2016

# Cross-sectional versus longitudinal designs for function estimation, with an application to cerebral cortex development

Philip T. Reiss

Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

# Cross-sectional versus longitudinal designs for function estimation, with an application to cerebral cortex development

**Philip T. Reiss***

**Motivated by studies of the development of the human cerebral cortex, we consider the estimation of a mean growth trajectory and the relative merits of cross-sectional and longitudinal data for that task. We define a class of relative efficiencies that compare function estimates in terms of aggregate variance of a parametric function estimate. These generalize the classical design effect for estimating a scalar with cross-sectional versus longitudinal data, and in particular cases are shown to be bounded above by it. Turning to nonparametric function estimation, we find that a longitudinal fits may tend to have higher aggregate variance than cross-sectional ones, but that this may occur because the former have higher effective degrees of freedom reflecting greater sensitivity to subtle features of the estimand. These ideas are illustrated with cortical thickness data from a longitudinal neuroimaging study.**


**Keywords:** Accelerated longitudinal design, Cortical thickness, Design effect, Effective degrees of freedom, Penalized splines

## 1. Introduction

Of the four objectives in the United States National Institute for Mental Health's current strategic plan for research (1), the second is to "chart mental illness trajectories to determine when, where, and how to intervene." One of the two proposed strategies to attain this objective is to "characterize the developmental trajectories of brain maturation and dimensions of behavior to understand the roots of mental illnesses across diverse populations." When studying developmental trajectories in psychiatry, or in other areas of biomedical research, a key question is the relative difficulty of estimating the process of interest on the basis of cross-sectional versus longitudinal data. Our aim here is to examine this fundamental question, from both theoretical and empirical perspectives.

The importance of longitudinal data for studying developmental processes in psychiatry is often noted (2; 3). In the study of the human cerebral cortex and its development in childhood and adolescence, the application area that motivated the present work, important discoveries have been gleaned from longitudinal studies (4; 5). Intuitively, if one wishes to

*University of Haifa, Israel, and New York University School of Medicine, USA*
** Correspondence to: Department of Statistics, University of Haifa, Haifa 31905, Israel*



*Prepared using* simauth.cls *[Version: 2010/03/10 v3.00]*

understand the trajectory of structural changes in the brain, then indeed one needs to follow the individuals under study over time.

On the other hand, studies of "neurodevelopmental trajectories" are often concerned not with individual change but rather with a population mean of a quantity of interest, such as cortical thickness, as a function of age (6). For estimating a mean, it is well known that the use of clustered observations, such as longitudinal observations with uniform correlation, results in inflated variance. Specifically, when estimating a scalar mean or effect based on clusters of $m$ observations having intraclass correlation $\rho$, the variance is inflated by

$$1 + (m-1)\rho \tag{1}$$

relative to a cross-sectional data set with the same total number of observations. The quantity (1) has been called the "design effect" in the survey sampling literature (7), or the "inflation factor" in the literature on cluster-randomized clinical trials (8), and it is often used for sample size planning. Intuitively, observing the same individuals repeatedly implies redundant data, and formula (1) represents the cost of this redundancy. When $\rho = 0$ the design effect equals 1 since there is no redundancy; in the limit as $\rho \to 1$ (complete redundancy) it tends to $m$; and for intermediate values it increases linearly with $\rho$.

In order to understand more fully the statistical advantages, and disadvantages, of longitudinal designs for studying developmental trajectories, in this paper we shall investigate whether an inflation factor such as (1) applies when the estimand is not a scalar but a function.

After introducing our formal framework in section 2, in section 3 we derive a general expression for the relative efficiency of cross-sectional versus longitudinal data for function estimation. This expression allows for varying numbers of observations per individual, as occurs in most realistic settings. In section 4 we consider the special case of a fixed number of observations per individual, and find that the relative efficiency generalizes the classical formula (1) in an interesting manner. Whereas our development up to section 4 concerns parametric models such as linear or polynomial regression, section 5 considers penalized regression such as spline smoothing. Our notion of relative efficiency is extended to penalized smoothing, but the fact that cross-sectional and longitudinal smooths can have different effective degrees of freedom makes the situation more complex. Some of the ideas are illustrated with a cortical thickness data set in section 6 and a simulation study in section 7. Section 8 offers some concluding remarks.

## 2. Setup

In the longitudinal setting we have $n$ individuals, the $i$th of whom is observed at ages $t_{i1} < \ldots < t_{im_i} \in \mathcal{T}$ for some interval $\mathcal{T}$, with responses $y_{i1}, \ldots, y_{im_i}$. Typically $\mathcal{T}$ represents a wide range of ages, and we sample individuals whose ages vary across this range and follow them over a much shorter time span; this is known as an accelerated longitudinal design (9; 10). We consider the simple random-intercept model

$$y_{ij} = \boldsymbol{x}^T(t_{ij})\boldsymbol{\beta} + u_i + \varepsilon_{ij} \tag{2}$$

for $i = 1, \ldots, n, j = 1, \ldots, m_i$. Here $\boldsymbol{x}(\cdot) = [x_1(\cdot), \ldots, x_K(\cdot)]^T$ denotes a set of basis functions, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$ is the corresponding coefficient vector; the random effects $u_i$ $(i = 1, \ldots, n)$ are independently and identically distributed (iid) as $N(0, \sigma_u^2)$; and the errors $\varepsilon_{ij}$ are iid $N(0, \sigma_\varepsilon^2)$. The mean function is then taken to be $f(t) = \boldsymbol{x}(t)^T\boldsymbol{\beta}$. In neurodevelopmental applications, polynomial models such as the quadratic model $\boldsymbol{x}(t) = (1, t, t^2)^T$ are popular. An alternative would be a spline basis $\boldsymbol{x}(t) = [b_1(t), \ldots, b_K(t)]^T$, but splines are usually combined with penalization, an approach that we consider below in section 5.

Let $N = \sum_{i=1}^{n} m_i$. To express (2) in matrix form, define the $N$-dimensional vectors $\boldsymbol{y} = (y_{11}, \ldots, y_{1m_1}, \ldots, y_{n1}, \ldots, y_{nm_n})^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \ldots, \varepsilon_{1m_1}, \ldots, \varepsilon_{n1}, \ldots, \varepsilon_{nm_n})^T$, and the $N \times K$ matrix $\boldsymbol{X}$ with rows $\boldsymbol{x}^T(t_{ij})$ ordered in the same way. Denote the $k$th column of $\boldsymbol{X}$ by

$$\boldsymbol{x}_{\cdot k} = [x_k(t_{11}), \ldots, x_k(t_{1m_1}), \ldots, x_k(t_{n1}), \ldots, x_k(t_{nm_n})]^T.$$

Let $\boldsymbol{Z}$ be the $N \times n$ matrix whose $k$th column consists of 1's for observations from the $k$th individual and 0's elsewhere, and let $\boldsymbol{u} = (u_1, \ldots, u_n)^T$. We can then write (2) in the standard linear mixed effects model form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}$. Our model implies $\mathrm{Var}(\boldsymbol{y}|\boldsymbol{X}) = \sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}^T + \sigma_\varepsilon^2 \boldsymbol{I}_N$. Let $\sigma_T^2 = \sigma_u^2 + \sigma_\varepsilon^2$ and define $\rho = \sigma_u^2/\sigma_T^2$; equivalently $\rho$ is the correlation (conditional on $\boldsymbol{x}$) among repeated observations. We can then write

$$\mathrm{Var}(\boldsymbol{y}|\boldsymbol{X}) = \sigma_T^2 \boldsymbol{\Sigma}(\rho) \tag{3}$$

where

$$\boldsymbol{\Sigma}(\rho) = \rho \boldsymbol{Z}\boldsymbol{Z}^T + (1-\rho)\boldsymbol{I}_N. \tag{4}$$

The best linear unbiased estimate (BLUE) of the coefficient vector, $\hat{\boldsymbol{\beta}} = [\boldsymbol{X}^T \boldsymbol{\Sigma}(\rho)^{-1} \boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma}(\rho)^{-1} \boldsymbol{y}$, has variance-covariance matrix

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma_T^2 [\boldsymbol{X}^T \boldsymbol{\Sigma}(\rho)^{-1} \boldsymbol{X}]^{-1}. \tag{5}$$

In practice, the variance components are estimated, usually by restricted maximum likelihood, giving an *empirical* BLUE; hence $\rho$ is also estimated, and thus the notation $\hat{\rho}$ would be more appropriate in (5). Nevertheless, it is conventional to ignore estimation error in relative efficiency measures such as (1), so in line with this convention we treat the asymptotic variance-covariance matrix (5) as exact (see (11) regarding small-sample correction).

Taking $\rho = 0$ in (3) and (5) gives the formulas $\mathrm{Var}(\boldsymbol{y}|\boldsymbol{X}) = \sigma_T^2 \boldsymbol{I}_N$, $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma_T^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$ that would obtain if, instead of longitudinal data, we were to estimate $\boldsymbol{\beta}$ by ordinary least squares using $N$ independent cross-sectional observations with error variance $\sigma_T^2$ and the same responses and design matrix as above. In this mathematical sense, cross-sectional data can be viewed as a special case of the above longitudinal setup, with $\rho = 0$. This observation plays a key role in what follows.

