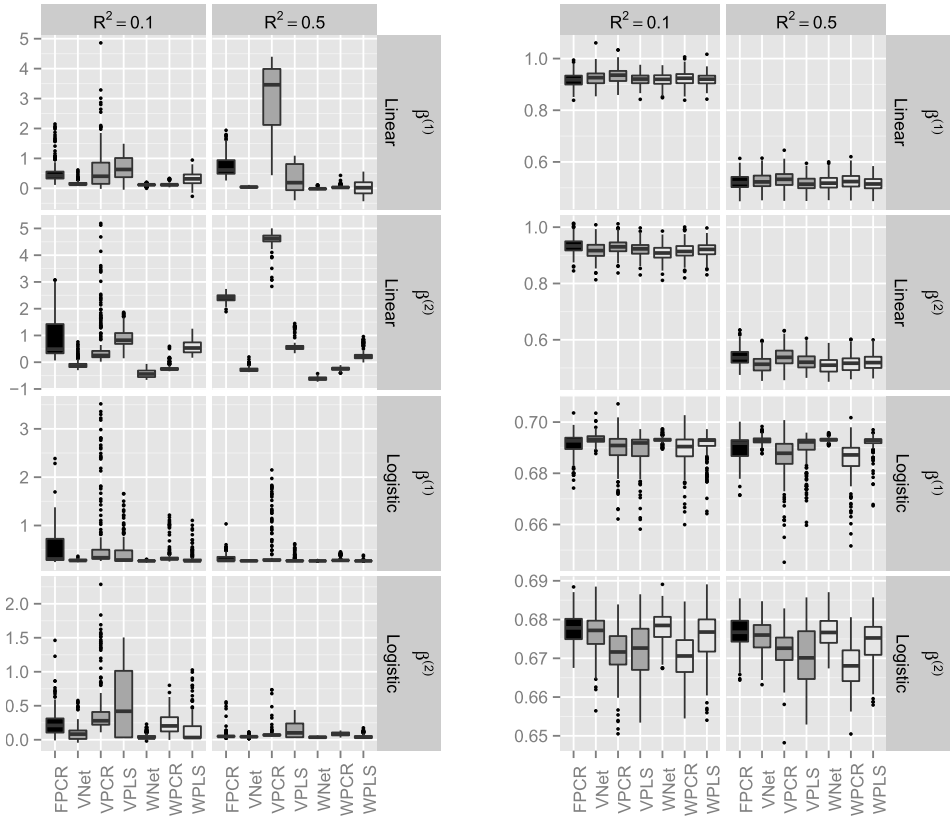

FIG. 3. *Estimation error, displayed as* log(*scaled MSE*) (*left subfigure*), *and prediction error* (*right subfigure*) *in the simulation study.*

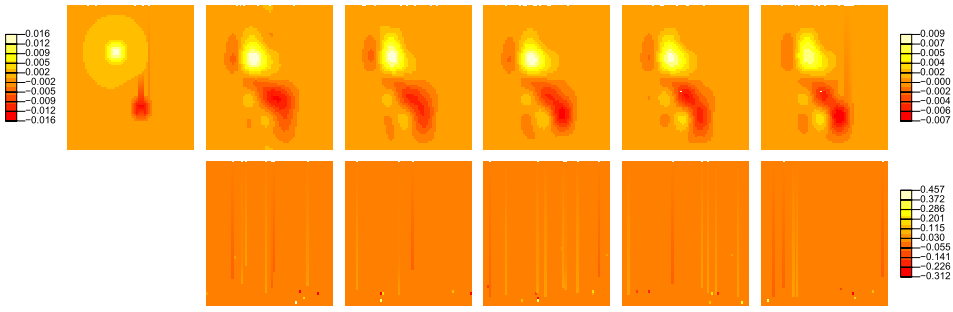

FIG. 4. *True coefficient function* $\boldsymbol{\beta}^{(1)}$ *from the comparative simulation study* (*top left*) *compared with five training-set coefficient function estimates* (*for data simulated under* $R^2 = 1$ *setting*) *based on wavelet-domain lasso* (*other top panels*) *and voxel-domain lasso* (*bottom panels*). *The wavelet-based estimates are reasonably accurate, while each of the voxel-domain estimates has about* 20–25 *scattered voxels with nonzero values. Note the unequal scales.*

Zhao, Chen and Ogden (2015), Zhao, Ogden and Reiss (2012)], the simulations and real-data analyses in the next two sections consider only wavelet-domain EN.

5. Inferential issues. We now turn to what the Introduction referred to as limitation (ii) of predictive analyses in neuroimaging: the need for methodology to assess the predictive value of image data, in particular, when scalar covariates are present.

5.1. *Permutation testing.* Consider testing the null hypothesis $\beta(\mathbf{s}) \equiv 0$ in the general model (1), (2), that is, testing the null parametric model $g(\mu_i) = \mathbf{t}_i^T \boldsymbol{\delta}$ versus the alternative (2). Informally, we are asking whether the images have predictive value beyond the information contained in the scalar predictors. We propose a permutation test procedure in which the CV criterion (13) or (14) serves as the test statistic. If the true-data CV falls in the left tail of the distribution of permuted-data CV values, significance is declared. Permutation techniques of this kind have previously appeared in the neuroimaging and machine learning literature [Golland and Fischl (2003), Ojala and Garriga (2010)].

