UNIVERSITY OF CALIFORNIA, SAN DIEGO

Conditional AIC Under General and Generalized Linear Mixed Models and Smoothing in the Presence of Random Effects with Application to fMRI Data Analysis

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics

by

Rosanna H. Overholser

Committee in charge:

Professor Ronghui (Lily) Xu, Chair
Professor Ian Abramson
Professor Thomas Liu
Professor Dimitris Politis
Professor Rema Raman

2013

The dissertation of Rosanna H. Overholser is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2013

TABLE OF CONTENTS

Signature Page iii

Table of Contents iv

List of Figures vi

List of Tables ix

Acknowledgements xii

Vita xiii

Abstract of the Dissertation xiv

Chapter 1 Conditional Akaike Information Under Generalized Linear Mixed Models 1
1.1 Introduction 1
1.2 Alternative estimation using the bootstrap 3
1.3 Generalized Linear Mixed Models 3
1.4 Simulations 5
1.5 Case Study: the Skin Cancer Prevention Study 7
1.6 Discussion 8
1.7 Proofs 9
1.8 Acknowledgements 13

Chapter 2 Effective Degrees of Freedom and Its Application to Conditional AIC for General Linear Mixed Models with Correlated Error Structures 17
2.1 Introduction 17
2.2 cAIC for General Linear Mixed Model 19
2.2.1 Model and notation 19
2.2.2 Effective degrees of freedom and cAI 20
2.2.3 Conditional AIC 21
2.3 Simulations 22
2.4 An Example 24
2.5 Discussion 25
2.6 Acknowledgements 31

Chapter 3 A Comparison of Smoothing Via L_1 and L_2 Penalization with Application to Group fMRI Data 44
3.1 Introduction 44
3.1.1 L_1 penalization 45

3.1.2 L_2 penalization 46
3.2 Penalized Mixed Effect Models 46
3.2.1 Notation 47
3.2.2 Penalized spline fit of the marginal mean 47
3.2.3 Standard errors 49
3.2.4 Computational aspects 51
3.3 Simulations 52
3.4 fMRI data 55
3.4.1 Model of hemodynamic response 55
3.4.2 Caffeine dataset 55
3.4.3 Model of pre and post caffeine sessions 56
3.4.4 Results 56
3.5 Discussion 61
3.6 Acknowledgements 61

Appendix A fMRI Plots 62

References 85

LIST OF FIGURES

Figure 2.1: Spaghetti plot (left) and group means (right) for the FEV data: solid - drug A, dashed - drug B, dotted - placebo. 35
Figure 2.2: Scatter plot of FEV1 over 8 hours 36

Figure 3.1: True Mean Curves from Simulation 53
Figure 3.2: Estimated (solid) and observed (dots) BOLD signals from three voxels per subject, by session. The estimated curves were obtained via L_2 penalization with smoothing parameter selection by AIC. The fits were obtained using the entire dataset; only the results for three voxels are shown for the purpose of display. 57
Figure 3.3: Estimated average BOLD signal by session (black) with observed average BOLD signals for 9 subjects, by session (colors). The estimated curves were obtained via L_2 penalization with smoothing parameter selection by AIC. 58
Figure 3.4: Pre (red) and Post (blue) caffeine session estimated average BOLD signals with pointwise 95% confidence intervals (shading). The estimated curves were obtained via L_2 penalization with smoothing parameter selection by AIC while the confidence intervals were computed using the sandwich estimator. 59
Figure 3.5: Estimated (black) difference in BOLD signals between sessions with pointwise 95% confidence intervals (shading). The difference is computed as the pre-caffeine signal minus the post-caffeine signal. The estimated curve was obtained via L_2 penalization with smoothing parameter selection by AIC while the confidence intervals were computed with the sandwich estimator. 60

Figure A.1: BOLD signals from the motor cortex of 9 subjects, before and after caffeine. Timeseries of the same color within a subject are from the same voxel. 63
Figure A.2: Randomly selected BOLD signals from the motor cortex of Subject 1, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel. . . . 64
Figure A.3: Randomly selected BOLD signals from the motor cortex of Subject 2, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel. . . . 65

Figure A.4: Randomly selected BOLD signals from the motor cortex of Subject 3, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel. . . . 66
Figure A.5: Randomly selected BOLD signals from the motor cortex of Subject 4, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel. . . . 67
Figure A.6: Randomly selected BOLD signals from the motor cortex of Subject 5, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel. . . . 68
Figure A.7: Randomly selected BOLD signals from the motor cortex of Subject 6, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel. . . . 69
Figure A.8: Randomly selected BOLD signals from the motor cortex of Subject 7, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel. . . . 70
Figure A.9: Randomly selected BOLD signals from the motor cortex of Subject 8, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel. . . . 71
Figure A.10: Randomly selected BOLD signals from the motor cortex of Subject 9, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel. . . . 72
Figure A.11: Percent change in BOLD signals, centered at 100, from the motor cortex of 9 subjects, before and after caffeine. Timeseries of the same color within a subject are from the same voxel. 73
Figure A.12: Percent change in BOLD signals from the motor cortex of Subject 1, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color. 74

Figure A.13: Percent change in BOLD signals from the motor cortex of Subject 2, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color. 75
Figure A.14: Percent change in BOLD signals from the motor cortex of Subject 3, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color. 76
Figure A.15: Percent change in BOLD signals from the motor cortex of Subject 4, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color. 77
Figure A.16: Percent change in BOLD signals from the motor cortex of Subject 5, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color. 78
Figure A.17: Percent change in BOLD signals from the motor cortex of Subject 6, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color. 79
Figure A.18: Percent change in BOLD signals from the motor cortex of Subject 7, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color. 80
Figure A.19: Percent change in BOLD signals from the motor cortex of Subject 8, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color. 81
Figure A.20: Percent change in BOLD signals from the motor cortex of Subject 9, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color. 82
Figure A.21: Average BOLD signals from the motor cortices of 9 subjects, before and after caffeine. 83
Figure A.22: Average BOLD signals from the motor cortices of 9 subjects, before and after caffeine. The top two plots show average BOLD signals for the pre (left) and post (right) caffeine sessions while the lower two plots display these signals as percent changes, centered at 100. Light gray shading indicates periods of fingertapping. 84

LIST OF TABLES

Table 1.1: Comparison of model selection procedures based on simulations from a Poisson generalized linear mixed model with log link. Data are generated from model (3,2): covariates x_1, x_2, x_3 are independent Bernoulli (0.5), $\beta = (1,1,1)^\top$, $z_1 = x_1, z_2 = x_2$, and $b_i \sim N(0, 0.25I_2)$. Model (2,1) includes x_1, x_2 and z_1; model (3,1) includes also x_3, and (3,3) includes also z_2 and $z_3 = x_3$. Average over 100 simulations are reported for cAI, cAIC, cAIC_b with 400 bootstrap samples, and $-2l(y \mid \hat{\beta}, \hat{b})$; a/b gives the number of times out of 100 a model was chosen using the rule of two (a) or simple minimum (b). 14

Table 1.2: Comparison of model selection procedures based on simulations from a Poisson mixed model with log link. We enumerate the candidate models according the number of fixed and random effects included. Data is generated from model (3,2) with three independent Bernoulli ($p = 0.5$) fixed effect covariates x_1, x_2, x_3, and two random effect covariates $z_1 = x_1, z_2 = x_2$. The fixed effect is $\beta = (1,1,1)^{\text{T}}$ and the random effects are $b_i \sim N(0, 0.25I_2)$. Model (2,1) includes x_1, x_2 and $z_1 = x_1$; model (3,1) includes x_1, x_2, x_3 and $z_1 = x_1$; and (3,3) includes x_1, x_2, x_3 and $z_1 = x_1, z_2 = x_2, z_3 = x_3$. Average (SE) over 100 simulations are reported for cAI, cAIC, cAIC_b from 400 bootstrap samples, and $-2l(y \mid \hat{\beta}, \hat{b})$. Below each average and SE are a/b: the number of times the model was chosen using the rule of two (a) or simple minimum (b). 15

Table 1.3: Skin cancer prevention study. Estimates of fixed effects (SE) and variance components, and cAIC from 5 Poisson mixed models differing in the fixed and random effects for year of follow-up. Age and number of skin cancers at baseline are continuous. The best models, marked with a *, cannot be ranked based on cAIC. 16

Table 2.1: Models used in simulation Tables 2.2 to 2.4. Model 4 is the true model that generated the data. 31

Table 2.2: Simulation Results for m=10, n=25. The models correspond to those in Table 2.1; the format '(p, q, r)' denotes the numbers of fixed and random effects, and the number of parameters in the error variance. (3,1,2) is the true model, and ρ is the correlation parameter under the AR(1) error. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details. . 32

Table 2.3: Simulation Results for m=20, n=25.The models correspond to those in Table 2.1; the format '(p, q, r)' denotes the numbers of fixed and random effects, and the number of parameters in the error variance. (3,1,2) is the true model, and ρ is the correlation parameter under the AR(1) error. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details. . 33

Table 2.4: Simulation Results for m=20, n=50. The models correspond to those in Table 2.1; the format '(p, q, r)' denotes the numbers of fixed and random effects, and the number of parameters in the error variance. (3,1,2) is the true model, and ρ is the correlation parameter under the AR(1) error. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details. . 34

Table 2.5: Simulation results comparing different error correlation structures, for m=10, n=10. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details. 37

Table 2.6: Simulation results comparing different error correlation structures, for m=25, n=10. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details. 38

Table 2.7: Simulation results comparing different error correlation structures, for m=75, n=10. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details. 39

Table 2.8: Simulation results comparing different error correlation structures, for m=10, n=25. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details. 40

Table 2.9: Simulation results comparing different error correlation structures, for m=20, n=25. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details. 41

Table 2.10: Simulation results comparing different error correlation structures, for m=20, n=50. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details. 42
Table 2.11: Parameters estimates (standard errors) and model selection for the FEV_1 data under three different error structures. 43

Table 3.1: Within group correlation $\rho = 0.4$ 53
Table 3.2: Error variance $\sigma^2 = 0.09$ 54
Table 3.3: MISE (se) = mean integrated square error 54

Table A.1: Number of motor cortex voxels per subject. 62

ACKNOWLEDGEMENTS

The contents of Chapter 1 have been reproduced from the published manuscript of "Conditional Akaike information under generalized linear and proportional hazards mixed models" by Michael Donohue, Rosanna Overholser, Ronghui Xu and Florin Vaida, which appeared in *Biometrika*, 2011; **98**(3): 685-700. Additional simulation results have been added and results on proportional hazards mixed models omitted. Oxford Press, publisher of Biometrika, is acknowledged for allowing me to reproduce published material.

The contents of Chapter 2 have been reproduced from the submitted manuscript of "Effective Degrees of Freedom and Its Application to Conditional AIC for General Linear Mixed Models with Correlated Error Structures" by Rosanna Overholser and Ronghui Xu. Additional simulation results have been added.

The contents of Chapter 3 are being prepared for submission as "A Comparison of L_1 and L_2 Smoothing with Application to Group fMRI Data" by Rosanna Overholser and Ronghui Xu. I thank Thomas Liu and Anna Leigh Rack-Gomer of the UCSD Center for fMRI for generously sharing an fMRI dataset.

I thank my advisor, Ronghui Xu, for her patient guidance during my studies and push to do better.

I am grateful to my husband Peter for his support and encouragement during my studies.

VITA

2006	B. S. in Mathematics, California Polytechnic State University, San Luis Obispo
2008	M. A. in Mathematics, University of California, San Diego
2013	Ph. D. in Mathematics, University of California, San Diego

PUBLICATIONS

Donohue MC, Overholser, R, Xu R, Vaida F. "Conditional Akaike information under generalized linear and proportional hazards mixed models". *Biometrika*, 2011; **98**(3): 685-700.

ABSTRACT OF THE DISSERTATION

Conditional AIC Under General and Generalized Linear Mixed Models and Smoothing in the Presence of Random Effects with Application to fMRI Data Analysis

by

Rosanna H. Overholser

Doctor of Philosophy in Mathematics

University of California, San Diego, 2013

Professor Ronghui (Lily) Xu, Chair

This thesis examines model selection for clustered data. Such data are often modeled using random effects. Conditional Akaike information, when cluster specific inference is desired, was proposed in Vaida and Blanchard (2005) and used to derive a corresponding model selection criterion under linear mixed models. We extend the approach to general and generalized linear mixed models. Exact calculations are not available outside linear mixed models so we resort to asymptotic approximations.

We show that under general linear mixed models with correlated errors, the number of effective degrees of freedom is equal to the trace of the usual 'hat' matrix plus the number of parameters in the error covariance matrix. Using it one can define a

crude version of the conditional AIC (cAIC), which is known to be inaccurate due to the estimation of unknown variance parameters. We show however, that a simple 'rule-of-thumb' correction performs nearly as well as an asymptotically unbiased cAIC counting for the unknown parameters, one which is difficult to compute without specific programming for each case of the error correlation structure. For generalized linear mixed models, we consider a bootstrap method in addition to the rule-of-thumb.

Finally, we investigate non-parametric estimation of a mean when data are clustered. We consider smoothing via splines with either L_1 or L_2 penalization. These models may be written as penalized general linear mixed models, thus allowing the use of existing software. We apply our methods to functional MRI time courses from multiple subjects.

Chapter 1

Conditional Akaike Information Under Generalized Linear Mixed Models

1.1 Introduction

Mixed effects models have been widely used to analyze clustered data, which arise in applications such as longitudinal studies, familial studies, and multi-center clinical trials. The focus of inference under such models can be on the population parameters, such as the fixed regression effects, the variance parameters, or on the cluster level parameters, often represented by the random effects themselves. As an example of the latter, in a multi-center clinical trial where the treatment effect is found to be heterogeneous among the trial centers, it is of scientific interest to estimate the treatment effects from individual centers, and to investigate the cause of the treatment differences. Other examples of cluster level focus occur in ecology, small-area estimation, and animal husbandry.

Specifically, assume that the data consist of outcomes from m clusters, with n_i observations in cluster i, $i = 1, \ldots, m$. Within a cluster the outcomes are dependent, but conditional on the cluster-specific $d \times 1$ vector of random effects b_i, the outcomes y_{ij} are independent and follow a generalized linear mixed model with mean

$$\mu_{ij} = E(y_{ij} \mid \beta, b_i) = g^{-1}(\beta^\top x_{ij} + b_i^\top z_{ij}), \tag{1.1}$$

where x_{ij} and z_{ij} are the covariate vectors for the fixed effects β and the random effects

b_i of cluster i, $b_i \sim p(b_i)$, and g is the link function. For cluster level inference, any future prediction takes place in the same clusters as the observed data, and the random effects for these clusters are held constant [66]. More specifically, let y^0 be independently replicated outcomes from the same conditional distribution as the original data y given the same random effects b. Here y, y^0 and b are random vectors consisting of elements y_{ij}, y^0_{ij} and b_i, respectively. For model selection purposes [66] defined the conditional Akaike information,

$$\text{cAI} = -2E_{(y,b)}E_{y^0|b}[l\{y^0 \mid \hat{\beta}(y), \hat{b}(y)\}], \tag{1.2}$$

where $l(\cdot \mid \cdot)$ is the conditional log-likelihood given the random effects, and $\hat{\beta}(y)$, $\hat{b}(y)$ are estimators of β and b based on the data y, for example, the maximum likelihood and the empirical Bayes estimators. The expectations are taken with respect to the true model generating the data. They proceeded to show that for the linear mixed model with known variance components, an unbiased estimator of (2.3) is

$$\text{cAIC} = -2l\{y \mid \hat{\beta}(y), \hat{b}(y)\} + 2\rho\,, \tag{1.3}$$

where the bias correction factor ρ is the effective degrees of freedom of the linear mixed model of [33] and [76]. Expression (2.4) is referred to as the conditional Akaike Information criterion. [66] and [48] give formulas for ρ in the more general case of unknown variance parameters in finite samples. The theory of the Akaike information criterion and its basis in model prediction are well understood in the literature, see, e.g., [1, 49, 7]. In [15], the conditional Akaike information and its criterion under generalized linear and proportional hazards mixed models are developed; we reproduce the results for generalized linear mixed models. Here, exact calculation is not available outside normal linear mixed models, and asymptotic approximations are necessary.

Generalized linear mixed models have been studied for the past two decades. For recent monographs see [37, 52, 64].

From a different perspective and not specifically for clustered data, the issue of focus of model selection was addressed in [11] and [32]. For an overview of the Akaike information criterion see the recent monographs of [7] and [12].

1.2 Alternative estimation using the bootstrap

The bootstrap has been used in the estimation of prediction errors [16, 17, 72] and for Akaike-type criteria [61], and has shown less bias in finite samples [8, 55, 59]. Our proposed estimate is

$$\mathrm{cAIC}_b = -2l(y \mid \hat{\beta}, \hat{b}) + 2\rho_b. \tag{1.4}$$

The correction factor ρ_b is given by

$$\rho_b = E^*\{l(y^* \mid \hat{\beta}^*, \hat{b}^*) - l(y \mid \hat{\beta}^*, \hat{b}^*)\}, \tag{1.5}$$

where E^* denotes the bootstrap expectation, i.e. average over the bootstrap datasets; the bootstrap datasets y^* are obtained by resampling first the clusters, then the observations within cluster, and $(\hat{\beta}^*, \hat{b}^*) = (\hat{\beta}(y^*), \hat{b}(y^*))$. For each cluster in y^*, the data y from the same cluster is used in calculating ρ_b. For applications with extremely small clusters, where some clusters have only one observation, resampling within the clusters might be infeasible. In this case parametric or model-based bootstrap can be used, where the bootstrap data are generated under a fitted large model, with estimated fixed and random effects. The formula (1.5) is derived similarly to [73], but adjusted to our conditional setting.

1.3 Generalized Linear Mixed Models

Consider the model given by (1.1). To set the notation, write $D^{-1} = -\partial^2 \log p(b)/\partial b \partial b'$, where $p(b) = \prod_{i=1}^m p(b_i)$ is the distribution of the independent random effects; when $b_i \sim N(0, \Sigma)$, $D = \mathrm{var}(b) = \mathrm{diag}_m(\Sigma)$, the block-diagonal matrix with m blocks equal to Σ. Let X_i and Z_i be the matrices with rows $x_{ij}^\top$ and $z_{ij}^\top$, and let $X = \mathrm{stack}(X_1, \ldots, X_m)$ and $Z = \mathrm{diag}(Z_1, \ldots, Z_m)$ be the $N \times p$ and $N \times q$ model matrices, where p , d and $q = md$ are the lengths of β, b_i and b, $N = n_1 + \cdots + n_m$, and the stack function stacks matrices on top of each other. Further, let $w_i = (w_{i1}, \ldots, w_{in_i})^\top$ be the vector of weights given by $w_{ij} = [\mathrm{var}(y_{ij} \mid b_i)\{g'(\mu_{ij})\}^2]^{-1}$, and $W = \mathrm{diag}(w_1, \ldots, w_m)$. The usual estimator for β is the maxi-

mizer of the marginal likelihood, $L(\beta) = \int \exp\{l_J(y, b \mid \beta)\}\, db$, where

$$l_J(y, b \mid \beta) = l(y \mid \beta, b) + \log p(b) \tag{1.6}$$

is the joint log-likelihood. Alternatively, (β, b) are estimated jointly as the maximizer of the joint log-likelihood l_J. The joint log-likelihood (1.6) and its maximizer $(\hat{\beta}, \hat{b})$ have been variously justified. [4], [71] and [67] show that under suitable conditions $\hat{\beta}$ is a first-order Laplace approximation to the maximum likelihood estimator. Given the fixed effects, the joint likelihood is proportional to the posterior distribution of the random effects, maximized by $\hat{b}$ [38]. [44] call (1.6) the hierarchical log-likelihood, and consider it as a basis of inference; see also [45]. In the smoothing literature l_J is seen as a penalized log-likelihood [see, e.g., 68]. The variance matrix Σ that is suppressed in $p(b)$ is estimated by maximum likelihood or residual maximum likelihood [4].

Let $U = (X, Z)$ and $\theta = \text{stack}(\beta, b)$, so that $U\theta = X\beta + Zb$. Let $s_J = \partial l_J / \partial \theta$ be the score function for the joint log-likelihood. Standard derivations show that

$$G = \text{var}\{s_J(y) \mid \theta\} = U^\top W U, \tag{1.7}$$

$$\Omega = E\{s_J(y) s_J^\top(y) \mid \theta\} = E\{-\partial^2 l_J(y)/\partial\theta\partial\theta^\top \mid \theta\} = U^\top W U + \text{diag}(0, D^{-1}) \tag{1.8}$$

Further, let

$$\rho = \text{tr}(G\,\Omega^{-1}) = \text{tr}[U^\top W U\, \{U^\top W U + \text{diag}(0, D^{-1})\}^{-1}]. \tag{1.9}$$

For the linear mixed model, $W = I$. In this case ρ is the effective degrees of freedom of [33], as well as the correction factor for cAIC (2.4) in Theorem 1 of [66]. [51] use a form similar to ρ as the effective degrees of freedom for the generalized linear mixed models. The following main result shows that ρ is asymptotically the relevant correction factor for the cAIC (2.4).

Theorem 1. *Assume that the data y are generated from the generalized linear mixed model (1.1). Let $\hat{\beta}$ be the maximum likelihood estimator, and $\hat{b}$ the maximizer of the joint likelihood given $\hat{\beta}$ and the maximum likelihood estimate of Σ. Under conditions A1-A11 given in Section 1.7, an asymptotically unbiased estimator of the conditional Akaike information (2.3) is given by the cAIC (2.4), with $\rho = tr(G\Omega^{-1})$ as in (1.9). That is, $E(\text{cAIC}) = \text{cAI} + o(1)$ for large m and n_i's.*

In addition, the effective degrees of freedom ρ satisfies $p \le \rho \le p + q$.

The proofs for this and for the following results are given in Section 1.7.

In practice, W is computed at $\theta = \hat{\theta}$. Using formulas for explicit computation of the inverse matrix in (1.9) [29, p.99] we get

$$\rho = (p+q) - \text{tr}[\{Z^\top WZ - Z^\top WX(X^\top WX)^{-1}X^\top WZ + D^{-1}\}^{-1}D^{-1}]. \quad (1.10)$$

Formulas (1.7) and (1.8) are a form of the Bartlett identities for the joint likelihood. The more general form is given below.

Proposition 1 (Bartlett identities for joint or penalized likelihood)**.** *Assume that the data y have log-likelihood $l(y \mid \theta)$, satisfying the standard regularity conditions which ensure differentiation with respect to parameter θ under the integral sign, as well as the first two Bartlett identities: $E\{s(y \mid \theta)\} = 0$ and $E\{-\ddot{l}(y \mid \theta)\} = E\{s(y \mid \theta)s(y \mid \theta)^\top\}$, where s and $\ddot{l}$ denote the first two derivatives of l with respect to θ. Further, assume that $p(\theta)$ is a non-negative function of θ not depending on y, twice differentiable with respect to θ and with continuous second derivatives. Let $l_J(y \mid \theta) = l(y \mid \theta) + \log p(\theta)$. Let s_J and $\ddot{l}_J$ be the first two derivatives of l_J. Then l_J satisfies the modified Bartlett identities:*

$$\begin{aligned} E(s_J \mid \theta) &= \partial \log p(\theta)/\partial\theta, \\ \text{var}(s_J \mid \theta) &= \text{var}(s \mid \theta), \\ E(-\ddot{l}_J \mid \theta) &= E(-\ddot{l} \mid \theta) - \partial^2 \log p(\theta)/\partial\theta\partial\theta^\top = \text{var}(s_J \mid \theta) - \partial^2 \log p(\theta)/\partial\theta\partial\theta^\top. \end{aligned}$$

In particular, if $\theta = (\beta, b)$ and $p(\theta) = p(b)$ is the $N(0, D)$ density, then $\partial \log p(\theta)/\partial\theta = A\theta$ and $-\partial \log p(\theta)/\partial\theta\partial\theta^\top = A = \text{diag}(0, D^{-1})$.

The proof follows directly from the definition of l_J and the Bartlett identities for l.

1.4 Simulations

We carried out a simulation experiment to evaluate the proposed criteria under the generalized linear mixed models. The emphasis is two-fold: on the criteria as estimators of the underlying Akaike information as well as on their success in selecting the correct model. We computed the conditional Akaike information by simulation, and

its criteria with correction factors ρ and ρ_b, The results are reported in Table 1.1. The numbers of fixed and random effects in each model are indicated as a pair at the top of each column; for example, (2,1) indicates that two fixed effects and one random effect are included in the model. The true model is indicated in bold. In each case we used a combination of small and large numbers of clusters m and observations within cluster n_i. We used two model selection rules: (1) the rule of two in which one selects the smallest model whose criterion value is within 2 of the minimum criterion value; or (2) select the model with the minimum criterion value. The rule of two acknowledges the variation in a criterion as an estimate of the underlying information, so that for models with close criterion values there is not enough evidence for a preference; in this situation a parsimony principle is applied. The estimation used the `lme4` package in `R`.

Table 1.1 reports the simulation under a log-link Poisson generalized linear mixed model. Overall cAIC provides a good estimate of cAI, within the statistical error range. In the first scenario of 10 clusters of 5 observations each, although the average of cAIC is minimized at the larger model (3,3), it is not significantly different from that of the true model (3,2). The rule of two chooses the correct model most often, and chooses the larger model (3.3) between 8% and 30% of the cases. Note, however, that the models (3,2) and (3,3) are very close to each other based on the cAI, to start with. The nonparametric bootstrap works well when the cluster sizes are reasonably large, e.g. $n_i \geq 10$, and when the model is not too far from the truth. While the true model is (3,2), under model (2,1) cAIC_b is typically more than twice the standard error away from cAI except when $n_i = 40$. Such inaccuracy seems to be attributable to the model fitting procedure by `lmer` and the high probability of duplicated data with small cluster resampling in the bootstrap.

In generating Table 1.1, we used a nested bootstrap in which we resample clusters, then resample observations within those resampled clusters. In addition, we also considered an alternative bootstrap approach in which clusters are not resampled. This approach is motivated by the rationale that the estimation of the inner expectation, conditional on the cluster-specific data and random effects (y, b), is crucial for estimating cAI. This alternative approach averages the estimated risk for each cluster, while the cAI is the expected risk over the distribution of the data including the distribution of the

clusters. Table 1.2 here appears to have more discrepancy from Table 1.1 in the paper, especially in the number of times the true model is selected. We think that it is caused by the PQL option used in the `lmer` function here, as compared to the Laplace option used previously.

1.5 Case Study: the Skin Cancer Prevention Study

The Skin Cancer Prevention Study was a randomized, double-blinded, placebo-controlled clinical trial of non-melanoma skin cancer prevention in high risk subjects [24, 23]; 1805 subjects were randomized to either 50mg of beta-carotene daily or placebo, for up to five years. The dataset consisted of the $m = 1683$ subjects with complete covariate information, with $N = 7081$ observations. We fitted Poisson mixed models with log link for the main outcome, the number of new skin cancers y_{ij} for subject i in year j. The covariates included skin type and gender as binary variables, and age at entry and number of skin cancers at entry as continuous variables. The year of follow-up was either omitted, or included a linear or quadratic effect. In some models a subject-specific year effect was fitted. All models included a subject-specific random intercept. The treatment effect was proven not significant in earlier analyses and was not included. The results in Table 1.3 show that the random year effect should not be included in the model; the three models with the year omitted, in linear, or quadratic form yield comparable conditional Akaike criteria and cannot be distinguished. On parsimony grounds, the model without year effect can be chosen. To determine whether the difference of conditional Akaike information between model was significant, a 95% confidence interval of this difference was computed by bootstrap for each pair of the models. The 95% confidence intervals for this difference were as follows: (5,1) versus (6,1) $= (-2.19, 18.55)$; (6,1) versus (7,1) $= (-1.96, 22.54)$; (5,1) versus (7,1) $= (-3.74, 28.61)$. This analysis confirms that the three models cannot be ranked on conditional Akaike information criterion alone, and that the simpler (5,1) model may be chosen.

1.6 Discussion

The analytic derivation and resampling method used in this paper follow a general approach that can be applied to other models, such as nonlinear mixed models. More importantly, the results apply to general distributions for the random effects, although in practice the normal distribution is often used. Our setting is of independent clusters. This is not the most general model for random effects. Much of the theory applies to the general setting. However, care is needed in establishing the asymptotic results, since they require adequate convergence for the random effects.

For generalized linear hierarchical models with only random intercepts, the ρ in (1.9) turns out to be equal to the effective degrees of freedom obtained in [51] equation (11), using their quasi-exact method; such a connection was in fact conjectured in their paper. In addition [63] gave an approximation to their Bayesian measure of model complexity, which is the same as (1.9). See also Ruppert et al. (2003, chapters 8 and 11). In the case of linear mixed models, [14] show that the part of ρ due to the random effects can be interpreted as the ratio of the random effects variance to the total variance. Our derivation provides a theoretical basis both for the information criteria and for the model degrees of freedom under generalized linear, that is, as an approximately unbiased estimate of the cAI defined in (2.3).

Although the bootstrap has been applied to risk estimation and has been shown to have good finite sample performance for independent and identically distributed data, in our investigation it did not substantially out-perform the analytic approximation. This may be because the cluster sizes are typically not large, and the resampling is done within the clusters. Given the wide range of possible implementations allowed by the bootstrap, further improvements are possible, and the topic deserves further exploration. As an alternative to the bootstrap, the numeric methods of [48] may be extended to our setting for evaluating ρ.

[22] recently established the asymptotic equivalence of cross-validation and AIC under linear mixed-models. It is argued that leave-one-cluster-out cross-validation is asymptotically equivalent to marginal AIC and leave-one-observation-out cross- validation is asymptotically equivalent to cAIC. In our simulations we used a nested bootstrap in which we resample clusters, then resample observations within those resampled clus-

ters. We also considered an alternative bootstrap approach in which clusters are not resampled, which was generally less effective at selecting the true model and unbiased estimation of cAI.

One limitation of the Akaike information is that it is model-inconsistent. That is, even in large samples it can select a wrong model with non-zero probability which is usually too large in dimension. This is the case when there is a fixed true model. It is now understood [60] that for linear model selection, various procedures, including the Akaike information criterion and the Bayesian information criterion, fall into three classes: valid if there exist fixed-dimension correct models, valid if no fixed-dimension correct model exists, or a compromise of the above two. Our work here on the conditional Akaike information, adjusted to mixed models, does not address these issues. When the more classical case is concerned where there is a fixed true model, a possible remedy is to include considerations of parsimony. One would choose the smaller model, unless the larger one has an Akaike information criterion that is significantly better. In another recent paper, [25] present a theorem suggesting that cAIC is almost surely biased in favor of models with additional random effects. Our simulations suggest, however, that cAIC does not always favor larger models, which should help allay concern. Our proposed rule of two also offers some protection against large model bias.