## 3. Relative efficiency for parametric function estimation

### *3.1. A class of aggregate variances*

When interest centers on estimating not a scalar quantity but an entire trajectory, it is appropriate to consider some sort of aggregate of variance along the function $f(\cdot) = \boldsymbol{x}(\cdot)^T \boldsymbol{\beta}$. In view of (5), a number of such aggregate indices can be expressed as

$$\mathrm{AV}_G(\rho) = \sigma_T^2 \mathrm{tr}[\boldsymbol{G}\{\boldsymbol{X}^T \boldsymbol{\Sigma}(\rho)^{-1} \boldsymbol{X}\}^{-1}] \tag{6}$$

where $\boldsymbol{G}$ is the $K \times K$ matrix whose $(k, \ell)$ entry equals a possibly degenerate inner product $\langle \cdot, \cdot \rangle$ evaluated for the $k$th and $\ell$th basis functions—i.e., $\boldsymbol{G}$ is the Gram matrix $[\langle x_k, x_\ell \rangle]_{1 \le k, \ell \le K}$. Three examples follow.

1. The example that will be our main focus is

$$\langle x_k, x_\ell \rangle = \boldsymbol{x}_{\cdot k}^T \boldsymbol{x}_{\cdot \ell} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} x_k(t_{ij}) x_\ell(t_{ij}). \tag{7}$$

Then (6) equals the variance of $\hat{f}(t_{ij})$ summed over the design points $t_{ij}$. To see this, first note that $\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{X}$; by (5), (6) then becomes

$$\begin{aligned} \text{AV}_{X^TX}(\rho) &= \sigma_T^2 \text{tr}[\boldsymbol{X}^T\boldsymbol{X}\{\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X}\}^{-1}] \\ &= \sigma_T^2 \text{tr}\left[\sum_{i=1}^n \sum_{j=1}^{m_i} \boldsymbol{x}(t_{ij})\boldsymbol{x}(t_{ij})^T\{\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X}\}^{-1}\right] \\ &= \sigma_T^2 \sum_{i=1}^n \sum_{j=1}^{m_i} \boldsymbol{x}(t_{ij})^T\{\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X}\}^{-1}\boldsymbol{x}(t_{ij}) \\ &= \sum_{i=1}^n \sum_{j=1}^{m_i} \boldsymbol{x}(t_{ij})^T \text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{x}(t_{ij}) \\ &= \sum_{i=1}^n \sum_{j=1}^{m_i} \text{Var}[\boldsymbol{x}(t_{ij})^T\hat{\boldsymbol{\beta}}] \\ &= \sum_{i=1}^n \sum_{j=1}^{m_i} \text{Var}[\hat{f}(t_{ij})]. \end{aligned} \tag{8}$$

2. Inner product (7) of the previous example equals $N$ times the $L^2$ inner product with respect to the empirical distribution of the $t_{ij}$'s. If instead $\langle \cdot, \cdot \rangle$ denotes $L^2$ inner product with respect to Lebesgue measure on $\mathcal{T}$, then by a similar argument, $\text{AV}_G(\rho) = \int_{\mathcal{T}} \text{Var}\hat{f}(t)dt$, the integrated variance.

3. If $\langle f, g \rangle$ is the $L^2$ inner product with respect to Lebesgue measure of the derivatives $f', g'$, then $\text{AV}_G(\rho) = \int_{\mathcal{T}} \text{Var}\hat{f}'(t)dt$ where $\hat{f}'(t) = \frac{d}{dt}\hat{f}(t)$. The inner product in this case is degenerate in the sense that $\langle f, f \rangle = 0$ does not imply $f = 0$.

Clearly $\text{AV}_G(\rho)$ depends not only on $\boldsymbol{G}$ but also on $\boldsymbol{X}$ and (via $\boldsymbol{\Sigma}(\rho)$) on $\boldsymbol{Z}$. But to keep the notation simple, in what follows we take $\boldsymbol{X}$ and $\boldsymbol{Z}$ as given, and the notation reflects only the dependence on $\boldsymbol{G}$ and $\rho$.

### *3.2. Relative efficiency defined, and the $\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{X}$ case*

From the last paragraph of section 2, the aggregate variance for the cross-sectional scenario is given by taking $\rho = 0$ in (6), i.e., by $\text{AV}_G(0)$. We can thus define the relative efficiency of a cross-sectional versus a longitudinal design, as a function of $\rho$,

$$r_G(\rho) = \text{AV}_G(\rho)/\text{AV}_G(0).$$

This can be thought of as a generalization of (1) from scalar to function estimation.

We assume henceforth that $\boldsymbol{G}$ is nonsingular. (Some of the development that follows, therefore, may not apply to the third example of $\boldsymbol{G}$ above.) We can then find a nonsingular $K \times K$ matrix $\boldsymbol{L}$ such that $\boldsymbol{G} = \boldsymbol{L}^T\boldsymbol{L}$, e.g., by Choleski decomposition. We then have a singular value decomposition (SVD)

$$\boldsymbol{X}\boldsymbol{L}^{-1} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{9}$$

where $\boldsymbol{D}$ is a $K \times K$ diagonal matrix with diagonal elements $d_1 \geq \ldots \geq d_K > 0$.

In most of what follows we take $\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{X}$, i.e., the first of the three example inner products listed in section 3.1. Let

$$\boldsymbol{\Xi}(\rho) = \text{Diag}\left\{\frac{\rho}{1+(m_1-1)\rho}, \ldots, \frac{\rho}{1+(m_n-1)\rho}\right\}, \tag{10}$$

and let $\gamma_1(\rho) \geq \ldots \geq \gamma_K(\rho)$ be the eigenvalues of

$$\boldsymbol{A}(\rho) = \boldsymbol{U}^T \boldsymbol{Z} \boldsymbol{\Xi}(\rho) \boldsymbol{Z}^T \boldsymbol{U}.$$

As shown in Appendix A,

$$\mathrm{AV}_{X^T X}(\rho) = \sigma_T^2 (1-\rho) \sum_{k=1}^{K} \frac{1}{1-\gamma_k(\rho)}. \tag{11}$$

Using either (8) or (11), we obtain

$$\mathrm{AV}_{X^T X}(0) = \sigma_T^2 K. \tag{12}$$

Thus the relative efficiency for $\boldsymbol{G} = \boldsymbol{X}^T \boldsymbol{X}$ is

$$r_{X^T X}(\rho) = \mathrm{AV}_{X^T X}(\rho) / \mathrm{AV}_{X^T X}(0) = \frac{1-\rho}{K} \sum_{k=1}^{K} \frac{1}{1-\gamma_k(\rho)}. \tag{13}$$

The following heuristic argument suggests that for most other Gram matrices $\boldsymbol{G}$, $r_G(\rho) \leq r_{X^T X}(\rho)$. We can assume without loss of generality that $\boldsymbol{G}$ is scaled so that $\det[\boldsymbol{G}(\boldsymbol{X}^T \boldsymbol{X})^{-1}] = 1$. Given positive definite $K \times K$ matrices $\boldsymbol{A}, \boldsymbol{B}$, the quantity

$$D_{ld}(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^{-1}) - \log\det(\boldsymbol{A}\boldsymbol{B}^{-1}) - K,$$

sometimes called the LogDet divergence, belongs to the class of non-negative dissimilarity measures known as Bregman divergences (12). Since $\mathrm{AV}_G(0) = \sigma_T^2[K + D_{ld}(\boldsymbol{G}, \boldsymbol{X}^T \boldsymbol{X})]$, we see that $\mathrm{AV}_G(0)$ attains its minimal value $\sigma_T^2 K$ when $\boldsymbol{G} = \boldsymbol{X}^T \boldsymbol{X}$. This does not imply that the ratio $r_G(\rho) = \mathrm{AV}_G(\rho)/\mathrm{AV}_G(0)$ is maximized when $\boldsymbol{G} = \boldsymbol{X}^T \boldsymbol{X}$, but it does suggest that this choice of $\boldsymbol{G}$ will tend to make $r_G(\rho)$ relatively high. We shall see an example of this below in section 4.3.

### 3.3. *Limit of $r_{X^T X}(\rho)$ as $\rho \to 1$*

As noted in the Introduction, the classical design effect (1) tends to $m$ as $\rho \to 1$. It is instructive to contrast this with the limit of $r_{X^T X}(\rho)$ as $\rho \to 1$, which we now derive. We assume that

$$\boldsymbol{X} \text{ is of rank } K, \tag{14}$$

as well as the following condition:

$$\mathrm{col}(\boldsymbol{X}) \cap \mathrm{col}(\boldsymbol{Z}) = \{c\mathbf{1}_N : c \in \mathbb{R}\}, \tag{15}$$

where $\mathrm{col}(\cdot)$ denotes column space. To see that this is a rather mild assumption, observe that $\mathbf{1}_N = \boldsymbol{Z}\mathbf{1}_n \in \mathrm{col}(\boldsymbol{Z})$, and $\mathbf{1}_N \in \mathrm{col}(\boldsymbol{X})$ when $\boldsymbol{X}$ is constructed from a polynomial basis including an intercept, or from a $B$-spline basis; hence in these cases $\{c\mathbf{1}_N : c \in \mathbb{R}\} \subseteq \mathrm{col}(\boldsymbol{X}) \cap \mathrm{col}(\boldsymbol{Z})$. For the former to be a *proper* subset of the latter, there must be a nonzero vector orthogonal to $\mathbf{1}_N$ that belongs to both column spaces, and this seems to be rare in practice.

We state first a lemma, whose proof appears in Appendix B, and then our main limit result. In what follows, let $\boldsymbol{v}_1(\rho), \ldots, \boldsymbol{v}_K(\rho)$ be orthonormal eigenvectors of $\boldsymbol{A}(\rho)$ corresponding to eigenvalues $\gamma_1(\rho) \geq \ldots \geq \gamma_K(\rho)$ respectively.