The way the permutation distribution is constructed depends on the null model under consideration. When $\mathbf{t}_i \equiv 1$ in (2) (no scalar covariates), one can simply permute the responses: that is, we repeatedly reorder the responses as $y_{\pi(1)}, \ldots, y_{\pi(n)}$ for some permutation π, refit the model, and record the CV value. For the linear model (3) with scalar covariates, a common approach is to permute the residuals from the null parametric model: that is, model (3) is refitted repeatedly with the ith response of the form $\hat{y}_i + \hat{\varepsilon}_{\pi(i)}$, where the hats refer to fitted values and residuals from the model $y_i = \mathbf{t}_i^T \boldsymbol{\delta} + \varepsilon_i$. For some GLMs, however, such pseudo-responses based on permuted residuals are not of the correct form (e.g., for logistic regression, they are not binary). One can instead form pseudo-*predictors*, by regressing the predictor of interest on the nuisance covariates and permuting the residuals from this fit. In other words, we replace the design matrix $(\mathbf{T}|\mathbf{X})$ with

$$[\mathbf{T}|\mathbf{P}_T\mathbf{X} + \boldsymbol{\Pi}(\mathbf{I} - \mathbf{P}_T)\mathbf{X}], \tag{15}$$

where $\mathbf{P}_T = \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T$ and $\boldsymbol{\Pi}$ is a permutation matrix. Although a similar idea was proposed by Potter (2005) for (ordinary) logistic regression, we have adopted it as our preferred permutation approach even for the linear case; see Supplementary Appendix B [Reiss et al. (2015)] for further discussion.

We conducted a simulation study, using the ADHD-200 image data analyzed in Section 6, to assess the type-I error rate and power of the permutation test procedure. Here we focus on logistic regression (see Supplementary Appendix C [Reiss et al. (2015)], for linear regression results) and the wavelet-domain lasso. We first considered the case without scalar covariates and generated binary responses $y_i \sim \text{Bernoulli}(p_i)$, $i = 1, \ldots, n = 333$, where

$$\log \frac{p_i}{1 - p_i} = \delta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \tag{16}$$

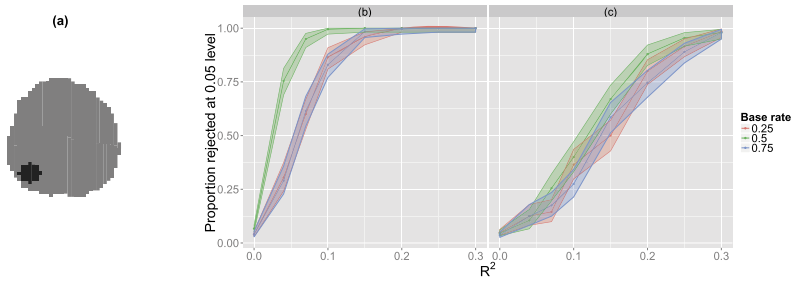

FIG. 5. (a) *True coefficient image* $\boldsymbol{\beta}$ *used in the power study*: *gray denotes* 0, *black denotes* 1. (b) *Estimated probability of rejecting the null hypothesis* $\boldsymbol{\beta} = 0$ *as a function of* R^2, *with* 95% *confidence intervals*, *for model* (16). (c) *Same*, *for model* (17).

where δ_0 is a constant used to adjust the base rate (probability of event); $\mathbf{x}_i \in \mathbb{R}^{64^2}$ is the ith image (expressed as a mean-zero vector); $\boldsymbol{\beta}$ is the true coefficient image shown in Figure 5(a) (similarly vectorized), multiplied by an appropriate constant to attain a specified value of R^2 (see Supplementary Appendix A [Reiss et al. (2015)], regarding the definition of R^2). For each of the base rates 0.25, 0.5, 0.75 and each of the R^2 values 0.04, 0.07, 0.1, 0.15, 0.2, 0.25, 0.3, we simulated 200 response vectors to assess power to reject $H_0 : \boldsymbol{\beta} = \mathbf{0}$ at the $p = 0.05$ level, as well as 1000 response vectors with $\boldsymbol{\beta} = \mathbf{0}$ ($R^2 = 0$) to assess the type-I error rate. Next we considered testing the same null hypothesis for the model

$$\log \frac{p_i}{1 - p_i} = \delta_0 + t_i \delta_1 + \mathbf{x}_i^T \boldsymbol{\beta}, \tag{17}$$

with a scalar covariate t_i such that R^2 for the submodel $E(y_i | t_i) = t_i \delta$ is approximately 0.2. We generated the same number of response vectors as above for each of the above R^2 values, but here R^2 refers to the partial R^2 adjusting for t_i (see Supplementary Appendix A.2 [Reiss et al. (2015)]).

The results, displayed in Figure 5(b) and (c), indicate that the nominal type-I error rate is approximately attained for both models. For a given $R^2 > 0$, the power is somewhat higher for model (16) than for model (17), and highest for either model when the base rate is 0.5. Evidently, for base rates closer to 0 or 1, the CV deviance under the null hypothesis tends to be lower, and thus a stronger signal is needed to reject the null.

