Statistics and Its Interface Volume 8 (2015) 379–390

Group analysis of fMRI data using L_1 and L_2 regularization

Rosanna Overholser and Ronghui Xu*

In clinical studies using functional magnetic resonance imaging (fMRI), it is of interest to compare multiple subjects from different groups. We investigate the analysis of such data using random effects and non-parametric estimation of mean activation curves. The random effects modeling replaces the existing approach in fMRI literature where each curve is 'normalized' by a percent change. For the mean curves we consider smoothing via splines using L_1 or L_2 regularization. Our general framework allows analysis of fMRI curves that are correlated, and with correlated within curve errors. We describe a unified algorithm that uses existing software to carry out the estimation. The different regularization approaches are compared using simulation. We apply the method to an fMRI study about the effects of caffeine on the motor cortex of the brain, and discuss the limitation on currently available computing resources for carrying out such analysis on very large data sets.

AMS 2000 subject classifications: Primary 62.
Keywords and phrases: Correlated curves, Correlated errors, Functional linear model, Penalized splines, Semiparametric mixed-effects model, Voxel level analysis.

1. INTRODUCTION

Functional magnetic resonance imaging (fMRI) is a popular method of studying brain activity by measuring blood flow to the brain. Researchers in a variety of fields rely on fMRI as a non-invasive technique to explore how minds are working (or not) in experiments. Introductions to fMRI data analysis can be found in [1, 21, 29]. However, the analysis of fMRI data remains challenging for several reasons. Data is often high-dimensional and noisy. Sample sizes are often small. In addition to variation between study subjects, there is also variation between two fMRI sessions of the same subject. Data from a single subject is correlated in both time and space.

In clinical studies it is often of interest to analyze fMRI data from groups of study subjects under different conditions. Since there is substantial between subject and session variation (see Figure 5) it is common to standardize by conversion to a percent change; that is, to divide the observed activation time series by some kind of baseline activation such as its mean over time. A side effect of doing so is that it artificially induces variation in the amplitudes of activation from subject to subject. Here we consider a different approach via modeling the between-subject variation using random effects.

Of clinical interest is the underlying mean activation functions from the groups, or their contrasts. In this paper we consider the semi-(non)parametric mixed effect model where the mean activation functions are unspecified. These are also referred to as functional linear models (see for example, [11, 38, 39]). In the functional data literature, the errors are often assumed to be i.i.d. This, however, is not the case for fMRI data. In addition in our application the curves (i.e. time series) are paired. We explore an approach that can accommodate potentially complex correlation among the curves. We consider the L_1 and L_2 regularization, together with splines to obtain nonparametric estimate of the mean function(s) in the presence of correlated errors. With L_1 penalty and a large number of basis functions, it is a way of knot selection with simultaneous penalized estimation of the spline coefficients. With L_2 penalty and penalized splines, we take advantage of the fact that the combined spline function and random effects can be fitted in a single mixed effects model framework.

1.1 A brief review of L_1, L_2 and mixed effects

For the L_1 regularization, the least absolute shrinkage and selection operator (LASSO) was introduced in [35] for the purposes of variable selection and estimation in linear regression. A discussion of the LASSO in comparison with other shrinkage techniques can be found in [15]. Knight and Fu [19] studied the asymptotics of LASSO estimates in the context of linear regression with i.i.d. errors; in a doctoral dissertation Gupta [12] studied the asymptotics of LASSO estimates in the presence of correlated errors and a large number of covariates. Wang, Li and Tsai [36] studied the problem of joint selection of covariates and order of autoregressive process via the LASSO. The LASSO has been applied to nonparametric setting with i.i.d. errors [26], and some related theory is presented in [4, Section 6.2.3].

Two applications of L_1 regularization in mixed effects models were recently proposed [2, 17], where the number of predictors is less than the number of observations. More recently, Schelldorfer *et al.* [32] proposed L_1 regularization for linear mixed models when the number of predictors is

*Corresponding author.

much larger than the number of observations, and where model selection only concerned the fixed effects.

The L_2 regularization has been widely used in the nonparametric setting including smoothing splines and penalized splines. The penalized splines (p-splines) have enjoyed popularity since its initial introduction in [7], mainly due to its relative computational ease. An overview of the p-splines can be found in [31]. The asymptotic theory for p-splines was first established in [14], in the case of independent errors and uniformly dense knots. It was then extended to explicitly account for the number of knots, non-normal outcomes, and to make connections with kernel smoothing [37, 18, 23, 6]. There is very little work on the theory of p-splines with correlated errors.

Recognizing the connection between the L_2 penalty and the random effects likelihood function [34, 31], Brumback and Rice in [3] applied smoothing splines to the analysis of nested curves. This mixed-effects model approach has become popular in the fMRI literature (see [24] and references therein).

2. MODEL AND ESTIMATION FOR THE FMRI DATA

Consider the functional data from a fMRI session as a curve over time. To describe the statistical methodology in this section while keeping the notation simple, let y_{ik} be the measurement at time t_k for curve i, $i = 1, \dots, m$ and $k = 1, \dots, n$. Here we assume that the measurements are taken at the same time points $t_1, \dots, t_n$ for all the curves, as is the case for our fMRI data. We assume that the curves are independent, coming from different study subjects. Later in the applications we generalize the model to correlated curves, but the statistical methodology remains the same. We assume the following model:

$$y_{ik} = \mu(t_k) + z_{ik}^\top b_i + \epsilon_{ik}, \tag{1}$$

where $\mu(t)$ is a smooth function of t, $z_{ik}^\top b_i$ is a curve specific deviation from $\mu(t_k)$ and $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in})^\top$ is distributed as $\mathrm{N}(0, V_1)$. We assume that the b_i's are vectors of random effects independently drawn from $\mathrm{N}(0, D_1)$, each associated with a covariate vector z_{ik}.

2.1 Regularized spline fit of $\mu(t)$

We approximate the function $\mu(t)$ over the interval $[t_1, t_n]$ by a linear combination of basis functions. B-spline basis are often preferred in conjunction with L_2 penalization, for their numerical stability and compact support [8]. For the L_1 penalty, however the minimal overlapping support of the B-spline basis is a disadvantage; when one spline is removed from the set via a zero coefficient, the entire basis must be re-calculated, or otherwise a 'dip' will appear in the estimated function. In the following we use the truncated power basis functions, which also makes the presentation simple. We start with a large number of evenly spaced knots in the interval $[t_1, t_n]$. An L_1 penalty simultaneously reduces the number of knots and estimates the spline coefficients that are non-zero. Alternately, an L_2 penalty keeps all the knots and penalizes the estimated spline coefficients.

