HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2017 December 28.

Published in final edited form as:

Biometrics. 2017 December ; 73(4): 1092–1101. doi:10.1111/biom.12697.

Pointwise Influence Matrices for Functional-Response Regression

Philip T. Reiss[1,2,3], **Lei Huang**[4], **Pei-Shien Wu**[1], **Huaihou Chen**[5], and **Stan Colcombe**[6]

[1]Department of Child & Adolescent Psychiatry, New York University School of Medicine, USA

[2]Department of Population Health, New York University School of Medicine, USA

[3]Department of Statistics, University of Haifa, Israel

[4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA

[5]Department of Biostatistics, University of Florida, USA

[6]Nathan S. Kline Institute for Psychiatric Research, USA

SUMMARY

We extend the notion of an influence or hat matrix to regression with functional responses and scalar predictors. For responses depending linearly on a set of predictors, our definition is shown to reduce to the conventional influence matrix for linear models. The pointwise degrees of freedom, the trace of the pointwise influence matrix, are shown to have an adaptivity property that motivates a two-step bivariate smoother for modeling nonlinear dependence on a single predictor. This procedure adapts to varying complexity of the nonlinear model at different locations along the function, and thereby achieves better performance than competing tensor product smoothers in an analysis of the development of white matter microstructure in the brain.

Keywords

Bivariate smoothing; Degrees of freedom; Fractional anisotropy; Function-on-scalar regression; Functional nonlinear regression; Neurodevelopmental trajectory; Tensor product spline

1. Introduction

The influence or hat matrix is a central concept in parametric regression (Hoaglin and Welsch, 1978), as well as in semiparametric regression (Ruppert et al., 2003), for which the term *smoother matrix* is often used. Given the growing importance of functional data analysis (Ramsay and Silverman, 2005), there is a need to extend the idea of a influence matrix to regression with functional responses. In this article we define influence matrices in a pointwise sense for functional-response regression, and to show that this definition, and a

Supplementary Materials

Web Appendix A, referenced in Sections 4, 5.2 and 7, and Web Appendix B, referenced in Section 5.1, are available with this paper at the Biometrics website on Wiley Online Library, along with the data and R code for the corpus callosum analyses.

related notion of pointwise degrees of freedom, can be useful tools for studying the performance of functional regression estimators.

Letting $y(s)$ denote a real-valued functional response, defined for s in an interval $\mathcal{S} \in \mathbb{R}$, we assume throughout that

$$E[y(s)|\boldsymbol{x}] = \boldsymbol{x}^T \beta(s), \quad (1)$$

where $\boldsymbol{x} \in \mathbb{R}^p$ is a predictor vector and $\boldsymbol{\beta}(s)$ is a vector-valued function on $\mathcal{S}$. This formulation will allow a unified presentation of linear and nonlinear dependence on scalar predictors, or more specifically, of the following two special cases of (1).

Problem 1: Linear dependence on several predictors

If $\boldsymbol{x} = (1, x_1, \ldots, x_{p-1})^T$, where $x_1, \ldots, x_{p-1}$ are scalar predictors, we have the varying-coefficient model

$$E[y(s)|x_1, \ldots, x_{p-1}] = \beta_0(s) + \sum_{j=1}^{p-1} x_j \beta_j(s) \quad (2)$$

where for each $j, \beta_j(\cdot)$ is a smooth function defined on $\mathcal{S}$. Model (2) is often called a functional linear model (e.g., Ramsay and Silverman, 2005); Reiss et al. (2010) coined the term *function-on-scalar* regression to distinguish (2) from linear models with functional *predictors*.

Problem 2: Nonlinear dependence on a single predictor

Let $p = K_t$ and $\boldsymbol{x} = \boldsymbol{b}_t(t)$, where $t \in \mathbb{R}$ is a predictor variable and $\boldsymbol{b}_t(t) = [b_{t1}(t), \ldots, b_{tK_t}(t)]^T$ is a vector of smooth basis functions defined on the domain of t; our main example will be a B-spline basis that is sufficiently rich to have small approximation error. Here the components $\beta_1(s), \ldots, \beta_{K_t}(s)$ of $\boldsymbol{\beta}(s)$ are again assumed to be smooth, but may not be individually of primary interest; instead, we are interested in estimating the bivariate smooth function

$$f(t, s) = E[y(s)|t] = \boldsymbol{b}_t(t)^T \boldsymbol{\beta}(s). \quad (3)$$

As suggested by the arguments t and s, in some settings (such as our motivating application), the predictor refers to a point in time, while the argument of the functional response denotes a location in space. Thus we sometimes refer to the temporal or spatial dimension in what follows.

For either of the above cases we shall assume that each component of $\boldsymbol{\beta}(s)$ can be represented with negligible approximation error as a linear combination of basis functions such as B-splines, so that (1) becomes $E[y(s)|\boldsymbol{x}] = \boldsymbol{x}^T \Theta \boldsymbol{b}_s(s)$ where $\boldsymbol{b}_s(s) = [b_{s1}(s), \ldots, b_{sK_s}(s)]^T$

is a basis function vector and $\boldsymbol{\Theta}$ is a $p \times K_s$ coefficient matrix. Consequently our estimators for model (1) will have the form

$$\hat{E}[y(s)|\boldsymbol{x}] = \boldsymbol{x}^T \hat{\boldsymbol{\Theta}} \boldsymbol{b}_s(s). \quad (4)$$

The present work was motivated by Problem 2 in the context of human brain development studies, in which the family of functions $\{f(\cdot, s) : s \in \mathscr{S}\}$ may be of particular interest. When t represents age, $f(\cdot,s)$ is the mean, as a function of age, of the quantity measured by y at point s. In the specific example that will be presented below, $\mathscr{S}$ is a set of locations in the brain, and $y(s)$ denotes fractional anisotropy, a measure of white matter integrity, at location s. Thus $f(t, s)$ denotes the mean fractional anisotropy for that location at age t, and the function $f(\cdot, s)$ is what neuroscientists often refer to as the developmental trajectory of fractional anisotropy at location s. As a brief example of the scientific meaning of such trajectories, suppose that for given s, $f(t, s)$ characteristically increases with t up to some point $t_s \equiv \text{argmax}_t f(t,s)$, then decreases. Then the peak age t_s can provide information about typical maturation for location s, and can be compared between diagnostic groups to study the links between psychiatric disorders and brain development (Shaw et al., 2007).

In our data set, fractional anisotropy was measured at 107 voxels, or 1 × 1 × 1 mm volume units, in 146 individuals age 7–48. These voxels, based on registration to a standard image from the FMRIB Software Library, trace a path along a midsagittal cross-section of the corpus callosum (see Fig. 1). We take s to represent arc length along this path, which ranges from $s_1 = 0$ mm (the leftmost point in the figure, toward the back of the brain) to $s_{107} = 110.55$ mm. At right in Fig. 1, a rainbow plot (Hyndman and Shang, 2010), with fractional anisotropy curves color-coded by age, is used to visualize the relationship between age and the functional response. This relationship is not easily discerned from the plot, and may be non-monotonic (and hence nonlinear) in some locations.