Basing a test of the hypothesis $\beta(\cdot) \equiv 0$ on the prediction performance of an estimation algorithm, rather than on an estimate of β, is admittedly somewhat unconventional. In neuroimaging specifically, inference typically proceeds by fitting separate models at each voxel, and then applying some form of multiple testing correction [Nichols (2012)]. In the present setting of a single model that uses the entire image to predict a scalar response, it might be possible to assign p-values to individual voxels as in Meinshausen, Meier and Bühlmann (2009). In practice,

however, predictive algorithms tend to produce rather unstable estimates, as a number of authors have acknowledged [e.g., Craddock et al. (2009), Honorio et al. (2012), Sabuncu, Van Leemput and Alzheimer's Disease Neuroimaging Initiative (2012)]. Our hypothesis testing approach thus sets the more modest inferential goal of verifying that the coefficient image as a whole yields better-than-chance prediction.

5.2. *Confounding*. For ordinary, as opposed to functional, regression, confounding is said to occur when (i) x appears predictive of y, but this relationship can be attributed to a third variable t such that (ii) t is predictive of y and (iii) t is correlated with x. For example, birth order (x) is associated with the occurrence of Down syndrome (y), but this is due to the effect of the confounding variable maternal age (t) [Rothman (2012)].

To extend the above definition to the case of a functional predictor $x(\cdot)$, suppose that (i) $x(\cdot)$ is ostensibly related to y, in the sense that $\beta(\cdot)$ is not identically zero when model (2) includes no scalar covariates, but (ii) the scalar variable t is also predictive of y. A functional-predictor analogue of point (iii) is to suppose that t is correlated with $\int x(s)\hat{\beta}(s)\,ds$, where $\hat{\beta}(\cdot)$ is an estimate obtained with t excluded from model (2). Aside from this "global" analogue of (iii), it may be useful to consider a "local" analogue which holds if t is correlated with $x(s)$, specifically for s such that $\beta(s) \neq 0$; but this is somewhat less straightforward to assess.

6. Application: fALFF and ADHD.

6.1. *ADHD*-200 *data set and candidate models*. We now apply the wavelet-domain elastic net to "predicting" ADHD diagnosis using maps of fractional amplitude of low-frequency fluctuations (fALFF) [Zou et al. (2008)] from a portion of the ADHD-200 sample referred to in the Introduction (http://fcon_1000.projects.nitrc.org/indi/adhd200/). fALFF is defined as the ratio of BOLD signal power spectrum within the 0.01–0.08 Hz range to total over the entire range. Yang et al. (2011) reported altered levels of fALFF in a sample of children with ADHD relative to controls, specifically in frontal regions. That study relied on the traditional analytic approach in neuroimaging, which regresses the imaged quantity (in this case fALFF) on diagnostic group, separately at each voxel. Here we employed scalar-on-image logistic regression, which reverses the roles of response and predictor, to regress diagnostic group on fALFF images. Our sample consisted of 333 individuals: 257 typically developing controls and 76 with combined-type ADHD. The sample included 198 males and 135 females, with age range 7–20 (see Supplementary Appendix D [Reiss et al. (2015)], for further details). We chose the 2D slice for which the mean across voxels of the SD of fALFF was highest. This was the axial slice located at $z = 26$ (just dorsal to the corpus callosum) in the coordinate space of the Montreal Neurological Institute's MNI152 template (4 mm

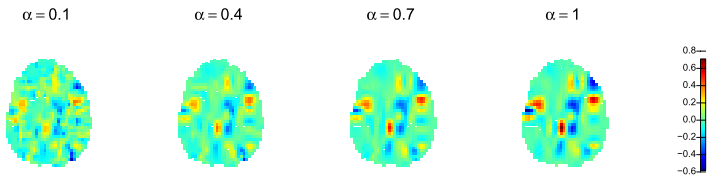

FIG. 6. *Coefficient image estimates for model* (18) *applied to the ADHD-*200 *data, using wavelet–domain elastic net with four different values of the mixing parameter* α.

resolution). We fitted two models. The first was

$$\text{logit Pr}(i\text{th subject has ADHD}) = \delta + \int_{\mathcal{S}} x_i(\mathbf{s})\beta(\mathbf{s})\,d\mathbf{s}, \tag{18}$$

where $x_i(s)$ denotes the ith subject's fALFF image. The second model was

$$\text{logit Pr}(i\text{th subject has ADHD}) = \mathbf{t}_i^T\boldsymbol{\delta} + \int_{\mathcal{S}} x_i(\mathbf{s})\beta(\mathbf{s})\,d\mathbf{s}, \tag{19}$$

where the vector $\mathbf{t}_i$ includes the ith subject's age, sex, IQ and mean FD, as well as a leading 1 for the intercept.

Figure 6 shows the coefficient images attained for model (18) with each value of the mixing parameter α. As expected, increasing values of α lead to more-sparse estimates in the wavelet domain, and hence in the voxel domain. Figure 7 shows the CV deviance as a function of λ for $\alpha = 0.1$, which had the lowest CV deviance overall, as well as for $\alpha = 1$.

