Contributions to Modeling the Dynamic Association Structure in Longitudinal Data Sets

by

Jason E. Goldstick

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2010

Doctoral Committee:

       Associate Professor Kerby A. Shedden, Chair
       Professor Tailen Hsing
       Professor Robert A. Zucker
       Assistant Professor Yves A. Atchade

ACKNOWLEDGEMENTS

I wish to express sincere gratitude to my advisor Kerby Shedden for his guidance and encouragement during my time in graduate school. Kerby has imparted to me extensive knowledge, particularly in the areas of modeling and computational statistics, and is more responsible for my development as a statistician than anyone else. I would also like to thank the other members of my committee – Tailen Hsing, Bob Zucker, and Yves Atchade – for their helpful suggestions. Special thanks to Bob Zucker for use of his data sets, and for providing me with my first interdisciplinary research experience.

I want to acknowledge the many instructors that contributed to my desire to do academic research. Many thanks to Ed Rothman, Kerby Shedden, Moulinath Banerjee, and George Michailidis for their unique teaching styles and encouragement early on in my statistics education. Special thanks to Ed for his long term friendship and for helping me improve my consulting and interdisciplinary research skills.

Finally, I want to acknowledge my family and friends. Most importantly, I appreciate my parents, whose unwaivering support throughout my life has made this work possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

1.1 Introduction and motivation

The analysis of clustered and multivariate longitudinal data arises in many situations. While mean structures dominate many applications involving these types of data, in some important applications the primary interest is in the associations between variables, or between individuals in a cluster. For example, when studying family dynamics, or the environmental or genetic basis of traits, it is important to understand how a particular trait covaries among family members. Here the family is viewed as a cluster comprising individuals whose traits are dependent within and across time. As another example, consider a collection of distinct but related traits measured over time on unrelated individuals. In this case, the dependencies among the traits describe patterns of comorbidity.

The motivating application for this work is data collected from a long term family study conducted jointly by the University of Michigan and Michigan State University, known as the Michigan Longitudinal Study (MLS) [Zucker et al. (1996)]. The families in the MLS all had at least one young child at the time of enrollment, and were ascertained following a complex strategy that enriched for families at high risk for substance abuse. Subsets of the MLS will be used to illustrate the various methodological contributions of this dissertation.

There are extensive research literatures on longitudinal data analysis, the analysis of clustered data, and multivariate data analysis. Much less work has been done to link these areas. For example, most of the literature on multivariate longitudinal data focuses on marginal and conditional mean structures, and the literature on clustered longitudinal data mainly focuses on the impact of within-cluster dependence on uncertainty levels for estimated mean structures. In both cases, the dependency structure enters primarily as a nuisance, rather than as the primary object of interest.

In this dissertation, two distinct, but related ideas are developed. The first is to explicitly model the dependency between variables in multivariate longitudinal data as a function of time. This allows the dependency between two variables to become stronger or weaker over time, and even to change in direction. This approach is shown to provide new insight into the pattern of comorbidities between related traits viewed over time. The second area developed here is motivated by the fact that all longitudinal data for human subjects is simultaneously longitudinal in two senses – it can be viewed as a longitudinal function of age, or as a longitudinal function of time. When individuals are related, these two ways of indexing the data may interact in subtle ways. We consider a generative model in which the unobserved events that drive within-cluster dependencies are either functions of age, or of time. A major focus is how and whether these two sources of covariation can be separated using observational data.

1.2 Background and review

For perspective, we now provide the necessary background on several fundamental concepts in the analysis of clustered and/or longitudinal data, with primary emphasis on methods used to characterize the association structure. This begins with a review of general random effects models and how they are used to describe the pattern of coherence in multivariate data. Generalized Estimation Equations (GEE) are another

approach which bears mentioning due to the frequency of use in the analysis of correlated data. Covariance structure modeling (CSM) will also be mentioned, due to the flexibile patterns of coherence they are capable of describing. Finally, log-linear models for multivariate binary data, which does not explicitly fall into any of the categories above, is mentioned.

1.2.1 Random effects models

Of fundamental importance in the field of modeling correlated responses is the random effects model. The commonality between all random effects models is that correlations between observed variables are assumed to arise from underlying unobserved variables that are completely or partially shared between observations. Classical work from the early 20th century by R.A. Fisher and his contemporaries [e.g. Fisher (1919)] utilized random effects models to estimate genetic and environmental contributions to an observed trait, where estimation was based on the comparison of identical and non-identical twin pairs. Much of the early work in random effects models is built upon the framework of Charles Roy Henderson, a good review of which is given in Searle (1968). The general model described by Searle was

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{U}_i \boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i \tag{1.1}$$

Here $\mathbf{Y}_i, \mathbf{X}_i, \mathbf{U}_i$ are the subject i response, predictor variables, and random effects design matrix, respectively. $\boldsymbol{\beta}$ are fixed effects coefficents, $\boldsymbol{\xi}_i$ are mean 0 random variables with covariance matrix $\sigma^2 \mathbf{A}$, and $\boldsymbol{\varepsilon}_i$ are iid errors. Several methods were described by Henderson to estimate the variance components, each of which involved solving a system of equations generated by equating some type of mean squares to their expectations. In addition, the methods of Henderson can be used to give best linear unbiased predictors for the random effects. Quite general structures for $\mathbf{A}$ may be used to describe very intricate patterns of correlation in a data set. One

3

major advantage of the approaches advocated by Searle and Henderson are they they do not require balance in the data and they do not explicitly assume an underlying distribution for the random effects, only the existence of the first two moments of the response variable. The other primary estimation technique contemporary to this was to assume a multivariate normal distribution for the random effects and use maximum likelihood (ML) estimation [e.g. Hartley and Rao (1967)].

Noting the computational difficulty of ML estimation of variance components, and the questionable theoretical basis for the ANOVA-type methods used by Searle and Henderson, Rao (1971) describes estimation of a linear combination of the variance components,

$$c_1 \sigma_1^2 + ... + c_k \sigma_k^2, \tag{1.2}$$

by a quadratic function of the responses, $\mathbf{Y'AY}$. Rao describes a method of determining $\mathbf{A}$ by minimizing a particular matrix norm under certain constraints, such as unbiasedness and invariance under translation of the covariates $\mathbf{X}$. This general approach is termed minimum norm quadratic unbiased estimation (MINQUE). It is argued by Rao that this approach is computational simpler to its predecessors and is appropriate for any experimental design, whereas with previous methods seemingly arbitrary choices must be made, depending on the data conditions.

Due to the desirable asymptotic properties of ML estimation, interest in determining methods for calculating MLEs for variance parameters became of greatest interest as computing power improved. Noting that ML estimators of variance components in model (1.1) are generally biased downward in finite samples and ignore the loss in degrees of freedom from estimation of $\boldsymbol{\beta}$, a restricted maximum likelihood (REML) approach to estimation was considered. This method was described by Harville (1977) as ML analysis of a data set based on $n - p^\star$ linearly independent error contrasts, where $p^\star = \mathrm{rank}(\mathbf{X})$. REML produces unbiased (and in some cases, lower MSE than

ML) estimates of variance components and still maintains the validity of the likelihood ratio test and Wald based confidence intervals under the standard regularity conditions.

A seminal work on the random effects model, Laird and Ware (1982), takes a different view. Instead of considering the marginal distribution, $\mathbf{Y}_i|\mathbf{X}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V})$, and maximizing the likelihood, the authors take a Bayesian view and consider the distribution of $\mathbf{Y}_i|\mathbf{X}_i, \boldsymbol{\xi}_i$. The unconditional likelihood is calculated by integrating over $\boldsymbol{\xi}_i$ and $\boldsymbol{\beta}$ and optimizing the resulting likelihood to get estimates of the random effect variances and covariances. Estimates of the fixed effects and random effects can be calculated by their posterior expectation, given the converged parameter values. The authors show that when a flat prior is placed on the fixed effects with variance approaching 0, this yields a REML estimator of the parameters in the model. This work laid the foundation for modern random effects models and paved the way for much subsequent progress.

Extending upon the work of Laird and Ware (1982), various models for mixed effects analysis of categorical data were developed. Stiratelli et al. (1984) use an exactly analogous approach to model the logit response probabilities $\boldsymbol{\lambda}_i = \log(\frac{\boldsymbol{p}_i}{1-\boldsymbol{p}_i})$ as

$$\boldsymbol{\lambda}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i\boldsymbol{\xi}_i. \tag{1.3}$$

Ochi and Prentice (1985) used the probit rather than logit link, noting its appropriateness for use on equicorrelated binary data. Gilmour et al. (1985) built mixed models for binomial outcomes by viewing them as thresholded normal random variables, which highlights an important equivalence. Lindstrom and Bates (1990) proposed a more general class of non-linear mixed models that is the basis of the most much mixed model software. Retaining notation from above and denoting the t-th entry of $\mathbf{Y}_i$ by $\mathbf{Y}_{it}$, this model is

$$\mathbf{Y}_{it} = f(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{U}_i\boldsymbol{\xi}_i, \mathbf{X}_{it}) + \boldsymbol{\varepsilon}_{it} \tag{1.4}$$

where $\mathbf{A}_i, \mathbf{U}_i$ are fixed and random effect design matrices. The authors assert that any non-linear function of fixed and random effects can be written as (1.4) for the properly chosen f. The estimation procedure is a combination of least-squares for non-linear fixed effects models and ML/REML for linear random effects models. Generalizations of such models for various types of data and analysis goals are discussed in Pinheiro and Bates (2000) and implemented in the R package `lme4`. A good survey of the some numerical techniques required for the approximation of the log-likelihood in mixed effects models, which often involves high dimensional numerical integration, is given by Pinheiro and Bates (1995).

Various other advancements have been made in recent years on the use of random effects models. Particularly with random effects models for longitudinal data, the view of responses as superpositions of random, smooth functions in the vein of Ramsey and Silverman (1997) has received some attention. As noted by Gao et al. (2001), several authors have used smooth functions to model fixed parameters in mixed effects models, but this alone does not provide an extension to the functional data case. The basic functional mixed effects model is the natural extension of Laird and Ware's formulation:

$$\mathbf{Y}_{it} = \mathbf{X}_{it}\boldsymbol{\beta}(t) + \mathbf{U}_{it}\boldsymbol{\xi}_i(t) + \boldsymbol{\varepsilon}_i(t), \tag{1.5}$$

where $\boldsymbol{\beta}(t)$ is a smooth function and each of $\boldsymbol{\xi}_i(t)$ are smooth random processes. The fixed effects are typically estimated by either some basis expansion– Guo (2002) used smoothing splines, Morris and Carroll (2006) used the wavelet basis– or by some penalty [e.g. Ke and Wang (2001)] to ensure the estimated function is smooth. Random effects are commonly envisioned as realizations of some Gaussian process

with smoothness quantified by some basis expansion or penalty [e.g. Guo (2002), Ke and Wang (2001), Morris and Carroll (2006)].

1.2.2 Generalized estimating equations

Another tool that has found applicability in models for multivariate data is generalized estimating equations (GEE). Zeger and Liang (1986) introduced GEE as a quasi-likelihood based method for estimating regression parameters in non-Gaussian GLMs. The only distributional assumption made is that the data arises from an exponential family, with the standard parameterization being that $\boldsymbol{\mu}_j = h(\mathbf{X}_j\boldsymbol{\beta})$, for some link function h. The GEE estimator solves

$$\sum_{j=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{\beta}}\right)' \mathbf{V}_j(\boldsymbol{\alpha}, \phi)^{-1}(\mathbf{Y}_j - \boldsymbol{\mu}_j) = \mathbf{0} \tag{1.6}$$

where $\mathbf{V}_j(\boldsymbol{\alpha}, \phi) = \mathbf{A}_j^{1/2}\mathbf{R}_j(\boldsymbol{\alpha})\mathbf{A}_j^{1/2}/\phi$, ϕ is a dispersion parameter and $\mathbf{A}_j$ is a diagonal matrix with elements $g(\mu_{ij})$, where g is a function linking the mean to the variance. $\mathbf{V}_j$ is a surrogate for the covariance matrix of $\mathbf{Y}_j$, where $\mathbf{R}_j(\boldsymbol{\alpha})$ is a working correlation matrix indexed by a parameter $\boldsymbol{\alpha}$. The authors show that, regardless of misspecification in $\mathbf{R}$, the $\widehat{\boldsymbol{\beta}}$ arrived at by iteratively solving (1.6) and plugging in empirical estimates of $\phi, \boldsymbol{\alpha}$ based on the residuals, is consistent. If the true covariance matrix is within the family defined by $\mathbf{R}_j$, then the resulting estimates are efficient.

The original formulation of GEE views the covariance structure as a nuisance parameter. As shown by Liang et al. (1992) the estimates of $\boldsymbol{\alpha}$ derived by ordinary GEE are very inefficient, thus when the association structure is of interest, ordinary GEE is not appropriate. Prentice (1988) extended GEE with a second set of estimating equations analogous to (1.6) for the pairwise correlations, with its own working correlation matrix. The estimator is then the simultaneous solution to the two sets of estimating equations. Lipsitz et al. (1991) did analogous work in binary data using

the odds ratio as a measure of association rather than correlation. Related work by Zhao and Prentice (1990) and Prentice and Zhao (1991) extend GEE to the simultaneous estimation of means and covariances in more general data settings in what is termed by Liang et al. (1992) to be 'GEE2'. Other work on extending GEE2 to more general settings with continuous and/or categorical responses can be found in Liang et al. (1992), Zhao et al. (1992). Molenberghs and Ritter (1996) show a connection between GEE2 and a class of likelihood-based marginal models using the quadratic exponential family.

GEE2 does represent an improvement when the association structure of interest, but some aspects are sacrificed. Depending on the size of the clusters, there may be a substantial number of squares and cross products used in the second order equations. In addition, the property that the estimates are consistent under misspecification is lost [McCulloch (2003)]. GEE, even when no misspecification is present, does not partition out variation into distinct sources the way random effects models do; in many inquiries, such as genetic analysis, the relative impact of certain sources of variation are the primary interest. Finally, when the overall goal of the analysis is to estimate the association structure, GEE may be quite inefficient [McCulloch (2003)].

1.2.3 Covariance structure models

Structured covariance models (CSM), originated by Bock (1960) and further explicated in Bock and Bargmann (1966), represent a very broad class of models used to represent a population covariance matrix, Σ as a function of a parameter vector $\boldsymbol{\theta}$. When the covariance structure is of primary interest, this model introduces an enticing possibility for readily interpretable components contributing to particular variances and covariances. CSM has primarily found application in social sciences where non-experimental data is very common. CSM is commonly used to measure some underlying construct only from observation of noisy data meant to measure that

construct, illuminating the relationship between CSM and factor analysis.

As noted by Joreskög (1978), much of the early work in covariance structure modeling involved parameterizing $\mathbf{\Sigma}$ by a sum or product of matrices, with largely a priori structures for each constituent matrix. A simple example of this would be

$$\mathbf{\Sigma}(\sigma_1, \sigma_2) = \sigma_1^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \sigma_2^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{1.7}$$

The work of Joreskög (1978) was the first to consider the possibility of modeling $\mathbf{\Sigma}(\boldsymbol{\theta})$ directly by minimizing, over $\boldsymbol{\theta}$, the discrepancy between $\mathbf{\Sigma}(\boldsymbol{\theta})$ and the sample covariance matrix, $\mathbf{S}$. This development largely gave birth to modern structural equation modeling packages such as LISREL. The primary methods for estimation of the CSMs described by Joreskög (1978) minimize:

1. *Least Squares:* $L(\boldsymbol{\theta}) = \text{Tr}\left(\mathbf{S} - \mathbf{\Sigma}(\boldsymbol{\theta})\right)$

2. *Generalized Least Squares:* $L(\boldsymbol{\theta}) = \text{Tr}\left(\mathbf{I} - \mathbf{S}^{-1}\mathbf{\Sigma}(\boldsymbol{\theta})\right)$

3. *Maximum (Normal) Likelihood:* $L(\boldsymbol{\theta}) = \text{Tr}\left(\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{S}\right) - \log\left|\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{S}\right| - p$

where p is the dimension of $\mathbf{S}$. Each criterion can be maximized using standard derivative based methods. When normality is a reasonable assumption, the maximum likelihood method is generally preferable, since twice the log of the likelihood ratio,

$$\Lambda(\hat{\boldsymbol{\theta}}) = 2\left(L(\mathbf{S}) - L(\mathbf{\Sigma}(\hat{\boldsymbol{\theta}}))\right), \tag{1.8}$$

has an asymptotic χ_k^2 distribution where L is the normal likelihood function, $k = \frac{1}{2}p(p+1) - D$, and D is the number of elements in $\hat{\boldsymbol{\theta}}$. $\Lambda(\hat{\boldsymbol{\theta}})$ is generally used as a measure of fit of a given CSM. In addition, nested CSMs can be compared by calculating $\Lambda(\hat{\boldsymbol{\theta}}_1) - \Lambda(\hat{\boldsymbol{\theta}}_2)$, which has a χ^2 distribution with degrees of freedom equal to the difference in dimension between the two models.

Building upon the work of Joreskög (1978), many developments have made in CSM. Covariate effects on the covariance structure can be included [e.g. Satorra and Saris (1985)]. When the response variable is categorical and can be viewed as a thresholded continuous variable, CSM analysis of thresholded responses [e.g. Muthén and Muthén (2004)] is possible. When the dependent variable is continuous and not directly observed, Muthén (1984) describes a method of discovering class membership, known now as *latent class analysis*. This belongs to a more general class of CSM models known as *exploratory factor analysis* where the user does not have a priori beliefs about the association structure in the data. In addition, non-random missingness has been investigated in [Muthén et al. (1987)]. A review of many of the capabilities of modern computing packages for CSMs, particularly MPLUS, is given in Skrondal and Rabe-Hesketh (2005).

CSM provides a general method of either discovering meaningful patterns of association in the data or confirming previously posited hypotheses. The method is attractive in that models can be constructed so parameters are interpretable and correlation structures corresponding to some substantive theory can be formulated and tested in a unified way. One drawback, as noted by Bentler and Dudgeon (1996), is that the distributional assumptions are very seldom tested and can have a major impact on inference. While inference procedures that do not depend on normality [e.g. Browne (1984)] exist, not many are implemented in statistical packages for practioner use [Bentler and Dudgeon (1996)]. In addition, approximate identifiability issues in CSM, particularly in complicated models, are often very subtle and can lead to erroneous conclusions if gone unnoticed. The practioner would be well advised to test for such subtle identifiability issues by examining the fisher information matrix for approximate singularity at the converged point or checking covergence points from different starting values.

1.2.4 Log-linear models for multivariate binary data

Since no convenient analogue to the normal distribution exists for binary data, methods which do not assume an underlying continuous construct (e.g. random effects models, CSM) are relatively sparse. Much of the material below is reviewed in greater depth by Pendergast et al. (1996). One common specification for the joint distribution is that which is used in Zhao and Prentice (1990); the commonly termed log linear model:

$$\log\left(P(\mathbf{Y}_i = \boldsymbol{y}_i)\right) = \boldsymbol{\Psi}_i'\boldsymbol{y}_i + \boldsymbol{\Omega}_i'\boldsymbol{w}_i - A(\boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i). \tag{1.9}$$

where $\boldsymbol{w}_i$ is a vector of all pairwise and higher order cross products of $\boldsymbol{y}_i$, $\boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i$ are parameters, and $A(\boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i)$ is a normalizing constant. The elements of $\boldsymbol{\Psi}_i$ are interpreted as conditional probabilities, while the elements of $\boldsymbol{\Omega}_i$ are interpreted as log conditional odds ratio contrasts and completely determine the association structure; each are typically modeled by regression on number of discrete [e.g. Koch et al. (1977)] or continuous [e.g. Fitzmaurice and Laird (1993)] predictors. This model gives a straightforward way of testing covariate effects on the association structure, as well as a framework for determining the presence of higher order associations. For example, one can exclude anything beyond pairwise associations by specifying the corresponding elements of $\boldsymbol{\Omega}_i$ to be 0, and compare this with the unconstrained model. As noted by Pendergast et al. (1996), various incarnations of this model have been used:

- Use (1.9) directly but set third and higher order interactions to zero. This yields a sort of autoregressive binary model and has been studied by Besag (1974), among others.

- Transform model (1.9) so as to model marginal, rather than conditional probabilities, and do not place any constraints on the association structure [e.g.

Fitzmaurice and Laird (1993)].

- Model the marginal probabilities and assume only pairwise associations are nonzero. This is equivalent to the GEE model of Zhao and Prentice (1990), and is the MLE when the higher order log conditional odds ratios truly are 0.

As noted by Laird (1991), the marginal models which result from transforming (1.9) do not admit a similar form. A closely related method detailed by Carey et al. (1993) describes estimation of marginal means and pairwise odds ratios by alternating between logistic regression of each response on a number of covariates and on each other response in a cluster.

1.3 Description and outline

Methodologies involving each of the two data structures described in the introduction represent the dichotomy of the work presented in this dissertation. In chapter 2 multiple binary variables observed on independent individuals over time is the setup. In the bivariate case, a framework based on reparameterization of the sequence of 2×2 tables into the corresponding marginal probabilities and odds ratio. Each of the parameters are modeled as smooth functions of time and are estimated with a penalized maximum likelihood approach. In the general multivariate case an estimator based on conditional composite likelihood is used to estimate each of the pairwise log odds ratios as smooth functions of time. This approach does not require specification of the marginal probabilities or higher order associations, which are considered to be nuisance parameters. Each method is demonstrated on a set of binary measurements made on children from the MLS.

In Chapter 3 the focus is shifted to the analysis of a single variable observed on independent clusters of individuals, with the "cluster" in mind being a family. Building on the conventional models for longitudinal family data, which are reviewed in detail,

a unified framework for characterization of familial correlation is built. A key feature of this framework is that longitudinal aspects of the correlation structure due to both age and time, which are both argued to be potentially important, are accommodated. The extent to which we can characterize and disaggregate these two effects is explored through empirical studies. The resulting formulation reflects a previously unexplored pattern of resemblance in longitudinal family data. ML estimation is described, as well as details on identifiability, hypothesis testing, and computation. Results are also displayed on quantitative measurements of child temperament from the MLS data.

Chapter 4 concerns the extension of the model in chapter 3 to the binary case. The primary reason for this extension is that many traits of interest that would be expected to cohere within families are binary. Complexities of the model are detailed, including an explanation of why ML estimation is not possible in the binary case. To mitigate this problem, an alternative criteria based on composite likelihood is used. In this model it is found that binary data does not support the estimation of the full model detailed in chapter 3 and a simplified model is settled for. Details on the computation of the estimates in the simplified model are given, as well as some results to display the ability to recover the data generating population structure.

CHAPTER II

Functional analysis of Odds Ratio trajectories in multivariate binary data

2.1 Introduction and illustrative examples

We consider multivariate binary responses made over multiple time points for a sample of individuals. For example, in the longitudinal family study discussed below we will analyze ratings of a child on a number of behavioral measures by his or her mother. Such data can be viewed as arising from a sequence $p \times p$ contingency table with cell probabilities $P_t(x_1, ..., x_p)$ indexed by time, t. A common line of investigation is to examine the cell marginal probabilities and their pattern of change over time. However, it may also be of interest to examine the association structure between variables and its change over time, which is our focus here.

As a concrete example, we consider a longitudinal study of children with ages ranging from 3 to 18. A number binary behavioral measurements are taken, mostly relating to characteristics like anxiety and aggression that are of interest to researchers studying children exposed to adverse environments. In the data analyzed below, children considered at high risk for substance abuse are observed at roughly 4 time points.

In this data setting the mean trajectory is often of substantial interest. However, the mean trajectory may only tell part of the story. For example, the some behaviors

are considered "normative" for children of a particular age, so higher frequencies can have different underlying meanings at different ages. Associations between variables indicate whether certain behaviors tend to appear or disappear simultaneously at specific ages. A set of behaviors with high pairwise association may be viewed as defining a response to environmental adversity (or lack thereof) that is shared by a subset of the study population. If this association appears for only, say, younger children, then we can infer that this particular response pattern is one that originates at a particular developmental stage.

There exists some previous work on the analysis of multivariate binary data. Of primary importance is the log-linear model described in the introduction. The log-linear model, which will be discussed more later, only models the odds ratio as implied by the parameters in the model, rather than by explicit parameterization. For this reason we seek an alternative. In other work, Wang (1997) looked at odds ratios in a regression framework using smoothing spline ANOVA [Wahba et al. (1995)]. However, the "odds ratio" studied by Wang only quantifies changes in prevalence from one time to another, rather than the measure of association we will consider below. In addition there are a number of previous applications of composite likelihood methods, which will be the focus of the second half of this chapter, to dependent binary data [Heagerty and Lele (1998)] and longitudinal analysis [Molenberghs and Verbeke (2005)]. Methods for dimension reduction have also been developed [e.g. de Leeuw (2006)].

We begin by considering the case where $p = 2$, and develop a method to estimate the marginal probabilities and log odds ratio as smooth functions of age. The approach is based on penalized maximum likelihood estimation and borrows from the regularized function estimation techniques described by Ramsey and Silverman (1997). Second, we leave p unconstrained and described a conditional composite likelihood approach that is tractable and flexible in terms of how pairwise associations

15

and their temporal change is specified, and is not burdened with the need to model marginal rates and higher order associations. The higher order associations may have their own complex structure, but are of lesser or separate interest in many applications. The first method is demonstrated on the MLS data, with this analysis omitted on the second method due to substantial content overlap.

2.2 Penalized ML estimation for bivariate binary trajectories

A natural starting point is a cross-sectional analysis of a specific pair of behavioral measures "Item 1" and "Item 2" at a particular child age a. These ratings can be viewed as arising from a 2×2 table

$$
\begin{array}{cc}
 & \text{Item 2} \\
 & \begin{array}{cc} \text{Y} & \text{N} \end{array}
\end{array}
$$

$$
\text{Item 1} \quad \begin{array}{c|cc}
 & \text{Y} & \text{N} \\
\hline
\text{Y} & P_{11}(a;X) & P_{10}(a;X) \\
\text{N} & P_{01}(a;X) & P_{00}(a;X)
\end{array}
$$

where $P_{11}(a;X)$ denotes the probability that a child at age a is rated as symptomatic on both items, $P_{10}(a;X)$ denotes the probability that a child at age a is rated as symptomatic on item 1 but not on item 2, and so on. The symbol X denotes a vector of covariates.

Focusing on the longitudinal structure of these contingency tables, an important question is whether there are meaningful temporal patterns in the log-odds ratio

$$
\text{LOR}(a;X) \equiv \log P_{11}(a;X) + \log P_{00}(a;X) - \log P_{10}(a;X) - \log P_{01}(a;X) \quad (2.1)
$$

between two items over time. Such patterns reflect comorbidities between distinct behaviors that may vary in strength for children of different ages. Changes in the marginal frequencies

$$M_1(a; X) \equiv P_{11}(a; X) + P_{10}(a; X) \tag{2.2}$$

$$M_2(a; X) \equiv P_{11}(a; X) + P_{01}(a; X) \tag{2.3}$$

may also be of interest, but these could be evaluated in a univariate setting, while the bivariate setting is especially suited for studying associations between items. The marginal probabilities are considered nuisance parameters in this analysis. We also consider covariate effects. In particular, we will consider how covariates may either explain, or modulate the comorbidity between two items.