**Lemma 1** *Assume* (14) *and* (15)*. Then*

*(a)* $\gamma_1(1) = 1$*, and this eigenvalue is attained by the eigenvector*

$$\boldsymbol{v}_1(1) = N^{-1/2} \boldsymbol{U}^T \mathbf{1}_N; \tag{16}$$

*(b) for* $\rho \in [0, 1)$, $\gamma_2(\rho) \leq \gamma_2(1) < 1$*;*

*(c)* $\gamma_1(\cdot)$ *is continuously differentiable at* $\rho = 1$*, and* $\gamma_1'(1) = 1/\bar{m}$ *where* $\bar{m} = N/n$*, i.e., the mean of* $m_1, \ldots, m_n$*.*

**Theorem 1** *If* (14) *and* (15) *hold then* $\lim_{\rho\to 1} r_{X^TX}(\rho) = \bar{m}/K$*.*

**Proof of Theorem 1.** Rewriting (11) as $\text{AV}_{X^TX}(\rho) = \sigma_T^2 \sum_{k=1}^K \frac{1-\rho}{1-\gamma_k(\rho)}$, Lemma 1(b) implies that the $k \geq 2$ summands vanish as $\rho \to 1$. Combining this with l'Hôpital's rule and Lemma 1(c) yields

$$\lim_{\rho\to 1} \text{AV}_{X^TX}(\rho) = \sigma_T^2 \lim_{\rho\to 1} \frac{1-\rho}{1-\gamma_1(\rho)} = \frac{\sigma_T^2}{\lim_{\rho\to 1}\gamma_1'(\rho)} = \frac{\sigma_T^2}{\gamma_1'(1)} = \sigma_T^2\bar{m}.$$

Dividing by (12) gives the result. □

Theorem 1 shows that when $\bar{m} < K$, the use of longitudinal data deflates the aggregate variance, rather than inflating it, at least for sufficiently large $\rho$. Moreover, when all individuals have a fixed number $m$ of observations (henceforth, "the fixed-$m$ case"), the theorem says that as $\rho \to 1$, the limit of $r_{X^TX}(\rho)$ equals the limit of design effect (1) divided by $K$. The next section examines the relationship between $r_{X^TX}(\rho)$ and (1) for general $\rho$ in the fixed-$m$ case.

## 4. The fixed-$m$ case

When $m$ is fixed, $r_{X^TX}(\rho)$ has a simpler form and can be written as the classical design effect (1) times an adjustment factor that is less than or equal to 1. This suggests that the redundancy cost of longitudinal data is lower for function estimation than for estimating a scalar. In this section we derive the adjustment factor and show that it can depend quite strongly on the specific longitudinal design that is chosen.

### *4.1. An adjustment factor*

When $m_1 = \ldots = m_n = m$, (10) reduces to $\boldsymbol{\Xi}(\rho) = \frac{\rho}{1+(m-1)\rho}\boldsymbol{I}_n$ and so $\gamma_k(\rho) = \frac{\rho}{1+(m-1)\rho}\zeta_k$ where $\zeta_1 \geq \ldots \geq \zeta_K \geq 0$ are the eigenvalues of $\boldsymbol{U}^T\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{U}$. By (13),

$$\begin{aligned} r_{X^TX}(\rho) &= \frac{1}{K}\sum_{k=1}^K \frac{1-\rho}{1-\frac{\rho}{1+(m-1)\rho}\zeta_k} \\ &= [1+(m-1)\rho]\left[\frac{1}{K}\sum_{k=1}^K \frac{1-\rho}{1+(m-\zeta_k-1)\rho}\right], \end{aligned}$$

i.e., $r_{X^TX}(\rho)$ equals the classical design effect (1) times the adjustment factor

$$a(\rho) \equiv \frac{1}{K}\sum_{k=1}^K \frac{1-\rho}{1+(m-\zeta_k-1)\rho}. \tag{17}$$

The adjustment factor has derivative $a'(\rho) = \frac{1}{K}\sum_{k=1}^K \frac{\zeta_k - m}{[1+(m-\zeta_k-1)\rho]^2}$. Since, as shown below, $\zeta_1 \leq m$, we see that $a(\rho)$ is a decreasing function of $\rho$.

### *4.2. Bounds on the adjustment factor and* $r_{X^TX}(\rho)$

We can derive bounds on $a(\rho)$ under a weaker assumption than (15), namely

$$\mathbf{1} \in \text{col}(\boldsymbol{X}) \cap \text{col}(\boldsymbol{Z}). \tag{18}$$

In the present fixed-$m$ case $\boldsymbol{U}^T\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{U} = m\boldsymbol{U}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{U}$, so

$$\zeta_1 = m \max_{\|v\|=1} \boldsymbol{v}^T\boldsymbol{U}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{U}\boldsymbol{v} \leq m \max_{\|w\|=1} \boldsymbol{w}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{w} = m.$$

When (18) holds, equality obtains in the above with $\boldsymbol{v} = N^{-1/2}\boldsymbol{U}^T\boldsymbol{1}_N$ (cf. Lemma 1(a)), so $\zeta_1 = m$. Since $a(\rho)$ is an increasing function of $\zeta_k$ for $k = 1, \ldots, K$, it follows that its highest possible value is attained when

$$\zeta_1 = \ldots = \zeta_K = m; \tag{19}$$

we then obtain $a(\rho) = 1$ and so $r_{X^TX}(\rho) = 1 + (m-1)\rho$. Indeed we then have $r_G(\rho) = 1 + (m-1)\rho$ for general Gram matrices $\boldsymbol{G}$, as shown in Appendix C.

At the opposite extreme, when $\zeta_2 = \ldots = \zeta_k = 0$, the adjustment factor attains its lowest possible value, $a(\rho) = \frac{1}{1-\rho} + \frac{K-1}{1+(m-1)\rho}$, and some algebra yields the lower bound

$$r_{X^TX}(\rho) = 1 + (m/K - 1)\rho. \tag{20}$$

The limit of this last quantity as $\rho \to 1$ is $\frac{m}{K}$. Thus, under the assumptions of Theorem 1, $r_{X^TX}(\rho)$ approaches its lower bound as $\rho \to 1$.

The above upper bound is attained when for $i = 1, \ldots, n$, the $i$th individual is observed $m$ times at some fixed time point $t_i$, so that $\boldsymbol{X} = \boldsymbol{S} \otimes \boldsymbol{1}_m$ where $\boldsymbol{S} = [b_k(t_i)]_{1\leq i\leq n, 1\leq k\leq K}$ and $\boldsymbol{Z} = \boldsymbol{I}_n \otimes \boldsymbol{1}_m$. In this case, given an SVD $\boldsymbol{S}\boldsymbol{L}^{-1} = \boldsymbol{U}_*\boldsymbol{D}_*\boldsymbol{V}_*^T$, we can take $\boldsymbol{U} = \boldsymbol{U}_* \otimes (m^{-1/2}\boldsymbol{1}_m)$ in (9), so

$$\begin{aligned}
\boldsymbol{U}^T\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{U} &= [\boldsymbol{U}_*^T \otimes (m^{-1/2}\boldsymbol{1}_m^T)](\boldsymbol{I}_n \otimes \boldsymbol{1}_m)(\boldsymbol{I}_n \otimes \boldsymbol{1}_m^T)[\boldsymbol{U}_* \otimes (m^{-1/2}\boldsymbol{1}_m)] \\
&= m\boldsymbol{U}_*^T\boldsymbol{U}_* \\
&= m\boldsymbol{I}_K,
\end{aligned}$$

and thus (19) holds. Thus $r_{X^TX}(\rho) = 1 + (m-1)\rho$ for all $\rho \in [0, 1)$, resulting in a limit of $m$, rather than $\frac{m}{K}$, as $\rho \to 1$; but this does not contradict Theorem 1 since in this case (15) does not hold.

When $K = 1$, the lower bound (20) equals the upper bound $1 + (m-1)\rho$. This equality can in fact be used to deduce the classical formula (1). Consider a particular $K = 1$ case, namely $\boldsymbol{X} = \boldsymbol{1}_N = \boldsymbol{1}_{nm}$, so that $f$ is assumed to be a constant. Then our function estimation setting reduces to estimating a scalar mean; $\mathrm{AV}_{X^TX}(\rho)$ is just $N$ times the variance of the estimate; and thus $r_{X^TX}(\rho)$ reduces to the classical design effect, which must equal $1 + (m-1)\rho$ since it is bounded both below and above by that quantity.

### *4.3. Application to accelerated longitudinal designs*

As noted above in section 2, accelerated longitudinal designs allow us to estimate a growth trajectory over a wide age range by following individuals of various ages over a shorter time span. This time span, i.e. the length of the study, turns out to have a crucial impact on the adjustment factor (17). When the study period is very short, we have an approximate version of the upper bound scenario of the last paragraph of section 4.2. Thus $r_{X^TX}(\rho)$ remains near $1 + (m-1)\rho$ for all but the highest values of $\rho$, and then drops precipitously to the lower bound $\frac{m}{K}$. For a longer study, the adjustment factor is generally much less than 1 even for quite low $\rho$.