The left subfigure of Figure 8 shows that the CV deviance lies in the left tail of the permutation distribution for model (18), indicating a significant effect of the fALFF image predictors ($p = 0.015$). However, with the scalar covariate adjustment of model (19), this effect disappears. The next subsection examines more closely how the scalar covariates may be acting as confounders.

Our test of model (18) entailed 999 permuted-data fits with four candidate values of α and 100 of λ, requiring 14.25 hours on an Intel Xeon E5-2670 processor

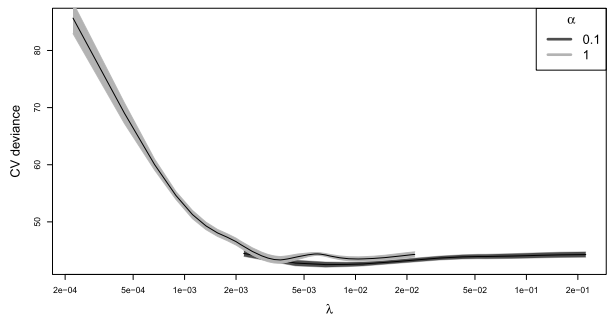

FIG. 7. *Cross-validated deviance +/− one approximate standard error, for the wavelet-domain elastic net models with* $\alpha = 0.1, 1$.

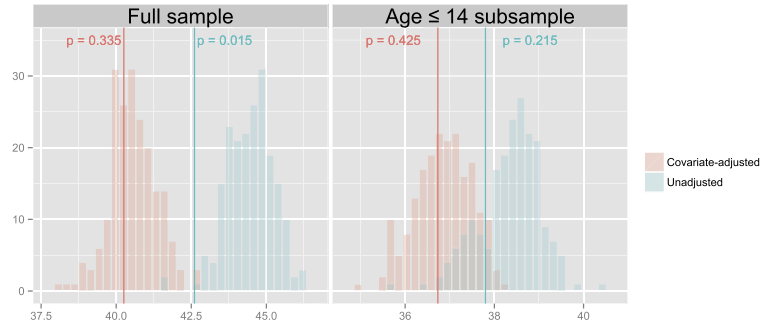

FIG. 8. *Permutation test results. For the full sample (left), a significant effect of the fALFF images is seen in model (18), but not in model (19), which adjusts for scalar covariates. When only younger individuals are included (right), neither model shows a significant fALFF effect.*

running at 2.6 GHz. In practice, we recommend parallelizing the permutations via cluster computing to make the computation time more manageable. In addition, truncated sequential probability ratio tests [Fay, Kim and Hachey (2007)] could in some cases reduce computation time via early stopping. We also explored fitting model (18) with the full 3D fALFF images as predictors; see Supplementary Appendix E [Reiss et al. (2015)].

6.2. *Assessing and remedying confounding.* As discussed in Section 5.2, the notion of confounding entails three elements (see Figure 9). Point (i), an apparent effect of the image predictor fALFF on diagnosis, was established by the above permutation test result for model (18). To check point (ii) of the definition for each of the four scalar covariates under consideration, we performed an ordinary logistic regression with diagnosis (1 = ADHD, 0 = control) as response and the above four scalar predictors. In Table 1 (at left), sex, age and IQ are all seen to be significantly related to diagnosis. See also Figure 10, which compares the fitted probabilities from this ordinary logistic regression with those resulting from models (18) and (19). The scalar-covariates model is seen to separate the two groups (black vs. gray dots) quite well; the image predictors increase the spread of the predicted probabilities without clearly improving the two groups' separation. Based on these results, each of these three variables may be acting as a confounder.

Next we consider point (iii), that is, the correlations of each scalar covariate with $\int_S x_i(\mathbf{s})\hat{\beta}(\mathbf{s})\,d\mathbf{s}$, where $\hat{\beta}$ is the coefficient image estimate from the fALFF-

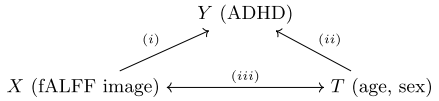

FIG. 9. *Relationships among a putative predictor X, outcome Y and confounder T (see Section 5.2), illustrated with respect to the ADHD-200 data.*

TABLE 1

To examine element (ii) *of confounding, an ordinary logistic regression was fitted with the four scalar predictors and with ADHD diagnosis as response; the resulting estimates are shown with* 95% *confidence intervals. For element* (iii), *we display the correlations of each predictor with the logit probabilities estimated by fitting model* (18)

	(ii)		(iii)	
	Log odds ratio	***p*-value**	**Correlation**	***p*-value**
Intercept	3.90 (1.11, 6.78)	0.007		
Sex (M–F)	1.26 (0.65, 1.91)	0.00008	0.14 (0.03, 0.24)	0.011
Age	−0.20 (−0.32, −0.09)	0.0005	−0.35 (−0.44, −0.25)	6×10^{-11}
IQ	−0.03 (−0.05, −0.01)	0.003	−0.09 (−0.19, 0.02)	0.10
Mean FD	−2.51 (−8.80, 3.56)	0.42	−0.04 (−0.15, 0.07)	0.47

only model (18) or, equivalently, with the predicted logit probability of ADHD from that model. The results, shown at right in Table 1, point to age and sex as the principal confounders. (Here sex was treated as a binary variable, with 1 for male and 0 for female; a t-test and a Mann–Whitney test yielded similar results.) "Local" examination in the sense of Section 5.2 reveals that the fALFF $x(\mathbf{s})$ tends to be higher in males and in younger individuals for many voxels $\mathbf{s}$; and such regions overlap considerably with those in which $\hat{\beta}(\mathbf{s}) > 0$. In other words, the ostensible association between fALFF and ADHD likely reflects the dependence of fALFF on age and sex, which in turn are related to ADHD in our sample.