As a motivating example, consider the plots shown in Figure 2.1, which show the sample log odds ratio trajectory for two item pairs from the Michigan Longitudinal Study (MLS), which is discussed in more detail below. These trajectories are constructed by calculating the sample log odds ratio for all children measured at a particular age (left panel) or for the subgroups of children defined by a measure of the father's alcohol use (right). Since the children are generally measured at three-year intervals, these cross-sectional estimates use approximately one-third of the sample, roughly 200 children. The first item pairs shown in the figure suggest a degree of positive association between the items. For example, for older children, the log odds ratio between "feels unloved" and "screams" varies around 2.5. These positive associations are expected given the overlapping content of the items being rated (being broadly in the anxiety domain). In the right panel, it appears there is a qualitatively different trajectory in the two strata defined by father alcohol use. The high level of variability between time points may be acting to obscure interesting features of these trajectories. Much of the year-to-year fluctuation is unlikely to be real. By estimating the model parameters as smooth functions of time, it becomes possible to estimate these trajectories more precisely, which greatly aids in making interpretations.
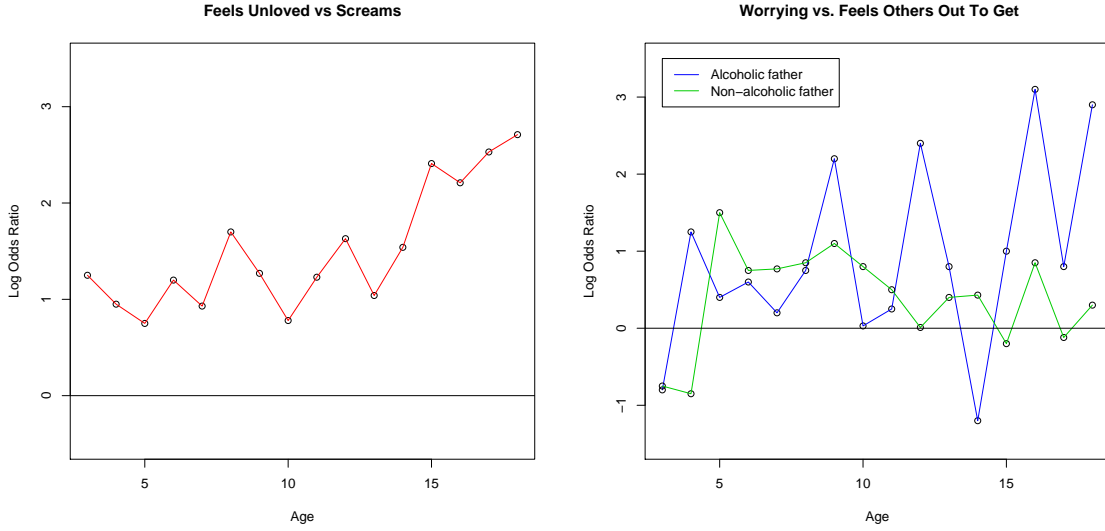
Figure 2.1: Sample log odds ratios using available data for each age, for two MLS item pairs. The left panel shows show log odds ratios for all children. In the right panel, the log odds ratios are calculated separately for two groups defined by their father's alcohol use level (blue: alcoholic father, green: non-alcoholic father).

2.2.1 Development of the model

The family of 2×2 tables given is in on-to-one correspondence with the three time varying parameters:

$$M_1(t;X) = P_{11}(t;X) + P_{10}(t;X) \tag{2.4}$$

$$M_2(t;X) = P_{11}(t;X) + P_{01}(t;X) \tag{2.5}$$

$$\text{LOR}(t;X) = \log P_{11}(t;X) + \log P_{00}(t;X) - \log P_{10}(t;X) - \log P_{01}(t;X). \tag{2.6}$$

Notice we have now substituted time (t) for age (a) for greater generality. The domains of M_1 and M_2 are $(0,1)$, while the domain of LOR is $(-\infty, \infty)$. We will parameterize these functions of time as

$$\mathrm{logit}\left(M_1(t;X)\right) \;=\; \mu_{1t} + \sum_k X_{tk}\phi_{1tk} \tag{2.7}$$

$$\mathrm{logit}\left(M_2(t;X)\right) \;=\; \mu_{2t} + \sum_k X_{tk}\phi_{2tk} \tag{2.8}$$

$$\mathrm{LOR}(t;X) \;=\; \theta_t + \sum_k X_{tk}\psi_{tk}. \tag{2.9}$$

When no covariates are present, there are no parametric constraints between (2.4), (2.5), (2.6) and (2.7), (2.8), (2.9). Covariate effects are assumed to be additive on either the logistic scale (for marginal probabilities) or the logarithmic scale (for the log odds ratio), with time-varying coefficients. For estimation, the trajectory parameters μ_{1t}, μ_{2t}, θ_t, ϕ_{1tk}, ϕ_{2tk}, and ψ_{tk} are viewed as smooth functions of t. We aim to construct estimators and inference procedures for these parameters that automatically adapt to the degree of smoothness in a particular population and to the size of the sample.

2.2.2 Relationship to log-linear modeling

The most widely used tool for studying complex patterns of association in binary data is the log-linear model discussed in the introduction. Suppose we let N_{ijt} denote the number of subjects responding at level i for item 1 and level j for item 2 at time t. Then

$$\log EN_{ijt} = \mu + \alpha_{it} + \beta_{jt} + \gamma_{ijt} \tag{2.10}$$

such that $\sum_i \alpha_i = \sum_j \beta_j = \sum_{ij} \gamma_{ijt} = 0$ defines a log-linear model. In this setting, the log odds ratio function is

$$\mathrm{LOR}(t) = \gamma_{11t} + \gamma_{00t} - \gamma_{10t} - \gamma_{01t}. \tag{2.11}$$

Analogous expressions for the marginal probabilities can be derived. Thus, the model

developed above is equivalent to this model, and it would be possible for either parameterization to work. However, our approach will directly model the marginal probabilities and log odds ratios as functions. Treating the α, β, γ as the primary functions of interest is less appealing, as these will not be the focus of our inference or interpretation. In addition, the penalized MLE we will describe below is not invariant to invertible re-parameterizations whereas the MLE is.

2.2.3 Computation of the estimates

First we show that each 2×2 table can be parameterized by the odds ratio and marginal probabilities. The log odds ratio at time t can be written in terms of $\Theta = (P_{11}(t; X), M_1(t; X), M_2(t; X))$:

$$
\begin{aligned}
\text{LOR}(t; X) &= \log P_{11}(t; X) + \log P_{00}(t; X) - \log P_{10}(t; X) - \log P_{01}(t; X) \\
&= \log P_{11}(t; X) + \log(1 + P_{11}(t; X) - M_1(t; X) - M_2(t; X)) - \\
&\quad \log(M_1(t; X) - P_{11}(t; X)) - \log(M_2(t; X) - P_{11}(t; X)), \quad (2.12)
\end{aligned}
$$

where

$$
\max(0, M_1(t; X) + M_2(t; X) - 1) \leq P_{11}(t; X) \leq \min(M_1(t; X), M_2(t; X)). \quad (2.13)
$$

As $P_{11}(t; X)$ moves through this range, $\text{LOR}(t; X)$ increases monotonically from $-\infty$ to $+\infty$. Therefore a specified value of the $\text{LOR}(t; X)$, $M_1(t; X)$, and $M_2(t; X)$ uniquely determines the response distribution at time t.

Let $R_{itk} \in \{0, 1\}$ be the response of subject i on item $k \in \{0, 1\}$ at time t. The bivariate outcome for subject i can be represented as

$$Y_{itab} = \mathcal{I}(R_{it1} = a) \cdot \mathcal{I}(R_{it2} = b). \tag{2.14}$$

The likelihood function can be explicitly parameterized in terms of Θ:

$$
\begin{aligned}
\mathbf{L}(\Theta) &= \sum_{itjk} Y_{it11} \log P_{11t}(X_{it}) + Y_{it10} \log P_{10t}(X_{it}) \\
&\quad + Y_{it01} \log P_{01t}(X_{it}) + Y_{it00} \log P_{00t}(X_{it}) \\
&= \sum_{itjk} Y_{it11} \log P_{11t}(X_{it}) + Y_{it10} \log(M_1(t; X) - P_{11t}(X_{it})) \\
&\quad + Y_{it01} \log(M_2(t; X) - P_{11t}(X_{it})) + Y_{it00} \log(1 + P_{11}(t; X_{it}) \\
&\quad - M_1(t; X) - M_2(t; X))
\end{aligned}
\tag{2.15}
$$

Since there is a one-to-one mapping between $P_{11}(t; X)$ and $\mathrm{LOR}(t; X)$ (for given values of $M_1(t; X)$ and $M_2(t; X)$), this likelihood is implicitly parameterized in terms of $\tilde{\Theta} = (\mathrm{LOR}(t; X), M_1(t; X), M_2(t; X))$. The parameterization in terms of $\tilde{\Theta}$ is of more interest to us since these are the quantities being estimated as functions. The details on calculating gradients of $\mathbf{L}$ with respect to $\tilde{\Theta}$ are given in Appendix A.

The integrated squared second derivative penalty is commonly used to regularize estimates of a functional parameter [Ramsey and Silverman (1997)]. We will discretize this penalty in the form of a second difference penalty, to adaptively smooth the estimates of all time varying parameters. For example, for the marginal probability intercept μ_{1t}, the penalty has the form

$$Q(\mu_1) \equiv \sum_t (\mu_{1,t+2} - 2\mu_{1,t+1} + \mu_{1,t})^2. \tag{2.16}$$

We will then use conjugate gradient optimization to maximize the penalized likelihood

$$L(\Theta) - \lambda \left(Q(\mu_1) + Q(\mu_2) + Q(\theta) + \sum_k Q(\phi_{1k}) + Q(\phi_{2k}) + Q(\psi_k) \right), \qquad (2.17)$$

where each parameter is vectorized over t. Each trajectory is penalized by the same weight λ. In fitting a single model, it would be just as easy to use different weights for different trajectories, however it is difficult to jointly define individual weights that perform well. Obtaining the derivatives of the penalty is straightforward, as they are simply quadratic functions.

To display the utility of this penalty for estimation, we generate simulated data from a population with demographic characteristics analogous to that of the MLS and log odds ratio function increasing from 0 and plateauing at 2. In Figure 2.2 we compare the functional estimate which maximizes (2.17) with the sample log odds ratios at each age connected with lines. We can see the functional estimate captures the true structure much more precisely. In particular, the key features of the true function are captured by the functional estimate, whereas the fact that the function plateaus at 2 is obscured by the simple empirical estimates.

As a second illustrative example we revisit the real data presented in Figure 2.1. As noted above, interesting features of the association patterns may be clouded by the "spikyness" of these plots. The functional estimates of these log odds ratio trajectories are given in Figure 2.3. We can see the increasing pattern in the item pair "Feels unloved", "Screams" is more clear with the functional estimate. In addition, the effect of the father alcoholism covariate on the association between "Worrying" and "Feels other out to get" is much more pronounced when the functional estimate is used. It appears that around age 12 there is a divergence between the strata, where children of alcoholic fathers experience an increasing level of comorbidity and the items become disassociated for children of non-alcoholic fathers.
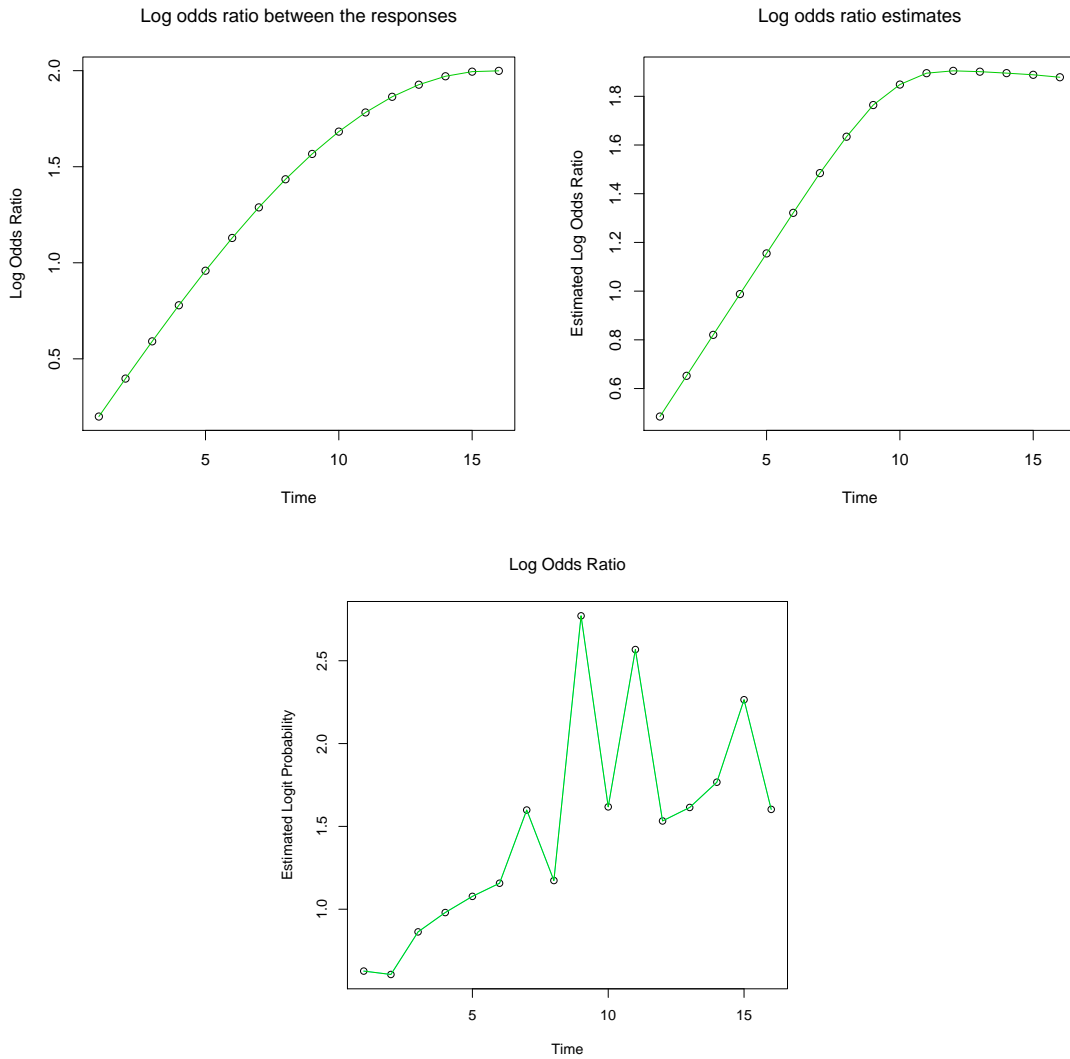
Figure 2.2: Ability to recover to the true data generating LOR(t) (top left) by estimating LOR(t) as a smooth function (top right) and by connecting the sample log odds ratios at each age (bottom).
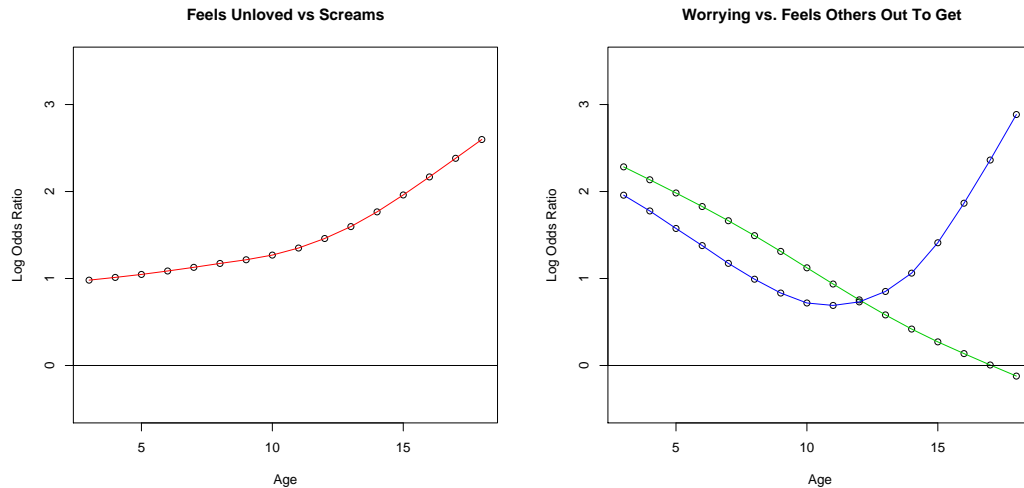
Figure 2.3: Functional log odds ratio estimates for two MLS item pairs. In the first item pair (left) we fit the model without covariates. In the second item pair (right), father alcohol use (blue: alcoholic father, green: non-alcoholic father) was used as a covariate.

2.2.4 AIC based Model Inference

The Akaike Information Criterion (AIC), adjusted for penalized estimation as in Shedden and Zucker (2008), will be used for identifying the smoothing parameter λ, for assessing covariate effects, and for assessing whether model parameters are time-invariant or time-varying. The standard AIC is equivalent to

$$L(\hat{\Theta}) - p \tag{2.18}$$

where p is the dimension of Θ, $\hat{\Theta}$ is the maximum likelihood estimate (MLE) of Θ, and $L(\cdot)$ is the log likelihood function. This estimates the out-of-sample expected value of the maximized log-likelihood

$$E_y E_x L(y|\hat{\Theta}(x)), \tag{2.19}$$

which is linearly related to the out-of-sample divergence between the fitted and true

likelihood functions. When a penalized estimate $\hat{\Theta}_\lambda$ is being used to estimate Θ, the number of free parameters must be reduced to account for the penalization. For example, in the setting described above, if the smoothing parameter $\lambda = 0$, then we are estimating $3 \cdot T$ parameters, where T is the number of distinct timepoints in the data set. On the other hand, if λ is large, the trajectories are effectively constrained to be linear, corresponding to only two parameters (the slope and the intercept) for each function, yielding only 6 effective parameters in the model.

It turns out that an adjusted degrees of freedom $\tilde{p}$ can be defined so that

$$L(\hat{\Theta}_\lambda) - \tilde{p} \tag{2.20}$$

estimates

$$E_y E_x L(y|\hat{\Theta}_\lambda(x)), \tag{2.21}$$

where $\hat{\Theta}_\lambda$ is the argmin of (2.17). Let $d = \lambda/n$, where n is the sample size, and let $T_1, \ldots, T_q$ be the eigenvalues of $F'\bar{I}^{-1}F$, where $\bar{I}$ is an estimate of the average Fisher information matrix across the sample, and F is the $p \times (p-2)$ second differencing matrix. The following adjusted degrees of freedom can be used in (2.20):

$$\tilde{p} = p - \sum_{j=1}^{q} \frac{2dT_i}{1 + 2dT_i}. \tag{2.22}$$

Based on (2.20), we can compare fits for different values of λ in a fixed model, and also compare across models. Our strategy will be to use a small set of λ values, typically $0, 10, 100, 1000$, which range from the MLE ($\lambda = 0$), to a fit that is effectively constrained to be linear ($\lambda = 1000$). After selecting the value of λ that maximizes (2.20) separately in each model, we then compare across models using these selected λ values.

There are two model comparisons of particular interest. First, using no covariates, we compare the model with a constant log odds ratio trajectory $\theta_t \equiv \theta$ to a model in which θ_t is allowed to vary with time. This is informative about whether the comorbidity between the two measured traits varies with a subject's age. Second, we compare the model with covariate effects for both the marginal probabilities and the log odds ratio, to the model in which only the marginal probabilities are linked to covariates (i.e., where $\psi_{tk} \equiv 0$). This is informative about whether apparent associations between traits are possibly due to known environmental factors.

For interpretation, it is useful to be able to calibrate differences in (2.20) between models. A rule of thumb stated by Burnham and Anderson (1998) is that AIC gaps of 1 or greater are indicative of "strong evidence" for the more complex model. To calibrate this against type 1 and type 2 error rates, we carried out a simulation study. Using the same sample size and pattern of observed measurements over time as the MLS, models with $\theta_t \equiv \theta$ were compared to models in which θ_t was allowed to vary as a smooth function of t. Based 1000 simulation replications, we calculated the empirical 95% range of the AIC gaps, and the proportion of simulated data sets in which the AIC gap exceeds one. The AIC gap is based on the AICs which correspond to the optimal choice of λ in each fitted model. The results, shown in Table 2.1, suggest that using an AIC gap of one unit in defining time-varying log odds ratio trajectories provides type 1 error rates below 0.05 with very high power, for the simulation models considered. In addition, the inference does not seem particularly sensitive to the structure of the marginal probabilities. Similar results (not shown) were observed for testing the covariate effects.

2.2.5 Delta method based pointwise confidence intervals

Following model selection, pointwise confidence bands for the log odds ratio and marginal probability trajectories can be calculated using an approximation based on

1	$\text{logit}\,M_1(t) \equiv \text{logit}\,M_2(t) \equiv 0$	$\text{LOR}(t) \equiv 1$	(-2.12,1.41)	.03
2	$\text{logit}\,M_1(t) \equiv -\text{logit}\,M_2(t) \equiv -3/2 + t/10$	$\text{LOR}(t) \equiv 1$	(-2.33,1.32)	.03
3	$\text{logit}\,M_1(t) \equiv \text{logit}\,M_2(t) \equiv 0$	$\text{LOR}(t) \equiv t/10$	(4.97,26.40)	1.00
4	$\text{logit}\,M_1(t) = -\text{logit}\,M_2(t) = -3/2 + t/10$	$\text{LOR}(t) \equiv t/10$	(1.83,21.50)	.99

Table 2.1: Calibration of AIC gaps to type I/II error rates. For each of four simulation models (column 1/2), the AIC gap (AIC's maximized over λ) was calculated comparing a model with time-varying log odds ratio to a model with time-invariant log odds ratio. The 95% range of the AIC gaps, and the proportion of AIC gaps exceeding 1 are shown in columns 3 and 4, respectively.

a Taylor expansion of the score function. Let $S(\Theta)$ denote the score function and let Θ_0 denote the true value of the parameter. The penalty functions of the form (2.16) can be written in matrix form $\mu' M \mu$, and the overall penalty function for Θ can be written $Q = \text{diag}(M, \cdots, M)$. A Taylor expansion of S around Θ_0 gives

$$S(\Theta) \approx S(\Theta_0) + \nabla S(\Theta_0)(\Theta - \Theta_0). \tag{2.23}$$

Let $\dot{Q}$ denote the derivative of the second differencing operator and λ the optimally chosen smoothing parameter. Evaluating (2.23) at $\hat{\Theta}_\lambda$, adding $\lambda \dot{Q}(\Theta_0 - \hat{\Theta}_\lambda)$ to each side, using that $S(\hat{\Theta}_\lambda) - \lambda \dot{Q}\hat{\Theta}_\lambda = 0$, and rearranging terms, we have

$$\hat{\Theta}_\lambda - \Theta_0 \approx [\nabla S(\Theta_0) - \lambda \dot{Q}]^{-1}[\lambda \dot{Q}\Theta_0 - S(\Theta_0)] \tag{2.24}$$

Replacing $\nabla S(\Theta_0)$ with its limit, $-I$, where I is the fisher information matrix at Θ_0, we have

$$\hat{\Theta}_\lambda - \Theta_0 \approx [I + \lambda \dot{Q}]^{-1}[S(\Theta_0) - \lambda \dot{Q}\Theta_0] \tag{2.25}$$

Taking the variance of both sides, we have that $\Sigma = \text{var}(\hat{\Theta}_\lambda)$ is approximately

$$\Sigma \approx [I + \lambda \dot{Q}]^{-1}\text{var}(S(\Theta_0))[I + \lambda \dot{Q}]^{-1}. \tag{2.26}$$

The fisher information is defined as the variance of the score function, thus $\text{var}(S(\Theta_0)) = I$ and we finally have

$$\Sigma \approx [I + \lambda\dot{Q}]^{-1}I[I + \lambda\dot{Q}]^{-1} \tag{2.27}$$

as a consistent estimator of Σ. A similar calculation can be done to derive the approximate bias but, in practice, we find that AIC selects λ values for which the bias is very small. In simulations (not shown), the 95% confidence intervals formed by ignoring the bias covered $> 90\%$ of the time. Therefore we will use non-bias-corrected intervals below.

2.2.6 Example: The Michigan Longitudinal Study

The Michigan Longitudinal Study [Zucker et al. (1996)], is a prospective long-term study of initially intact families at high risk of substance abuse. Behavioral assessments of children began at ages 3-5, and the study is ongoing today with the core group now at ages 18-20. The data used here are binary parent ratings on 34 behavioral items (coded "never true" or "sometime/always true") from the aggression and anxiety/depression subscales of the Child Behavior Checklist (CBCL) originated by Achenbach and Edelbrock (1983). The CBCL is the most widely used child behavioral assessment instrument in the United States. Using our model, we are able to estimate symptom endorsement probabilities and comorbidities for any two CBCL items, and understand how they vary with a child's environment, measured by the father's problem alcohol use.

The MLS includes data on 637 children in 323 families observed over some part of the age range 3 to 18. Interviews and questionnaire assessments with the mother of each child were carried out at 3-year intervals, with up to 6 assessments per child obtained, with a mean of 3.19 assessments. For the analysis presented here, the time variable t is taken to be the child's age at a given assessment (rounded to the

nearest year). Thus the fitted model parameters describe developmental trends in incidence and comorbidities among the behavioral items. The 637 children in our sample include some siblings. Dependencies resulting from family structure are not considered here.

The 34 behavioral items analyzed here cover a range of behavioral content spanning aggression to anxiety/depression. The variation in endorsement patterns among these items is complex but some prior expectations can be given. There is significant content overlap in certain items pairs, so some degree of comorbidity (i.e. positive odds ratios) is expected. Regarding marginal probabilities, some measurements are expected to decrease in frequency with age (e.g. "Demands attention", "Destroys own things") while other are expected to increase with age (e.g. "Unhappy", "Disobedient at school"). For all items, father alcoholism is expected to exacerbate symptoms.

Expectations regarding comorbidity patterns and covariate effects are harder to gauge. Co-occurrence can reflect a progression of related behaviors, such as when an argumentative child progresses to physical violence. But, since the ratings are given by the parent, co-occurrence can also reflect changes in the parent over time in terms of awareness and interpretation of their child's behavior. The father's alcohol use may create an environment which leads to greater rates of comorbidity, but may also lead to situation where the parents are less aware of their child's behavior. To minimize this last effect, we used mother ratings while assessing the effect of father alcoholism.

One possible explanation for an association between two items is population stratification. For example, if some stratifying variable which causes broadly different endorsement probabilities in different strata is unobserved (such as social status), an apparent association would result. If the stratifying variable was available and used as a covariate, this association may disappear or become less pronounced. In this inquiry, we only consider father alcohol use as a possible stratifying variable.

We began by considering all 561 item pairs, and identified the best fit for each pair

using the procedures described above. Specifically, fits were made for a sequence of values of the penalty parameter λ, and the value of λ that optimized (2.20) was chosen for each model. In this initial analysis, covariate data was not used. Consistent with expectations, the item pairs showed almost exclusively positive associations. While some item pairs were approximately independently endorsed for younger children, all showed positive associations for older children. Figure 2.4 summarizes the associations betweeen items. The lower left triangle of the plot indicates that most item pairs had estimated values $\text{LOR}(t)$ exceeding 1 for many time points. The upper triangle shows that a substantial number of item pairs had log odds ratio estimates exceeding 2.

Next, we compared models in which the $\text{LOR}(t)$ trajectory was time-invariant to models in which it was allowed to vary with time. Gaps in (2.20) exceeding one unit between the time-varying and time-invariant models were taken to indicate a significant time-dependence in the comorbidity trajectory. This threshold is supported by the simulation analysis described above. Following this approach, we found that $111/561$ item pairs showed a significant time varying pattern in the $\text{LOR}(t)$ trajectory. Nearly all of these trajectories followed increasing trends, indicating greater comorbidity between the items for older children. This analysis is based on the CBCL item ratings, without using the covariates. The results are given in the lower triangle of Figure 2.5.

Next we considered the effect of fathers alcohol use (assessed at entry to the study) as a covariate with possible links to the marginal probability trajectories and log odds ratio trajectories. We used an alcohol symptom measure coded as 0 (no alcohol dependence) to 3 (physical and psychological dependence). For parsimony, we parameterized the model using a single quantitative covariate taking on values 0, 1, 2, 3, so the fitted log odds ratio trajectories are θ_t, $\theta_t + \psi_t$, $\theta_t + 2\psi_t$, and $\theta_t + 3\psi_t$ for the four levels of alcohol use.

As an illustrative example, Figure 2.6 shows fitted marginal probability and log

odds ratio trajectories for groups 0 and 3 on two item pairs. It is important to note that there need not be a relationship between the marginal probabilities (rows 1 and 2 in the figure) and the degree of association (row 3 in the figure). While there is a weak tendency for children of alcoholics (COAs) to have higher symptom levels, the tendency of COAs to have higher comorbidity than non-COAs is much more pronounced. However the trend of this comorbidity is not consistent across the item pairs. For the nervous/stubborn pair, paternal alcohol use is most strongly linked to comorbidity in young children, while for worrying/fears someone might do something bad, paternal alcohol use is most strongly linked to comorbidity in older children. Note that while the pointwise confidence bands (the shaded region in the figure) largely overlap, the omnibus test for covariate effects indicates a significant difference in the trajectory patterns.