The truncated cubic spline basis can be written $\{1, t, t^2, t^3, (t-\tau_1)_+^3, \dots, (t-\tau_Q)_+^3\}$, where $(x)_+ = x$ if $x > 0$ and 0 otherwise, and $\tau_1, \dots \tau_Q$ are the knots. We approximate

$$\mu(t) \approx \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{l=1}^{Q} \beta_{l+3}(t-\tau_l)_+^3. \tag{2}$$

For each curve i, denote $y_i = (y_{i1}, \dots, y_{in})^\top$, and

$$X_i = \begin{bmatrix} 1 & t_1 & t_1^2 & t_1^3 & (t_1-\tau_1)_+^3 & \dots & (t_1-\tau_Q)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & t_n & t_n^2 & t_n^3 & (t_n-\tau_1)_+^3 & \dots & (t_n-\tau_Q)_+^3 \end{bmatrix}.$$

Let $Z_i = (z_{i1}^\top, ..., z_{in}^\top)^\top$. Let $X = (X_1^\top, \dots, X_m^\top)^\top$, $\beta = (\beta_0, \dots, \beta_{Q+3})^\top$, $Z = \mathrm{diag}(Z_1, \dots Z_m)$, $b = (b_1^\top, \dots, b_m^\top)^\top$, $\epsilon = (\epsilon_1^\top, \dots, \epsilon_m^\top)^\top$ and $y = (y_1^\top, \dots, y_m^\top)^\top$. Then model (1) becomes $y = X\beta + Zb + \epsilon$ in matrix form. Let $D = \mathrm{diag}(D_1, \dots, D_1)$, $V = \mathrm{diag}(V_1, \dots, V_n)$, and $\Sigma = V + ZDZ^\top$. Apart from a constant the log-likelihood of y is then

$$l(y|\beta, \Sigma) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y - X\beta)^\top \Sigma^{-1}(y - X\beta). \tag{3}$$

We estimate the parameters in (3) by maximizing a penalized log-likelihood. To make clear the distinction between the unpenalized and penalized coefficients, we partition $\beta = (\beta_u^\top, \beta_p^\top)^\top$ where the unpenalized coefficients are in $\beta_u = (\beta_0, ..., \beta_3)^\top$, and the penalized coefficients are in $\beta_u = (\beta_4, ..., \beta_{Q+3})^\top$. Let $\hat\beta$ and $\hat\Sigma$ jointly maximize the penalized log-likelihood

(4)
$$l_p(y|\beta, \Sigma) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y - X\beta)^T \Sigma^{-1}(y - X\beta) - p_\lambda(\beta_p),$$

where $p_\lambda(\beta_p)$ is a penalty term, and $\lambda > 0$ is a penalty parameter. We can choose λ by minimizing either the AIC or the BIC:

$$\hat\lambda = \underset{\lambda}{\mathrm{argmin}}\{-l(y|\hat\beta, \hat\Sigma) + a \cdot \mathrm{DF}\}, \tag{5}$$

where DF is the effective degrees of freedom. For the L_1 penalty, the DF is the number of non-zero parameters in the model; while for the L_2 penalty, it is the trace of the 'hat' matrix H plus the number of error covariance parameters [27, 28]. The hat matrix H is such that $\hat y = X\hat\beta + Z\hat b = Hy$. For AIC we let $a = 1$. For BIC [32] considered $a = \log(nm)$, i.e. nm is considered as the sample size. On the other hand, one may consider the number of independent curves, m, as the sample size, so that $a = \log(m)$.

L_1 regularization

If the L_1 penalty is used, $p_\lambda^1(\beta_p) = \lambda \sum_{q=4}^{Q+3} |\beta_q|$ in (4). In this case some of the coefficients in β_p will be estimated to be exactly zero. Therefore the L_1 regularization is a way of knot selection. However, this is different from the knot selection such as considered by [16, 40], in the sense that the non-zero coefficients are also penalized. In contrast, the more classical knot selection approach is followed by regression splines, where none of the spline coefficients are shrunk towards zero, given the selected knots.

L_2 regularization

When the L_2 penalty is used, $p_\lambda^2(\beta_p) = \lambda\sum_{q=4}^{Q+3}(\beta_q)^2$. It is recognized that (4) is the 'log-likelihood' with elements in β_p acting as i.i.d. normally distributed random effects with variance equal to $2/\lambda$; here the 'log-likelihood' refers to the joint log-likelihood of y and β_p. It has also been recognized that given λ, the estimated $\hat\beta_p$ from (4) is the best linear unbiased predictor (BLUP) in this mixed effects model formulation [34, 31, 3]. The equivalent mixed models formulation has been exploited in analyzing fMRI data.

Unified computational framework

The criterion in (4) may be non-convex in β and the parameters in Σ. In the following we iterate between the penalized and the non-penalized parameters until convergence. In other words, given the penalty parameter λ, at the $(j+1)$th step we iterate between

$$(\beta_p^{j+1}|\beta_u^j,\Sigma^j) = \operatorname*{argmin}_{\beta_p}\{(y - X_u\beta_u^j - X_p\beta_p)^\top(\Sigma^j)^{-1} (y - X_u\beta_u^j - X_p\beta_p) + p_\lambda(\beta_p)\} \tag{6}$$

and

$$(\Sigma^{j+1},\beta_u^{j+1}|\beta_p^j) = \operatorname*{argmin}_{\Sigma,\beta_u}\left\{\frac{1}{2}\log|\Sigma| + \frac{1}{2}(y - X_u\beta_u - X_p\beta_p^j)^\top \Sigma^{-1}(y - X_u\beta_u - X_p\beta_p^j)\right\}. \tag{7}$$

When the L_1 penalty $p_\lambda^1(\beta_p)$ is used, β_p^{j+1} may be obtained from (6) using any of the standard methods for LASSO in linear models (LARS, cyclic or greedy coordinate descent, homotopy); we found that the R package 'lars' worked well. When the L_2 penalty $p_\lambda^2(\beta_p)$ is used, (6) is simply ridge regression, and we use the function 'lm.ridge' from the R package 'MASS'. Finally (7) is equivalent to fitting a linear mixed model on the outcome $y^* = y - X_p\beta_p^j$, so one can use any software for such models, and we use the R package 'nlme'.

Once the estimates of β and Σ are obtained, predictions of the random effects b can be obtained by the best linear unbiased predictor (BLUP)

$$\hat b = \hat D Z^\top\hat\Sigma^{-1}(y - X\hat\beta). \tag{8}$$

As mentioned earlier, the estimated b_i's help to capture the between-subject variation among the curves.

2.2 Standard errors

To make inference about the estimated mean function $\mu(t)$, we consider the sandwich estimator of [9]. Denote the vector of all non-zero elements of the estimated β and the parameters in Σ by θ. Let ∇l and $\nabla^2 l$ be the first and second derivatives of l in (3) with respect to θ. The sandwich estimate for the covariance of $\hat\theta$ is

$$\widehat{\text{cov}}(\hat\theta) = m\{\nabla^2 l(\hat\theta) - \Lambda\}^{-1}\widehat{\text{cov}}\{\nabla l(\hat\theta)\}\{\nabla^2 l(\hat\theta) - \Lambda\}^{-1}, \tag{9}$$

where the (k,s)-element of Λ is $\partial^2 p_\lambda(\theta)/\partial\theta_k\partial\theta_s$, and the (k,s)-element of $\widehat{\text{cov}}\{\nabla l(\hat\theta)\}$ is

$$\left\{\frac{1}{m}\sum_{i=1}^m \frac{\partial l_i(\hat\theta)}{\partial\theta_k}\frac{\partial l_i(\hat\theta)}{\partial\theta_s}\right\} - \left\{\frac{1}{m}\sum_{i=1}^m \frac{\partial l_i(\hat\theta)}{\partial\theta_k}\right\}\left\{\frac{1}{m}\sum_{i=1}^m \frac{\partial l_i(\hat\theta)}{\partial\theta_s}\right\}. \tag{10}$$

When $p_\lambda(\theta)$ is the L_1 penalty, following [10] we approximate $\partial^2 p_\lambda(\theta)/\partial\beta_k\partial\beta_k$ by $1/|\beta_k|$. The formulas for ∇l and $\nabla^2 l$ are given in the Appendix. Once $\widehat{\text{cov}}(\hat\beta)$ is obtained, a 95% pointwise confidence interval for $\mu(t_k)$ may be formed by $\hat\mu(t_k) \pm 1.96c_k$, where c_k is the kth element of $\sqrt{\text{diag}(X\widehat{\text{cov}}(\hat\beta)X^\top)}$.

