March, 2010

Functional Generalized Linear Models with Images as Predictors

Philip T. Reiss, *New York University*
R. Todd Ogden, *Columbia University*

BIOMETRICS ??, 1–20 DOI: 10.1111/j.1541-0420.2005.00454.x
under review

Functional Generalized Linear Models with Images as Predictors

Philip T. Reiss

Department of Child and Adolescent Psychiatry, New York University, New York, New York 10016, U.S.A.

and Nathan Kline Institute for Psychiatric Research, Orangeburg, New York 10962, U.S.A.

email: phil.reiss@nyumc.org

and

R. Todd Ogden

Department of Biostatistics, Columbia University, New York, New York 10032, U.S.A.

email: to166@columbia.edu

SUMMARY: Functional principal component regression (FPCR) is a promising new method for regressing scalar outcomes on functional predictors. In this paper we present a theoretical justification for the use of principal components in functional regression. FPCR is then extended in two directions: from linear to the generalized linear modeling, and from univariate signal predictors to high-resolution image predictors. We show how to implement the method efficiently by adapting generalized additive model technology to the functional regression context. A technique is proposed for estimating simultaneous confidence bands for the coefficient function; in the neuroimaging setting, this yields a novel means to identify brain regions that are associated with a clinical outcome. A new application of likelihood ratio testing is described for assessing the null hypothesis of a constant coefficient function. The performance of the methodology is illustrated via simulations and real data analyses with positron emission tomography images as predictors.

KEY WORDS: B-splines; Functional principal component regression; Positron emission tomography; Simultaneous confidence bands; Smoothing parameter.

1. Introduction

The past decade has seen heightened interest in generalized linear models (GLMs) with scalar outcomes and functional predictors (Ramsay and Silverman, 2005). Whereas ordinary multiple regression estimates a vector $\boldsymbol{\beta}$ whose inner product with the predictor vector equals the expectation of the outcome (or is related to it by a link function, in the generalized linear case), the functional regression models to which we refer estimate a coefficient *function* whose L^2 inner product with the functional predictors determines the fitted values. The functional predictors in some applications may be viewed as longitudinal data, e.g., the number of eggs laid each day for a month by fruit flies (Müller and Stadtmüller, 2005). In other applications, the predictors are densely observed curves or signals, such as near-infrared spectra acquired from chemical samples. Applications of the latter type may require greater flexibility for the coefficient function, to enable detection of features at high levels of detail. In this paper we are concerned with the even higher-dimensional challenge of images as predictors. This work was motivated by psychiatric applications in which it is of interest to predict outcomes from positron emission tomography (PET) images of subjects' brains. Our modeling strategy builds on the functional principal component regression (FPCR) method proposed by Reiss and Ogden (2007) for linear models with one-dimensional signal predictors.

The novel contributions of this paper are as follows. First, we present an optimality result (Proposition 1) justifying the FPCR approach of using principal components for reduction of the predictor dimension. Second, we extend FPCR to multivariate signal (image) predictors, and from linear to generalized linear models. Third, we describe how existing computationally optimized software for generalized additive models can be adapted to fit the FPCR model. Fourth, we propose a novel approach to detecting regions of the image exerting significant influence on the outcome, via resampling-based simultaneous confidence bands for the coefficient function. Finally, we show how a test recently developed for mixed

models can be applied to evaluate the null hypothesis of a constant coefficient function. The proposed methods are applied to real data sets from psychiatric neuroimaging data, and simulations show their favorable performance compared with existing approaches.

2. Functional Principal Component Regression

In the linear regression form of the problem, a vector $\boldsymbol{y}$ of n scalar responses is modeled as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{S}\boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{X}$ is an $n \times p$ covariate matrix; $\boldsymbol{S}$ is an $n \times N$ matrix, each row of which represents a signal predictor defined at points $v_1, \ldots, v_N$; and $\boldsymbol{\varepsilon}$ denotes iid errors. We assume throughout the paper that $\boldsymbol{S}$ has mean-zero columns. The matrix $\boldsymbol{X}$ may consist of only a row of ones, but extending the methodology to include genuine covariates presents little difficulty. The difficult problem is estimation of the (discretized) coefficient function $\boldsymbol{f} = [f(v_1), \ldots, f(v_N)]^T$, since in general $n \ll N$ and thus the model is overdetermined. It is therefore necessary to reduce the dimension of the signal matrix $\boldsymbol{S}$.

One approach to dimension reduction restricts the coefficient function to the span of a B-spline basis and adds a roughness penalty to enforce smoothness (Marx and Eilers, 1999, 2005; Cardot, Ferraty and Sarda, 2003). Letting $\boldsymbol{B}$ be an $N \times K$ matrix whose columns form a set of B-spline basis functions (evaluated at the same N points as the signal predictors), we choose $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ to minimize

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{S}\boldsymbol{B}\boldsymbol{\gamma}\|^2 + \lambda\boldsymbol{\gamma}^T\boldsymbol{P}\boldsymbol{\gamma}, \tag{2}$$

where $\boldsymbol{P}$ is chosen so that $\boldsymbol{\gamma}^T\boldsymbol{P}\boldsymbol{\gamma}$ provides a measure of the roughness of the function $\boldsymbol{f} = \boldsymbol{B}\boldsymbol{\gamma}$, and the constant $\lambda > 0$ determines the extent to which such roughness is penalized.

In some applications, in order to capture the relevant detail, the dimension K of the B-spline basis needs to be substantially larger than the sample size n. This is particularly so for imaging applications: typically, K must be in the hundreds to provide a sufficiently

rich basis, but n is in the dozens. Provided $\lambda > 0$, criterion (2) will still have a unique minimum if $n < K$. Nevertheless, this involves inversion of very large matrices; furthermore, a basis that is larger than the sample size is evidently too rich to estimate $\boldsymbol{\gamma}$ well. Thus a further dimension reduction step is indicated. Optimally, this second step should attain minimax perturbation of the fitted value, relative to the penalized B-spline expansion of (2) (see Proposition 1 below). This goal is attained by FPCR, which first projects onto a B-spline basis (while adding a roughness penalty as above), then reduces to leading principal components. Thus we minimize

$$\|\boldsymbol{y} - \boldsymbol{X\alpha} - \boldsymbol{SBV}_q\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{V}_q^T\boldsymbol{PV}_q\boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{V}_q$ now consists of the leading columns of the matrix $\boldsymbol{V}$ from the singular value decomposition $\boldsymbol{UDV}^T$ of $\boldsymbol{SB}$. Given the minimizing spline coefficient vector $\hat{\boldsymbol{\beta}}$, the coefficient function is estimated as $\hat{\boldsymbol{f}} = \boldsymbol{BV}_q\hat{\boldsymbol{\beta}}$. Reiss and Ogden (2007) provided real- and simulated-data evidence that this second dimension reduction step improves performance even in the $n > K$ case. They also obtained asymptotic convergence results.