Using a gap of one unit in the AIC, as described above, we compared models with covariate effects on both the marginal probability trajectories and the log odds ratio trajectory to models with covariate effects only on the marginal probabilities. Formally, this is a test of the null hypothesis $\psi_{tk} = 0$. We found that 64/561 item pairs showed significant relationships between paternal alcohol use and the degree of comorbidity between the two items in the pair. For nearly all of these pairs, the greater comorbidity occurred in the families with alcoholic fathers. However, the timing of peak comorbidity varied across the items pairs. These results are summarized in the upper triangle of Figure 2.5.

2.3 Conditional likelihood estimation for multivariate binary trajectories

We now shift focus to an alternative model for the same data structure, but p is no longer constrained to be 2. Below we will describe a method which exploits
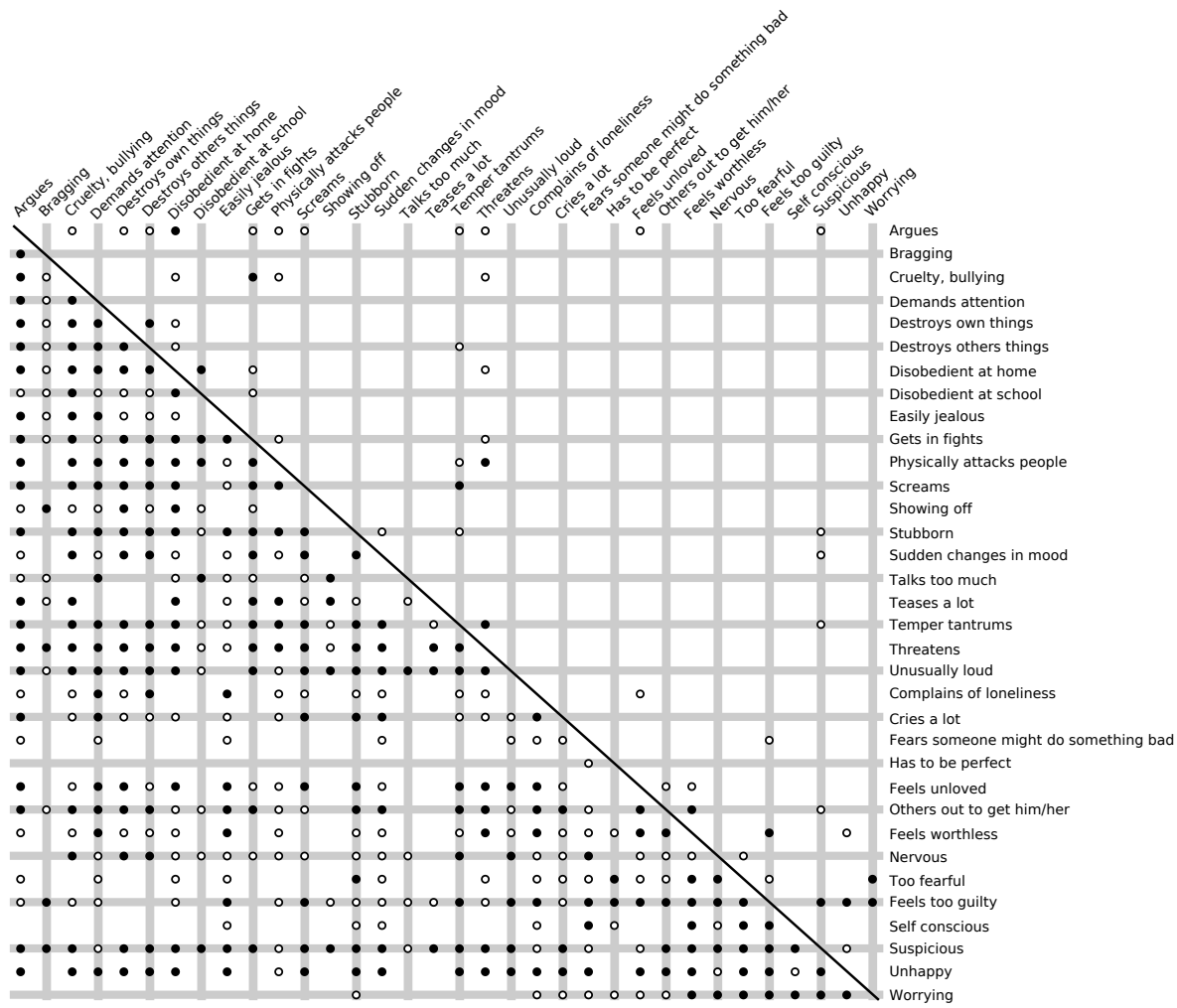
Figure 2.4: Summary of associations for all item pairs. The item pairs for which the specified number of ages exceed certain thresholds are marked. The lower triangle corresponds to LOR > 1 while the upper triangle is LOR > 2.
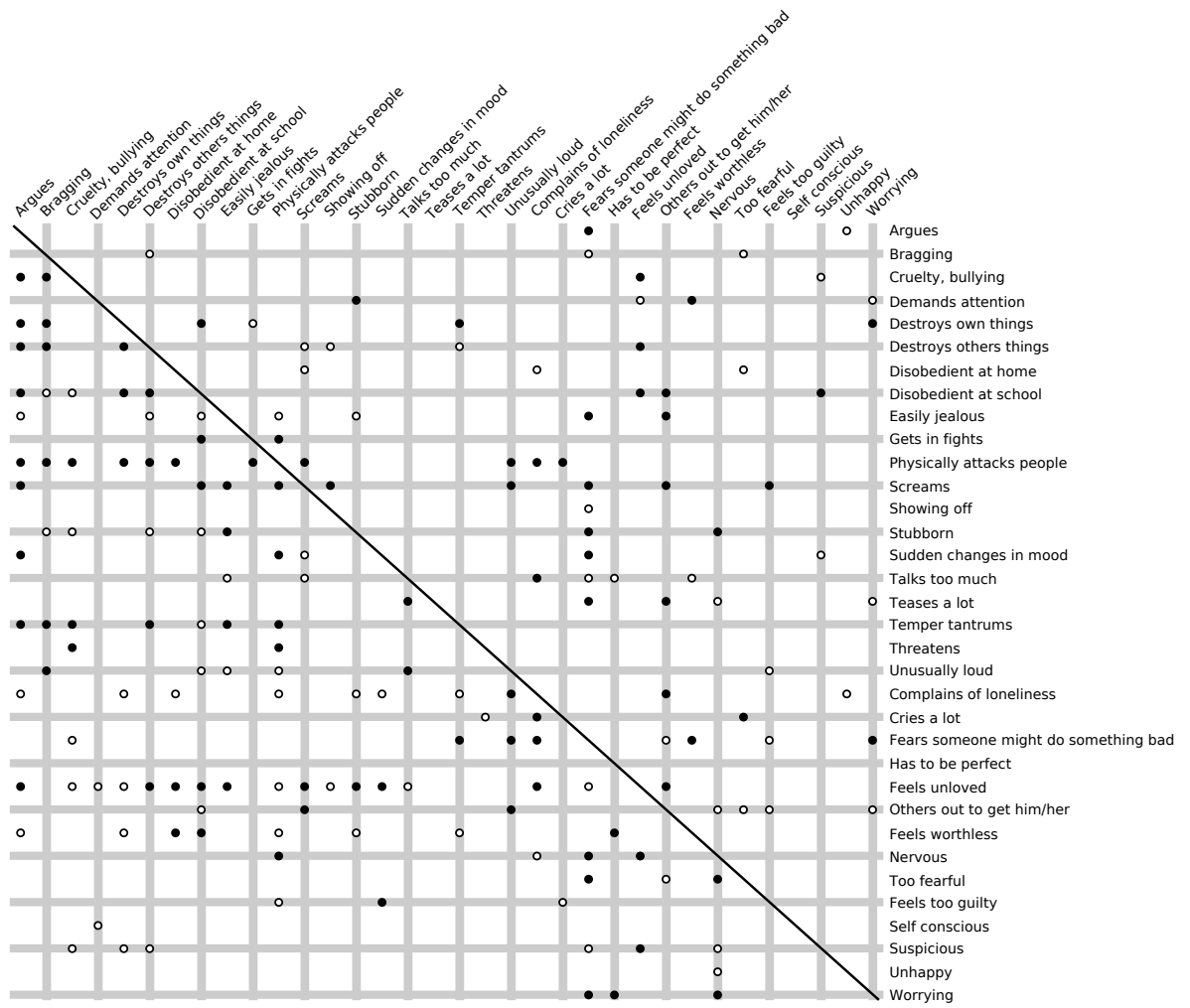
Figure 2.5: Summary of trajectory analysis for all item pairs. The lower triangle shows results for testing time-varying patterns in the log odds ratio. Pairs with moderate ($1 \leq \Delta\text{AIC} \leq 2$) and strong ($\Delta\text{AIC} > 2$) evidence are indicated with the symbols designated in the legend. The upper triangle shows the results for testing effects of father alcoholic status on the log odds ratio function.

33
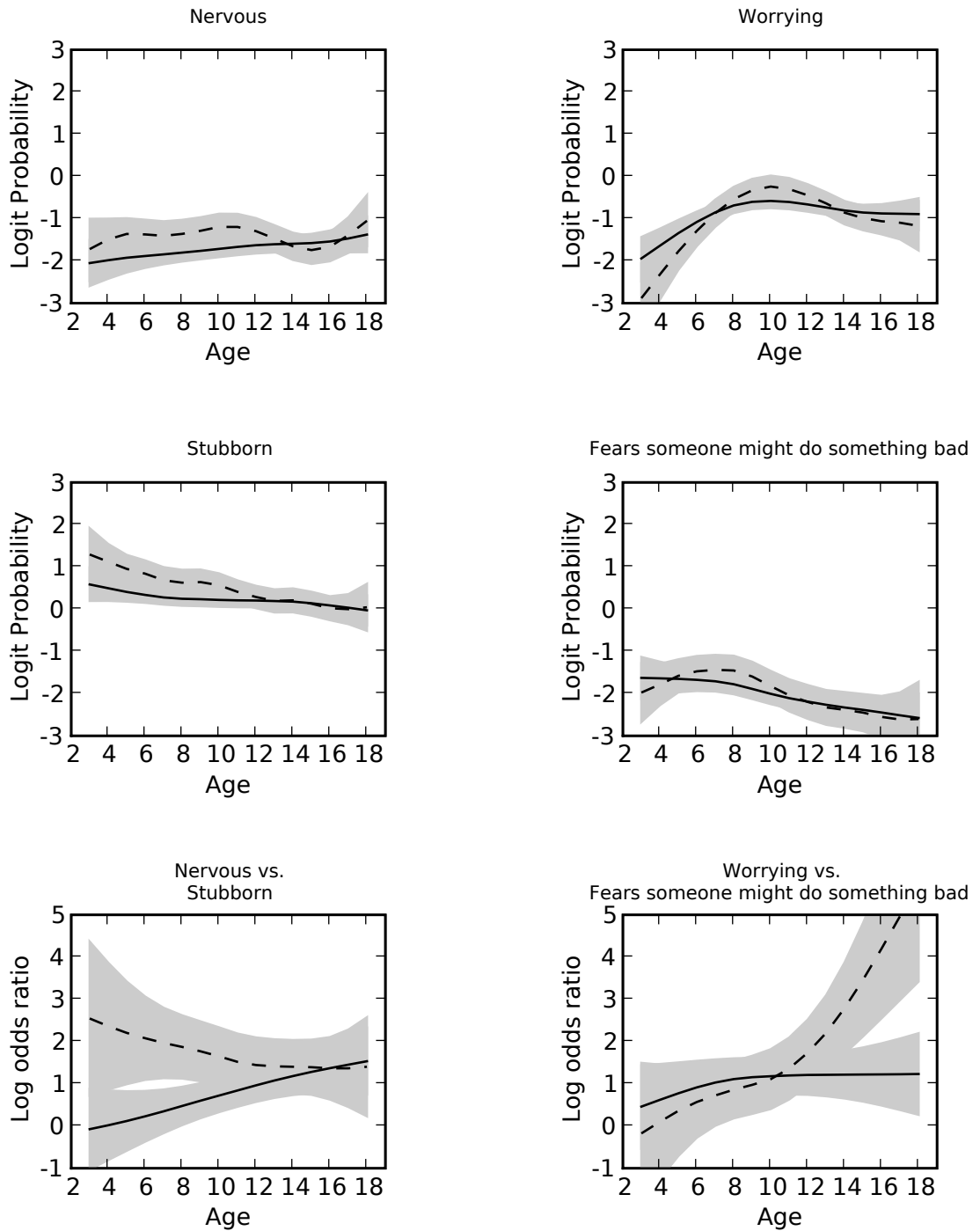
Figure 2.6: Fitted functions for two MLS item pairs. The top two rows show the fitted marginal logistic probabilities and the last row shows the fitted log odds ratio functions. The solid line is the fit for children of non-alcohlic fathers and the dashed line is for children of alcohol dependent fathers. Asymptotic pointwise confidence bands are shown in grey.

both conditional likelihood and composite likelihood to eliminate the need to model the marginal probabilities or the higher order associations, beyond the pairwise odds ratios.

2.3.1 The model and estimation framework

It is well known that in a 2×2 table each cell count, conditioned on the total, follows a hypergeometric distribution when the two binary variables are independent. This is the basis of Fisher's exact test. When there is association present, the cell counts conditioned on the total follow Fisher's non-central hypergeometric distribution.

If the log odds ratio for the 2×2 table is LOR, then the mass function for the count of the 1,1 cell, denoted C, conditioned on the row and column totals, r and c respectively, is

$$P(C = k|r, c) = \frac{\exp(\text{LOR} \cdot k)\binom{r}{k}\binom{n-r}{c-k}}{\sum_k \exp(Rk)\binom{r}{k}\binom{n-r}{c-k}}. \tag{2.28}$$

From this expression, we can derive the expected value $E(C|r, c)$. This expectation is not analytically tractable, but can be easily evaluated numerically. Since the subjects are iid, the marginal probability that any one subject contributes to the 1,1 cell is equal to $E(C|r, c)/n$, where n is the number of subjects.

To place this in the context of the data described above, let $Z_{i12} = \mathcal{I}(Y_{ij_1}(t) \cdot Y_{ij_2}(t) = 1)$, where $Y_{ij}(t)$ is the j'th variable measured on the i'th subject at time t. Let $n(t)$ be the total sample size at time t and $m_j(t) = \sum_{i=1}^{n(t)} \mathcal{I}(Y_{ij}(t) = 1)$ be the marginal total of variable j at time t. Then

$$P(Z_{i12} = 1|m_{j_1}(t), m_{j_2}(t)) = \frac{\sum_k k \exp(\text{LOR}_{12}(t) \cdot k)\binom{m_{j_1}(t)}{k}\binom{n(t)-m_{j_1}(t)}{m_{j_2}(t)-k}}{n(t)\sum_k \exp(\text{LOR}_{12}(t) \cdot k)\binom{m_{j_1}(t)}{k}\binom{n(t)-m_{j_1}(t)}{m_{j_2}(t)-k}}, \tag{2.29}$$

where $\text{LOR}_{12}(t)$ is the pairwise log odds ratio between variables j_1 and j_2 at time t.

We propose a method which treats $Z_{i12}(t)$ as the data, and uses composite likeliood methods to estimate the unknown parameter $\text{LOR}_{12}(t)$, using the product over all pairs of variales to define the composite likelihood contribution of a single subject. This leads to the following overall composite likelihood

$$\mathbf{CL}(\boldsymbol{\theta}) = \prod_i \prod_{j_1 < j_2} P(Z_{i12}(t)|m_{j_1}(t), m_{j_2}(t)), \tag{2.30}$$

where $\boldsymbol{\theta}$ denote a concatenation of all parameters underlying the set of log odds ratio trajectories. $\mathbf{CL}(\boldsymbol{\theta})$ is maximized over $\boldsymbol{\theta}$ to produce point estimates.

$\mathbf{CL}(\boldsymbol{\theta})$ may be thought of as a product of bivariate mass functions of the original data. The use of product of univariate or bivariate distributions that are compatible with an overall joint distribution is one of the standard ways to construct composite likelihoods [Cox and Reid (2004)] . There is precedent for using conditional distributions in forming the composite likelihood. For example, in spatial data it is common to form composite likelihoods by conditioning points on the values of those nearby [Besag (1974)] . In most applications all points are of interest, so each points enters as both a random quantity and a point being conditioned on. In our application, the marginal probabilities are not of particular interest and are thought to be uninformative for the estimation of log odds ratio, thus they only appear as conditioning variables in this analysis.

The general model allows each odds ratio trajectory, $\text{LOR}_{12}(t)$, to vary freely across the variable pairs over time. To reduce the number of parameters and impose smoothness, the trajectories can be parameterized in terms of splines or other basis functions. For example,

$$\text{LOR}_{12}(t) = \sum_{m=1}^{Q} \beta_{12m}\phi_m(t) \tag{2.31}$$

for some set of basis functions ϕ_m. The number of parameters grows linearly in the number of variables, thus the number of parameters may become impractically large for higher dimensional datasets. In that case we propose a dimensionally constrained model where the variation of each $\text{LOR}_{12}(t)$ is constrained to a fixed number q dimensions. Specifically,

$$\text{LOR}_{12}(t) = \eta(t) + \sum_{m=1}^{q} \gamma_{12m}\xi_m(t), \tag{2.32}$$

where an analogous basis expansion is used to model $\eta(t)$. The interpretation of this constrained parameterization is analogous to PCA. In the case where $q = 2$, which is the extent to which we demonstrate this model here, the constraint that $\sum_{m=1}^{q} \gamma_{12m} = 0$ suffices to ensure identifiability of the resulting coefficients. For $q > 2$, a potential identifiability issue exists. However, any solution will result in a smooth estimate of $\text{LOR}_{12}(t)$ for each item pair, which is the goal of this model.

Since (2.29) only depends on $\text{LOR}_{12}(t)$, there is no need to parameterize or estimate the other properties of the joint distribution of Y. In addition, full likelihood methods may be difficult or impractical for multiple timepoints or variables. The optimization of (2.30) is straightforward and inference procedures for composite likelihoods can be applied for model selection.

To fit the model, it is not necessary for the data to be balanced. Since the log odds ratio are parameterized as smooth functions of time, the model implies a likelihood for any observed point within the range of the data. Treating the timepoints as random, and thus the "missing" timepoints as "missing at random" is necessary for this procedure to give meaningful results [Little and Rubin (1987)]. An additional concern arises if the data are very irregularly observed and one is forced to condition of small marginal counts, which may lead to lower efficiency.

First derivatives of the log composite likelihood with respect to the model parameters are straightforward to calculate and implement on a computer. Therefore, we

can use gradient-based methods to optimize (2.30). When considering the saturated model in which in which the log odds ratio trajectories $\text{LOR}_{12}(t)$ are treated as free variables over all j_1, j_2, t values, the gradients of (2.30) can be used directly for optimization. This corresponds to separate estimation for each of the 2×2 tables defined by each item pair. When the basis representation or dimensionally-constrained representation described in (2.31) and (2.32) is used, gradients are easily calculated using the chain rule.

2.3.2   Inference

The key inferential questions considered regard determining whether a given trajectory constant at 0, and whether it is constant over time. As mentioned previously, it is expected that many of these items will be associated due to their overlapping content. Documenting these dependencies may be of interest, but how they change over time is much less appreciated. Thus, the primary interest is in assessing whether a given trajectory is constant.

Familiar inference procedures including likelihood ratio tests, Wald tests, and score tests have been extended to the composite likelihood setting [e.g. Varin and Vidoni (2005), Cox and Reid (2004)]. Here we will focus on nested model comparisons using the log likelihood ratio test statistic (LR) constructed from the optimized log composite likelihood functions. The use of composite likelihoods and the use of conditional likelihoods both affect the sampling behavior of the LR statistic. To account for these effects, we will use simulation to calibrate the LR statistic values.

Simulation based calibration will be most effective when the sampling distribution does not depend strongly on the structures of the generating model and the two working models being compared, except through the numbers of free parameters in the working models. To explore this, we considered generating models with marginal probabilities following

$$\text{logit}\left(P(Y_{ij}(t) = 1|\alpha_{ij})\right) = \theta + \alpha_{ij}, \tag{2.33}$$

and with log odds ratio trajectories $\text{LOR}_{12}(t) \equiv E$, where E is a given effect size. The θ parameter in (2.33) controls the marginal mean of $Y_{ij}(t)$, and the α_{ij} are iid $N(0, \sigma_\alpha^2)$ random effects introducing within-subject correlations. We then used simulation (with 500 replications per population) to estimate the increase in the optimized log composite likelihood when comparing a constant model for $\text{LOR}_{12}(t)$ to a linear model, and from a linear model to a quadratic model. Since the constant model is true, these results reflect the null sampling distribution of the LR statistic when adding either one or two irrelevant parameters. The results of this simulation are shown in Table 2.2.

Simulation results suggest that the calibration of the LR statistic is more sensitive to the marginal probabilities, and not to serial dependencies, the sample size, or to the size of the log odds ratio being estimated. This suggests that as long as the data set is large enough so that the estimates of marginal probabilities that determine the sampling distribution are accurate, simulation can be relied upon to calibrate the LR statistic.

2.3.3 Simulation studies

We first considered the following generating model for $p = 2$ variables that uses random effects to induce associations between variables and between time points:

$$\text{logit}\left(P(Y_{ij}(t) = 1|\alpha_{ij}, \gamma_{it})\right) = \mu_j + \alpha_{ij} + \gamma_{it}. \tag{2.34}$$

where $\boldsymbol{\alpha}_i \sim N(0, \sigma_\alpha^2 \mathbf{I})$, $\gamma_{it} \sim N(0, \sigma_{\gamma t}^2)$, and all random effects are independent. The between-variable associations are controlled by the γ_{it}, with the possibility of having

E	θ	σ_α^2	250 CL	250 LQ	500 CL	500 LQ
0	-1	0	0.32	0.33	0.35	0.32
0	-1	1	0.33	0.33	0.35	0.30
0	-1	2	0.35	0.33	0.28	0.33
0	0	0	0.65	0.65	0.75	0.60
0	0	1	0.60	0.63	0.66	0.66
0	0	2	0.63	0.61	0.58	0.54
1	-1	0	0.36	0.32	0.31	0.33
1	-1	1	0.28	0.34	0.37	0.37
1	-1	2	0.32	0.39	0.36	0.41
1	0	0	0.55	0.60	0.48	0.52
1	0	1	0.58	0.55	0.68	0.50
1	0	2	0.59	0.48	0.53	0.54

Table 2.2: Monte Carlo estimates of 95th percentiles of the log composite likelihood ratio statistic when comparing a constant to a linear model (CL) and when comparing a linear to a quadratic model (LQ) for $\text{LOR}_{12}(t)$, when the truth is constant. The generating models have various effect sizes E, and various intercepts θ and random effect variances σ_α^2.

different strengths of association at different time points. The α_{ij} induce within-subject associations over time.

We considered $\sigma_\alpha^2 = 0$, 0.5, and 1, sample sizes $N = 250$ and 500, and either 5 or 10 time points. The intercepts were always $\mu_1 = -0.5$ and $\mu_2 = 0.5$, and $\sigma_\gamma^2(t)$ always varied linearly from 0 to 1 over the observed time interval. The log odds ratio trajectory $\text{LOR}_{12}(t)$ was modeled using (2.31) as a linear function. The simulation results are based on 500 simulation replications. A separate simulation based on a sample size of 10^6 values was used to numerically approximate the true value of $\text{LOR}_{12}(t)$ for these populations.

Based on the simulation results, we estimated the bias and standard error of the log odds ratio estimate at each T_j for which data was observed. The results suggest very low bias, with the median estimated bias over all time points and all generating models being less than 0.01, and the maximum bias being 0.04. The standard errors of the estimates are shown in Table 2.3. The standard errors for sample size $n = 250$

n	σ_α^2						SD					
		t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}
250	0	0.19	0.14	0.11	0.11	0.14	0.19					
500	0	0.14	0.10	0.08	0.08	0.11	0.14					
250	0.5	0.19	0.15	0.12	0.12	0.15	0.20					
500	0.5	0.14	0.10	0.08	0.08	0.11	0.15					
250	1	0.20	0.15	0.12	0.11	0.14	0.19					
500	1	0.14	0.10	0.08	0.08	0.10	0.13					
250	0	0.15	0.13	0.11	0.10	0.09	0.08	0.08	0.09	0.11	0.13	0.15
500	0	0.11	0.09	0.08	0.07	0.06	0.06	0.06	0.07	0.08	0.09	0.10
250	0.5	0.15	0.13	0.11	0.10	0.09	0.08	0.09	0.10	0.12	0.14	0.16
500	0.5	0.10	0.09	0.08	0.07	0.06	0.06	0.06	0.07	0.08	0.09	0.11
250	1	0.15	0.13	0.11	0.10	0.09	0.08	0.09	0.10	0.12	0.13	0.15
500	1	0.10	0.09	0.08	0.07	0.06	0.06	0.06	0.06	0.07	0.09	0.10

Table 2.3: Standard deviations of $\text{LOR}_{12}(t)$ parameters for model (2.34). The sets of either five or ten SD values are the standard errors of the $\text{LOR}_{12}(t)$ estimates for the observed values of t.

are very nearly 40% higher than those in the same generating model using sample size $n = 500$, as expected. Also, note that the standard errors are considerably smaller than what would be obtained from an analysis using sample log odds ratios at each time point. These standard errors would be at least $4/\sqrt{250} \approx 0.25$ for sample size 250 and $4/\sqrt{500} \approx 0.18$ for sample size 500.

The random effects α_{ij} introduce associations that are not modeled in the composite likelihood (2.30). As intended, the within variable/between time point log odds ratio tracked closely with the value of σ_α^2, with values approximately equal to 0, 0.25, and 0.5, for σ_α^2 equal to 0, 0.5, and 1. The results in Table 2.3 show that introduction of this type of unmodeled dependence does not introduce substantial additional variance (or bias) into the estimation of the log odds ratios of interest.

Next we considered a generating model with $p = 5$ variables and sample size $n = 500$, using the constrained model (2.32) with $q = 1$ directions of deviation from the mean.

$$\text{logit}\left(P(Y_{ij}(t) = 1|\beta_{ij})\right) = 1 + \theta_{ij}t, \tag{2.35}$$

where the vector $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip}) \sim N(0, \mathbf{R})$, and t ranges over a grid of five points in $(0,1)$. The matrix $\mathbf{R}$ had unit diagonal values, and pairwise correlation 0.8 among the first three variables with all other off-diagonal entries equal to 0. Given this structure, we expect all pairs among the first three variables to have increasing log odds ratio trajectories, with all other pairs independent. All other correlations were zero. Thus we expect all pairs among the first three variables to show increasing associations over time, and all other variable pairs should be independent.

In the estimated model, the trajectories $\eta(t)$ and $\xi_1(t)$ are parameterized as linear functions of time. Since there are $\binom{5}{2} = 10$ pairs of variables, there are 10 γ_{12m} values in (2.32), for a total of 14 parameters. The results are shown in Table 2.4. The bias is seen to be quite low, and the standard errors tend to be considerably less than the best value of 0.18 that could be obtained from a simple model-free analysis, as discussed above.

Next we examine the utility of the LR test to distinguish constant from time-varying odds ratio trajectories. We simulated data with $p = 2$ variables, 5 time points, and sample sizes of 250 or 500, for which the population value of $\text{LOR}_{12}(t)$ had the form $E \cdot t$. Various values for the effect size E were considered, and marginal frequencies were always fixed at $1/2$. Using the LR statistic derived from the optimized log composite likelihoods, we compared models for which $\text{LOR}_{12}(t)$ was modeled as either a constant, linear, or quadratic function of time.