3. SIMULATIONS

It has been known that proper estimation of the error variance-covariance is important for the selection of penalty parameter [33, 20]. In order to understand the performance of the different penalties in the setting of nonparametric smoothing with correlated errors, we carry out simulation studies in this section. The simulations also investigate the different criteria (AIC or BIC) in choosing the penalty parameter λ.

For each of 100 simulation runs, we generate $m = 5$, 10 or 40 curves under model (1), each of length $n = 70$ points. The true curve is either $\mu(t) = \sin(10\pi t/70)$ ('sine') or one that mimics hemodynamic response ('fmri'). For the latter we use the R package *neuRosim*, which allows for three models of hemodynamic response function and various types of noise. We use the following double gamma model of hemodynamic response:

$$h(t) = \left(\frac{t}{d_1}\right)^{a_1}\exp\left(-\frac{t-d_1}{e_1}\right) - c\left(\frac{t}{d_2}\right)^{a_2}\exp\left(-\frac{t-d_2}{e_2}\right) \tag{11}$$

with the default setting for the parameters: $a_1 = 6$, $a_2 = 12$, $e_1 = e_2 = 0.9$, $c = 0.35$, $d_i = a_ie_i$, $i = 1,2$. This hemodynamic response is then convoluted with an experiment design of 10 seconds of rest, followed by 30 seconds of an activity, followed by 30 seconds of rest. The resulting true mean curve is shown in the right panel of Figure 1. Data are generated from either the 'sine' or 'fmri' curve with errors ϵ_i distributed as $\text{N}(0,\sigma^2R_1(\rho))$ where $R_1(\rho)$ is the correlation matrix of a first order autoregressive process AR(1) with

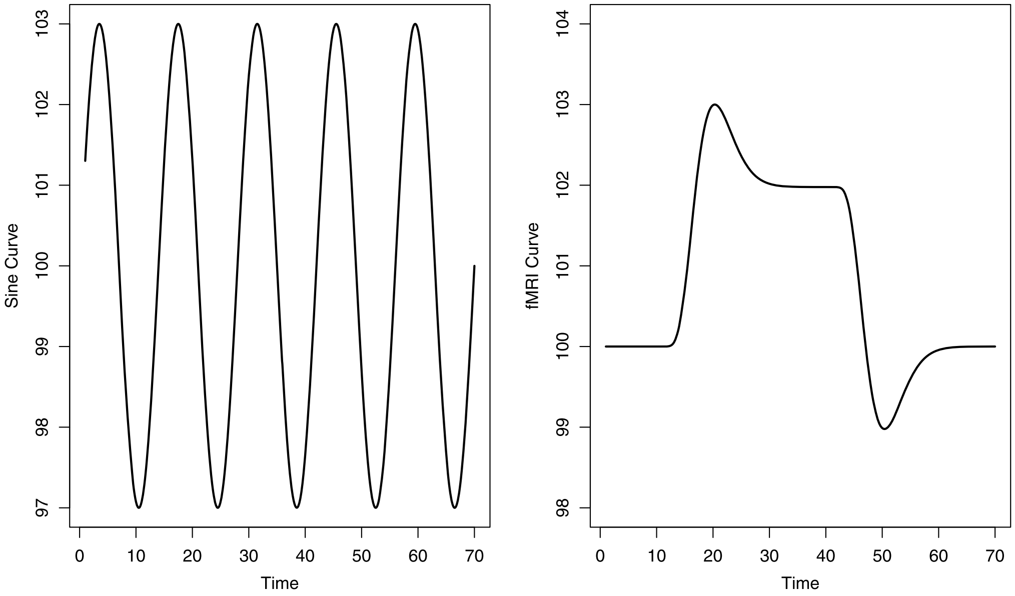

Figure 1. True mean curves used in simulation.

correlation parameter $\rho = 0.4$. We carry out simulations both with and without random effects.

For each simulated data set, we estimate the mean curve using the regularization methods described in the previous section. For the L_1 regularization the initial knots are placed at every data point, and for the L_2 regularization the knots are placed at every other data point. For each method, the penalty parameter λ is chosen to minimize the AIC, and the BIC with the sample size m or nm. Convergence is declared if the sum of the absolute changes in the parameters is less than 10^{-4}.

Figure 2 left panel shows the boxplots of the integrated squared error (ISE) of the estimated $\mu(t)$ from the 100 simulation runs, when no random effects were simulated or fitted, i.e. $z_{ik} = 0$ in model (1). For the right panel we simulated random intercept for each curve, i.e. $z_{ik} = 1$ and $b_i \sim N(0, 100)$. We then used the estimated b_i to 'center' each curve, and the ISE's are averaged over the m curves. Comparing the left and right panels it is seen that there is relatively little difference with or without the random effects; this makes sense since we have a large 'cluster' size of 70 for each curve to estimate the b_i's. From the plots we see that the ISE decreases with the number of curves. BIC(nm) leads to underperformance compared to AIC and BIC(m), in particular for the 'fmri' true curve. AIC and BIC(m) are roughly comparable, with AIC performing slightly better when $m = 5$ and under L_1 regularization for the 'fmri' curve. Overall there is not a clear preference between the L_1 and L_2 regularization methods in term of ISE. Supplement Figure 1 (http://www.intlpress.com/SII/p/2015/8-3/SII-8-3-overholser-supplement.pdf) contains 3 randomly selected simulation runs with $m = 10$, no random effects, and using AIC; upon close inspection the fits using the L_2 regularization appear slightly less smooth.

Supplement Figure 2 shows the boxplots of the estimated correlation ρ, Supplement Figure 3 shows the boxplots of the estimated error variance σ^2, and Supplement Figure 4 shows the boxplots of the estimated random effect variance. It is seen that all these parameters are underestimated the majority of the times, though the bias reduces with increasing number of curves. For correlation and error variance BIC(nm) appear to provide less biased though more varied estimates. Also L_1 regularization appears to provide less biased estimates of σ^2.

At the suggestion of a reviewer we contrast in Figure 3 the L_1 and L_2 fits with no penalty fits, where knots were placed at every other data point. We note that the necessity of regularization is very much data dependent, in particular dependent on the amount of data as compared to the number of parameters, as well as the signal-to-noise ratio. For $m = 5$ and 10 we see that in terms of ISE there is little advantage of regularization under the true 'sine' curve; but when the true curve is 'fmri' L_1 regularization clearly reduces the ISE. When $m = 40$ (data not shown) even L_1 regularization no longer appears to have a clear advantage. On the other hand, the variance parameters appear to be better estimated with regularization in general (see Supplement Figures 5–7). Tabulated simulation results are given in the Supplement Tables 1–4.