1.7 Proofs

Setup and conditions for Theorem 1

We assume the following. Given the number of clusters m, let $\theta_m = \text{stack}(\beta, b_1, \ldots, b_m)$, and let $\theta = \text{stack}(\beta, b_1, b_2, \ldots)$ be the corresponding infinite-dimensional parameter as m increases; θ_m contains the first $m+1$ elements, or $p+q$ components, of θ. The true value of θ is θ_0, with first $m+1$ elements $\theta_{0m} = \text{stack}(\beta_0, b_{01}, \ldots, b_{0m})$. For fixed m and cluster size $n = n_1 = \ldots = n_m$, the data y is generated from model (1.1), with parameter $\theta_m = \theta_{m0}$. Further, θ_m is estimated by $\hat{\theta}_{nm}$, the maximizer of the joint likelihood $l_J(y \mid \theta_m)$. Note that $N = mn$.

[53] partitions β into $\beta = (\beta_1, \beta_2)$, where the covariates x_{ij1} of β_1 do not have random effects, and the covariates x_{ij2} of β_2 have random effects, i.e., $x_{ij2} = z_{ij}$. [53]

shows that the maximum likelihood estimates of these two components have different rates of convergence as $m, n \to \infty$. For simplicity we will assume that $\beta = \beta_1$; the more general case follows with straightforward modifications.

In the following, unless explicitly stated, all expectations are conditional on θ, and therefore on the random effects b. Let $\Delta(\theta_m) = -E\{l_J(\theta_m; y)\}$, and $\hat{\Delta}(\theta_m) = -l_J(\theta_m; y)$. We do not include indices m, n for Δ and $\hat{\Delta}$ unless necessary. Let Δ', Δ'' denote the derivative and the Hessian of Δ with respect to θ, with a similar notation for $\hat{\Delta}$. Write $l_J = \sum_{i=1}^{m} l_{Ji}$, where l_{Ji} is the component for the ith cluster. From (1.8) we have that $\Delta'' = U^\top W U + \text{diag}(0, D^{-1})$. Let $\Delta''_{\beta\beta}$, $\Delta''_{\beta b_i}$, $\Delta''_{b_i b_j}$ be the corresponding matrix blocks from the Hessian matrix Δ'', and similarly for $\hat{\Delta}''$. We assume that the following conditions hold:

A1. The true parameter θ_0 is unique, and is in the interior of a convex closed bounded set $\Theta \subset \mathcal{R}^\infty$ equipped with the sup norm.

A2. The fixed effects component $\hat{\beta}$ of θ_{mn} satisfies $\hat{\beta} \to \beta_0$ almost surely as $m, n \to \infty$.

A3. The random effects components $\hat{b}_1, \ldots, \hat{b}_m$ of θ_{mn} satisfy $\max_{i=1,\ldots,m} ||\hat{b}_i - b_{0i}|| \to 0$ almost surely as $m, n \to \infty$.

A4. For any m, n, the first and second derivatives $\Delta', \hat{\Delta}'$ and $\Delta'', \hat{\Delta}''$ exist, and are continuous on Θ.

A5. The ratio $n/m \to \infty$.

A6. As $m, n \to \infty$, $\{\hat{\Delta}_{\beta\beta}(\theta)'' - \Delta_{\beta\beta}(\theta)''\}/N$, $\sum_{i=1}^{m}\{\hat{\Delta}_{b_i b_i}(\theta)'' - \Delta_{b_i b_i}(\theta)''\}/n$, and $\sum_{i=1}^{m}\{\hat{\Delta}_{\beta b_i}(\theta)'' - \Delta_{\beta b_i}(\theta)''\}/(nm^{1/2})$ converge almost surely to 0 uniformly on Θ.

A7. As $m, n \to \infty$, $N^{1/2}(\hat{\beta} - \beta_0) \to N(0, v_1)$ in distribution, and $N||\hat{\beta} - \beta_0||_2^2$ is uniformly integrable.

A8. As $n \to \infty$, $n^{1/2}(\hat{b}_i - b_{i0}) \to N(0, v_{bi})$ in distribution uniformly over i, and $n||\hat{b}_i - b_{0i}||_2^2$ is uniformly integrable for all i.

A9. The quantity $\hat{\Delta}''_{\beta\beta}/N$ is bounded for all θ, m and n, and $\lim_{m,n\to\infty} \hat{\Delta}''_{\beta\beta}/N$ is positive definite.

A10. The quantity $\hat{\Delta}''_{b_i b_i}/n$ is bounded for all θ, m and n, and $\lim_{n\to\infty} \hat{\Delta}''_{b_i b_i}/n$ is positive definite.

A11. The quantity $\sum_{i=1}^{m} \hat{\Delta}''_{\beta b_i}/(nm^{1/2})$ is bounded for all θ, m and n.

Under the generalized linear mixed model the distributional convergences in A7 and A8 are established in Nie [53], under conditions A9-11 and some additional conditions that are discussed in details in [53]; they can be interpreted as non-collinearity among the covariates under the mixed model, for example. The uniform integrability conditions in A7 and A8, and the boundedness conditions in A9-11 are for the uniform integrability of R_{nm} which leads to $E(R_{nm}) = o(1)$ in the proof below; these are not always easy to establish in general. However conditions A9-11 can be directly verified under specific models such as the generalized linear mixed model, since the derivatives can be explicitly calculated.

of Theorem 1. A second order Taylor expansion of $\hat{\Delta}$ yields

$$\begin{aligned}\hat{\Delta}(\hat{\theta}_{nm}) &= \hat{\Delta}(\theta_{0m}) + (\hat{\theta}_{nm} - \theta_{0m})^\top \hat{\Delta}'(\theta_{0m}) + \frac{1}{2}(\hat{\theta}_{nm} - \theta_{0m})^\top \hat{\Delta}''(\bar{\theta}_{nm})(\hat{\theta}_{nm} - \theta_{0m}) \\ &= \hat{\Delta}(\theta_{0m}) - \frac{1}{2}(\hat{\theta}_{nm} - \theta_{0m})^\top \Delta''(\theta_{0m})(\hat{\theta}_{nm} - \theta_{0m}) + R_{nm}, \qquad (1.11)\end{aligned}$$

where

$$R_{nm} = (\hat{\theta}_{nm} - \theta_{0m})^\top \{\hat{\Delta}''(\bar{\theta}_{nm}) + \Delta''(\theta_{0m}) - 2\hat{\Delta}''(\tilde{\theta}_{nm})\}(\hat{\theta}_{nm} - \theta_{0m})/2,$$

$\bar{\theta}_{nm}, \tilde{\theta}_{nm}$ are measureable functions such that $||\bar{\theta}_{nm} - \theta_{0m}|| \leq ||\hat{\theta}_{nm} - \theta_{0m}||$ almost surely, $||\tilde{\theta}_{nm} - \theta_{0m}|| \leq ||\hat{\theta}_{nm} - \theta_{0m}||$ almost surely, and $\hat{\Delta}'(\theta_{0m}) = -\hat{\Delta}''(\tilde{\theta}_{nm})(\hat{\theta}_{nm} - \theta_{0m})$. Write $\hat{Q}_{nm}(\theta) = (\hat{\theta}_{nm} - \theta_{0m})^\top \hat{\Delta}''(\theta)(\hat{\theta}_{nm} - \theta_{0m})$, $Q_{nm}(\theta) = (\hat{\theta}_{nm} - \theta_{0m})^\top \Delta''(\theta)(\hat{\theta}_{nm} -$

θ_{0m}). Then $R_{nm} = \{\hat{Q}(\bar{\theta}_{nm}) + Q(\theta_{0m}) - 2\hat{Q}(\tilde{\theta}_{nm})\}/2$. We have that

$$\begin{aligned}
\hat{Q}_{nm}(\theta) &= \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{b}_1 - b_{01} \\ \vdots \\ \hat{b}_m - b_{0m} \end{pmatrix}^\top \begin{pmatrix} \frac{\partial^2 l_J}{\partial\beta\partial\beta^\top} & \frac{\partial^2 l_{J1}}{\partial\beta\partial b_1'} & \cdots & \frac{\partial^2 l_{Jm}}{\partial\beta\partial b_m'} \\ \frac{\partial^2 l_{J1}}{\partial b_1\partial\beta^\top} & \frac{\partial^2 l_{J1}}{\partial b_1\partial b_1'} & & 0 \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 l_{Jm}}{\partial b_m\partial\beta^\top} & 0 & \cdots & \frac{\partial^2 l_{Jm}}{\partial b_m\partial b_m'} \end{pmatrix} \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{b}_1 - b_{01} \\ \vdots \\ \hat{b}_m - b_{0m} \end{pmatrix} \\
&= (\hat{\beta} - \beta_0)^\top \hat{\Delta}''_{\beta\beta}(\hat{\beta} - \beta_0) + 2(\hat{\beta} - \beta_0)^\top \sum_{i=1}^m \hat{\Delta}''_{\beta b_i}(\hat{b}_i - b_{0i}) \\
&\quad + \sum_{i=1}^m (\hat{b}_i - b_{0i})^\top \hat{\Delta}''_{b_i b_i}(\hat{b}_i - b_{0i}),
\end{aligned}$$

with an analogous formula for $Q_{nm}(\theta)$. Under the above conditions A2, A3, A6-A8 it is seen that $R_{nm} = o_p(1)$. In addition, conditions A7-A11 imply that $E(R_{nm}) = o(1)$.

Now, take expectations conditional on b on both sides of equation (1.11):

$$\begin{aligned}
E(\hat{\Delta}(\hat{\theta}_{nm})) &= E\{\hat{\Delta}(\theta_{0m})\} - \frac{1}{2}E\{(\hat{\theta}_{nm} - \theta_{0m})^\top \Delta''(\theta_{0m})(\hat{\theta}_{nm} - \theta_{0m})\} + E(R_{nm}) \\
&= \Delta(\theta_{0m}) - \frac{1}{2}\text{tr}(\Omega\,\text{var}(\hat{\theta}_{nm})) + o(1).
\end{aligned}$$

In a similar manner we can show that $E(\Delta(\hat{\theta}_{nm})) = \Delta(\theta_{0m}) + \frac{1}{2}\text{tr}\{\Omega\,\text{var}(\hat{\theta}_{nm})\} + o(1)$. Replacing $\Delta(\theta_{0m})$ in the last two equations above we get that

$$E\{\Delta(\hat{\theta}_{nm})\} = E\{\hat{\Delta}(\hat{\theta}_{mn})\} - \text{tr}\{\Omega\,\text{var}(\hat{\theta}_{nm})\} + o(1).$$

Finally, it can be seen that $\text{tr}\{\Omega\,\text{var}(\hat{\theta}_{nm})\} = \text{tr}(G\,\Omega^{-1}) + o(1) = \rho + o(1)$. After the simplification of $\log p(\hat{b})$ terms, and taking expectations over b, we get that

$$-2E_{(y,b)}E_{y^0|b}[l\{y^0 \mid \hat{\theta}(y)\}] = -2E_y[l\{y \mid \hat{\theta}(y)\}] + 2\rho + o(1).$$

Now we show $p \le \rho \le p + q$. The semi-positive definite matrix W admits a square root, so we can write $Z_1 = W^{1/2}Z$, $X_1 = W^{1/2}X$. In formula (1.10) we have $Z^\top WZ - Z^\top WX(X^\top WX)^{-1}X^\top WZ = Z_1^\top(I - P)Z_1 = Z_2^\top Z_2$, where $P = X_1(X_1^\top X_1)^{-1}X_1^\top$ and therefore $I - P$ are both projection matrices, and $Z_2 = (I - P)Z_1$. Then $\rho = p + q - \text{tr}\{(Z_2^\top Z_2 + D^{-1})^{-1}D^{-1}\} = p + q - \text{tr}\{(I_q + Z_3^\top Z_3)^{-1}\} = p + q - \sum_{i=1}^q (1 + u_i)^{-1}$, where $Z_3 = Z_2 D^{1/2}$, and $0 \le u_1 \le \ldots \le u_q$ are the eigenvalues of $Z_3^\top Z_3$. It follows that $p \le \rho \le p + q$. □

1.8 Acknowledgements

The contents of Chapter 1 have been reproduced from the published manuscript of "Conditional Akaike information under generalized linear and proportional hazards mixed models" by Michael Donohue, Rosanna Overholser, Ronghui Xu and Florin Vaida, which appeared in *Biometrika*, 2011; **98**(3): 685-700. Additional simulation results have been added and results on proportional hazards mixed models omitted. Oxford Press, publisher of Biometrika, is acknowledged for allowing me to reproduce published material.

Table 1.1: Comparison of model selection procedures based on simulations from a Poisson generalized linear mixed model with log link. Data are generated from model (3,2): covariates x_1, x_2, x_3 are independent Bernoulli (0.5), $\beta = (1,1,1)^\top$, $z_1 = x_1, z_2 = x_2$, and $b_i \sim N(0, 0.25 I_2)$. Model (2,1) includes x_1, x_2 and z_1; model (3,1) includes also x_3, and (3,3) includes also z_2 and $z_3 = x_3$. Average over 100 simulations are reported for cAI, cAIC, cAIC_b with 400 bootstrap samples, and $-2l(y \mid \hat\beta, \hat b)$; a/b gives the number of times out of 100 a model was chosen using the rule of two (a) or simple minimum (b).

(fixed, random)	(2,1)	(3,1)	**(3,2)**	(3,3)
$m = 10, n_i = 5$				
cAI	318 (3.8)	244 (1.9)	230 (1.1)	232 (1.1)
	0/0	9/2	89/74	2/24
cAIC	321 (4.4)	244 (2.1)	229 (1.3)	227 (1.8)
	0/0	14/10	73/65	13/25
cAIC_b	369 (7.1)	252 (3.0)	235 (2.3)	242 (3.7)
	0/0	17/14	72/66	11/20
$-2pl(y \mid \hat\beta, \hat b)$	302 (4.4)	224 (2.1)	203 (1.2)	200 (1.2)
$m = 10, n_i = 10$				
cAI	674 (7.0)	494 (3.7)	448 (1.6)	450 (1.6)
	0/0	0/0	99/85	1/15
cAIC	675 (7.7)	495 (4.1)	447 (2.4)	448 (3.7)
	0/0	5/5	84/72	11/23
cAIC_b	746 (9.7)	509 (5.0)	450 (2.8)	454 (3.0)
	0/0	3/2	86/83	11/15
$-2pl(y \mid \hat\beta, \hat b)$	654 (7.7)	473 (4.1)	414 (2.2)	411 (2.2)
$m = 10, n_i = 40$				
cAI	2732 (28.1)	1929 (14.3)	1727 (6.1)	1729 (6.1)
	0/0	0/0	100/97	0/3
cAIC	2733 (29.0)	1933 (14.7)	1730 (6.9)	1741 (13.6)
	0/0	0/0	92/83	8/17
cAIC_b	2810 (31.0)	1943 (15.2)	1729 (7.0)	1733 (7.0)
	0/0	0/0	95/89	5/11
$-2pl(y \mid \hat\beta, \hat b)$	2711 (29.0)	1909 (14.7)	1691 (6.9)	1688 (6.8)
$m = 50, n_i = 5$				
cAI	1612 (8.1)	1240 (4.9)	1137 (2.4)	1140 (2.4)
	0/0	0/0	100/98	0/2
cAIC	1606 (10.0)	1238 (5.5)	1135 (2.8)	1135 (3.5)
	0/0	0/0	70/60	30/40
cAIC_b	1821 (14.2)	1276 (7.3)	1131 (3.8)	1137 (3.8)
	0/0	0/0	85/80	15/20
$-2pl(y \mid \hat\beta, \hat b)$	1519 (10.0)	1151 (5.4)	1001 (2.7)	996 (2.5)
$m = 50, n_i = 10$				
cAI	3352 (14.9)	2475 (9.1)	2219 (3.6)	2221 (3.6)
	0/0	0/0	100/96	0/4
cAIC	3354 (17.3)	2479 (10.7)	2218 (4.5)	2222 (6.3)
	0/0	0/0	79/69	21/31
cAIC_b	3670 (20.8)	2535 (12.9)	2210 (5.5)	2225 (5.6)
	0/0	0/0	99/99	1/1
$-2pl(y \mid \hat\beta, \hat b)$	3258 (17.3)	2382 (10.7)	2052 (4.4)	2046 (4.2)

Table 1.2: Comparison of model selection procedures based on simulations from a Poisson mixed model with log link. We enumerate the candidate models according the number of fixed and random effects included. Data is generated from model (3,2) with three independent Bernoulli ($p = 0.5$) fixed effect covariates x_1, x_2, x_3, and two random effect covariates $z_1 = x_1, z_2 = x_2$. The fixed effect is $\beta = (1, 1, 1)^{\mathrm{T}}$ and the random effects are $b_i \sim N(0, 0.25I_2)$. Model (2,1) includes x_1, x_2 and $z_1 = x_1$; model (3,1) includes x_1, x_2, x_3 and $z_1 = x_1$; and (3,3) includes x_1, x_2, x_3 and $z_1 = x_1, z_2 = x_2, z_3 = x_3$. Average (SE) over 100 simulations are reported for cAI, cAIC, cAIC_b from 400 bootstrap samples, and $-2l(y \mid \hat{\beta}, \hat{b})$. Below each average and SE are a/b: the number of times the model was chosen using the rule of two (a) or simple minimum (b).

(fixed, random)	(2,1)	(3,1)	**(3,2)**	(3,3)
$m = 10, n_i = 5$				
cAI	345.5 (8.1)	241.8 (2.1)	229.6 (1.2)	237.5 (1.6)
	0/0	28/4	70/77	2/19
cAIC	349.2 (8.9)	242.1 (2.3)	227.2(1.5)	236.5 (2.1)
	0/0	13/1	77/73	10/26
cAIC_b	403.1 (11.5)	256.3 (3.3)	246.4 (2.8)	309.8 (3.3)
	0/0	33/32	66/67	1/1
$-2l(y \mid \hat{\beta}, \hat{b})$	337.1(9.1)	223.8 (2.2)	202.4 (1.7)	215.6 (2.3)
$m = 10, n_i = 10$				
cAI	704.5 (10.9)	481.0 (2.9)	444.5 (1.8)	446.2 (1.9)
	0/0	0/0	86/65	14/35
cAIC	707.9 (12.1)	482.4 (3.2)	443.3 (2.1)	440.9 (2.0)
	0/0	2/0	59/23	39/77
cAIC_b	779.3 (15.6)	495.5 (3.9)	449.5 (2.5)	457.6 (2.7)
	0/0	3/3	86/85	11/12
$-2l(y \mid \hat{\beta}, \hat{b})$	692.8 (12.2)	461.3 (3.2)	411.0 (2.2)	407.4 (2.2)
$m = 10, n_i = 40$				
cAI	2973.0 (44.5)	1924.9 (11.1)	1728.5 (5.6)	1730.3 (5.2)
	0/0	0/0	81/70	19/30
cAIC	2967.5 (45.1)	1919.0 (11.3)	1724.0 (5.1)	1722.0 (3.6)
	0/0	0/0	67/27	33/73
cAIC_b	3062.0 (48.5)	1929.8 (11.8)	1722.8 (5.7)	1728.1 (5.7)
	0/0	0/0	94/88	6/12
$-2l(y \mid \hat{\beta}, \hat{b})$	2947.2 (45.1)	1895.6 (11.3)	1684.7 (5.5)	1682.0 (5.5)
$m = 50, n_i = 5$				
cAI	1760.0 (28.0)	1233.6 (7.5)	1138.4 (3.8)	1139.6 (3.9)
	0/0	0/0	73/60	27/40
cAIC	1755.5 (30.6)	1230.6 (8.3)	1134.6 (4.6)	1133.4 (4.6)
	0/0	0/0	77/28	23/72
cAIC_b	1979.0 (40.6)	1272.7 (11.4)	1138.2 (6.6)	1166.5 (6.6)
	0/0	0/0	97/97	3/3
$-2l(y \mid \hat{\beta}, \hat{b})$	1697.4 (22.7)	1151.2 (6.4)	998.8 (3.4)	992.3 (3.6)
$m = 50, n_i = 10$				
cAI	3637.5 (46.6)	2458.4 (11.2)	2212.7 (4.5)	2214.9 (4.4)
	0/0	0/0	77/61	23/39
cAIC	3658.1 (45.6)	2467.5 (11.9)	2216.6 (5.4)	2214.8 (5.5)
	0/0	0/0	76/33	24/67
cAIC_b	3990.6 (57.1)	2526.1 (14.7)	2219.0 (6.5)	2251.5 (7.2)
	0/0	0/0	98/97	2/3
$-2l(y \mid \hat{\beta}, \hat{b})$	3585.6 (35.4)	2375.8 (9.9)	2051.1 (4.8)	2045.9 (5.0)

Table 1.3: Skin cancer prevention study. Estimates of fixed effects (SE) and variance components, and cAIC from 5 Poisson mixed models differing in the fixed and random effects for year of follow-up. Age and number of skin cancers at baseline are continuous. The best models, marked with a *, cannot be ranked based on cAIC.

(fixed, random)	(5,1)	(6,1)	(7,1)	(6,2)	(7,2)
Estimates of β					
Intercept	-4.28	-4.35	-4.18	-4.79	-5.36
	(0.37)	(0.37)	(0.39)	(0.50)	(0.54)
Age (years)	0.02	0.02	0.02	0.02	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Skin that burns	0.33	0.33	0.33	0.33	0.32
	(0.10)	(0.12)	(0.11)	(0.14)	(0.15)
Male gender	0.63	0.63	0.63	0.69	0.70
	(0.12)	(0.12)	(0.12)	(0.16)	(0.17)
Baseline cancers	0.18	0.18	0.18	0.19	0.19
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Year	–	0.02	-0.13	-0.03	0.38
	–	(0.02)	(0.08)	(0.04)	(0.11)
Year^2	–	–	0.03	–	-0.08
	–	–	(0.01)	–	(0.02)
Estimates of σ^2					
Intercept	2.37	2.38	2.39	9.97	13.12
Year	–	–	–	0.85	1.22
$-2l(y \mid \hat{\beta}, \hat{b})$	6072.10	6068.57	6063.50	4942.28	4825.32
cAIC	7824.69*	7824.28*	7823.13*	8023.39	8038.86

Chapter 2

Effective Degrees of Freedom and Its Application to Conditional AIC for General Linear Mixed Models with Correlated Error Structures

2.1 Introduction

The general linear mixed model is a useful approach [13] to analyzing a wide variety of data structures in the applications of statistics, including longitudinal or repeated measures data, growth and dose-response curves, clustered or nested data, and multivariate data. While model fitting and parameter estimation is now broadly available in major statistical softwares, methods for model comparison and selection under mixed-effects models have only received attention very recently. Model complexity is an important related concept, often captured by the number of degrees of freedom. For complex models such as in the presence of mixed effects, this has commonly been referred to as the effective degrees of freedom, reflecting the fact that it is not straightforward to count the number of parameters as the number of degrees of freedom.

Although the number of degrees of freedom can be defined geometrically both under classical linear regression models, and under some richly parametrized models as

in [33], a more general definition seems to be related to the concept of Akaike information. The Akaike information is defined using the expected predictive log-likelihood given the estimated parameters. Since the population expectation is unknown, it is approximated by the corresponding observed data log-likelihood, plus a bias correction term. For the classical AIC [2], this bias correction is given by the number of unknown parameters. For more complex models such as in the presence of random effects, and when the conditional Akaike information (cAI) is of interest, the bias correction has been given by the effective degrees of freedom, which coincides with the trace of the so-called hat matrix under the mixed models [66, 15].

The mixed models considered in the literature typically assume independently and identically distributed (i.i.d.) errors. In this paper we consider the general linear mixed models, where the errors are not i.i.d., and model complexity should count for the complex error covariance structures. We prove a simple expression for the effective degrees of freedom, which is the trace of the hat matrix plus the number of parameters in the error covariance. Although some related work has been done as summarized below, this simple expression has not been known in the literature to the best of our knowledge.

The number of effective degrees of freedom typically involves unknown parameters. In practice for the purposes of model selection, the unknown parameters have to be replaced by their estimates. Using such a plug-in estimate one can define a crude version of the conditional AIC (cAIC), which is not guaranteed to be consistent or asymptotically unbiased for the cAI. For linear mixed model with i.i.d. errors, [48] proposed data perturbation [76], and [26] derived a complex expression assuming known variance, to obtain a cAIC that is asymptotically unbiased for the cAI. For general linear mixed models with correlated errors, [42] used an asymptotic expansion of the parameter estimates, and their formulas must be explicitly derived for each model in question via differentiation of the error covariance matrix with respect to the parameters.

2.2 cAIC for General Linear Mixed Model

2.2.1 Model and notation

Suppose that data $y = (y_1^\top, y_2^\top, \ldots, y_m^\top)^\top$ follow a general linear mixed model:

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{2.1}$$

for $i = 1 \ldots m$, where X_i is a $n_i \times p$ matrix of covariates for the fixed effects, Z_i is a $n_i \times q$ matrix of covariates for the random effects, β is a $p \times 1$ vector of fixed effects, b_i is a $q \times 1$ vector of random effects with distribution $\mathrm{N}(0, D)$, ϵ_i is a $n_i \times 1$ vector with distribution $\mathrm{N}(0, V_i(\alpha))$, and the covariance $V_i(\alpha)$ is known up to a $r \times 1$ vector of parameters α. In the following we will also refer to each i value as a cluster, and let $N = \sum_{i=1}^m n_i$. We can express the model in the larger matrix form by defining $X = (X_1^\top, X_2^\top, \ldots, X_m^\top)^\top$, $Z = \mathrm{diag}(Z_1, Z_2, \ldots, Z_m)$, $b = (b_1^\top, b_2^\top, \ldots, b_m^\top)^\top$, and $\epsilon = (\epsilon_1^\top, \epsilon_2^\top, \ldots, \epsilon_m^\top)^\top$. Then $y = X\beta + Zb + \epsilon$.

Given α and D, β and b are typically estimated by the best linear unbiased estimator (BLUE) and the best linear unbiased predictor (BLUP), respectively. Both estimators are linear in y so the predicted y can be expressed as $\hat{y} = Hy$, with trace

$$\mathrm{tr}(H) = \mathrm{tr}\left(\begin{bmatrix} X^\top W^{-1} X & X^\top W^{-1} Z \\ Z^\top W^{-1} X & Z^\top W^{-1} Z + \Sigma \end{bmatrix}^{-1} \begin{bmatrix} X^\top W^{-1} X & X^\top W^{-1} Z \\ Z^\top W^{-1} X & Z^\top W^{-1} Z \end{bmatrix} \right), \tag{2.2}$$

$\Sigma = \mathrm{diag}(D^{-1}, \ldots, D^{-1})$ and $W = \mathrm{diag}(V_1, \ldots, V_m)$. H is often called the "hat" matrix. In practice the variance parameters are typically estimated by maximum likelihood (ML) or restricted maximum likelihood (REML); in this case, $\hat{y}$ is no longer a linear function of y.

For our subsequent purposes denote $\theta = (\beta^\top, \alpha^\top, b^\top)^\top$, and the density of b by $p(b)$ where we suppress the variance parameters of b. The conditional log-likelihood of y given the random effects is

$$\begin{aligned} l_c(y|b) &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^m \{\log|V_i(\alpha)| \\ &\quad + (y_i - X_i\beta - Z_i b_i)^\top V_i(\alpha)^{-1} (y_i - X_i\beta - Z_i b_i)\}, \end{aligned}$$

where $|\cdot|$ denotes the determinant of a matrix. The joint log-likelihood of y and b is $l_J(\theta) = l_c(y|b) + \log p(b)$, which can sometimes be viewed as a penalized log-likelihood; see for example [4].

2.2.2 Effective degrees of freedom and cAI

The conditional Akaike information (cAI) was originally defined in [66] as

$$\text{cAI} = -2E_{y,b}E_{y^0|b}l_c(y^0|b;\hat{\theta}(y)), \tag{2.3}$$

where y^0 is an independent replicate of y from the distribution of y given b, and $\hat{\theta}$ is an estimator of θ. This definition of the Akaike information focuses on prediction given the random effects, or clusters. In other words, we are interested in predicting the outcome of a new observation for an existing cluster. As a contrast, the classical AIC using the marginal observed data log-likelihood under (2.1) is applicable when we are interested in predicting the outcome of a new observation for a new cluster.

Note that the modeling assumption of the covariance structure D of the random effects does not enter into the conditional log-likelihood (2.3) used in the definition of cAI. On the other hand, the variance parameters α of the error terms are explicitly counted for in the conditional likelihood. In this way the two sets of variance parameters are treated very differently. The focus of cAI is on the estimation of the random effects themselves, and their covariance structure becomes secondary in our view, although it does affect the overall modeling assumption and hence the parameter estimates. This distinction between the two sets of variance parameters also reflects the distinction between the idea of a marginal Akaike information versus that of a conditional Akaike information [66].

The Akaike information is typically approximated by the observed $-2l_c(y|b;\hat{\theta})$ plus a bias correction term. When $V_i = \sigma^2 I$ where σ^2 is known, $-2l_c(y|b;\hat{\theta}) + 2\text{tr}(H)$ is unbiased for cAI, where $\hat{\theta}$ is the maximum likelihood estimate of θ and $\text{tr}(H)$ is given in (2.2) [66]. In the following (mostly for technical reasons) we consider $\hat{\theta}$ as maximizes the joint log-likelihood $l_J(\theta)$, which under suitable conditions is a first-order Laplace approximation to the maximum likelihood estimator [4, 71, 67].

Theorem 2. *Under Conditions B1-B11 in the Appendix,*

$$-2l_c(y|b;\hat{\theta}) + 2\{tr(H) + r\} \tag{2.4}$$

is asymptotically unbiased for cAI under the general linear mixed model (2.1), where $r = dim(\alpha)$ *is the number of unknown parameters in the error covariance matrix. That is, the expectation of (2.4) with respect to the joint distribution of* y *and* b *equals cAI* $+$ $o(1)$ *for large* m *and* n_i*'s.*

The proof of the theorem is given in the Appendix. We refer to $\{\text{tr}(H) + r\}$ as the effective degrees of freedom under model (2.1). Intuitively this should fall within the range of two extremes: the number of fixed effects plus the number of error variance parameters, which corresponds to zero variances of the random effects, and the sum of these two plus the number of random effects, corresponding to infinite variances of the random effects [33]. Where it lies within the range reflects the amount of shrinkage in the estimation of the random effects.

We note that the previously mentioned results of [42] were derived via direct computation of expectations under the linear mixed model. It is curious to note that their degrees of freedom was only the $\text{tr}(H)$ part of ours, while the additional r cannot be obviously identified in their other term h_c.

2.2.3 Conditional AIC

It is apparent from (2.2) that $\text{tr}(H)$ involves the variance parameters α and those in D, which are unknown in practice. In the following we let

$$\text{cAIC} = -2l_c(y|b;\hat{\theta}) + 2[\text{tr}\{H(\hat{\alpha},\hat{D})\} + r], \tag{2.5}$$

where $\hat{\alpha}$ and $\hat{D}$ are estimates of α and D, respectively. Even though each element of H may be estimated consistently, because the dimension of H is on the order of the sample size, $\text{tr}\{H(\hat{\alpha},\hat{D})\}$ may have error greater than $o_p(1)$. In the next section we will compare cAIC with the proposed asymptotic correction of [42] through simulations.