The results of the simulation study described above are shown in Figure 2.7. The "LC", "QC", and "QL" columns show results for comparison of linear to constant models, quadratic to constant models, and quadratic to linear models, respectively.

	t_1	t_2	t_3	t_4	t_5		t_1	t_2	t_3	t_4	t_5
	-0.07	0.08	0.23	0.38	0.53		-0.01	-0.01	-0.00	0.01	0.01
1/2	-0.01	0.05	0.19	0.36	0.57	3/4	0.00	-0.00	-0.00	-0.00	0.00
	0.12	0.08	0.08	0.12	0.17		0.09	0.06	0.07	0.11	0.15
	-0.07	0.08	0.22	0.37	0.51		-0.01	-0.00	-0.00	0.00	0.01
1/3	0.00	0.05	0.18	0.36	0.58	1/5	0.00	-0.00	-0.00	0.00	-0.01
	0.11	0.08	0.09	0.12	0.16		0.09	0.06	0.08	0.12	0.17
	-0.07	0.08	0.23	0.37	0.52		-0.01	-0.01	0.00	0.01	0.02
2/3	0.00	0.04	0.18	0.36	0.57	2/5	-0.00	0.01	0.01	-0.00	0.00
	0.11	0.08	0.09	0.12	0.17		0.08	0.07	0.09	0.13	0.17
	-0.00	0.00	0.01	0.01	0.02		-0.01	-0.01	-0.01	-0.01	-0.01
1/4	0.00	0.00	-0.00	0.00	-0.01	3/5	0.00	-0.01	0.00	-0.01	0.00
	0.10	0.07	0.08	0.12	0.17		0.09	0.07	0.09	0.13	0.18
	-0.00	0.00	0.01	0.01	0.02		-0.00	-0.01	-0.01	-0.01	-0.01
2/4	-0.01	0.01	-0.00	-0.00	-0.00	4/5	-0.00	0.00	-0.01	0.00	0.00
	0.08	0.07	0.09	0.13	0.18		0.10	0.07	0.08	0.12	0.17

Table 2.4: Results for each odds ratio trajectory $\text{LOR}_{12}(t)$ are shown in the three lines preceded by j_1/j_2. The first, second, and third rows in each group correspond to the expected values of the estimated log odds ratios, the true log odds ratios, and the standard errors of the estimated log odds ratios, respectively. The five columns in each group of lines correspond to the five time points.
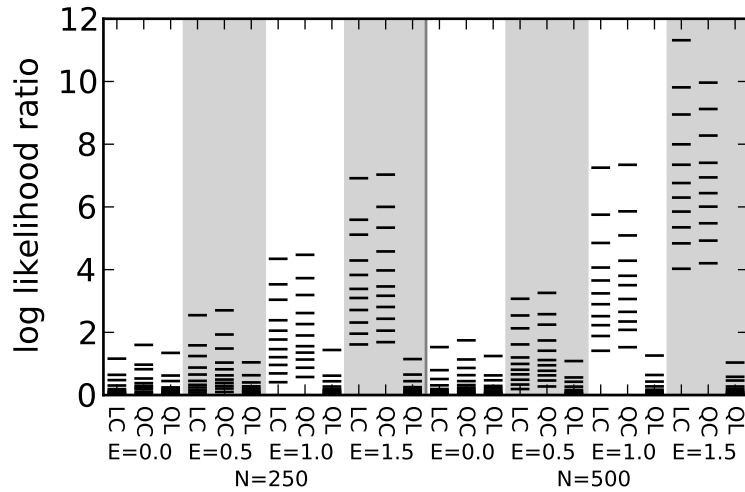
Figure 2.7: The LR statistic values for comparison of linear and constant models (LC), quadratic and constant models (QC), and quadratic and linear models (QL) are shown for 300 simulation replications. The horizontal lines show the deciles of the sampling distribution from 0.1 to 0.9, along with the 95th and 99th percentiles.

When $E = 0$, the constant model is true, and the 95th percentile of the LR statistic falls at around 0.6, consistent with Table (2.2). For larger effect sizes, the linear model becomes favored over the constant and quadratic models.

2.4   Discussion

When presented with multivariate measurement vectors on each member of a sample, a natural starting point is to consider how the univariate mean responses vary with respect to covariates of interest. However it can happen that the strongest covariate effects occur at the level of associations between responses rather than in their mean levels. In a longitudinal setting, the degree of association and its relation to covariates may change over time. The aim of this work is to capture all of these relationships in a function-based estimation framework. While we did restrict our attention to the analysis of binary data in this setting, there is no reason why

analogous models could not be formulated for continuous data, using the correlation (on some transformed scale) as the measure of association.

There is a long tradition of studying complex associations between categorical responses using log-linear modeling techniques, but relatively little work has been done to adapt these methods to the context where the model parameters are appropriately viewed as functions. On the other hand, in functional data analysis the main emphasis has been placed on efficient estimation of functions, but most of this work is in the context of regression models describing the mean response and covariate effects for a univariate dependent variable. Each of the methods used here represent an attempt at integrating these two approaches.

The first approach was based on penalized ML estimation for bivariate trajectories. The marginal probability functions, as well as the log odds ratio function, were modeled as non-parametric functions of age, with regularity imposed by ensuring the integrated squared second derivative of each function was not too large. Using log-linear modeling notation, it would be possible to extend this to more than two binary responses, or more generally to ordered or unordered categorical responses. For more than two binary responses, the number of parameters being optimized over in the above model may be prohibitively large. For cases such as this, an alternative approach may be more suitable.

The second method presented was based on a composite conditional likelihood estimator. One concern in the use of such methods is the potential loss of efficiency when compared to full likelihood based approaches. For complex data, this loss may be offset by the potential biased introduced by incorrect specification of nuisance parameters (such as higher order associations, in our case). Due to the parameter explosion alluded to above, practical methods are required to constrain certain aspects of the joint distribution. An advantage of our approach is that we can model associations of interest without making arbitrary constraints on nuisance parameters.

In the second approach discussed here, there is an interplay between the use of conditional and composite likelihoods. Loss of efficiency in composite likelihood estimation is largely driven by the dependency between the blocks of data being explicitly modeled. Conditioning on the table margins presumably leads to substantially lower dependency among these blocks than would occur with an unconditional likelihood. The results shown here suggest that the loss of efficiency when compared to full ML estimation is not substantial. Using the traditional unconditional log odds ratio estimate, the standard errors at a particular time point are $4/\sqrt{250} \approx 0.25$ and $4/\sqrt{500} \approx 0.18$ for the two sample sizes considered here. Even if the data for six time points were observed independently at a single point, the standard errors would be $4/\sqrt{250 \cdot 6} \approx 0.1$ and $4/\sqrt{500 \cdot 6} \approx 0.07$. For the upper rows of Table 2.3, the standard errors from the composite/conditional likelihood are seen to be quite similar to these values at the center of the time range, and are no worse than the standard errors for single time point analysis at the edges of the time range.

CHAPTER III

Multiple timing variable models for longitudinal family data

3.1   Introduction

Certain traits ranging from behaviors, to physical characteristics, to disease processes cannot be summarized by a single number. In particular, some traits change as a function of some independent and continuous variable. Such traits were first called *infinite dimensional* in Kirkpatrick and Heckman (1989) and have also been referred to as *function valued* in other literature [e.g. Kingsolver et al. (2001)]. Function valued traits can only be properly characterized by observation along a range of the independent variable. One common example of a function valued trait in ecology and genetics involves so-called *norms of reaction* [e.g. Scheiner (1993)], where a phenotypic expression is a function of particular environmental variables is of interest. Another readily understandable example, which will be the focus of this inquiry, is when a trait varies as a function of a longitudinal quantity, such a age and/or observation time.

When studying function valued traits, analysis of longitudinal family data is crucial to a more precise understanding of the mechanisms driving human traits and how they cohere within a family. It is well documented that factors affecting familial correlation on a trait can accumulate or diminish as a subject ages [e.g Gauderman and

Morrison (2000), Lindenberger et al. (2008)]. In addition, the fact that researchers study reaction norms reflects the fact that many trait expressions are modulated by events occurring in the subject's environment. In either case, as noted by Gauderman and Conti (2005), non-longitudinal observation would be insufficient to capture the etiology of the measured trait. Existing literature on the modeling of family data, which we will now review, is primarily comprised of quantitative genetics models. However, it should be noted that we are building a general framework for modeling familial association where the separation of genetic from environmental sources of correlation is: a) not of particular interest, and b) not identifiable with the data structure we have.

3.1.1 Background on models for longitudinal family data

The issue of modeling longitudinal family data was the subject the Genetic Analysis Workshop 13 [Almasy et al. (2003)], where several approaches to analysis were presented. Typical approaches fall into one of two classes. The first class is comprised of two step procedures which reduce the longitudinal data to one response per subject, and then analyze the reduced data in the second stage. This type of data reduction occurs in many fields; for example, in the area of addiction research "age of first drunk" is used as a predictor of future addictive behavior [e.g. Corte and Zucker (2008)]. A drawback of these approaches is that they require all longitudinal features of interest to be pre-specified rather than determined by the data. For this reason we give no further attention to such approaches. The second class of models explicitly takes into account the longitudinal structure in the data. Before hashing out these approaches in greater detail we digress with a discussion of classical genetics models.

Consider observation of a quantitative trait for a single family with n children $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$. According to classical genetics [e.g. Kempthorne (1969), Falconer (1981)], we model $\mathbf{Y}$ as

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{g} + \mathbf{e} \tag{3.1}$$

where $\boldsymbol{\mu}$ are fixed effects and $\mathbf{g}, \mathbf{e}$ are genetic and environmental random effects, respectively. To illustrate the use of this model in the simplified case of no gene-by-environment interaction, we can decompose the variance in the responses as

$$\text{var}(\mathbf{Y}) = \sigma_G^2 \mathbf{R} + \sigma_E^2 \mathbf{1}_n + \sigma_\varepsilon^2 \mathbf{I}_n \tag{3.2}$$

where $\mathbf{1}_n$ in an n-by-n matrix of 1's, $\mathbf{I}_n$ is the n-by-n identity matrix, and $\mathbf{R}$ is the matrix of relatedness (e.g. $\mathbf{R}_{ij} = 1/8$ for two cousins). The intuition here is that all covariance between family members is due to shared environment effects (E), genetic (G) similarity effects. Additional variance is absorbed by unshared effects (ε). The shared environment contributes to the everyone's variance/covariance equally, and the genetic effects contribute to the covariance based on how closely related the individuals are. In principle $\mathbf{e}$ can be broken into a number of identifiable parts [Visscher et al. (2008)]. Here we have only split $\mathbf{e}$ into shared and unshared environment but, for example, if we had repeated measurements there should be an additional term for individual environment, shared across time points.

A primary quantity of interest in classical genetics is the *broad sense heritability*:

$$H^2 = \frac{\sigma_G^2}{\sigma_E^2 + \sigma_G^2 + \sigma_\varepsilon^2}. \tag{3.3}$$

That is, the proportion of phenotypic variance due to genetic variation. When only data on subjects of a fixed level of relatedness, r, is available, there is an identifiability issue with H^2, since

$$\phi \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} = \phi(1-r) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \phi r \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{3.4}$$

for any number ϕ. Since the partitioning of variance is not unique, one has a somewhat arbitrary choice to make regarding the values of σ_E^2, and σ_G^2. In this case, H^2 is not uniquely identified; only

$$\tilde{H}^2 = \frac{\sigma_G^2 + \sigma_E^2}{\sigma_E^2 + \sigma_G^2 + \sigma_\varepsilon^2}. \tag{3.5}$$

is identified, which is the intra-class correlation coefficient with family as the grouping variable. When σ_E^2 can be assumed to be 0, $\tilde{H}^2$ reflects genetics; otherwise it is nothing more than a measure of familial resemblance.

The above setup is the basis for the second stage in the two-step models alluded to above, and provide the necessary perspective to extend family data models to a longitudinal setting. Commonly, observations on function valued traits at various values of the independent variable are treated as distinct, correlated characters [e.g. Tatar et al. (1996)]. Traditional multivariate techniques are used to extend model (3.1) to the case of correlated traits [e.g. Lande and Arnold (1983), Riska et al. (1985)]. To illustrate this suppose you had n siblings observed at T timepoints with data $\mathbf{Y} = (Y_{11}, Y_{12}, ..., Y_{1T}, Y_{21}, ..., Y_{2T}, ..., Y_{nT})$. Then the natural extension of (3.2) is

$$\operatorname{var}(\mathbf{Y}) = \mathbf{R} \otimes \mathbf{G} + \mathbf{1}_n \otimes \mathbf{E} + \mathbf{I}_n \otimes \boldsymbol{\sigma}^2 \tag{3.6}$$

$\mathbf{G}, \mathbf{E}$ are the $T \times T$ covariance matrices of the genetic and environment random effects, $\boldsymbol{\sigma}^2$ is the T-length vector of residual variances for the trait at each time, and $\otimes$ denotes the Kronecker product. This model is very general but the parameter explosion as the number of time points increases coupled with potential interpretation problems make the multivariate approach impractical. In addition, as noted by Kirk-

patrick and Heckman (1989), multivariate techniques do not necessarily exploit the ordering inherent in function valued traits and typically have less statistical power than methods which do account for the ordering.

Many methods exist that go beyond the multivariate approach and exploit the time ordering to borrow information across observations. For notational simplicity we consider the trait value for a single subject i in family j at time t, $Y_{ij}(t)$. We use t here to denote time, typically measured in age, but it can be, in principle, any variable. For example, in the case where the function valued trait in some type of reaction norm, the input may be some environmental measure, such as temperature. Much of the information below is reviewed in greater detail by Jaffrezic and Pletcher (2000). The basic model can we written as

$$Y_{ijt} = \text{ fixed effects } + g_{ij}(t) + e_{ij}(t) \tag{3.7}$$

where $g_j(t), e_j(t)$ are the genetic and environmental random *processes* for subject j at time t. In the discrete setting, the statistical goal was to estimate the covariance matrices of the genetic and environmental random effects; in the continuous setting, our goal is to estimate the genetic and environmental *covariance functions*:

$$G(s, t) = \text{cov}(g_{ij}(s), g_{ij}(t)) \tag{3.8}$$

$$E(s, t) = \text{cov}(e_{ij}(s), e_{ij}(t)), \tag{3.9}$$

which can be thought of as continuous analogs to a covariance matrix. Assuming that the genetic and environmental random processes are uncorrelated, the covariance between related individuals i, k at times t, s is then modeled as

$$\text{cov}(Y_{ijt}, Y_{kjs}) = \text{cov}(g_{ij}(t), g_{kj}(s)) + \text{cov}(e_{ij}(t), e_{kj}(s)) \tag{3.10}$$

The first term in (3.10) will have the form

$$\text{cov}(g_{ij}(t), g_{kj}(s)) = \mathbf{R}_{ik}G(s,t). \tag{3.11}$$

Regarding the second term in (3.10), we can generally write

$$\text{cov}(e_{ij}(t), e_{kj}(s)) = \mathbf{C}_{ik}E(s,t), \tag{3.12}$$

where $\mathbf{C}_{ik}$ is defined according to how environment is shared between individuals. For example, $\mathbf{C}_{ik} = \mathcal{I}(i = k)$ is equivalent to supposing environment is completely unshared between individuals. As in the univariate case, $e_{ij}(t)$ can potentially be broken in any number of identifiable parts. For example, if there are repeated measurements on individuals as well as multiple family members, environment may be partitioned into familial common environment, individual common environment, and transient environment or measurement error:

$$e_{ij}(t) = e_j(t) + e_i(t) + \varepsilon_{ij}(t) \tag{3.13}$$

In this case, the covariance between the environmental processes for two observations would be the sum of the covariances between each of the constituent processes. The statistical goal in this inquiry is to estimate the functional forms of the covariance functions.

The first principle approach to modeling G, E was originally presented by Kirkpatrick and Heckman (1989). The genetic and environmental covariance matrix in the traditional multivariate techniques can be thought of as discrete observations of the covariance functions. This seminal paper was presented as a non-parametric method of smoothing a previously estimated covariance matrix. The authors choose a non-parametric estimate for each of the covariance functions of the form:

$$G(s,t) = \sum_{q=0}^{m}\sum_{r=0}^{m} \phi_q(t)\phi_r(s)K_{qr} \qquad (3.14)$$

where $\phi_q(t)$ are an orthogonal basis, m is the order of the polynomial, and and the K_{qr} terms are parameters to be estimated. In a later publication Kirkpatrick et al. (1990) suggests taking $\phi_q(t)$ to be Legendre polynomials, and describes a generalized least squares approach to estimating the coefficients with fit assessed relative to a previously estimated covariance matrix. Addressing the fact that Kirkpatrick's approach requires the factoring of a multivariate mixed model whose complexity grows with the number of unique ages in the dataset, a publication by Meyer and Hill (1997) gives a method of estimating the parameters in the above model directly by REML. One problem with these models is that, as noted by Pletcher and Geyer (1999), they do not definitionally yield positive definite covariance functions. Another major concern is that the individual parameters in the model lack any meaningful scientific interpretation. Finally, the use of Legendre polynomials is largely arbitrary and may not be appropriate.

The second approach, advocated by Meyer (1998) uses random regression to model g and e, and then works with the implied covariance functions. The method relies on an equivalence between the class of covariance functions and that of random regression models. That is, any covariance function can be written as the implied covariance of a properly structure random regression model. Conversely, all random regression models imply a valid covariance function. The author's justification for this formulation is primarily as a means to lower the complexity of the Meyer and Hill (1997) method, which grows proportional to the number of unique ages (or times) observed in the dataset. The simplest example of Meyer's random regression formulation is

$$g_{ij}(t) = \theta_{j1} + \theta_{j2}t. \qquad (3.15)$$

The covariance function implied by (3.15) is

$$G(s,t) = \text{var}(\theta_{j1}) + st\text{var}(\theta_{j2}) + (s+t)\text{cov}(\theta_{j1}, \theta_{j2}) \qquad (3.16)$$

with an analogous model for $e(t)$. In models of this form estimation of the relevant variance parameters can be estimated by REML. With this approach the problem of estimating G, E is reduced to that of determining an optimal form for the random regression model. This is typically done by the use of likelihood ratio tests, since various forms of interest in the random regression models are often nested. This approach does provide positive definite covariance functions but, like its predecessors, suffers from a lack of interpretability of individual parameters.

The third approach, given by Pletcher and Geyer (1999) is similar to the first method in that it attempts to explicitly parameterize the covariance functions, rather than the latent processes themselves. They largely abandon the lead taken by their predecessors, citing the critiques mentioned above, among others. They find a non-parametric approach to be overly optimistic, as there is not likely to be enough information about the latent processes for this to be feasible with realistic sample sizes.

The basic approach uses the fact that given any covariance function of a second order stationary process, $f(s,t) = f(|s-t|)$, it necessarily follows that

$$r(s,t) = v(s)v(t)f(|s-t|) \qquad (3.17)$$

is a valid covariance function for any choice of the function v. For example, if G were modeled in this way, then v would describe the way genetic variation changes with age and the parameters underlying f would display how correlation changes with age. The authors indicate that low order polynomials often suffice to capture interest patterns of change. There are many potential choices of f, and the authors

recommend the use of various well known characteristic functions viewed only as a function of $|s - t|$. The rationale for this is that an arbitrary function is a valid covariance function if and only if the function is positive definite, and characteristic functions are guaranteed to be positive definite. Pletcher and Geyer (1999) note that the assumption of stationarity is very likely to be erroneous, but found from experience that it can serve as a effective approximation in most situations.

Jaffrezic and Pletcher (2000) offer a slight extension of this approach to relax the assumption of stationarity. They instead advocate an approach based on a result from Nunez-Anton (1998), and Nunez-Anton and Zimmerman (2000). They showed, by using a properly chosen nonlinear transformation of time, the assumption of correlation stationarity can be relaxed. In particular,

$$f(s, t) = f(|h(s) - h(t)|), \qquad (3.18)$$

for a properly chosen h, will still yield a positive definite function as above, but will not impose stationarity. One particular h that will suffice is the well known Box-Cox transformation [Box and Cox (1964)]:

$$h(t) = (t^\lambda - 1) \cdot \mathcal{I}(\lambda \neq 0) + \log(t) \cdot \mathcal{I}(\lambda = 0). \qquad (3.19)$$

An appealing property of this choice is that if $\lambda = 1$, then $|h(t) - h(s)| = |t - s|$, resulting in a stationary covariance function. Thus, in some sense, the data can indicate the level of correlation stationarity present.

We have reviewed the seminal work for each of the main approaches to estimation of genetic and environmental contributions to function valued traits, which fall into three main categories: non-parametric estimation using a basis expansion; random regression; explicit parametric models. Many variations of non-parametric covariance function estimation exist: Yang et al. (2003) and Macgregor et al. (2003) which mod-

eled covariance matrices between family members as functions of age using splines; Kirkpatrick and Meyer (2004) propose a method of directly estimating the eigenfunctions of G, E which effectively reduces the dimension of the problem. Other methods involving parametric approaches have been published. Work by Soler and Blangero (2003) used stationary processes with exponential covariance functions to model G and E. A review of some other methods for longitudinal family data is given in Gauderman et al. (2003).

3.1.2 Rationale for this work

Models which explicitly model the longitudinal structure are a step in the right direction, but have one major shortcoming in all of their current incarnations. This alludes to the primary theme of this work: familial correlations are potentially modulated by more than a single timing variable. Each existing method requires an arbitrary specification of how time is indexed in their model; in most cases time is specified to be indexed by age rather than chronological time. This may be sensible with regard to genetic latent processes, since one could argue that age is the only timing variable relevant to genetically induced effects on the response. However, with respect to the environmental latent process parameterized by a stationary correlation function, this specification would indicate that the environmental covariance is maximized when two siblings are at similar ages, regardless of how far apart the observations are made in chronological time. If, for example, two siblings are born 10 years apart, then the environment they are experiencing at a common age is likely to be quite different, thus it does not make sense for their environmental covariance to be definitionally maximized there.

In addition to interpretability problems, it has been noted in many cases that the indexing specification can have a considerable impact on the conclusions of the analysis. For example, it has been shown in various contexts that the apparent

level of correlation between outcomes on family members over time differs when the data is aligned by age as opposed to time (Diego et al, 2003). This makes sense because, as mentioned earlier, it is well documented that many diseases are impacted by age-specific genetic predispositions. In addition, mechanisms which depend only on calendar time, such as catastrophic event occurring in the shared environment, can induce correlations between individuals which share that environment. Another, considerably more subtle, observation is that the amount by which an individual is affected by their environment is likely to depend on the person's age. Thus, multiple timing variables may be simultaneously modulating a single underlying latent process, and may be the driving force for distinct devices that are inducing familial correlation.

As mentioned before, no existing methods for longitudinal family data take more than a single timing variable into account. In particular, the concept that correlations induced by the family environment at a particular chronological time point are dependent on the individuals' ages reflects a previously unexplored pattern of resemblance between family members. The goal of this work is to investigate the extent to which age-specific and temporal mechanisms modulate the association between family members, and to what extend we can disaggregate the two effects. The motivating application for this work is the Michigan Longitudinal Study (MLS), Zucker et al. (1996), a prospective long term prospective study of initially intact families at high risk for substance abuse. For this work we focus on a subset of the data comprised of DOTS-R exam measurements. The DOTS-R exam is a quantitative tool used for the identification of age-continuous features of temperment across an age span from early childhood to young adulthood.

3.1.3   Chapter Outline

To begin, we present a parametric framework for modeling processes modulated by multiple timing variables based on a data generating model, rather than implicit

specification through covariance functions. In this section details on the intuition underlying the model specification and the practical interpretation of key parameters in the model are mentioned. This is followed by a discussion on how to parameterize the model and some considerations which must be made. The correlation structure implied by our parameterization is shown for some nominal parameter values. A description of how the parameters in our model are identified is given, as well as a rough diagnostic for pinpointing when non-identifiability may be an issue in a particular data set.

In the next section, details on computation of the estimates and inference for parameters is given. A maximum likelihood based estimation procedure is described and some simulations to display the ability to recover the data generating population structure are shown. The performance of the estimation under a few types of model misspecification is also given. When non-identifiability may be an issue, an approach to estimation of one of the two confounded parameters is described; this method is based on maximum likelihood analysis of a reduced data set. This alternative approach is shown with simulations to improve estimation when the structural form in one of the confounded parameters is misspecified.

In section 4, we move on the the issue of hypothesis testing. Most of the scientifically meaningful hypotheses we are interested in testing are nested, making the likelihood ratio test (LRT) a convenient choice. This discussion begins by noting the difficulty in constructing a measure of effect size, and giving the restricted setting where a particular measure is apt. Since we are doing LRT in a potentially non-standard setting, various simulations are conducted to investigate the level and power of the test. How the level and power are affected by potential non-identifiability issues is also shown.

In section 5, we apply the methods described to DOTS-R exam measurements, which is a subset of the MLS data. The DOTS-R exam is a quantitative tool used

for the identification for age-continuous features of temperament across an age span from early childhood to young adulthood. The exam assumes that temperament is composed of "characteristic behavior styles", such as rhythmicity, response patterns to new situations, and patterns of adjustment to environment [Windle and Lerner (1986)]. A complete analysis including assessment of fit on a number of candidate models, as well as interpretation of the final model for some interesting cases. The chapter than concludes with a discussion of the method as well as some avenues for future work.

3.2   Model Formulation

3.2.1   A general conceptual framework

Recalling the discussion in the introduction, we seek to build a framework which induces time-dependent correlation between family members, and incorporates the possibility of this correlation being modulated by the subjects' ages. In many realistic data scenarios mostly nuclear families are available, making genetics weakly identified or not indentified at all from environment, thus we do not attempt to make this discomposition, although it is hypothetically possible.

On a general level we can write such a model for subject i in family j at time t and age a_{ijt} as

$$Y_{ijt} = \mu(a_{ijt}) + \eta_j(t, a_{ijt}) + \gamma_i(a_{ijt}) + \varepsilon_i(t). \tag{3.20}$$

In this model μ is the age specific mean, η_j should be viewed broadly as a combination of all possible shared sources of variation in the family at time t, and age a_{ijt}. The γ_i term is individual specific and models sources of correlation not shared with the rest of the family; $\varepsilon_i(t)$ is an individual and occasion specific perturbation analogous to the residual in a regression model.

The model formulation given above asserts that an individual response is a combination of

- certain baseline fixed effects, which are a function of age

- a family specific latent process, $\eta_j(t, a)$, which captures longitudinal aspects of the data due to chronological family time and/or age.

- an individual specific latent variable γ_i, which models self-correlations not shared with the rest of the family

- individual and occasion specific perturbations, $\varepsilon_i(t)$, which may be thought of as measurement error

In all inquiries of this type only a single timing variable is allowed to modulate the familial correlation, which is captured by η_j. Typical approaches correspond to specifying either

$$\eta_j(t, a_{ijt}) = \eta_j(a_{jit}) \tag{3.21}$$

or

$$\eta_j(t, a_{ijt}) = \eta_j(t). \tag{3.22}$$

The choice in all models of this type is typically (3.21). For reasons discussed in the introduction we advocate the accommodation of multiple timing variables. For simplicity we choose to decompose η_j into two parts which have distinct interpretations:

$$\eta_j(t, a_{ijt}) = \eta_{j1}(t) + \eta_{j2}(a_{jit}). \tag{3.23}$$

The term η_{j1} can be interpreted as events occurring in the family environment at a specific time. For example, a stressful event in the family that effects trait measure-

ments would correspond to a spike or dip in η_{j1}. The second term, η_{j2} would be more thought of as environmental or genetics effects that tend to occur at similar ages within a family. For example, in many families the freedom a child has is dependent on their age, and tends to be similar between siblings.

For each of the random effects we allow an age specific variance, but do not allow the variance function to depend on chronological time; to do so would be to model global secular trends in the data and that is not the purpose of this analysis. Bearing this in mind, we now modify our formulation so that each random effect has mean 0 and unit variance, and has a functional parameter to capture the variance:

$$Y_{ijt} = \mu(a_{ijt}) + \psi_1(a_{ijt})\eta_{j1}(t) + \psi_2(a_{ijt})\eta_{j2}(a_{ijt})$$
$$+ \varphi(a_{ijt})\gamma_i + \xi(a_{ijt})\varepsilon_i(t) \tag{3.24}$$

Here $\mu, \psi_1, \psi_2, \varphi, \xi$ are all parameters that are functions of age. The functional coefficients, ψ_1, ψ_2, are the primary components of interest in this model and represent age specific sensitivity to chronological time and age indexed family environment, respectively. The individual specific random effect, γ_i, is specified as a scalar random effect rather than a random process; this term should be interpreted as stable within-individual characteristics. In Figure 3.1 a schematic plot is given to further clarify the role of the familial sensitivity functions, using ψ_1 as an example.