In ordinary nonparametric estimation of a function $g(t) = E(y|t)$ based on data pairs $(t_1,y_1),\ldots,(t_n,y_n)$, an important role is played by the $n \times n$ *influence* or *smoother matrix* $\boldsymbol{H}$ such that the response vector $\boldsymbol{y} = (y_1,\ldots, y_n)^T$ is related to a fitted value vector $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T = [\hat{g}(t_1), \ldots, \hat{g}(t_n)]^T$ by

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}. \quad (5)$$

(Since its effect is to add a "hat" to $\boldsymbol{y}$, $\boldsymbol{H}$ is also known as the *hat matrix.)* In particular, the trace of $\boldsymbol{H}$ serves as a popular definition of the effective degrees of freedom, a measure of the complexity of a smoother (Wahba, 1983; Buja et al., 1989). When trajectory estimates $t \mapsto \hat{f}(t, s)$ are of interest, it would be very useful to extend these concepts to the curve estimates $\hat{f}(\cdot, s)$ for each point $s \in \mathscr{S}$ at which the functional responses are observed. However, for estimators of the type considered here, which share information among points $s \in \mathscr{S}$, it is not clear how to extend the definition of $\boldsymbol{H}$ to a *pointwise* influence matrix for a given $s \in \mathscr{S}$.

In this paper we develop a notion of pointwise influence matrices for estimators of the form (4), and demonstrate its conceptual and methodological utility. After introducing some needed notation and assumptions in Section 2, we define the pointwise influence matrix, and a corresponding notion of pointwise effective degrees of freedom, in Section 3. In Section 4 we provide some support for our definition by showing that, for Problem 1 (the linear case), it reduces to the influence matrix as ordinarily defined. Section 5 presents a result about pointwise influence matrices for "post-smoothed" estimators which suggests that a simple but apparently novel method should work well for the motivating class of applications; and this is corroborated by the brain imaging data analysis in Section 6. Section 7 offers some concluding remarks.

2. Further notation and assumptions

In what follows, $\mathbf{1}_m$ denotes a vector of m 1's, while $\boldsymbol{e}_{r;m}$ is a vector in $\mathbb{R}^m$ with 1 in the rth position and 0 elsewhere. We assume independent data pairs $[\boldsymbol{x}_i, y_i(\cdot)]$, $i = 1,\ldots,n$, where $n > p$, with each of the functional responses $y_i(\cdot)$ observed at a common set of points $s_1,\ldots, s_L$, giving rise to an $n \times L$ response matrix

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1L} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nL} \end{pmatrix} = \begin{pmatrix} y_1(s_1) & \cdots & y_1(s_L) \\ \vdots & \ddots & \vdots \\ y_n(s_1) & \cdots & y_n(s_L) \end{pmatrix}.$$

Irregularly sampled functions can be accommodated by adding a presmoothing step (cf. Chiou et al., 2003). Let $\boldsymbol{y}_i^T$ and $\boldsymbol{y}_\ell$ denote the ith row and ℓth column of $\boldsymbol{Y}$, respectively, and let $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y}) \in \mathbb{R}^{nL}$.

An estimator of the general type considered here, i.e., having the form (4), yields an $n \times L$ fitted value matrix

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\Theta}}\boldsymbol{B}_s^T, \quad (6)$$

where $\boldsymbol{X}$ is the matrix with ith row $\boldsymbol{x}_i^T\ (i=1,\ldots,n)$ and $\boldsymbol{B}_s$ is the $L \times K_s$ matrix with ℓth row $[b_{s1}(s_\ell),\ldots, b_{sK_s}(s_\ell)]$. A standard result on the vec operator and Kronecker products yields

$$\hat{\boldsymbol{y}} = (\boldsymbol{B}_s \otimes \boldsymbol{X})\hat{\theta} \quad (7)$$

where $\hat{\boldsymbol{y}} = \mathrm{vec}(\hat{\boldsymbol{Y}})$ and $\hat{\theta} = \mathrm{vec}(\hat{\boldsymbol{\Theta}})$.

When the above K_s basis functions are B-splines, these are typically used in conjunction with a roughness penalty (Green and Silverman, 1994), which is implemented by means of a a symmetric positive semidefinite $K_s \times K_s$ matrix $\boldsymbol{P}_s$ such that, for a given function $g(s) = \boldsymbol{\gamma}^T\boldsymbol{b}_s(s)$, we have $\boldsymbol{\gamma}^T\boldsymbol{P}_s\boldsymbol{\gamma} = r_s(g)$ where $r_s(g)$ is some measure of the roughness of g. Popular choices include the second-derivative penalty matrix $\boldsymbol{P}_s = [\int b_{si}''(s)\, b_{sj}''(s)\, ds]_{i,j=1,\ldots,K_s}$, for

which $\gamma^T \boldsymbol{P}_s \gamma = r_s(g) \equiv \int_{\mathcal{T}} g''(s)^2 ds$, or difference penalties as in Eilers and Marx (1996).

We can then define the $L \times L$ smoother matrix $\boldsymbol{H}_S^{\lambda_s} = \boldsymbol{B}_s (\boldsymbol{B}_s^T \boldsymbol{B}_s + \lambda_s \boldsymbol{P}_s)^{-1} \boldsymbol{B}_s^T$, where $\lambda_s \geq 0$ is the tuning parameter that governs the extent of the smoothing. For Problem 2, we have $\boldsymbol{X} = \boldsymbol{B}_t$, defined analogously to $\boldsymbol{B}_s$ but with temporal-dimension smooth basis functions $\boldsymbol{b}_t(t)$ as in (3); we likewise define the $K_t \times K_t$ penalty matrix $\boldsymbol{P}_t$, and the $n \times n$ smoother matrix $\boldsymbol{H}_t^{\lambda_t} = \boldsymbol{B}_t (\boldsymbol{B}_t^T \boldsymbol{B}_t + \lambda_t \boldsymbol{P}_t)^{-1} \boldsymbol{B}_t^T$.

We shall require two assumptions regarding smoothing in the the spatial direction:

$$\boldsymbol{B}_s \mathbf{1}_{K_s} = \mathbf{1}_L, \quad (8)$$

$$\boldsymbol{P}_s \mathbf{1}_{K_s} = 0. \quad (9)$$

In particular, assumption (8) holds for a *B*-spline basis in one dimension (Schumaker, 2007), while (9) holds for a derivative or difference penalty (Eilers and Marx, 1996).

3. The pointwise influence matrix

All of the estimation procedures presented below (in Sections 4 and 5) produce fitted values that can be written as $\hat{\boldsymbol{y}} = \mathscr{H} \boldsymbol{y}$ for some $nL \times nL$ matrix

$$\mathscr{H} = \begin{pmatrix} \mathscr{H}_{11} & \cdots & \mathscr{H}_{1L} \\ \vdots & \ddots & \vdots \\ \mathscr{H}_{L1} & \cdots & \mathscr{H}_{LL} \end{pmatrix}, \quad (10)$$

where each of the blocks $\mathscr{H}_{\ell_1 \ell_2}$ is $n \times n$. Moreover, for all of the procedures we consider one can write $\hat{\theta} = \boldsymbol{C} \boldsymbol{y}$, for some $pK_S \times nL$ matrix $\boldsymbol{C}$, in (7), so the influence matrix has the form

$$\mathscr{H} = (\boldsymbol{B}_s \otimes \boldsymbol{X}) \boldsymbol{C}. \quad (11)$$

While an explicit estimate $\hat{\theta}$ is required to estimate the conditional mean of $y(s)$ for arbitrary s, some procedures are more conveniently described by merely specifying $\hat{\boldsymbol{Y}}$.