The optimality of principal components for the second dimension reduction step follows from the following result, which is proved in Web Appendix A.

PROPOSITION 1: Let $\boldsymbol{UDV}^T$ be the singular value decomposition of the $n \times K$ matrix $\boldsymbol{Z}$, let $\boldsymbol{U}_q$ be the matrix consisting of the first $q < \min(n, K)$ columns of $\boldsymbol{U}$, and let $\boldsymbol{D}_q$ be the $q \times q$ upper left submatrix of $\boldsymbol{D}$. Then $\boldsymbol{M}_0 = \boldsymbol{U}_q\boldsymbol{D}_q$ minimizes

$$\max_{\mathbf{w}\in\mathbf{R}^K, \|\mathbf{w}\|=1} \|\boldsymbol{Zw} - \mathrm{proj}_{\mathbf{M}}\boldsymbol{Zw}\|$$

over all $n \times q$ matrices $\boldsymbol{M}$, where $\mathrm{proj}_{\mathbf{M}}$ denotes projection onto the column space of $\boldsymbol{M}$.

The import of Proposition 1 can be seen most clearly by taking $\boldsymbol{y}$ to be centered, so that it is appropriate to fit a no-intercept model $\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$, with $n \times K$ design matrix $\boldsymbol{Z} = \boldsymbol{SB}$ (cf. (2)). Consider replacing $\boldsymbol{Z}$ with an $n \times q$ design matrix $\boldsymbol{M}$ with $q < K$, and define the

resulting perturbation in fitted values to be the norm of the difference between $\hat{\boldsymbol{y}} = \boldsymbol{SB}\hat{\boldsymbol{\gamma}}$ and its projection $\text{proj}_{\mathbf{M}}\hat{\boldsymbol{y}}$ onto the column space of $\boldsymbol{M}$. The maximum perturbation $\|\hat{\boldsymbol{y}} - \text{proj}_{\mathbf{M}}\hat{\boldsymbol{y}}\|$ is minimized by taking $\boldsymbol{M}$ to be $\boldsymbol{U}_q\boldsymbol{D}_q$, where $\boldsymbol{UDV}^T$ is the singular value decomposition of $\boldsymbol{SB}$. But in this case $\boldsymbol{U}_q\boldsymbol{D}_q$ equals $\boldsymbol{SBV}_q$, the FPCR design matrix in (3). In this sense FPCR attains minimax perturbation of the fitted values.

Proposition 1 bears comparison with Wood's (2003) development of thin plate regression splines, a low-rank approximation to the thin plate spline basis for nonparametric regression with multidimensional predictors. Wood showed that reducing the design matrix to its leading principal components—a second dimension-reduction step, analogous to that in FPCR—results in minimax perturbation of both the fitted values and the smoothing penalty. In the present setting, we seek to minimize the former type of change, but not the latter.

3. Extending Functional Principal Component Regression

3.1 *From Signals to Images*

We sought to generalize our smoothing strategy for one-dimensional signals (cubic splines with integrated squared second derivative penalty) to the case of d-dimensional image predictors (for most of this paper $d = 2$) while retaining radial symmetry (Ruppert, Wand, and Carroll, 2003) of both the basis functions and the roughness penalty. Thus, for the basis functions, we chose radial cubic B-splines (Saranli and Baykal, 1998) centered at each of an equally spaced grid of knots $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K \in \mathcal{R}^d$. The ith basis function (represented in discretized form by the ith column of $\boldsymbol{B}$) is given by $b_i(\boldsymbol{v}) = g(\|\boldsymbol{v} - \boldsymbol{\kappa}_i\|)$ with

$$g(r) = \frac{1}{4h^2}\begin{cases} h^3 + 3h^2(h-r) + 3h(h-r)^2 - 3(h-r)^3, & r \leqslant h; \\ (2h-r)^3, & h < r \leqslant 2h; \\ 0 & r > 2h, \end{cases} \tag{4}$$

where $h > 0$ is the distance between adjacent knots. The above function g is (within a constant) that used to define a univariate B-spline. For $d = 2$, then, the radial B-spline can

be thought of as the univariate B-spline function rotated around a normal to the plane at the point at which the function peaks. Note that (4) can serve as the definition of a radial cubic B-spline in any number of dimensions.

To penalize roughness in a radially symmetric manner we use the thin plate penalty (Green and Silverman, 1994). For functions f of d variables, if $2m > d$, we can define a mth-order roughness functional by $J_{md}(f) = \int \ldots \int_{\mathbf{R}^d} \sum_{\nu_1+\ldots+\nu_d=m} \frac{m!}{\nu_1!\ldots\nu_d!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1}\ldots\partial x_d^{\nu_d}}\right)^2 dx_1 \ldots dx_d$.Taking $m = 2, d = 1$ yields the familiar univariate roughness penalty $J_{21}(f) = \int_{-\infty}^{\infty}[f''(x_1)]^2 dx_1$. For bivariate smoothing the standard choice is $J_{22}(f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\left(\frac{\partial^2 f}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2}\right)^2\right] dx_1 dx_2$. Upon calculating the mth-order partial derivatives of the basis functions $b_1, \ldots, b_K$ as determined by (4), it is straightforward to obtain a penalty matrix $\boldsymbol{P}$ such that the penalty term $\boldsymbol{\beta}^T \boldsymbol{V}_q^T \boldsymbol{P} \boldsymbol{V}_q \boldsymbol{\beta}$ in (3) equals $J_{md}(f)$, where f is given in discretized form by $\boldsymbol{f} = \boldsymbol{B}\boldsymbol{V}_q\boldsymbol{\beta}$.

3.2 *From Linear to Generalized Linear Models*

We wish to generalize FPCR to generalized linear models, i.e., we shall consider a vector $\boldsymbol{y}$ of iid responses with density of the form $f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$, whose expectation $\boldsymbol{\mu}$ satisfies $g(\boldsymbol{\mu}) \equiv (g(\mu_1)\ldots g(\mu_n))^T = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{S}\boldsymbol{f}$ for some link function g, where, as before, $\boldsymbol{X}$ is an $n \times p$ covariate matrix and $\boldsymbol{S}$ is an $n \times N$ signal matrix. Specifically we shall focus on the binary logistic regression model. As in the one-dimensional linear case, we propose to restrict the coefficient function to the span of a B-spline basis (but in this case using radial B-splines), then reduce to the leading q principal components to obtain the model

$$y_i \sim \text{Bernoulli}\left(\frac{\exp[(\boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{S}\boldsymbol{B}\boldsymbol{V}_q\boldsymbol{\beta})_i]}{1 + \exp[(\boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{S}\boldsymbol{B}\boldsymbol{V}_q\boldsymbol{\beta})_i]}\right). \tag{5}$$