3.2.2 Parameterization and implied correlation structure

Let $\boldsymbol{Y}_j$ denote the N_j-length vector of family j responses observed at times $\boldsymbol{t}_j$ and ages $\boldsymbol{a}_j$. The mean of $\boldsymbol{Y}_j$ will be characterized by a single functional parameter μ. That is,
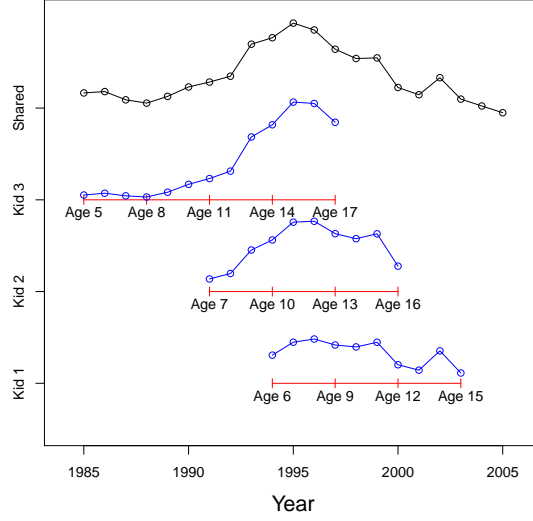
Figure 3.1: Simulated family trajectories when $\psi_1(a)$ is a linearly increasing function and all other random effect coefficients are 0 for three siblings separated in age by four years. The top trajectory is the time indexed familial environment.

$$\boldsymbol{\mu}_j = \left(\mu([\boldsymbol{a}_j]_1), ..., \mu([\boldsymbol{a}_j]_{N_j}) \right). \tag{3.25}$$

Covariate effects can easily be incorporated by including additional arguments to μ.

Without making any parametric specifications, we can write the family j covariance matrix:

$$\boldsymbol{\Sigma}_j = \boldsymbol{\psi}_1' \boldsymbol{\Sigma}_{\eta_1} \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2' \boldsymbol{\Sigma}_{\eta_2} \boldsymbol{\psi}_2 + \boldsymbol{\varphi}' \boldsymbol{\Sigma}_{\gamma} \boldsymbol{\varphi} + \boldsymbol{\xi}' \boldsymbol{\xi} \tag{3.26}$$

where

$$\boldsymbol{\psi}_1 = \begin{pmatrix} \psi_1([\boldsymbol{a}_j]_1) & & & \\ & \psi_1([\boldsymbol{a}_j]_2) & & \\ & & \ddots & \\ & & & \psi_1([\boldsymbol{a}_j]_{N_j}) \end{pmatrix} \tag{3.27}$$

62

The matrices $\boldsymbol{\psi}_2$, $\boldsymbol{\varphi}$, and $\boldsymbol{\xi}$ are defined analogously. $\boldsymbol{\Sigma}_{\eta 1}, \boldsymbol{\Sigma}_{\eta 2}, \boldsymbol{\Sigma}_\gamma$ denote discretized version of the covariance functions for the latent processes; for example $[\boldsymbol{\Sigma}_{\eta 1}]_{k,\ell} = \mathrm{cov}(\eta_{j1}([\boldsymbol{t}_j]_k), \eta_{j1}([\boldsymbol{t}_j]_\ell))$.

For this model to be identifiable we must place certain constraints on the covariance structure of the latent processes. If the latent processes are allowed to have non-constant variance, the variance functions would not be identified. This tacitly implies that we must choose a stationary process for each latent variable, or else certain mean and variance parameters in the model will not be identified. Thus η_{j1}, for example, can be viewed as some standardized quantification of how different a given family is from the average at a given time, and analogously for $\eta_{j2}(a)$.

We now examine choices for the covariance functions for the latent processes, η_{j1}, η_{j2}. We will refer to $\eta(t)$ but we intend to use the same parameterization for η_{j1} and η_{j2}, so we will suppress the subscripts throughout and similarly with ψ. Characteristic functions are reasonable candidates for covariance functions since they are guaranteed to be positive definite; a table of characteristic functions modified in form for use as covariance function is given in Pletcher and Geyer (1999). Consider the family of covariance functions indexed by two parameters, (τ, κ) of the form

$$r(s,t) = \tau^{|t-s|^\kappa} \tag{3.28}$$

Notice this can be rewritten as

$$r(s,t) = \exp\left(\log(\tau)|t-s|^\kappa\right), \tag{3.29}$$

which is a valid class of covariance functions for $\kappa \in (1,2], \tau \in (0,1)$ [Schoenberg (1938)] and is known as the κ-exponential covariance function. For $\kappa < 2$ the sample paths of the corresponding process are nowhere differentiable, but have infinitely many derivatives for $\kappa = 2$ [Rasmussen and Williams (2006)]; this special case is
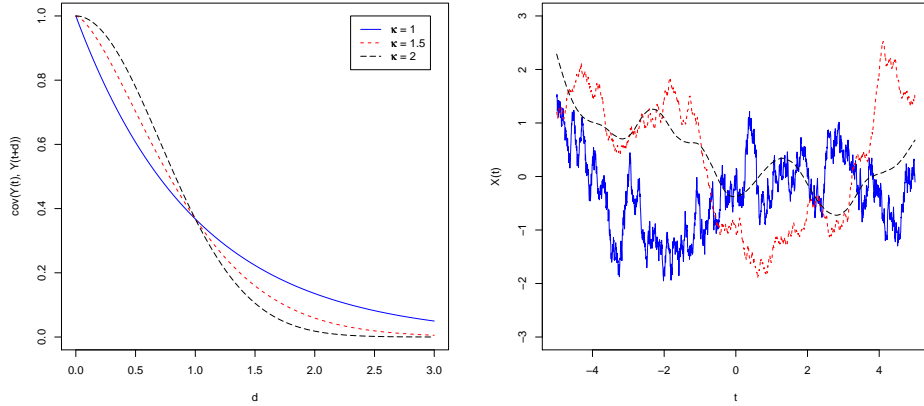
Figure 3.2: **Left**: κ-exponential covariance functions for $\kappa = 1, 1.5, 2$ with $\tau = e^{-1}$. **Right:** Sample paths of processes with a κ-exponential covariance function for $\kappa = 1, 1.5, 2$

the well known squared exponential covariance. Larger values of τ indicate that stronger correlation exists over longer time lags, and larger values of κ indicate a more smoothly varying process; this is illustrated by Figure 3.2.

When the data supports such estimation, we propose taking this one step further by modeling the parameter τ as a family specific random effect. The simplest choice for this parameterization is the Beta distribution since the support is the same of that for τ. Specifying τ as a random effect is equivalent to a belief that the level of stability within a family process depends on the family. Assuming the Beta(α_1, α_2) distribution for τ, this family of covariance functions has the following properties (proofs in Appendix A):

1. The covariance function, unconditional on τ, is

$$\mathrm{cov}(\eta(t), \eta(s)) = \frac{\Gamma(\alpha_1 + \alpha_2)\Gamma(\alpha_1 + |t - s|^\kappa)}{\Gamma(\alpha_1 + \alpha_2 + |t - s|^\kappa)\Gamma(\alpha_1)}$$

2. The exchangeable correlation structure is a limiting case. That is, if $\rho = \alpha_1/(\alpha_1 + \alpha_2)$ remains fixed, then as $\alpha_1, \alpha_2 \downarrow 0$,

64

$$\text{cov}(\eta(t), \eta(s)) \to \rho$$

independently of t, s.

3. The κ-exponential covariance is a limiting case. In particular, if $\rho = \alpha_1/(\alpha_1+\alpha_2)$ remains fixed, then as $\alpha_1, \alpha_2 \to \infty$

$$\text{cov}(\eta(t), \eta(s)) \to \exp\left(\log(\rho)|t - s|^\kappa\right)$$

This parameterization can be difficult to fit when both ψ_1 and ψ_2 are in the model. When it is evident that this is an issue, the regular κ-exponential covariance is used instead.

A simple parameterization will be chosen for each of the functional parameters. In the work of Pletcher and Geyer (1999), they chose to model the variance functions as linear in some number of parameters. That is,

$$v(a)^2 = \sum_{q=1}^{Q} \beta_q h_q(a), \tag{3.30}$$

where $h_q(a)$ is typically some monotone function of a. Alternatively, one could model $v(a)$ linearly, which eliminates the need to constrain $\boldsymbol{\beta}$ to form a function that is positive for any a; this is the choice we use. As noted by the Pletcher and Geyer (1999), we have no a priori expectations about the shapes of these functions, but this parameterization can capture many plausible structures and noting the latent nature of the processes of interest, more complicated parameterizations are not likely to pay off. In addition, more highly parameterized functions are likely to display more peaks and valleys that are not interpretable and are most likely overfitting the data. The mean is specified as a low order polynomial in each age (with a similar specification if covariates were to be included).

To make this parametric specification more clear we now examine the implied correlation structure. It is convenient to define two quantities now. First,

$$\sigma^2(a) = \psi_1(a)^2 + \psi_2(a)^2 + \varphi(a)^2 + \xi(a)^2 \tag{3.31}$$

which is the total variance in an age a subject's response. One quantity that is intuitively linked to the role of ψ_1, ψ_2 is

$$\lambda_k(a) = \frac{\psi_k(a)^2}{\sigma^2(a)} \tag{3.32}$$

which is the proportion of total variance at a given age that is due to ψ_k. We will revisit this quantity later when we look to define a measure of effect size for particular hypothesis tests. The correlation between two subjects i, j in a family at times t, s can be written as

$$\text{cor}(Y_{ijt}, Y_{kjs}|\tau_j) = S_1 + S_2 + S_3 + S_4 \tag{3.33}$$

where

$$S_1 = \tau_{j1}^{|t-s|^{\kappa_1}} \text{sign}(\psi_1(a_{ijt}))\text{sign}(\psi_1(a_{kjs})) \cdot \sqrt{\lambda_1(a_{ijt})\lambda_1(a_{kjs})} \tag{3.34}$$

$$S_2 = \tau_{j2}^{|a_{ijt}-a_{kjs}|^{\kappa_2}} \text{sign}(\psi_2(a_{ijt}))\text{sign}(\psi_2(a_{kjs})) \cdot \sqrt{\lambda_2(a_{ijt})\lambda_2(a_{kjs})} \tag{3.35}$$

$$S_3 = \varphi(a_{ijt})\varphi(a_{kjs})\mathcal{I}(i=k)/\sqrt{\sigma^2(a_{ijt})\sigma^2(a_{kjs})} \tag{3.36}$$

$$S_4 = \xi(a_{ijt})\xi(a_{kjs})\mathcal{I}(i=k)\mathcal{I}(t=s)/\sqrt{\sigma^2(a_{ijt})\sigma^2(a_{kjs})} \tag{3.37}$$

We now graphically display the implied correlation structure for two populations. In each case we look at the implied covariance matrix as a function of observation time (1999-2009) for two siblings born in 1989 and 1993. In each population all variance functions other than ψ_1, ψ_2 are fixed at nominal constants. The upper 3×3

block of Figure 3.3 is a population where ψ_1 is linearly increasing, while ψ_2 is linearly decreasing, each with equal magnitude (averaged across observed ages). The lower 3×3 block corresponds to a structure where ψ_1 is linearly decreasing while ψ_2 is increasing. The differences in the implied correlation structures highlight where the information to identify these parameters comes from. In addition, when both of the latent processes have high levels of autocorrelation, decoupling these sources of variation becomes a more difficult task, as can be seen in the bottom right corner of each 3-by-3 block. In each case the κ-exponential covariance is used with $\kappa = 2$ and $E(\tau)$ incremented across $.1, .5, .9$ for both random processes to display the effect of different autocorrelation magnitudes.

3.2.3 Identifiability issues

The variance components corresponding to φ, ξ are identified from each other and from ψ_1, ψ_2 by the blocking structure in the data. These two components are shared within individual and within timepoint, respectively, which identifies them from the components that are shared between family members. However, there is no blocking structure that explicitly identifies ψ_1 from ψ_2; the only thing distinguishing them is the differential staggering of age and time gaps between observations made on family members. The reason that the gaps are the point of interest is that we assume a correlation stationary distribution for the corresponding latent processes.

Let $\mathcal{D}$ denote the joint distribution of age and time gaps between observations made on family members. There are two ways in which identifiability issues can be linked to $\mathcal{D}$. The first is when $\mathcal{D}$ is a distribution such that the time and age gaps, Δt, Δa, respectively, are frequently very similar to each other. That is, $E_{\mathcal{D}}(|\Delta t - \Delta a|)$ is very small. For example, in a twin study, $\Delta a = \Delta t$ for every pair of assessments. In that case, the covariance between a pair of sibling measurements is
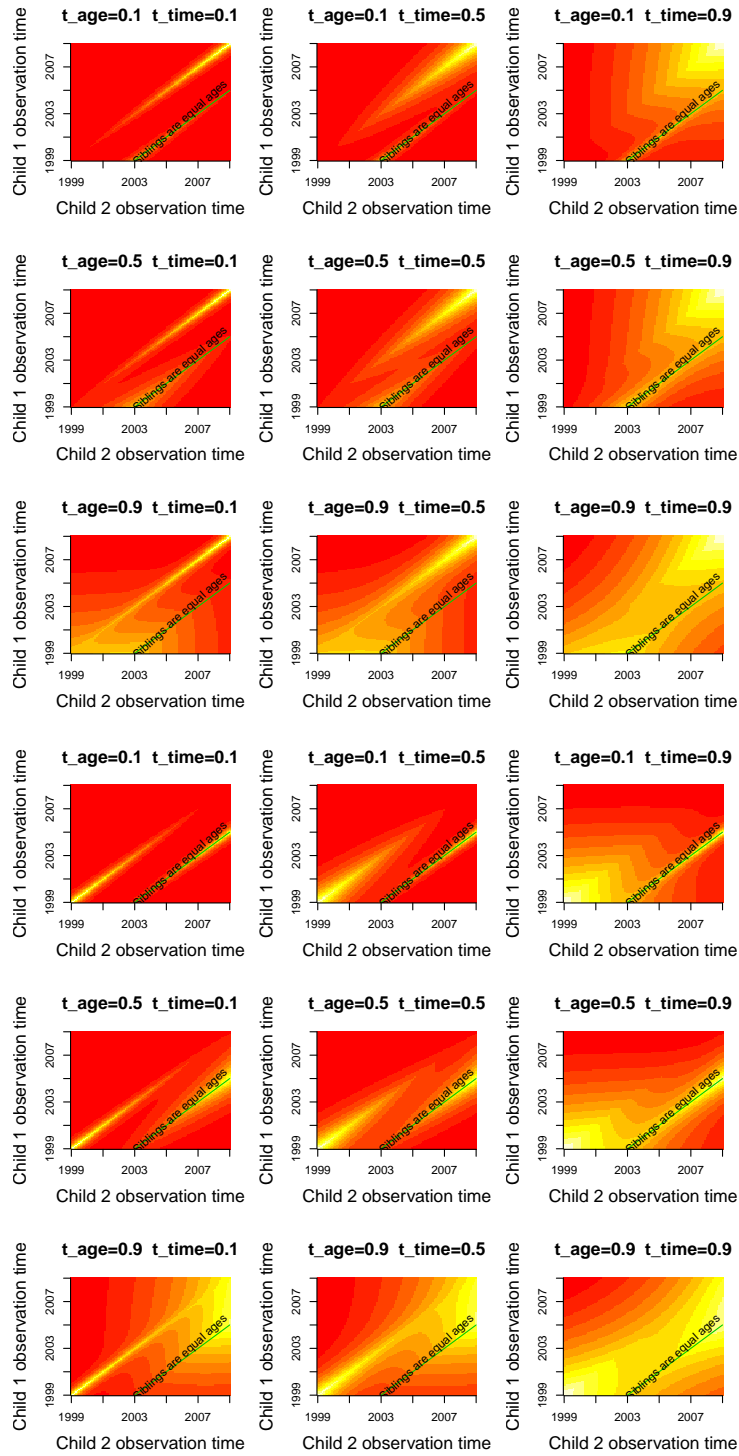
Figure 3.3: Implied covariance matrix between two siblings born 1989 and 1993. **Left**: ψ_1 linearly increasing, ψ_2 linearly decreasing. **Right**: ψ_1 linearly decreasing, ψ_2 linearly increasing. In both panels $E(\tau_{j1}), E(\tau_{j2})$ are incremented across $.1, .5, .9$, indicated by the heading of each plot.

$$\psi_1(a_1)\psi_1(a_2)r_1(\Delta a) + \psi_2(a_1)\psi_2(a_2)r_2(\Delta a) \tag{3.38}$$

Here it is clear that it is impossible to tell which r is the covariance function for the age and time indexed latent processes, respectively. Thus, ψ_1, ψ_2 are not identified. Usually when such a structure for $\mathcal{D}$ is present in a data set, it is clear. When this is not the case, a scatter plot of the observed age vs. time gaps can used as a diagnostic.

When the joint distribution of ages is not structured in such a pathological way, the identifiability of ψ_1, ψ_2 from each other depends primarily on how different the covariance functions for η_{j1}, η_{j2} across $\mathcal{D}$. In the extreme case where $r_1(\Delta t) \approx r_2(\Delta a)$ for all $\Delta t, \Delta a$, the covariance between pairs is approximately

$$\psi_1(a_1)\psi_1(a_2)r_1(\Delta t) + \psi_2(a_1)\psi_2(a_2)r_1(\Delta t). \tag{3.39}$$

Then ψ_1, ψ_2 enter the covariance matrix as multiples of $\psi_1(a_1)\psi_1(a_2) + \psi_2(a_1)\psi_2(a_2)$ in every case, leaving the two clearly unidentified. A quantity closely linked to the degree to which this phenomena is occurring is

$$M_{\mathcal{D}} = \mathrm{E}_{\mathcal{D}}\left(\left|r_1(\Delta t) - r_2(\Delta a)\right|\right). \tag{3.40}$$

If $M_{\mathcal{D}}$ is not too small, then there is information in the data to distinguish ψ_1, ψ_2. The primary situation where this is a concern is when both latent processes are highly autocorrelated so that $r_1(\Delta t) \approx r_2(\Delta a) \approx 1$ for any $\Delta t, \Delta a \neq 0$. Then, $M_{\mathcal{D}} \approx 0$ and there is nothing to distinguish the two processes from each other.

While we do not know the forms of r_1, r_2 a priori, and therefore cannot know the exact value for $M_{\mathcal{D}}$, we do have a sample from $\mathcal{D}$. Therefore we can calculate $M_{\mathcal{D}}$ across a grid of values for the parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ underlying r_1, r_2, and calculate the matrix with structure

$$[\boldsymbol{M}_{\mathcal{D}}]_{k,\ell} = \mathrm{E}_{\mathcal{D}}\left(\left|r_{1,\boldsymbol{\theta}_{1k}}(\Delta t) - r_{2,\boldsymbol{\theta}_{2\ell}}(\Delta a)\right|\right) \tag{3.41}$$

where the entries are estimated by the sample moments in the data. If $\widehat{\boldsymbol{M}}_{\mathcal{D}}$ has many very small values, this is an indicator to the practitioner that for many of the admissible parameter values, ψ_1, ψ_2 are poorly identified from each other. Analogous measures can be constructed for each of the other functional parameters in the model to assess their identifiability. For example, if we only have repeated measures on a small subset of the sample, φ, ξ will be poorly identified from each other. In Figure 3.4 we have heatmaps of $\widehat{\boldsymbol{M}}_{\mathcal{D}}$ for several different joint distributions of age and time gaps when r_1, r_2 are each squared exponential covariance functions. We can see that, universally, ψ_1, ψ_2 are most identified from each other when one is highly autocorrelated and the other is nearly white noise. In addition, it appears that having large age spacing diversity is more beneficial to identifiability than observation time diversity. In Figure 3.5, a scatter plot of Δa vs. Δt is shown (with jittering to prevent overlapping) as well as the measure of identifiability for the DOTS-R data. There appears to be a healthy mixture of age and time gap combinations, and across a wide range of the parameters underlying r_1, r_2, there are large values of $M_{\mathcal{D}}$ in the DOTS-R data.

3.3 Computation of the estimates

3.3.1 Maximum Likelihood Estimation

To estimate the mean and covariance structure parameters, denoted by $\boldsymbol{\theta}$, we use maximum likelihood. The log likelihood for family j has the form

$$\mathbf{L}_j(\boldsymbol{\theta}) = -\frac{1}{2}\left(\log(|\boldsymbol{\Sigma}_j(\boldsymbol{\theta})|) + (\mathbf{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}))'\boldsymbol{\Sigma}_j(\boldsymbol{\theta})^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}))\right) \tag{3.42}$$

70
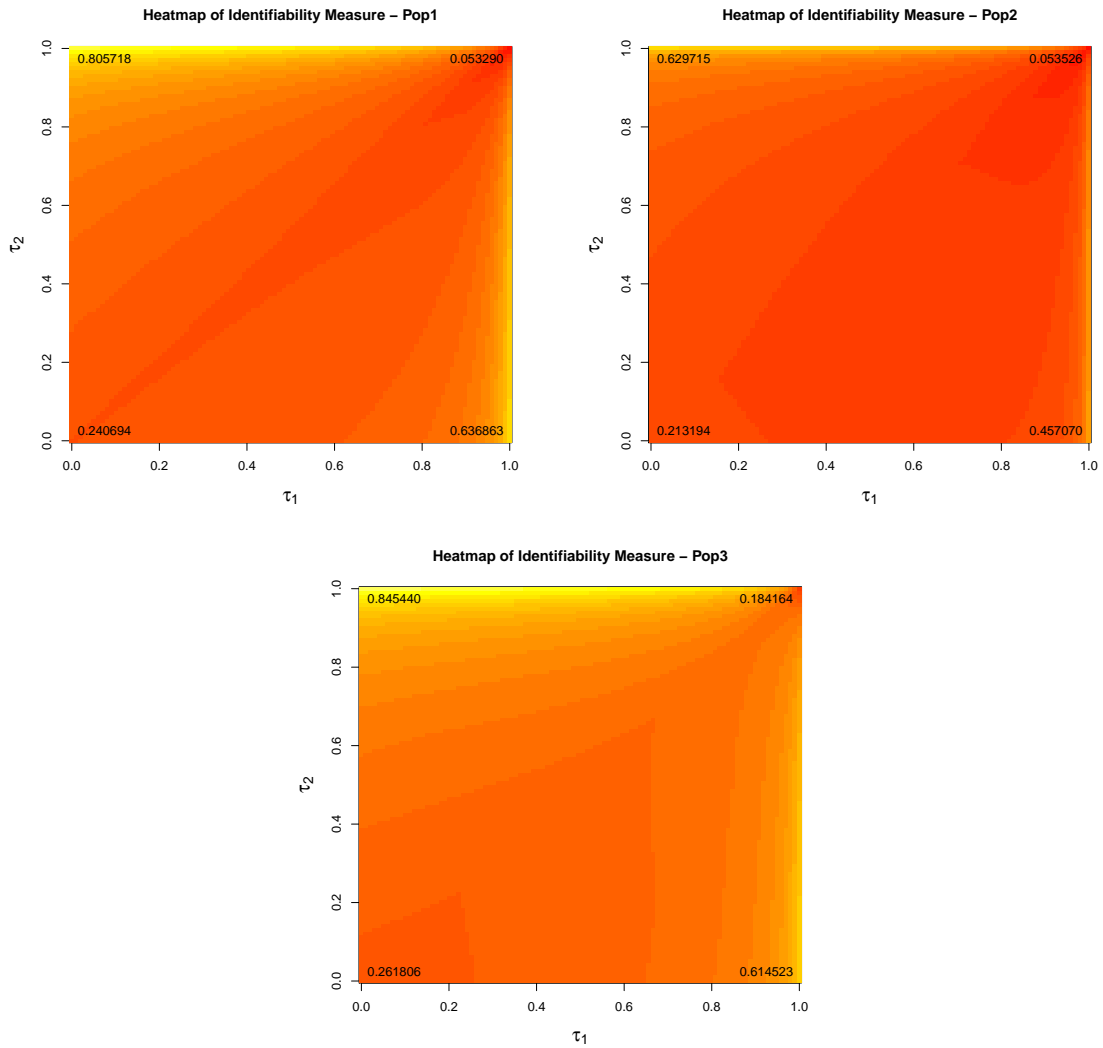
Figure 3.4: Measure of identifiability of ψ_1 from ψ_2 for families with three children, three timepoints, and initial ages a and and observation times t such that **Left:** $a \sim U(5, 10)$, $t \sim U(0, 10)$. **Middle:** $a \sim U(5, 10)$, $t \sim U(0, 20)$. **Right:** $a \sim U(5, 20)$, $t \sim U(0, 10)$.
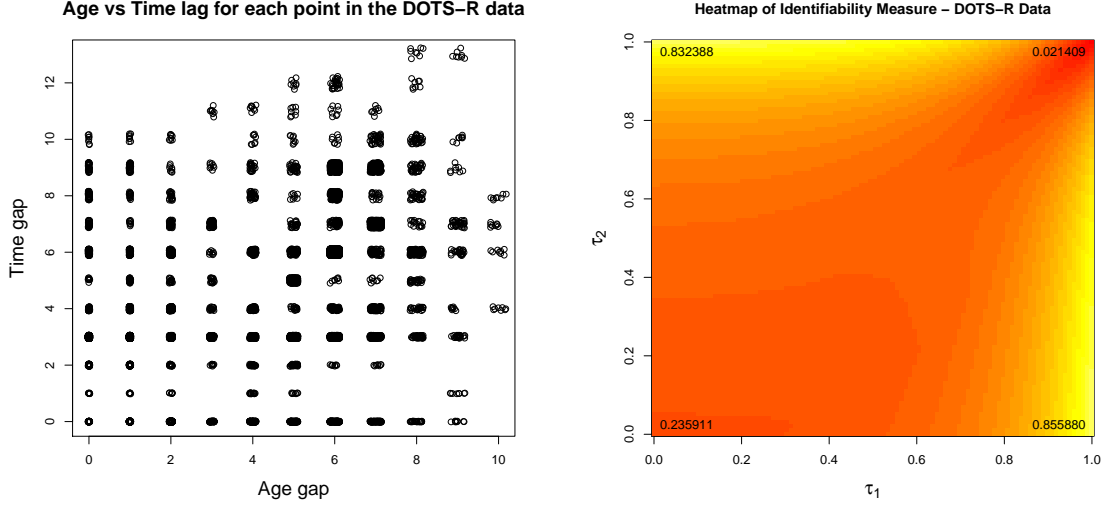
Figure 3.5: **Left:** Scatterplot of age vs. time gaps for each pair of sibling measurements in the DOTS-R data. **Right:** Measure of identifiability for the DOTS-R data.

where $\boldsymbol{\mu}_j(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_j(\boldsymbol{\theta})$ are given by (3.25) and (3.26). Assuming independence between families, the log likelihood based on a sample of F families is

$$\mathbf{L}(\boldsymbol{\theta}) = \sum_{j=1}^{F} \mathbf{L}_j(\boldsymbol{\theta}) \tag{3.43}$$

In principle we can differentiate this likelihood and carry out some derivative based optimization algorithm such as conjugate gradient optimization. In practice, a derivative calculation takes orders of magnitude longer than a function evaluation, making such an algorithm less useful. We have found the Nelder-Mead simplex algorithm [Nelder and Mead (1965)] effective in optimizing $\mathbf{L}$ without derivatives.

We now show some results on our ability to recover the correct population structure when the model is correctly specified, as well as the variance associated with the estimates, presented in Table 3.1. Under each sample size 250 datasets were generated with parameter value equal to the first column of the table, three time points per family uniformly generated from the discrete interval $\{0, 1, ..., 15\}$ and initial ages generated uniformly from $\{5, 6, ..., 15\}$, scaled by 100 for greater numerical stability.