Some intuition regarding the block influence matrix (10) can be gained from Fig. 2. Subfigure (a) displays a portion of $\mathscr{H}$ obtained by fitting a tensor product smooth (as in Section 5) to a subset of the corpus callosum data. Had we fitted separate models at each voxel, the nonzero entries in the influence matrix, representing influence of the responses on the fitted values, would be confined to diagonal blocks such as those outlined in black. The sharing of information across locations is expressed as a blockwise blurring in the horizontal

direction. Formally, whereas with separate models the fitted values at the ℓth location would be $\hat{\boldsymbol{y}}_{\cdot\ell} = \mathscr{H}_{\ell\ell}\boldsymbol{y}_{\cdot\ell}$, here we have

$$\hat{\boldsymbol{y}}_{\cdot\ell} = \sum_{\ell^*=1}^{L} \mathscr{H}_{\ell\ell^*}\boldsymbol{y}_{\cdot\ell^*}, \quad (12)$$

implying that the fitted values are influenced not just by the ℓth location data via $\mathscr{H}_{\ell\ell}$, but by all of the response data via $\mathscr{H}_{\ell 1}, \ldots, \mathscr{H}_{\ell L}$. This motivates the following two definitions.

Definition 1

The *pointwise influence matrix* or *pointwise hat matrix* at location ℓ is the $n \times n$ matrix

$$\mathscr{H}_{\ell\cdot} = \sum_{\ell^*=1}^{L} \mathscr{H}_{\ell\ell^*} \quad (13)$$

$$= \left(\boldsymbol{e}_{\ell;L}^{T} \otimes \boldsymbol{I}_n\right) \mathscr{H} \left(\mathbf{1}_L \otimes \boldsymbol{I}_n\right). \quad (14)$$

Definition 2

The *pointwise effective degrees of freedom*, or simply *pointwise degrees of freedom*, at location ℓ is

$$\mathrm{DF}_\ell = \mathrm{tr}\left(\mathscr{H}_{\ell\cdot}\right) = \sum_{\ell^*=1}^{L} \mathrm{tr}\left(\mathscr{H}_{\ell\ell^*}\right) = \sum_{i=1}^{n}\sum_{\ell^*=1}^{L} \frac{\partial \hat{y}_{i\ell}}{\partial y_{i\ell^*}}.$$

As an aid to intuition regarding Definition 2, consider a toy example with $n = 5$ observations and $L = 4$ locations, so that the overall influence matrix (10) comprises a 4×4 grid of 5×5 blocks. The conventional degrees of freedom for the 2nd-location model, which would be appropriate if separate models were fitted at each location, is the sum of the shaded values in Fig. 2(b). The pointwise degrees of freedom DF_2, which takes into account the influence of neighboring locations, is the sum of the shaded values in Fig. 2(c).

A formal justification for Definition 1 is provided below by Theorem 1, but for now we offer a heuristic interpretation that is valid when $\hat{\boldsymbol{y}}_{\cdot\ell}$ lies in the column space of $\mathscr{H}_{\ell\cdot}$. On that assumption we can rewrite (12) in a form resembling (5), namely as

$$\hat{\boldsymbol{y}}_{\cdot\ell} = \mathscr{H}_{\ell\cdot}\boldsymbol{y}_{\cdot\ell}^{\dagger} \quad (15)$$

where

$$\boldsymbol{y}^{\dagger}_{\cdot\ell}=\left(\sum_{\ell^*=1}^{L}\mathscr{H}_{\ell\ell^*}\right)^{-}\sum_{\ell^*=1}^{L}\mathscr{H}_{\ell\ell^*}\boldsymbol{y}_{\cdot\ell^*}, \quad (16)$$

with $(\cdot)^-$ denoting a generalized inverse. The ℓth-location fitted value vector $\hat{\boldsymbol{y}}_{\cdot\ell}$ can thus be viewed as resulting from a two-step process: first, in (16), we share information among the $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_L$ by forming their weighted average $\boldsymbol{y}^{\dagger}_{\cdot\ell}$, with matrix-valued weights $\mathscr{H}_{\ell 1},\ldots,\mathscr{H}_{\ell L}$; then, in (15), we premultiply by the pointwise influence matrix to obtain $\hat{\boldsymbol{y}}_{\cdot\ell}$.

4. Application to varying-coefficient models

In this section we study the pointwise influence matrix in the context of Problem 1, i.e., the varying coefficient linear model (2). We show that for two natural approaches to estimating this model, the pointwise influence matrix reduces to the usual hat matrix for linear regression. We refer to these two approaches as *contemporaneous smoothing* and *post-smoothing*. In contemporaneous smoothing one minimizes a penalized criterion and thereby simultaneously estimates the linear relationship at each point s and maintains smoothness across points, as opposed to the post-smoothing approach of first fitting separate initial models at each point and then smoothing the results. Post-smoothing can be more computationally efficient than contemporaneous smoothing, but its favorable performance relies on being able to fit the initial models with a modicum of accuracy.

A general expression for the contemporaneous-smoothing coefficient estimate is

$$\hat{\theta}=\arg\min_{\theta}\left[\{\boldsymbol{y}-(\boldsymbol{B}_s\otimes\boldsymbol{X})\theta\}^T(\hat{\sum}^{-1}\otimes\boldsymbol{I}_n)\{\boldsymbol{y}-(\boldsymbol{B}_s\otimes\boldsymbol{X})\theta\}+\theta^T(\boldsymbol{P}_s\otimes\Lambda_s)\theta\right], \quad (17)$$

where $\hat{\sum}$ is an $L\times L$ matrix and $\boldsymbol{\Lambda}_s=\mathrm{Diag}(\lambda_{s,1},\ldots,\lambda_{s,p})$; more transparently, the penalty $\boldsymbol{\theta}^T(\boldsymbol{P}_s\otimes\boldsymbol{\Lambda}_s)\boldsymbol{\theta}$ in (17) can be written as $\sum_{j=1}^{p}\lambda_{s,j}\theta_{j\cdot}^T\boldsymbol{P}_s\theta_{j\cdot}$ where $\theta_{j\cdot}^T$ is the jth row of $\boldsymbol{\Theta}$, i.e., as the sum of separate penalties for the p coefficient functions $\beta_k(s)=\theta_{k\cdot}^T\boldsymbol{b}_s(s)$, $k=1,\ldots,p$. If we take $\hat{\sum}=\boldsymbol{I}_L$ then the first term on the right side of (17) reduces to the squared Frobenius norm $||\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Theta}\boldsymbol{B}_s^T||_F^2$, and $\hat{\theta}$ is a penalized ordinary least squares (OLS) estimate as in Chapter 13 of Ramsay and Silverman (2005). On the other hand, if $\hat{\sum}^{-1}$ is an estimate of the inverse of the covariance matrix $\boldsymbol{\Sigma}=[\mathrm{Cov}\{y(s_\ell),y(s_{\ell^*})\}|\boldsymbol{x}]_{\ell,\ell^*=1,\ldots,L}$, typically derived from the residuals of an OLS fit, then $\hat{\theta}$ is a penalized generalized least squares (GLS) estimate (Reiss et al., 2010).