Further inspection revealed that, of the 67 individuals with age above 14.0, only 8 had ADHD, with maximum age 17.43—whereas the controls had ages as high

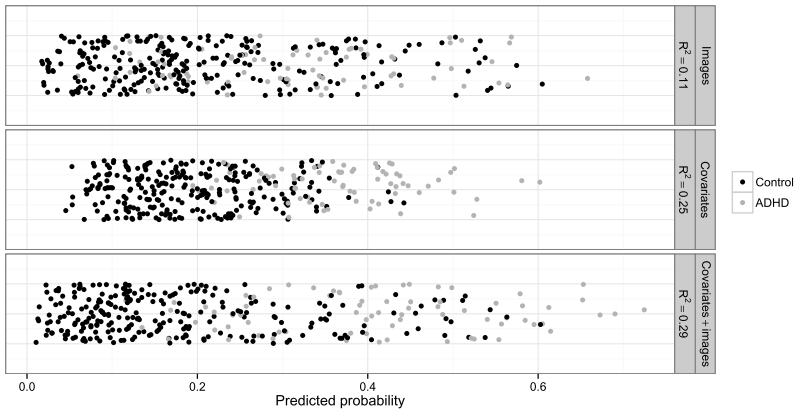

FIG. 10. *Predicted probabilities of ADHD diagnosis, according to the images-only model* (18); *an ordinary logistic regression with the four scalar covariates*; *and model* (19), *which includes both. Also shown are the* R^2 *values, as defined in Supplementary Appendix* A.1, *for the three models.*

as 20.45. This led us to suspect that these older individuals might be driving the confounding with age that results in a spurious effect of fALFF on diagnosis. To investigate this possibility, we repeated the analysis using only the 266 individuals of age 14.0 or lower. Figure 8 shows that in this subsample, the fALFF effect is no longer significant, even without adjusting for the scalar covariates. Moreover, given how far the test statistic is from the left tail of the permutation distribution, it seems unlikely that the loss of significance is due merely to the lower sample size.

In general, absent careful matching at the design stage, it would be advisable to match the two diagnostic groups optimally on a complete set of clearly relevant variables, via algorithms such as those described in Rosenbaum (2010). Our aim here, however, was to show how a straightforward new notion of confounding for functional predictors can be used to identify a principal scalar confounder, whose impact can be removed by the crude device of simply truncating the age range.

7. Discussion. Our analysis in Section 6 included only one imaging modality and only a subset of the individuals from the ADHD-200 Global Competition database. At any rate, our essentially negative result is consistent with the finding [Brown et al. (2012)] that diagnostic accuracy was optimized by basing prediction on scalar predictors, while ignoring the image data. In a blog comment on that outcome, cited both by ADHD-200 Consortium (2012) and by Brown et al. (2012), the neuroscientist Russ Poldrack suggested that "any successful imaging-based decoding could have been relying upon correlates of those variables rather than truly decoding a correlate of the disease." Stated a bit differently, the competing teams' successes in using the image data to predict diagnosis may have been brought about by confounding. But there appear to have been few attempts, if any, to study systematically how confounding may give rise to spurious relationships between quantitative image data and clinical variables. Similarly, analyses of the ADHD-200 data, and related work on brain "decoding," have devoted little attention to formally testing the contribution of imaging data to prediction of scalar responses [but see Reiss (2015)].

As we have shown, these two interrelated issues—testing the effect of image predictors and investigating possible confounders—can be handled straightforwardly within our scalar-on-image regression framework. The permutation test procedure of Section 5.1 found a statistically significant relationship between fALFF images and ADHD diagnosis, but this disappeared when four scalar covariates were adjusted for. Further examination, in light of our extension of the notion of confounding to functional/image predictors in Section 5.2, pointed to age and sex as the key confounders.

The ADHD-200 project is one of a number of recent initiatives to make large samples of neuroimaging data publicly available [Milham (2012)]. These initiatives have been a boon for statistical methodology development, but it must be borne in mind that even as neuroimaging sample sizes increase rapidly, they remain much smaller than the data dimension. No approach to scalar-on-image regression

can completely escape the ensuing nonidentifiability of the coefficient image. We can, however, (i) put forth assumptions, likely to hold approximately in practice, that reduce the effective dimension of the coefficient image; and (ii) employ multiple methods in the hope that these will converge upon similar coefficient image estimates, at least when the signal is sufficiently strong.