Parameter	True	Estimate(se): $F = 200$	Estimate(se): $F = 400$
μ_0	0	-.008(.085)	-.007(.061)
μ_1	0	.034(.537)	.012(.358)
ψ_{10}	.1	-.001(.219)	.073(.179)
ψ_{11}	2	2.014(1.337)	2.081(1.08)
ψ_{20}	-.125	-.152(.323)	-.116(.214)
ψ_{21}	2.5	2.56(1.839)	2.48(1.664)
φ_0	.5	.474(.132)	.495(.089)
φ_1	0	.125(.787)	.094(.502)
ξ_0	.1	.994(.074)	1.001(.068)
ξ_1	0	-.018(.441)	.015(.344)
τ_{j1}	.5	.424(.254)	.458(.165)
τ_{j2}	.8	.833(.222)	.787(.173)
κ_1	1	1.121(.409)	1.152(.378)
κ_2	1	1.215(.397)	1.179(.341)

Table 3.1: Simulation based bias and variance estimates for the model parameters, with 1000 simulations under each setting.

The model was fit to each dataset and the average and standard standard deviation of the estimates are reported. In this table ψ_{10}, ψ_{11}, for example, refer to the intercept and slope of ψ_1, respectively. Since the variance parameters are only identified up to multiplication by -1, we perform this correction before constructing the table. We can see that our estimates to not appear to have much bias, with the possible exception of the κ parameters. In addition, it is clear that the standard errors decrease approximately with the expected scaling factor, $\sqrt{2}$ in this case, since the sample size is doubled.

Next we display the fitted trajectories under two data scenarios: the first corresponding exactly to the results in the table above with $F = 200$ families. The second case has the same parameter values as the first, except we know ψ_1, ψ_2 to be highly confounded: $r_1(\Delta a) = .95^{\Delta a}$, and $r_2(\Delta t) = .95^{\Delta t}$. In the "confounded" model, $\hat{\varphi}(a) = .489 + .056a$ and $\hat{\xi}(a) = .994 - .018a$, indicating very little sensitivity of these estimates to the confounding of ψ_1, ψ_2. We can see in Figure 3.6 that the estimates of $\psi_1^2(a), \psi_2^2(a)$ are far more variable under the confounding data setting, while the bias

is left apparently unaffected. This makes sense since we expect non-orthogonality of parameter estimates to result in greater variance, not bias.

3.3.2 Computation when non-identifiability is a concern

As displayed by the simulation above, approximate non-identifiability can affect estimation. In this situation, one can simply fit the model while leaving out one of these two parameters, but this would not result in unbiased estimate of the corresponding parameter. This would result in an estimate that would be some sort of combination of ψ_1 and ψ_2. In addition, if the structure of ψ_2 is misspecified, then strong confounding with ψ_1 may cause contamination of the estimation of ψ_1, and actually result in bias in $\hat{\psi}_1$. By performing maximum likelihood estimation on an altered data set, we can estimate ψ_1 independently of ψ_2, regardless of the level of confounding.

Consider two measurements taken on family members $Y_{i_1jt_1}, Y_{i_2jt_2}$ at a common age. Define a new variable

$$Z_{12} = Y_{i_1jt_1} - Y_{i_2jt_2}. \tag{3.44}$$

Letting a denote the common age, it follows that

$$
\begin{aligned}
Z_{12} &= \left(\mu(a) + \psi_1(a)\eta_{j1}(t_1) + \psi_2(a)\eta_{j2}(a) + \varphi(a)\gamma_{i_1} + \xi(a)\varepsilon_{i_1}(t_1)\right) \\
&\quad - \left(\mu(a) + \psi_1(a)\eta_{j1}(t_2) + \psi_2(a)\eta_{j2}(a) + \varphi(a)\gamma_k + \xi(a)\varepsilon_k(t_2)\right) \\
&= \psi_1(a)\left(\eta_{j1}(t_1) - \eta_{j1}(t_2)\right) + \varphi(a)\left(\gamma_{i_1} - \gamma_{i_2}\right) + \xi(a)\left(\varepsilon_{i_1}(t_1) - \varepsilon_{i_2}(t_2)\right) \tag{3.45}
\end{aligned}
$$

We can see that ψ_2 drops out of Z_{12}, thus the joint distribution of the altered data does not depend on ψ_2. Carrying out maximum likelihood analysis on this new data set gives a means of estimating ψ_1 that is independent of its confounding with ψ_2.
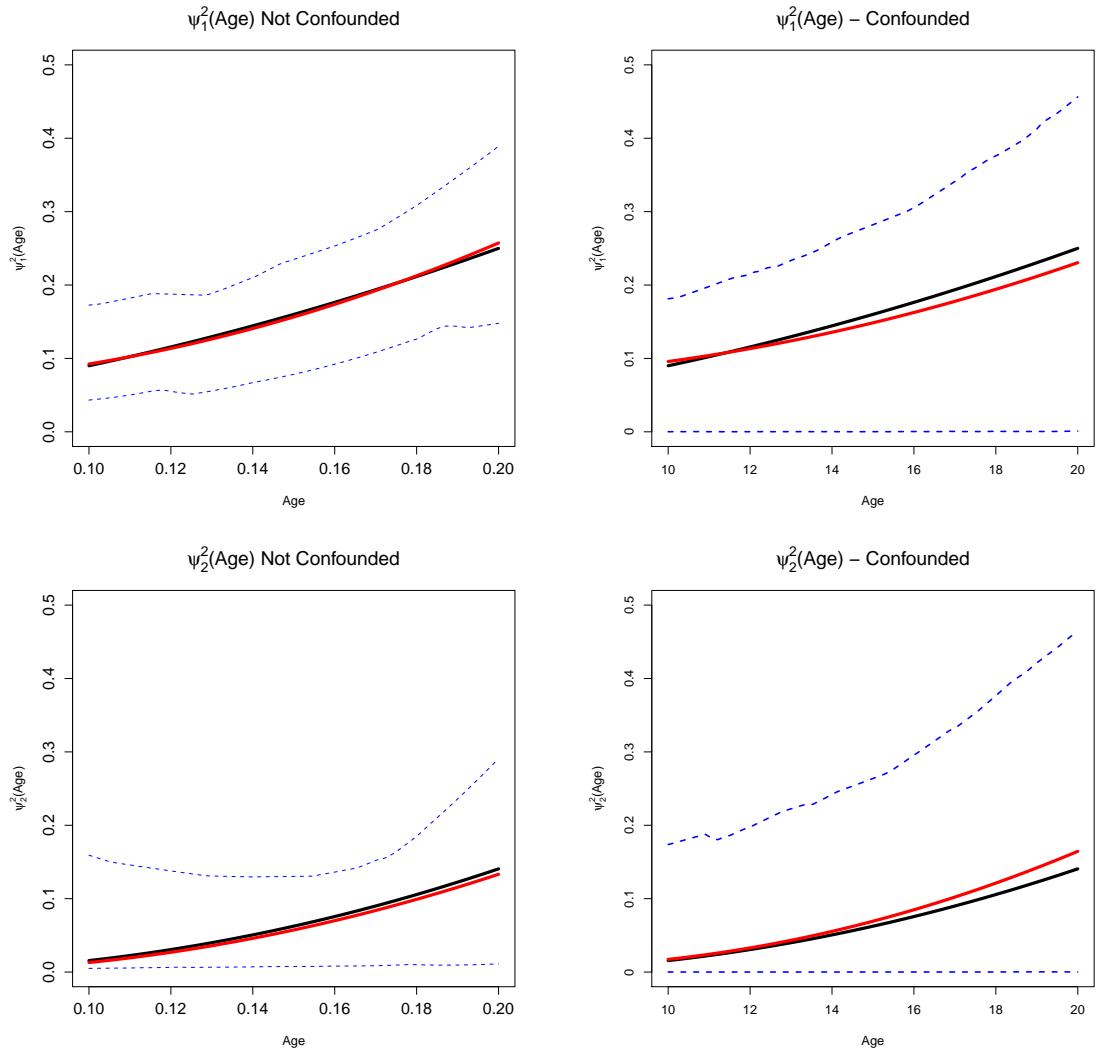
Figure 3.6: Estimation precision of $\hat{\psi}_1$ and $\hat{\psi}_2$ under a population where the estimates are expected to be confounded (right) and not confounded (left). The solid black line is the true function in each case, the red line is the average estimate (at each age) and the dotted lines are 2.5 and 97.5 quantiles on the estimated functions

Each of the new data points are marginally normal with expectation 0 and variance

$$\text{var}(Z_{12}) = \psi_1(a)^2 \left(2 - 2r_1(t_1, t_2)\right) + 2\varphi(a)^2 + 2\xi(a)^2. \tag{3.46}$$

To fully specify the joint distribution of the Z's we need the covariance terms. Defining Z_{34} analogously to above, with common age b,

$$\text{cov}(Z_{12}, Z_{34}) = A + B + C. \tag{3.47}$$

The first term is

$$A = \text{cov}\left(\psi_1(a)\left(\eta_{j1}(t_1) - \eta_{j1}(t_2)\right), \psi_1(b)\left(\eta_{j1}(t_3) - \eta_{j1}(t_4)\right)\right)$$
$$= \psi_1(a)\psi_1(b)\left(r(t_1, t_3) - r(t_1, t_4) - r(t_2, t_3) + r(t_2, t_4)\right). \tag{3.48}$$

The second term is

$$B = \varphi(a)\varphi(b)\text{cov}\left(\gamma_{i_1} - \gamma_{i_2}, \gamma_{i_3} - \gamma_{i_4}\right)$$
$$= \varphi(a)\varphi(b)\left(\mathcal{I}(i_1 = i_3) - \mathcal{I}(i_1 = i_4) - \mathcal{I}(i_2 = i_3) + \mathcal{I}(i_2 = i_4)\right). \tag{3.49}$$

The final term is

$$C = \xi(a)\xi(b)\left(I_{13} - I_{14} - I_{23} + I_{24}\right) \tag{3.50}$$

where $I_{k,\ell} = \mathcal{I}(i_k = i_\ell) \cdot \mathcal{I}(t_k = t_\ell)$. Defining $\mathbf{Z}_j$ to be the vector of non-overlapping pairwise differences at common ages, $\mathbf{Z}_j \sim N(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{Z}_j})$, where the elements of the covariance matrix can be built up from (3.46) and (3.47). From there maximum likelihood estimation can be carried out directly on the altered data.

A particular case where this tool is useful is when ψ_1, ψ_2 are highly confounded and the structural form of ψ_2 is misspecified. In this case the misspecification in ψ_2 can contaminate the estimate of ψ_1 due to the high correlation between their estimates. To demonstrate this we generate data from a population where $r_1(t, s) = .95^{|t-s|}$, $r_2(a_1, a_2) = .95^{|a_1-a_2|}$. To ensure there are plenty of measurement pairs at concurrent ages, we generated child ages at $t = 0$ as 3,6,9 and the observation times were 1,4,7,10,13,16 in each of the $F = 100$ families. The data configuration yields a small value for $M_{\mathcal{D}}$, indicating a poorly identified data condition. In this case we have incorrectly specified ψ_2 to be a linear function, when it is in fact a quadratic peaking at $a = 12$. Plots of the maximum likelihood estimate of ψ_1 from the altered data vs. the full data is given in Figure 3.7, indicating a reduction in bias. Estimates and error bars are based on the mean and standard deviations from 500 simulation replications, where non-converged runs are eliminated before constructing the plots.

In this case performing ML on the reduced data set proved to reduce the bias, but one should use caution when using this technique. First, as can be seen in (3.46), ψ_1 enters the variance as a coefficient on $2(1 - r(t_1, t_2))$, which is small at many data points in the settings we advocate this approach, since η_1 and η_2 are highly autocorrelated. A similar observation can be made from inspection of (3.48). This is very likely part of the explanation for the substantial variance in this estimate, and leads to less numerical stability in the optimization.

Another issue with this approach is that the sample size is significantly reduced with this method, which also contributes to the substantial increase in variance. In particular, since we are constrained to looking at measurements taken at concurrent ages, the range of the data is significantly reduced. In data sets of this type, it is unlikely that pairs of sibling observations at very young or old ages will be available. Therefore, the estimate of ψ_1 acquired from this technique will often be restricted to a possibly very small subset of the age range. The rigid age and timing structure we
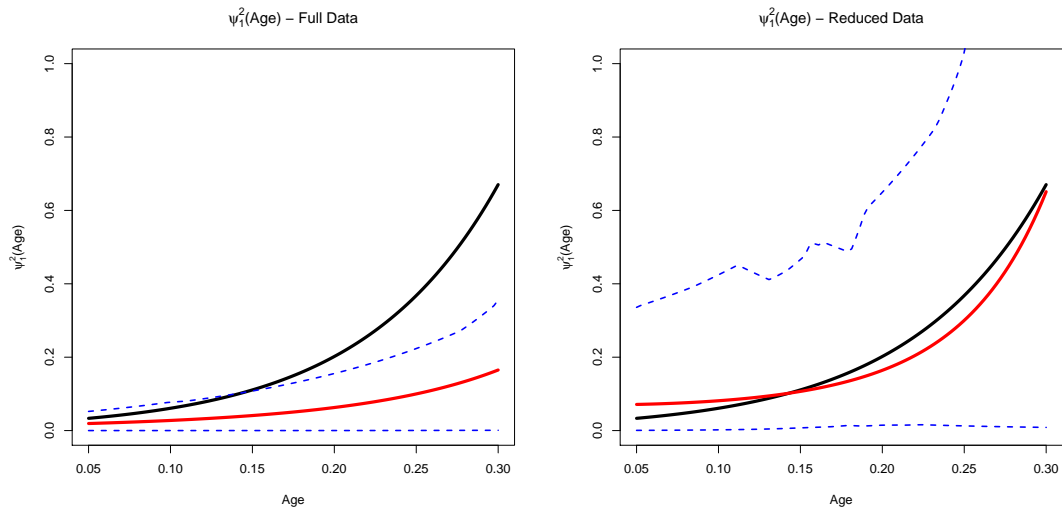
Figure 3.7: ψ_1 estimation precision using the MLE from the full data (Left) and from the reduced data (Right)

imposed for this simulation is, in some sense, optimal for this approach, which may not perform this well in other settings.

3.3.3 Computing techniques

If reduced down to a certain level, the model we have proposed here is a complicated random effects model that almost certainly fits into the Pinheiro/Bates framework of (1.4). However, it is the limit of the computing interface for specifying random effects that keeps model (1.4) from living out its full generality. Using the R package lme4, the user is constrained to, basically, specifying models that can be written as linear functions of fixed and univariate random intercepts and slopes, shared by some pre-specified blocks in the data set. Certain pre-packaged link functions can be used but only those utilized for traditional purposes, such as logistic regression with random effects, are available.

Two aspects of our model do not appear to be possible within a package such as lme4. First, specifying a latent process with *any* desired covariance structure is not possible. The most sophisticated incarnation of this available in lme4 would be

something like the random regression model described by Meyer (1998). A model like (3.15), for example, may be specified in `lme4` by

```
lmer(y ~ (1+t | Family) )
```

however, nothing like specifying an arbitrary parametric covariance function appears possible. Secondly, there does not appear to be a way to specify an arbitrary interplay between fixed and random effects. Even relatively simple interactions, such as the $\psi_1(a) \cdot \eta_1(t)$ terms in our model are not readily implementable. This are both likely due in large part to syntactic issues, since writing a method to translate to user input to a model specification for arbitrary model specifications may be impossible.

Our approach circumvents the problems with computer syntax and manipulates a code generation scheme which allows an arbitrary covariance structure. Through a somewhat cumbersome interface, the user can specify any entry of the covariance matrix to be an arbitrary mathematical function of parameters and/or data points. For example, to write out the contribution of a linear ψ_1 to an entry of the covariance matrix in our model with corresponding ages `a1,a2` and times `t1,t2` with a squared exponential covariance for η_{j1}, one would type

```
"(P[0] + P[1]*%f)*(P[0] + P[1]*%f)*exp(-P[3]*%f)" % (a1,a2, abs(t1-t2))
```

Strings of this form, which depend on the parameterization chosen, are then combined to build up the mean and covariance structure of every cluster in the data set. This process can usually be automated if the parameterization chosen is fairly regular between clusters, like our model is. Next the lists of means, covariances, and data points are run through an intermediary program which writes them to a template C script that contains nothing more than a generic function to call the multivariate normal log-likelihood, a function to carry out Nelder-Mead optimization of an arbitrary objective function, and necessary variable storage. This newly produced C program

is then compiled and run to optimize the likelihood as a function of the vector P and produce point estimates.

The algorithm described above bypasses the need to include superfluous functionality in the model fitting program. In addition, after shifting from a pre-compiled general function built to fit our model to this approach, we saw runtimes drop from several minutes to seconds. The longest part of this process is generally the compilation step, because if the data set is moderately large, then the number of lines in the final C script to fill in the mean vectors and covariance matrices can be substantial. This highlights the primary shortcoming with this approach, because no more than approximately 1000-2000 clusters can be accommodated with taking a prohibitively long time to compile. It may be possible to handle this problem using some sort of looping to reduce the size of the compiled file.

3.4 Model inference using the Likelihood Ratio Test

3.4.1 Some hypotheses of interest

A key feature of our modeling framework is that the parameters of interest have meaningful scientific interpretations. One of the primary hypotheses that we seek to make inference about is whether or not ψ_1, ψ_2 are constant functions. For example, a time-varying ψ_1 would indicate that the level to which a person's responses are modulated by the shared family environment at a particular time depends on their age. In this case, the variance function ψ_1, is interpreted as age-specific sensitivity to family environment. As an example, suppose a stressful event, such as the provider losing his/her job, occurs in the family environment, leading to a spike in η_1. A very young child is not likely to be affected much by this. On the other hand, a young adolescent may be more sensitive to such an event, since it would have implications to his/her material possessions and/or relationship with their parent. A

similar hypothesis may be formulated for ψ_2. The most natural example of this would be an age-specific onset of a genetic predisposition which is related to the response variable.

A second hypotheses that may be of interest to researchers are whether or not there is any familial correlation at all. Specifically, is it the case that $\psi_1(a) = \psi_2(a) = 0$ for every age? This can be accomplished by testing whether each parameter is a constant function at 0. While this hypothesis can be tested using simpler random effects models, it is not clear how this would be affected if, for example, the true correlation was non-constant but crossed 0 near the median age in the data set. A hypothesis of secondary interest that we give little attention to is to assess whether time-indexed factors are inducing more familial resemblance that age-indexed factors. This question may be answered by use of AIC or some other model selection criteria.

The first two hypotheses mentioned above can be formulated as sub-models of the full model, (3.24). Therefore, the likelihood ratio test (LRT) statistic,

$$\Lambda = 2 \left(\max_{\boldsymbol{\theta} \in \Omega} \mathbf{L}(\boldsymbol{\theta}) - \max_{\boldsymbol{\theta} \in \Omega_0} \mathbf{L}(\boldsymbol{\theta}) \right), \tag{3.51}$$

can be applied for inference. Here Ω denotes the full parameter space, and Ω_0 is the subspace defined by the null hypothesis. It is well known that, under the null hypothesis, $\Lambda \sim \chi_d^2$ where d is the difference in dimensionality between Ω and Ω_0. When testing whether either of the functions are identically 0, this presents a non-standard LRT problem. Much work has been done to study the behavior of the LRT statistic under non-standard conditions; the most well known work is that of Self and Liang (1987). They show that in many traditional cases, such as testing whether a single variance parameter is 0, the LRT is found to be conservative.

3.4.2 Problems defining a measure of effect size

We now attempt to investigate properties of the study population that lead to improved power for hypothesis testing. Clearly study design characteristics such as sample size, family size, and number of time points are relevant, but these are of less interest here since these may be constrained by things out of a researcher's control.

When considering ordinary likelihood ratio testing of the parameters relevant to ψ, certain measures of effect size are necessary. Recall from above the quantity $\lambda(a)$ defined in (3.32), the proportion of variation at a given age due to ψ_1. A natural measure of effect size for detecting non-zero ψ_k is

$$E_a(\lambda_k(a)) \qquad k \in \{1, 2\}. \qquad (3.52)$$

$E_a(\lambda_k(a))$ is the average contribution of $\psi_k(a)$ to the total variation across ages. Clearly, the larger this quantity, the greater the relative magnitude of $\psi_k(a)$. It is somewhat less clear to define an effect size for testing constancy of $\psi_k(a)$. One possibility is

$$Q_k = \text{var}_a(\lambda_k(a)) \qquad k \in \{1, 2\}. \qquad (3.53)$$

If Q_k is large, this indicates that the relative contribution of $\psi_k(a)$ changes a lot over the age range. One way in which this can happen is if $\psi_k(a)$ is a non-constant function. However, there are other possibilities:

- $Q_k = 0$ and $\text{var}_a(\sigma^2(a)) = 0$

    - Both $\psi_k(a)$ and $\xi(a)$ are constant

- $Q_k > 0$ and $\text{var}_a(\sigma^2(a)) = 0$ or $Q_k = 0$ and $\text{var}_a(\sigma^2(a)) > 0$

    - Both $\psi_k(a)$ and $\xi(a)$ are not constant

- $Q_k > 0$ and $\text{var}_a(\sigma^2(a)) > 0$

    - At least one of $\psi_k(a)$ and $\xi(a)$ are not constant

- If all variance functions are constant, then $\psi_k(a)$ is non-constant $\iff Q_k > 0$.

An important thing to notice here is that age-varying correlation does not necessarily imply that $\psi_k(a)$ is time varying, which makes this measure problematic. However, in some cases one may be able to rationalize that the impact of stable individual characteristics, $\varphi(a)$, and measurement error, $\xi(a)$, are constant over time. If this is the case, $\text{var}_a(\lambda_k(a))$ is a sensible measure of effect size.

In Figure 3.8 the value of Q_k is given as a function of the slope when all other functional parameters are constant, as well as the power (using the nominal LRT critical value of 3.84) as a function Q_1 for testing constancy of ψ_1 for various values of $E(\tau_1)$ when ψ_2 is fixed (in the data generation and model fitting) to be 0. In these simulations the there were 200 families with two kids in every family with $t = 0$ ages uniform on (5,10), observed three times, uniformly distributed on (0,20). We can see that Q_k is a strong determinant of power in the ideal case where all other functions are constant. However, in general, age-varying correlation does not necessarily indicate age-specific sensitivity to family environment. Thus Q_k does not provide a perfect measure of power to detect non-constant ψ_k.

As an example, suppose, under the same data conditions as above, $\psi_1(a) = .042 + .21a$ and $\xi(a) = .1 + .5a$, with all other functional parameters identically 0. In this case $Q_1 = 0$ but $\psi_1(a)$ is clearly non-constant. In this particular case, we have approximately .65 power to detect non-constant ψ. With the exact same ψ_1 and $\xi(a) = 1.6$ (the best constant approximation of $1 + .5a$), we have approximately .82 power. This indicates, in addition to the value of Q_1, other age-varying qualities of the data have an impact on power. This presumably because these two factors are mildly confounded when it comes to modeling the marginal variances, and the

marginal variances of the estimates of ψ and ξ increase in this case, leading to a loss in power.

Related to this discussion, there are other conceivable quantities related to the power to correctly decide in hypothesis tests involving ψ_1. As in ordinary the Pinheiro and Bates style mixed effects model, the number and size of clusters in highly relevant. When the sample size is fixed and one varies the cluster sizes (and implicitly the number of clusters) certain settings provide estimation precision of the variance parameters. For example having 5 families of size 20 will result in a less optimal situation for variance component estimation than having 20 families of size 5. Another relevant possibly related to power for ψ_k is the magnitude of the expected autocorrelation parameter $E(\tau_k)$. When $E(\tau_k)$ is very large, pairs of data points spread apart farther in time still yield non-zero correlations and therefore those data points will give more information about ψ_1 than if $E(\tau_k)$ were smaller; larger $E(\tau_k)$ effectively increases the sample size for estimating ψ_k. On the other hand, larger values of $E(\tau_k)$ also effectively reduces the total sample size. Perhaps these two factors are both at work, because $E(\tau_k)$ seems to have a negligible effect on power in Figure 3.8.

3.4.3 Power analysis in the full model

We now present a series of simulations to display the approximate level and power of tests our primary hypotheses of interest. The main hypotheses addressed here concern determining constancy of ψ_1, ψ_2 and determining whether either are identically 0. In addition, we intend to investigate the degree to which the confounding of ψ_1, ψ_2 discussed in section 3.2.3 contaminates the properties of the tests. Each of the simulation results presented are the result of 500 simulations, each with $F = 200$ families, 3 assessments per family, 2 children per family with initial ages uniformly distributed on (5,10) and assessment times uniformly distributed on (0,20). These values were chosen as possible demographics of a realistic, relatively sparse, long term
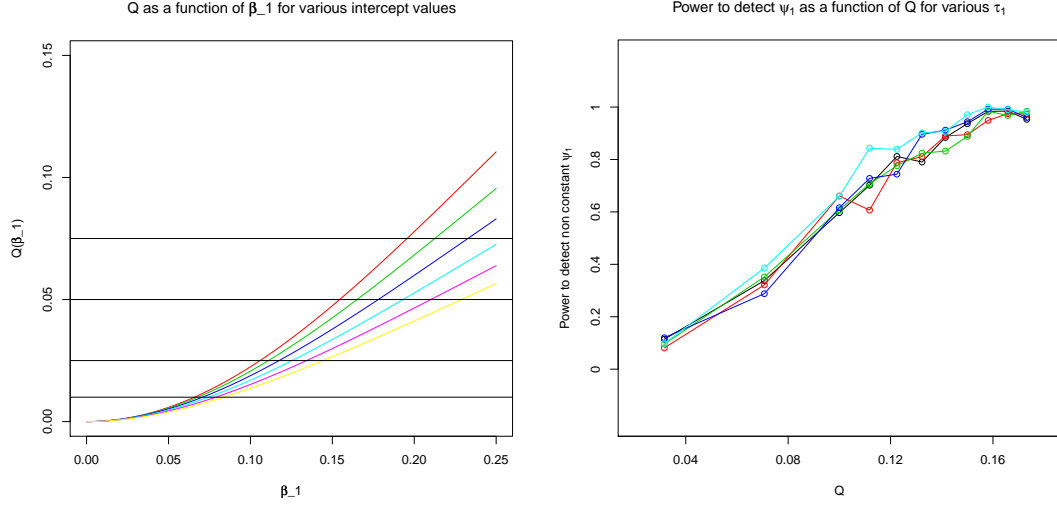
Figure 3.8: **Left**: Size of Q_k as a function of the slope with each line corresponding to $\beta_0 = 0, .05, .1, ..., .25$. **Right:** Power to correctly conclude the ψ_1 is non-constant as a function of Q_1 for $E(\tau_1) = 0$ (black), .25(red), .5(green), .75(dark blue), .9(light blue).

family study. This is reflected by the moderate number of children per family and average birth spacing being 5/3. All simulations below have $\varphi(a) = .5$ and $\xi(a) = 1$.

In the first set of simulations we test the null hypothesis that $\psi_k(a) \equiv c$. First focusing on testing ψ_1, we generate data from three population structures:

1. $\psi_1(a) = .1 + 1.25a$

2. $\psi_1(a) = .1 + 2a$

3. $\psi_1(a) = .1 + 3a$

with $\psi_2(a) = -.3 + 2.5a$ in each population. These three structures correspond to ψ_1 accounting for approximately 5%, 10%, and 15% of the total variation (averaged across ages), respectively. Note that age is scaled by 100 for greater numerical stability.

Each simulation compares two models: the first with $\psi_1(a)$ parameterized as a linear function, and the second with $\psi_1(a)$ constrained to be constant. To approximate

	Not Confounded		Confounded	
Population	Level	Power	Level	Power
1	.057	.223	0.067	0.262
2	.070	.526	0.128	0.621
3	.078	.775	0.160	0.821

Table 3.2: Monte Carlo estimates of the level and power of the test of the null hypothesis that ψ_1 is a constant function for three population structures and two degrees of confounding.

the finite sample level of the test, we also did analogous simulations where $\psi_1(a)$ truly was a constant function. In the second set of simulations, $\psi_1(a)$ was generated as constant at $\sqrt{.05}, \sqrt{.1}, \sqrt{.15}$, respectively to maintain the same proportion of variation due to ψ_1. Each of the two simulations were run when $E(\tau_1) = .5$, $E(\tau_2) = .9$ and $E(\tau_1) = .95$, $(\tau_2) = .95$ to determine the effect of confounding on the power. The results of this first simulation are given in Table 3.2.