A post-smoothing estimator via local polynomial smoothing is studied by Fan and Zhang (2000) (see also Zhu et al., 2014). An analogous penalized approach proceeds as follows:

1. Obtain a $p\times L$ matrix of initial pointwise least squares estimates

$$\tilde{\beta}=\begin{pmatrix}\tilde{\beta}_1(s_1) & \cdots & \tilde{\beta}_1(s_L)\\ \vdots & \ddots & \vdots\\ \tilde{\beta}_p(s_1) & \cdots & \tilde{\beta}_p(s_L)\end{pmatrix}=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

2. For $j = 1,\ldots, p$, obtain final estimates $\hat{\beta}_j(s_1),\ldots,\hat{\beta}_j(s_L)$ by applying a smoother matrix $\boldsymbol{H}_s^{\lambda_{s,j}}=\boldsymbol{B}_s\left(\boldsymbol{B}_s^T\boldsymbol{B}_s+\lambda_{s,j}\boldsymbol{P}_s\right)^{-1}\boldsymbol{B}_s^T$ to the jth row of $\tilde{\beta}$.

This heuristic description suggests the more formal specification

$$\hat{\theta}_{j\cdot}^T=e_{j;p}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{B}_s\left(\boldsymbol{B}_s^T\boldsymbol{B}_s+\lambda_{s,j}\boldsymbol{P}_s\right)^{-1} \quad (18)$$

for the j th row of the $p \times K_s$ coefficient estimate matrix $\hat{\Theta}$.

We can now state the following result, which is proved in Web Appendix A.

Theorem 1

Assume that $\boldsymbol{X}$ is of rank p and that (8) and (9) hold. For either the contemporaneous smoothing estimator (17) or the post-smoothing estimator (18), the pointwise influence matrix (13) equals $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ for $\ell=1,\ldots,L$.

In short, Theorem 1 says that the pointwise influence matrix for the varying-coefficient model (2) directly generalizes the usual influence matrix for linear regression. This property serves as a form of validation for Definition 1; indeed, we can conceive of no other property that would more cogently validate the definition.

5. Post-smoothing and nonlinear trajectories

5.1 Pointwise influence matrices for a class of post-smoothing procedures

In the above post-smoothed estimator for model (2) (Problem 1), if $\lambda_{s,1} = \ldots = \lambda_{s,p} = \lambda_s$, i.e., a common spatial smoother matrix $\boldsymbol{H}_s^{\lambda_s}$ is used for each of the p coefficients, then the final fitted value matrix is

$$\hat{\boldsymbol{Y}}=\tilde{\boldsymbol{Y}}\boldsymbol{H}_s^{\lambda_s} \quad (19)$$

where $\tilde{\boldsymbol{Y}}$ is the initial fitted value matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. More generally, consider the class of post-smoothed model fits of the form (19), i.e., those obtainable by forming an initial fit $\tilde{\boldsymbol{Y}}$ and then postmultiplying by a spatial smoother matrix $\boldsymbol{H}_s^{\lambda_s}$. For example, the fast sandwich smoother

$$\hat{\boldsymbol{Y}}=\boldsymbol{H}_t^{\lambda_t}\boldsymbol{Y}\boldsymbol{H}_s^{\lambda_s} \quad (20)$$

(Xiao et al., 2013) is of this form, with $\tilde{\boldsymbol{Y}}=\boldsymbol{H}_t^{\lambda_t}\boldsymbol{Y}$. Given a fitted value matrix (19) such that $\tilde{\boldsymbol{y}}=\mathrm{vec}\left(\tilde{\boldsymbol{Y}}\right)=\tilde{\mathscr{H}}\boldsymbol{y}$ for some matrix $\tilde{\mathscr{H}}$, we can define the initial pointwise influence matrix $\tilde{\mathscr{H}}_\ell$ analogously to Definition 1, but with $\tilde{\mathscr{H}}$ replacing $\mathscr{H}$ in (13), as well as the initial pointwise degrees of freedom $\tilde{\mathrm{DF}}_\ell=\mathrm{tr}\left(\tilde{\mathscr{H}}_{\ell\cdot}\right)$. We then have the following result, which is proved in Web Appendix B.

Theorem 2—Using the subscript *ij* to denote the *(i, j)* entry of a matrix, we have

$$\begin{pmatrix} \mathscr{H}_{1\cdot,ij} \\ \vdots \\ \mathscr{H}_{L\cdot,ij} \end{pmatrix} = \boldsymbol{H}_s^{\lambda_s} \begin{pmatrix} \tilde{\mathscr{H}}_{1\cdot,ij} \\ \vdots \\ \tilde{\mathscr{H}}_{L\cdot,ij} \end{pmatrix}, \tag{21}$$

for $i, j = 1,\ldots, n$. Moreover,

$$\begin{pmatrix} \mathrm{DF}_1 \\ \vdots \\ \mathrm{DF}_L \end{pmatrix} = \boldsymbol{H}_s^{\lambda_s} \begin{pmatrix} \tilde{\mathrm{DF}}_1 \\ \vdots \\ \tilde{\mathrm{DF}}_L \end{pmatrix}. \tag{22}$$

Theorem 2 says, first, that the pointwise influence matrices $\mathscr{H}_{1\cdot},\ldots,\mathscr{H}_L$ can be obtained by elementwise smoothing of the initial pointwise influence matrices $\tilde{\mathscr{H}}_{1\cdot},\ldots,\tilde{\mathscr{H}}_L$ with smoother matrix $\boldsymbol{H}_s^{\lambda_s}$; and second, that applying that smoother to the initial pointwise degrees of freedom vector yields the final pointwise degrees of freedom vector.

5.2 An adaptive post-smoothing procedure for nonlinear trajectories

Theorem 2 will turn out to have interesting implications for Problem 2, estimating the bivariate smooth function $f(t, s)$ (3) given a set of functional responses associated with a scalar predictor. Problem 2 admits of the same two types of approaches to spatial smoothness that we described above for Problem 1 (see Table 1). A *contemporaneous smoothing* approach is the tensor product smoother $\hat{f}(t,s)=\boldsymbol{b}_t(t)^T\hat{\boldsymbol{\Theta}}\boldsymbol{b}_{\mathbf{s}}(\mathrm{s})$ with

$$\hat{\boldsymbol{\Theta}}=\underset{\Theta}{\arg\min}\left[\left\|\left(\boldsymbol{Y}-\boldsymbol{B}_{\mathbf{t}}\boldsymbol{\Theta}\boldsymbol{B}_{\mathbf{s}}^{\mathbf{T}}\right)\hat{\sum}^{-1/2}\right\|_{\mathbf{F}}^{\mathbf{2}}+\lambda_{\mathbf{t}}\mathbf{p}_{\mathbf{t}}(\boldsymbol{\Theta})+\lambda_{\mathbf{s}}\mathbf{p}_{\mathbf{s}}(\boldsymbol{\Theta})\right], \tag{23}$$

where p_t, p_s penalize roughness in the *t*-direction and the *s*-direction, respectively (e.g., Eilers and Marx, 2003; Wood, 2006b). As for Problem 1, one can set $\hat{\sum}=\boldsymbol{I}$ in (23), resulting in an OLS smoother, or let $\hat{\sum}^{-1}$ be an estimate of the inverse of $[\mathrm{Cov}\{y(s_\ell), y(s_{\ell^*})\}|t]_{\ell,\ell^*=1,\ldots,L}$, giving a GLS smoother. An example of a *post-smoothing*

approach is the sandwich smoother (20), although that method was not conceived as such by Xiao et al. (2013).