With these considerations in mind, we have introduced three methods for scalar-on-image regression, each relying on a different set of assumptions to achieve dimension reduction in the wavelet domain. Implementations of these three methods, for 2D and 3D image data, are provided in the `refund.wave` package [Huo, Reiss and Zhao (2014)] for R [R Development Core Team (2012)], available at http://cran.r-project.org/web/packages/refund.wave. This new package, a spinoff of the `refund` package [Crainiceanu et al. (2014)], relies on the `wavethresh` package [Nason (2013)] for wavelet decomposition and reconstruction.

As discussed in Section 2.2, the three methods described here are merely three instances of a meta-algorithm for scalar-on-image regression. The `refund.wave` package allows for straightforward incorporation of alternative penalties, and other extensions may allow for more refined wavelet-domain algorithms, which may improve the stability and reproducibility of the coefficient image estimates [Rasmussen et al. (2012)]. For instance, in wavelet-based nonparametric regression, thresholding is often performed in a level-specific manner. Analogously, it might be appropriate to modify criterion (11) so as to differentially penalize coefficients at different levels. One might also employ resampling techniques [cf. Meinshausen and Bühlmann (2010)] to select those wavelet basis elements that are consistently predictive of the outcome. Finally, wavelets whose domain is anatomically customized, such as the wavelets defined on the cortex by Özkaya and Van De Ville (2011), offer a promising new way to confine the analysis to relevant portions of the brain.

Acknowledgments. The authors are grateful to the Editor, Karen Kafadar, and to the Associate Editor and referees, whose feedback led to major improvements in the paper; to Adam Ciarleglio, for contributions to software implementation; to Xavier Castellanos, Samuele Cortese, Cameron Craddock, Brett Lullo, Eva Petkova, Fabian Scheipl and Victor Solo for helpful discussions about our methodology and its application; to Jeff Goldsmith, Lei Huang and Ciprian Crainiceanu for sharing their insights as well as a preprint of Goldsmith, Huang and Crainiceanu (2014); and to the ADHD-200 Consortium (http://fcon_1000.projects.nitrc.org/indi/adhd200/) and the Neuro Bureau (http://neurobureau.projects.nitrc.org/) for making the fMRI data set publicly available. In addition to the funding sources listed on the first page, the first author thanks the National Science Foundation for its support of the Statistical and Applied Mathematical Sciences Institute, whose Summer 2013 Program on Neuroimaging Data Analysis provided a valuable opportunity to present part of this

research. This work utilized computing resources at the High Performance Computing Facility of the Center for Health Informatics and Bioinformatics at New York University Langone Medical Center.

SUPPLEMENTARY MATERIAL

Supplementary appendices (DOI: 10.1214/15-AOAS829SUPP; .pdf). Description of simulation details, permutation of residuals for the proposed test procedure, a power study, selection of a subsample from the ADHD-200 data set, and results with 3D predictors.

REFERENCES

ADHD-200 CONSORTIUM (2012). The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Front. Syst. Neurosci.* **6** 62.

BERLINET, A., BIAU, G. and ROUVIÈRE, L. (2008). Functional supervised classification with wavelets. *Ann. I.S.U.P.* **52** 61–80. MR2435041

BROWN, P. J., FEARN, T. and VANNUCCI, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Amer. Statist. Assoc.* **96** 398–408. MR1939343

BROWN, M. R. G., SIDHU, G. S., GREINER, R., ASGARIAN, N., BASTANI, M., SILVERSTONE, P. H., GREENSHAW, A. J. and DURSUN, S. M. (2012). ADHD-200 global competition: Diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Front. Syst. Neurosci.* **6** 69.

CAFFO, B., ELOYAN, A., HAN, F., LIU, H., MUSCHELLI, J., NEBEL, M. B., ZHAO, T. and CRAINICEANU, C. (2012). SMART thoughts on the ADHD 200 Competition. Available at http://www.smart-stats.org/?q=content/repost-our-document-adhd-competition.

CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496

CARDOT, H., FERRATY, F. and SARDA, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45** 11–22. MR1718346

CARDOT, H., FERRATY, F. and SARDA, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13** 571–591. MR1997162

CHANG, C., CHEN, Y. and OGDEN, R. T. (2014). Functional data classification: A wavelet approach. *Comput. Statist.* **29** 1497–1513.

CHI, E. C. and SCOTT, D. W. (2014). Robust parametric classification and variable selection by a minimum distance criterion. *J. Comput. Graph. Statist.* **23** 111–128.

CHUN, H. and KELEŞ, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 3–25. MR2751241

COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22** 1–26. MR2408655

CRADDOCK, R. C., HOLTZHEIMER III, P. E., HU, X. P. and MAYBERG, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magn. Reson. Med.* **62** 1619–1628.

CRAINICEANU, C. M., REISS, P. T., GOLDSMITH, J., HUANG, L., HUO, L. and SCHEIPL, F. (2014). refund: Regression with functional data. R package version 0.1-10.