4. FMRI DATA ANALYSIS

The Center for Functional MRI at UCSD performed a study to examine the effect of caffeine on the blood oxygenation level dependent (BOLD) signal from fMRI sessions

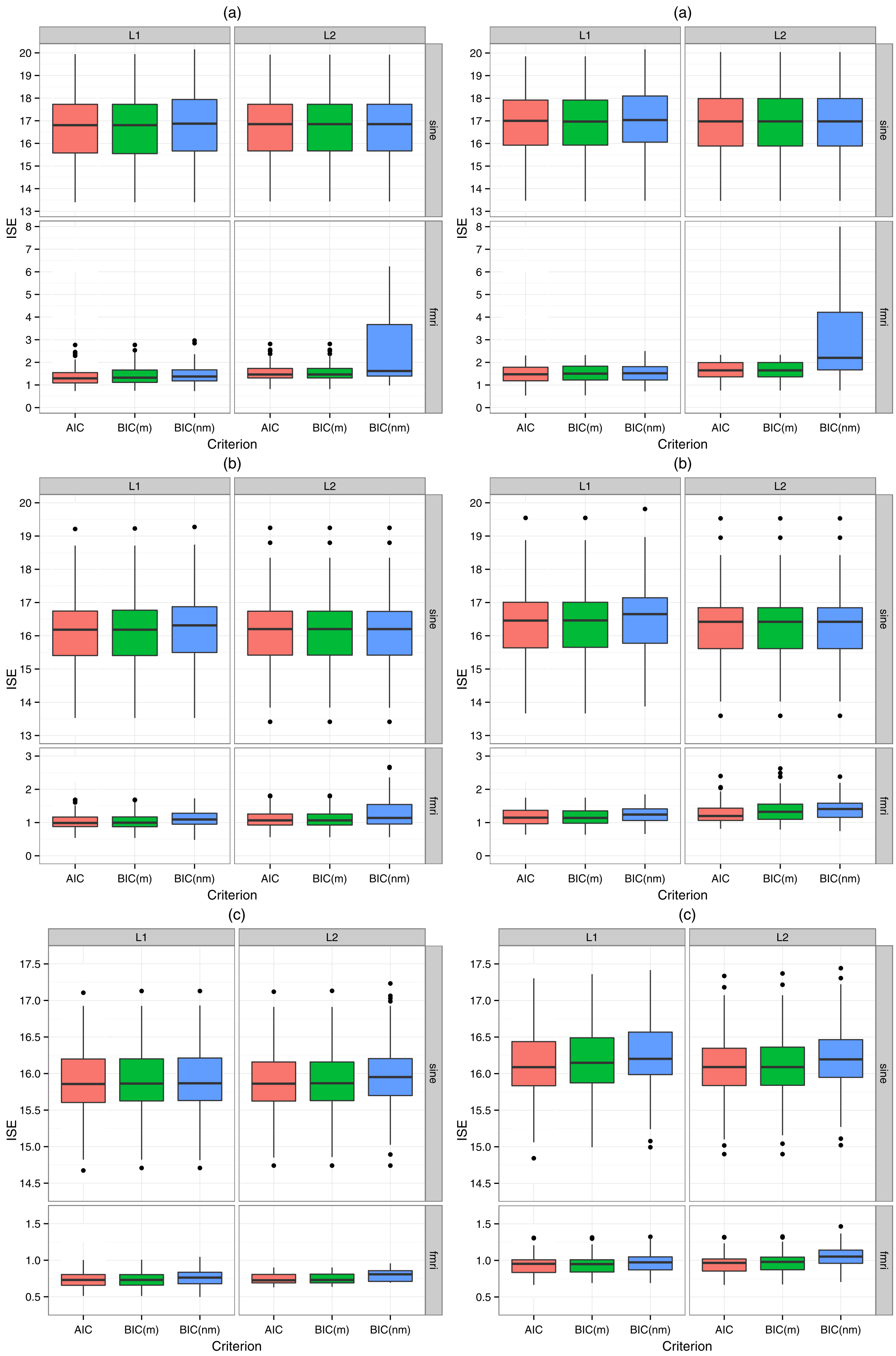

Figure 2. Boxplots of (average) integrated squared error (ISE) of the estimated curves. Left panel: without random effects; right panel: with random effects. (a) $m = 5$*, (b)* $m = 10$*, (c)* $m = 40$*.*

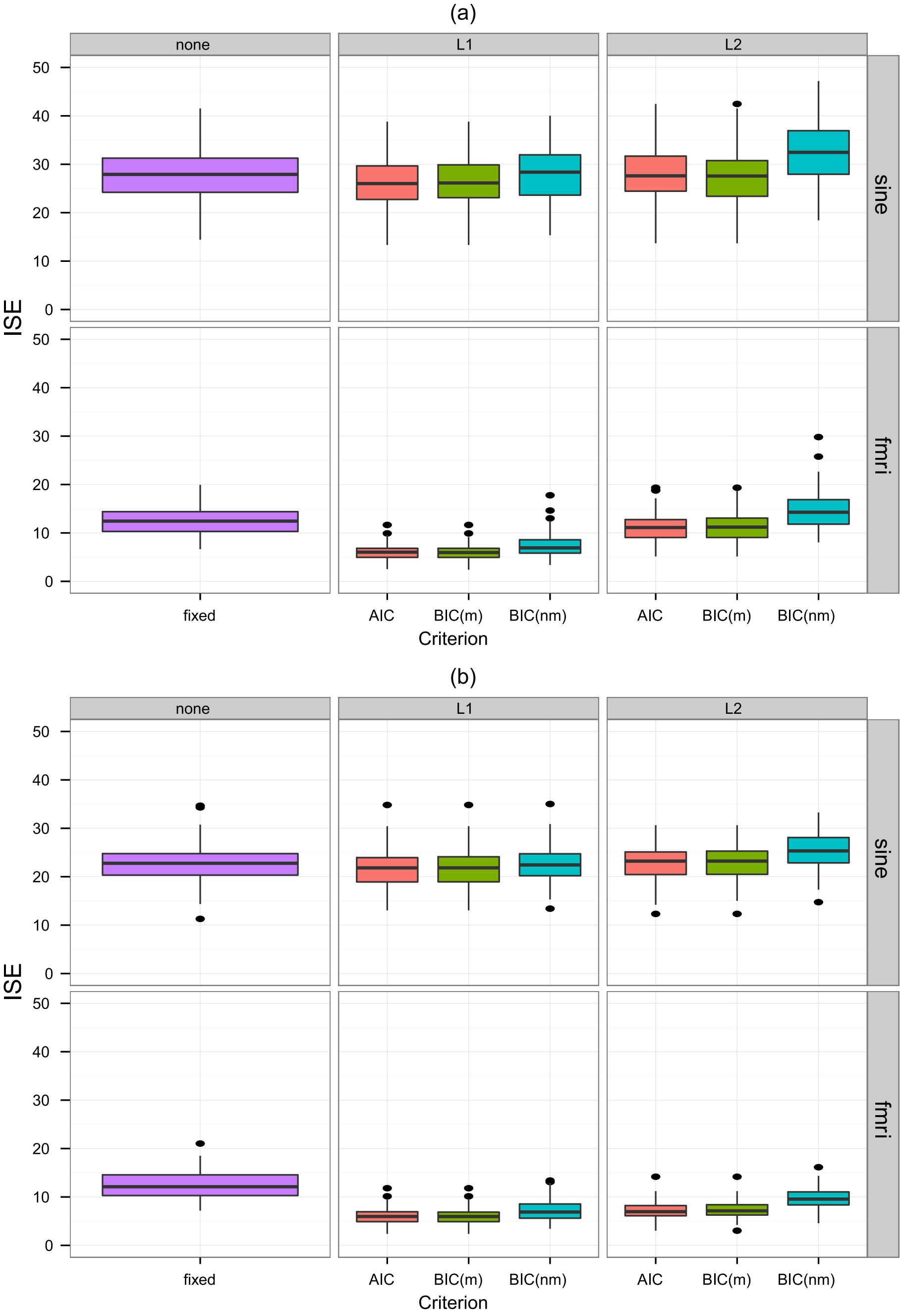

Figure 3. Boxplots of average integrated squared error (ISE) of the estimated curves, compared with no penalty (left panel). (a) $m = 5$, (b) $m = 10$.