For the sandwich smoother, the initial pointwise influence matrix $\tilde{\mathscr{H}}_{\ell\cdot}$ equals $\boldsymbol{H}_t^{\lambda_t}$ for each ℓ. This, combined with Theorem 2 and Lemma 1 in Web Appendix A, implies that the sandwich smoother has final pointwise influence matrix $\boldsymbol{H}_t^{\lambda_t}$, and thus $\mathrm{DF}_\ell = \mathrm{tr}(\boldsymbol{H}_t^{\lambda_t})$, for each t. Recall, however, that our motivating class of applications involves estimating trajectories $f(\cdot, s)$ with possibly varying complexity. For this purpose it seems suboptimal to have the same pointwise degrees of freedom at each location. We therefore propose an alternative post-smoothing procedure, which generalizes the sandwich smoother to allow for separate temporal smoothing parameters at each location:

1. Derive the initial estimate $\tilde{\boldsymbol{Y}} = \left(\tilde{\boldsymbol{y}}_{\cdot 1} \cdots \tilde{\boldsymbol{y}}_{\cdot L}\right)$ with $\tilde{\boldsymbol{y}}_{\cdot\ell} = \boldsymbol{H}_t^{\lambda_{t,\ell}} \boldsymbol{y}_{\cdot\ell}$ where, for each ℓ, $\lambda_{t,\ell}$ is chosen by restricted maximum likelihood (Ruppert et al., 2003; Reiss and Ogden, 2009) for an optimal ℓth-location smooth. Reiss et al. (2014) propose a fast procedure for this.

2. Obtain the final estimate $\hat{\boldsymbol{Y}} = \tilde{\boldsymbol{Y}} \boldsymbol{H}_s^{\lambda_s}$, where the spatial smoothing parameter λ_s may be chosen by multifold cross-validation.

This procedure avoids inversion of $nL \times nL$ matrices and is thus much faster than the contemporaneous methods, as will be seen below. In this two-step procedure, $\tilde{\mathrm{DF}}_\ell = \mathrm{tr}(\boldsymbol{H}_t^{\lambda_{t,\ell}})$, for $\ell = 1, \ldots, L$. If we think of $\left(\tilde{\mathrm{DF}}_1, \ldots, \tilde{\mathrm{DF}}_L\right)^T$ as a vector of noisily measured complexity indices for $\{f(\cdot, s_\ell): \ell = 1, \ldots, L\}$, then by Theorem 2, the final pointwise degrees of freedom vector $(\mathrm{DF}_1, \ldots, \mathrm{DF}_L)^T$ can be viewed as a denoised version of these measurements, with the denoising performed by application of the smoother matrix $\boldsymbol{H}_s^{\lambda_s}$. In this sense the final fit adapts in a smooth manner to the varying complexity of $f(\cdot, s_\ell)$. The application described next illustrates how such adaptivity can be beneficial in practice.

6. Application: Development of corpus callosum microstructure

We now return to the corpus callosum fractional anisotropy data described in the introduction. A full description of the image data processing, as well as an illuminating set of analyses, can be found in Imperati et al. (2011). Here we aim to estimate the mean fractional anisotropy $f(t,s)$ where t denotes age and s denotes location, expressed as arc length along the path depicted in Fig. 1.

Before presenting our modeling results, let us consider some evidence for nonlinear change in mean fractional anisotropy with respect to age. For $\ell = 1, \ldots, 107$ we performed a restricted likelihood ratio test (Crainiceanu and Ruppert, 2004), implemented by the vows package (Reiss et al., 2016) for R (R Core Team, 2016), to test the null hypothesis that $f(\cdot, s_\ell)$ is linear—that is, mean fractional anisotropy for the ℓth voxel changes linearly with age—against the alternative of nonlinear change. Figure 3 shows the resulting p-values for each voxel. These p-values are not adjusted for multiple testing, since our aim here is descriptive

rather than to test the global null hypothesis that $f(\cdot,s)$ is linear for each ℓ. Also shown are separate penalized spline smooths at three voxels, with the smoothing parameter chosen by restricted maximum likelihood. The first of these voxels is located in the sensorimotor portion of the brain; the second, in the posterior (rear) portion of the prefrontal lobes, with projections into the inferior and middle frontal gyri; and the third, in the anterior (front) portion of the prefrontal lobes. In the first and third voxels, it appears that mean fractional anisotropy attains a peak in young adulthood, and then declines. These nonlinear trajectories are consistent with the p-values of .019 and .001, respectively, for the two voxels. But for the second voxel, there is no strong evidence for nonlinear change. Features of these curves, such as peaks, are of biological interest, as they may provide insight into characteristic patterns of development for different cognitive abilities. However, the data are quite noisy, suggesting that estimation might be improved by sharing information across voxels.

We applied the post-smoothing procedure of Section 5.2 to this data set, and compare its performance to that of the OLS and GLS smoothers based on (23) with automatic selection of λ_t and λ_s provided by the R package mgcv (Wood, 2006a, 2011). For OLS, we chose these tuning parameters by restricted maximum likelihood, ignoring within-function dependence. While this approach may be suboptimal, it finds some support in recent work (e.g., Krivobokova and Kauermann, 2007; Crainiceanu et al., 2012). For GLS, we treated the prewhitened response vectors $\hat{\sum}^{-1/2} y_{i\cdot}\ (i=1,\ldots,n)$ as the data, and again chose λ_t and λ_s by restricted maximum likelihood, as in Reiss et al. (2010). We used 15-dimensional B-spline bases in the temporal direction, but since the amount of detail to capture is much greater in the spatial direction (i.e., differences among brain regions), we used more B-splines in that direction: 30 for the post-smoothing method, and slightly fewer, 25, for the contemporaneous methods to mitigate their much higher computation time. The postsmoothing procedure requires a grid search for λ_s; we selected $\log(\lambda_s)$ by five-fold crossvalidation from the candidate values −20, −19,…, 10. The post-smoothing fit required 2.1 seconds, versus 23.4 seconds for OLS and 42.6 seconds for GLS.

The rainbow plots (Hyndman and Shang, 2010) in the upper panels of Fig. 4 make it easier to examine the shape of the estimates $\hat{f}(\cdot, s)$ for different locations s. Vertical lines are drawn at the same three voxels as in Fig. 3, and vertical progression from blue to red at a given arc length indicates that the estimate $\hat{f}(\cdot, s)$ is monotonic for that s. Recall that separate nonparametric regressions suggested that fractional anisotropy attained a peak for the first and third voxels, but not for the second. The OLS estimates $\hat{f}(\cdot, s)$ are very erratic, with apparently spurious fluctuations with respect to age. The GLS estimates, on the other hand, indicate that mean fractional anisotropy changes monotonically in all three regions, suggesting oversmoothing with respect to age. Only the post-smoothed estimates appear consistent with the scatterplots in Fig. 3: they capture the peak during young adulthood in the sensorimotor and anterior prefrontal regions, whereas in the posterior prefrontal region, the rainbow pattern suggests linear decrease with age.

These impressions are borne out by the lower left panel of Fig. 4, in which the pointwise degrees of freedom for OLS is seen to be uniformly high while that for GLS is uniformly

low. In line with Theorem 2, the pointwise degrees of freedom for the post-smoothed method is a smoothed version of the degrees of freedom for the separate smooths, so this method adapts to spatial variation in temporal-direction complexity. The lower right panel of Fig. 4 presents functional R^2 values (using a definition similar to that of Müller and Yao, 2008) for the three methods against prediction error estimates, based on repeated five-fold crossvalidation (Burman, 1989) with 10 different splits of the observations into five validation sets. This figure provides further evidence that OLS and GLS overfit and underfit the data, respectively, but suggests that in this case underfitting is not very detrimental to predictive performance. It might be objected that these methods' over-and underfitting is attributable to poor automatic selection of the smoothing parameter λ_t in (23). What these results suggest, however, is that the use of a fixed λ_t, in itself, severely limits the ability of these contemporaneous smoothers to adapt to the varying complexity of $f(\cdot, s)$ for different s.