We can see that the power is very poor for population 1 in both cases, and increases as the slope becomes more steep. In the confounded population the nominal critical value of 3.84 does not seem to be appropriate, as the null distribution seems to exceed it far more than 5% of the time. This is very likely due to the fact that, since ψ_1 and ψ_2 are confounded, ψ_2 can compensate for the lack of slope in ψ_1, thus making the null hypothesis less likely to be rejected when ψ_1 truly is constant.

An exactly analogous simulation study was carried out with the role of ψ_1 and ψ_2 reversed and is summarized in Table 3.3. In this case the power is lower in every case than the corresponding tests of ψ_1, although the significance levels are not inflated as they were before. While this is disconcerting, we may still be confident that a significant result encountered in real data is indeed significant. This may be due that the asymptotics of the LRT statistic have not kicked in yet. Non-orthogonal predictors can have the effect on slowing down convergence and this may be happening here to a greater degree with ψ_2 than ψ_1.

Our second set of simulations focuses on determining whether ψ_1, ψ_2 are identically

	Not Confounded		Confounded	
Population	Level	Power	Level	Power
1	.037	.168	0.057	0.184
2	.037	.208	0.025	0.226
3	.045	.422	0.035	0.233

Table 3.3: Monte Carlo estimates of the level and power of the test of the null hypothesis that ψ_2 is a constant function for three population structures and two degrees of confounding.

zero. To do this we simulate data with φ, ξ the same as above with $\psi_2(a) = .1 + 2a$ and generate ψ_1 from three populations that are each qualitatively quite different:

1. $\psi_1(a) = -.125 + 2.5a$

2. $\psi_1(a) = -.7 + 4a$

3. $\psi_1(a) = .447$

The first population structure is meant as a simple linear structure that accounts for, on average, about 8% of the total variation. The second population is one where ψ_1 crosses zero around the median point of the age range. We would like to investigate whether this adversely affects the power. Finally, the third is a scenario we expect to do quite well, since ψ_1 is constantly very far from 0, accounting for about 20% of the total variation. The results of this simulation are shown in Table 3.4. Notice that the simulations under the null hypothesis do not depend on the three populations, thus only one number is reported. Once again we carry out the same simulations with the roles of ψ_1 and ψ_2 reversed, the results of which can be found in Table 3.5.

We can see in these simulations that the nominal critical value for the LRT is very conservative. This agrees with the standard theory of Self and Liang (1987), since we are testing on the boundary of the parameter space. In this case we see an astounding loss of power when comparing the confounded data setting with the non-confounded.

	Not Confounded		Confounded	
Population	Level	Power	Level	Power
1	.006	.343	.007	.077
2	-	.317	-	.329
3	-	.807	-	.172

Table 3.4: Monte Carlo estimates of the level and power of the test of the null hypothesis that ψ_1 is identically 0 for three population structures and two degrees of confounding.

	Not Confounded		Confounded	
Population	Level	Power	Level	Power
1	.012	.301	.006	.181
2	.008	.511	.003	.428
3	.011	.583	.002	.153

Table 3.5: Monte Carlo estimates of the level and power of the test of the null hypothesis that ψ_2 is identically 0 for three population structures and two degrees of confounding.

This gives a legitimate cause for concern, but will only result in conservatism. We can still be confident that a significant result is legitimate in practice.

We have shown with these simulations that the LRT statistic is, at worst, conservative when we are not in a highly confounded data setting. The loss of power once entering the confounded data setting is not surprising, but the magnitude of the effect is troublesome. This is particularly a concern, since there is no more than a vague diagnostic for determining whether a given data set could be a problem. Secondly, the LRT test is largely conservative for testing the null hypothesis of one of the functional coefficients being identically 0. This is consistent with the conventional wisdom and should be noted when something appears near the border of significance in such tests. Finally, the power of the inference for ψ_2 is considerably smaller than analogous tests for ψ_1. While we do not have a particular explanation for this, this does seem to agree with the results reported in Table 3.1, which showed that the estimates of the ψ_2 parameters had higher sampling variance.

	Testing ψ_1		Testing ψ_2	
Population	Level	Power	Level	Power
1	.089	.120	.102	.156
2	.067	.535	.114	.418
3	.032	.930	.061	.830

Table 3.6: Monte Carlo estimates of the level and power for LRT tests of constancy in ψ_1(left), ψ_2(right) for three population structures.

3.4.4 Power analysis in the reduced model

The purpose here is first to determine whether this loss in power for testing ψ_2 is a property of the joint model, or whether ψ_2 is simply more difficult to estimate than ψ_1, generally, which would agree with the results in Table 3.1, where the estimated slope of ψ_2 appears to have considerably greater variance. The second reason for this inquiry is regarding the conventional approach in these types of models to choose age as the time index, rather than chronological time. In some applications, such as genetic analysis where the environmental covariance is assumed to be 0, age indexing may be the only sensible thing, since chronological time clearly does not affect genetics. However, in applications where there is a judgment to be made, these results may be of interest.

To investigate the properties of these tests we carry out simulations exactly analogous to those above with only one of ψ_1, ψ_2 in the model. To compensate for the decrease in marginal variance, the value of ξ, while still be left constant, is increased by the appropriate amount. The tests of the null hypothesis of $\psi_k \equiv c$ is summarized by Table 3.6. The results of testing $\psi_k \equiv 0$ are given in Table 3.7.

We can see that the tests are much more reasonably behaved since each of the models are much more well identified than the full model. The drastic advantage that ψ_1 appeared to have in power has largely disappeared, although there does seem to be a moderate difference in the power for testing constancy of the functional parameters. In addition, the significance levels for testing $\psi_2 \equiv c$ appear slightly

	Testing ψ_1		Testing ψ_2	
Population	Level	Power	Level	Power
1	.031	.781	.001	.745
2	-	.534	-	.512
3	-	.973	-	.943

Table 3.7: Monte Carlo estimates of the level and power for LRT tests of ψ_1(left), ψ_2(right) being identically 0 for three population structures.

inflated. For testing whether either function is identically 0, both seem to struggle a bit when the true, non-constant, function crosses 0 near the median age.

Coming back to the purpose for this inquiry, it does not appear that ψ_2 is inherently difficult to estimate or underpowered for hypothesis testing, but it does not outperform ψ_1 in any case. In addition, in the joint model, the inference on ψ_1 and estimation precision appears to perform generally better and is less affected by the confounding than ψ_2. This calls into question the default choice of age indexing in models of the type with a single timing variable. Given the more favorable properties of ψ_1 when the true model contains ψ_1 (and not ψ_2), a practioner may be well served to do some model selection first to determine which indexing fits better.

3.5   Sensitivity to misspecification of η covariance structure

Since the choice for the covariance functions of the η processes is largely arbitrary we carry out some simulations to investigate the sensitivity of this specification to estimation of the parameters of primary interest– ψ_1, ψ_2. In other settings where a covariance matrix or function is considered a nuisance parameter, the AR1 is the default parameterization. In our model, we go slightly beyond that, but our choice is still susceptible to misspecification. In this section we test our model's sensitivity to three types of misspecification in the covariance structure of η: sample paths of the process are not identically distributed, non-stationary, or non Gaussian.

In the first two simulations we generated a population structure such that each

of $F = 100$ families had four kids, observed six equally spaced times from 0 up to 15 and initial ages with initial ages $4, 7, 10, 13$. In the third simulation there are $F = 200$ families observed only 3 times at initial ages $4, 7, 10$. Purely arbitrarily, these simulations are displayed for ψ_2 rather than ψ_1. The estimation of ψ_1, whose covariance was correctly specified, is not shown. For all three studies 500 simulations were done.

The first misspecification in $\text{cov}(\eta)$ we analyze critically is the "iid" assumption. Technically our choice, (3.29) with the autocorrelation parameter specified as a random effect, may be thought of as not identically distributed, since the autocorrelation parameter varies family to family, but we consider a different type here. The true structure is that $\text{cov}(\eta$ has an exchangeable correlation structure where the off diagonal entries are Uniform$(0, 1)$ drawn by family. In the left panel of Figure 3.9, we can see that somewhat substantial downward bias occurs from this misspecification, while the variance of the estimate doesn't seem particularly large.

One explanation for the observed bias that an exchangeable correlation can induce relatively large correlations in $\eta_2(a)$ for ages quite far apart. In the generated data, for example, there is a 4 year old and a 28 year old in each family (among other sizeable gaps), so in order to accomodate such large age lags, the autocorrelation parameter must be overestimated. This means that the random coefficient on $\psi_2(a)$ is too large at many entries in the covariance matrix, thus $\psi_2(a)$ must be underestimated to compensate.

The second covariance assumption we challenge is that of stationarity. Previous authors [e.g. Pletcher and Geyer (1999)] have noted that the assumption of stationarity is almost certainly false, but is likely to provide a reasonable approximation. To test this we generated $\eta_2(a)$ so that

$$\eta_2(a_2) = \tau(a_2) \cdot \eta_2(a_1) + \sqrt{1 - \tau(a_2)^2} \cdot \epsilon \tag{3.54}$$

where $\epsilon \sim N(0,1)$, and $\tau(a) = e^{-(a-17.5)^2/50}$. This covariance function makes η_2 highly autocorrelated near 17.5, but essentially white noise elsewhere. Such a structure would not make much sense when indexed by time, which is a rationale for demonstrating this on ψ_2. The results of this simulation are shown in the top right of Figure 3.9 and indicate that this type of non-stationary has little if any effect on the estimation. When there are enough timepoints, as there are here, it appears that the model is fairly robust to non-stationarity.

The final covariance misspecification we investigate is non-Gaussianity. The Gaussian assumption is almost automatic in many statistical problems with continuous data as a convenience. In our model this stipulation is convenient because we can neglect higher order dependencies beyond covariances. To test this assumption we generated $\eta_2(t)$ as the sign of an AR1 process with $\tau = .99$. The type of misspecification is very far from the parametric assumption and should yield a problem if the estimation is at all sensitive to non-normality. This results in what appears to be small bias, but significantly increases the variance, as can in the bottom panel of Figure 3.9.

Each of the three types of misspecification investigated here can be difficult to assess a priori, so a researcher may have difficulty knowing whether the results shown here should be of concern. The stationary assumption is probably the most difficult to assess, but is also less likely to be a major problem. While pathological examples of non-stationary processes may seriously damage estimation in our model, we do not expect the data we have in mind to have such properties.

The non identically distributed assumption, which presented the largest problem in our simulations, is probably also the easiest to spot. In some data settings, such as stratified populations, a researcher may have good reason to believe there is between cluster heterogeneity in the temporally indexed cluster dynamics. When this is a concern stratifying the sample and performing this analysis on each strata separately

92
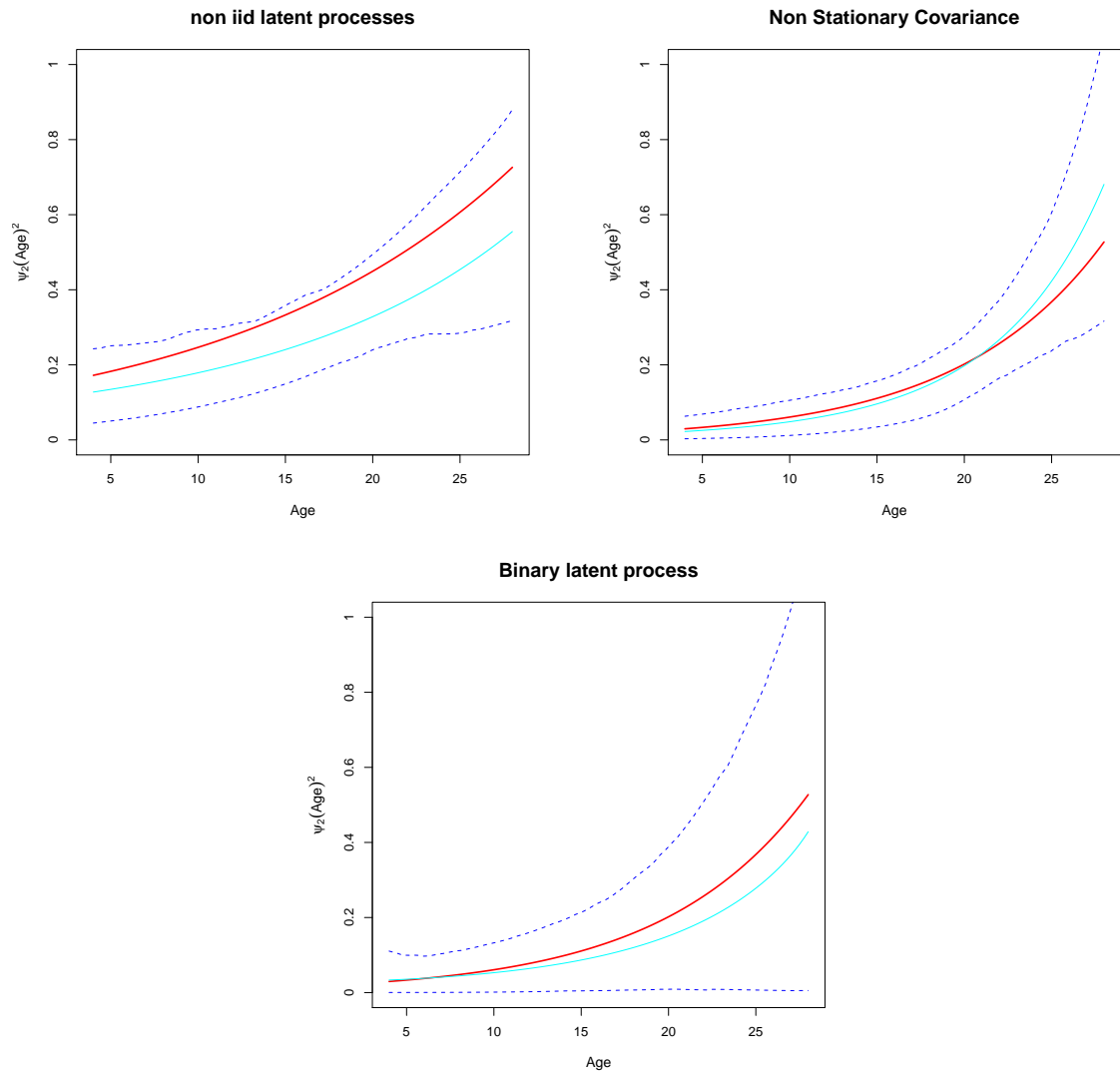
Figure 3.9: Estimate of ψ_2 (red) with empirical 95% pointwise confidence bands (dashed blue lines) along with the true ψ_2(light blue). The covariance structure of η is non identically distributed (top left), non-stationary (top right), and non-Gaussian (bottom).

may be appropriate.

The Gaussian assumption is the most straightforward to check in our model. In the case where η_2 is a binary process, as above, a histogram of the data would most likely raise a red flag, since the sum of normal variables should be normal. In addition, non-normality, particularly of the extreme type simulated here may be more obvious to spot. For example, a genetic predisposition which turns on a different ages in different families (if at all) may be give rise to a latent process similar in structure to the one generated above. However, even the extreme non-normality simulated here did not cause the machinery to break down, and most realistic violations of this assumption are likely to be less serious.

3.6   Analysis of the DOTS-R data

Our motivating application for this work are longitudinal behavioral measurements made on individuals in families gathered as part of the Michigan Longitudinal Study (MLS) [Zucker et al. (1996)]. The particular section of the MLS data we analyze here are eight DOTS-R measurements. The DOTS-R exam is a widely used quantitative tool used for characterizing temperament as individuals progress from childhood through early adulthood. The eight DOTS-R items we have here are sums of responses to various scales constructed to capture distinct characteristics described in Table 3.8.

We begin by describing the demographic characteristics of the dataset. There are a total of 579 children from 286 families, for an average of about 2 kids per family; the full breakdown is given in Table 3.9. The DOTS-R is very sparse longitudinal data, with only 51.1% of children observed twice and the rest only observed once. Our first goal with this data is to investigate the extent to which familial resemblance is present. If there is, we would then like to determine whether this is arising from temporal or age indexed factors, and if there is age-specific sensitivity to these factors.

94

Item	Description
1	Self Activity-General
2	Self Activity-Sleep
3	Approach/Withdrawal
4	Self Flexibility/Rigidity
5	Self Mood
6	Rhythmicity-Sleep
7	Rhythmicity-Eating
8	Rhythmicity-Daily Habits

Table 3.8: Descriptions of the DOTS-R items

Number of children	Percentage of families
1	26.5%
2	50.6%
3	16.8%
4	5.6%
5	.3%

Table 3.9: Number of children per family in the DOTS-R data

A practical issue which arises with real data analysis is starting values. For the simulation studies we have the luxury of knowing the true population parameters, and can start there to presumably arrive at the global mode. In real data we must make some considerations for this problem. A simulation study was carried out (not shown) and it was found that by starting from 10 random starting values, our program finds a converged point that is within .2 log likelihood points of the global mode over 90% of the time. More random starts does not generally payoff, as increasing from 10 to 50 only raises this rate to about 96%. We use 10 random starts and retain the point with the highest likelihood as the estimate for each DOTS-R item.

The handling of the nuisance parameters is addressed first. Due to the very sparse nature of the data, we do not expect the fitting of the three parameters underlying each of the covariance functions for η_1, η_2 to be estimated well, so we simply use the squared exponential covariance. The nuisance functions, φ, ξ were parameterized as linear functions to prevent ψ_1, ψ_2 from being identified through variances.

To address our first objective, we formalize the problem by testing $H_0 : \psi_1 = \psi_2 = 0$ against

1. ψ_1, ψ_2 both non-zero

2. ψ_1 zero, ψ_2 non-zero

3. ψ_1 non-zero, ψ_2 zero.

If there is suggestive evidence against H_0 by comparison with the the first alternative, each individual function can be tested marginally in an attempt to determine where the familial correlation is coming from. Table 3.10 contains LRT statistics of each hypothesis. there is strong evidence for familial correlation in items 4, 6, which correspond to Flexibility and Sleep Rhythmicity. In addition, due to the known conservative nature of the test, as displayed in section 3.4.3, there is suggestive evidence of familial correlation in items 3, 7, and 8 as well.

Item 4 is the only item where there is unequivocal evidence that time indexed causes of familial correlation are having a greater impact than age indexed causes. For this item the AIC difference is about 1.5, which is beyond the rule of thumb stated by Burnham and Anderson (1998). The point estimates in this case are $\widehat{\psi}_1 = 1.424$ and $\widehat{\psi}_2 = 1.064$ in the marginal models and $\widehat{\psi}_1 = 1.229$, $\widehat{\psi}_2 = .860$ in the joint model. ψ_1 also appears to be having a greater impact on suggestive items 3 and 8.

Now that we have established that there is some familial resemblance in the data, we look to test whether there is evidence for time varying sensitivity to family environment in any of the items. We accomplish this by comparing the model with ψ_1, ψ_2 constrained to be constant with three other models:

1. ψ_1, ψ_2 both linear

2. ψ_1 constant, ψ_2 linear

3. ψ_1 linear, ψ_2 constant.

Item	1	2	3
1	2.984	2.980	2.972
2	0.026	0.026	0.026
3	3.698	2.172	3.353
4	**11.802**	**10.856**	**8.223**
5	3.170	2.138	2.527
6	**13.318**	**12.370**	**12.092**
7	3.948	3.636	3.636
8	5.684	4.286	5.118

Table 3.10: LRT statistics testing $\psi_1 \equiv \psi_2 \equiv 0$, and $\psi_1 \equiv 0$, $\psi_2 \equiv 0$ separately with the other constrained to be 0. DOTS-R items

Item	1	2	3
1	1.154	1.126	0.271
2	0.235	0.079	0.231
3	**1.836**	0.374	**1.581**
4	0.052	0.008	0.034
5	0.552	0.497	0.110
6	0.498	0.441	0.274
7	**3.435**	**2.835**	0.227
8	**1.967**	0.122	**1.759**

Table 3.11: LRT statistics testing for significant slope in ψ_1, ψ_2 simultaneously (column 1), and ψ_1 and ψ_2 separately (columns 2,3) for each of the 8 DOTS-R items

The LRT statistics corresponding to these comparisons are given in Table 3.11. While none of these results are nominally significant, there is suggestive evidence in item 7 of some age-varying sensitivity to time indexed environment. To a lesser extent, items 3, 8 show evidence of non-constant ψ_2. In each of these three cases, it seems that the age-specific sensitivity to family environment is occurring with respect to only a single one of ψ_1, ψ_2 and not both. In each item, the level of familial correlation is decreasing as a function of age.

3.7 Discussion

Random process models for the analysis of temporally dependent traits were originated in Kirkpatrick and Heckman (1989), Kirkpatrick et al. (1990) and were later extended by the Meyer and Hill (1997), Meyer (1998), Pletcher and Geyer (1999), Jaffrezic and Pletcher (2000), among others. Such models make intuitive sense and reduce the problem to estimation of the covariance functions underlying the genetic and environmental random processes. Our contribution to the field is a method of encorporating effects on the correlations which are modulated by both age and chronological time by borrowing and extending methods most closely resembling those of Pletcher and Geyer (1999). In addition, our method is the first to introduce clustered data models that are not definitionally identified by a particular blocking structure in the data, but rather by the differential staggering of age and time gaps in the data. This disaggregation is an ambitious estimation task, and we have partially characterized the situations where the decomposition of the two latent processes is possible.

The framework described in this chapter has many attractive features. The parameters in the model are interpretable and are potentially of scientific interest. While ψ_1, ψ_2 are related and are targeting similar characteristics, they are distinct entities that can be separated, as demonstrated herein. ψ_1 may be thought of as a broad measure characterization of a subject's sensitivity to the environment shared within a family (or, more generally, a cluster) at a particular chronological time. ψ_2 is a measure of sensitivity to age-dependent sources of correlation within a family, which would include genetics. Having the availability to simultaneously model these two components, as opposed to making an arbitrary specification of how time is indexed, is desirable property of our framework.

One primary shortcoming of our approach is that the assumptions of the model may be exceedingly difficult to test. For example, there is no straightforward way

of determining whether our parametric specification of the covariance structure for η is reasonable. We have shown in simulations that this specification can be crucially important, and can result in sizeable bias and/or additional variance in some cases. In addition, testing other types of misspecification is problematic, such as the additivity of familial correlation inducing processes. A non-parametric approach would be necessary to circumvent these issues, but, as noted by Pletcher and Geyer (1999), this is simply not feasible with the types of sample sizes likely to occur with such data. In such an estimation setting, the higher variance non-parametric estimates may not balance favorably with the bias induced by a misspecified parametric model.

More work needs to be done on how to characterize the level of confounding present in a given data set and potential strategies to mitigate this. The massive negative impact of the confounding on the significance levels and, more severely, the power of the LRT is very troublesome. An alternative method which is not this sensitive must be sought. In addition, noting the infeasibility of the most intuitive measure of effect size, some other quantification is necessary to characterize the power of hypothesis tests under particular conditions.

Some promising directions for future work may be to explore an alternative estimation procedure the does not necessitate the estimation of so many nuisance parameters. One possibility here is to use composite likelihood. For example, there may be some way of using composite likelihood to make the nuisance parameters φ, ξ drop out of the model. Possible gains in computational simplicity would be a primary reason for this inquiry.

Other avenues of future work are related to modeling phenomena likely to present themselves in real data. For one, researchers may be interested in analyzing multiple traits simultaneously. This could potentially be accounted for in this framework by including cross covariance functions as in the multivariate character process model of Jaffrezic et al. (2004). In this way, age and chronological time effects on the

99

multivariate process evolution can be explored. Another natural extension of our model would be to include modifications to analyze genetically informative data. Finally, many longitudinal family studies include data from multiple generations, which would be straightforward to include in this framework.

A second avenue for future work regards including some constructs likely to affect the response which are not currently accounted for by our framework. Our current model assumes that all subjects are governed by a single environmental sentivity function, ignoring the fact that birth order is relevant. It seems unlikely that the first born sibling would respond in the same way as the middle or youngest sibling in the family to the familial environment. In some more traditional work [e.g. Pawitan et al. (2005)], an indicator of "first born" is used as a fixed effect in the model, but this does not modulate the correlations in any way. Potentially using such an indicator as an additional argument to the functions ψ_1, ψ_2 is a simple modification that could be interesting. More generally, variables other than just age may modulate sensitivity to the shared latent process. Investigation of general covariate effects on sensitivity is a possibility.

Along similar lines, we have assumed that effects which occur at a particular time in the family affect the response immediately. In reality, it is likely that some sort of age specific warping would occur. In particular, subjects of a particular age may take a different amount of time to respond to an environmental stimuli than subjects of another age. Roughly, this would correspond to something like

$$\eta(t) \longrightarrow \eta(t + \Delta(a)) \tag{3.55}$$

Research on what type of data is required to properly identify $\Delta(a)$ and the precision with which it could be estimated represents an interesting possiblity. $\Delta(a)$ would give some quantification to how individuals react to events and may be of interest to social science researchers.

CHAPTER IV

Point estimation in multiple timing variable models for binary longitudinal family data

4.1 Introduction

Many traits in which familial association structure is of interest are binary in nature. For example, the presence or absence of a disease or smoker/non-smoker status are binary traits whose coherence within a family may be of substantive interest to a researcher. In this chapter we seek to extend the methods of chapter 3 for use on binary data.

In the analysis of quantitative traits that are discrete, natural extensions of the continuous time models are used. In many approaches [e.g. Pawitan et al. (2005)], the discrete trait is viewed as arising from some underlying *liability* that is a continuous variable. Keeping with the analogy to the classical continuous trait genetics model, the discrete response, $\mathbf{Y}$ is modeled as

$$\phi(\mathbf{Y}) = \boldsymbol{\mu} + \mathbf{g} + \mathbf{e} \tag{4.1}$$

where ϕ is some link function. The model implied correlations of the liabilities are known as *polychortic correlations* and are a standard way of quantifying familiality in binary family data. Model (4.1) amounts to fitting a generalized linear mixed model, which is discussed chapter 1. Other approaches exist that do not assume an

underlying liability [e.g. Hopper et al. (1984)], and work directly with the discrete distribution. These approaches largely make us of some form of log-linear model, which is also discussed in chapter 1.

The literature of models for discrete longitudinal family data are very sparse. Little work on extending the "character process" model of Pletcher and Geyer (1999) to the discrete setting has been done. The primary reference in this area is is Pletcher and Jaffrezic (2002), where the authors model the discrete variables as arising from an underlying unobserved continuous variable that follows the model of Pletcher and Geyer (1999), Jaffrezic and Pletcher (2000). The link between the observations and the underlying continuous variable is governed by a second set of parameters and the model is fit using a hybrid MC/EM algorithm.

We begin by viewing the binary responses as thresholded normal variables and discuss why certain modifications to the model formulation must be made to ensure identifiability. Estimation in this model presents a more formidable task, since MLE calculation is more cumbersome. We discuss why ML presents a numerically intractable problem and discuss an alternative approach to estimation based on composite likelihood (CL). Simulation results are given to show point estimation accuracy in this model. We apply this method to simulated data and to each of the CBCL items used in Chapter 2 and interpret the results for some particular items of interest. The chapter concludes with a brief discussion and some suggestions for extensions of the model to more general categorical data.

4.2   Model Formulation

Let $Y_{ijt} \in \{0,1\}$ denote the response of subject i in family j at time t and age a_{ijt}. We choose a parameterization analogous to the continuous case; for a rationale of this model formulation see Chapter 3. We assume unobserved Gaussian random variables X_{ijt} following model (3.24) underlie the observed binary random variables.