In addition to the original AIC, the corrected AIC or, AIC_c, was derived for variable selection in linear regression and order selection in autoregressive processes

by [35] when the sample sizes are small. It was later extended for smoothing parameter selection by [34]. Although no general theoretical results exist beyond the initial parametric linear and autoregressive cases, studies have shown that AIC_c has impressive finite-sample properties [7, 31, 43, 46, 9]. Following previous proposals for AIC_c, we also consider the following criterion under model (2.1):

$$\text{cAIC}_c = -2l_c(y|b;\hat{\theta}) + \frac{2N[\text{tr}\{H(\hat{\alpha},\hat{D})\}+r]}{N-\text{tr}\{H(\hat{\alpha},\hat{D})\}-r-1}. \tag{2.6}$$

2.3 Simulations

In this section we describe the simulation experiments conducted to study cAIC, cAIC_c, and cAIC_k from [42]. The purposes of these simulations are to examine the accuracy of the criteria as estimates of cAI, as well as their performances in model selection. We considered various sample sizes, from $m = 10$ to 75 clusters, each containing $n = 10$ to 25 observations. For display purposes only $m = 10$ and $n = 25$ are given in the tables; other results are available upon request. One hundred data sets were generated for each case.

For each candidate model, we can approximate the 'true' cAI using definition (2.3) by taking the inner expectation over 100 "new" values of y^0 drawn from the same model as the original data, albeit with a new ϵ. This alone does not give the population quantity cAI, but instead is the approximate risk for each simulated data set. The average of the above over the 100 simulations gives the approximate cAI. In the tables the performance of cAIC, cAIC_c and cAIC_k were measured in two ways: by comparing their means (standard error) over 100 simulations to the approximate cAI, and by counting the number of times each model was chosen by a given criterion. A model was chosen in two different ways which is expressed in the format '-,-': the first count is the number of times out of 100 that a model had the minimum value of the criterion among all the models, and the second is the number of times out of 100 that a model was chosen by the 'rule-of-the-thumb: choose the smaller model if the difference in criterion is less than 2' [7, p.131]. If two models were not nested, the 'smaller' model was taken to be the one with the smaller estimated effective degrees of freedom, $\text{tr}(H)+r$. This 'rule-of-the-thumb' reflected a preference towards parsimonious modeling, as well as took

into account the statistical variation in the computed criteria [15]. When there were no random effects, the classical AIC was used for both cAIC and cAIC$_k$, while cAIC$_c$ was the corrected AIC of [35].

We first considered the six models in Table 2.1, with Model 4 being the true model. For the results in Tables 2.2 to 2.4 the models are in the orders of 1 - 6, but are described in the column names using the format '(p, q, r)'. This format helps to visualize the dimensions and nestedness of the models; among the models, those with the same total number of parameters $p + q + r$ are not nested within each other. Data were generated under Model 4, which was (2.1) with a random intercept, $X_{ij} = (1, x_i, t_j)^T$ where x_i was drawn from Uniform (1.5, 3.5), and $t_j = \log j$. We set $\beta = (1, 1, 1)'$. The random intercept b_i and errors ϵ_i were independently drawn from $N(0, 0.25)$ and $N(0, 0.05R)$, respectively, where R was a AR(1) correlation matrix with the correlation parameter ρ as given in the table. Finally for Tables 2.5 to 2.10 we focused more extensively on different error correlation structures, which was also the case for our application example below. Data were generated as $y_{ij} = 1 + 4x_j + b_i x_j + \epsilon_{ij}$, where $x_j = j$, $b_i \sim N(0, 4)$, and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in})' \sim N(0, 2R)$ where R was a AR(1) correlation matrix as before. The effective degrees of freedom under both models are around 13; in fact exactly 13 under the second model since the covariates are fixed. They fall within the range of 5-15 under the first model, and 4-14 under the second, the ranges obtained by counting the number of fixed effects plus the error variance parameters, with or without the number of random effects. The fact that they are close to the upper bounds reflects relatively little shrinkage in estimating the random effects, perhaps given the reasonably large cluster sizes.

From the tables we see that as estimates of the population quantity cAI, under the true models cAIC$_c$ tends to be the most accurate when the error correlation is weak, and cAIC$_k$ becomes the most accurate when the correlation is strong. When the models are wrong, there is no theory to support any of them to be a good estimate of cAI, and indeed in the simulations their values can be quite far from cAI. In terms of model selection, cAIC$_k$ performs the best in selecting the true model. Note that $\rho = 0.1$ makes the AR(1) structure close to independent, as reflected in the tables difficulty for any criterion to distinguish between the two error structures with these moderate sample sizes. As noted

in [26], cAIC has a tendency to select additional random effects, which is the case here for model (3,2,2). This also appears to be the case for cAIC$_c$. It is curious, however, that this is also the case for the simulated cAI which is the approximate risk as discussed above, and which does not involve tr(H) with the estimated α and D. cAIC$_k$ selects (3,2,2) sometimes, but much less frequently than the others.

It is important to note that with the 'rule-of-the-thumb', cAIC and cAIC$_c$ perform nearly as well as cAIC$_k$. This is important because cAIC and cAIC$_c$ are much easier to compute and readily available after fitting the mixed models. On the other hand, cAIC$_k$ requires additional programming and we had to do it specifically for each model considered. Our attempt to implement an R package using a general method (symbolic differentiation) without specifically programming for each case, ran into the prohibitive problem of intense computation with only moderate cluster sizes (and the corresponding sizes of error covariance matrices).

While data not shown, we note reduced performance of all three criteria when the sample sizes are smaller, especially when the cluster size n is small. In fact when $n = 10$ for example, cAIC$_k$ can perform slightly worse than cAIC and cAIC$_c$ in model selection. The reason that the larger cluster sizes are advantageous is because the conditional likelihood involves estimated b's, which are more accurately estimated when the clusters are larger.

2.4 An Example

[50] analyzed a data set on the effect of two drugs on forced expiratory volume (FEV). In the study, 72 subjects were randomly assigned to one of three treatment groups: drug A, drug B, or a placebo. The maximum amount of breath exhaled in 1 second, FEV$_1$, was measured at the beginning of the trial and each hour afterwards for 8 hours. A summary of the data is plotted in Figure 2.1, which indicates an approximately linear trend of FEV$_1$ over time. Figure 2.2 contains the scatter plots of the measurements against each other at different hours, which shows that the FEV$_1$ values closer in time are more correlated than those farther away. The data set is available at http://www.uvm.edu/ abh/stat295/datasets/.

We consider three general linear mixed models for the data. Each model had the same eight fixed effects including intercept, baseline FEV_1, treatment, hour, and treatment by hour interaction. A random intercept is also included in each model. Here we consider different error structures for the three models: independent, AR(1) and an AR(2). Table 2.11 shows the parameter estimates and cAIC, $cAIC_c$ and $cAIC_k$ under each of the three error structures. In the table τ is the standard deviation of the random intercept, and σ is the error standard deviation. Under the AR(1) structure the single correlation parameter is denoted by ρ_1, and under AR(2) the two correlation parameters are denoted by ρ_1 and ρ_2, i.e. $\epsilon_{ij} = \rho_1\epsilon_{i,j-1} + \rho_2\epsilon_{i,j-2} + e_{ij}$, where the e_{ij}'s are i.i.d $N(0, \sigma^2)$. It is seen that both cAIC and $cAIC_c$ have clearly selected the model with AR(1) error structure, with the 'rule-of-the-thumb' or not. $cAIC_k$ also selects the AR(1) model, although its value under AR(1) is only slightly lower than under independence. Notice that a test of significance, however, would have concluded that ρ_2 is significantly different from zero. An interpretation of this might be that if we are interested in predicting individual FEV_1 outcomes over time, the AR(1) model would do a better job.

2.5 Discussion

Measures of model complexity was discussed in [63], both in the classical sense and in the Bayesian context. Using bias correction for the (conditional) Akaike information to define the effective degrees of freedom, is the same as their approach using 'excess of true over the estimated residual information'. The number of effective degrees of freedom was derived this way in [66], as a theoretical justification to that of [33] under linear mixed models with known error variance. Under the general linear mixed models in this paper, because estimation of the error covariance parameters is involved, the theoretical development is asymptotic, and parallels that of our previous work [15]. In contrast, [66] and [25] involved mostly finite-sample matrix algebra under i.i.d. errors with known variances.

The AIC is commonly used for model selection in practice. It is well-known that AIC can be inconsistent in selecting the true model if the model dimensions are fixed [60]; further discussion on this can also be found in the literature, see for example

[74, 75]. For cAIC both the model dimension and corresponding sample sizes are more complex. We expect that some of characteristics of the classical AIC might carry over, and this is certainly an area for future research.

Proof of Theorem

Given the number of clusters m, let $\theta_m = (\beta^\top, \alpha^\top, b_1^\top, \ldots, b_m^\top)^\top$, and let $\theta = (\beta^\top, \alpha^\top, b_1^\top, b_2^\top, \ldots)^\top$ be the corresponding infinite-dimensional parameter; θ_m contains the first $m+2$ elements of θ. The "true" value of θ is θ_0, with first $m+2$ elements $\theta_{0m} = (\beta_0^\top, \alpha_0^\top, b_{01}^\top, \ldots, b_{0m}^\top)^\top$. Suppose data y is generated from the model with parameter $\theta_m = \theta_{0m}$. For ease of notation we assume $n_1 = ... = n_m = n$, so that $N = mn$. We will also denote $V_i = V$. For the proof and conditions below, $\hat{\theta}_{nm}$ maximizes the joint likelihood $l_J(\theta_m)$. Condition B5, however, can be used to establish that approximate likelihood methods such as $\hat{\theta}$ here is asymptotically equivalent to the MLE [67, 53], so that the theoretical results hold when MLE is used in calculating the cAIC.

Let $\Delta(\theta_m) = -E\{l_J(\theta_m)\}$, and $\hat{\Delta}(\theta_m) = -l_J(\theta_m)$. [53] distinguishes the components of β between those without a corresponding random effect, $\beta^{(1)}$, and those with a corresponding random effect, $\beta^{(2)}$, because the maximum likelihood estimate of the two sets of components have different rates of convergence: $1/\sqrt{N}$ for the former and $1/\sqrt{m}$ for the latter. For ease of description let us assume that $\beta = \beta^{(1)}$ in the following. Denote $\phi = (\beta, \alpha)$. Let $\Delta''_{\beta\beta}$, $\Delta''_{\beta\alpha_k}$, $\Delta''_{\alpha_k,\alpha_s}$, Δ''_{α_k,b_i}, Δ''_{α_k,b_i}, $\Delta''_{\beta b_i}$, $\Delta''_{b_i b_j}$, etc. be the corresponding matrix blocks from the Hessian matrix Δ'', with a similar notation for $\hat{\Delta}''$.

We assume that the following conditions hold:

B1. θ_0 is unique, and is in the interior of a convex closed bounded set $\Theta \subset \mathcal{R}^\infty$ equipped with the $\sup$ norm, eigenvalues of $V(\alpha_0)$ are bounded away from 0.

B2. $\hat{\beta} \to \beta_0$ almost surely (a.s.) and $\hat{\alpha} \to \alpha_0$ a.s. as $m, n \to \infty$

B3. $\max_{i=1,\ldots,m} \|\hat{b}_i - b_{0i}\|_2^2 \to 0$ almost surely as $m, n \to \infty$.

B4. For any m, n, the first and second derivatives Δ', $\hat{\Delta}'$ and Δ'', $\hat{\Delta}''$ exist, and $\hat{\Delta}''$ and Δ'' are continuous on Θ.

B5. $n/m \to \infty$.

B6. $(\hat{\Delta}-\Delta)/N$, $(\hat{\Delta}'-\Delta')/N$ and $(\hat{\Delta}''-\Delta'')/N$ converge almost surely to 0 uniformly on Θ as $m, n \to \infty$.

B7. $\sqrt{N}(\hat{\phi} - \phi) \to N(0, v_\phi)$ in distribution for some positive definite matrix v_ϕ.

B8. $\sqrt{n}(\hat{b}_i - b_{i0}) \to N(0, v_{bi})$ in distribution for some positive definite matrix v_{bi}, uniformly over i.

B9. $\lim_{m,n\to\infty} \Delta''_{\phi\phi}/N$ is positive definite.

B10. $\lim_{n\to\infty} \Delta''_{b_i b_i}/n$ is positive definite. In addition, $(\hat{b}_i-b_i)^\top \hat{\Delta}''_{b_i b_i}(\hat{b}_i-b_i) = O_p(1/m)$ as $m \to \infty$.

B11. The remainder term as described below is uniformly integrable.

A discussion of similar conditions can be found in our previous work [15].

Proof of Theorem 1

Inspiration: Linhart and Zucchini (1986) give a set of conditions under which

$$\Delta(\theta_0) \approx E\hat{\Delta}(\hat{\theta}) + \text{tr}\left(\Omega^{-1}G\right)/2,$$

where $\hat{\Delta}(\theta)$ is a general discrepancy measure whose expectation is given by Δ. We extended this result to general linear mixed models by

- changing to an ∞ dimensional parameter space.
- stating a new set of conditions.
- computing tr $(\Omega^{-1}G)$.

Let $\theta = (\beta^T, b^T, \sigma^T)^T$, $\Delta(\theta) = -E\{l_J(\theta; y)\}$, and $\hat{\Delta}(\theta) = -l_J(\theta; y)$ where $l_J(\theta; y) = \log f(y|\theta) + \log p(b)$. Our proof follows the lines of [15], which is a technical extension of the approach described in [49] to the 'estimation' (prediction) of the random effects

b, whose dimension $q \times m$ grows with the number of clusters m. The proof itself is mainly on consistency, i.e. convergence in probability, and the asymptotic unbiasedness follows under the condition of uniform integrability.

Similar to the proof in [15], we can show that

$$-2E\{l_c(y^0; \hat{\theta}(y))\} = -2E\{l_c(y; \hat{\theta}(y)\} + 2\text{tr}(\Omega^{-1}G) + o(1), \tag{2.7}$$

where $G = \text{var}\{\partial l_J(\theta)/\partial\theta\}$, and $\Omega = E\{-\partial^2 l_J(\theta)/\partial\theta\partial\theta^T\}$. The main idea of the proof lies in recognizing the difference between the joint log-likelihood and conditional log-likelihood being the log-density of b, and the fact that since the future new data y^0 is sampled using the same b as the original data y, the log-density term is a constant in the calculation of $E_{y^0|b}$. The rest of the proof of (2.7) follows the general outline of [49], albeit with conditions above for the mixed-effects model.

Taylor Expansion of $\hat{\Delta}$ about θ:

$$\begin{aligned}
\hat{\Delta}(\hat{\theta}_{nm}) &= \hat{\Delta}(\theta_{0m}) + (\hat{\theta}_{nm} - \theta_{0m})'\hat{\Delta}'(\theta_{0m}) + \frac{1}{2}(\hat{\theta}_{nm} - \theta_{0m})'\hat{\Delta}''(\bar{\theta}_{nm})(\hat{\theta}_{nm} - \theta_{0m}) \\
&= \hat{\Delta}(\theta_{0m}) - \frac{1}{2}(\hat{\theta}_{nm} - \theta_{0m})'\Delta''(\theta_{0m})(\hat{\theta}_{nm} - \theta_{0m}) + R_{nm},
\end{aligned}$$

where $R_{nm} = (\hat{\theta}_{nm} - \theta_{0m})'\{\hat{\Delta}''(\bar{\theta}_{nm}) + \Delta''(\theta_{0m}) - 2\hat{\Delta}''(\tilde{\theta}_{nm})\}(\hat{\theta}_{nm} - \theta_{0m})/2$.
Now, take expectations (conditional on b) on both sides:

$$\begin{aligned}
E(\hat{\Delta}(\hat{\theta}_{nm})) &= E\{\hat{\Delta}(\theta_{0m})\} - \frac{1}{2}E\{(\hat{\theta}_{nm} - \theta_{0m})'\Delta''(\theta_{0m})(\hat{\theta}_{nm} - \theta_{0m})\} + E(R_{nm}) \\
&= \Delta(\theta_{0m}) - \frac{1}{2}\text{tr}(\Omega \cdot \text{var}(\hat{\theta}_{nm})) + o(1),
\end{aligned}$$

if $R_{nm} = o_p(1)$ and is uniformly integrable. In a similar manner, we can show that

$$E(\Delta(\hat{\theta}_{nm})) = \Delta(\theta_{0m}) + \frac{1}{2}\text{tr}(\Omega \cdot \text{var}(\hat{\theta}_{nm})) + o(1).$$

Replacing $\Delta(\theta_{0m})$ in the last two equations, we arrive at

$$E(\Delta(\hat{\theta}_{nm})) = E(\hat{\Delta}(\hat{\theta}_{mn})) - \text{tr}(\Omega \cdot \text{var}(\hat{\theta}_{nm})) + o(1).$$

Write

$$\begin{aligned}
R_{nm} &= (\hat{\theta}_{nm} - \theta_{0m})'\{\hat{\Delta}''(\bar{\theta}_{nm}) + \Delta''(\theta_{0m}) - 2\hat{\Delta}''(\tilde{\theta}_{nm})\}(\hat{\theta}_{nm} - \theta_{0m})/2 \\
&= \{\hat{Q}(\bar{\theta}_{nm}) + Q(\theta_{0m}) - 2\hat{Q}(\tilde{\theta}_{nm})\}/2.
\end{aligned}$$

where $\hat{Q}_{nm}(\theta)$ is

$$\begin{pmatrix} \hat{\beta}-\beta_0 \\ \hat{\alpha}-\alpha_0 \\ \hat{b}_1-b_{01} \\ \vdots \\ \hat{b}_m-b_{0m} \end{pmatrix}^T \begin{pmatrix} \hat{\Delta}''_{\beta\beta} & \hat{\Delta}''_{\beta\alpha} & \hat{\Delta}''_{\beta b_1} & \cdots & \hat{\Delta}''_{\beta b_m} \\ \hat{\Delta}''_{\alpha\beta} & \hat{\Delta}''_{\alpha\alpha} & \hat{\Delta}''_{\alpha b_1} & \cdots & \hat{\Delta}''_{\alpha b_m} \\ \hat{\Delta}''_{b_1\beta} & \hat{\Delta}''_{b_1\alpha} & \hat{\Delta}''_{b_1 b_1} & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ \hat{\Delta}''_{b_m\beta} & \hat{\Delta}''_{b_m\alpha} & 0 & \cdots & \hat{\Delta}''_{b_m b_m} \end{pmatrix} \begin{pmatrix} \hat{\beta}-\beta_0 \\ \hat{\alpha}-\alpha_0 \\ \hat{b}_1-b_{01} \\ \vdots \\ \hat{b}_m-b_{0m} \end{pmatrix}.$$

Multiply each term in R_{nm} by the appropriate factor: $R_{nm} = \sqrt{N}(\hat{\beta}-\beta_0)[\{\hat{\Delta}''_{\bar{\beta}\bar{\beta}} + \Delta''_{\beta_0\beta_0} - 2\hat{\Delta}''_{\tilde{\beta}\tilde{\beta}}\}/N]\sqrt{N}(\hat{\beta}-\beta_0) + \ldots$

The rest of the proof of Theorem 2 involves explicitly computing $\text{tr}(\Omega^{-1}G)$. Note that

$$\begin{aligned} -2l_J(\theta) &= mn\log 2\pi + m\log|V| + \sum_{i=1}^{m}(y_i - X_i\beta - Z_i b_i)^\top V^{-1}(y_i - X_i\beta - Z_i b_i) \\ &\quad + mq\log 2\pi + m\log|D| + \sum_{i=1}^{m} b_i^\top D^{-1} b_i. \end{aligned}$$

We have

$$\begin{aligned} \hat{\Delta}'_\beta &= -\sum_{i=1}^{m}(y_i - X_i\beta - Z_i b_i)^\top V^{-1} X_i, \\ \hat{\Delta}'_{\alpha_k} &= -\frac{m}{2}\cdot\text{tr}\left(V^{-1}\frac{\partial V}{\partial\alpha_k}\right) \\ &\quad + \frac{1}{2}\sum_{i=1}^{m}(y_i - X_i\beta - Z_i b_i)^\top V^{-1}\frac{\partial V}{\partial\alpha_k}V^{-1}(y_i - X_i\beta - Z_i b_i), \\ \hat{\Delta}'_{b_i} &= -(y_i - X_i\beta - Z_i b_i)^\top V^{-1} Z_i + b_i D^{-1}, \end{aligned}$$

and, using the matrix notation of Section 2.1,

$$\hat{\Delta}'' = \begin{bmatrix} X^\top W^{-1} X & X^\top W^{-1} M_\alpha W^{-1} e & X^\top W^{-1} Z \\ e^\top W^{-1} M_\alpha W^{-1} X & M_{\alpha\alpha} & e^\top W^{-1} M_\alpha W^{-1} Z \\ Z^\top W^{-1} X & Z^\top W^{-1} M_\alpha W^{-1} e & Z^\top W^{-1} Z + \Sigma \end{bmatrix}$$

where $e_i = y_i - X_i\beta - Z_i b_i$, $e = (e_1^\top, \ldots, e_m^\top)^\top$, $M_\alpha = \text{diag}(\partial V/\partial\alpha_1, \ldots, \partial V/\partial\alpha_r)$,

and the elements of $M_{\alpha\alpha}$ are

$$\begin{aligned}\{M_{\alpha\alpha}\}_{ks} &= -\frac{m}{2}\text{tr}\left(-V^{-1}\frac{\partial V}{\partial \alpha_s}V^{-1}\frac{\partial V}{\partial \alpha_k}+V^{-1}\frac{\partial^2 V}{\partial \alpha_s \partial \alpha_k}V^{-1}\right)\\ &\quad+\sum_{i=1}^{m} e_i^\top\left[-V^{-1}\frac{\partial V}{\partial \alpha_s}V^{-1}\frac{\partial V}{\partial \alpha_k}V^{-1}-V^{-1}\frac{\partial V}{\partial \alpha_k}V^{-1}\frac{\partial V}{\partial \alpha_s}V^{-1}\right]e_i\\ &\quad+\sum_{i=1}^{m} e_i^\top\left[V^{-1}\frac{\partial^2 V}{\partial \alpha_s \partial \alpha_k}V^{-1}\right]e_i.\end{aligned}$$

Using the fact that $E(e_i)=0$ and formulas from McCulloch, Searle and Neuhaus (2008) for expectation of $M_{\alpha\alpha}$, we have

$$\begin{aligned}\Omega &= E\left(\hat{\Delta}''\right)\\ &= \begin{bmatrix} X^\top W^{-1}X & 0 & X^\top W^{-1}Z \\ 0 & E(M_{\alpha\alpha}) & 0 \\ Z^\top W^{-1}X & 0 & Z^\top W^{-1}Z+\Sigma \end{bmatrix},\\ G &= \text{var}\left(\hat{\Delta}'\right)\\ &= \begin{bmatrix} X^\top W^{-1}X & 0 & X^\top W^{-1}Z \\ 0 & E(M_{\alpha\alpha}) & 0 \\ Z^\top W^{-1}X & 0 & Z^\top W^{-1}Z \end{bmatrix},\end{aligned}$$

where $\{E(M_{\alpha\alpha})\}_{ks} = -\text{tr}(V^{-1}\cdot \partial V/\partial\alpha_k \cdot V^{-1}\cdot \partial V\partial\alpha_s)/2$. From the above we see that $\text{tr}\,(\Omega^{-1}G) = \text{tr}\,(H)+r$.

Table 2.1: Models used in simulation Tables 2.2 to 2.4. Model 4 is the true model that generated the data.

Model	Fixed effects	Random effects	Error structure
1	1, x	1	AR(1)
2	1, x, t	-	AR(1)
3	1, x, t	1	i.i.d.
4	1, x, t	1	AR(1)
5	1, x, t	1, t	i.i.d.
6	1, x, t	1, t	AR(1)

2.6 Acknowledgements

The contents of Chapter 2 have been reproduced from the submitted manuscript of "Effective Degrees of Freedom and Its Application to Conditional AIC for General Linear Mixed Models with Correlated Error Structures" by Rosanna Overholser and Ronghui Xu. Additional simulation results have been added.

Table 2.2: Simulation Results for m=10, n=25. The models correspond to those in Table 2.1; the format '(p, q, r)' denotes the numbers of fixed and random effects, and the number of parameters in the error variance. (3,1,2) is the true model, and ρ is the correlation parameter under the AR(1) error. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details.

	quantity	(2,1,2)	(3,0,2)	(3,1,1)	**(3,1,2)**	(3,2,1)	(3,2,2)
$\rho = 0.1$							
	-2l$_c$	220.9(1.9)	91.3(2.7)	-54.9(2.3)	-55.9(2.3)	-57.6(2.2)	-57.7(2.3)
		0	0	0	21	54	25
	cAI	225.2(0.2)	97.1(1.6)	-23.1(0.7)	-26.4(0.7)	-22.4(0.7)	-26.3(0.7)
		0, 0	0, 0	9, 12	42, 62	9, 7	40, 19
	cAIC	228.9(1.9)	101.3(2.7)	-29.1(2.3)	-28.2(2.3)	-30.1(2.3)	-28.8(2.3)
		0, 0	0, 0	32, 82	10, 5	53, 13	5, 0
	cAIC$_c$	229.0(1.9)	101.5(2.7)	-27.6(2.3)	-26.4(2.3)	-28.4(2.3)	-26.8(2.3)
		0, 0	0, 0	44, 84	9, 5	43, 11	4, 0
	cAIC$_k$	288.2(1.4)	101.3(2.7)	-28.7(2.3)	-27.2(2.3)	-26.3(2.3)	-24.2(2.3)
		0, 0	0, 0	77, 88	11, 6	12, 6	0, 0
$\rho = 0.4$							
	-2l$_c$	159.1(1.7)	1.0(2.5)	-69.8(2.5)	-96.7(2.1)	-82.3(2.7)	-99.5(2.3)
		0	0	0	23	3	74
	cAI	162.9(0.2)	7.3(1.1)	-9.2(1.5)	-66.3(0.9)	-1.9(1.9)	-65.0(1.0)
		0, 0	0, 0	0, 0	55, 81	0, 0	45, 19
	cAIC	167.1(1.7)	11.0(2.5)	-44.0(2.5)	-69.2(2.1)	-50.3(2.6)	-70.3(2.2)
		0, 0	0, 0	0, 0	39, 81	1, 1	60, 18
	cAIC$_c$	167.3(1.7)	11.2(2.5)	-42.5(2.5)	-67.4(2.1)	-47.9(2.6)	-68.3(2.2)
		0, 0	0, 0	0, 0	45, 86	0, 1	55, 13
	cAIC$_k$	265.5(1.6)	11.0(2.5)	-43.6(2.5)	-67.6(2.1)	-47.0(2.6)	-65.3(2.2)
		0, 0	0, 0	0, 0	86, 93	0, 1	14, 6
$\rho = 0.7$							
	-2l$_c$	77.5(1.3)	-153.7(2.3)	-112.8(3.1)	-214.7(2.2)	-149.1(3.2)	-218.8(2.2)
		0	0	0	22	0	78
	cAI	76.8(0.1)	-150.7(0.7)	42.3(3.5)	-183.9(0.9)	85.5(4.9)	-181.8(1.1)
		0, 0	1, 1	0, 0	56, 81	0, 0	43, 18
	cAIC	85.5(1.3)	-143.7(2.3)	-87.0(3.1)	-187.6(2.2)	-111.7(3.1)	-189.5(2.2)
		0, 0	0, 0	0, 0	23, 74	0, 0	77, 26
	cAIC$_c$	85.7(1.3)	-143.4(2.3)	-85.4(3.1)	-185.9(2.2)	-108.4(3.1)	-187.5(2.2)
		0, 0	0, 0	0, 0	25, 76	0, 0	75, 24
	cAIC$_k$	283.3(2.7)	-143.7(2.3)	-86.4(3.2)	-184.8(2.2)	-108.6(3.2)	-183.2(2.2)
		0, 0	0, 0	0, 0	78, 85	0, 0	22, 15
$\rho = 0.9$							
	-2l$_c$	-5.7(1.1)	-418.5(2.4)	-230.2(5.6)	-450.6(2.9)	-318.3(5.0)	-458.1(2.9)
		0	0	0	24	0	76
	cAI	-5.0(0.1)	-408.3(0.7)	339.1(17.7)	-413.8(2.2)	643.3(26.1)	-411.3(2.6)
		0, 0	15, 19	0, 0	50, 58	0, 0	35, 23
	cAIC	2.3(1.1)	-408.5(2.4)	-204.3(5.6)	-426.8(2.7)	-277.3(5.0)	-430.5(2.7)
		0, 0	1, 2	0, 0	25, 65	0, 0	74, 33
	cAIC$_c$	2.5(1.1)	-408.3(2.4)	-202.8(5.6)	-425.5(2.6)	-273.4(5.0)	-428.7(2.7)
		0, 0	1, 3	0, 0	26, 64	0, 0	73, 33
	cAIC$_k$	390.9(5.2)	-408.5(2.4)	-203.8(5.6)	-399.9(6.7)	-275.0(4.9)	-415.7(3.2)
		0, 0	18, 24	0, 0	61, 59	0, 0	21, 17

Table 2.3: Simulation Results for m=20, n=25.The models correspond to those in Table 2.1; the format '(p, q, r)' denotes the numbers of fixed and random effects, and the number of parameters in the error variance. (3,1,2) is the true model, and ρ is the correlation parameter under the AR(1) error. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details.

	quantity	(2,1,2)	(3,0,2)	(3,1,1)	**(3,1,2)**	(3,2,1)	(3,2,2)
$\rho = 0.1$							
	-2l$_c$	451.1(2.8)	200.6(3.9)	-100.6(3.1)	-101.9(3.1)	-107.2(3.1)	-105.3(3.1)
		0	0	0	5	79	16
	cAI	449.1(0.2)	198.2(1.7)	-52.0(0.8)	-59.1(0.8)	-49.6(1.0)	-58.4(0.8)
		0, 0	0, 0	4, 4	53, 75	3, 3	40, 18
	cAIC	459.1(2.8)	210.6(3.9)	-54.9(3.1)	-54.3(3.1)	-56.9(3.1)	-55.3(3.1)
		0, 0	0, 0	25, 68	5, 8	57, 22	13, 2
	cAIC$_c$	459.2(2.8)	210.7(3.9)	-52.7(3.1)	-51.9(3.1)	-54.1(3.1)	-52.5(3.1)
		0, 0	0, 0	35, 78	7, 9	48, 11	10, 2
	cAIC$_k$	517.3(2.4)	210.6(3.9)	-54.6(3.1)	-53.4(3.1)	-53.3(3.1)	-50.7(3.1)
		0, 0	0, 0	72, 84	15, 6	11, 8	2, 2
$\rho = 0.4$							
	-2l$_c$	327.9(2.5)	20.5(3.5)	-128.6(3.7)	-182.5(3.3)	-157.9(3.8)	-186.9(3.3)
		0	0	0	20	1	79
	cAI	324.6(0.2)	20.3(1.3)	-27.0(1.6)	-138.6(0.8)	-12.0(1.9)	-136.8(0.8)
		0, 0	0, 0	0, 0	60, 79	0, 0	40, 21
	cAIC	335.9(2.5)	30.5(3.5)	-82.9(3.7)	-135.2(3.3)	-96.1(3.7)	-136.6(3.2)
		0, 0	0, 0	0, 0	32, 80	0, 0	68, 20
	cAIC$_c$	336.0(2.5)	30.6(3.5)	-80.7(3.7)	-132.8(3.3)	-91.9(3.7)	-133.8(3.2)
		0, 0	0, 0	0, 0	37, 83	0, 0	63, 17
	cAIC$_k$	434.7(2.0)	30.5(3.5)	-82.5(3.7)	-133.7(3.3)	-92.9(3.7)	-131.6(3.2)
		0, 0	0, 0	0, 0	84, 89	0, 0	16, 11
$\rho = 0.7$							
	-2l$_c$	154.7(2.2)	-301.1(3.0)	-197.2(5.0)	-420.1(3.0)	-281.1(5.2)	-426.2(3.3)
		0	0	0	23	0	77
	cAI	153.1(0.2)	-301.3(1.0)	68.9(4.6)	-377.2(0.9)	165.1(7.0)	-373.6(1.3)
		0, 0	0, 0	0, 0	58, 84	0, 0	42, 16
	cAIC	162.7(2.2)	-291.1(3.0)	-151.5(5.0)	-373.8(3.0)	-208.1(5.1)	-376.0(3.1)
		0, 0	0, 0	0, 0	30, 73	0, 0	70, 27
	cAIC$_c$	162.8(2.2)	-291.0(3.0)	-149.2(5.0)	-371.5(3.0)	-202.2(5.1)	-373.2(3.1)
		0, 0	0, 0	0, 0	33, 76	0, 0	67, 24
	cAIC$_k$	359.3(2.3)	-291.1(3.0)	-151.1(5.0)	-370.9(3.0)	-205.6(5.1)	-369.5(3.1)
		0, 0	0, 0	0, 0	79, 87	0, 0	21, 13