The fitting procedure for our generalized linear FPCR model is a modification of the iteratively reweighted least squares (IRLS) method for ordinary GLMs. Letting $\boldsymbol{Z} = \boldsymbol{S}\boldsymbol{B}\boldsymbol{V}_q$, the updating steps have the form

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \lambda\boldsymbol{V}_q^T\boldsymbol{P}\boldsymbol{V}_q \end{bmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{z} \end{pmatrix}, \tag{6}$$

where $\boldsymbol{W}$ is the weight matrix and $\boldsymbol{z}$ is the "working" dependent variable vector, both defined as for unpenalized GLMs. But this penalized IRLS procedure can often fail to converge, and may be numerically unstable even when it does converge due to numerical rank deficiency of the model matrix (Wood, 2004, 2006). Although the references just cited focus on generalized additive models, Wood (2000) notes that the same methods apply to a broad class of penalized GLMs, of which our functional GLMs are an example. Consequently, as the next subsection discusses, smoothing parameter selection criteria for generalized additive models can be readily adapted to our context. Moreover, FPCR can be implemented using `mgcv` (Wood, 2006), a package for `R` (R Development Core Team, 2008), which addresses the computational difficulties encountered by penalized IRLS by means of efficient orthogonal matrix decompositions. We have found that this implementation, specifically for logistic-regression FPCR, largely removes the problem of model divergence.

Model (5) is similar to the "PCA1/Method I" form of functional principal component logistic regression considered by Escabias, Aguilera, and Valderrama (2004). These authors consider regressing on principal components of $\boldsymbol{A\Psi}$, where $\boldsymbol{A}$ is a matrix of basis coefficients (flexibly defined) and, assuming the above basis is used, $\boldsymbol{\Psi} = \boldsymbol{B}^T\boldsymbol{B}$. If the coefficients in $\boldsymbol{A}$ are obtained by projecting the images onto the given basis then it is easy to show that $\boldsymbol{A\Psi} = \boldsymbol{SB}$, i.e., the matrix whose principal components are regressed on is the same as above. The key difference is that their procedure excludes the roughness penalty in (6). The simulation results of Section 5 suggest that such a penalty improves performance.

3.3 *Smoothing Parameter Selection*

Letting $\boldsymbol{H}_\lambda$ be the hat matrix of the weighted regression in the last penalized IRLS step, we define the degrees of freedom of the fit (df_λ) as $\mathrm{tr}\boldsymbol{H}_\lambda$. We can then define the generalized cross-validation score $GCV(\lambda) = \frac{nD(\lambda)}{(n-df_\lambda)^2}$, where $D(\lambda)$ is the deviance, defined as 2ϕ times the difference in log likelihood between the saturated model and the model obtained with

smoothing parameter value λ (recall that $\phi = 1$ for the binary logistic model). Similarly, the Akaike information criterion (AIC) can be defined as $n^{-1}\left[D(\lambda) + 2\phi df_\lambda\right]$. Wood (2006) argues that the extension of AIC to the generalized case is better justified than that of GCV. Hurvich, Simonoff, and Tsai (1998) propose a corrected AIC which, in the generalized linear setting, can be written as $\frac{D(\lambda)}{n} + \frac{2\phi(df_\lambda+1)}{n-df_\lambda-2}$.

Cai and Hall (2006) show that predicting future outcomes in functional regression calls for a lesser degree of smoothing than is appropriate for estimating the coefficient function, since, in the prediction context, extra smoothness is conferred by taking the inner product of the coefficient function with a new signal. These authors consider a setup like that of Escabias et al. (2004), in which the smoothing parameter is the number of principal components rather than our parameter λ. Nevertheless, the underlying intuition remains valid for our setup. Indeed, although we have found that the correction to AIC results in improved estimation in logistic regression, the uncorrected AIC, which is prone to undersmoothing, worked better for prediction (see below, Section 5).

An alternative to GCV and AIC is to observe that (3) is proportional to the negative of the best linear unbiased predictor (BLUP) criterion for a linear mixed model, with λ equal to the ratio of the error variance and the random effects variance (Ruppert et al., 2003). One can thus, in the linear case, apply the restricted maximum likelihood (REML) method for linear mixed models to obtain an estimate of λ. Reiss and Ogden (2007) found that this REML method outperformed GCV, perhaps due to the latter's greater tendency to have multiple optima. In the GLM case, one could analogously choose λ by the penalized quasilikelihood (PQL) method associated with generalized linear mixed models. However, our experience with logistic FPCR suggests that PQL tends to severely oversmooth, although this is perhaps not surprising since PQL is known to perform poorly for binary logistic regression with mixed effects (e.g., Ruppert et al., 2003).

In addition to choosing λ, FPCR requires selecting the number of principal components q. For the results reported below in Sections 5 and 6, we chose a number of components that seemed reasonably large and thus adequate to capture the details of the coefficient function. A more computationally intensive approach (but not necessarily a more effective one; see Reiss and Ogden, 2007) would choose the number of components by multifold cross-validation.

4. Inferential Aspects

4.1 *Simultaneous Testing*

4.1.1 *Statistical Parametric Maps.* In the neuroimaging context, the methodology we have described departs radically from the standard approach to identifying brain regions associated with a clinical variable of interest. FPCR fits a single regression of clinical variable y on a brain map of some quantity s. The standard "mass-univariate" approach, as implemented in widely used software packages such as Statistical Parametric Mapping (Wellcome Department of Imaging Neuroscience, University College London), fits a model separately for each voxel, and in the opposite direction: s is regressed on y, producing a "statistic image" or map of test statistics at each voxel.

The two paradigms lead to very different approaches to inferring regions of "significance." With the mass-univariate approach one typically declares as significant those regions in which the statistic image values are more extreme than would be expected for fluctuations of an appropriate random field; this approach performs simultaneous inference, i.e., controls the family-wise error rate across all voxels. With FPCR, on the other hand, it is natural to base simultaneous inference upon interval estimation for the coefficient image, as we discuss next.

4.1.2 *Testing with Nonparametric Bootstrap-Derived Simultaneous Confidence Bands.* From a functional GLM perspective, the key inferential problem is to identify $j \in \{1, \ldots, N\}$ such that $f(v_j)$, the coefficient function value at the jth location, differs significantly from

zero. As noted above, brain mapping investigations in particular define significance in a simultaneous sense. In this context, then, testing by inversion of confidence intervals for $\boldsymbol{f}$ defined as in Marx and Eilers (1999) would not suffice—not only because such intervals do not account for the choice of the smoothing parameter (a limitation which may be removed by Bayesian alternatives), but because they are pointwise intervals. Our approach, following Buja and Rolke (2007), is to perform simultaneous tests by inverting simultaneous confidence bands for $\boldsymbol{f}$ derived from nonparametric bootstrapping. The coefficient function is then declared significantly positive at those sites at which the lower band is positive, and significantly negative wherever the upper band is negative.