These ideas are illustrated with simulated data in Figure 1. For different values of $\rho \in [0, 1)$ and for $m = 2, 3, 4, 5$, we simulated $m$ observation ages $t_{i1}, \ldots, t_{im}$ from 10–30 years for each of $n = 100$ individuals (see Appendix D for details). We considered quadratic fits, i.e., $K = 3$ and basis function vector $\boldsymbol{x}(t) = (1, t, t^2)^T$, and computed $r_G(\rho)$ for the first two examples of $\boldsymbol{G}$ given in section 3.1: (1) variance summed over the design points ($\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{X}$) and (2) variance integrated over the age range. The left panel shows $r_G(\rho)$ values for a short study of length 0.5 years, i.e., the maximum value of

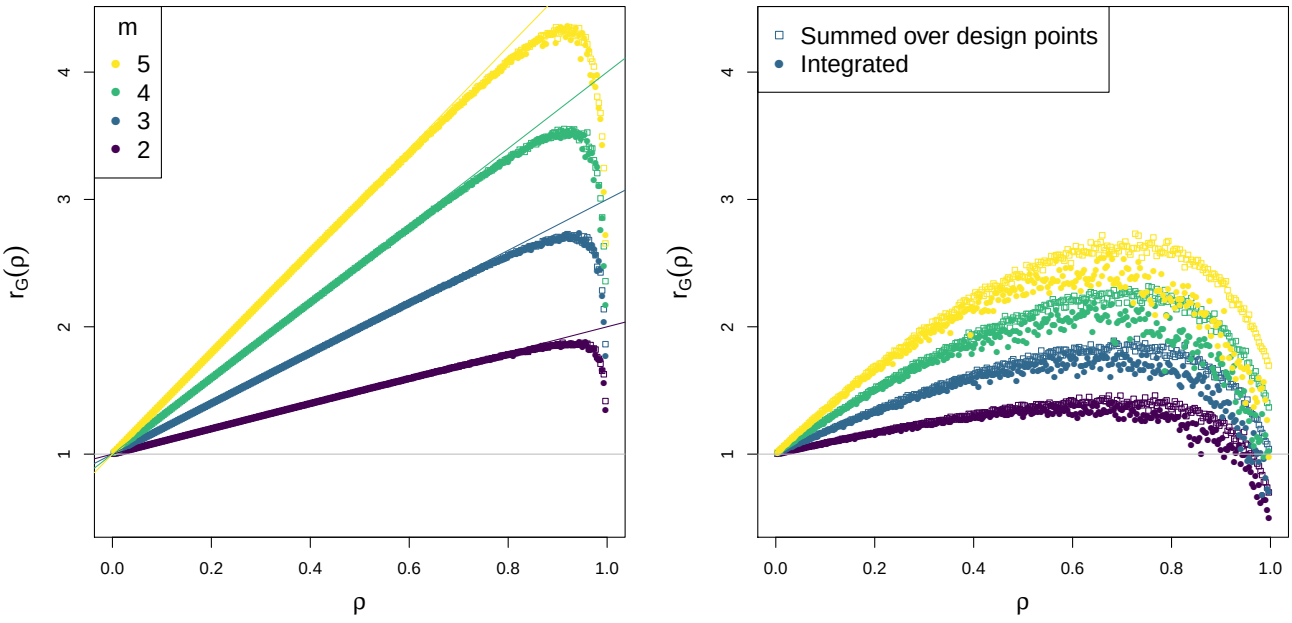


**Figure 1.** Relative efficiency $r_G(\rho)$ for accelerated longitudinal designs with $m = 2, 3, 4, 5$ observations per individual, where $\boldsymbol{G}$ corresponds to (1) variance summed over the design points ($\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{X}$) and (2) variance integrated over the age range. Left: a short (0.5-year) study; the linear function $1 + (m-1)\rho$ for each $m$ is shown for comparison. Right: a longer (3-year) study.

$t_{im} - t_{i1}$ is 0.5; the right panel assumes a longer study, of length 3 years. The relative efficiency tracks $1 + (m-1)\rho$ up to high values of $\rho$ for the short study depicted in the left panel, but is much lower for the longer study depicted at right. In accordance with Theorem 1, $r_{X^TX}(\rho)$ tends to $m/3$ as $\rho \to 1$, although this is less clearly seen in the left panel. Since the upper bound scenario discussed in section 4.2 applies to arbitrary $\boldsymbol{G}$, it is not surprising that the results in the left panel are virtually identical for summed versus integrated variance. On the other hand, as the end of section 3.2 would lead us to expect, for given $m$, $r_G(\rho)$ tends to be somewhat lower for integrated than for summed variance. This figure is reminiscent of the relative efficiency curves of (13) for designed experiments, although he considered generalized variance of polynomial parameters rather than aggregate variance.

To summarize, $r_{X^TX}(\rho)$ is bounded by two linear functions of $\rho$: above by the classical design effect $1 + (m-1)\rho$ and below by $1 + (\frac{m}{k} - 1)\rho$. For very short studies it is near the upper line for most values of $\rho$, dropping toward the lower bound only for $\rho$ near 1. For longer studies it is closer to the lower bound line throughout the range of $\rho$ values.

## 5. Penalized spline smoothing

For estimating trajectories of brain development, many authors take $K$ =2, 3 or 4 and fit a linear, quadratic or cubic model, respectively. But lifespan changes in the brain are often not well represented by polynomials, and moreover polynomial fits are highly dependent on the age range of the data (14; 15). In view of these limitations of polynomial models, it is increasingly common for neurodevelopmental trajectories to be estimated by penalized splines (16; 17; 6).

Let $\boldsymbol{x}(t) = [b_1(t), \ldots, b_K(t)]^T$ be a basis of $K$ spline functions, such as $B$-splines, and let $\boldsymbol{P}$ be a positive semidefinite $K \times K$ matrix such that $\boldsymbol{\beta}^T\boldsymbol{P}\boldsymbol{\beta}$ is an index of the "roughness" of $f(t) = \boldsymbol{x}(t)^T\boldsymbol{\beta}$. For example, if $\boldsymbol{P} = (p_{ij})_{1\le i,j\le K} = [\int x_i''(t)x_j''(t)dt]_{1\le i,j\le K}$ then the above quadratic form equals $\int f''(t)^2dt$. Here $K$ is typically quite high; to avoid an overfitted estimate $\hat{f}(\cdot) = \boldsymbol{x}(\cdot)^T\hat{\boldsymbol{\beta}}$, we add a roughness penalty $\lambda\boldsymbol{\beta}^T\boldsymbol{P}\boldsymbol{\beta}$, for some $\lambda > 0$, to the criterion used to estimate

$\boldsymbol{\beta}$: the sum of squared errors in the cross-sectional case, or the best linear unbiased prediction (BLUP) criterion (16) in the longitudinal case. We now define summed variance $\mathrm{AV}^{\lambda}_{X^TX}(\rho)$ for the penalized case, and examine the relative efficiency

$$r^{\lambda}_{X^TX}(\rho) = \mathrm{AV}^{\lambda}_{X^TX}(\rho)/\mathrm{AV}^{\lambda}_{X^TX}(0). \tag{21}$$

For simplicity, in this section we take $\lambda$ to be fixed, as opposed to being optimized in a data-driven manner.

To proceed we must decide how to define the (estimated) variance-covariance matrix of the fixed effects, a somewhat subtle matter due to the bias in penalized estimation of $\boldsymbol{\beta}$ (16; 18). Here we employ the posterior variance of (17), which in our notation is

$$\mathrm{Var}(\boldsymbol{\beta}|\boldsymbol{y}) = \sigma^2_T \left[\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X} + \frac{\lambda}{1-\rho}\boldsymbol{P}\right]^{-1}, \tag{22}$$

a direct generalization of (5); this motivates the definition

$$\mathrm{AV}^{\lambda}_{X^TX}(\rho) = \sigma^2_T \mathrm{tr}\left[\boldsymbol{X}^T\boldsymbol{X}\left\{\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X} + \frac{\lambda}{1-\rho}\boldsymbol{P}\right\}^{-1}\right]. \tag{23}$$

This is less than or equal to $\sigma^2_T\mathrm{tr}[\boldsymbol{X}^T\boldsymbol{X}\{\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X}\}^{-1}]$. Under the assumptions of Theorem 1, the proof of the theorem shows that the latter quantity converges to $\sigma^2_T\bar{m}$ as $\rho \to 1$. Taking $\rho = 0$ in (23) yields

$$\mathrm{AV}^{\lambda}_{X^TX}(0) = \sigma^2_T\mathrm{tr}[\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{P})^{-1}].$$

But the trace in the above expression equals that of the smoother or "hat" matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{P})^{-1}\boldsymbol{X}^T$ for the cross-sectional fit, and that trace is a standard way to define the effective degrees of freedom $\mathrm{DF}_\lambda$ of the fit (19; 20), so $\mathrm{AV}^{\lambda}_{X^TX}(0) = \sigma^2_T\mathrm{DF}_\lambda$. Substituting into (21), we conclude that

$$\lim_{\rho\to 1} r^{\lambda}_{X^TX}(\rho) \leq \bar{m}/\mathrm{DF}_\lambda.$$

This result resembles Theorem 1, but $\mathrm{DF}_\lambda$ replaces $K$ and the equality becomes an inequality.

## 6. Smoothing with cross-sectional versus longitudinal data in practice

### *6.1. Relative efficiencies for cortical thickness trajectories*

The design effect $r_G(\rho)$ assumes a hypothetical scenario in which we have cross-sectional and longitudinal data sets, each with the exact same design matrix $\boldsymbol{X}$. Such a pair of data sets would essentially never arise in practice, at least in a biomedical study for which the ages determining $\boldsymbol{X}$ depend to some extent on random sampling. We can, however, illustrate some of the above ideas approximately by considering cross-sectional and longitudinal subsamples of a larger data set.