Table 2.4: Simulation Results for m=20, n=50. The models correspond to those in Table 2.1; the format '(p, q, r)' denotes the numbers of fixed and random effects, and the number of parameters in the error variance. (3,1,2) is the true model, and ρ is the correlation parameter under the AR(1) error. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details.

	quantity	(2,1,2)	(3,0,2)	(3,1,1)	**(3,1,2)**	(3,2,1)	(3,2,2)
$\rho = 0.1$							
	-2l$_c$	687.3(4.0)	364.9(6.2)	-183.2(4.9)	-189.0(4.9)	-190.4(4.9)	-192.7(4.9)
		0	0	0	11	32	57
	cAI	688.3(0.3)	366.6(3.2)	-129.5(1.0)	-143.1(0.9)	-127.6(1.0)	-142.1(0.9)
		0, 0	0, 0	1, 1	55, 61	0, 0	44, 38
	cAIC	695.3(4.0)	374.9(6.2)	-137.4(4.9)	-141.2(4.9)	-139.4(4.9)	-142.2(4.9)
		0, 0	0, 0	5, 31	32, 53	17, 12	46, 4
	cAIC$_c$	695.3(4.0)	375.0(6.2)	-136.3(4.9)	-140.0(4.9)	-137.9(4.9)	-140.8(4.9)
		0, 0	0, 0	7, 32	33, 53	16, 11	44, 4
	cAIC$_k$	735.2(3.8)	374.9(6.2)	-137.2(4.9)	-140.7(4.9)	-135.9(4.9)	-138.2(4.9)
		0, 0	0, 0	15, 37	72, 54	8, 7	5, 2
$\rho = 0.4$							
	-2l$_c$	380.5(4.3)	6.0(5.4)	-203.1(5.7)	-345.8(5.1)	-234.2(5.8)	-349.3(5.1)
		0	0	0	14	0	86
	cAI	379.1(0.3)	4.6(2.3)	-102.9(1.6)	-303.1(0.8)	-88.1(2.1)	-302.3(0.8)
		0, 0	0, 0	0, 0	55, 74	0, 0	45, 26
	cAIC	388.5(4.3)	16.0(5.4)	-157.3(5.7)	-298.2(5.1)	-171.3(5.7)	-299.1(5.1)
		0, 0	0, 0	0, 0	47, 84	0, 0	53, 16
	cAIC$_c$	388.5(4.3)	16.1(5.4)	-156.2(5.7)	-297.0(5.1)	-169.1(5.7)	-297.8(5.1)
		0, 0	0, 0	0, 0	51, 84	0, 0	49, 16
	cAIC$_k$	462.5(3.8)	16.0(5.4)	-157.1(5.7)	-297.4(5.1)	-168.3(5.7)	-294.8(5.1)
		0, 0	0, 0	0, 0	90, 93	0, 0	10, 7
$\rho = 0.7$							
	-2l$_c$	-84.4(3.7)	-657.4(4.4)	-291.1(6.8)	-845.5(4.4)	-374.0(7.7)	-850.5(4.4)
		0	0	0	18	0	82
	cAI	-82.2(0.3)	-650.3(1.2)	-25.5(3.9)	-794.1(0.8)	46.6(6.0)	-792.2(0.9)
		0, 0	0, 0	0, 0	60, 76	0, 0	40, 24
	cAIC	-76.4(3.7)	-647.4(4.4)	-245.2(6.8)	-798.3(4.4)	-301.2(7.6)	-799.9(4.4)
		0, 0	0, 0	0, 0	33, 76	0, 0	67, 24
	cAIC$_c$	-76.3(3.7)	-647.3(4.4)	-244.1(6.8)	-797.2(4.4)	-298.4(7.5)	-798.5(4.4)
		0, 0	0, 0	0, 0	34, 76	0, 0	66, 24
	cAIC$_k$	113.1(3.1)	-647.4(4.4)	-245.0(6.8)	-796.9(4.4)	-298.9(7.6)	-795.0(4.4)
		0, 0	0, 0	0, 0	79, 88	0, 0	21, 12

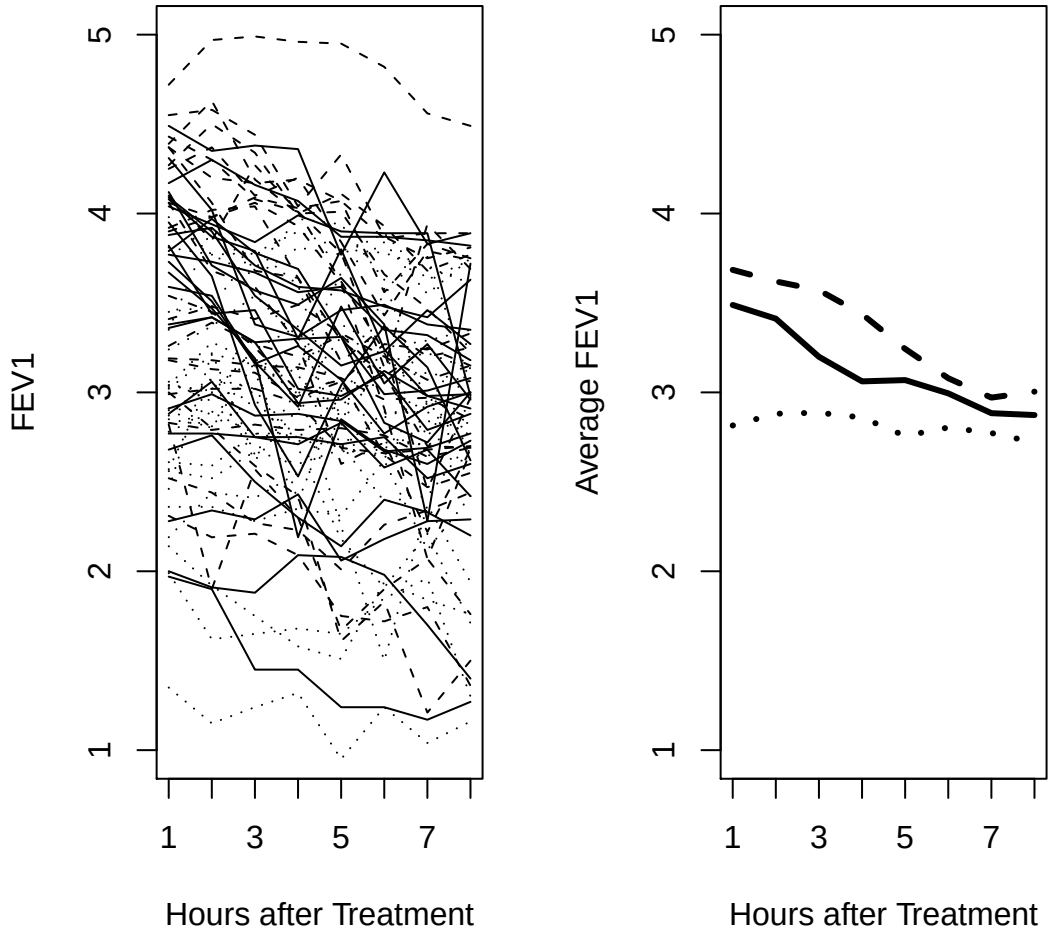

Figure 2.1: Spaghetti plot (left) and group means (right) for the FEV data: solid - drug A, dashed - drug B, dotted - placebo.

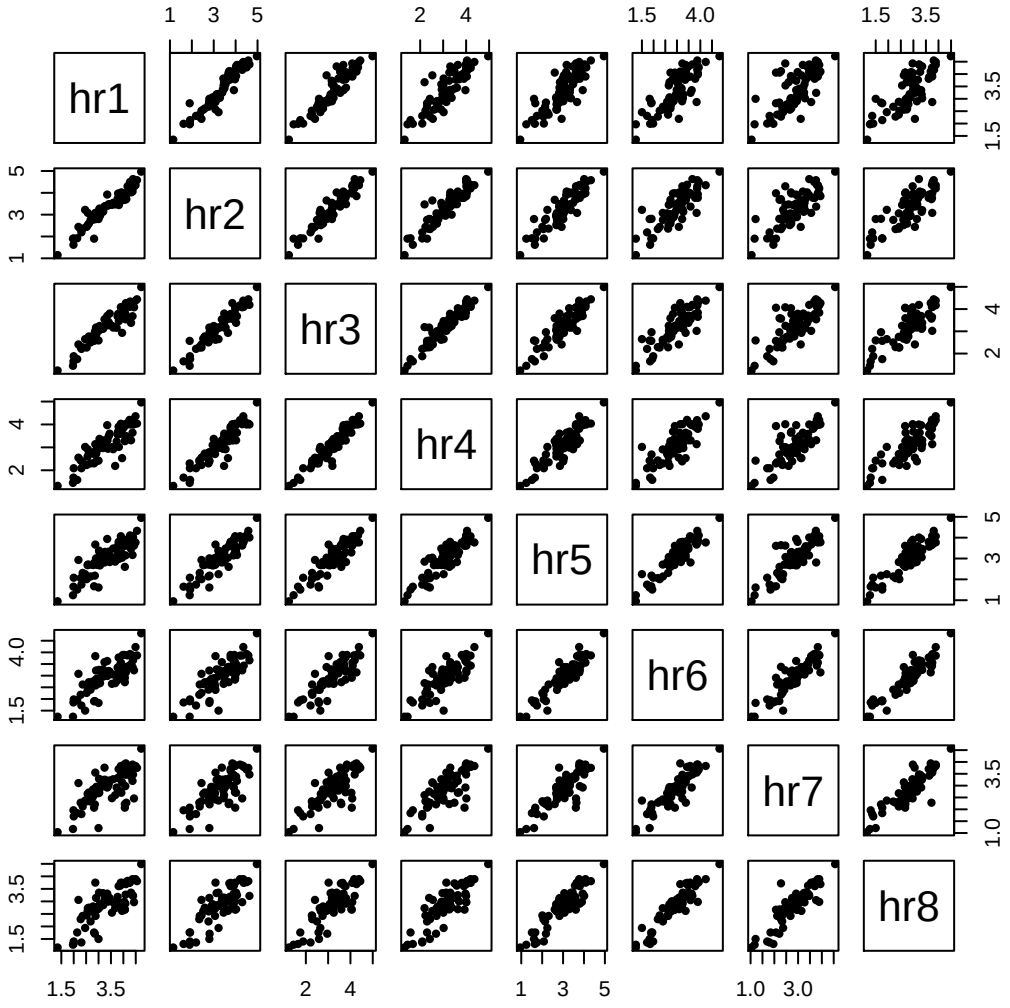

Figure 2.2: Scatter plot of FEV1 over 8 hours

Table 2.5: Simulation results comparing different error correlation structures, for m=10, n=10. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details.

	quant.	iid	ar1	ar2	comp sym
$\rho = 0.1$					
	-2l$_c$	337.1(1.6)	337.3(1.6)	335.5(1.7)	339.2(1.7)
		27	11	60	2
	cAI	367.9(0.7)	366.6(0.7)	370.0(1.0)	368.4(0.7)
		14, 43	50, 40	19, 1	17, 16
	cAIC	361.0(1.6)	363.2(1.6)	363.4(1.7)	365.2(1.7)
		80, 91	5, 1	15, 8	0, 0
	cAIC$_c$	364.6(1.6)	367.4(1.6)	368.4(1.7)	369.4(1.7)
		84, 93	5, 1	11, 6	0, 0
	cAIC$_k$	363.7(1.6)	369.1(1.6)	373.1(1.7)	435.9(3.5)
		95, 100	1, 0	4, 0	0, 0
$\rho = 0.4$					
	-2l$_c$	327.3(1.5)	323.5(1.4)	320.7(1.4)	331.1(1.6)
		9	26	65	0
	cAI	383.9(1.4)	354.3(1.2)	359.6(1.6)	375.2(1.4)
		0, 0	68, 88	30, 9	2, 3
	cAIC	351.2(1.5)	349.4(1.4)	348.6(1.4)	357.0(1.6)
		28, 45	28, 24	43, 29	1, 2
	cAIC$_c$	354.8(1.5)	353.6(1.4)	353.5(1.4)	361.2(1.6)
		33, 52	31, 23	35, 24	1, 1
	cAIC$_k$	353.9(1.5)	356.5(1.4)	359.3(1.4)	386.5(3.0)
		69, 79	15, 11	15, 9	1, 1
$\rho = 0.7$					
	-2l$_c$	299.8(2.2)	277.7(1.6)	276.2(1.7)	292.4(1.9)
		0	39	59	2
	cAI	432.3(3.9)	310.5(1.2)	314.6(1.8)	384.8(2.7)
		0, 0	70, 91	30, 9	0, 0
	cAIC	323.8(2.2)	303.6(1.6)	304.1(1.7)	318.3(1.9)
		0, 0	71, 82	26, 16	3, 2
	cAIC$_c$	327.4(2.2)	307.8(1.6)	309.0(1.7)	322.5(1.9)
		0, 0	78, 84	18, 14	4, 2
	cAIC$_k$	326.5(2.2)	312.9(1.6)	316.9(1.7)	326.7(1.8)
		2, 6	74, 72	15, 10	9, 12

Table 2.6: Simulation results comparing different error correlation structures, for m=25, n=10. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details.

	quant.	iid	ar1	ar2	comp sym
$\rho = 0.1$					
	-2l$_c$	847.0(2.7)	849.3(2.7)	848.0(2.8)	855.6(2.9)
		66	4	28	2
	cAI	916.2(1.0)	909.4(1.0)	911.7(1.1)	914.5(1.1)
		2, 5	49, 61	30, 5	19, 29
	cAIC	900.9(2.7)	905.2(2.7)	905.9(2.8)	911.5(2.9)
		90, 91	0, 0	10, 9	0, 0
	cAIC$_c$	907.7(2.7)	912.5(2.7)	913.8(2.8)	918.8(2.9)
		91, 93	0, 0	9, 7	0, 0
	cAIC$_k$	903.6(2.7)	911.1(2.7)	915.8(2.8)	966.7(3.5)
		96, 98	0, 0	4, 2	0, 0
$\rho = 0.4$					
	-2l$_c$	824.9(2.9)	815.4(2.6)	813.6(2.6)	836.9(2.8)
		6	33	61	0
	cAI	945.7(2.0)	873.8(1.1)	876.3(1.3)	928.4(2.2)
		0, 0	65, 89	35, 11	0, 0
	cAIC	878.9(2.9)	871.3(2.6)	871.5(2.6)	892.7(2.8)
		16, 24	43, 51	41, 25	0, 0
	cAIC$_c$	885.7(2.9)	878.6(2.6)	879.4(2.6)	900.0(2.8)
		16, 25	48, 55	36, 20	0, 0
	cAIC$_k$	881.6(2.9)	878.4(2.6)	882.4(2.6)	910.9(2.8)
		40, 49	46, 40	14, 11	0, 0
$\rho = 0.7$					
	-2l$_c$	767.5(3.4)	707.4(2.5)	705.1(2.5)	745.9(3.0)
		0	41	58	1
	cAI	1050.9(6.1)	764.0(1.6)	767.7(1.7)	946.3(4.4)
		0, 0	68, 88	32, 12	0, 0
	cAIC	821.4(3.4)	763.2(2.5)	762.8(2.5)	801.7(3.0)
		0, 0	59, 70	40, 29	1, 1
	cAIC$_c$	828.2(3.4)	770.5(2.5)	770.7(2.5)	809.0(3.0)
		0, 0	65, 76	34, 23	1, 1
	cAIC$_k$	824.1(3.4)	772.8(2.5)	776.0(2.5)	808.5(3.0)
		0, 0	77, 81	21, 16	2, 3

Table 2.7: Simulation results comparing different error correlation structures, for m=75, n=10. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details.

	quant.	iid	ar1	ar2	comp sym
$\rho = 0.1$					
	-2l$_c$	2557.3(4.1)	2564.9(4.1)	2562.4(4.2)	2582.5(4.6)
		77	2	21	0
	cAI	2741.8(1.7)	2721.1(1.6)	2724.5(1.8)	2728.6(1.5)
		0, 1	50, 52	33, 26	17, 21
	cAIC	2711.1(4.1)	2720.7(4.1)	2720.2(4.2)	2738.2(4.6)
		86, 92	1, 1	13, 7	0, 0
	cAIC$_c$	2728.9(4.1)	2739.0(4.1)	2739.0(4.2)	2756.5(4.6)
		89, 95	1, 0	10, 5	0, 0
	cAIC$_k$	2711.1(4.1)	2720.7(4.1)	2720.2(4.2)	2760.3(5.0)
		86, 92	1, 1	13, 7	0, 0
$\rho = 0.4$					
	-2l$_c$	2476.9(4.9)	2448.5(4.4)	2447.3(4.4)	2521.0(4.7)
		1	44	55	0
	cAI	2836.1(3.5)	2612.5(1.8)	2614.3(1.9)	2768.2(3.0)
		0, 0	55, 70	45, 30	0, 0
	cAIC	2630.7(4.9)	2604.1(4.4)	2604.9(4.4)	2676.6(4.7)
		1, 2	62, 71	37, 27	0, 0
	cAIC$_c$	2648.5(4.9)	2622.3(4.4)	2623.7(4.4)	2694.8(4.7)
		1, 2	63, 73	36, 25	0, 0
	cAIC$_k$	2633.4(4.9)	2611.2(4.4)	2616.0(4.4)	2688.0(4.7)
		4, 4	74, 79	22, 17	0, 0
$\rho = 0.7$					
	-2l$_c$	2485.3(4.7)	2456.1(4.2)	2454.5(4.4)	2529.1(4.7)
		0	48	52	0
	cAI	2828.3(3.0)	2610.1(1.5)	2612.6(1.8)	2766.9(3.4)
		0, 0	52, 70	48, 30	0, 0
	cAIC	2639.1(4.7)	2611.7(4.2)	2612.1(4.4)	2684.6(4.7)
		0, 1	54, 67	46, 32	0, 0
	cAIC$_c$	2656.9(4.7)	2630.0(4.2)	2630.9(4.4)	2702.8(4.7)
		0, 1	56, 69	44, 30	0, 0
	cAIC$_k$	2641.8(4.7)	2618.8(4.2)	2623.2(4.4)	2696.4(4.6)
		2, 5	77, 83	21, 12	0, 0

Table 2.8: Simulation results comparing different error correlation structures, for m=10, n=25. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details.

	quant.	iid	ar1	ar2	comp sym
$\rho = 0.1$					
	-2l$_c$	867.9(2.6)	867.1(2.6)	866.1(2.5)	870.4(2.6)
		22	26	52	0
	cAI	897.3(0.7)	893.9(0.6)	895.1(0.7)	897.5(0.7)
		3, 14	62, 73	28, 0	7, 13
	cAIC	891.9(2.6)	893.1(2.6)	894.1(2.5)	896.4(2.6)
		76, 89	15, 6	9, 5	0, 0
	cAIC$_c$	893.2(2.6)	894.6(2.6)	895.9(2.5)	897.9(2.6)
		80, 92	11, 5	9, 3	0, 0
	cAIC$_k$	894.1(2.6)	897.8(2.6)	901.4(2.6)	949.3(3.7)
		96, 98	4, 2	0, 0	0, 0
$\rho = 0.4$					
	-2l$_c$	852.5(2.7)	824.9(2.2)	823.8(2.2)	855.7(2.7)
		0	42	58	0
	cAI	911.5(1.3)	855.3(0.8)	856.9(0.9)	903.5(1.4)
		0, 0	62, 96	38, 4	0, 0
	cAIC	876.5(2.7)	850.9(2.2)	851.8(2.2)	881.7(2.7)
		0, 1	79, 93	21, 6	0, 0
	cAIC$_c$	877.8(2.7)	852.4(2.2)	853.6(2.2)	883.3(2.7)
		0, 1	81, 94	19, 5	0, 0
	cAIC$_k$	878.7(2.7)	856.0(2.2)	859.5(2.2)	905.3(3.3)
		1, 2	94, 93	5, 5	0, 0
$\rho = 0.7$					
	-2l$_c$	818.8(3.7)	704.1(2.2)	703.3(2.2)	804.5(3.4)
		0	44	56	0
	cAI	958.5(3.6)	736.8(0.8)	737.9(1.0)	916.5(2.5)
		0, 0	57, 93	43, 7	0, 0
	cAIC	842.8(3.7)	730.1(2.2)	731.3(2.2)	830.5(3.4)
		0, 0	84, 93	16, 7	0, 0
	cAIC$_c$	844.1(3.7)	731.7(2.2)	733.1(2.2)	832.0(3.4)
		0, 0	86, 95	14, 5	0, 0
	cAIC$_k$	845.0(3.7)	736.5(2.2)	740.4(2.3)	838.2(3.6)
		0, 0	95, 97	5, 3	0, 0

Table 2.9: Simulation results comparing different error correlation structures, for m=20, n=25. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details.

	quant.	iid	ar1	ar2	comp sym
$\rho = 0.1$					
	-2l$_c$	1731.8(3.2)	1730.6(3.2)	1729.1(3.2)	1737.3(3.2)
		18	26	54	2
	cAI	1790.3(0.8)	1782.8(0.8)	1784.7(0.9)	1789.8(1.0)
		1, 4	50, 82	39, 1	10, 13
	cAIC	1775.8(3.2)	1776.6(3.2)	1777.1(3.2)	1783.3(3.2)
		60, 82	16, 10	24, 8	0, 0
	cAIC$_c$	1778.0(3.2)	1778.9(3.2)	1779.6(3.2)	1785.6(3.2)
		63, 82	17, 10	20, 8	0, 0
	cAIC$_k$	1778.1(3.2)	1781.2(3.2)	1784.4(3.2)	1834.6(4.0)
		87, 94	9, 4	4, 2	0, 0
$\rho = 0.4$					
	-2l$_c$	1722.1(3.6)	1661.2(3.1)	1660.0(3.1)	1731.4(3.7)
		0	44	56	0
	cAI	1814.5(1.7)	1704.2(0.8)	1705.6(0.8)	1800.0(1.7)
		0, 0	62, 95	38, 5	0, 0
	cAIC	1766.1(3.6)	1707.2(3.1)	1708.0(3.1)	1777.4(3.7)
		0, 0	72, 89	28, 11	0, 0
	cAIC$_c$	1768.2(3.6)	1709.5(3.1)	1710.6(3.1)	1779.7(3.7)
		0, 0	74, 90	26, 10	0, 0
	cAIC$_k$	1768.3(3.6)	1712.4(3.1)	1715.8(3.1)	1794.0(3.8)
		0, 0	91, 95	9, 5	0, 0
$\rho = 0.7$					
	-2l$_c$	1651.2(4.7)	1418.4(2.9)	1417.3(2.9)	1617.3(4.4)
		0	45	55	0
	cAI	1902.3(4.5)	1467.6(1.0)	1469.1(1.1)	1841.8(4.4)
		0, 0	64, 94	36, 6	0, 0
	cAIC	1695.2(4.7)	1464.4(2.9)	1465.3(2.9)	1663.2(4.4)
		0, 0	71, 85	29, 15	0, 0
	cAIC$_c$	1697.3(4.7)	1466.7(2.9)	1467.8(2.9)	1665.6(4.4)
		0, 0	72, 85	28, 15	0, 0
	cAIC$_k$	1697.5(4.7)	1470.8(2.9)	1474.4(2.9)	1669.2(4.4)
		0, 0	86, 98	14, 2	0, 0

Table 2.10: Simulation results comparing different error correlation structures, for m=20, n=50. AR(1) is the true model, and ρ is the AR correlation parameter. Every second row shows the number of times out of 100 a model is selected by the minimum of a criterion, and by the 'rule-of-the-thumb of difference of 2'; see the text for more details.

	quant.	iid	ar1	ar2	comp sym
$\rho = 0.1$					
	-2l$_c$	3508.7(4.7)	3502.3(4.6)	3501.0(4.6)	3514.1(4.7)
		1	41	58	0
	cAI	3558.1(0.8)	3544.8(0.7)	3546.2(0.8)	3556.8(0.8)
		0, 0	59, 99	40, 0	1, 1
	cAIC	3552.7(4.7)	3548.3(4.6)	3549.0(4.6)	3560.1(4.7)
		21, 28	55, 61	24, 11	0, 0
	cAIC$_c$	3553.8(4.7)	3549.4(4.6)	3550.2(4.6)	3561.2(4.7)
		21, 30	55, 59	24, 11	0, 0
	cAIC$_k$	3552.7(4.7)	3548.3(4.6)	3549.0(4.6)	3587.9(5.4)
		21, 28	55, 61	24, 11	0, 0
$\rho = 0.4$					
	-2l$_c$	3490.8(5.8)	3342.6(4.8)	3341.5(4.8)	3499.9(5.8)
		0	38	62	0
	cAI	3581.3(1.5)	3383.5(0.8)	3384.7(0.8)	3566.1(1.7)
		0, 0	65, 100	35, 0	0, 0
	cAIC	3534.8(5.8)	3388.6(4.8)	3389.5(4.8)	3545.9(5.8)
		0, 0	80, 94	20, 6	0, 0
	cAIC$_c$	3535.8(5.8)	3389.7(4.8)	3390.7(4.8)	3547.0(5.8)
		0, 0	80, 94	20, 6	0, 0
	cAIC$_k$	3536.9(5.8)	3393.1(4.8)	3396.3(4.8)	3560.5(5.9)
		0, 0	94, 96	6, 4	0, 0
$\rho = 0.7$					
	-2l$_c$	3413.9(7.5)	2853.2(4.2)	2852.2(4.2)	3383.3(7.5)
		0	37	63	0
	cAI	3661.6(4.9)	2896.0(1.0)	2897.1(1.0)	3594.9(4.2)
		0, 0	68, 99	32, 1	0, 0
	cAIC	3457.9(7.5)	2899.2(4.2)	2900.2(4.2)	3429.3(7.5)
		0, 0	80, 91	20, 9	0, 0
	cAIC$_c$	3458.9(7.5)	2900.4(4.2)	2901.5(4.2)	3430.4(7.5)
		0, 0	82, 91	18, 9	0, 0
	cAIC$_k$	3460.0(7.5)	2904.4(4.2)	2907.7(4.2)	3435.4(7.5)
		0, 0	91, 99	9, 1	0, 0

Table 2.11: Parameters estimates (standard errors) and model selection for the FEV_1 data under three different error structures.

Parameters	Independent	AR(1)	AR(2)
Intercept	0.515 (0.284)	0.514 (0.285)	0.516 (0.284)
Baseline FEV_1	0.903 (0.101)	0.895 (0.100)	0.892 (0.100)
Hour	-0.018 (0.008)	-0.016 (0.012)	-0.015 (0.012)
Treatment A	0.603 (0.140)	0.633 (0.149)	0.656 (0.152)
Treatment B	0.943 (0.140)	0.926 (0.149)	0.927 (0.152)
Hour : treatment A	-0.071 (0.011)	-0.073 (0.016)	-0.075 (0.018)
Hour : treatment B	-0.097 (0.011)	-0.090 (0.016)	-0.089 (0.018)
τ (random intercept)	0.454 (0.044)	0.429 (0.048)	0.390 (0.067)
σ (error SD)	0.253 (0.008)	0.292 (0.019)	0.341 (0.052)
ρ_1	–	0.546 (0.051)	0.544 (0.008)
ρ_2	–	–	0.184 (0.063)
cAIC	130.1	125.8	178.2
$cAIC_c$	153.4	149.4	201.1
$cAIC_k$	130.8	130.5	171.9

Chapter 3

A Comparison of Smoothing Via L_1 and L_2 Penalization with Application to Group fMRI Data

3.1 Introduction

Clustered data, such as that arising from repeated measurements on several subjects, is often analyzed with a mixed effect model. In a linear mixed effect model, the population, or marginal, mean is modeled by a linear combination of covariates; the coefficients of these covariates are called 'fixed effects'. The mean for a given cluster, or conditional mean, is modeled by the fixed effects plus a linear combination of covariates; the coefficients corresponding to these additional covariates are called 'random effects' and are specific to each cluster in the model. The random effects are assumed to come from some distribution, typically normal distribution with mean 0, and induce correlation within each cluster in the marginal model. In the general linear mixed effect model, the within cluster correlation is additionally modeled through a parametric correlation structure for the errors.

In this chapter, we consider a variation of the general linear mixed effect model where the parametric model of the marginal mean is replaced by a non-parametric model. As in a standard general linear mixed effect model, the correlation structure

is modeled through a combination of random effects and an error correlation structure. We chose to use splines to obtain a non-parametric estimate of the marginal mean since, as noted by Speed in [62], a spline model can be written as a mixed effect model and thus we can write the combined spline and mixed effect model as a single general linear mixed effects models. Two types of splines are considered, regression splines and penalized splines (p-splines). We start with a large number of basis function in either method and use penalization to prevent over-fitting. The L_2 penalty is used for p-splines while an L_1 penalty is used to select knot locations for the regression splines.

3.1.1 L_1 penalization

The LASSO (least absolute shrinkage and selection operator) was introduced by Tibshirani [65] for the purposes of estimation and variable selection in linear regression. In the usual linear regression model with centered covariates $y = X\beta + \epsilon$ the parameter β is estimated by minimizing the residual sum of squares and an L_1 penalty term:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ (y - X\beta)^\top (y - X\beta) + \lambda \sum_{j=1}^{s} |\beta_j| \right\}.$$

Model selection is performed by the choice of the parameter λ in the penalty term: as λ increases from zero, the number of covariates in the model may change, since $\hat{\beta}_j$'s are allowed to be set to exactly zero. A discussion of the LASSO in comparison with other shrinkage techniques is presented by Hastie, Tibshirani and Friedman [30]. Knight and Fu [40] studied the asymptotics of LASSO estimates in the context of linear regression with iid errors.