The simultaneous confidence bands are constructed as follows (Mandel and Betensky, 2008). We draw B random samples with replacement of n data points $(y^*, \boldsymbol{x}^*, \boldsymbol{s}^*)$ from the n observations, and obtain coefficient function estimates $\hat{\boldsymbol{f}}^*_1, \ldots, \hat{\boldsymbol{f}}^*_B$. Ordering the B bootstrap estimates $\hat{f}^*_{(1)}(v) \leqslant \ldots \leqslant \hat{f}^*_{(B)}(v)$ at each point v, we define (for $k \geqslant 1$) an "envelope" $E(k) = \prod_v [\hat{f}^*_{(k)}(v), \hat{f}^*_{(B+1-k)}(v)]$, the Cartesian product of symmetric $100(1-\frac{2k}{B+1})\%$ pointwise confidence intervals at each point. The simultaneous coverage of $E(k)$ can then be estimated by a cross-validation-type procedure (cf. Davison and Hinkley, 1997, Section 4.2.4). Letting $E_{-b}(k)$ be the envelope formed by all except the bth resample, the non-coverage rate of $E(k)$ (i.e., 1 minus the simultaneous coverage rate) is estimated as the proportion of estimates $\hat{\boldsymbol{f}}^*_b$ which exit $E_{-b}(k)$ at some point. This can be calculated as the proportion of $b \in \{1, \ldots, B\}$ such that, for some v, the rank of $\hat{f}^*_b(v)$ among $\hat{f}^*_1(v), \ldots, \hat{f}^*_B(v)$ is $\leqslant k$ or $\geqslant B + 1 - k$.

With high-dimensional, noisy bootstrap coefficient function estimates such as we have encountered with neuroimaging data, this procedure may need to be modified in two ways. First, if more than $100\alpha\%$ of the function estimates attain the (strictly) highest or lowest rank at some point, then even $E(1)$, the band described by all B function estimates, has estimated simultaneous coverage lower than $100(1-\alpha)\%$. This difficulty could theoretically

be surmounted by increasing B, but the required B might be too large to be practical. Instead, when $E(1)$ has inadequate coverage, we propose to use the stretched envelope

$$E^{c_\alpha} = \prod_v \left[\hat{f}(v) + \min\left\{0, c_\alpha(v)[\hat{f}^*_{(1)}(v) - \hat{f}(v)]\right\}, \hat{f}(v) + \max\left\{0, c_\alpha(v)[\hat{f}^*_{(B)}(v) - \hat{f}(v)]\right\}\right],$$

where c_α is a "stretching function" defined in Web Appendix B. Let $E^{c_\alpha}_{-b}$ denote the above simultaneous interval, constructed using the same function c_α, but with $\hat{f}^*_{(1)}(v)$ and $\hat{f}^*_{(B)}(v)$ replaced by the minimum and maximum, respectively, of $\{\hat{f}^*_1(v), \ldots, \hat{f}^*_{b-1}(v), \hat{f}^*_{b+1}(v), \ldots, \hat{f}^*_B(v)\}$. By the construction of c_α, under mild conditions spelled out in Web Appendix B, at most $100\alpha\%$ of the $\hat{\boldsymbol{f}}^*_b$ exit $E^{c_\alpha}_{-b}$. This justifies taking E^{c_α} as our modified confidence band.

Second, automatic smoothing parameter selection methods tend to choose very high-df models for the bootstrap samples, which can result in very unstable estimates for the confidence bands. A detailed analysis of this problem and how to resolve it will be presented in a paper currently in preparation. An approximate remedy is available when adapting the `mgcv` package to implement FPCR. This package allows the user to multiply the df in GCV or AIC by a constant $\gamma > 1$ to counteract overfitting. For the bootstrap samples described below in Section 5, setting this constant to 1.3 appeared to work well.

4.1.3 *Advantages of FPCR Over the Standard Neuroimaging Approach.* Our approach might be preferable to the standard paradigm in some neuroimaging data settings, for both conceptual and practical reasons. Conceptually, to the extent that the mapped quantity s is seen as possibly causing the outcome measured by y, we might prefer to regress y on s rather than vice versa. For instance, levels of the glucose analogue [^{18}F]-FDG are detectable by PET and serve as an index of glucose metabolism throughout the brain. If depression is thought to result from low glucose metabolism, then it might make sense to regress a depression score such as the Hamilton rating scale on [^{18}F]FDG, rather than vice versa.

Practically, the single model (1) enables us to predict y based on a given subject's map of the quantity s. Patient-specific predictions of this kind are unavailable with statistical

parametric mapping, but are of great interest in applications such as those described in the Introduction. Moreover, our paradigm allows one to assess the relative contributions of s and other predictors of interest ($\boldsymbol{x}$ in our notation) toward explaining the variation in y.

4.2 *Testing a Constant-Effect Null vs. a Smooth Alternative*

Another inferential problem is to determine whether the smooth FPCR estimate is preferable to a constant coefficient estimate—i.e., simply regressing the outcome on the mean of each subject's image. We consider the linear case and, for simplicity, assume there are no covariates. Let us refer to the basic FPCR model $\boldsymbol{y} = \alpha\mathbf{1}_n + \boldsymbol{SBV}_q\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, fitted by minimizing (3), as Model 1, and to the simpler model

$$\boldsymbol{y} = \alpha\mathbf{1}_n + (\boldsymbol{S}\mathbf{1}_N/N)\beta_0 + \boldsymbol{\varepsilon} \tag{7}$$

as Model 0, where $\mathbf{1}_n$ and $\mathbf{1}_N$ are columns of 1's. Model 0 is not nested within Model 1, unless $\mathbf{1}_n$ lies in the column space of $\boldsymbol{BV}_q$; but it is nested in Model 2,

$$\boldsymbol{y} = \alpha\mathbf{1}_n + (\boldsymbol{S}\mathbf{1}_N/N)\beta_0 + \boldsymbol{S}^*\boldsymbol{BV}_q^*\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}. \tag{8}$$

Here $\boldsymbol{S}^*$ is obtained from $\boldsymbol{S}$ by replacing each column with the residuals upon regressing on $\boldsymbol{S}\mathbf{1}_N/N$, i.e., decorrelating the images from each image's mean; and $\boldsymbol{V}_q^*$ is derived as above, but using the singular value decomposition of $\boldsymbol{S}^*\boldsymbol{B}$ rather than of $\boldsymbol{SB}$. Model 2 separates the constant or "across-the-board" effect, represented by β_0, from the locally specific effects captured by $\boldsymbol{\beta}_1$. The roughness penalty affects only the latter coefficients.