Our data consist of cortical thickness measurements derived by magnetic resonance imaging in a sample of individuals age 3–31. As described elsewhere, the images were acquired as part of a longitudinal study of typical brain development at the U.S. National Institute of Mental Health (21), and cortical thickness was estimated at approximately 80 000 brain locations, known as vertices, using the Montreal Neurological Institute's CIVET pipeline (22). The participants were grouped in families, and some were scanned repeatedly. In total, 1181 images were acquired for 615 individuals belonging to 398 families. We obtained a cross-sectional subsample by randomly selecting one image from each of the 398 families, and a longitudinal subsample (characterized by within-person, but not within-family, dependence) by taking all observations for 197 of the individuals in the cross-sectional subsample, the total number of images being again 398.

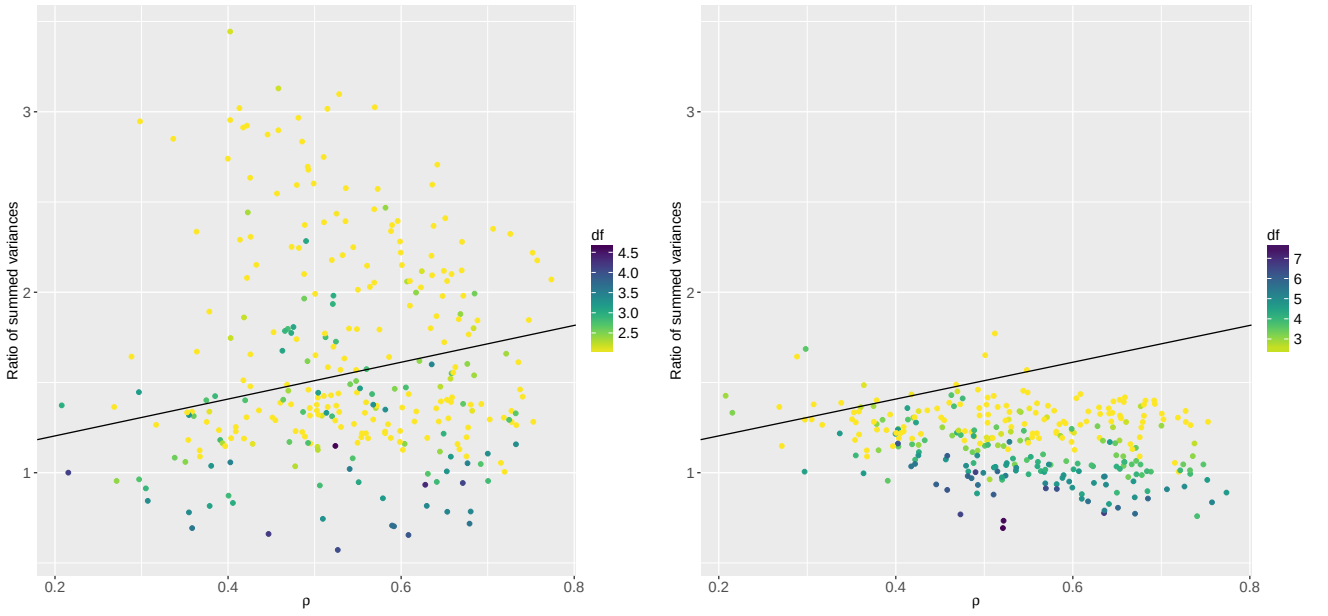


**Figure 2.** Ratios $\tilde{r}$ of posterior variances summed over the design points, for longitudinal versus cross-sectional nonparametric models fitted to the cortical thickness data. These ratios are approximate versions of $r_{X^TX}(\rho)$ and accordingly are plotted against estimated $\rho$. Each of the 300 points represents a vertex (brain location). Left: Ratios obtained with REML-based smoothness selection. Color legend refers to effective df of the cross-sectional model. The putative approximate upper bound $1 + (\bar{m} - 1)\rho$, shown as a black line, is exceeded by a wide margin for many of the points. Right: Ratios obtained with fixed degrees of freedom, as explained in the text.

For each of 300 randomly chosen vertices we performed penalized spline smoothing for both subsamples: nonparametric regression with a 15-dimensional $P$-spline basis (23) for the cross-sectional subsample, and a nonparametric mixed effects model with the same spline basis for the longitudinal subsample. The design matrices $\boldsymbol{X}_c$ and $\boldsymbol{X}_\ell$ for the cross-sectional and longitudinal samples, respectively, are not identical for a given vertex, but they share 197 observations, so we expect that $\boldsymbol{X}_c^T\boldsymbol{X}_c \approx \boldsymbol{X}_\ell^T\boldsymbol{X}_\ell$ and we let $\bar{\boldsymbol{G}} = \frac{1}{2}(\boldsymbol{X}_c^T\boldsymbol{X}_c + \boldsymbol{X}_\ell^T\boldsymbol{X}_\ell)$ take the place of $\boldsymbol{X}^T\boldsymbol{X}$ in the approximate analogue of $r_{X^TX}(\rho)$ that we define next.

More specifically, for each vertex we compute an approximate relative efficiency

$$\tilde{r} = \frac{\mathrm{tr}(\bar{\boldsymbol{G}}\boldsymbol{V}_p^\ell)}{\mathrm{tr}(\bar{\boldsymbol{G}}\boldsymbol{V}_p^c)}. \tag{24}$$

Here $\boldsymbol{V}_p^\ell$ is the posterior variance-covariance matrix for the longitudinal model, given by (22) where $\boldsymbol{X} = \boldsymbol{X}_\ell$ and model-based estimates are used for $\rho$ and $\lambda$; and $\boldsymbol{V}_p^c$ is the variance-covariance matrix for the cross-sectional model, which is likewise derived from (22) with $\rho = 0$. These variance-covariance matrices are included in the output from the packages `mgcv` (17; 24) and `gamm4` (25) for R (26), which were used to fit the cross-sectional and longitudinal models respectively. Ratio (24) is a real-data analogue of $r_{X^TX}(\rho)$, but differs from it in several respects:

1. The matrix $\boldsymbol{X}^T\boldsymbol{X}$ appears in both the numerator and denominator of the ratio (13) defining $r_{X^TX}(\rho)$. But in our case, as already noted, the design matrices differ for the two subsamples, so we use $\bar{\boldsymbol{G}}$, the average of $\boldsymbol{X}_c^T\boldsymbol{X}_c$ and $\boldsymbol{X}_\ell^T\boldsymbol{X}_\ell$.

2. In computing (24), the total variance $\sigma_T^2$ on which the posterior variance depends is estimated separately for the longitudinal and cross-sectional subsamples, and thus does not cancel as in $r_{X^TX}(\rho)$.

3. The smoothing parameter $\lambda$ is also estimated separately for the two subsamples, using restricted maximum likelihood (REML) (27; 24).

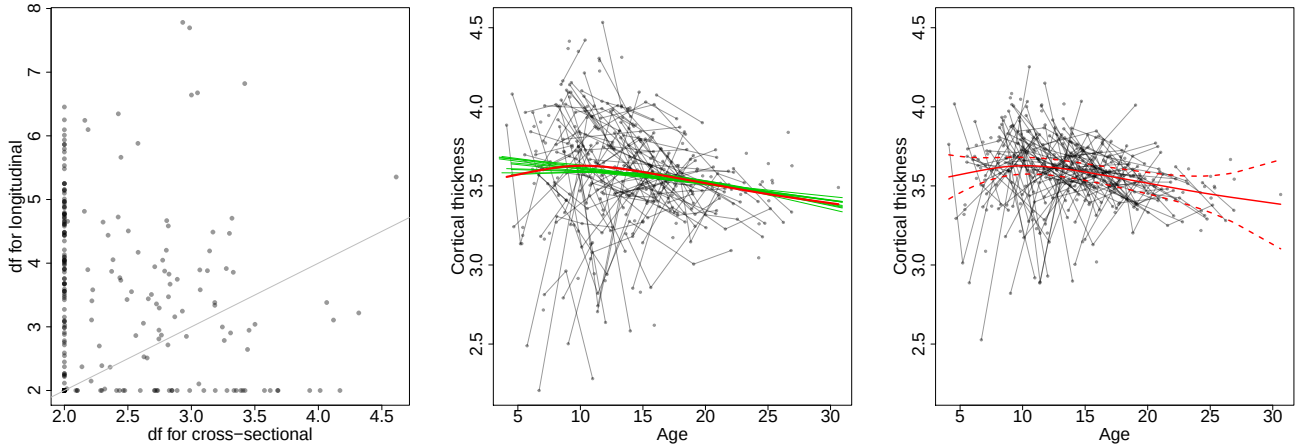


**Figure 3.** Left: Effective df of cross-sectional versus longitudinal fits for cortical thickness at each of 300 randomly chosen vertices; the line of identity is displayed for reference. Center: Spaghetti plot for one of the 300 vertices, along with curve fits based on the longitudinal data (red) and on 10 cross-sectional subsets thereof (green). Right: Same data with random effect estimates subtracted off (see the text), along with the curve fit based on the longitudinal data as well as $\pm 2$ standard error bands (dashed curves).