While much of the theory of the LASSO is derived under the assumption of a parametric model, the LASSO has been applied to non-parametric problems: the problem of knot selection in regression splines by Osborne, Presnell and Turlach [54] is one example and some theory is presented in Bühlmann and van de Geer [6].

Limited results have been obtained for correlated data. In a doctoral dissertation, Gupta [27] studied the asymptotics of LASSO estimates in the presence of correlated errors and a large number of covariates. In particular, results were obtained for the

following estimate of β:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ (y - X\beta)^T \hat{\Sigma}^{-1} (y - X\beta) + \lambda \sum_{j=1}^{s} |\beta_j| \right\}$$

where the errors ϵ are assumed to be $N(0, \Sigma)$ and $\hat{\Sigma}$ is assumed to be consistent estimator of Σ. Wang, Li and Tsai [69] study the problem of joint selection of covariates and type of autoregressive process via the LASSO for linear regression with autoregressive errors.

3.1.2 L_2 penalization

An overview of p-splines can be found in [57]; here, the L_2 penalization is used with the truncated power basis. P-splines with difference penalties and the B-spline basis are introduced by Marx and Eilers in [18]. Asymptotics were first considered in [28]; here, expressions were derived for asymptotic variance and bias in the case of independent errors and uniformly dense knots. Later references are [70, 39, 47, 10]. While the results in these papers extend [28] to explicitly account for the number of knots, non-normal outcomes, and make connections with kernel smoothing, there is limited work on the asymptotics of p-splines with correlated errors. Smoothing parameter selection when the error structure is mispecified is covered in [41].

3.2 Penalized Mixed Effect Models

Two applications of L_1 penalization for the selection of both random and fixed effects have recently been proposed. Ibrahim et al. [36] used one penalty to select fixed effects and another penalty on the cholesky decomposition of the random effects covariance matrix to select random effects. Their method extends to generalized linear mixed models. Bondell, Krishna and Ghosh [3] purposed methods similar to that of Ibrahim et al [36]. In both works, the variance components of the model are estimated using REML and the number of covariates is less than the number of datapoints. Most recently, Schelldorfer, Bühlmann and van de Geer [58] proposed L_1 penalization for linear mixed models when the number of covariates to be much larger than the number of observations. Their verison is appropriate for situations where it is known which covariates will be used as random effects, such as the random intercept model.

Smoothing splines with with cubic B-spline basis have been proposed by Brumback and Rice in [5] for the analysis of nested curves. Their model uses a smoothing spline to fit the marginal mean as well as cluster-specific deviations and may be written as a linear mixed effect model where both the fixed and random effect covariates are basis functions.

We first describe our approach to estimate single mean curve in the presence of random effects and correlated errors.

3.2.1 Notation

Let y_{ik} be the measurement at time t_k in cluster i for $i = 1, \ldots, m$ and $k = 1, \ldots, n_i$. Assume that each measurement follows the model

$$y_{ik} = \mu(t_k) + z_{ik} b_i + \epsilon_{ik}, \tag{3.1}$$

where $\mu(t)$ is a smooth function of t, $z_{ik} b_i$ is a cluster specific deviation from $\mu(t_k)$ and $\epsilon_i = \text{stack}(\epsilon_{i1}, \ldots, \epsilon_{in_i})$ is distributed $\text{N}(0, V_i(\sigma_v))$ for some correlation matrix $V_i(\sigma_v)$ known up to a $s \times 1$ vector of parameters σ_v. We assume that the b_i's are $r \times 1$ vectors of random effects independently drawn from $\text{N}(0, D_1(\sigma_d))$ for $i = 1, \ldots, m$, each associated with the $1 \times r$ covariate vector z_{ik}. Let Z_i be the matrix with rows z_{ik}.

For simplicity, we consider the case where $k = 1, \ldots, n_i = n$ for all $i = 1, \ldots, m$.

3.2.2 Penalized spline fit of the marginal mean

We estimate the function μ over the interval $[t_1, t_n]$ by a linear combination of basis functions. One might consider the use of the B-spline basis here; these are often preferred for their numerical stability and compact support [19]. For the L_1 penalty, the minimal amount of overlapping support is actually a disadvantage: when one spline is removed from the set via a zero coefficient then the entire basis must be re-calculated, otherwise a 'dip' will be appear in the estimated function. For L_2 penalization, either set of basis functions may be used; often B-splines are preferred for the aforementioned reasons. We stick with truncated basis functions for ease of presentation. With either

basis, the number and position of knots are essential for obtaining a good estimate of μ: we start with a large Q and evenly space the knots in the interval $[t_1, t_n]$: from this initial set of basis functions we simultaneously estimate μ and reduce the number of knots used by placing an L_1 penalty on the elements of β that correspond to basis functions with knots. Alternately, in the p-spline approach, we keep all knots but penalize the coefficients of basis functions that correspond to knot locations with an L_2 penalty.

We approximate μ by a linear combination of basis functions. We present the model in terms of the truncated cubic spline basis, $\{1, t, t^2, t^3, (t-\tau_1)^3_+, \ldots, (t-\tau_Q)^3_+\}$ where $(x)_+$ is x if $x > 0$ and 0 otherwise. The points $\tau_1, \ldots \tau_Q$ are called knots. For each i, denote $y_i = (y_{i1}, \ldots, y_{in})^\top$ and

$$X_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & t_{i1}^3 & (t_{i1}-\tau_1)^3_+ & \ldots & (t_{i1}-\tau_Q)^3_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & t_{in} & t_{in}^2 & t_{in}^3 & (t_{in}-\tau_1)^3_+ & \ldots & (t_{in}-\tau_Q)^3_+ \end{bmatrix}.$$

Our model can be written in matrix form as $y = X\beta + Zb + \epsilon$ where $X =$ stack$(X_1, \ldots, X_m)$, $\beta =$ stack$(\beta_0, \ldots, \beta_{S+3})$, $Z =$ diag$(Z_1, \ldots Z_m)$, $b =$ stack$(b_1, \ldots, b_m)$, $\epsilon =$ stack$(\epsilon_1, \ldots, \epsilon_m)$ and $y =$ stack$(y_1, \ldots, y_m)$. To make the distinction between unpenalized and penalized elements clear, we partition β as $\beta =$ stack(β_u, β_p) and $X = (X_u, X_p)$ so that $X\beta = X_u\beta_u + X_p\beta_p$. Let $D =$ diag$(D_1(\sigma_d), \ldots, D_1(\sigma_d))$, $V(\sigma_v) =$ diag$(V_1(\sigma_v), \ldots, V_n(\sigma_v))$ and $\Sigma = V + ZDZ^\top$. Since V, Z and D are all block diagonal matrices, Σ is as well. Denote the ith block of Σ, which corresponds to cluster i, by Σ_i.

Let $\mu = stack(\mu(t_1), \ldots, \mu(t_n))$. Under the model in (1), the log likelihood of y is (up to a constant)

$$l(y|\mu, \Sigma) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y-\mu)^\top\Sigma^{-1}(y-\mu). \tag{3.2}$$

We approximate μ by $X\beta$, a linear combination of basis functions, and choose $\hat{\mu} = X\hat{\beta}$ by replacing μ with $X\beta$ in (2) and penalizing the coefficients of basis functions that correspond to knot locations:

$$(\hat{\beta}, \hat{\Sigma}) = \underset{\beta,\Sigma}{\operatorname{argmin}} \left\{ \frac{1}{2}\log|\Sigma| + \frac{1}{2}(y-X\beta)^T\Sigma^{-1}(y-X\beta) + p_\lambda(\beta_p) \right\}. \tag{3.3}$$

If Σ was known and the L_1 penalty used, $p^1_\lambda(\beta_p) = \lambda \sum_{q=1}^{Q} |\beta_{pq}|$, then this procedure would be exactly that of Osbourne, Presnell and Turlach used to select knot locations for regression splines[54]. If the L_2 penalty were used, $p^2_\lambda(\beta_p) = \lambda \sum_{q=1}^{Q} (\beta_{pq})^2$, and Σ known, then this method would result in a p-spline estimate of $\mu(t)$ in the style of [57]. Once the estimates of μ and Σ are obtained, predictions of the random effects b can be by maximizing the 'joint' likelihood of μ and b with μ replaced by $X\beta$. This maximization results in

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

We choose λ to minimize either BIC, as in Schelldorfer, Bühlmann and van de Geer [58]:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \left[l(y|X\beta) + \log(nm) DF \right]$$

or AIC. For either AIC or BIC, the degrees of freedom (DF) depends on the penalty used: for the L_1 penalty, the degrees of freedom 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 covariance parameters. The hat matrix H is such that $\hat{y} = X\hat{\beta} + Z\hat{b} = Hy$.

3.2.3 Standard errors

Standard errors via the Sandwich Estimator

In [20], Fan and Li proposed a sandwich estimator to obtain standard errors for nonzero elements of $\hat{\beta}$ when penalization is used to estimate β in the usual linear regression setting. It was later shown that this estimator is consistent when the number of parameters grows at a certain rate with the sample size [21]. The sandwich estimator for penalized likelihoods has an additional term in the second derivative due to the penalty.

Denote the vector of all non-zero elements of $\hat{\beta}$ and the parameters $\sigma = \text{stack}(\sigma_v, \sigma_d)$ in Σ by θ. Let ∇l and $\nabla^2 l$ be the first and second derivatives of l with respect to θ and $p_\lambda(\theta)$ be the penalty term. The sandwich estimate for the covariance of $\hat{\theta}$ in [20] is

$$\widehat{\text{cov}}(\hat{\theta}) = m\{\nabla^2 l(\hat{\theta}) - \Lambda\}^{-1} \widehat{\text{cov}}\{\nabla l(\hat{\theta})\} \{\nabla^2 l(\hat{\theta}) - \Lambda\}^{-1} \tag{3.4}$$

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

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

We follow [21] by using a local quadratic approximation to approximate $\partial^2 p_\lambda(\theta)/\partial\beta_k\partial\beta_k$ with $1/|\beta_k|$ when $p_\lambda(\theta)$ is the L_1 penalty. Differentiating the log-likelihood (2) leads to the following formulas for ∇l and $\nabla^2 l$:

$$\begin{aligned}
\frac{\partial l}{\partial\beta} &= -\sum_{i=1}^{m} X_i^\top\Sigma_i^{-1}(y_i - X_i\beta)\\
\frac{\partial l}{\partial\sigma_k} &= -\frac{1}{2}\left\{m\text{tr}\left(\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\sigma_k}\right)-\sum_{i=1}^{m}(y_i - X_i\beta)^T\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\sigma_k}\Sigma_i^{-1}(y_i - X_i\beta)\right\}\\
\frac{\partial^2 l}{\partial\beta\partial\beta^T} &= \sum_{i=1}^{m} X_i^\top\Sigma_i^{-1}X_i\\
\frac{\partial^2 l}{\partial\beta\partial\sigma_k} &= \sum_{i=1}^{m} X_i^\top\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\sigma_k}\Sigma_i^{-1}(y_i - X_i\beta)
\end{aligned}$$

and

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

Once $\widehat{\text{cov}}(\hat{\beta})$ is obtained, a 95% confidence interval for $\mu(t_k)$ may be formed by $\hat{\mu}(t_k)\pm 1.96\sigma_k$ where σ_k is the kth element of $\sqrt{\text{diag}(X\widehat{\text{cov}}(\hat{\beta})X^\top)}$.

Parametric bootstrap

In addition to the sandwich estimator, we also considered confidence intervals for $\mu(t_k)$ based on a parametric bootstrap. We generated $\hat{\mu}^*(t)$ as follows:

1. draw m b_i^*'s from $\text{N}(0,\hat{D}_1^2)$,

2. draw m ϵ_i^*'s from $\text{N}(0,\hat{V}_i)$,

3. form $y_i^* = X_i\hat{\beta} + Z_i b_i^* + \epsilon_i^*$ for $i = 1, \ldots, m$,

4. obtain $\hat{\mu}^*(t)$ via equation (3).

Once B such $\hat{\mu}^*(t)$ are obtained, we created a percentile 95% confidence interval by taking the 2.5th and 97.5th highest $\hat{\mu}^*(t)$'s.

We also considered using estimates from a full model in the above procedure. In a small number of simulations, both bootstrap procedures worked well; in practice the amount of computation was undesirable and so we did not pursue these methods further.

3.2.4 Computational aspects

The criterion may be non-convex in β and the parameters in Σ. Following Wang, Li and Tsai [69] but breaking up β into its penalized and unpenalized components, we iterate between

$$(\beta_p^{j+1}|\beta_u^j, \Sigma^j) = \underset{\beta_p}{\operatorname{argmin}} \left\{ (y - X_u\beta_u^j - X_p\beta_p)^\top (\Sigma^j)^{-1} (y - X_u\beta_u^j - X_p\beta_p) + p_\lambda(\beta_p) \right\} \tag{3.5}$$

and

$$\begin{aligned} (\Sigma^{j+1}, \beta_u^{j+1}|\beta_p) \quad = \quad & \underset{\Sigma, \beta_u}{\operatorname{argmin}} \left\{ \frac{1}{2} \log|\Sigma| + \frac{1}{2} (y - X_u\beta_u - X_p\beta_p^j)^\top \Sigma^{-1} \right. \\ & \left. \times (y - X_u\beta_u - X_p\beta_p^j) \right. \end{aligned} \tag{3.6}$$

(3.8)

until convergence. In each step, β_p^j may be obtained from 3.5 using any of the standard methods for LASSO in linear regression models (LARS, cyclic or greedy coordinate descent, homotopy) when the L_1 penalty $p_\lambda^1(\beta_p)$ is used. We found that the 'lars' function in the R package 'lars' worked well. When the L_2 penalty $p_\lambda^2(\beta_p)$ is used, this step is simply ridge regression. Here we used the function 'lm.ridge' from the 'MASS' R package. The covariance parameters in Σ may be obtained from 3.7 by numerical optimization. By noting that 3.7 is equivalent to fitting a general linear mixed model on $y^{**} = y - X_p\beta_p^j$, one can use any standard software for maximum likelihood estimation

of such models. We used the function 'lme' in the R package 'nlme'. An alternative to iteration between 3.5 and 3.7 would be an EM algorithm as in Ibrahim et al, [36] and Bondell, Krishna and Ghosh [3].

3.3 Simulations

For each of 100 simulations, we generated $m = 5$, 10 or 40 curves, each of length $n_i = 70$ points. The generating model had either an overall mean of $\mu(t) = \sin(10\pi t/70)$ ('sine') or a model of the hemodynamic response ('fMRI'). We used the R package *neuRosim* to generate the latter curve. This package allows for three models of hemodynamic response function and various types of noise. We used the double gamma model of hemodynamic response:

$$h(t) = \left(\frac{t}{d_1}\right)^{a_1} \exp\left(-\frac{t-d_1}{b_1}\right) - c\left(\frac{t}{d_2}\right)^{a_2} \exp\left(-\frac{t-d_2}{b_2}\right) \tag{3.9}$$

with the default setting for the parameters: $a_1 = 6$, $a_2 = 12$, $b_1 = b_2 = 0.9$, $d_i = a_i b_i$, $c = 0.35$. We assume an experiment design of 10 seconds of rest, followed by 30 seconds of an activity, followed by 30 seconds of rest, with a TR of 1. The assumed signal for this setting is shown in the right of Figure 3.1. Data generated from either the sine or fMRI curve had within curve errors ϵ_i from $\mathrm{N}(0, \mathrm{R}_1(0.4))$ and and $R_1(0.4)$ is the autocorrelation matrix of a first order autoregressive process with parameter 0.4.

For each simulation set, we fit three models of the mean curve: a regression spline model with L_1 penalization for knot selection ('L_1'), a penalized spline with knots at every other other time point ('L_2 by 2') and a penalized spline with knots at every time point ('L_2'). For each model, the smoothing parameter was selected to minimize AIC, BIC with the sample size as the number of clusters ('BIC (m)') or BIC with the sample size as the number of observations ('BIC (nm)'). The R package 'lars' was used when fitting the regression spline; for both the regression and penalized splines the covariance parameters were found using the function 'gls' from the R package 'nlme'. The criterion used to determine convergence was sum of absolute values of differences in parameters less than 10^{-4}.

Tables 3.1 and 3.2 and shows the mean (se) of the covariance parameters from the 100 simulation runs while Table 3.3 has the mean integrated squared error (se).

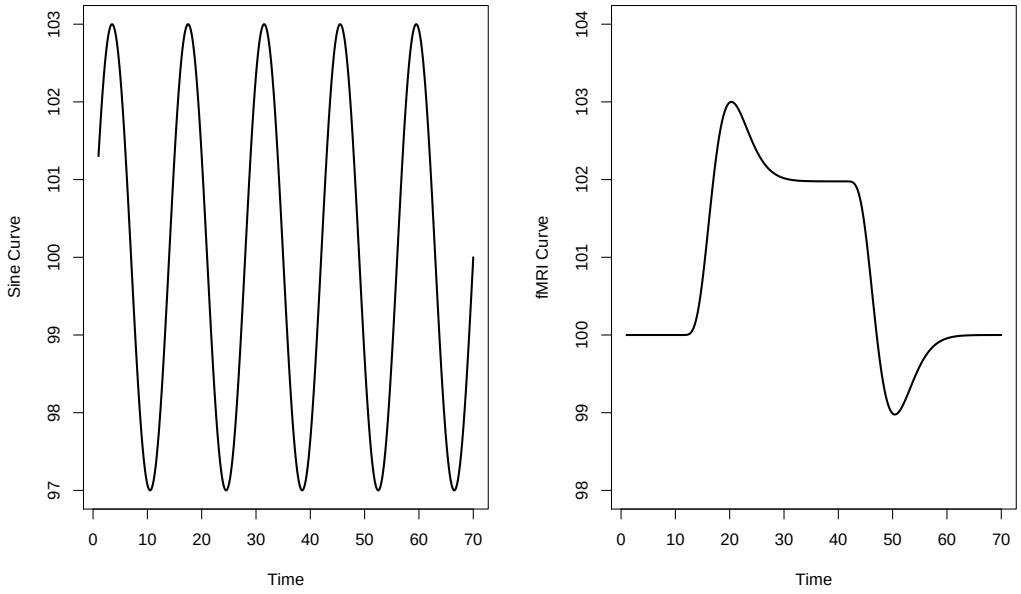

Figure 3.1: True Mean Curves from Simulation

Table 3.1: Within group correlation $\rho = 0.4$

	Sine			**fMRI**		
	AIC	BIC(m)	BIC(nm)	AIC	BIC(m)	BIC(nm)
$m = 5$						
L1	0.337 (0.005)	0.337 (0.005)	0.340 (0.005)	0.337 (0.006)	0.341 (0.006)	0.358 (0.006)
L2 (by 2)	0.335 (0.005)	0.335 (0.005)	0.335 (0.005)	0.335 (0.005)	0.335 (0.005)	0.386 (0.011)
L2	0.388 (0.005)	0.389 (0.005)	0.420 (0.005)	0.382 (0.005)	0.384 (0.005)	0.425 (0.007)
$m = 10$						
L1	0.371 (0.003)	0.371 (0.003)	0.371 (0.003)	0.366 (0.003)	0.367 (0.003)	0.377 (0.003)
L2 (by 2)	0.370 (0.003)	0.370 (0.003)	0.370 (0.003)	0.370 (0.003)	0.370 (0.003)	0.379 (0.004)
L2	0.394 (0.003)	0.392 (0.003)	0.407 (0.004)	0.392 (0.003)	0.391 (0.003)	0.406 (0.003)
$m = 40$						
L1	0.395 (0.002)	0.394 (0.002)	0.394 (0.002)	0.394 (0.002)	0.394 (0.002)	0.396 (0.002)
L2 (by 2)	0.370 (0.003)	0.370 (0.003)	0.370 (0.003)	0.391 (0.002)	0.392 (0.002)	0.395 (0.002)
L2	0.394 (0.003)	0.392 (0.003)	0.407 (0.004)	0.398 (0.002)	0.398 (0.002)	0.401 (0.002)

Table 3.2: Error variance $\sigma^2 = 0.09$

	Sine			**fMRI**		
	AIC	BIC(m)	BIC(nm)	AIC	BIC(m)	BIC(nm)
$m = 5$						
L1	0.079 (0.001)	0.079 (0.001)	0.081 (0.001)	0.079 (0.001)	0.077 (0.001)	0.084 (0.001)
L2 (by 2)	0.076 (0.001)	0.076 (0.001)	0.076 (0.001)	0.076 (0.001)	0.076 (0.001)	0.090 (0.003)
L2	0.072 (0.001)	0.072 (0.001)	0.087 (0.001)	0.074 (0.001)	0.073 (0.001)	0.091 (0.002)
$m = 10$						
L1	0.084 (0.001)	0.084 (0.001)	0.086 (0.001)	0.085 (0.001)	0.085 (0.001)	0.088 (0.001)
L2 (by 2)	0.084 (0.001)	0.084 (0.001)	0.084 (0.001)	0.084 (0.001)	0.084 (0.001)	0.086 (0.001)
L2	0.083 (0.001)	0.083 (0.001)	0.089 (0.001)	0.083 (0.001)	0.084 (0.001)	0.089 (0.001)
$m = 40$						
L1	0.089 (0.000)	0.089 (0.000)	0.089 (0.000)	0.089 (0.000)	0.089 (0.000)	0.090 (0.000)
L2 (by 2)	0.084 (0.001)	0.084 (0.001)	0.084 (0.001)	0.088 (0.000)	0.088 (0.000)	0.089 (0.000)
L2	0.083 (0.001)	0.083 (0.001)	0.089 (0.001)	0.089 (0.000)	0.089 (0.000)	0.090 (0.000)

Table 3.3: MISE (se) = mean integrated square error

	Sine			**fMRI**		
	AIC	BIC(m)	BIC(nm)	AIC	BIC(m)	BIC(nm)
$m = 5$						
L1	0.744 (0.022)	0.741 (0.022)	0.786 (0.025)	0.726 (0.023)	0.784 (0.026)	0.855 (0.030)
L2 (by 2)	0.906 (0.023)	0.906 (0.023)	0.905 (0.023)	0.902 (0.023)	0.902 (0.023)	1.756 (0.139)
L2	2.100 (0.149)	2.101 (0.149)	2.586 (0.146)	2.062 (0.148)	2.085 (0.149)	2.641 (0.151)
$m = 10$						
L1	0.404 (0.012)	0.402 (0.012)	0.415 (0.013)	0.377 (0.012)	0.375 (0.012)	0.482 (0.017)
L2 (by 2)	0.454 (0.013)	0.454 (0.013)	0.453 (0.013)	0.454 (0.013)	0.454 (0.013)	0.614 (0.035)
L2	1.401 (0.120)	1.399 (0.119)	1.626 (0.118)	1.405 (0.119)	1.406 (0.118)	1.587 (0.115)
$m = 40$						
L1	0.111 (0.003)	0.112 (0.003)	0.113 (0.003)	0.105 (0.003)	0.110 (0.003)	0.142 (0.005)
L2 (by 2)	0.454 (0.013)	0.454 (0.013)	0.453 (0.013)	0.113 (0.004)	0.120 (0.003)	0.167 (0.002)
L2	1.401 (0.120)	1.399 (0.119)	1.626 (0.118)	0.318 (0.029)	0.336 (0.029)	0.377 (0.029)

3.4 fMRI data

3.4.1 Model of hemodynamic response

The hemodynamic response $h_i(t_k)$ to brain activation, observed in a voxel at time t_k, may be assumed to have some common form:

$$h_i(t_k) = a_i + b_i h(t_k) + \epsilon_{ik} \tag{3.10}$$

where a_i and b_i are constants, $h(t)$ is a particular function, and ϵ_{it} are errors that are correlated over time. Typically the baseline activation a_i varies by subject and so, to facilitate between-subject comparisons, it is removed by modeling the percent change $PC_i(t) = (h_i(t) - a_i)/a_i$. In practice, this is accomplished by dividing the observed activation at each measurement time by $\hat{a}$, where $\hat{a}$ is some estimate of the baseline activation (either the mean observed time series or the mean of the observed time series during 'rest' blocks). The model may be expanded to include allow other types of subject specific deviations from $h(t)$. For example, a time shift c_i may be included via $h(t + c_i) \approx h(t) + h'(t)c_i$.

3.4.2 Caffeine dataset

Functional MRI is a popular method of estimating brain activity by measuring blood flow to the brain. 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 (Rack-Gomer, Liau, and Liu, [56]). The study had 11 subjects, but 2 were dropped due to head movement during the scans. A block design was used to observe fingertapping: after an initial period of 20 seconds, the subjects were told to alternate fingertapping (30 seconds) and not fingertapping (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. These 156 points of 2 second intervals will be referred to as the 'time' variable. The block design was performed twice for each subject, once for a 'pre-caffeine' session and again after ingested 200 mg of caffeine (the 'post-caffeine' session). During fingertapping periods, some of the voxels in the motor-cortex region of the brain become 'activated'

- more oxygen was sent to this part of the brain. Following standard preprocessing techinques in the field, the voxels in the motor cortex for each subject were selected that were common to both the pre and post caffiene sessions.

3.4.3 Model of pre and post caffeine sessions

Let y_{ivjk} be the BOLD signal at time k, during session j, in voxel v for subject i for $i = 1, \ldots, m$, $j = 1, 2$ and $t = 1, \ldots, 156$ and $v = 1, \ldots, n_{vi}$ Assume that each measurement follows the model

$$y_{ivjk} = \beta_{ij} 1_{[i \neq 1]} + \mu_j(t_k) + b_{ivj} + \epsilon_{ivjk} \tag{3.11}$$

where $\mu_j(t)$ is smooth functions of t, $\epsilon_{ivj} = (\epsilon_{ivj1}, \ldots, \epsilon_{ivjn})^T$ is distributed $N(0, \sigma^2 R_1)$ for some correlation matrix R_1 (known up to a vector of parameters ρ) and b_{iv} is independently distributed $N(0, D_1)$ for each iv.

3.4.4 Results

In the analysis of the pre and post-caffeine scans over all subjects for the finger-tapping sessions, Rack-Gomer, Liau, and Liu [56] compared four measures. They found a significant difference in 1) time to reach 50% of peak response and 2) time to fall to 50% of peak response but not a significant difference in 3) the full width-half maximum (difference between the previous two times) or 4) the maximum amplitude of the BOLD response.

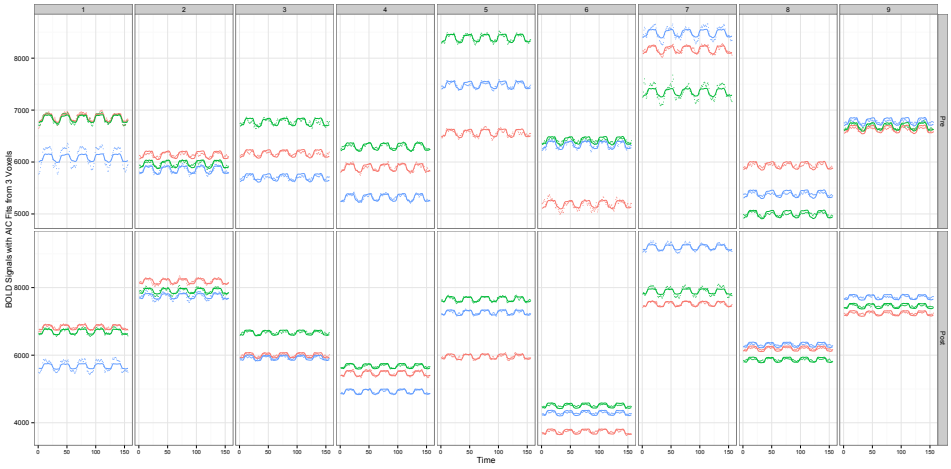

Figure 3.2: Estimated (solid) and observed (dots) BOLD signals from three voxels per subject, by session. The estimated curves were obtained via L_2 penalization with smoothing parameter selection by AIC. The fits were obtained using the entire dataset; only the results for three voxels are shown for the purpose of display.

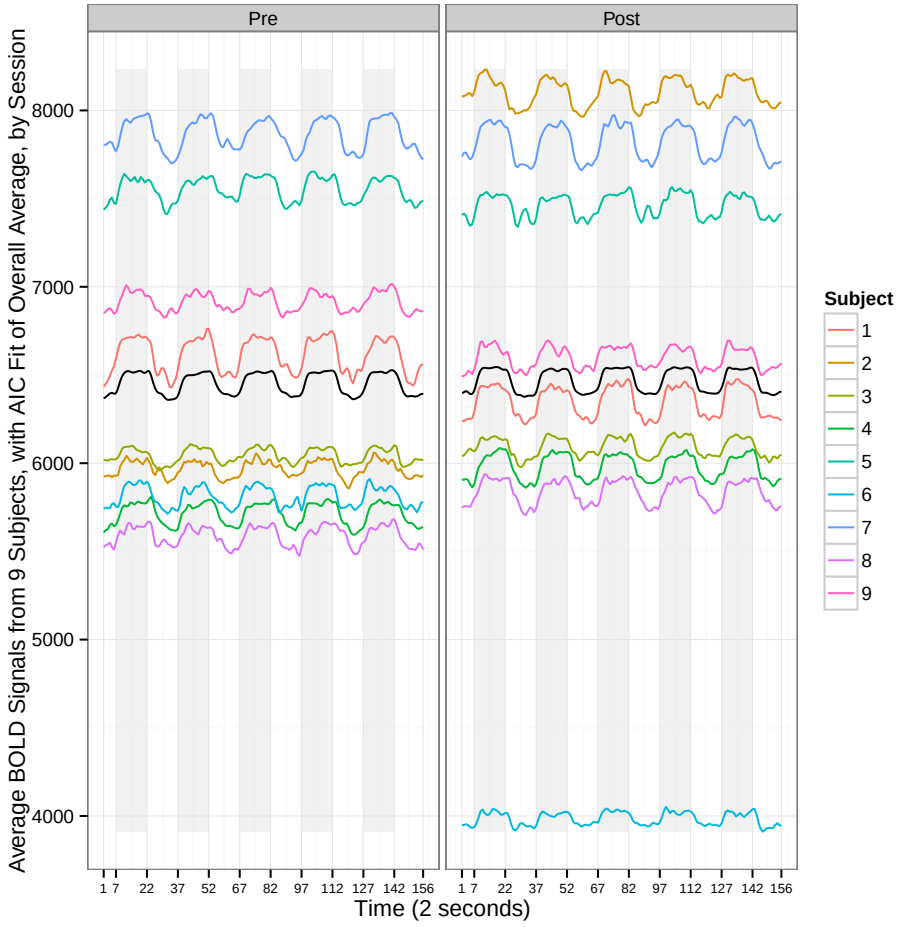

Figure 3.3: Estimated average BOLD signal by session (black) with observed average BOLD signals for 9 subjects, by session (colors). The estimated curves were obtained via L_2 penalization with smoothing parameter selection by AIC.