Since Model 0 is nested in Model 2, they can be compared formally via a restricted likelihood ratio test for a zero variance component, using the null distribution derived by Crainiceanu and Ruppert (2004). The applicability of this test follows from the mixed model viewpoint of Section 3.3, which treats λ as the ratio of the error variance to the random effects variance; Model 0 then translates into a null hypothesis of zero random effects variance, or $\lambda = \infty$. (This infinite null parameter value is avoided in Crainiceanu and Ruppert's application of the test to nonparametric regression, by letting λ denote the *reciprocal* of the

smoothing parameter, thus converting the null hypothesis to $\lambda = 0$.) This test provides a novel way to assess the significance of fluctuations in the coefficient function.

5. Simulations

5.1 *PET Image Data Set*

This section describes simulations designed to test the predictive efficacy of FPCR, and the method's capacity to identify nonzero regions of the coefficient function via the bootstrap-based simultaneous confidence bands discussed in Section 4. Outcomes were generated using an artificial coefficient image (described below in Section 5.2) together with an actual set of PET images. These images, obtained by Parsey et al. (2006) from 27 subjects with major depressive disorder and 41 controls, are maps of binding potential of 5-HT_{1A} receptors, which are thought to play an important role in the disorder. Binding potential, an index of how many receptors bind to a radioligand, provides a measure of the receptors' availability in the brain. Using Statistical Parametric Mapping software (version SPM2), the images were co-registered to a template image in Montreal Neurological Institute standard space, resulting in a set of 69 79×95 transaxial slices. FPCR was applied to slice 12 (i.e., 12th from the bottom). Binding potential BP was replaced by $\log(BP + 1)$ to ensure a less skewed distribution of image values, but one restricted to positive numbers. The knots used to define the radial B-spline basis, both here and in Section 6, were spaced 4 voxels apart in each direction.

5.2 *Simulation Methods*

Outcomes were simulated using the 5-HT_{1A} images and an artificial coefficient function $\boldsymbol{f}_0$ equal to 1 in the circular posterior (i.e., lower) region in Figure 2, 0.5 in the other two circular regions, and 0 elsewhere. Our continuous-outcome simulations assume a true linear model $\boldsymbol{y} = \boldsymbol{S}\boldsymbol{f} + \boldsymbol{\varepsilon}$ (cf. (1)) with coefficient image $\boldsymbol{f} = c\boldsymbol{f}_0$ for various values of $c > 0$ and with $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, I_n)$. For the true model to have a desired coefficient of determination R^2, c is

chosen such that $[\mathrm{Cor}(\boldsymbol{Sf}, \boldsymbol{Sf} + \boldsymbol{\varepsilon})]^2 = R^2$. To simulate binary outcomes analogously, one can define a coefficient of determination for logistic regression as the log likelihood ratio

$$R_L^2 = \frac{\log(L_0) - \log(L_M)}{\log(L_0)}, \tag{9}$$

where L_M is the likelihood of the given model while L_0 is the likelihood of the model containing only an intercept (Menard, 2000). A true logistic model with zero parametric part and coefficient image $c\boldsymbol{f}_0$ implies $E(\boldsymbol{y}) = \exp(c\boldsymbol{Sf}_0)/[1 + \exp(c\boldsymbol{Sf}_0)]$; given a value of R_L^2, this expectation and some algebra enable us to find $c > 0$ such that $\boldsymbol{y}$ simulated from coefficient image $c\boldsymbol{f}_0$ should yield approximately that value of R_L^2.

Using the above approach to simulating outcomes, we performed two studies. The first compared the performance of three methods for both linear and logistic regression:

(a) The penalized B-spline expansion (see (2) for the criterion being minimized in the linear case; the logistic case is analogous), with λ chosen by REML for linear models and by AIC for logistic models.

(b) A method related to FPCR, but without a roughness penalty (i.e., $\lambda = 0$). Here the smoothing parameter is the number of components, which we chose by 5-fold cross-validation from among the values 1–10. As noted in Section 3.2, Escabias et al. (2004) considered a similar approach to functional logistic regression, but without the choice of the number of components by cross-validation. The loss function for cross-validation was squared error in the linear case and classification error in the logistic case.

(c) FPCR with 35 principal components (accounting for 96% of the variation in $\boldsymbol{SB}$), with λ chosen by REML or AIC as above. More components were used than for the previous method, since the roughness penalty effects further dimension reduction.

Each of these methods begins with restriction to a spline basis; this is followed in method (a) by roughness penalization, in (b) by reduction to leading PCs, and in (c) by both.

For linear regression, we generated 100 continuous outcome vectors with equally spaced R^2

values from .2 to .95, and computed each method's scaled prediction error $(\hat{\boldsymbol{f}} - \boldsymbol{f})^T \boldsymbol{S}^T \boldsymbol{S}(\hat{\boldsymbol{f}} - \boldsymbol{f})/\widehat{\mathrm{Var}}(\boldsymbol{S}\boldsymbol{f})$. For logistic regression, we simulated 100 binary outcome vectors with equally spaced R_L^2 values from .2 to .95. Performance was again measured by the above quantity, which in this case can be interpreted as prediction error on the linear scale.

For the second study, 20 binary outcome vectors were generated with each of the R_L^2 values 0.5, 0.6, 0.7, 0.8, and 0.9. Logistic FPCR models with 35 components were then fitted to each simulated data set, and significantly nonzero regions of the coefficient image were inferred from 95% simultaneous confidence bands derived from 2000 bootstrap replicates. Setting the df multiplier γ (see above, Section 4.1.2) to 1.4 yielded an approximate version of Hurvich et al.'s (1998) corrected AIC for smoothing parameter selection in this second set of simulations, but uncorrected AIC was used for the logistic regressions in the first simulation study, for the reasons discussed above in Section 3.3.

5.3 *Prediction Error Results*

Scatterplots of scaled prediction error, for linear and logistic models using the three methods, are displayed in Figure 1. Spline smoothers were fitted for each method using the `gam` function in the R package `mgcv`. In one case shown at the top right of Figure 1(b), logistic FPCR produced a poor result; but all in all, FPCR is seen to outperform the other two methods for both linear and logistic regression, across the range of coefficient of determination values. Moreover, FPCR was much faster than the other two methods, as can be seen from the following processing times, on a MacBook Pro with a 2.16 GHz Intel Core Duo processor. For linear regression, the FPCR models required an average of 0.93 second, compared with 3.3 seconds for the B-spline expansion and 5.8 for the $\lambda = 0$ method; for logistic regression, the mean times in seconds were 1.4, 3.2, and 18.0, respectively.

[Figure 1 about here.]