[30]. The study had 11 subjects, but 2 were dropped due to head movement during the scans. A block design was used for finger tapping: after an initial period of 20 seconds, the subjects were told to alternate finger tapping (30 seconds) and not finger tapping (30 seconds) for five cycles. The BOLD signal was measured every 2 seconds and the first 4 seconds were dropped from each scan, giving a total of 156 time points for the duration of each scan. Two fMRI sessions were performed for each subject: once for a 'pre-caffeine' session and again after ingested 200 mg of caffeine (the 'post-caffeine' session). During finger tapping periods, some of the voxels in the motor-cortex region of the brain became 'activated', as more oxygen was sent to that part of the brain. Following standard preprocessing, the voxels in the motor cortex for each subject were selected that were activated in both the pre and post caffeine sessions. Figure 4 shows the BOLD signals from the motor cortex of the 9 study subjects, before and after caffeine, where each curve

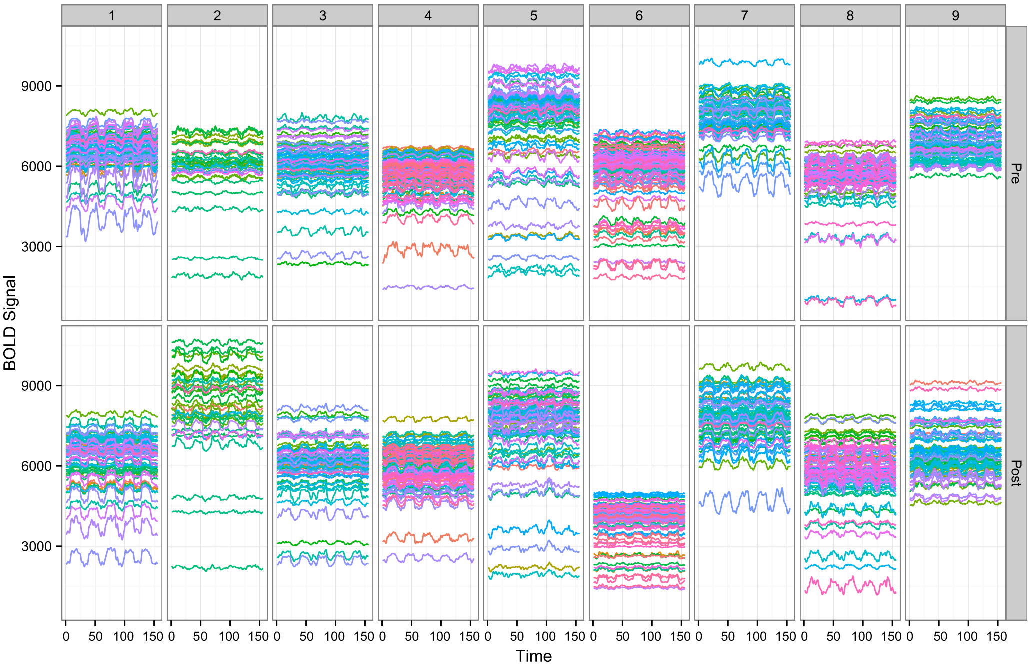

Figure 4. BOLD signals from the motor cortex of 9 subjects, before and after caffeine.

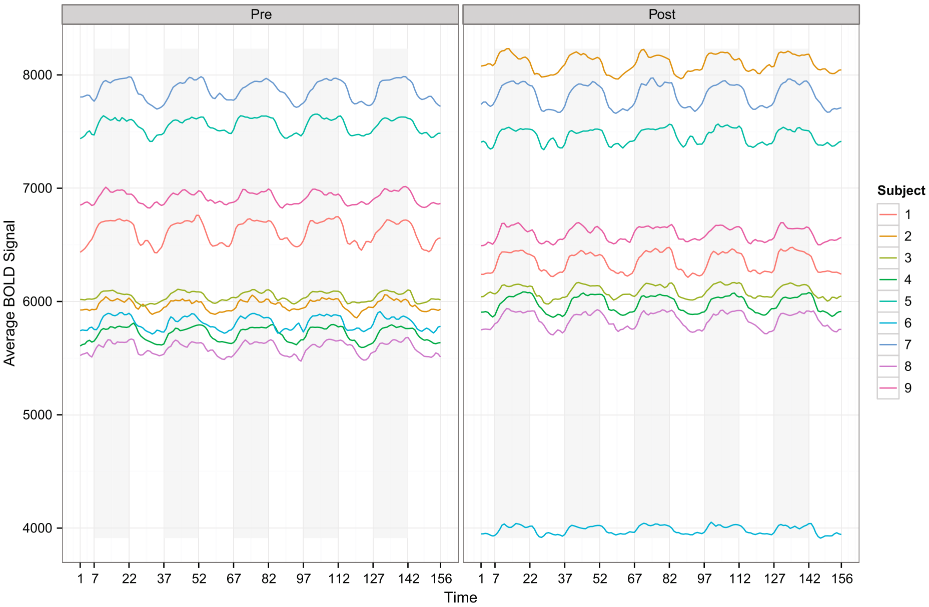

Figure 5. Average BOLD signals for the 9 subjects by session. (Color figure online)

represents a voxel. Numbers of activated voxels per subject ranged from 35 to 124.

4.1 Voxel averaged data analysis

For this analysis, we average over the BOLD signals of activated voxels at each time point for each subject and session. This average is referred to as the 'signal', and is shown in Figure 5. Denote $i = 1, \dots, 9$ subjects, $j = 1, 2$ sessions ('1' for pre-caffeine), and $k = 1, \dots, 156$ time points. Instead of model (1) where the curves are independent, here the two sessions within the same study subject are paired. We assume

$$y_{ijk} = \mu_j(t_k) + b_{ij} + \epsilon_{ijk}, \tag{12}$$

where $(b_{i1}, b_{i2})^\top \sim \mathrm{N}(0, D_1)$ with D_1 an unrestricted 2×2 covariance matrix, counts for the shift (centering) of each