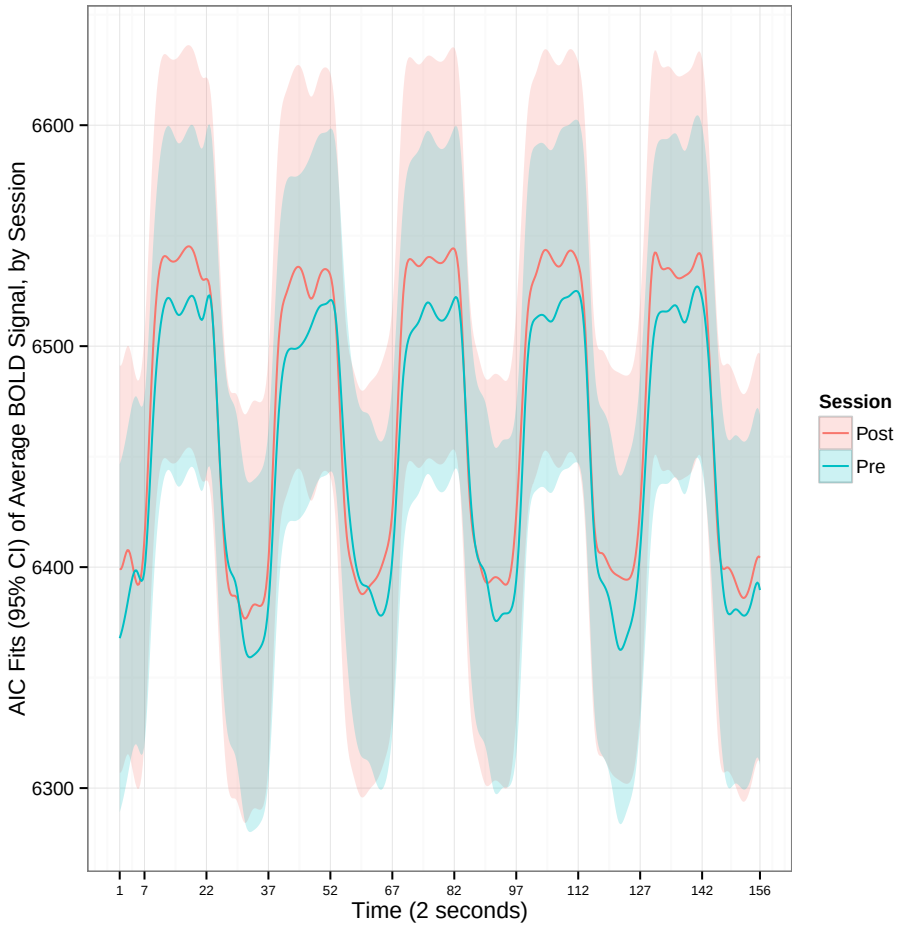

Figure 3.4: Pre (red) and Post (blue) caffeine session estimated average BOLD signals with pointwise 95% confidence intervals (shading). The estimated curves were obtained via L_2 penalization with smoothing parameter selection by AIC while the confidence intervals were computed using the sandwich estimator.

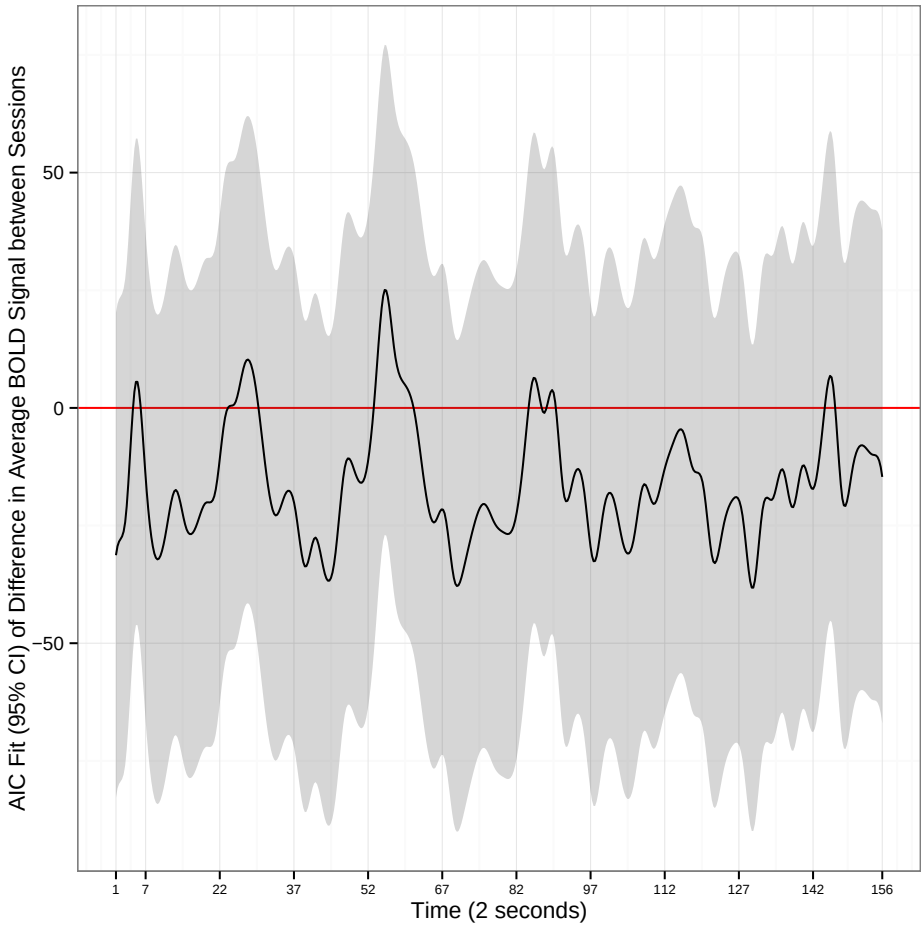

Figure 3.5: Estimated (black) difference in BOLD signals between sessions with pointwise 95% confidence intervals (shading). The difference is computed as the pre-caffeine signal minus the post-caffeine signal. The estimated curve was obtained via L_2 penalization with smoothing parameter selection by AIC while the confidence intervals were computed with the sandwich estimator.

3.5 Discussion

In this paper, we have combined several methods to obtain a non-parametric estimate of a mean from correlated data. The resulting model can easily be fit using existing software.

3.6 Acknowledgements

The contents of Chapter 3 are being prepared for submission as "A Comparison of L_1 and L_2 Smoothing with Application to Group fMRI Data" by Rosanna Overholser and Ronghui Xu. I thank Thomas Liu and Anna Leigh Rack-Gomer of the UCSD Center for fMRI for generously sharing an fMRI dataset.

Appendix A

fMRI Plots

Table A.1: Number of motor cortex voxels per subject.

Subject	1	2	3	4	5	6	7	8	9
Number of Voxels	77	35	73	124	78	106	67	89	72

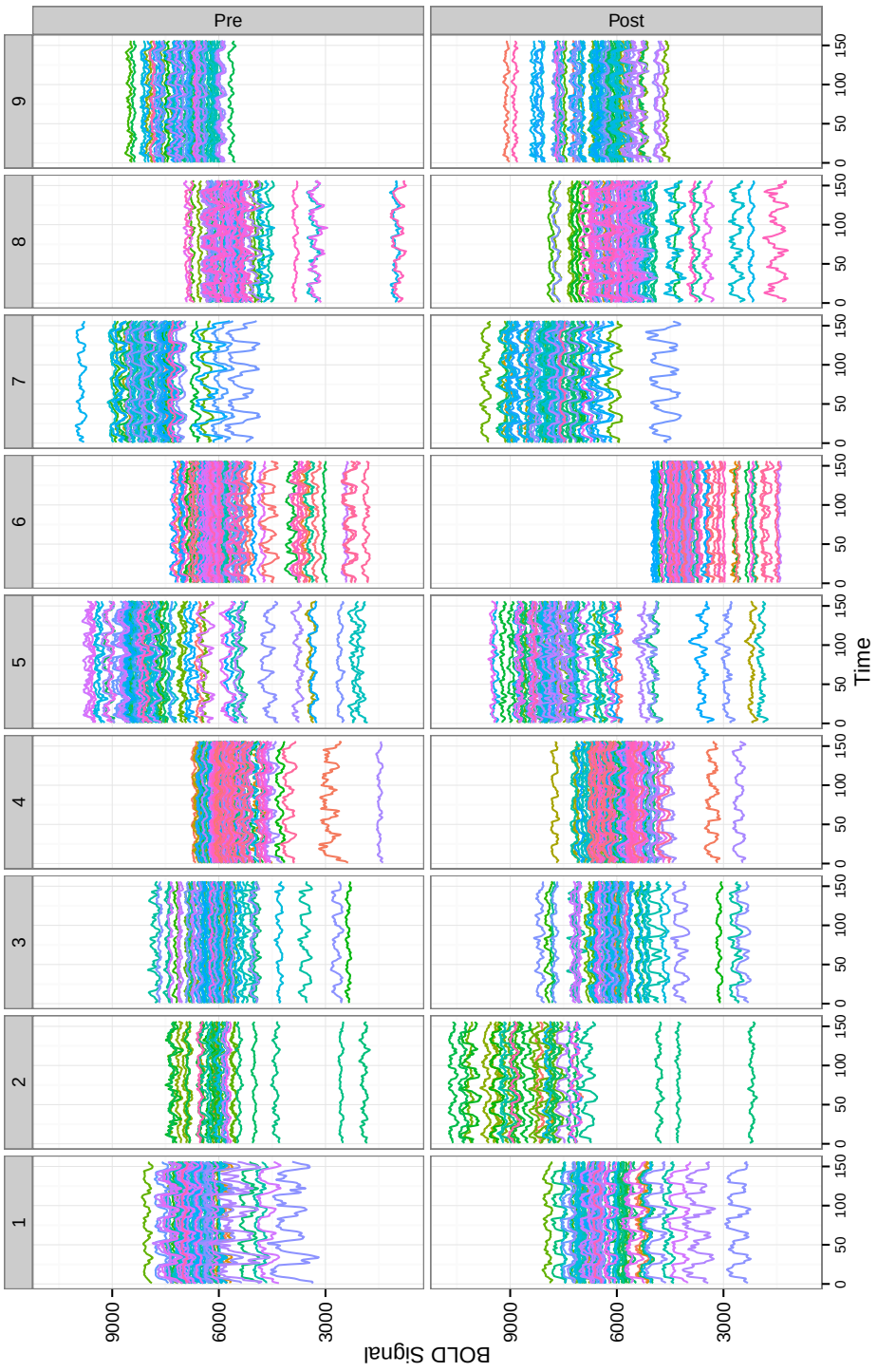

Figure A.1: BOLD signals from the motor cortex of 9 subjects, before and after caffeine. Timeseries of the same color within a subject are from the same voxel.

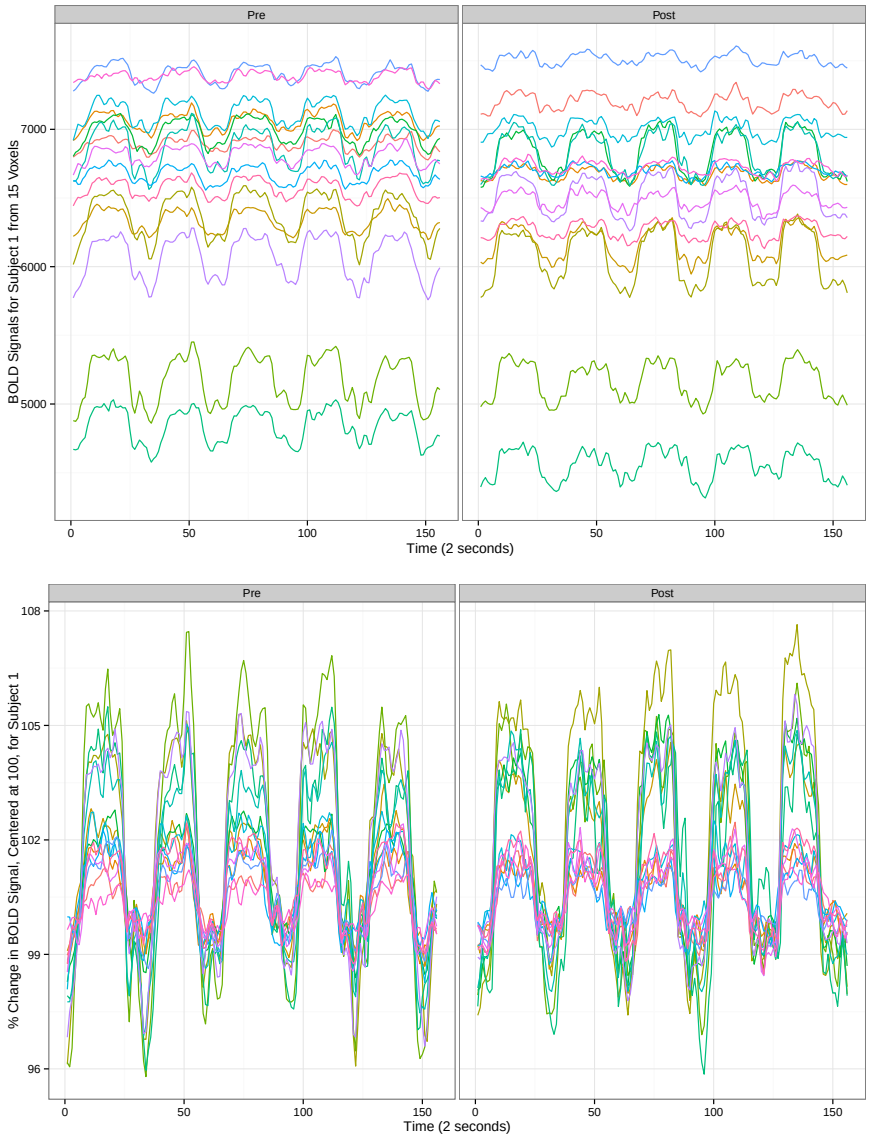

Figure A.2: Randomly selected BOLD signals from the motor cortex of Subject 1, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel.

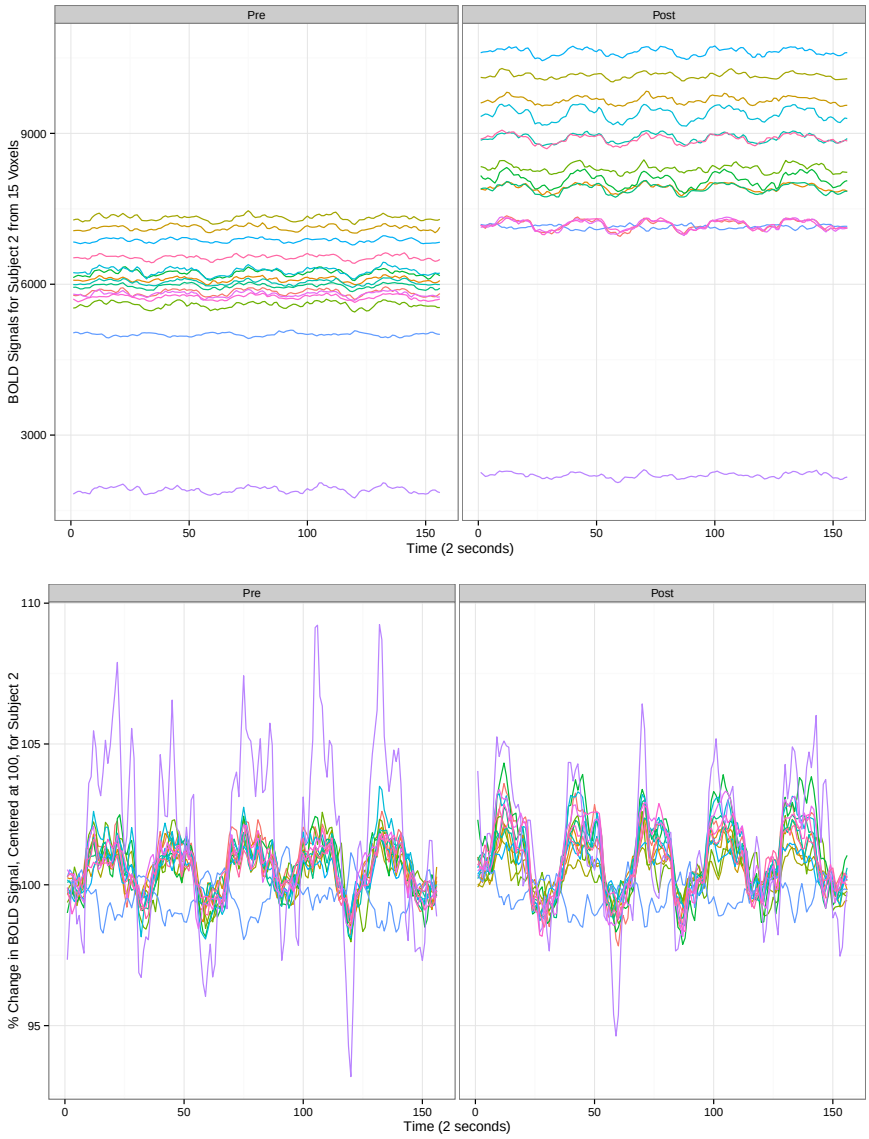

Figure A.3: Randomly selected BOLD signals from the motor cortex of Subject 2, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel.

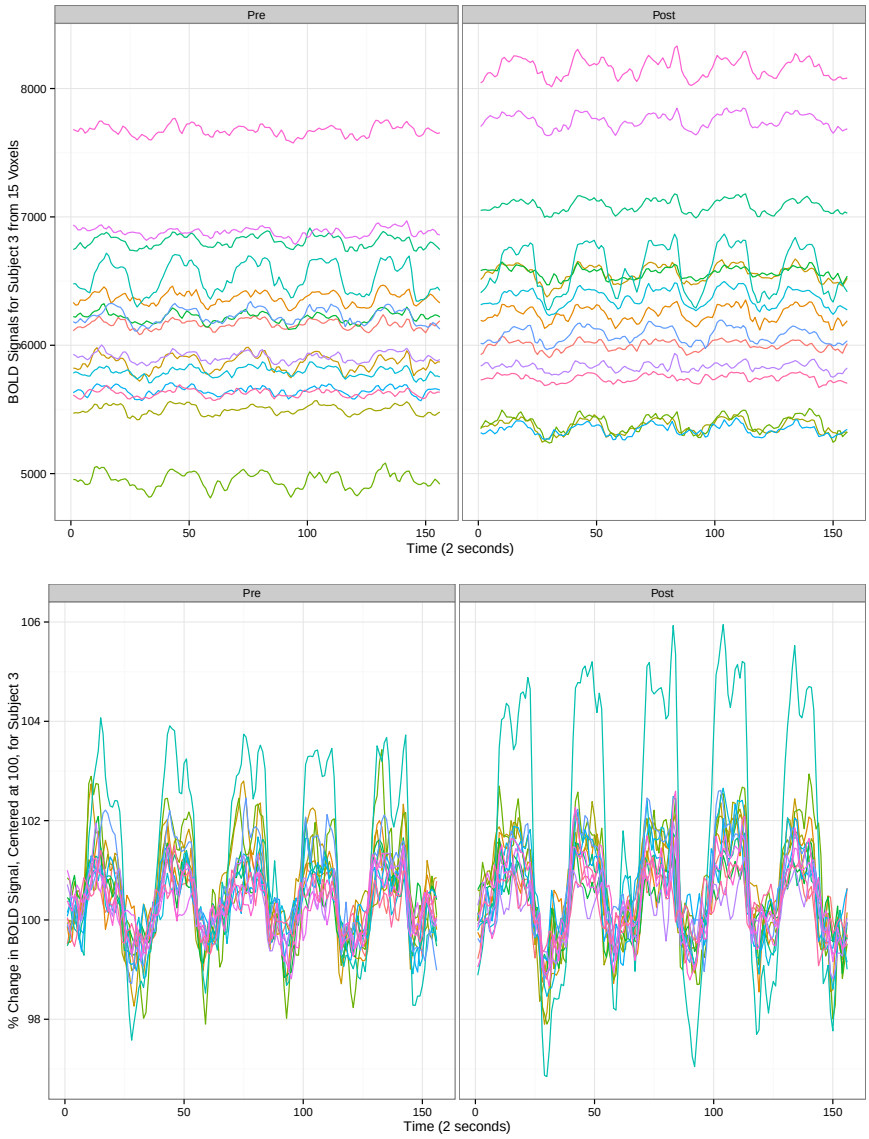

Figure A.4: Randomly selected BOLD signals from the motor cortex of Subject 3, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel.

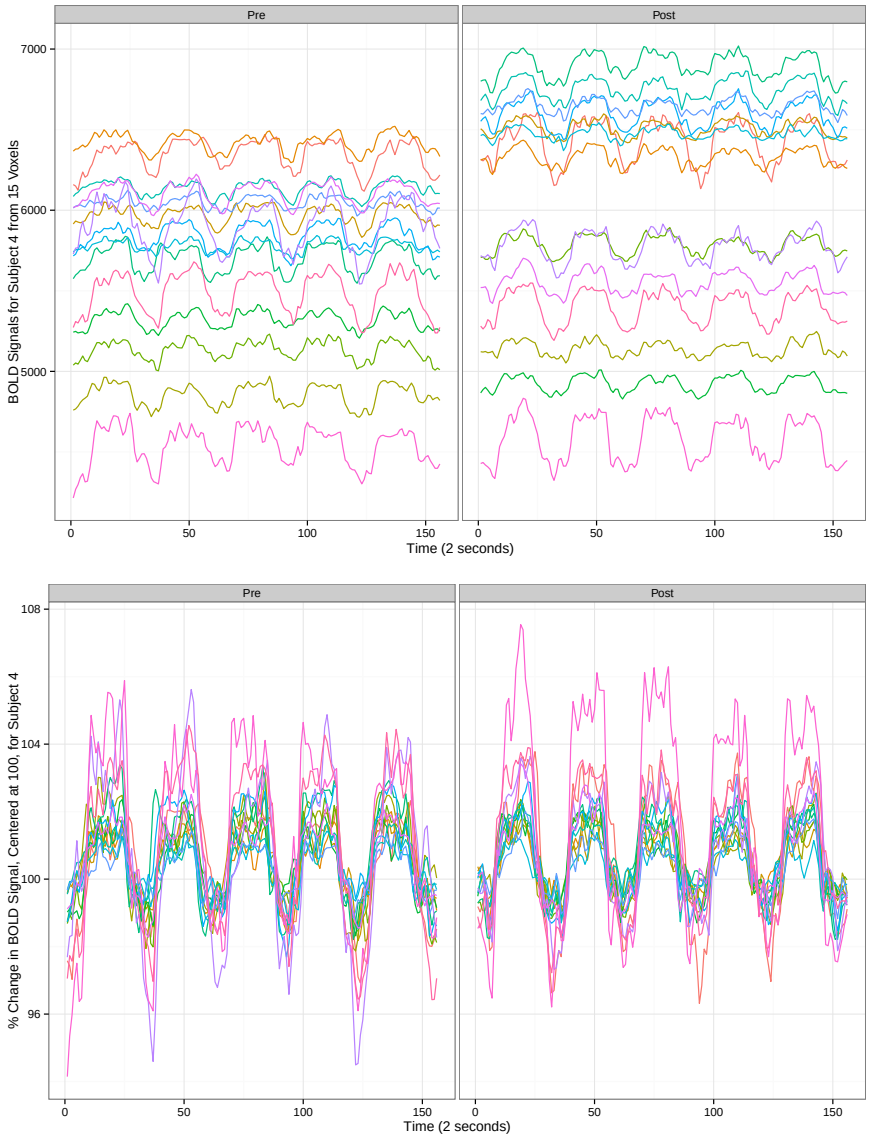

Figure A.5: Randomly selected BOLD signals from the motor cortex of Subject 4, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel.

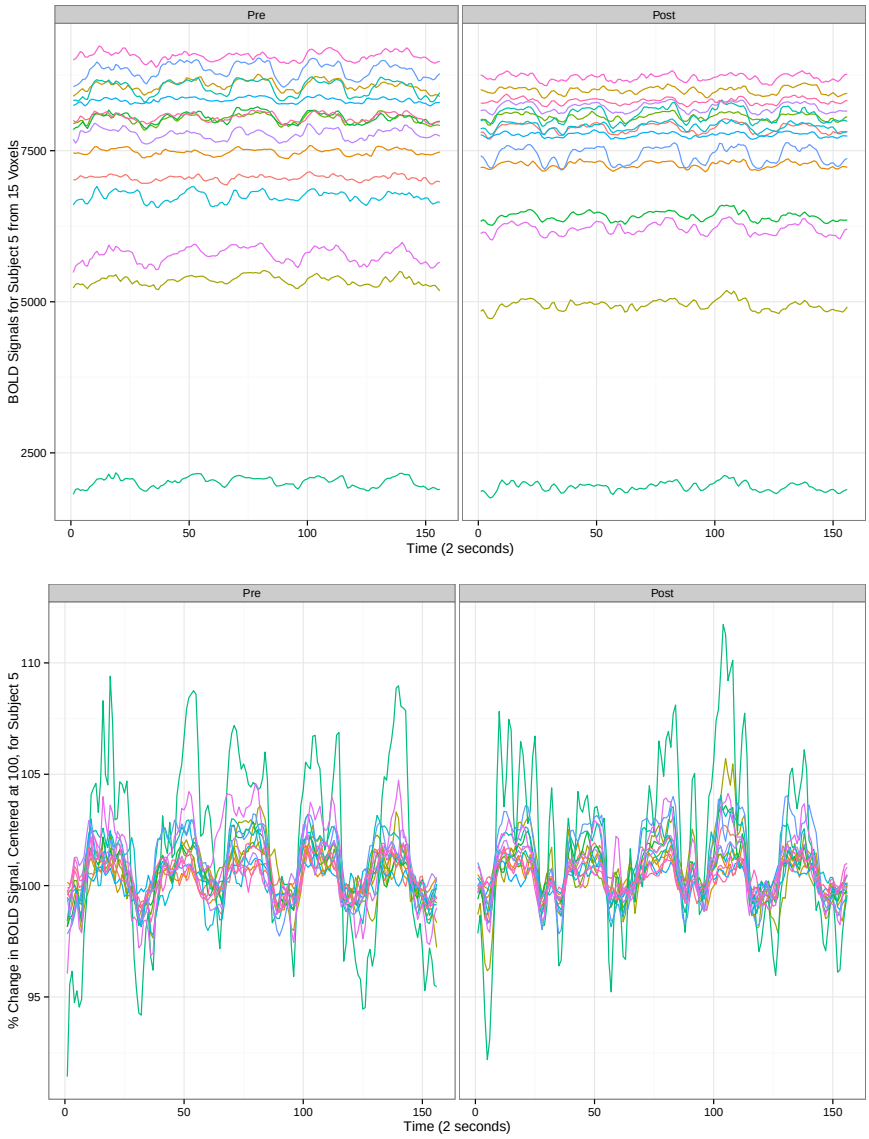

Figure A.6: Randomly selected BOLD signals from the motor cortex of Subject 5, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel.

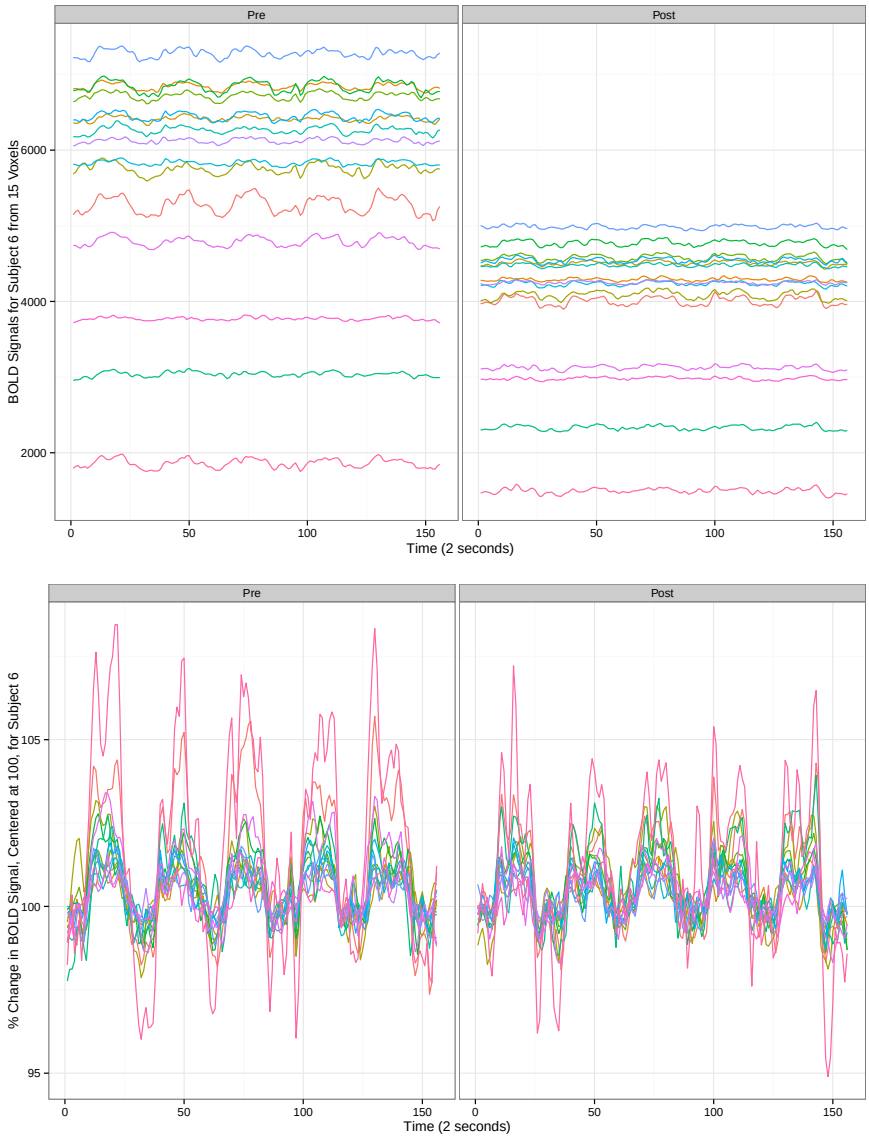

Figure A.7: Randomly selected BOLD signals from the motor cortex of Subject 6, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel.