Figure 5 presents the estimates $\hat{f}(\cdot, s_\ell)$, for $\ell = 91,\ldots, 98$, of the mean fractional anisotropy as a smooth function of age. These eight voxels correspond to arc lengths 94.1–101.5, a range whose right terminus corresponds to the rightmost peak observed in Fig. 4, and to the most prominent trough in the *p*-value plot of Fig. 3. The upper left subfigure displays separate curve estimates for the eight voxels. In line with the results shown in Figures 4, the OLS and GLS curves clearly overfit and underfit, respectively, whereas the post-smoothed estimates appear to do a reasonable job of denoising the separate curve estimates, as Theorem 2 might lead us to expect.

7. Discussion

In introducing the pointwise influence matrix, we have to a large extent focused on the trace of that matrix, i.e., the pointwise degrees of freedom. We offer here a few remarks on some related (and not-so-related) work on the notion of degrees of freedom.

For linear modeling procedures in which the fitted value matrix cannot be expressed in the linear form (5) for fixed $\boldsymbol{H}$, Ye (1998) proposed the generalized degrees of freedom $\mathrm{GDF} = \sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i)/\sigma^2$, where we assume uncorrelated mean-zero errors with common variance σ^2. Following Efron (2004) and Efron et al. (2004), GDF is now often referred to as simply the degrees of freedom. While GDF indeed generalizes the degrees of freedom of Wahba (1983), i.e. tr($\boldsymbol{H}$) in (5), to a much broader class of fitting procedures (e.g., Tibshirani and Taylor, 2012), it is *less* general than tr($\boldsymbol{H}$) in that the latter entails no assumptions about the error structure. Notwithstanding the caveats recently issued by Kaufman and Rosset (2014) and Janson et al. (2015) about GDF, we believe that the degrees of freedom of Wahba (1983) remains useful as a description of the complexity of smooth model fits, and our concern here has been to extend it in a completely different direction: from scalar to functional responses. Extending GDF to functional responses would be a worthy goal for future research.

The ordinary degrees of freedom are sometimes partitioned among the parameters in the model (Hodges and Sargent, 2001; Cui et al., 2010). By (14), influence matrices of the

general form (11) imply the expression $\mathscr{H}_{\ell\cdot} = \left[\boldsymbol{b}_s(s_\ell)^T \otimes \boldsymbol{X}\right] \boldsymbol{C}\,(\mathbf{1}_L \otimes \boldsymbol{I}_n)$ for the point-wise influence matrix, which suggests dividing the degrees of freedom into contributions of the individual columns of $\boldsymbol{X}$, i.e.,

$\mathrm{DF}_\ell(j) = \mathrm{tr}\left[\{\boldsymbol{b}_s(s_\ell)^T \otimes \boldsymbol{X} e_{j;p} e_{j;p}^T\}\boldsymbol{C}\,(\mathbf{1}_L \otimes \boldsymbol{I}_n)\right] = \left[\boldsymbol{b}_s(s_\ell)^T \otimes e_{j;p}^T\right] \boldsymbol{C}\,(\mathbf{1}_L \otimes \boldsymbol{X}_{\cdot j})$ for $j = 1,\ldots, p$, where $\boldsymbol{X}_{\cdot j} = \boldsymbol{X}\boldsymbol{e}_{j;p}$ is the jth column of $\boldsymbol{X}$. This directly extends the degrees of freedom per parameter in section 4.4 of Wood (2006a). By plugging in the particular values of $\boldsymbol{C}$ that are given implicitly in Web Appendix A for the functional linear model estimators of Theorem 1, one can readily show that $\mathrm{DF}_\ell(j) \equiv 1$ for each j in the linear case. For nonlinear models, $\mathrm{DF}_\ell(j)$ can be interpreted as a pointwise parameter-specific shrinkage factor.

More straightforwardly, $\mathrm{DF}_\ell = \mathrm{tr}\,(\mathscr{H}_{\ell\cdot})$ can also be partitioned into contributions of the individual observations, given by the n elements of the main diagonal of $\mathscr{H}_{\ell\cdot}$. These elements provide a novel definition of pointwise leverage that could be used to examine which of the functional observations are influential in a pointwise sense.

While we have focused on independent functional observations and the simple settings of Problems 1 and 2, the pointwise influence matrix is readily extended to more complex cases such as longitudinal functional responses and pointwise additive models, as well as to scalar responses observed on a grid or array. Finally, the Bayesian notion of priors on degrees of freedom (Hodges and Sargent, 2001; Hodges, 2013) might be extended to the pointwise setting. We hope to pursue these developments in future work.

An earlier unpublished report by the authors (Reiss et al., 2014) includes additional estimators, comparative simulations, and details not covered here. An R package implementing the methods described here is available at https://github.com/philreiss/vsm.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the reviewers as well as Yin-Hsiu Chen, Ciprian Crainiceanu, Yair Goldberg, Jeff Goldsmith, Lan Huo, Mike Milham, Todd Ogden, Eva Petkova, David Ruppert, Fabian Scheipl, and Simon Wood for very helpful advice and feedback. The first author's work was supported by U.S. National Science Foundation grant DMS-0907017, and the work of the first four authors was supported by U.S. National Institutes of Health grant 1R01MH095836.

References

Buja A, Hastie T, Tibshirani R. Linear smoothers and additive models. Ann Statist. 1989; 17:453–510.

Burman P. A comparative study of ordinary cross-validation, *v*-fold cross-validation and the repeated learning-testing methods. Biometrika. 1989; 76:503–514.

Chiou JM, Müller HG, Wang JL. Functional quasi-likelihood regression models with smooth random effects. J R Statist Soc B. 2003; 65:405–423.

Crainiceanu CM, Ruppert D. Likelihood ratio tests in linear mixed models with one variance component. J R Statist Soc B. 2004; 66:165–185.

Crainiceanu CM, Staicu AM, Ray S, Punjabi N. Bootstrap-based inference on the difference in the means of two correlated functional processes. Statistics in Medicine. 2012; 31:3223–3240. [PubMed: 22855258]

Cui Y, Hodges JS, Kong X, Carlin BP. Partitioning degrees of freedom in hierarchical and other richly parameterized models. Technometrics. 2010; 52:124–136. [PubMed: 20559456]

Efron B. The estimation of prediction error (with Discussion). J Am Statist Assoc. 2004; 99:619–642.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression (with Discussion). Ann Statist. 2004; 32:407–499.

Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. Stat Sci. 1996; 11:89–102.

Eilers PHC, Marx BD. Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. Chemometr Intell Lab. 2003; 66:159–174.

Fan J, Zhang JT. Two-step estimation of functional linear models with applications to longitudinal data. J R Statist Soc B. 2000; 62:303–322.

Green, PJ., Silverman, BW. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman & Hall; Boca Raton, Florida: 1994.

Hoaglin DC, Welsch RE. The hat matrix in regression and ANOVA. The American Statistician. 1978; 32:17–22.