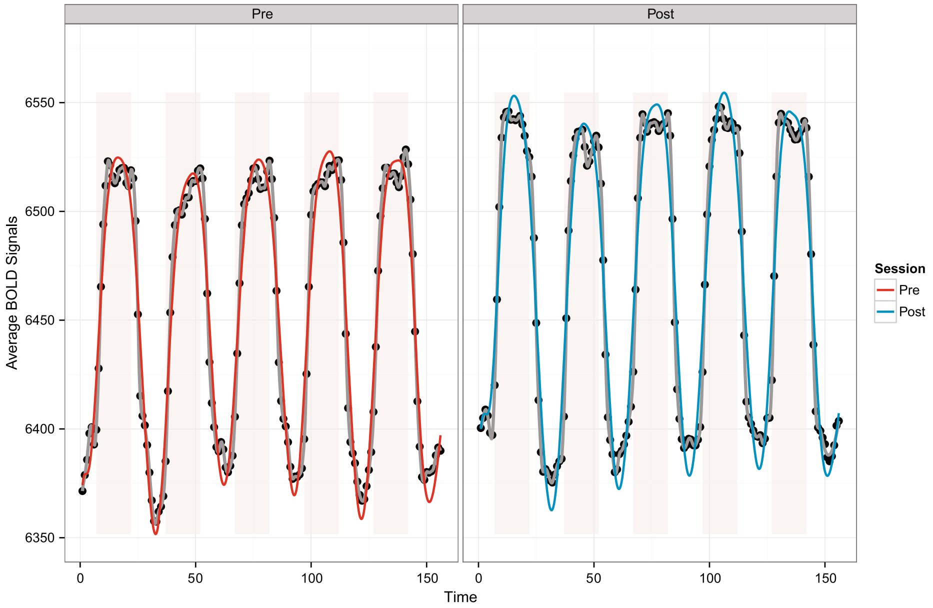

Figure 6. Pre- and post-caffeine L_1 fits. The black dots are the average signals of the 9 subjects, and the grey lines are fits with no penalty.

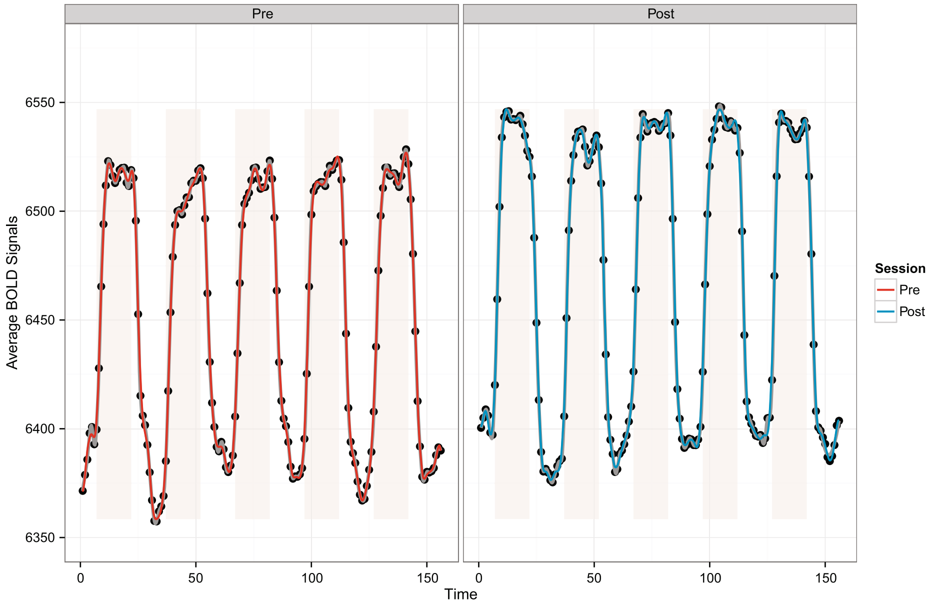

Figure 7. Pre- and post-caffeine L_2 fits. The black dots are the average signals of the 9 subjects, and the grey lines are fits with no penalty.

subject and session, as well as induces the correlation between pre- and post-caffeine sessions within a subject. $(\epsilon_{ij1}, \ldots, \epsilon_{ij156})^\top \sim \mathrm{N}(0, V_1)$ where V_1 has an AR(1) correlation structure, which was used in [30].

Following the approaches described in Section 2, Figure 6 shows the L_1 estimate of the mean activation curves $\mu_j(t)$, and Figure 7 shows the L_2 estimate of $\mu_j(t)$, $j = 1, 2$. As in the simulations, for the L_1 regularization the initial knots were placed at every data point, and for the L_2 regularization the knots were placed at every other data point. AIC was used to choose the penalty parameter λ, since it appears to work well overall relative to the BIC's in our simulations. Consistent with our simulations of last section, the L_2 fit appears less smooth and follows the data closer than the L_1 fit, and is almost indistinguishable from the fits with no penalty (grey lines). Supplement Figure 8 plots the

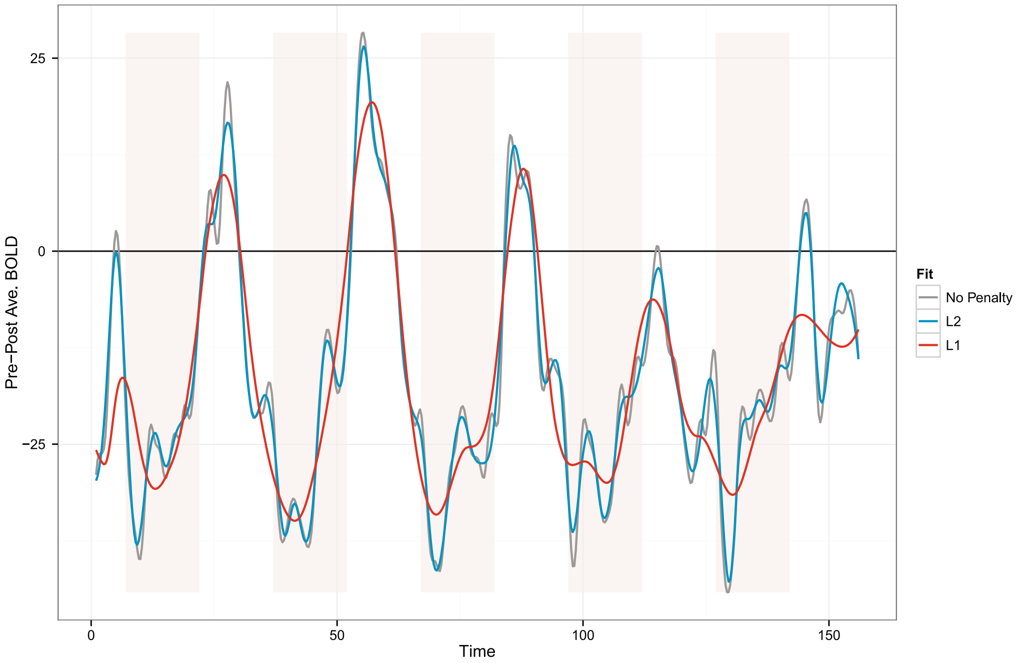

Figure 8. Mean difference $\mu_1(t) - \mu_2(t)$ (red: L_1; blue: L_2 fit; grey: no penalty). The shaded areas are when finger-tapping was prompted. (Color figure online)

individual activation difference between the pre- and post-caffeine sessions $y_{i1k} - y_{i2k}$, super-imposed with the fitted $\mu_1(t_k) + b_{i1} - \mu_2(t_k) - b_{i2}$. The fact that the fitted lines are almost invisible reflects the fact that the individual random effects b_{ij}'s appear to be well estimated, capturing the variation among the subjects.