5.4 *Bootstrap Results*

We discarded a small number of attempted bootstrap fits which generated error messages, all of which appeared to arise from internal errors in `mgcv` involving Cholesky decomposition of penalty matrices. The number of such errors averaged 0.82, with a maximum value of 6, out of the 2000 bootstrap replicates in each simulation. In 99 of the 100 simulations, $E(1)$ had less than 95% simultaneous coverage, and thus we stretched the bands as described in Section 4.1.2. The average value of $\max_v c_\alpha(v)$ was approximately 1.14 (i.e., stretching the interval by up to 14%), with a maximum of 1.37.

The heat maps in Figure 2 show the frequency with which $f(v)$ was deemed significantly positive for each v; we shall refer to this as "detecting" the voxel. (False detections of $f(v)$ as significantly negative did not occur for any v in any of the simulations.) Of the three regions with true $\boldsymbol{f} > 0$, voxels in the region at right proved most difficult to detect, evidently because this region has the lowest binding potential variation and hence contributed least to the outcome. Moreover, this region is highly correlated with, and hence difficult to distinguish from, the corresponding left-hemisphere region, where there were many false detections. The rates of (true) detections for voxels in the three $\boldsymbol{f} > 0$ regions, and of (false) detections for all other voxels, are given in Table 1.

[Figure 2 about here.]

[Table 1 about here.]

In applications, one might be more interested to know how frequently a *region* with $\boldsymbol{f} > 0$ is detected, in the sense that $f(v)$ is found significantly positive for at least one voxel v in the region. The rate of such regional detections, pooled across the five R_L^2 values, was 90% for the posterior region, 84% for the anterior region, and 40% for the region at right.

6. Real Data Example: Amyloid Beta Maps and Alzheimer's Disease

We applied the testing framework of Section 4.2 to an Alzheimer's disease data set. The data consisted of maps of [^{11}C]PIB—a radioligand binding specifically to amyloid beta (Aβ), a protein associated with the disorder—along with scores on the Mini-Mental State Examination (MMSE), a widely used cognitive screening test, obtained from a sample of 30 subjects. The imaging methodology was broadly similar to that for the 5-HT$_{1A}$ images (see Mikhno et al., 2008, for more details), but in this case we used slice 14. An example image is displayed in Figure 3(a). We began by simply regressing the MMSE scores on the mean [^{11}C]PIB value taken over the image; this is Model 0 given by (7), where $\boldsymbol{y}$ is the vector of MMSE scores and $\boldsymbol{S}$ has as its rows the [^{11}C]PIB maps (mean-centered at each voxel). As expected, the mean-signal predictor had a very significant negative effect, with $p = .00014$. We wished to determine whether linear FPCR might improve on Model 0 by pinpointing areas in which Aβ levels were especially consequential for cognitive performance. To that end we fitted Models 1 and 2 of Section 4.2 with $q = 20$ principal components.

Model 1 with λ chosen by REML had df 2.99, as compared to 2 df for Model 0, and attained a slightly better fit than Model 0 as judged by the GCV criterion (11.08 vs. 11.27) (see Figure 3(d)). Using GCV to choose λ, however, led to anomalous findings. The global minimum of the GCV criterion for Model 1 was 9.59, but this was attained at a value of λ giving a very bumpy fit with 14.5 df (see Figure 3(b)); given the sample size of 30, this is quite clearly an overfitted model. The GCV criterion also had a local minimum of 11.06 at a much higher value of λ, resulting in a 2.59-df model (see Figure 3(e)). For Model 2, the REML criterion chose $\lambda = \infty$, i.e., $\boldsymbol{\beta}_1 = 0$ in (8), so that Model 2 was reduced to Model 0. Consistent with this result, the restricted likelihood test cited above in Section 4.2 test failed to reject (p=.36). The GCV criterion once again chose an overfitted model (with 13.9 df), but had a local minimum at $\lambda = \infty$.

[Figure 3 about here.]

In summary, there is some indication that Model 1 with 2.5–3 degrees of freedom may improve slightly on the 2-df Model 0, but this is difficult to judge given the non-nested character of the comparison. Model 2, within which Model 0 is nested, seems not to improve on the latter. These results illustrate the difficulty in fitting FPCR with such small samples, and the care required in choosing the smoothing parameter (Reiss and Ogden, 2009).

7. Conclusions and Future Work

To fully realize its potential usefulness in neuroimaging, our methodology will have to be extended to three-dimensional images. Such an extension will not necessarily entail any conceptual advances: the radial B-spline function is still given by (4), although the distance is now measured in $\mathcal{R}^3$ rather than $\mathcal{R}^2$. Intuitively, using all slices of an image should result in more informative and useful coefficient images than using just one slice. However, the computational burden will rise considerably. Furthermore, the effective dimensionality might be so much higher that much larger sample sizes are needed to obtain useful results. Functional magnetic resonance imaging (fMRI) data sets, which include time-varying scalar outcomes alongside repeated brain images, could more readily fulfill this requirement.

We are currently working on methods for improved smoothing parameter estimation. The `mgcv` package allows one to estimate a spatially adaptive smoothing parameter function $\lambda(v)$ rather than the constant smoothing parameter λ. Alternatively, separate smoothing parameters might be chosen for different components of the image data, such as those chosen by independent component analysis.

In future work we intend to develop a wavelet version of FPCR. Here the roughness penalty smoothing would be replaced by wavelet thresholding. Wavelets would be more amenable than splines to prescreening of coefficients, an often useful first step in dimension reduction.

Acknowledgements

The authors thank the Associate Editor and referees for their very insightful suggestions; Arthur Mikhno and Ramin Parsey for their multifaceted assistance with the PET data; and Brian Caffo, Chung Chang, Ciprian Crainiceanu, Martin Lindquist, Marianthi Markatou, Ian McKeague, Martina Pavlicova, Eva Petkova, Simon Wood, and Hongtu Zhu for helpful discussions. Dr. Reiss's research was supported in part by a Kirschstein-NRSA Predoctoral Fellowship, Grant Number F31 MH073379, from the National Institute of Mental Health.

Supplementary Materials

Web Appendices A and B, referenced in Sections 2 and 4.1.2 respectively, are available under the Paper Information link at the Biometrics website `http://www.biometrics.tibs.org`.

References

Buja, A., and Rolke, W. (2007). Calibration for simultaneity: (re)sampling methods for simultaneous inference with applications to function estimation and functional data. Under revision.

Cai, T. T., and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics* **34,** 2159–2179.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica* **13,** 571–591.

Crainiceanu, C. M., and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component, *Journal of the Royal Statistical Society, Series B* **66,** 165–185.

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge and New York: Cambridge University Press.

Escabias, M., Aguilera, A. M., and Valderrama M. J. (2004). Principal component estimation

of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics* **16,** 365–384.

Green, P. J., and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman and Hall.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60,** 271–293.

Mandel, M., and Betensky, R. A. (2008). Simultaneous confidence intervals based on the percentile bootstrap approach. *Computational Statistics and Data Analysis* **52,** 2158–2165.

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.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician* **54,** 17–24.