Hodges, JS. Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects. CRC Press; Boca Raton, Florida: 2013.

Hodges JS, Sargent DJ. Counting degrees of freedom in hierarchical and other richly-parameterised models. Biometrika. 2001; 88:367–379.

Hyndman RJ, Shang HL. Rainbow plots, bagplots, and boxplots for functional data. J Comp Graph Statist. 2010; 19:29–45.

Imperati D, Colcombe S, Kelly C, Di Martino A, Zhou J, Castellanos FX, Milham MP. Differential development of human brain white matter tracts. PLoS ONE. 2011; 6:e23437. [PubMed: 21909351]

Janson L, Fithian W, Hastie TJ. Effective degrees of freedom: a flawed metaphor. Biometrika. 2015; 102:479–485. [PubMed: 26977114]

Kaufman S, Rosset S. When does more regularization imply fewer degrees of freedom? Sufficient conditions and counterexamples Biometrika. 2014; 101:771–784.

Krivobokova T, Kauermann G. A note on penalized spline smoothing with correlated errors. J Am Statist Assoc. 2007; 102:1328–1337.

Müller HG, Yao F. Functional additive models. J Am Statist Assoc. 2008; 103:1534–1544.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2016.

Ramsay, JO., Silverman, BW. Functional Data Analysis. 2nd. Springer; New York: 2005.

Reiss, P., Chen, YH., Huang, L., Huo, L., Tan, R., Jiao, R. vows: Voxelwise Semiparametrics. 2016. R package version 0.5

Reiss, PT., Huang, L., Chen, H., Colcombe, S. Varying-smoother models for functional responses. arXiv:1412.0778 [stat.ME]. 2014. available at http://arxiv.org/abs/1412.0778

Reiss PT, Huang L, Chen YH, Huo L, Tarpey T, Mennes M. Massively parallel nonparametric regression, with an application to developmental brain mapping. J Comp Graph Statist. 2014; 23:232–248.

Reiss PT, Huang L, Mennes M. Fast function-on-scalar regression with penalized basis expansions. Int J Biostat. 2010; 6 article 28.

Reiss PT, Ogden RT. Smoothing parameter selection for a class of semiparametric linear models. J R Statist Soc B. 2009; 71:505–523.

Ruppert, D., Wand, MP., Carroll, RJ. Semiparametric Regression. Cambridge University Press; New York: 2003.

Schumaker, L. Spline Functions: Basic Theory. 3rd. Cambridge University Press; Cambridge, UK: 2007.

Shaw P, Eckstrand K, Sharp W, Blumenthal J, Lerch JP, Greenstein D, Clasen L, Evans A, Giedd J, Rapoport JL. Attention-deficit/hyperactivity disorder is characterized by a delay in cortical maturation. Proc Natl Acad Sci USA. 2007; 104:19649–19654. [PubMed: 18024590]

Tibshirani RJ, Taylor J. Degrees of freedom in lasso problems. Ann Statist. 2012; 40:1198–1232.

Wahba G. Bayesian "confidence intervals" for the cross-validated smoothing spline. J R Statist Soc B. 1983; 45:133–150.

Wood, SN. Generalized Additive Models: An Introduction with R. Chapman & Hall; Boca Raton, Florida: 2006a.

Wood SN. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics. 2006b; 62:1025–1036. [PubMed: 17156276]

Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Statist Soc B. 2011; 73:3–36.

Xiao L, Li Y, Ruppert D. Fast bivariate P-splines: the sandwich smoother. J R Statist Soc B. 2013; 75:577–599.

Ye J. On measuring and correcting the effects of data mining and model selection. J Am Statist Assoc. 1998; 93:120–131.

Zhu H, Fan J, Kong L. Spatially varying coefficient model for neuroimaging data with jump discontinuities. J Am Statist Assoc. 2014; 109:1084–1098.

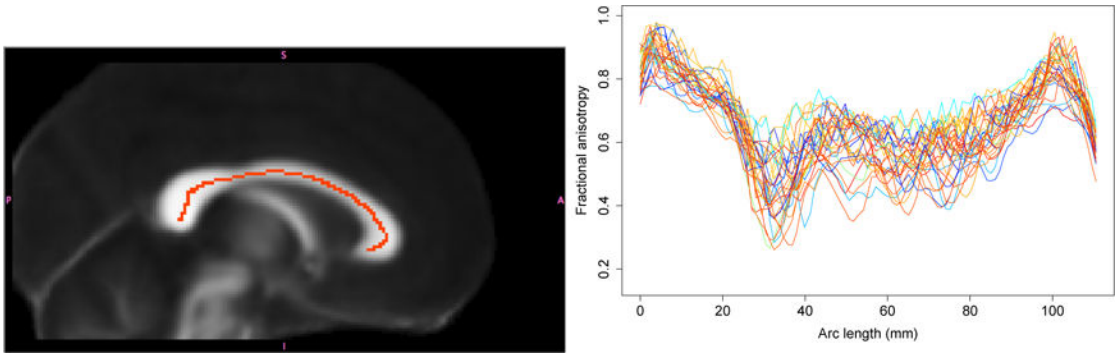

Figure 1.

Left: Sequence of 107 corpus callosum voxels at which fractional anisotropy was recorded. Right: Rainbow plots displaying how the resulting fractional anisotropy profiles (functional responses) vary with age, which ranges from 7 (dark blue) to 48 (red).

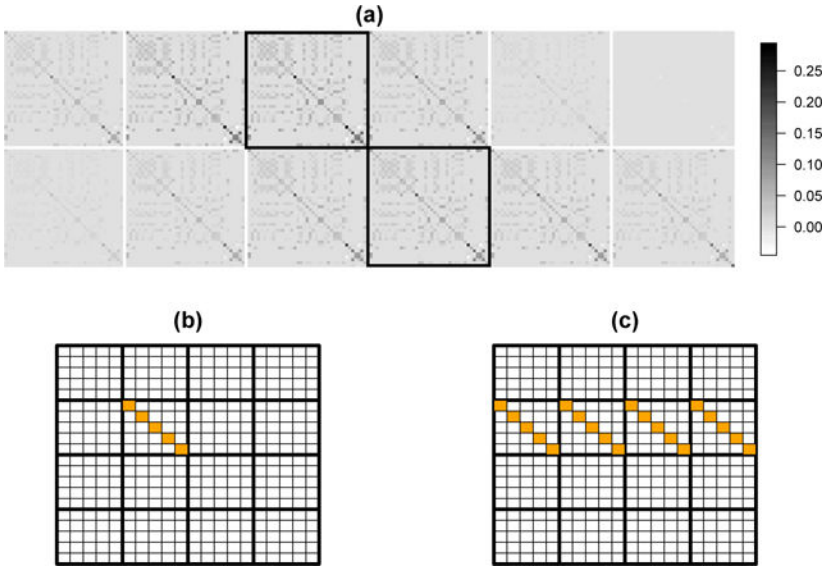

Figure 2.

(a) Excerpt from the block influence matrix (10) for a fit to the corpus callosum data, with diagonal blocks outlined in black. (b) Schematic illustration of the usual definition of degrees of freedom for the 2nd location, in a toy example with separate models at each of $L = 4$ locations. (c) Proposed pointwise degrees of freedom DF_2 for the 2nd location. This figure appears in color in the electronic version of this article.