Finally, Figure 8 shows the estimated difference $\mu_1(t) - \mu_2(t)$ from the L_1 and L_2 fits. Super-imposed on the plot are the shaded intervals when finger tapping was prompted. [30] found significantly shortened time to reach 50% of peak response after ingesting caffeine. The positions of negative dips relative to the finger-tapping prompts, i.e. towards the early part of the interval, also reflect faster activation post-caffeine compared to pre-caffeine. In this way our analysis confirms the findings of [30] which was done using direct comparisons of times to reach 50% of peak responses.

4.2 Voxel level data analysis

For the analysis using all of the data points in Figure 4, let y_{ivjk} be the BOLD signal at time k, during session j, in voxel v for subject i for $i = 1, \ldots, 9$, $j = 1, 2$, $k = 1, \ldots, n$ (=156), and $v = 1, \ldots, n_i$ where n_i the number of activated voxels varies from 35 to 124. We assume the following model:

$$y_{ivjk} = \beta_{ij} 1_{[i \neq 1]} + \mu_j(t_k) + b_{ivj} + \epsilon_{ivjk}. \tag{13}$$

Note that we use fixed effects β_{ij} for the subject and session specific shift of the activation. $\epsilon_{ivj} = (\epsilon_{ivj1}, \ldots, \epsilon_{ivjn})^T$ again has an AR(1) structure, and $b_{iv} = (b_{iv1}, b_{iv2})^\top$ is independently distributed $N(0, D_1)$ for each iv, inducing the correlation between pre- and post-caffaine sessions within a voxel.

To apply the methods described in Section 2 to the voxel level data, which consist of 224,952 data points, we used the San Diego Super Computer (SDSC), mainly to ease the memory problem that made it impossible to fit the model on an imac, for example. Table 1 gives the approximate computing time on various machines, contrasting the voxel averaged versus the voxel level analysis. Unfortunately we were not able to perform the L_1 fit on SDSC, because the 'lars()' function failed to converge. It would be of interest in the future to identify an L_1 algorithm that is suitable for data sets of this size. Supplement Figure 5 shows for three randomly picked voxels per subject and session, the fitted and the observed BOLD signals, where the fitted curves are obtained using L_2 regularization with AIC to choose the penalty parameter λ. The fact that the fitted and the observed curves are hard to distinguish once again shows that the parameters and in particular the random effects are very well estimated. Figure 9 shows the estimated $\mu_j(t)$ ($j = 1, 2$), together with the pointwise 95% confidence intervals obtained using the sandwich estimator of Section 2.2. Supplement Figure 6 shows the estimated difference $\mu_1(t) - \mu_2(t)$ and its corresponding pointwise 95% confidence intervals.

5. DISCUSSION

In this paper, we have explored using existing statistical methodology to analyze group fMRI data arising in clinical settings. Our project was motivated by a clinical study

Table 1. Approximate computation times in statistical environment R

	Analysis	
	Voxel averaged	Voxel level
Number of data points	2,808	224,952
2011 imac (Intel Core i5 2.5 GHz CPU)	2 min	–
Dept server (Intel Core i7 970 3.20 GHz CPU)	2 min	?*
SDSC Triton resource (single comp node Intel Nahalem E5530, 2.4 GHz)	2 min	1.5 days

*The analysis repeatedly ran out of memory when the server had other users.

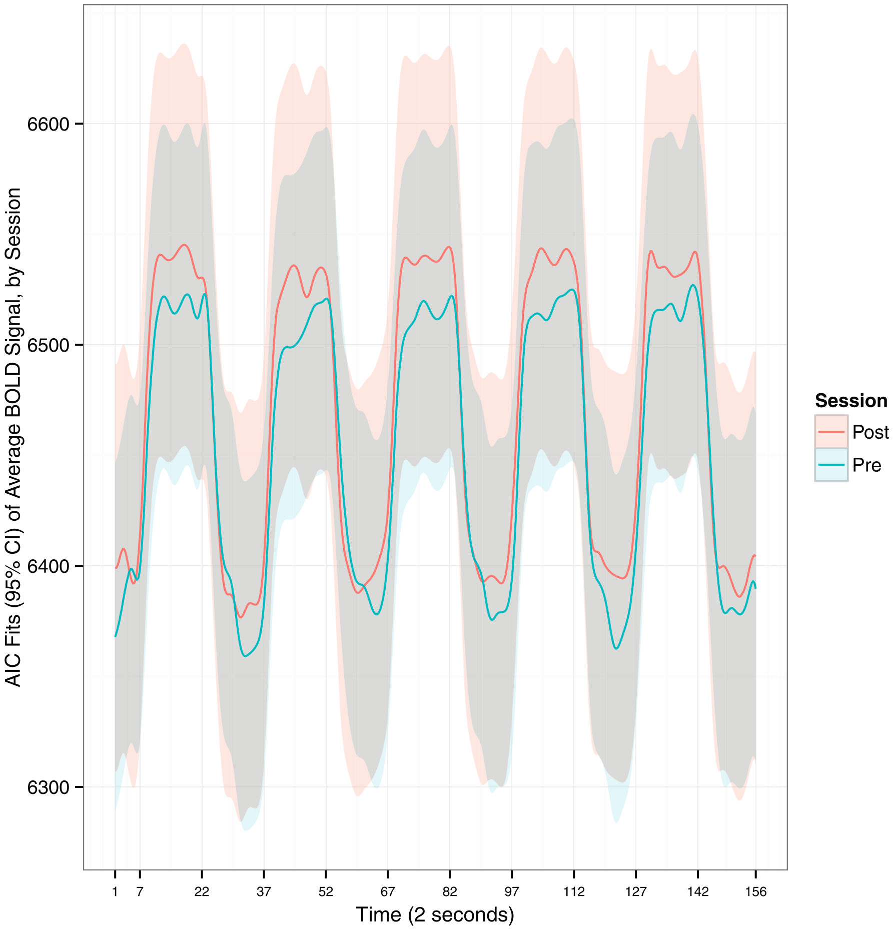

Figure 9. Estimated mean BOLD signal with pointwise 95% confidence intervals (shaded). (Color figure online)

to treat chronic pain and how the different therapies affect the regions of interest in the brain [22]. We use nonparametric estimates of mean activation functions allowing for correlated errors, which is typically the case for fMRI data. We account for the multiple subjects using random effects modeling, which is an alternative approach to the 'percent change' method often used by fMRI researchers. One main advantage of the random effects modeling is that it does not introduce additional between-subject variation in the amplitudes of the activation, which might be of interest in fMRI studies [30]. The fitting algorithms can be relatively easily carried out using existing statistical software.

As seen from our fMRI voxel level data analysis, challenges still exist in today's academic computing environment, perhaps in connection with the use of R. The memory problem appears to be a main one. Cloud computing might be a solution without access to a supercomputer, but additional resources (funding, in our case) is often required in order to use cloud computing. Admittedly the LARS algorithm is most suitable when the sample size is smaller than the number of predictors, which is not the case for the voxel level data. We are working to investigate other numerical methods for L_1 regularization, that will work at least with the memory capacity of the supercomputer.

There is very little theory to guide simultaneous estimation of correlation structure and choice of smoothing parameter. In our simulations we saw that the variance parameters were underestimated. Recently we became aware of the work of [25], which suggested a two-step procedure where one first overfits the unknown mean curve, in order to obtain a consistent estimate of the variance parameters. Another topic that seems to need more theoretical work is simultaneous confidence bands based on nonparametric estimators, in the presence of correlated errors. Work has been done on this problem by, for example, [5, 13], but only under i.i.d. errors.