Mikhno, A., Devanand, D., Pelton, G. H., Cuasay, K., Gunn, R., Upton, N., Lai, R. Y., Libri, V., Mann, J. J., and Parsey, R. V. (2008). Voxel-based analysis of [^{11}C]PIB scans for diagnosing Alzheimer's disease. *Journal of Neuroscience Methods*, to appear.

Müller, H.-G., and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics* **33,** 774–805.

Parsey, R. V., Oquendo, M. A., Ogden, R. T., Olvet, D. M., Simpson, N., Huang, Y., Van Heertum, R. L., Arango, V., and Mann, J. J. (2006). Altered serotonin 1A binding in major depression: A [carbonyl-C-11]WAY100635 positron emission tomography study. *Biological Psychiatry* **59,** 106–113.

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, 2nd ed. New York: Springer.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reiss, P. T., and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association* **102,** 984–996.

Reiss, P. T., and Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society, Series B* **71**.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge and New York: Cambridge University Press.

Saranli, A., and Baykal, B. (1998). Complexity reduction in radial basis function (RBF) networks by using radial B-spline functions. *Neurocomputing* **18,** 183–194.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B* **62,** 413–428.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B* **65,** 95–114.

Wood, S. N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* **99,** 673–686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.

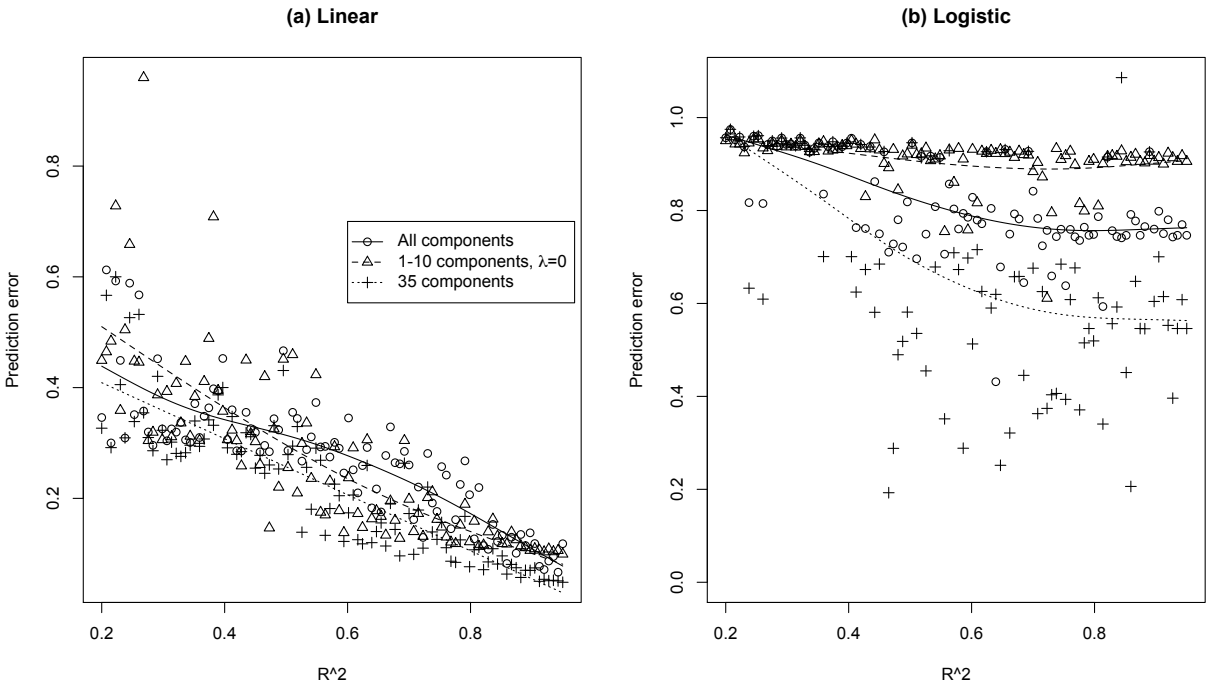

Figure 1. Scaled prediction error as a function of R^2 (or its logistic regression analogue R_L^2; see (9)) of the true model, for (1) the penalized B-spline expansion method (i.e., all principal components included); (2) regression on 1–10 leading principal components of $\boldsymbol{SB}$ without a roughness penalty ($\lambda = 0$); (3) FPCR with 35 components. Linear regression results are shown in (a), and logistic regression results in (b).