Despite these differences between $\tilde{r}$ and $r_{X^TX}(\rho)$, in view of our upper bound result (section 4.2) one might expect that the points $(\hat{\rho}, \tilde{r})$ for the 300 vertices, where $\hat{\rho}$ is determined by the variance component estimates for the given longitudinal fit, would generally lie below the linear function

$$\rho \mapsto 1 + (\bar{m} - 1)\rho \tag{25}$$

where $\bar{m} = 398/197$ is the mean number of observations per participant in the longitudinal subsample. But the left panel of Figure 2 shows that many of these points lie far above the line. The main reason for this is discrepancy 3 above. The differences between the two fits' REML-optimizing smoothing parameters are such that the effective df of the longitudinal fits tends to be higher than that of the cross-sectional fits, as seen in the left panel of Figure 3. When we instead fix the smoothing parameter for the cross-sectional model so as to yield the same df as for the longitudinal model, most of the points lie below, or at least not far above, line (25), as seen in the right panel of Figure 2.

### *6.2. The role of effective degrees of freedom*

The df for a nonparametric mixed-effects model is given by (17, p. 318) as

$$\operatorname{tr}\left[\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X}\left\{\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X} + \frac{\lambda}{1-\rho}\boldsymbol{P}\right\}^{-1}\right],$$

which is very similar to expression (23) for $\mathrm{AV}^{\lambda}_{X^TX}(\rho)$. Thus high df and high aggregate variance go hand-in-hand. In terms of our example, then, we would expect the *ratios* of df for the longitudinal versus the cross-sectional fits, for the 300 vertices, to resemble the corresponding values of $\tilde{r}$; and indeed the correlation between the two is 0.935.

The lesson here may be this: when automatic (e.g., REML-optimal) smoothness selection is used for both cross-sectional and longitudinal models, as in the left panel of Figure 2, the whole notion of a relative efficiency generalizing (1) is largely irrelevant. The longitudinal fit may indeed have higher variance, but only as a side effect of higher df—and the higher df may arise because the longitudinal fit detects features of the mean function that a cross-sectional fit would miss.

A possible example of this phenomenon is displayed in the middle panel of Figure 3, where cortical thickness values are plotted against age for a vertex for which the longitudinal fit (red curve) had markedly higher df than the cross-sectional. Specifically, the longitudinal fit (df=3.58) detects a peak around age 10, whereas the cross-sectional fit (df=2.22) indicates near-linear decline throughout the age range studied. In addition to the cross-sectional sample of size 398 as described

in section 6.1, we estimated the mean function using 10 cross-sectional subsamples of the longitudinal sample, each of which comprised one randomly sampled observation for each of the 197 individuals. The resulting estimates, shown in green in the middle panel of Figure 3, all appear quite similar: in each case the mean is found to decrease approximately linearly, rather than attaining a peak. Thus the monotonic decrease, as opposed to detecting a peak, appears to be a stable "non-feature" of cross-sectional fits to these data.

Writing the responses as $y_{ij} = \hat{f}(t_{ij}) + \hat{u}_i + \hat{\varepsilon}_{ij}$ where hats denote estimates from the longitudinal model, we see that the responses with the estimated (or "predicted") random effects removed are given by $\hat{f}(t_{ij}) + \hat{\varepsilon}_{ij}$. The plot of these adjusted responses in the right panel of Figure 3 offers stronger evidence of an initial increase to a peak than can be gleaned from the middle panel. This serves to illustrate the intuitive notion that longitudinal data makes function estimation more precise by separating within- from between-subject variation. Still, the wide pointwise credible intervals suggest that the conclusion of an initial increase should be adopted only very tentatively.

If the peak found only by the longitudinal fit is real, and if discrepancies of this sort between longitudinal and cross-sectional results are common, then the implications for the study of cortical development may be profound. The general finding that mean cortical thickness increases to a peak and then declines has been attributed to a process of synaptic pruning (4). Disparities in the estimated timing of the peak between children with attention deficit/hyperactivity disorder (ADHD) and controls gave rise to an influential theory that attributes ADHD to a neurodevelopmental delay (5). But a recent review of the evidence (28) pointed out that many studies find continuous thinning of the cortex from very early ages, rather than detecting a peak. (It has also been suggested that putative early peaks may be an artifact resulting from motion effects (29)). Thus the very existence of a peak in mean cortical thickness is a key open question in developmental neuropsychiatry. Related to this substantive question is the methodological question of whether an early peak is apt to be detected with longitudinal data but missed with cross-sectional data, as Figure 3 might lead one to suspect. To shed some light on the latter question, we turn to a small simulation study.

## 7. A comparative simulation study

To compare the efficacy of cross-sectional versus longitudinal data for detecting a subtle peak of the sort suggested by Figure 3, we simulated data resembling the cortical thickness data set (see Appendix E for details). The true function,

$$f(t) = 3.1 - 0.02t - 0.001 \exp[0.53(15 - t)], \tag{26}$$

is given by the black curve in each of the first two panels of Figure 4. This was chosen to have $f(t) \approx 3.1 - 0.02t$, i.e., essentially linear decline, for $t \geq 15$, but a somewhat subtle increase to a peak for younger ages.

Cross-sectional and longitudinal models were fitted to each of 294 data sets. A plot of the respective df for each of the 294 pairs of models (not shown) is similar to the left panel of Figure 3. For 91 of the model pairs, the longitudinal df exceeded the cross-sectional df by at least 0.5. The function estimates for a random sample of 30 of these 91 pairs are shown in the first two panels of Figure 4. Most of the cross-sectional fits are at least approximately linear and miss the peak. The longitudinal fits fare somewhat better, but still not terribly well, at detecting the peak. The pointwise mean squared estimation error comparison in the lower panel shows that the longitudinal fits significantly outperform the cross-sectional ones at the left end of the age range where the peak occurs. This clear superiority of the longitudinal estimates is observed only for this subset of the data set, i.e., those pairs of models for which the longitudinal fits have markedly higher df.

These results imply that a subtle feature such as an early peak may indeed be more readily detected with a longitudinal sample than with a cross-sectional sample of the same size; and that this occurs specifically when a longitudinally-based smooth has higher df than a cross-sectional one. As discussed in section 6.2, such a higher-df fit is apt to have higher aggregate variance, but that may be a price worth paying for greater sensitivity to features of the function of interest.

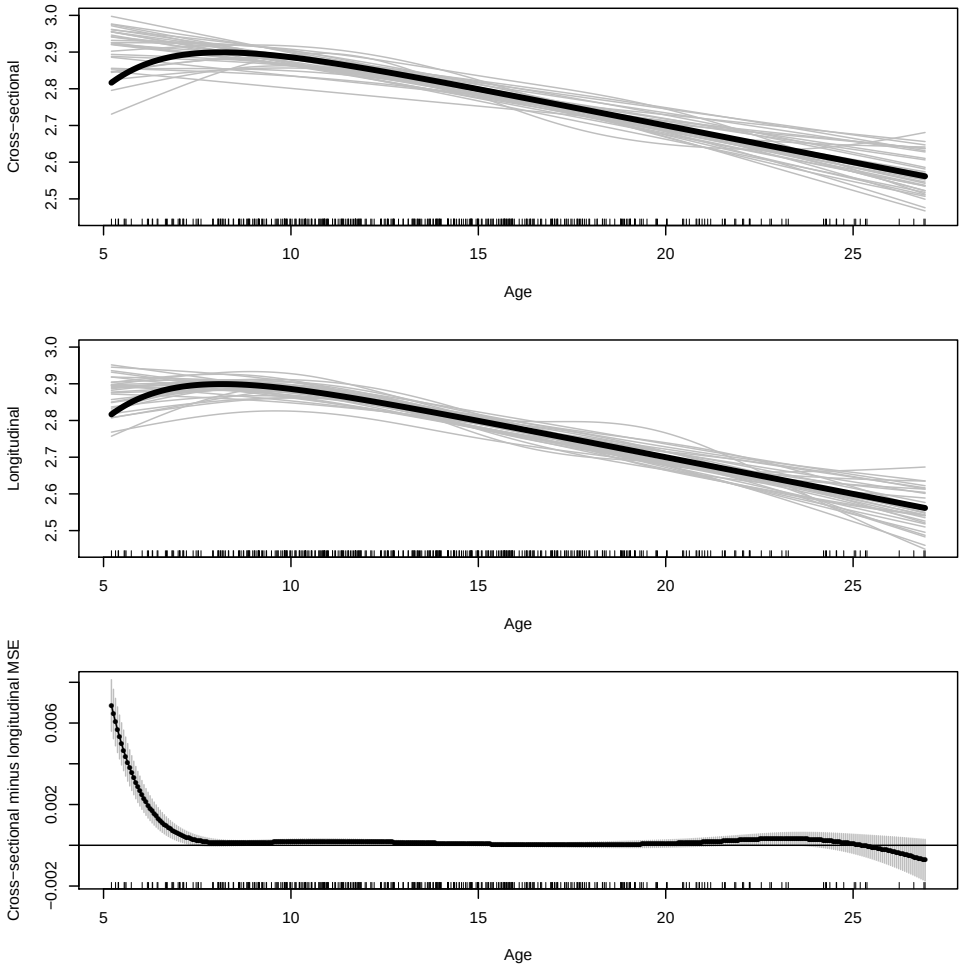


**Figure 4.** Test function (thick black curve) along with 30 estimates thereof based on cross-sectional data (top panel) and longitudinal data (middle panel), selected from among 91 simulation replicates in which the longitudinal fit markedly higher df. The bottom panel displays pointwise differences in mean squared estimation error for these 91 replicates, along with $\pm 2$ standard error bars for the difference.