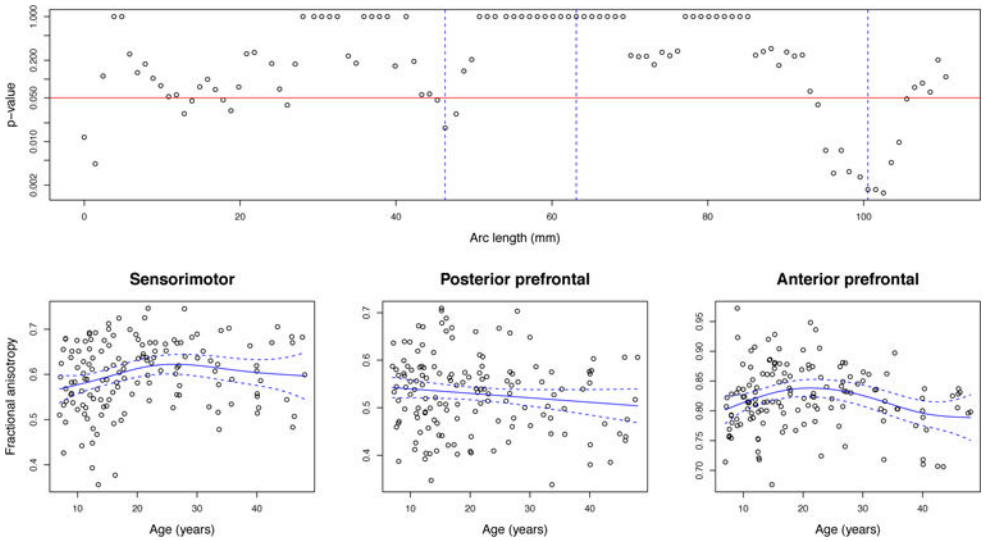

Figure 3.

Above, *p*-values from restricted likelihood ratio tests of the null hypothesis that mean fractional anisotropy changes linearly with age. Below, curve estimates, ±2 approximate standard errors, for the three voxels indicated above by dashed vertical lines; headings refer to the brain regions where these voxels are located. This figure appears in color in the electronic version of this article.

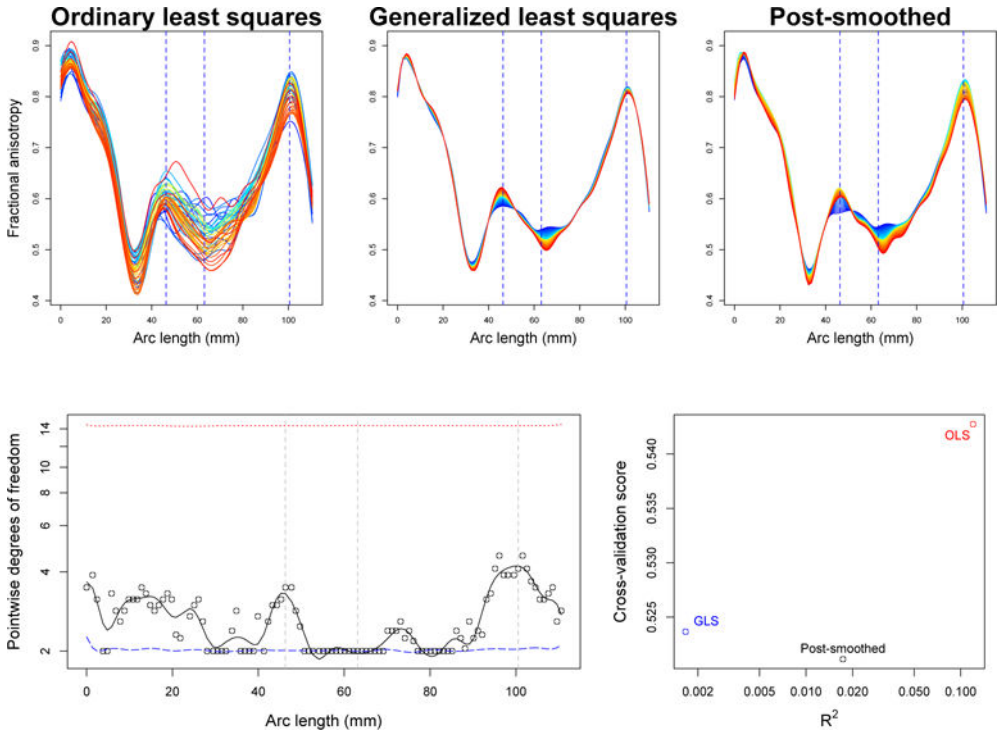

Figure 4.

Above: Rainbow plots of fitted value functions $\hat{f}(t, \cdot)$ with $t = 7, 8, \ldots, 48$, with 7 in dark blue and 48 in red. Lower left: Pointwise degrees of freedom for the corpus callosum data—separate smooths at each of the 107 voxels (circles), ordinary least squares (dotted curve), generalized least squares (dashed curve), and post-smoothed (solid curve). The vertical axis is on a log scale for clarity. Lower right: Functional R^2 plotted against prediction error estimate from repeated five-fold cross-validation.

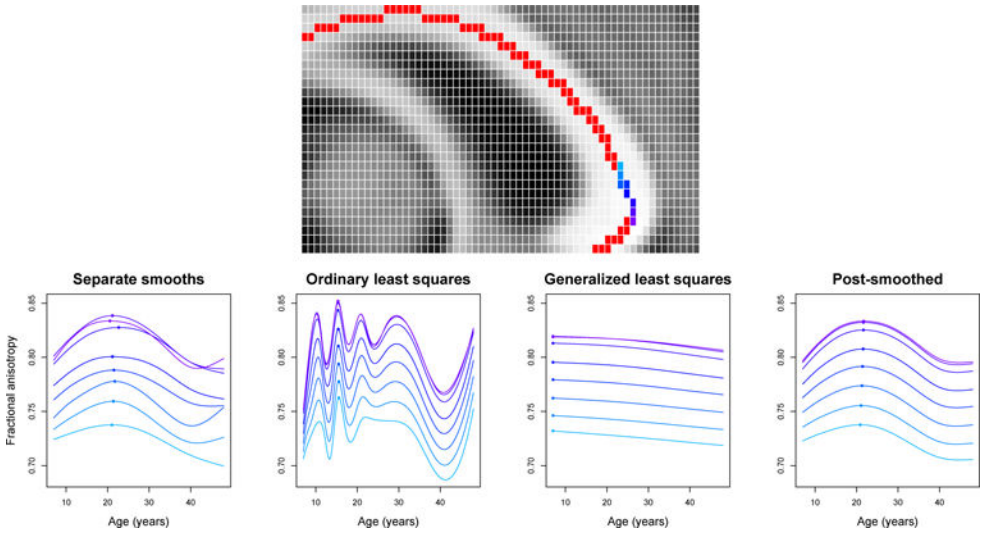

Figure 5.

Estimated mean fractional anisotropy as a function of age for eight voxels in the prefrontal cortex, which are displayed in the image above. Dots indicate estimated age of peak fractional anisotropy. This figure appears in color in the electronic version of this article.

Table 1

Representative references for some existing approaches to fitting function-on-scalar regression models of the general form (1).

Problem	*Dependence of y on predictor(s)*	*Contemporaneous smoothing*	*Post-smoothing*
1	Linear	Ramsay and Silverman (2005), Reiss et al. (2010)	Fan and Zhang (2000), Zhu et al. (2014)
2	Nonlinear	Wood (2006b)	Xiao et al. (2013)