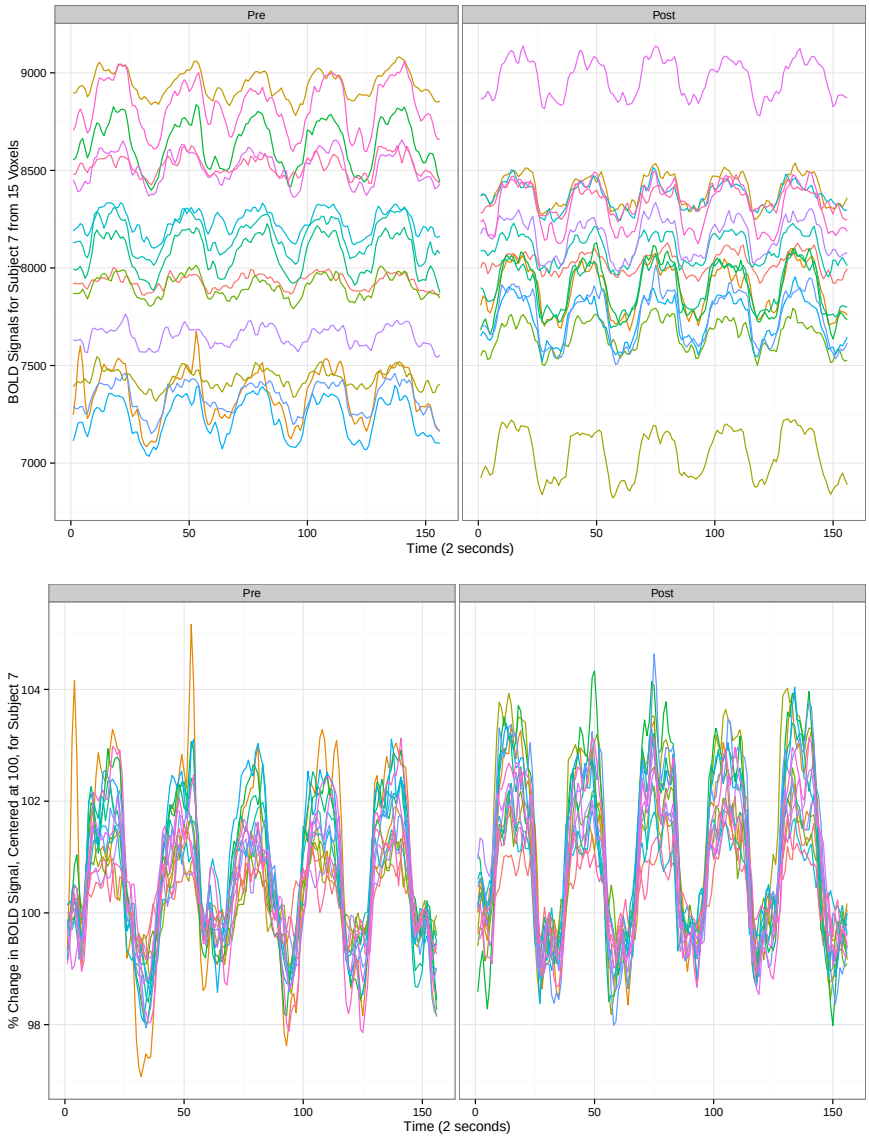

Figure A.8: Randomly selected BOLD signals from the motor cortex of Subject 7, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel.

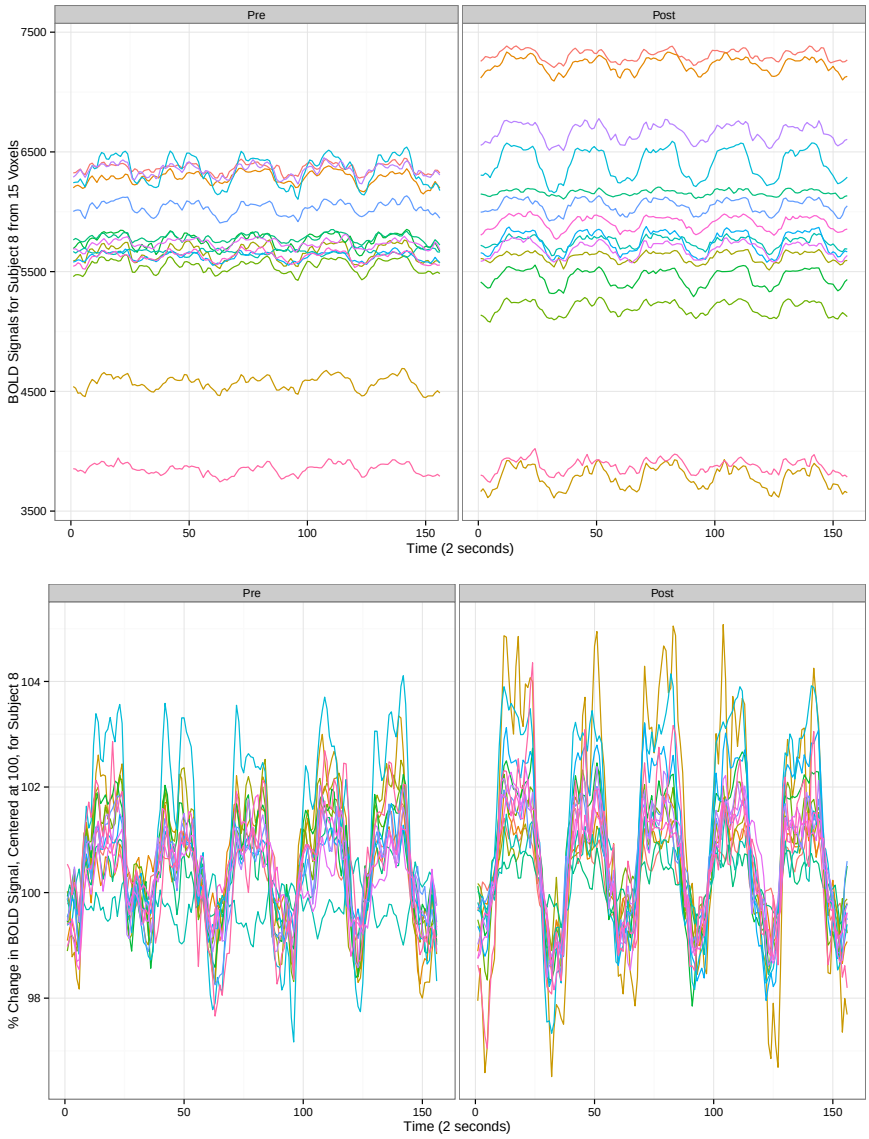

Figure A.9: Randomly selected BOLD signals from the motor cortex of Subject 8, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel.

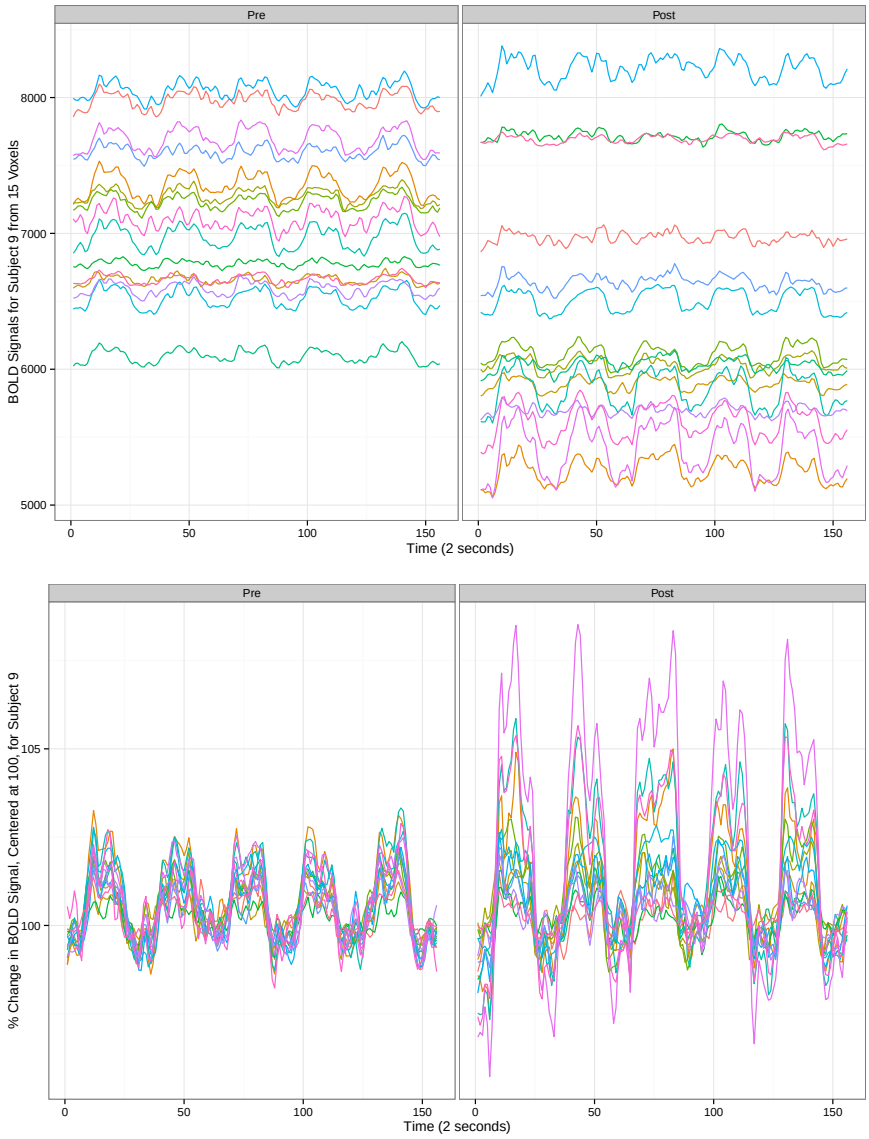

Figure A.10: Randomly selected BOLD signals from the motor cortex of Subject 9, before and after caffeine. The top two plots show the BOLD signals during the pre (left) and post (right) caffeine sessions while the bottom two plots display these signals as percent changes centered at 100. Time-series of the same color are from the same voxel.

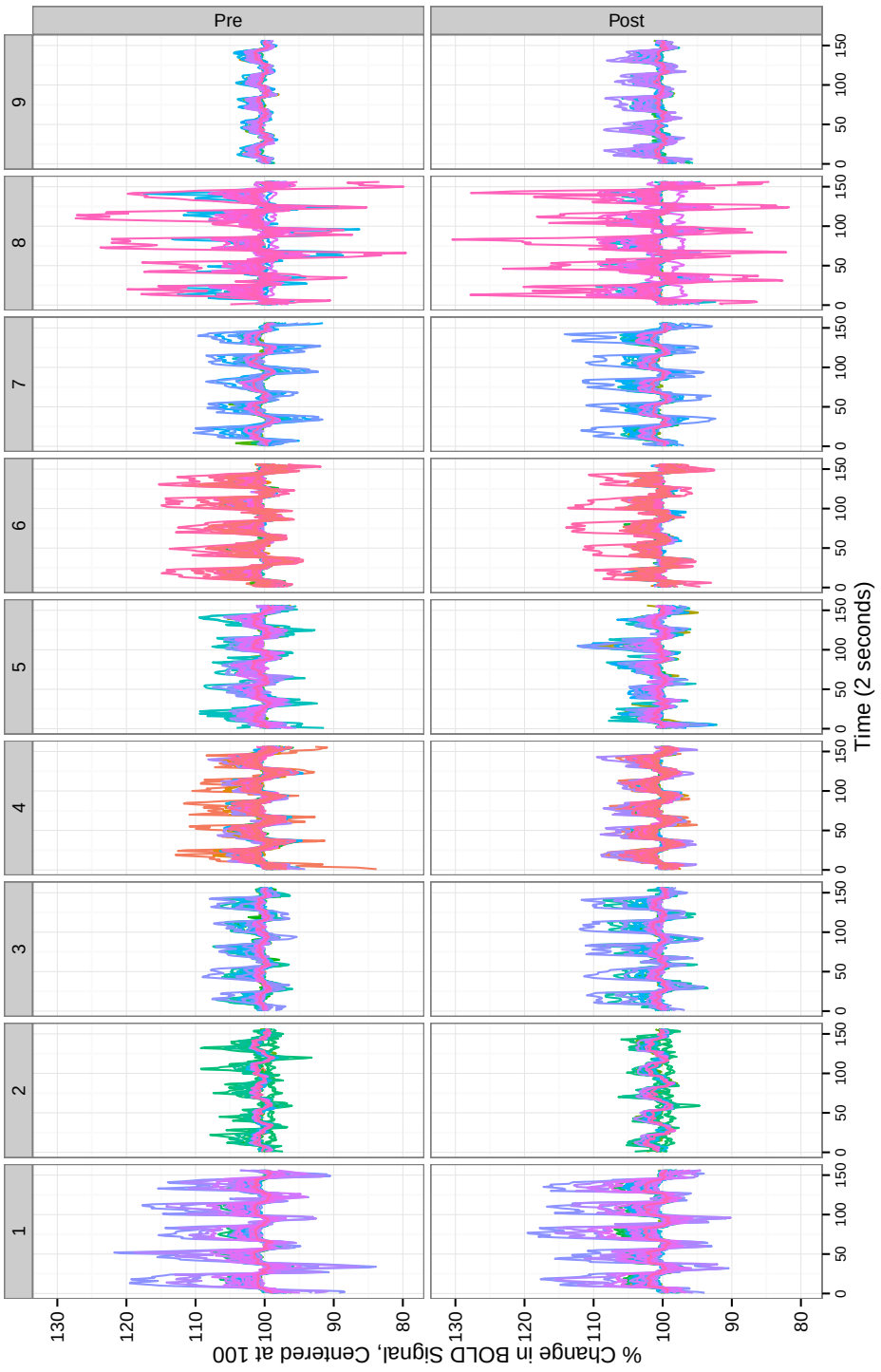

Figure A.11: Percent change in BOLD signals, centered at 100, from the motor cortex of 9 subjects, before and after caffeine. Timeseries of the same color within a subject are from the same voxel.

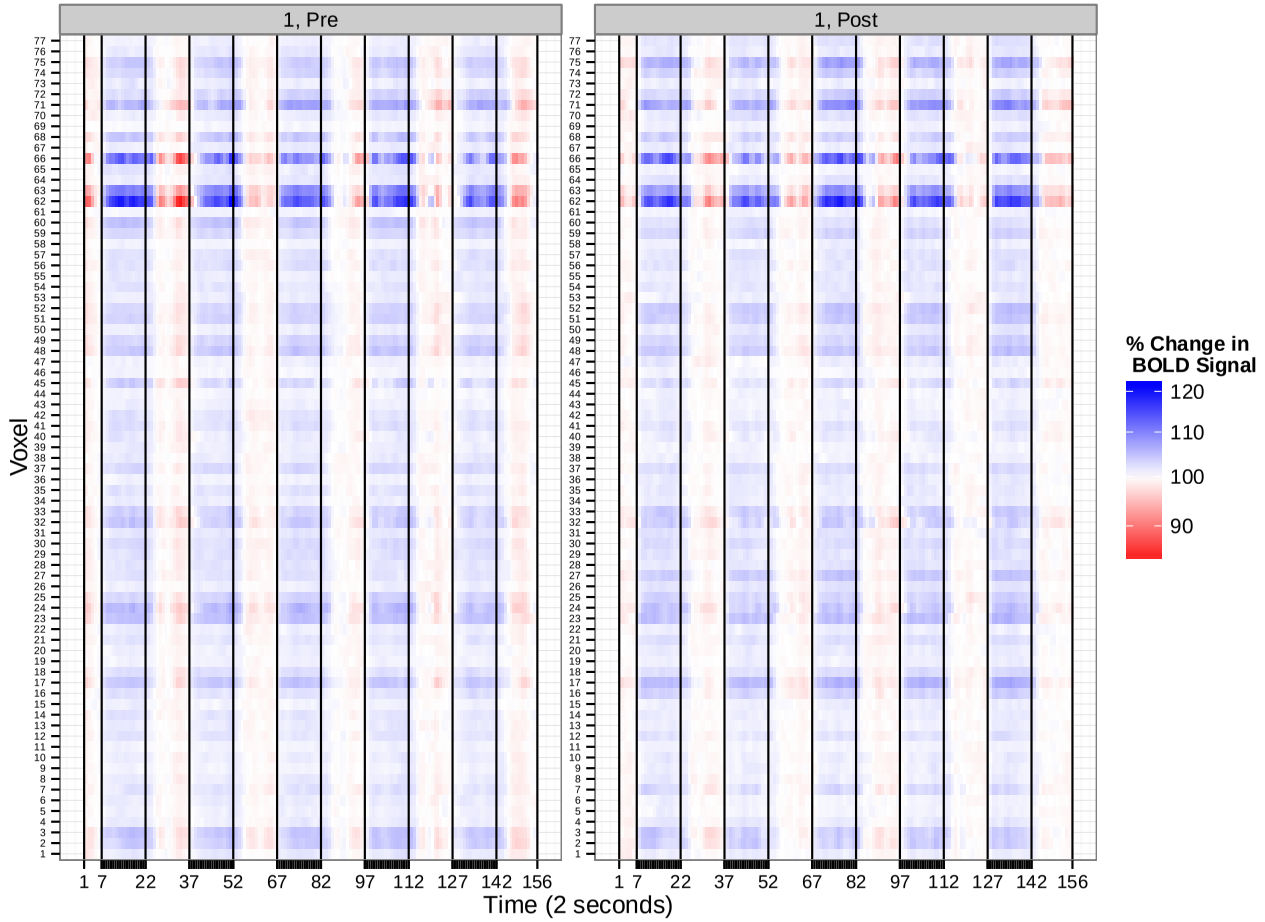

Figure A.12: Percent change in BOLD signals from the motor cortex of Subject 1, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color.

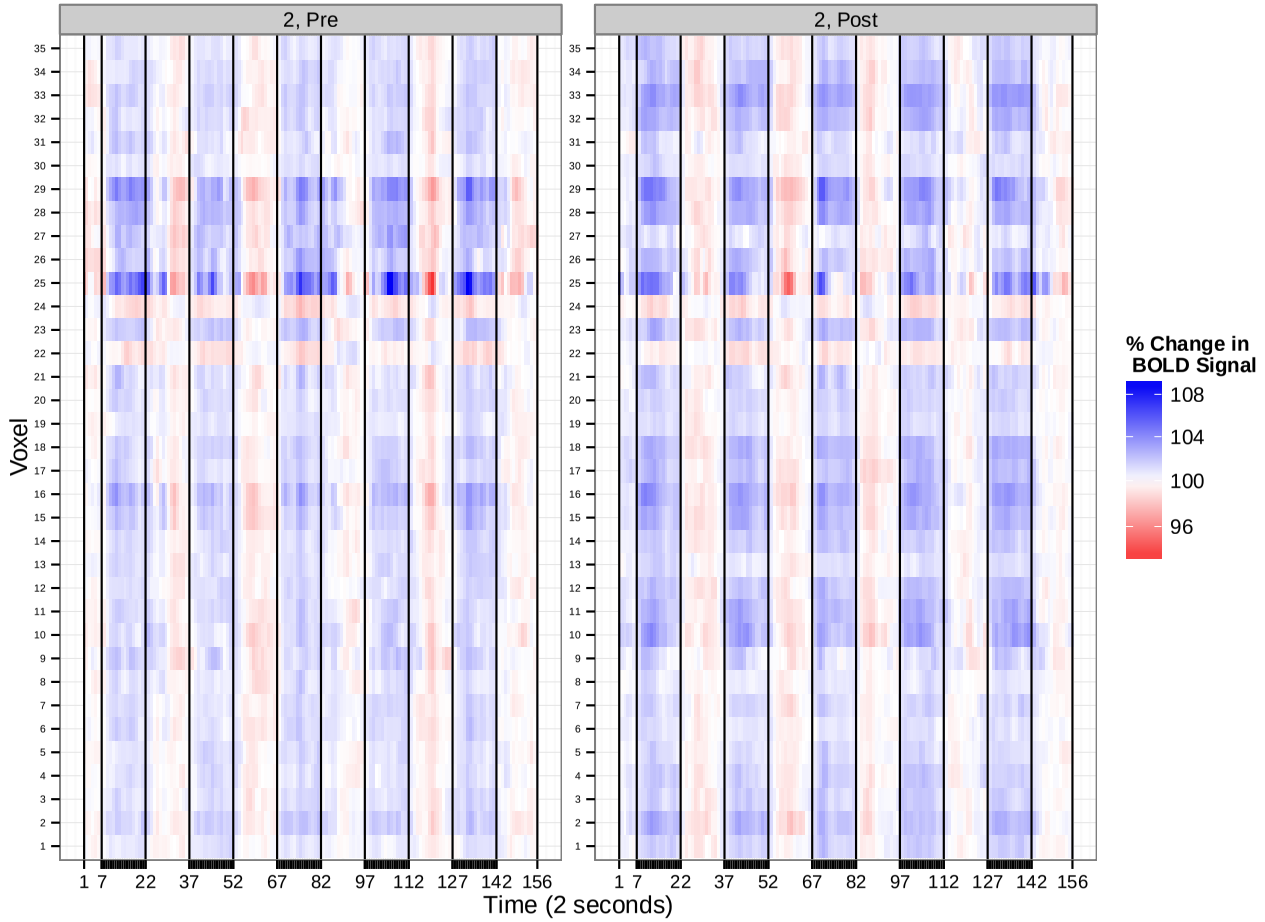

Figure A.13: Percent change in BOLD signals from the motor cortex of Subject 2, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color.

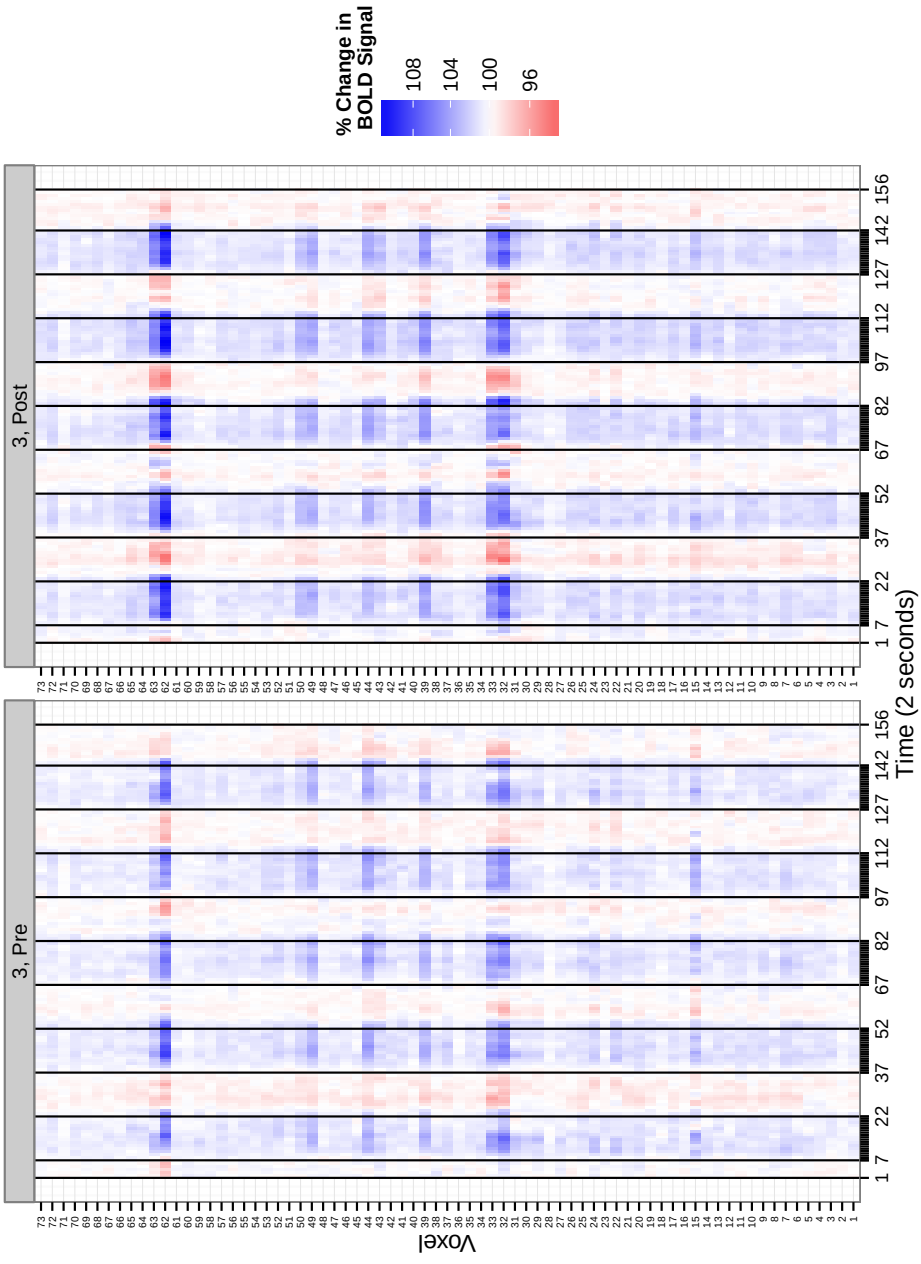

Figure A.14: Percent change in BOLD signals from the motor cortex of Subject 3, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color.

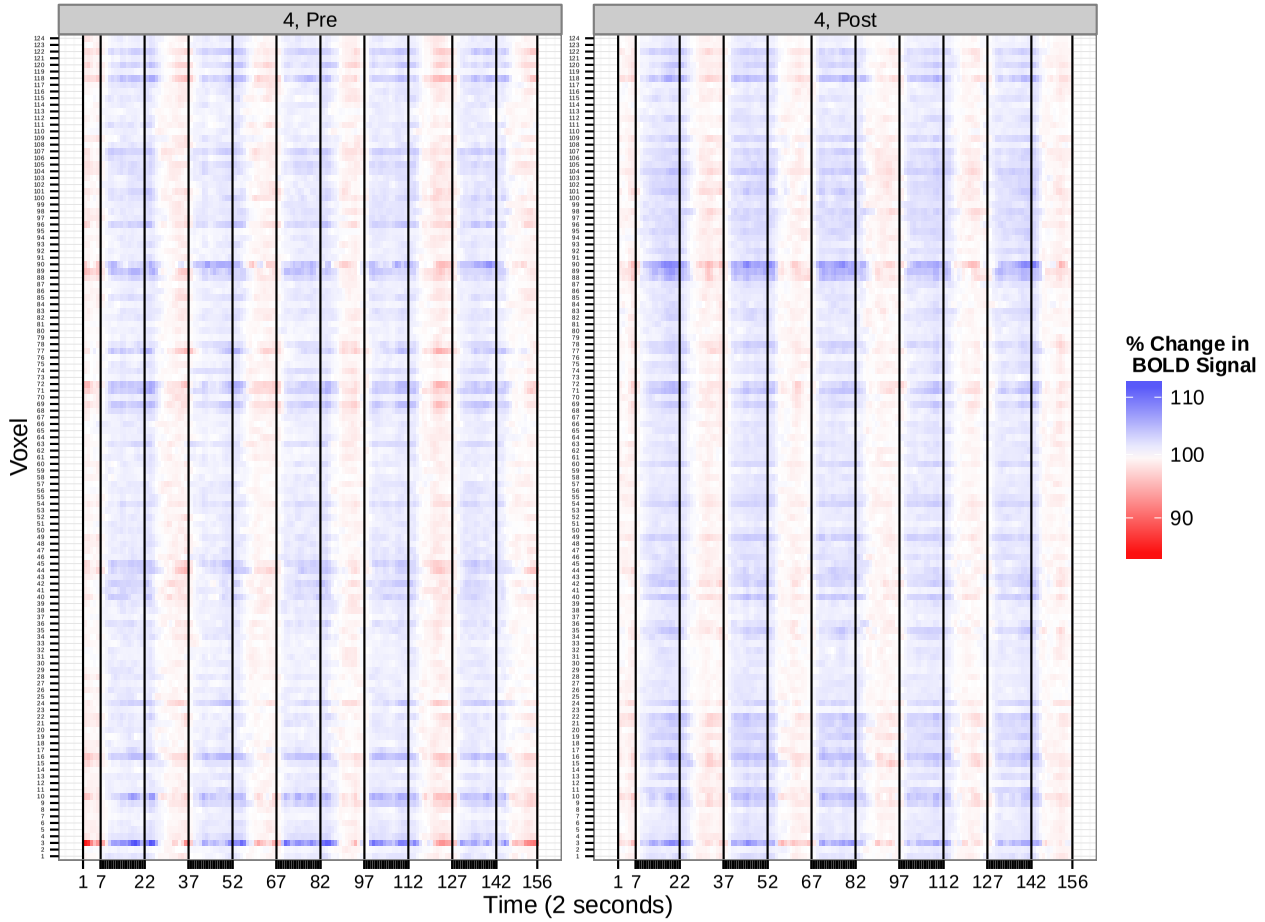

Figure A.15: Percent change in BOLD signals from the motor cortex of Subject 4, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color.

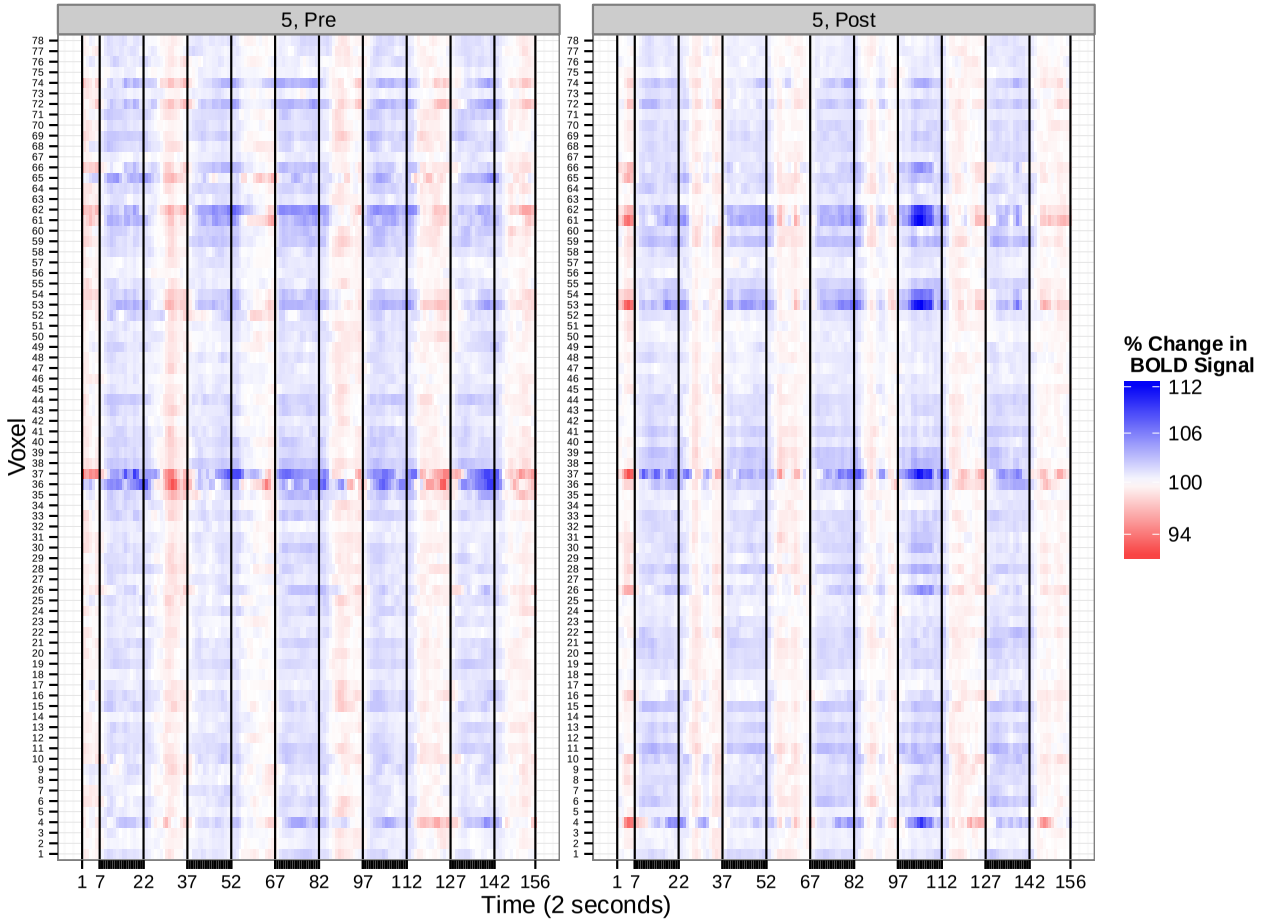

Figure A.16: Percent change in BOLD signals from the motor cortex of Subject 5, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color.