## 8. Discussion

The "trajectory" $f(\cdot)$ that we have been concerned with estimating is a population mean, say of cortical thickness, as a function of age. It might be objected that such a mean can be representative of few if any individuals' trajectories of change (2; 30; 6). Nevertheless, the mean function may often be the most reasonable estimand available, especially when, as in most longitudinal brain imaging studies, there are too few observations per subject to estimate individual change with any precision.

Whereas our study of relative efficiency has taken the design matrix $\boldsymbol{X}$ (and $\boldsymbol{Z}$) as given and considered $\text{AV}_G(\rho)$ as a function of $\rho$, the field of optimal design is concerned with the choice of $\boldsymbol{X}$. Depending on the choice of $\boldsymbol{G}$ and of nomenclature, $\text{AV}_G(0)$ might be the criterion for an $I$-, $V$- or $IV$-optimal (31) or $Q$-optimal (32) cross-sectional design. In neuroimaging and other biomedical studies of growth trajectories, we generally cannot "optimally design" the participants' ages. But one lesson for study planning that emerges from our work is that the comparative efficiency of a longitudinal design may be improved by increasing the duration of follow-up (see Figure 1).

The relative efficiency $r_{X^TX}(\rho)$, a function-estimation analogue of the design effect (1), can be extended to nonparametric regression via penalized smoothing (section 5). But the very notion of relative efficiency presupposes a fixed model dimension $K$, whereas with penalized smoothing the effective model dimension depends on the smoothing

parameter and may differ for cross-sectional versus longitudinal models. We have seen that a longitudinally-based smooth may have higher effective dimension than a comparable fit based on cross-sectional data. When this occurs, the longitudinal function estimate may have higher aggregate variance, but this disadvantage may be offset by a more accurate fit. More work is needed to understand when and why a longitudinal design leads to improved accuracy in function estimation.

## Acknowledgements

Many thanks to Jay Giedd, Armin Raznahan and Aaron Alexander-Bloch for providing the cortical thickness data. The author gratefully acknowledges the support of the U.S. National Institutes of Health (grant R01 MH095836) and the Israel Science Foundation (grant 1777/16).

*Appendix A. Proof of* (11)

We can rewrite (6) as

$$\begin{aligned} \text{AV}_G(\rho) &= \sigma_T^2 \text{tr}[\boldsymbol{L}\{\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X}\}^{-1}\boldsymbol{L}^T] \\ &= \sigma_T^2 \text{tr}[\{\boldsymbol{L}^{-T}\boldsymbol{X}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{X}\boldsymbol{L}^{-1}\}^{-1}] \\ &= \sigma_T^2 \text{tr}[\boldsymbol{D}^{-2}\{\boldsymbol{U}^T\boldsymbol{\Sigma}(\rho)^{-1}\boldsymbol{U}\}^{-1}]. \end{aligned}$$

By (4), (10) and the Sherman-Morrison-Woodbury identity, we obtain that for $0 \leq \rho < 1$,

$$\boldsymbol{\Sigma}(\rho)^{-1} = (1-\rho)^{-1}\left[\boldsymbol{I}_N - \boldsymbol{Z}\boldsymbol{\Xi}(\rho)\boldsymbol{Z}^T\right],$$

and thus

$$\text{AV}_G(\rho) = \sigma_T^2(1-\rho)\text{tr}[\boldsymbol{D}^{-2}\{\boldsymbol{I}_K - \boldsymbol{U}^T\boldsymbol{Z}\boldsymbol{\Xi}(\rho)\boldsymbol{Z}^T\boldsymbol{U}\}^{-1}]. \tag{A.1}$$

When $\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{X}$, (9) leads to $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^T$. On the other hand, $\boldsymbol{X}$ and $\boldsymbol{U}$ have the same column space, so the projections onto the two column spaces are the same, i.e., $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{U}\boldsymbol{U}^T$. It follows that $\boldsymbol{D}^2 = \boldsymbol{I}_K$ and thus $\boldsymbol{D} = \boldsymbol{I}_K$, whence (A.1) becomes

$$\begin{aligned} \text{AV}_{X^TX}(\rho) &= \sigma_T^2(1-\rho)\text{tr}[\{\boldsymbol{I}_K - \boldsymbol{U}^T\boldsymbol{Z}\boldsymbol{\Xi}(\rho)\boldsymbol{Z}^T\boldsymbol{U}\}^{-1}] \\ &= \sigma_T^2(1-\rho)\sum_{k=1}^{K}\frac{1}{1-\gamma_k(\rho)}, \end{aligned}$$

proving (11).

*Appendix B. Proof of Lemma 1*

For (a), first note that $\boldsymbol{\Xi}(1) = \text{Diag}\{m_1^{-1},\ldots,m_n^{-1}\} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$, so $\boldsymbol{A}(1) = \boldsymbol{U}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{U}$ and thus

$$\gamma_1(1) = \max_{\|v\|=1}\boldsymbol{v}^T\boldsymbol{U}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{U}\boldsymbol{v} \leq \max_{\|w\|=1}\boldsymbol{w}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{w} = 1.$$

To see that this upper bound is attained by $\boldsymbol{v}_1(1) = N^{-1/2}\boldsymbol{U}^T\boldsymbol{1}_N$, observe that (15) implies

$$\text{col}(\boldsymbol{U}) \cap \text{col}(\boldsymbol{Z}) = \{c\boldsymbol{1}_N : c \in \mathbb{R}\} \tag{A.2}$$

  

and that $\boldsymbol{\Xi}(1) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$. Thus

$$\boldsymbol{A}(1)(\boldsymbol{U}^T\mathbf{1}_N) = \boldsymbol{U}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{U}\boldsymbol{U}^T\mathbf{1}_N = \boldsymbol{U}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{1}_N = \boldsymbol{U}^T\mathbf{1}_N.$$

For (b), if $\gamma_2(1) = 1$ then

$$1 = \boldsymbol{v}_2(1)^T\boldsymbol{A}(1)\boldsymbol{v}_2(1) = [\boldsymbol{U}\boldsymbol{v}_2(1)]^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T[\boldsymbol{U}\boldsymbol{v}_2(1)].$$

Since $\|\boldsymbol{U}\boldsymbol{v}_2(1)\| = 1$, the above implies $\boldsymbol{U}\boldsymbol{v}_2(1) \in \mathrm{col}(\boldsymbol{Z})$. By (A.2) we must have $\boldsymbol{U}\boldsymbol{v}_2(1) = \pm N^{-1/2}\mathbf{1}_N$; premultiplying by $\boldsymbol{U}^T$ gives $\boldsymbol{v}_2(1) = \pm\boldsymbol{v}_1(1)$, a contradiction. It follows that $\gamma_2(1) < 1$.

Let $\boldsymbol{M}(\rho)$ be the $K \times (K-2)$ matrix with columns $\boldsymbol{v}_3(\rho), \ldots, \boldsymbol{v}_K(\rho)$. By Theorem 3.17 of (33),

$$\begin{aligned} \gamma_2(\rho) &= \min_{\substack{\|w\|=1,\\ M(\rho)^T w=0}} \boldsymbol{w}^T\boldsymbol{A}(\rho)\boldsymbol{w} \\ &\leq \min_{\substack{\|w\|=1,\\ M(\rho)^T w=0}} \boldsymbol{w}^T\boldsymbol{A}(1)\boldsymbol{w} \end{aligned} \tag{A.3}$$

since $\boldsymbol{A}(1) - \boldsymbol{A}(\rho)$ is positive definite for $\rho \in [0, 1)$. By the Courant-Fischer theorem (e.g., Theorem 3.18 of (33)),

$$\gamma_2(1) = \max_{\substack{C\in\mathbb{R}^{K\times(K-2)}:\\ C^TC=I_{K-2}}} \min_{\substack{\|w\|=1,\\ C^Tw=0}} \boldsymbol{w}^T\boldsymbol{A}(1)\boldsymbol{w} \geq \min_{\substack{\|w\|=1,\\ M(\rho)^Tw=0}} \boldsymbol{w}^T\boldsymbol{A}(1)\boldsymbol{w}.$$

Combining this with (A.3) gives $\gamma_2(\rho) \leq \gamma_2(1) < 1$.

For (c), the fact that 1 is a simple eigenvalue of $\boldsymbol{A}(1)$ allows us to apply the implicit function theorem to deduce that $\gamma(\cdot)$ is continuously differentiable at 1 (34; 35). In particular, using (16) and letting $\boldsymbol{A}'(\cdot), \boldsymbol{\Xi}'(\cdot)$ denote elementwise differentiation with respect to $\rho$,

$$\begin{aligned} \gamma_1'(1) &= \boldsymbol{v}_1(1)^T\boldsymbol{A}'(1)\boldsymbol{v}_1(1) \\ &= N^{-1}\mathbf{1}_N^T\boldsymbol{U}[\boldsymbol{U}^T\boldsymbol{Z}\boldsymbol{\Xi}'(1)\boldsymbol{Z}^T\boldsymbol{U}]\boldsymbol{U}^T\mathbf{1}_N \\ &= N^{-1}\mathbf{1}_N^T\boldsymbol{Z}\boldsymbol{\Xi}'(1)\boldsymbol{Z}^T\mathbf{1}_N. \end{aligned} \tag{A.4}$$

But $\boldsymbol{Z}^T\mathbf{1}_N = (m_1, \ldots, m_n)^T$ and

$$\boldsymbol{\Xi}'(1) = \mathrm{Diag}\left\{\frac{1}{[1+(m_1-1)\rho]^2}, \ldots, \frac{1}{[1+(m_n-1)\rho]^2}\right\}\Bigg|_{\rho=1} = \mathrm{Diag}\{m_1^{-2}, \ldots, m_n^{-2}\},$$

so (A.4) leads to $\gamma_1'(1) = n/N = 1/\bar{m}$.