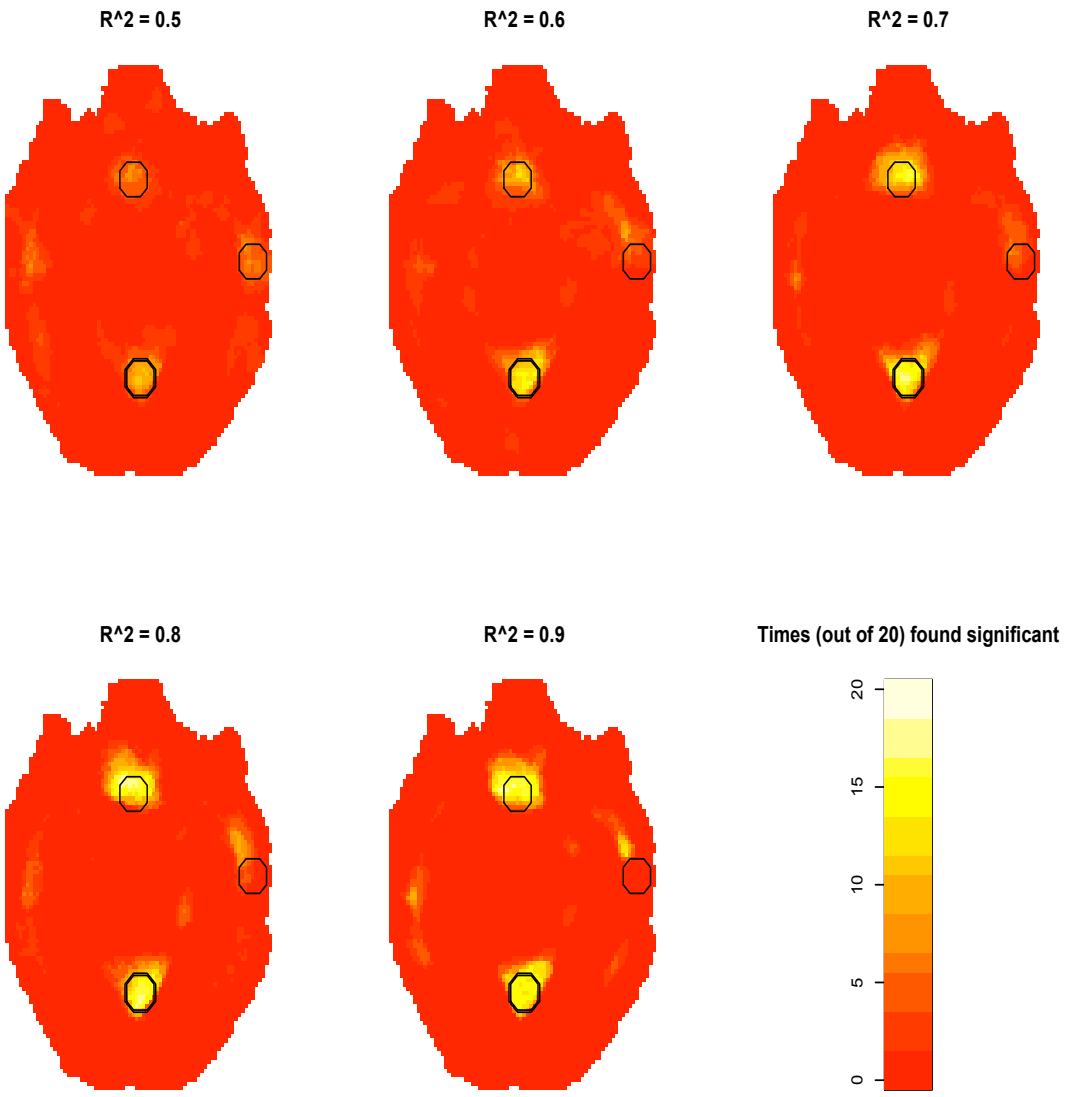

Figure 2. Heat maps indicating how many times, out of 20 simulations at each of the indicated values of R_L^2, each voxel was found to have significantly positive coefficient function value, using the bootstrap procedure of Section 4. The true coefficient function equals $c/2$ within the anterior (upper) and rightmost black circles, c within the posterior circle, and zero elsewhere, where c is a positive number chosen to attain the desired value of R_L^2.

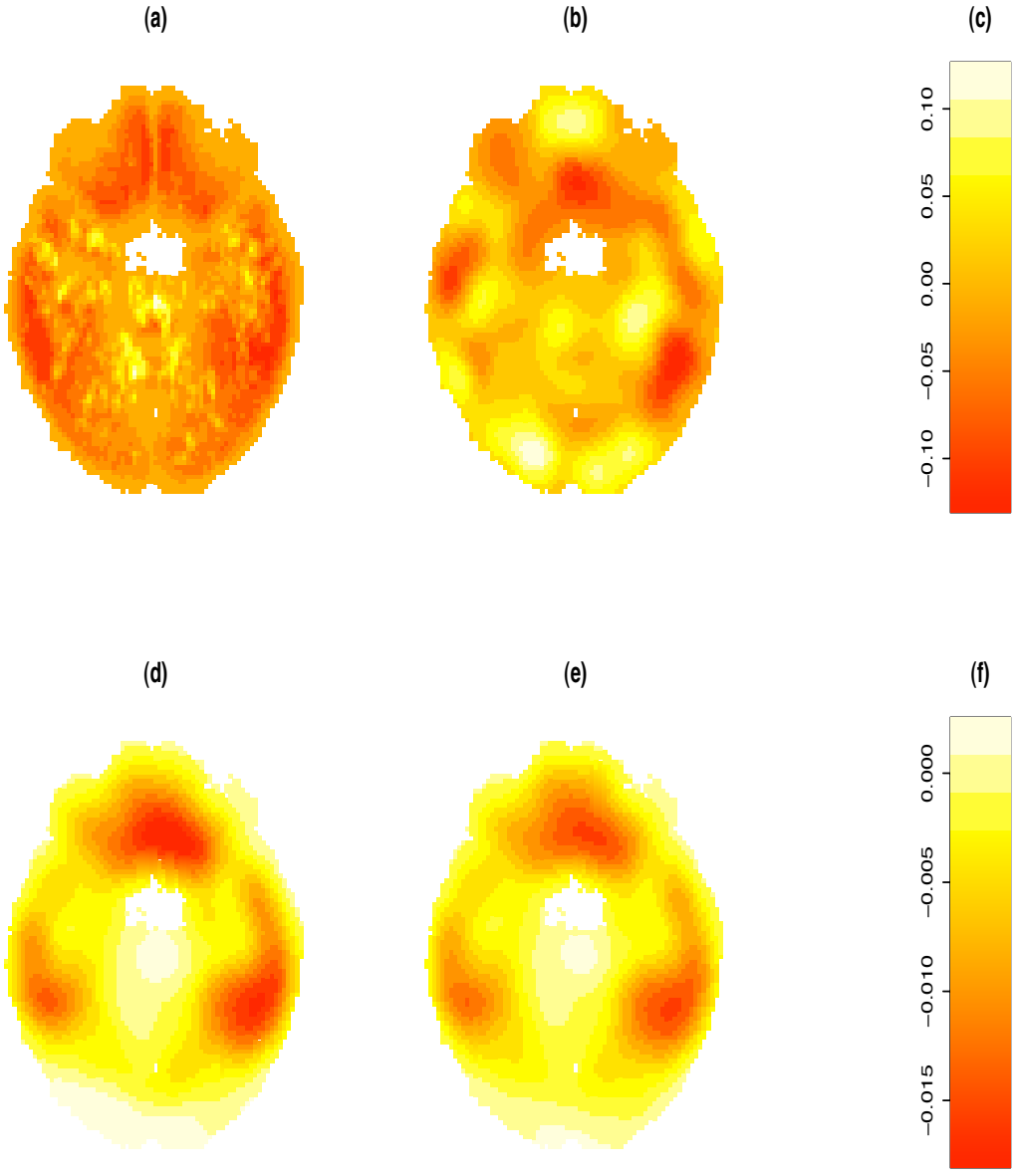

Figure 3. (a) The slice-14 [^{11}C]PIB binding potential image for one of the 30 subjects. (The "hole" just above the center of the image represents a brain stem region with negligible [^{11}C]PIB binding.) (b) The coefficient image estimate obtained by 20-component FPCR with λ chosen by GCV, with color scale given in (c); as argued in the text, this appears to overfit the data. (d)-(e) The estimate with λ maximizing the REML criterion and at the local minimum of the GCV criterion, respectively. These agree quite closely with each other and seem to share major features with the global-minimum-GCV estimate in (b), but have much smaller magnitude than the latter, as indicated by the color scale in (f).

Table 1
Average proportion of simulations in which a given voxel in the regions indicated had significantly positive coefficient image value.

	Region			
R_L^2	Posterior	Anterior	Right	All other
0.5	0.38	0.25	0.24	0.02
0.6	0.55	0.38	0.12	0.02
0.7	0.63	0.51	0.12	0.02
0.8	0.64	0.57	0.09	0.02
0.9	0.64	0.61	0.01	0.02