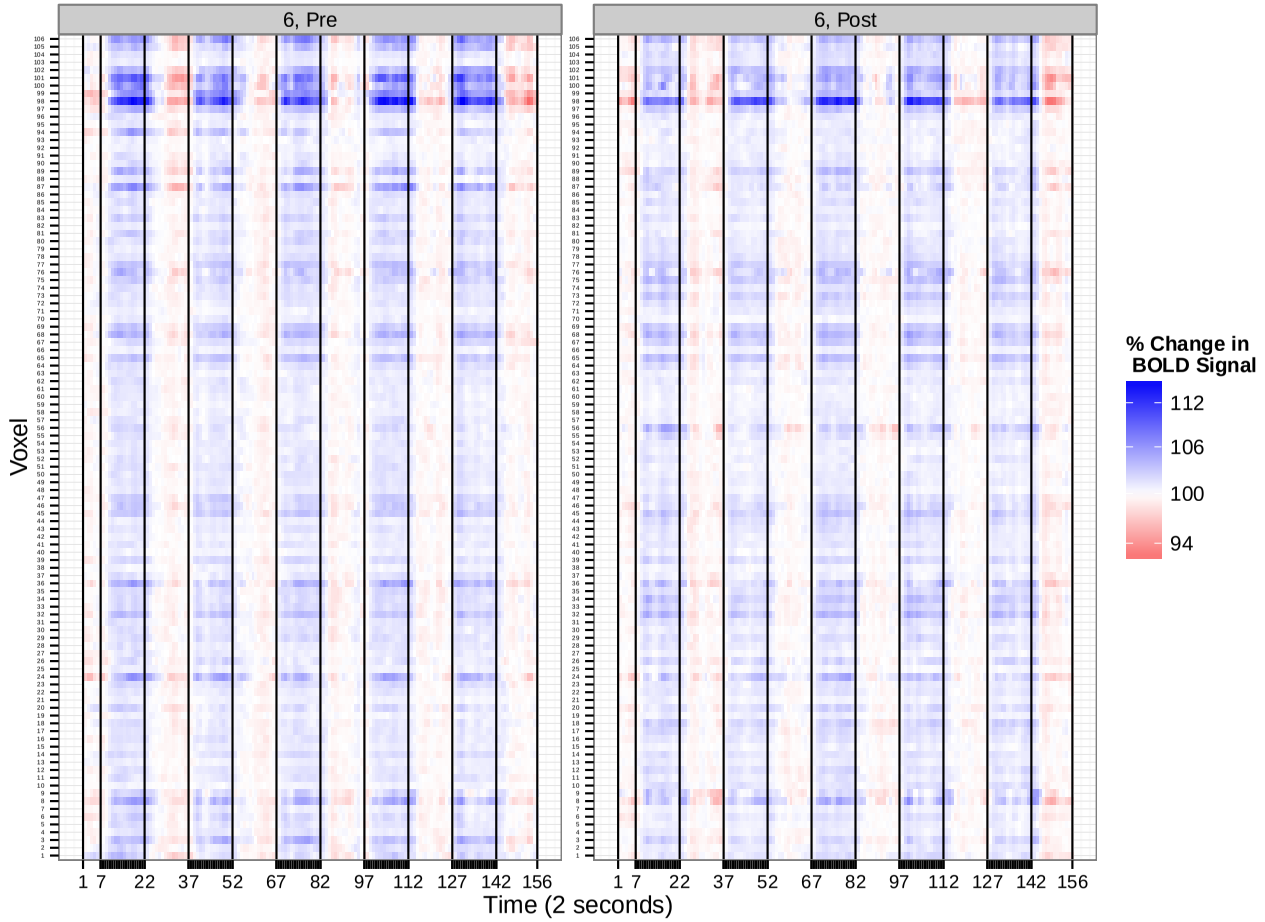

Figure A.17: Percent change in BOLD signals from the motor cortex of Subject 6, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color.

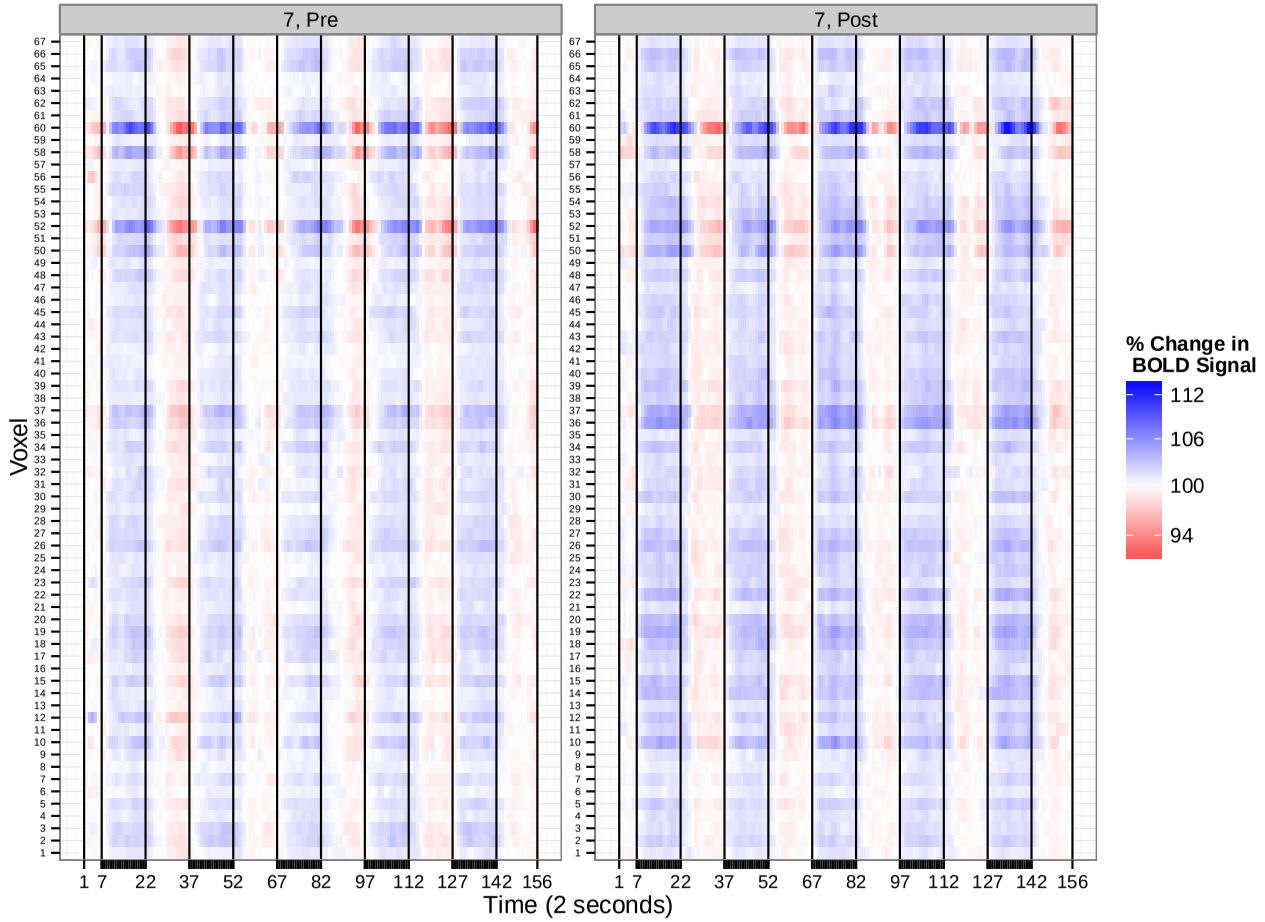

Figure A.18: Percent change in BOLD signals from the motor cortex of Subject 7, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color.

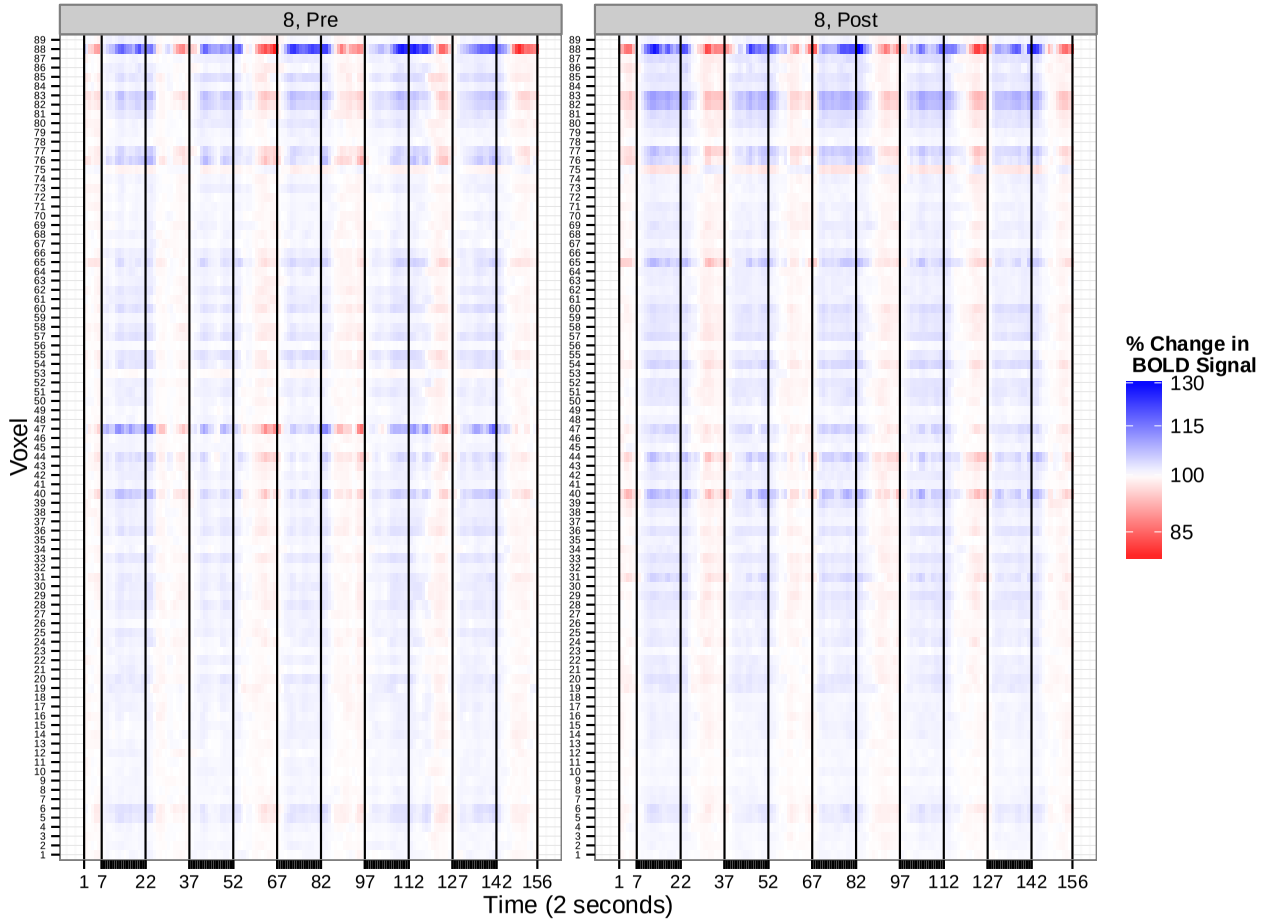

Figure A.19: Percent change in BOLD signals from the motor cortex of Subject 8, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color.

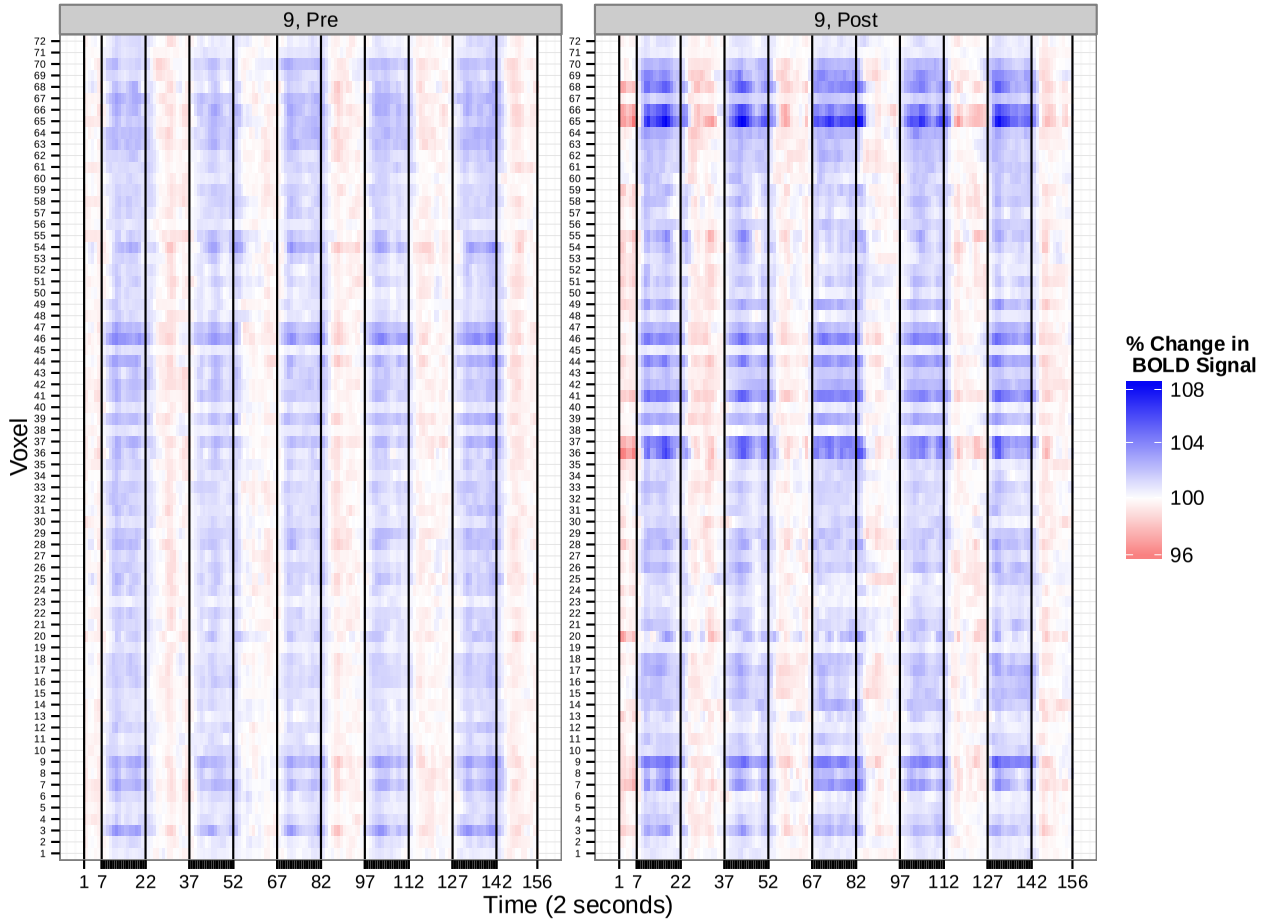

Figure A.20: Percent change in BOLD signals from the motor cortex of Subject 9, before and after caffeine. Periods of finger tapping are denoted by bars along the time axis, while the amount of change in activation is shown by intensity of color.

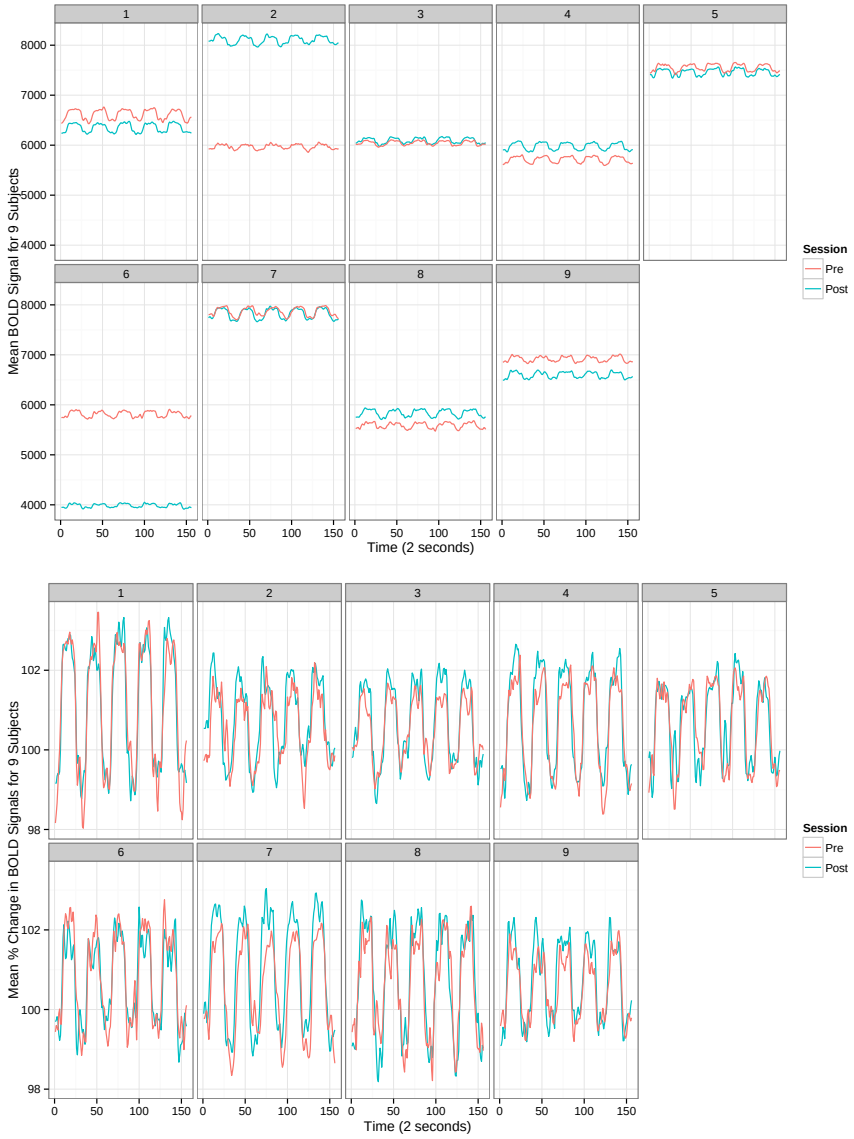

Figure A.21: Average BOLD signals from the motor cortices of 9 subjects, before and after caffeine.

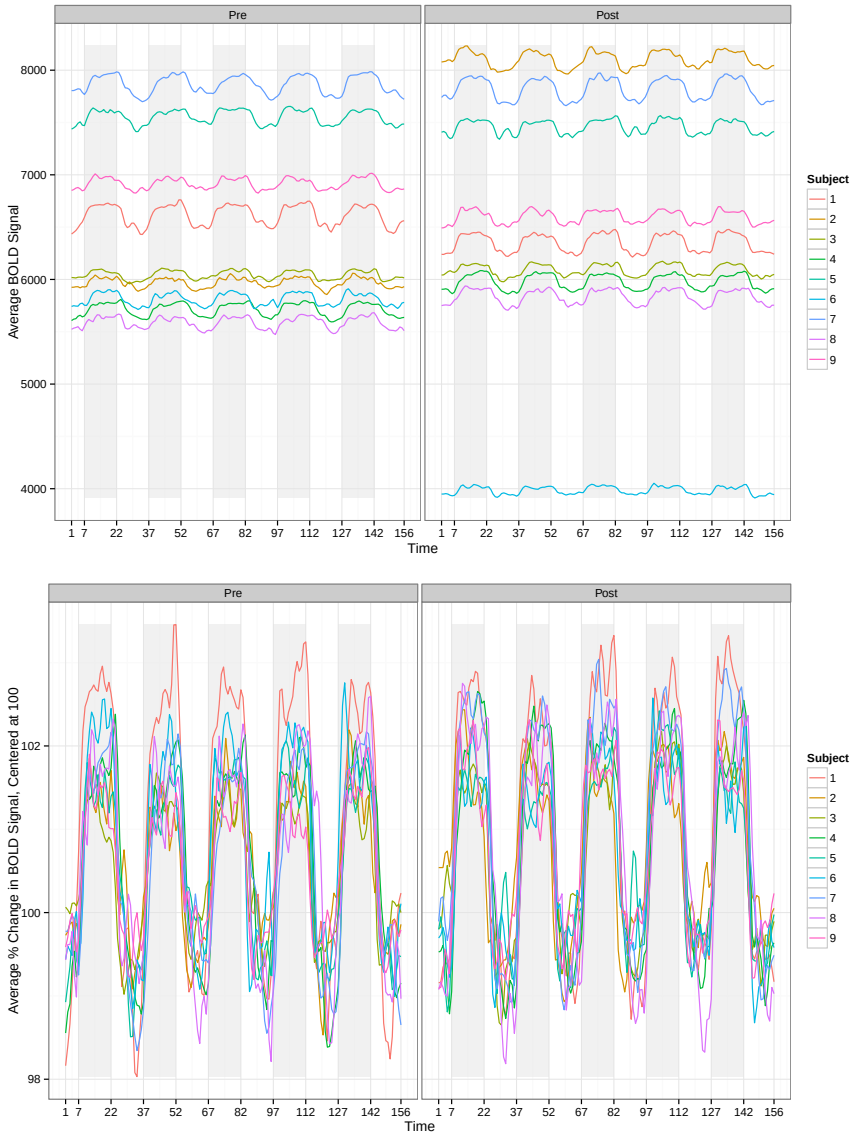

Figure A.22: Average BOLD signals from the motor cortices of 9 subjects, before and after caffeine. The top two plots show average BOLD signals for the pre (left) and post (right) caffeine sessions while the lower two plots display these signals as percent changes, centered at 100. Light gray shading indicates periods of fingertapping.

Bibliography

[1] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics (1992)*, volume 1, pages 610–624. Springer-Verlag, 1973.

[2] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceeding of the Second International Symposium on Information Theory*, pages 267–281. B.N. Petrov and F. Caski eds. Akademiai Kiado, Budapest, 1973.

[3] HD Bondell, A Krishna, and SK Ghosh. Joint Variable Selection for Fixed and Random Effects in Linear MixedEffects Models. *Biometrics*, 2010.

[4] NE Breslow and DG Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

[5] BA Brumback and JA Rice. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical . . .* , 1998.

[6] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications (Springer Series in Statistics)*. Springer, 2011.

[7] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information - Theoretic Approach*. Springer, second edition, 2002.

[8] J E Cavanaugh and R H Shumway. A bootstrap variant of AIC for state-space model selection. *Statistica Sinica*, 7:473–496, 1997.

[9] Jeng-Min Chiou and Chih-Ling Tsai. Smoothing parameter selection in quasi-likelihood models. *Journal of Nonparametric Statistics*, 18(3):307–314, April 2006.

[10] G Claeskens, T Krivobokova, and JD Opsomer. Asymptotic properties of penalized spline estimators. *Biometrika*, 2009.

[11] Gerda Claeskens and Nils Lid Hjort. Focused information criterion (with discussion). *Journal of the American Statistical Association*, 98:900–945, 2003.

[12] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008.

[13] A. Cnaan, N.M. Laird, and P. Slasor. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16:2349–2380, 1997.

[14] Y Cui, J S Hodges, X Kong, and B P Carlin. Partitioning degrees of freedom in hierarchical and other richly-parameterized models. *Technometrics*, 52:124–136, 2010.

[15] M. C. Donohue, R. Overholser, R. Xu, and F. Vaida. Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, 98(3):685–700, July 2011.

[16] B Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

[17] B Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81:461–470, 1986.

[18] PHC Eilers and BD Marx. Flexible smoothing with B-splines and penalties. *Statistical science*, 1996.

[19] PHC Eilers and BD Marx. Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: ...*, 2010.

[20] J Fan and R Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 2001.

[21] J Fan and H Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 2004.

[22] Yixin Fang. Asymptotic equivalence between cross-validations and akaike information criteria in mixed-effects models. *Journal of data science*, 9:15, 2011.

[23] G M Fitzmaurice, N M Laird, and Ware J H. *Applied Longitudinal Analysis*. Wiley, Hoboken, New Jersey, 2004.

[24] E. R. Greenberg, J. A. Baron, T. A. Stukel, M. M. Stevens, J. S. Mandel, S. K. Spencer, P. M. Elias, N. Lowe, D. W. Nierenberg, G. Bayrd, J. C. Vance, D. H. Freeman, W. E. Clendenning, T. Kwan, and the Skin Cancer Prevention Study Group. A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin. *New England Journal of Medicine*, 323:789–795, 1990.

[25] S. Greven and T. Kneib. On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, 97(4):773–789, July 2010.

[26] Sonja Greven and Thomas Kneib. On the behaviour of marginal and conditional aic in linear mixed models. *Biometrika*, 97(4):773–789, 2010. ID: 692680154.

[27] Shuva Gupta. *A Study Of The Asymptotic Properties Of Lasso Estimates For Correlated Data*. PhD thesis, 2009.

[28] P Hall and J D Opsomer. Theory for penalised spline regression. *Biometrika*, 92:105–118, 2005.

[29] David A. Harville. *Matrix Algebra From a Statistician's Perspective*. John Wiley, New York, 1996.

[30] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. This is page v Printer: Opaque this. *Elements*, 27(2):745, 2009.

[31] Nicolas W. Hengartner, Marten H. Wegkamp, and Eric Matzner-Lober. Bandwidth selection for local linear regression smoothers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):791–804, October 2002.

[32] Nils Lid Hjort and Gerda Claeskens. Focused information criterion and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association*, 101:1449–1464, 2006.

[33] JS Hodges and DJ Sargent. Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 88(2):367–379, 2001.

[34] Clifford M. Hurvich, Jeffrey S. Simonoff, and Chih-Ling Tsai. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293, May 1998.

[35] CLIFFORD M. Hurvich and CHIH -LING Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[36] JG Ibrahim, H Zhu, RI Garcia, and R Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 2011.

[37] J Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York, 2007.

[38] Jiming Jiang. Maximum posterior estimation of random effects in generalized linear mixed models. *Statistica Sinica*, 11:97–120, 2001.

[39] Göran Kauermann, Tatyana Krivobokova, and Ludwig Fahrmeir. Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):487–503, 2009.

[40] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, October 2000.

[41] Tatyana Krivobokova and Göran Kauermann. A Note on Penalized Spline Smoothing With Correlated Errors. *Journal of the American Statistical Association*, 102(480):1328–1337, 2007.

[42] T Kubokawa. Conditional and unconditional methods for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, 102:641–660, 2011.

[43] Thomas C.M. Lee. Smoothing parameter selection for smoothing splines: a simulation study. *Computational Statistics & Data Analysis*, 42(1-2):139–148, February 2003.

[44] Y. Lee and J. A. Nelder. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society.Series B (Methodological)*, 58(4):619–678, 1996. ID: 484505123.

[45] Y. Lee, J.A. Nelder, and Y. Pawitan. *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall/CRC, Boca Raton, Florida, USA, 2006.

[46] Q Li and J Racine. Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14:485–512, 2004.

[47] Y Li and D Ruppert. On the asymptotics of penalized splines. *Biometrika*, 2008.

[48] H. Liang, H. L Wu, and G. H. Zou. A note on conditional AIC for linear mixed-effects models. *Biometrika*, 95:773–778, 2008.

[49] H Linhart and W Zucchini. *Model Selection*. Wiley, New York, 1986.

[50] R. Littell, W. Stroup, and R Freund. *SAS for Linear Models (4th Edition)*. SAS Publishing, 2002.

[51] H Lu, J S Hodges, and B P Carlin. Measuring the complexity of generalized linear hierarchical models. *The Canadian Journal of Statistics*, 35:69–87, 2007.

[52] C E McCulloch, S R Searle, and J M Neuhaus. *Generalized, Linear, and Mixed Models*. Wiley, New York, 2008.

[53] L. Nie. Convergence rate of mle in generalized linear and nonlinear mixed-effects models: Theory and applications. *Journal of Statistical Planning and Inference*, 137(6):1787–1804, 2007. ID: 442804694.

[54] MR Osborne, B Presnell, and BA Turlach. Knot selection for regression splines via the lasso. *Computing Science and ...*, 1998.

[55] W Pan. Bootstrapping likelihood for model selection with small samples. *Journal of Computational and Graphical Statistics*, 8(4):687–698, 1999.

[56] Anna Leigh Rack-Gomer, Joy Liau, and Thomas T Liu. Caffeine reduces resting-state BOLD functional connectivity in the motor cortex. *NeuroImage*, 46(1):56–63, 2009.

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

[58] Jürg Schelldorfer, Peter Bühlmann, GEER DE, and SARA VAN. Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

[59] J Shang and J E Cavanaugh. Bootstrap variants of the Akaike information criterion for mixed model selection. *Computational Statistics and Data Analysis*, 52:2004–2021, 2008.

[60] J Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[61] R Shibata. Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica*, 7:375–394, 1997.

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

[63] D J Spieglehalter, N G Best, Carlin B P, and A van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–639, 2002.

[64] Walter W. Stroup. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications* . CRC Press, Boca Raton, FL, 2012.

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

[66] Florin Vaida and Suzette Blanchard. Conditional Akaike information for mixed-effects models. *Biometrika*, 92:351–370, 2005.

[67] E. F. Vonesh. A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, 83(2):447–452, 1996. ID: 92221435.

[68] Carrie Wager, Florin Vaida, and Goeran Kauermann. Model selection for penalized spline smoothing using Akaike information criteria. *Australian and New Zealand Journal of Statistics*, 49:173–190, 2007.

[69] Hansheng Wang, Guodong Li, and Chih-Ling Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):63–78, February 2007.

[70] X Wang, J Shen, and D Ruppert. On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics*, 2011.

[71] Russ Wolfinger. Laplace's approximation for nonlinear mixed models. *Biometrika*, 80:791–795, 1993.

[72] R. Xu and A Gamst. Risk estimation. In *High Dimensional Data Analysis in Oncology*, pages 63–88. Springer, New York, 2008.

[73] Akifumi Yafune, Takashi Funatogawa, and Makio Ishiguro. Extended information criterion approach for linear mixed effects models under restricted maximum likelihood estimation. *Statistics in Medicine*, 24:3417–3429, 2005.

[74] Y Yang. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92:937–950, 2005.

[75] Y Yang. Prediction/estimation with simple linear models: is it really that simple? *Econometric Theory*, 23:1–36, 2007.

[76] J Ye. On meausuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131, 1998.