*Appendix C. Proof that (19) implies $r_G(\rho) = 1 + (m-1)\rho$ for general $\boldsymbol{G}$*

By Theorem 3.34 of (33), (A.1) implies

$$\sum_{k=1}^K \frac{1}{d_k^2[1-\gamma_k(\rho)]} \leq \frac{\mathrm{AV}_G(\rho)}{\sigma_T^2(1-\rho)} \leq \sum_{k=1}^K \frac{1}{d_{K+1-k}^2[1-\gamma_k(\rho)]}$$

(a generalization of (11)). When (19) holds, $\frac{1}{1-\gamma_k(\rho)} = \frac{1+(m-1)\rho}{1-\rho}$ for each $k$, so the first and last expressions above are both equal to $\frac{1+(m-1)\rho}{1-\rho}\sum_{k=1}^K \frac{1}{d_k^2}$ and hence $\mathrm{AV}_G(\rho) = \sigma_T^2[1+(m-1)\rho]\sum_{k=1}^K \frac{1}{d_k^2}$, leading to $r_G(\rho) = \mathrm{AV}_G(\rho)/\mathrm{AV}_G(0) = 1 + (m-1)\rho$.

*Appendix D. How $r_G(\rho)$ was computed in section 4.3*

For $m = 2, 3, 4, 5$, we simulated $100m \times 3$ design matrices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{299}$ with rows of the form $(1, t_{ij}, t_{ij}^2)^T$. The ages $t_{ij}$ for each design matrix were generated by (i) simulating age in years at study entry ($t_{i1}$) for 100 participants from the uniform distribution on $[10, 30 - L]$, where $L$ is the study length in years (0.5 or 3), and (ii) obtaining follow-up times $t_{i2}, \ldots, t_{im}$ by drawing $t_{i2} - t_{i1}, \ldots, t_{im} - t_{i1}$ from the uniform distribution on $[0, L]$. Then, for $a = 1, \ldots, 299$, we computed $r_G(\rho)$ with $\boldsymbol{X} = \boldsymbol{X}_a$ and $\rho = \frac{a}{300}$, and these are plotted against $\rho$ in Figure 1. Since the simulated ages are marginally approximately uniform on $[10, 30]$, $\boldsymbol{G}$ corresponding to integration over $[10, 30]$ is roughly proportional to $\boldsymbol{X}^T\boldsymbol{X}$. This helps explain why the differences between $r_G(\rho)$ and $r_{X^TX}(\rho)$ in both panels of the figure are minor. For age distributions that are farther from uniform, larger differences might be observed.

*Appendix E. Details of the simulation study in section 7*

The real data analysis included overlapping cross-sectional and longitudinal data sets, each consisting of 300 vertices from each of 398 brain images. For the simulation study we used the ages $t_{ij}$ from the longitudinal data set, excluding 3 observations with age$< 5$ and 1 observation with age$> 30$, for a total of 394 ages for 196 individuals. Each set of responses was generated from the model $y_{ij} = f(t_{ij}) + u_i + \varepsilon_{ij}$ with $f$ given by (26) and $u_i, \varepsilon_{ij}$ drawn independently from the $N(0, \hat{\sigma}_u^2)$ and $N(0, \hat{\sigma}_\varepsilon^2)$ distributions, respectively, where $\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2$ are the estimates for one of the vertices. This was repeated for 294 of the 300 vertices studied in section 6.1; the other 6 were excluded due to unusually high $\hat{\sigma}_\varepsilon^2$ values. For each of the 294 sets of responses, we fitted a "cross-sectional" nonparametric regression model treating the 394 observations as independent, and a "longitudinal" nonparametric model with random subject effects. As in section 6.1, the models were implemented with the `mgcv` and `gamm4` packages using a 15-dimensional $P$-spline basis.

## References

[1] US National Institute of Mental Health. NIMH Strategic Plan for Research 2016. Retrieved from https://www.nimh.nih.gov/about/strategic-planning-reports/index.shtml.

[2] Kraemer HC, Yesavage JA, Taylor JL, Kupfer D. How can we learn about developmental processes from cross-sectional studies, or can we? *American Journal of Psychiatry* 2000; **157**(2):163–171.

[3] Horga G, Kaur T, Peterson BS. Annual Research Review: Current limitations and future directions in MRI studies of child- and adult-onset developmental psychopathologies. *Journal of Child Psychology and Psychiatry* 2014; **55**(6):659–680.

[4] Sowell ER, Thompson PM, Leonard CM, Welcome SE, Kan E, Toga AW. Longitudinal mapping of cortical thickness and brain growth in normal children. *Journal of Neuroscience* 2004; **24**(38):8223–8231.

[5] 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. *Proceedings of the National Academy of Sciences* 2007; **104**(49):19 649–19 654.

[6] Reiss PT, Crainiceanu CM, Thompson WK, Huo L. Modeling change in the brain: methods for cross-sectional and longitudinal data. *Handbook of Neuroimaging Data Analysis*, Ombao H, Lindquist M, Thompson W, Aston J (eds.). Chapman and Hall/CRC, 2016.

[7] Kish L. *Survey Sampling*. New York: Wiley Classics Library, 1995.

  

[8] Donner A, Birkett N, Buck C. Randomization by cluster: Sample size requirements and analysis. *American Journal of Epidemiology* 1981; **114**(6):906–914.

[9] Bell RQ. Convergence: An accelerated longitudinal approach. *Child Development* 1953; **24**(2):145–152.

[10] Harezlak J, Ryan LM, Giedd JN, Lange N. Individual and population penalized regression splines for accelerated longitudinal designs. *Biometrics* 2005; **61**(4):1037–1048.

[11] Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**(3):983–997.

[12] Kulis B, Sustik MA, Dhillon IS. Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research* 2009; **10**(Feb):341–376.

[13] Berger MPF. A comparison of efficiencies of longitudinal, mixed longitudinal, and cross-sectional designs. *Journal of Educational and Behavioral Statistics* 1986; **11**(3):171–181.

[14] Fjell AM, Walhovd KM, Westlye LT, Østby Y, Tamnes CK, Jernigan TL, Gamst A, Dale AM. When does brain aging accelerate? Dangers of quadratic fits in cross-sectional studies. *NeuroImage* 2010; **50**(4):1376–1383.

[15] Reiss PT, Huang L, Chen YH, Huo L, Tarpey T, Mennes M. Massively parallel nonparametric regression, with an application to developmental brain mapping. *Journal of Computational and Graphical Statistics* 2014; **23**(1):232–248.

[16] Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: New York, 2003.

[17] Wood SN. *Generalized Additive Models: An Introduction with R*. Chapman & Hall: Boca Raton, Florida, 2006.

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

[19] Wahba G. Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B* 1983; **45**:133–150.

[20] Buja A, Hastie T, Tibshirani R. Linear smoothers and additive models. *Annals of Statistics* 1989; **17**(2):453–510.

[21] Giedd JN, Raznahan A, Alexander-Bloch A, Schmitt E, Gogtay N, Rapoport JL. Child Psychiatry Branch of the National Institute of Mental Health Longitudinal Structural Magnetic Resonance Imaging Study of Human Brain Development. *Neuropsychopharmacology* 2015; **40**(1):43–49.

[22] Kim JS, Singh V, Lee JK, Lerch J, Ad-Dab'bagh Y, MacDonald D, Lee JM, Kim SI, Evans AC. Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *NeuroImage* 2005; **27**(1):210–221.

[23] Eilers PHC, Marx BD. Flexible smoothing with $B$-splines and penalties (with discussion). *Statistical Science* 1996; **11**(2):89–121.

[24] Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B* 2011; **73**(1):3–36.

[25] Wood S, Scheipl F. *gamm4: Generalized additive mixed models using mgcv and lme4* 2014. URL https://CRAN.R-project.org/package=gamm4, r package version 0.2-3.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2016. URL https://www.R-project.org/.

[27] Reiss PT, Ogden RT. Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B* 2009; **71**(2):505–523.

[28] Walhovd KB, Fjell AM, Giedd J, Dale AM, Brown TT. Through thick and thin: a need to reconcile contradictory results on trajectories in human cortical development 2016. *Cerebral Cortex*, in press.

[29] Alexander-Bloch A, Clasen L, Stockman M, Ronan L, Lalonde F, Giedd J, Raznahan A. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human Brain Mapping* 2016; **37**:2385–2397.

[30] Thompson WK, Hallmayer J, O'Hara R. Design considerations for characterizing psychiatric trajectories across the lifespan: Application to effects of APOE-e4 on cerebral cortical thickness in Alzheimer's disease. *American Journal of Psychiatry* 2011; **168**(9):894–903.

[31] Myers RH, Montgomery DC, Anderson-Cook CM. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. 4th edn., John Wiley & Sons: Hoboken, NJ, 2016.

[32] Fedorov VV. *Theory of Optimal Experiments*. Academic Press: New York, 1972.

[33] Schott JR. *Matrix Analysis for Statistics*. 2nd edn., Wiley: New York, 2005.

[34] Magnus JR. On differentiating eigenvalues and eigenvectors. *Econometric Theory* 1985; **1**(2):179–191.

[35] Harville DA. *Matrix Algebra from a Statistician's Perspective*. Springer: New York, 1997.

  