DAUBECHIES, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41** 909–996. MR0951745

DELAIGLE, A. and HALL, P. (2012a). Achieving near perfect classification for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 267–286. MR2899863

DELAIGLE, A. and HALL, P. (2012b). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* **40** 322–352. MR3014309

DING, B. and GENTLEMAN, R. (2005). Classification using generalized partial least squares. *J. Comput. Graph. Statist.* **14** 280–298. MR2160814

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

ELOYAN, A., MUSCHELLI, J., NEBEL, M. B., LIU, H., HAN, F., ZHAO, T., BARBER, A. D., JOEL, S., PEKAR, J. J., MOSTOFSKY, S. H. and CAFFO, B. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Front. Syst. Neurosci.* **6** 61.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAY, M. P., KIM, H.-J. and HACHEY, M. (2007). On using truncated sequential probability ratio test boundaries for Monte Carlo implementation of hypothesis tests. *J. Comput. Graph. Statist.* **16** 946–967. MR2412490

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GOLDSMITH, J., HUANG, L. and CRAINICEANU, C. M. (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection. *J. Comput. Graph. Statist.* **8** 1045–1064.

GOLDSMITH, J., BOBB, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2011). Penalized functional regression. *J. Comput. Graph. Statist.* **20** 830–851. MR2878950

GOLLAND, P. and FISCHL, B. (2003). Permutation tests for classification: Towards statistical significance in image-based studies. In *Information Processing in Medical Imaging*: *Proceedings of the* 18*th International Conference* (C. J. Taylor and J. A. Noble, eds.) 330–341. Springer, Berlin.

GROSENICK, L., KLINGENBERG, B., KATOVICH, K., KNUTSON, B. and TAYLOR, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage* **72** 304–321.

GUILLAS, S. and LAI, M.-J. (2010). Bivariate splines for spatial functional regression models. *J. Nonparametr. Stat.* **22** 477–497. MR2662608

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

HONORIO, J., TOMASI, D., GOLDSTEIN, R. Z., LEUNG, H.-C. and SAMARAS, D. (2012). Can a single brain region predict a disorder? *IEEE Trans. Med. Imaging* **31** 2062–2072.

HUANG, L., GOLDSMITH, J., REISS, P. T., REICH, D. S. and CRAINICEANU, C. M. (2013). Bayesian scalar-on-image regression with application to association between intracranial DTI and cognitive outcomes. *NeuroImage* **83** 210–223.

HUO, L., REISS, P. and ZHAO, Y. (2014). refund.wave: Wavelet–domain regression with functional data. R package version 0.1.

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

KAPUR, S., PHILLIPS, A. G. and INSEL, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* **17** 1174–1179.

MALLAT, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11** 674–693.

MALLAT, S. (2009). *A Wavelet Tour of Signal Processing*: *The Sparse Way*, 3rd ed. Academic Press, Burlington, MA. MR2479996

MALLOY, E. J., MORRIS, J. S., ADAR, S. D., SUH, H., GOLD, D. R. and COULL, B. A. (2010). Wavelet-based functional linear mixed models: An application to measurement error-corrected distributed lag models. *Biostatistics* **11** 432–452.

MARX, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics* **38** 374–381.

MARX, B. D. and EILERS, P. H. C. (1999). Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics* **41** 1–13.

MARX, B. D. and EILERS, P. H. C. (2005). Multidimensional penalized signal regression. *Technometrics* **47** 13–22. MR2135789

MASSY, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Statist. Assoc.* **60** 234–256.

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584

MILHAM, M. P. (2012). Open neuroscience solutions for the connectome-wide association era. *Neuron* **73** 214–218.

MORRIS, J. S., BALADANDAYUTHAPANI, V., HERRICK, R. C., SANNA, P. and GUTSTEIN, H. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *Ann. Appl. Stat.* **5** 894–923. MR2840180

MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805. MR2163159

NADLER, B. and COIFMAN, R. R. (2005). The prediction error in CLS and PLS: The importance of feature selection prior to multivariate calibration. *J. Chemom.* **19** 107–118.

NASON, G. P. (2008). *Wavelet Methods in Statistics with R*. Springer, New York. MR2445580

NASON, G. (2013). wavethresh: Wavelets statistics and transforms. R package version 4.6.2.

NGUYEN, D. V. and ROCKE, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18** 39–50.

NICHOLS, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage* **62** 811–815.

OGDEN, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, Boston, MA. MR1420193

OJALA, M. and GARRIGA, G. C. (2010). Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* **11** 1833–1863. MR2660654

ÖZKAYA, S. G. and VAN DE VILLE, D. (2011). Anatomically adapted wavelets for integrated statistical analysis of fMRI data. In 2011 *IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 469–472. IEEE, New York.

POTTER, D. M. (2005). A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Stat. Med.* **24** 693–708. MR2134534

PREDA, C. and SAPORTA, G. (2005). PLS regression on a stochastic process. *Comput. Statist. Data Anal.* **48** 149–158. MR2134488

R DEVELOPMENT CORE TEAM (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

RASMUSSEN, P. M., HANSEN, L. K., MADSEN, K. H., CHURCHILL, N. W. and STROTHER, S. C. (2012). Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition* **45** 2085–2100.

REISS, P. T. (2006). Regression with signals and images as predictors. Ph.D. Thesis, Dept. of Biostatistics, Columbia Univ., New York.