Another approach to be explored is the elastic net, i.e. combined L_1 and L_2 regularization of [41]. The elastic net is known to enjoy the grouping effect property (i.e. regression coefficients of a group of highly correlated variables tend to be equal) of a strictly convex penalty function such as L_2. The combined penalty was shown to outperform the L_1 penalty but retains its sparsity feature. In the case of splines, the basis functions are highly correlated (Pearson correlation coefficient > 0.9). We hope to report our findings using this approach in the near future.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Thomas Liu for providing the imaging data set. Xu's research was partly supported by the NIH Clinical and Translational Science Award 1UL1RR031980-01 to UC San Diego.

APPENDIX

Since V, Z and D are all block diagonal matrices, Σ is as well. Denote the ith block of Σ, which corresponds to curve i, by Σ_i. Also let $(\sigma_1, ..., \sigma_d)$ be the unknown parameters in Σ. We have for $1 \leq s, k \leq d$

$$
\begin{aligned}
\frac{\partial l}{\partial \beta} &= -\sum_{i=1}^{m} X_i^\top \Sigma_i^{-1}(y_i - X_i\beta),\\
\frac{\partial l}{\partial \sigma_k} &= -\frac{1}{2}\biggl\{ m\mathrm{tr}\left(\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_k}\right) - \sum_{i=1}^{m}(y_i - X_i\beta)^T\Sigma_i^{-1}\\
&\qquad \frac{\partial \Sigma_i}{\partial \sigma_k}\Sigma_i^{-1}(y_i - X_i\beta)\biggr\},\\
\frac{\partial^2 l}{\partial \beta \partial \beta^\top} &= \sum_{i=1}^{m} X_i^\top \Sigma_i^{-1} X_i,\\
\frac{\partial^2 l}{\partial \beta \partial \sigma_k} &= \sum_{i=1}^{m} X_i^\top \Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_k}\Sigma_i^{-1}(y_i - X_i\beta),\\
\frac{\partial^2 l}{\partial \sigma_s \partial \sigma_k} &= -\frac{1}{2}\biggl\{ m\mathrm{tr}\left(-\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_s}\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_k} + \Sigma_i^{-1}\frac{\partial^2 \Sigma_i}{\partial \sigma_s \partial \sigma_k}\right)\\
&\qquad + \sum_{i=1}^{m}(y_i - X_i\beta)^\top \biggl(-\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_s}\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_k}\Sigma_i^{-1}\\
&\qquad + \Sigma_i^{-1}\frac{\partial^2 \Sigma_i}{\partial \sigma_s \partial \sigma_k}\Sigma_i^{-1} - \Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_k}\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_s}\Sigma_i^{-1}\biggr)\\
&\qquad (y_i - X_i\beta)\biggr\}.
\end{aligned}
$$

Received 28 February 2014

REFERENCES

[1] Ashby, F. G. (2011). *Statistical analysis of fMRI data*. The MIT Press.

[2] Bondell, H., Krishna, A., and Ghosh, S. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, **66**, 1069–1077. MR2758494

[3] Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**, 961–983. MR1649194

[4] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer. MR2807761

[5] Cao, G., Yang, L., and Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, **24**, 359–377. MR2921141

[6] Claeskens, G., Krivobokova, T., and Opsomer, J. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, **96**, 529–544. MR2538755

[7] Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121. MR1435485

[8] Eilers, P. H. C. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 637–653.

[9] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360. MR1946581

[10] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, **32**, 928–961. MR2065194

[11] Fan, J. and Zhang, J. (2000). Two-step estimation of functional linear models with application to longitudinal data. *Journal of the Royal Statistical Society, Series B*, **62**, 303–322. MR1749541

[12] Gupta, S. (2009). *A Study of the Asymptotic Properties of Lasso Estimates for Correlated Data*. Ph.D. thesis, Department of Statistics, Florida State University. MR2714079

[13] Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *Annals of Statistics*, **41**, 1892–1921. MR3127852

[14] Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, **92**, 105–118. MR2158613

[15] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. MR2722294

[16] He, X., Shen, L., and Shen, Z. (2001). A data-adaptive knot selection scheme for fitting splines. *IEEE Signal Processing Letters*, **8**, 137–139.

[17] Ibrahim, J., Zhu, H., Garcia, R., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, **67**, 495–503. MR2829018

[18] Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), 487–503. MR2649606

[19] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, **28**(5), 1356–1378. MR1805787

[20] Krivobokova, T. and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, **102**(480), 1328–1337. MR2412553

[21] Lazar, N. (2008). *The Statistical Analysis of Functional MRI Data*. Springer.

[22] Leung, A., Duann, J., McGreevy, K., Li, E., Xu, R., and Donohue, M. (2009). The supraspinal pain pathway of the thermal grill illusion. *NeuroImage*, **47**, S61–S61.

[23] Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, **95**, 415–436. MR2521591

[24] Lindquist, M. A. (2008). The statistical analysis of fmri data. *Statistical Science*, **23**, 439–464. MR2530545

[25] Ma, S. (2012). Two-step spline estimating equations for generalized additive partially linear models with large cluster sizes. *Annals of Statistics*, **40**, 2943–2972. MR3097965

[26] Osborne, M., Presnell, B., and Turlach, B. (1998). Knot selection for regression splines via the lasso. *Computing Science and Statistics*, **30**, 44–49.

[27] Overholser, R. (2013). *Conditional AIC Under General and Generalized Linear Mixed Models and Smoothing in the Presence of Random Effects with Application to fMRI Data Analysis*. Ph.D. thesis, Department of Mathematics, University of California, San Diego. MR3167258

[28] Overholser, R. and Xu, R. (2014). Effective degrees of freedom and its application to conditional AIC for linear mixed-effects models with correlated error structures. *Journal of Multivariate Analysis*, **132**, 160–170. MR3266267

[29] Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge University Press. MR2839490

[30] Rack-Gomer, A. L., Liau, J., and Liu, T. T. (2009). Caffeine reduces resting-state BOLD functional connectivity in the motor cortex. *NeuroImage*, **46**(1), 56–63.

[31] Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York. MR1998720

[32] Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, **38**(2), 197–214. MR2829596

[33] Shen, X. and Ye, J. (2002). Adaptive model selection. *Journal of the American Statistical Association*, **97**, 210–221. MR1947281

[34] Speed, T. P. (1991). Comment on “That BLUP is a good thing: the estimation of random effects” by G.K. Robinson. *Statistical Science*, **6**, 42–44. MR1108815

[35] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B Methodological*, **58**(1), 267–288. MR1379242

[36] Wang, H., Li, G., and Tsai, C. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **69**, 63–78. MR2301500

[37] Wang, X., Shen, J., and Ruppert, D. (2011). On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics*, **5**, 1–17. MR2763795

[38] Wu, H. and Zhang, J. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, **97**, 883–897. MR1941417

[39] Zhang, J. and Chen, J. (2007). Statistical inferences for functional data. *Annals of Statistics*, **35**, 1052–1079. MR2341698

[40] Zhou, S. and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association*, **96**, 247–259. MR1952735

[41] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320. MR2137327

Rosanna Overholser
E-mail address: rosanna.haut@gmail.com

Ronghui Xu
9500 Gilman Drive, MC 0112
La Jolla, CA 92093-0112
USA
E-mail address: rxu@ucsd.edu