REISS, P. T. (2015). Cross-validation and hypothesis testing in neuroimaging: An irenic comment on the exchange between Friston and Lindquist et al. *NeuroImage*. To appear.

REISS, P. T. and OGDEN, R. T. (2007). Functional principal component regression and functional partial least squares. *J. Amer. Statist. Assoc.* **102** 984–996. MR2411660

REISS, P. T. and OGDEN, R. T. (2010). Functional generalized linear models with images as predictors. *Biometrics* **66** 61–69. MR2756691

REISS, P. T., HUO, L., ZHAO, Y., KELLY, C. and OGDEN, R. T. (2015). Supplement to "Wavelet-domain regression and predictive inference in psychiatric neuroimaging." DOI:10.1214/15-AOAS829SUPP.

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

ROTHMAN, K. J. (2012). *Epidemiology*: *An Introduction*, 2nd ed. Oxford Univ. Press, New York.

RUTTIMANN, U. E., UNSER, M., RAWLINGS, R. R., RIO, D., RAMSEY, N. F., MATTAY, V. S., HOMMER, D. W., FRANK, J. A. and WEINBERGER, D. R. (1998). Statistical analysis of functional MRI data in the wavelet domain. *IEEE Trans. Med. Imaging* **17** 142–154.

SABUNCU, M. R., VAN LEEMPUT, K. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2012). The relevance voxel machine (RVoxM): A self-tuning Bayesian model for informative image-based prediction. *IEEE Trans. Med. Imaging* **31** 2290–2306.

SHEN, H. and HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99** 1015–1034. MR2419336

STONE, M. and BROOKS, R. J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. Roy. Statist. Soc. Ser. B* **52** 237–269. MR1064418

SUN, D., VAN ERP, T. G. M., THOMPSON, P. M., BEARDEN, C. E., DALEY, M., KUSHAN, L., HARDT, M. E., NUECHTERLEIN, K. H., TOGA, A. W. and CANNON, T. D. (2009). Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: Classification analysis using probabilistic brain atlas and machine learning algorithms. *Biological Psychiatry* **66** 1055–1060.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN DE VILLE, D., SEGHIER, M. L., LAZEYRAS, F., BLU, T. and UNSER, M. (2007). WSPM: Wavelet-based statistical parametric mapping. *NeuroImage* **37** 1205–1217.

VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. Wiley, New York. MR1681904

WAND, M. P. and ORMEROD, J. T. (2011). Penalized wavelets: Embedding wavelets into semiparametric regression. *Electron. J. Stat.* **5** 1654–1717. MR2870147

WANG, X., RAY, S. and MALLICK, B. K. (2007). Bayesian curve classification using wavelets. *J. Amer. Statist. Assoc.* **102** 962–973. MR2354408

WANG, X., NAN, B., ZHU, J., KOEPPE, R. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2014). Regularized 3D functional regression for brain image data via Haar wavelets. *Ann. Appl. Stat.* **8** 1045–1064.

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

WOLD, H. (1966). Nonlinear estimation by iterative least square procedures. In *Research Papers in Statistics* (*Festschrift J. Neyman*) (F. N. David, ed.) 411–444. Wiley, London. MR0210250

YANG, H., WU, Q.-Z., GUO, L.-T., LI, Q.-Q., LONG, X.-Y., HUANG, X.-Q., CHAN, R. C. K. and GONG, Q.-Y. (2011). Abnormal spontaneous brain activity in medication-naïve ADHD children: A resting state fMRI study. *Neurosci. Lett.* **502** 89–93.

ZHAO, Y., CHEN, H. and OGDEN, R. T. (2015). Wavelet-based weighted LASSO and screening approaches in functional linear regression. *J. Comput. Graph. Statist.* To appear.

ZHAO, Y., OGDEN, R. T. and REISS, P. T. (2012). Wavelet-based LASSO in functional linear regression. *J. Comput. Graph. Statist.* **21** 600–617. MR2970910

ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. MR3174640

ZHU, H., BROWN, P. J. and MORRIS, J. S. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics* **68** 1260–1268. MR3040032

ZHU, J. and HASTIE, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5** 427–443.

ZHU, H., VANNUCCI, M. and COX, D. D. (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* **66** 463–473. MR2758826

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

ZOU, Q.-H., ZHU, C.-Z., YANG, Y., ZUO, X.-N., LONG, X.-Y., CAO, Q.-J., WANG, Y.-F. and ZANG, Y.-F. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. *J. Neurosci. Methods* **172** 137–141.

P. T. REISS
L. HUO
Y. ZHAO
C. KELLY
DEPARTMENT OF CHILD AND ADOLESCENT PSYCHIATRY
NEW YORK UNIVERSITY SCHOOL OF MEDICINE
1 PARK AVE., 7TH FLOOR
NEW YORK, NEW YORK 10016
USA
E-MAIL: phil.reiss@nyumc.org
huolanlan@gmail.com
yihong.zhao@nyumc.org
amclarekelly@gmail.com

R. T. OGDEN
DEPARTMENT OF BIOSTATISTICS
COLUMBIA UNIVERSITY
722 W. 168TH ST., 6TH FLOOR
NEW YORK, NEW YORK 10032
USA
E-MAIL: to166@cumc.columbia.edu