That is,

$$X_{ijt} = \mu(a_{ijt}) + \psi_1(a_{ijt})\eta_{j1}(t) + \psi_2(a_{ijt})\eta_{j2}(a_{ijt})$$

$$+ \varphi(a_{ijt})\gamma_i + \xi(a_{ijt})\varepsilon_{ij}(t). \tag{4.2}$$

Then the observed binary outcome is

$$Y_{ijt} = \mathcal{I}(X_{ijt} > 0). \tag{4.3}$$

The correlation structure of the underlying continuous variables $\boldsymbol{X}_j$ induces association between the observed binary variables. As observed by other authors [e.g. Rabe-Hesketh and Skrondal (2001)], the observed variables are thresholded versions of the continuous responses, and thus the scale of the latent processes can not be deduced from the observations. More specifically,

$$P(Y_{ijt} = 1) = P(X_{ijt} > 0) = P\left(\frac{X_{ijt} - \mu(a_{ijt})}{\sqrt{\sigma^2(a_{ijt})}} > \frac{-\mu(a_{ijt})}{\sqrt{\sigma^2(a_{ijt})}}\right) = \Phi\left(\frac{\mu(a_{ijt})}{\sqrt{\sigma^2(a_{ijt})}}\right) \tag{4.4}$$

where $\sigma^2(a)$ is defined in (3.31) and Φ is the standard normal CDF. The threshold value, $\frac{\mu(a_{ijt})}{\sqrt{\sigma^2(a_{ijt})}}$, is identified uniquely, but not each term in the ratio. Furthermore, equation (3.33) shows that the correlation between X_{ijt} and X_{kjs} (known as the *polychortic correlation* between Y_{ijt} and Y_{kjs}) only depends on

$$\psi_1^\star(a) = \frac{\psi_1(a)}{\sqrt{\sigma^2(a)}} \qquad (4.5)$$

$$\psi_2^\star(a) = \frac{\psi_2(a)}{\sqrt{\sigma^2(a)}} \qquad (4.6)$$

$$\varphi^\star(a) = \frac{\varphi(a)}{\sqrt{\sigma^2(a)}} \qquad (4.7)$$

$$\xi^\star(a) = \frac{\xi(a)}{\sqrt{\sigma^2(a)}} \qquad (4.8)$$

When $\sigma^2(a)$ is fixed to a constant, it is known that the polychortic correlations are identified in thresholded gaussian models [e.g. Pearson (1900)]. In this model we fixed $\sigma^2(a) = 1$ and estimate $\mu(a)$ and functions (4.5)-(4.8). In practice this is done by letting the functions $\psi_1, \psi_2, \varphi, \xi$ be unconstrained, but scaling them by the sum of their squares when they enter the covariance matrix.

These parameters defined by (4.5)-(4.8) will carry a similar interpretation to the analogous terms in the continuous case, but are somewhat different. For example, it is possible for $\psi_1^\star(a)$ to be a constant function when the underlying $\psi_1(a)$ which determined it is a non-constant function. This is related to the discussion of effect size in chapter 3. $\psi_1^\star(a)$, $\psi_2^\star(a)$ should now be interpreted as the contribution of the time and age indexed sources of familial association to the correlation rather than covariance.

4.3 Estimation

4.3.1 Maximum Likelihood Estimation

To begin we describe the theoretical possiblity of maximum likelihood estimation and describe why it is not feasible for this problem. Let $\boldsymbol{\theta}$ denote the concatenation of all parameters in the model, respectively. Also let $\boldsymbol{t}_j$, $\boldsymbol{a}_j$ be the list of observation

times and ages for the N_j length family j data, consisting of K_j siblings. Let

$$\eta_{j1}(\boldsymbol{t}_j) \sim N(0, \boldsymbol{\Sigma}_{\eta_1}) \tag{4.9}$$

$$\eta_{j2}(\boldsymbol{a}_j) \sim N(0, \boldsymbol{\Sigma}_{\eta_2}) \tag{4.10}$$

$$\boldsymbol{\varepsilon}_j = (\varepsilon_1, ..., \varepsilon_{N_j}) \sim N(0, I) \tag{4.11}$$

$$\boldsymbol{\gamma}_j = (\gamma_1, ..., \gamma_{K_j}) \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}). \tag{4.12}$$

As in the continous case, $\boldsymbol{\Sigma}_{\eta_1}, \boldsymbol{\Sigma}_{\eta_2}$ are covariance matrices known up to a finite number of parameters contained in $\boldsymbol{\theta}$; $\boldsymbol{\gamma}$ has a block diagonal covariance matrix and ε has a diagnonal covarance matrices; all random effects have unit variance.

The mean vector for the underlying normal variables is, of course, $\boldsymbol{\mu}_j(\boldsymbol{\theta}) = \mu(\boldsymbol{a}_j)$. The covariance matrix is

$$\boldsymbol{\Sigma}_j(\boldsymbol{\theta}) = \boldsymbol{\psi}^{\star\prime}_1 \boldsymbol{\Sigma}_{\eta_1} \boldsymbol{\psi}^{\star}_1 + \boldsymbol{\psi}^{\star\prime}_2 \boldsymbol{\Sigma}_{\eta_2} \boldsymbol{\psi}^{\star}_2 + \boldsymbol{\varphi}^{\star\prime} \boldsymbol{\Sigma}_{\gamma} \boldsymbol{\varphi}^{\star} + \boldsymbol{\xi}^{\star\prime} \boldsymbol{\xi}^{\star}. \tag{4.13}$$

Now letting

$$\boldsymbol{R} = \{R_1 \times R_2 \times \ldots \times R_{N_j}\} \tag{4.14}$$

where $R_k = (0, \infty)$ if the k-th element of $\boldsymbol{Y}_j$ is 1, and $R_k = (-\infty, 0)$ otherwise. Then the family j log-likelihood can be written as

$$\mathbf{L}_j(\boldsymbol{\theta}) = \log \int_{\boldsymbol{R}} \phi(\mathbf{x} \; ; \; \boldsymbol{\mu}_j(\boldsymbol{\theta}), \boldsymbol{\Sigma}_j(\boldsymbol{\theta})) d\mathbf{x} \tag{4.15}$$

where $\phi(\mathbf{x} \; ; \; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ evaluated at $\mathbf{x}$. Thus a likelihood calculation for a single family amounts to an N_j-dimensional numerical integration. Even in relatively sparse

datasets, N_j could easily be 10 or more, in which case time consuming numerical integration or monte carlo techniques must be used to approximate the likelihood. In either case there are many tuning parameters related to speed and accuracy of the resulting likelihood approximation that make ML an unappealing option here.

4.3.2 Composite Likelihood Estimation

As noted by Lindsay (1988), composite likelihood is a viable substitute when the maximum likelihood estimate is difficult to calculate. With this being the exact situation we find ourselves in, we now describe a composite likelihood estimation approach to estimation in the model defined by (4.2), (4.3). Lindsay (1988) also mentions that modeling the distribution of pairs of variables is a common way to construct composite likelihoods. Taking this approach, we can construct an estimation criterion which only requires the calculation of one and two-dimensional integrals.

Consider a pair of observations between family members i, k in family j at times t, s and ages a_{ijt}, a_{kjs}. The pair of liability variables, X_{ijt}, X_{kjs} are distributed as bivariate normal with mean $\mu = (\mu(a_{ijt}), \mu(a_{kjs}))$ and covariance matrix Σ with diagonal elements equal to 1 (from the constraint imposed by (4.5)-(4.8)), and off diagonal entry

$$
\begin{aligned}
\Sigma_{12} = {} & \psi_1^\star(a_{ijt})\psi_1^\star(a_{kjs})[\Sigma_{\eta_1}]_{t,s} + \psi_2^\star(a_{ijt})\psi_2^\star(a_{kjs})[\Sigma_{\eta_2}]_{a_{ijt},a_{kjs}} \\
& + \mathcal{I}(i_1 = i_2)\left(\varphi^\star(a_{ijt})^2 + \xi^\star(a_{ijt})^2\mathcal{I}(s = t)\right)
\end{aligned}
\tag{4.16}
$$

Letting $\phi_{ikts}(z_1, z_2|\boldsymbol{\theta})$ denote the bivariate normal density with the corresponding parameters defined as functions of i, k, t, s and $\boldsymbol{\theta}$ as above. Then the log-likelihood for the given pair of observations, Y_{ijt}, Y_{kjs} is

$$\mathbf{L}_{ikts}(\boldsymbol{\theta}) = \log \int_{R_{ik}} \phi_{ikts}(z_1, z_2|\boldsymbol{\theta}) dz_1 dz_2 \tag{4.17}$$

where R_{ik} is defined analogously to $\mathbf{R}$ in (4.14). Define

$$J_i = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-(z-\mu(a_{ijt}))^2/2} dz \tag{4.18}$$

and define J_k analogously. Also let

$$J_{ik} = \int_{-\infty}^{0} \int_{-\infty}^{0} \phi_{ikts}(z_1, z_2|\boldsymbol{\theta}) dz_1 dz_2. \tag{4.19}$$

Then the joint probabilities $p_{m,\ell} = P(Y_{ijt} = m, Y_{kjs} = \ell)$ can be written as

$$p_{00} = J_{ik} \tag{4.20}$$

$$p_{01} = J_i - J_{ik} \tag{4.21}$$

$$p_{10} = J_k - J_{ik} \tag{4.22}$$

$$p_{11} = 1 - (J_i + J_k - J_{ik}) \tag{4.23}$$

The joint likelihood for a pair, (4.17), can be easily evaluated using (4.20)-(4.23). The composite likelihood for the entire family j is:

$$\mathbf{CL}_j(\boldsymbol{\theta}) = \sum_{i=1}^{N_j} \sum_{k \neq i} \mathbf{L}_{ikts}(\boldsymbol{\theta}) \tag{4.24}$$

The composite likelihood estimator maximizes

$$\mathbf{CL}(\boldsymbol{\theta}) = \sum_{j=1}^{F} \mathbf{CL}_j(\boldsymbol{\theta}). \tag{4.25}$$

Since gradient calculation of this criteria may be very cumbersome, we again use the

simplex algiorithm of Nelder and Mead.

While this estimator is most naturally written as in (4.25), a different representation should be used to avoid calculating the same integrals multiple times. Pairs of responses are related to the parameters by the response values, the time lag between the pair of measurements, the ages of the two subjects, and whether or not the pair of observations are taken on the same subject. The are a fixed number, say k, unique combinations of these factors in a given data set. Letting N_q, L_q denote the number of pairs of configuration q, and the corresponding likelihood for pair q, respectively. It follows that

$$\mathbf{CL}(\boldsymbol{\theta}) = \sum_{q=1}^{k} L_q \cdot N_q \tag{4.26}$$

Using this approach can potentially save a significant amount of computing time if there are any pairs that occur many times.

In practice we have found that, with realistic sample sizes, binary data does not support joint estimation of $\psi_1^\star$ and $\psi_2^\star$ using (4.25). Even under data conditions found to be quite advantageous in the continuous case, the complete lack of estimation precision when both functions are in the model necessitates a simplification by omitting one of the two. In principal, one may do some sort of model selection to determine which function leads to a model which fits better. However, to maintain the spirit of a longitudinal family model with which allows for correlation modulated by multiple timing variables, we complete all analysis with only ψ_1 in the model going forward.

To show that this estimator is reasonable we demonstrate its performance on some simulated data. A complex covariance structure of η_1 has proven to be difficult to characterize, thus we simply use the squared exponential covariance. To demonstrate estimation precision, a sample size of $F = 200$ families, with 3 people per family and 5 observations per subject was used. Initial ages were generated as Uniform$(5, 10)$ and observation times were sampled uniformly on $(0, 10)$. The estimate of the auto-

correlation parameter, truly .9, was .899 on average with a standard error of .055. We can see in Figure 4.1 that the functional parameters are recovered very accurately.

4.4 Model Inference

Much like the continuous case, most of the relevant testing problems will involve nested hypotheses. As mentioned in Chapter 2, the LRT has been extended to the case of CL estimators [Varin and Vidoni (2005), Cox and Reid (2004)]. We will do a small simulation study to investigate the level of the test using the nominal critical value. The two primary hypotheses we will test involve concluding that a particular function is either constant or identically zero.

To investigate the power for testing $H_0 : \psi_k \equiv c$ we generate data from two populations. In both cases $\xi(a) = \sqrt{.8}$, $\varphi(a) = \sqrt{.33}$; population 1 for this experiment will be $\psi_1(a) = -.04 + 2a$ and $\psi_1(a) = -.08 + 4a$ is population 2. For determining the significance level, we conduct simulations with $\psi_1(a) = .2, .35$ in the two populations, respectively, as this is the best constant approximation to the functions used in the power simulations. The corresponding functions $\psi_1^\star, \varphi_1^\star, \xi_1^\star$ are ultimately used by applying (4.5), (4.7), (4.8). Recall that a is scaled by 100 for greater numerical stability. The results of this simulation are shown in the left half of Table 4.1. In each case the nominal critical value of 3.84 proved to be too small. The "corrected" rows are those where the .95 quantiles of the simulations run under the null hypothesis were used as critical values. These quantiles were 5.94 and 7.49 in Populations 1 and 2, respectively.

To examine the power and significance level using the nominal likelihood ratio critical value for testing $H_0 : \psi_k \equiv 0$ we use the same population parameter values as above. When assessing significance level, $\psi_1(a)$ was fixed at 0. In this case the significance levels using the nominal critical value of 7.81 were not inflated as much. The .95 quantiles of the null simulation test statistics were 9.14 and 8.26, respectively.
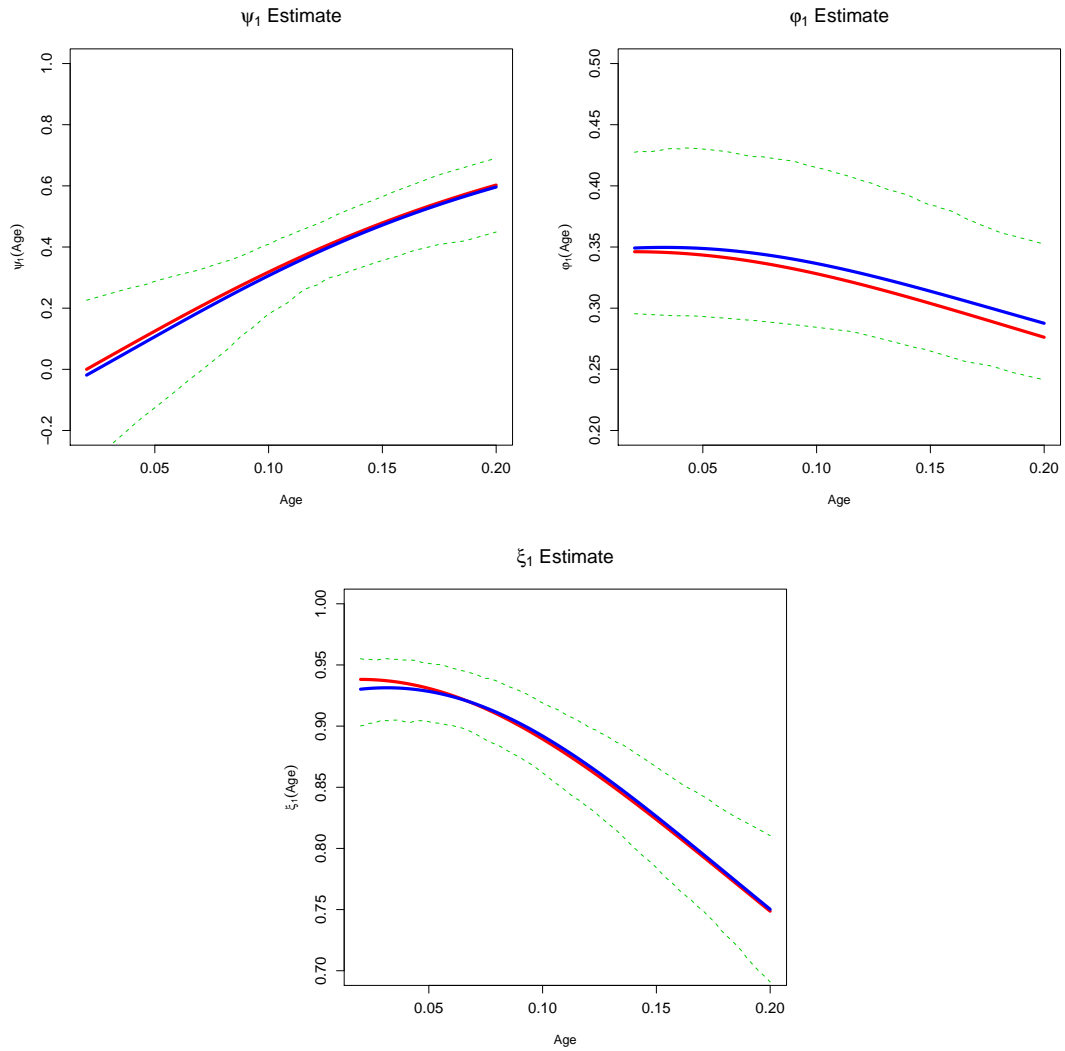
Figure 4.1: Estimates of $\psi_1^\star$, $\varphi^\star$, and $\xi^\star$. True population structure is the solid blue line, estimated is the solid red line, with empirical error bars in light green

	Testing $\psi_1 \equiv c$		Testing $\psi_1 \equiv 0$	
Population	Level	Power	Level	Power
1 - Nominal	.116	.426	0.070	0.607
2 - Nominal	.176	.983	0.056	1.000
1 - Corrected	.05	.353	.05	0.641
2 - Corrected	.05	.921	.05	1.000

Table 4.1: Monte Carlo estimates of the level and power of the likelihood ratio test for CL estimators for testing $\psi_1 \equiv c$ (left) and $\psi_1 \equiv 0$ (right) under two population structures.

The results of this study can be found on the right side of Table 4.1. For both simulations 500 replications were used.

Unfortunately, the implications of this inquiry is that tests of the constancy of ψ_1 are subject to potentially high Type I error rates. Therefore, when using the LRT has a decision rule caution should be exercised. The empirically corrected power estimates are still impressive, but this is not useful in practice. Instead something more sophisticated, like a parametric bootstrap could be useful, but will be very costly since a single model is relatively complex to fit.

4.5   Analysis of the CBCL item data

For our real data application we revisit the Child Behavioral Checklist (CBCL) data analyzed in Chapter 2. Recall that the CBCL is one of the most widely used child behavioral assessment tools. The CBCL data is far less sparse than the DOTS-R data examined in Chapter 3. The average number of assessments in the CBCL is 3.19, with nearly all children being seen at least twice. The names for the CBCL items are given in Table 4.2. Notice that most of these items are negative behaviors and all fall into one of a few broad categories: aggression/rule breaking(1-20) and anxiety/depression(21-34).

Our goal is to glean information on patterns of association between children on

1	Argues	Bragging	Cruelty, Bullying
4	Demands attention	Destroys own things	Destroys others' things
7	Disobedient at home	Disobedient at school	Easily jealous
10	Gets in fights	Physically attacks people	Screams
13	Showing off	Stubborn	Sudden changes in mood
16	Talks too much	Teases a lot	Temper tantrums
19	Threatens	Unusually loud	Complains of loneliness
22	Cries a lot	Fears someone might do something bad	Has to be perfect
25	Feels unloved	Others out to get him/her	Feels worthless
28	Nervous	Too fearful	Feels too guilty
31	Self conscious	Suspicious	Unhappy
34	Worrying		

Table 4.2: CBCL Item list

these items and thus provide evidence for the coherence of such behaviors in children at risk for substance abuse. In addition, we hope to see whether this pattern of resemblance accumulates or diminishes as a child ages. As discussed briefly in chapter 2, we expect to see siblings correlations in these data partially because the same person (the mother in these data) is rating all of the children on each item. Therefore, idiosyncracies of the mother may artificially inflate the apparent level of sibling correlation and should be considered when interpreting these results. On the other hand, we still do expect children to be correlated on several of these measures. For example, if one child demands attention frequently, this may cause his/her sibling to react by doing the same.

To assess age-varying sensitivity to family environment we fit a model to each of the CBCL items with a slope in ψ_1, a constant model in ψ_1 and a model without ψ_1 altogether. In each model φ, ξ were both parameterized as linear functions. Converged negative log composite likelihood values for each of these three items on all 34 items are given in Table 4.3. Although inference is ambiguous, since we do not know the exact level of the test, extremely large LRT statistics (e.g. > 12) is probably a conservative cut-off for concluding ψ_1 is time-varying.

Regarding testing of $\psi_1 \equiv 0$, 32 of 34 items have a sizeable likelihood ratio test statistic far beyond any doubtful range when testing the hypothesis of no familiality,

Item	$\psi_1(a) = \beta_0 + \beta_1 a$	$\psi_1(a) \equiv c$	$\psi_1(a) \equiv 0$	
1	4397.639	4397.986	4414.131	⋆
2	5630.328	5631.410	5650.201	⋆
3	4354.742	4355.412	4365.127	⋆
4	5574.057	5580.182	5617.201	⋆⋆
5	3104.876	3110.491	3138.735	⋆⋆
6	3622.950	3625.650	3661.318	⋆
7	5486.420	5487.625	5534.424	⋆
8	3565.174	3565.512	3572.171	⋆
9	5174.638	5174.727	5204.292	⋆
10	2690.275	2690.711	2702.153	⋆
11	3031.331	3031.370	3064.412	⋆
12	4174.818	4176.180	4192.301	⋆
13	5496.243	5506.661	5538.583	⋆⋆
14	5481.630	5482.142	5518.576	⋆
15	4898.249	4901.506	4912.151	⋆
16	5335.094	5336.081	5345.226	⋆
17	5355.907	5357.167	5373.608	⋆
18	5498.980	5499.235	5537.019	⋆
19	2427.701	2429.099	2446.808	⋆
20	4293.323	4293.735	4312.593	⋆
21	3640.154	3640.734	3651.661	⋆
22	3860.002	3860.931	3886.308	⋆
23	2886.643	2889.092	2894.181	⋆
24	5349.938	5355.752	5371.755	⋆⋆
25	4283.602	4285.501	4291.306	⋆
26	2583.137	2583.559	2593.117	⋆
27	3584.259	3585.081	3596.576	⋆
28	3718.568	3719.631	3733.591	⋆
29	2690.313	2690.755	2693.563	
30	1521.967	1522.237	1530.890	⋆
31	5583.594	5584.019	5611.391	⋆
32	1702.855	1703.119	1710.150	⋆
33	3698.175	3968.502	3701.327	
34	4975.894	4979.847	4999.008	⋆

Table 4.3: Negative log composite likelihood values for models fit with ψ_1 linear (left), constant (middle) and 0 (right). ⋆ in the margin indicates strong evidence for non-zero familial correlation; ⋆⋆ indicates strong evidence for age-varying sensitivity to family environment.

many of which are > 30. As mentioned before, this is not a surprise since all children are being rated by the same person. Four items, coming both from the aggression and anxiety domains, give suggestive evidence for non-constant sensitivity to familial environment.

In Figure 4.2 we have "Demands attention" and "Has to be perfect", two items with significantly time-varying familial sensitivity. The estimation autocorrelation parameter for each of these items was approximately .971 and .937, respectively, indicating that families whose environments exacerbate these symptoms tend to remain so over time. It appears that the tendency for siblings to resemble each other, with regard to demanding attention, decreases with age. The makes sense since children tend to become more independent and are less likely to be competing for a parent's attention as they age, making it less likely that they are both demanding attention. On the other hand, it appears that siblings tend to become more similar as they age in their penchants for feeling the need to be perfect. Perhaps this is because, if one has a sibling with this tendency, then a need to compete with them induces the same behavior in them, and this realization is not made until an older age.

4.6   Discussion and Future Work

While the literature on longitudinal family data is relatively sparse, that on binary longitudinal family data is practically non-existent. In this chapter, we have partially extended the methods of Chapter 3 to the case of binary data. While the disaggregation of superimposed processes of interest is not a theme in this chapter, we still have presented a model with a nice interpretation that reflects a previously unexplored pattern of coherence between family members in longitudinal data, binary or otherwise. It is of interest to determine how to estimate ψ_1, ψ_2 concurrently; perhaps larger sample sizes than those used in simulations are required to do so. If it is the case that an unrealistic sample size is required for the model of chapter 3 to
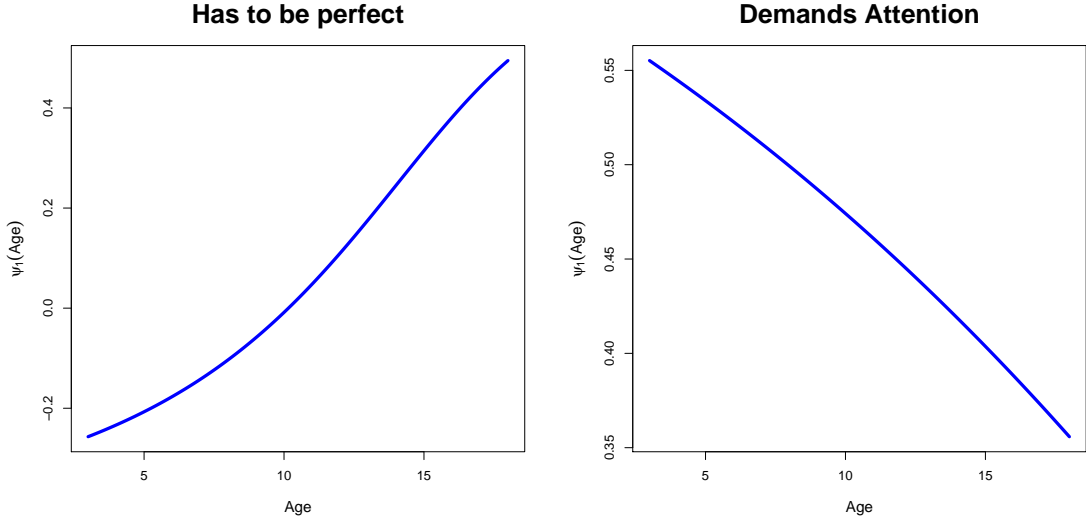
Figure 4.2: Estimated ψ_1 trajectories for two items, "Has to be perfect" and "Demands attention", which display evidence for non-constant sensitivity to familial environment.

work, that would also be an interesting finding and is a subject of future work.

Since maximum likelihood estimation is not feasible in this setting, we described a method based on summing over all pairwise likelihoods, appealing to the popular method of composite likelihood. The CL estimator provides a significant improvement over ML in terms of computational efficiency, and displayed a reasonable ability to recover the data generating population structure. Future work on determining a reliable inference technique in this setting is crucial. Even information based confidence intervals are difficult in this model, since the fitted parameters are actually quotients of a number of other functions in the model.

The extension of this model to more general categorical data is straightforward, if the observed response are ordinal and can reasonably be viewed as a thresholded continuous variable. For example, consider a pair of three level categorical responses with underlying continuous thresholds δ_1, δ_2. Then the region of integration in (4.17) would now be $R_{ik} = R_i \times R_k$ where

$$R_i = (-\infty, \delta_1) \qquad \text{if } Y_{ijt} = 1$$

$$R_i = (\delta_1, \delta_2) \qquad \text{if } Y_{ijt} = 2$$

$$R_i = (\delta_2, \infty) \qquad \text{if } Y_{ijt} = 3, \qquad (4.27)$$

and similarly for R_k. The corresponding table probabilities can be calculated from this expression and the composite likelihood given as in (4.25). Estimation in such models, as well as determining how much additional information you get about the underlying continuous variable from, say, a 3-level ordinal variable vs. a binary variable is an interesting subject of future work.

CHAPTER V

Conclusion

The main contribution of this work is to the study of clustered longitudinal data when the dynamic association structure is of primary interest. This dissertation focused on clustering brought about either by measurement of multiple variables on a single individual, or by single observations made on subjects grouped in some way (e.g. families). These two scenarios are similar in that at any single point in time one can characterize the associations within a cluster. However, the way the transition is made to modeling the dynamically changing coherence structure is very different.

In the case of multiple measurements taken on a single individual at each time point, which was the subject of chapter 2, there is only a single timing variable, t. Our approaches to modeling the dynamic association structure in this case may be broadly thought of as "regressing" a univariate measure of association on t, and examining the estimated function. The first approach we outlined for this was with bivariate binary trajectories, where the log odds ratio was modeled as a smooth function of time using penalized maximum likelihood. This represents a novel semi-parameteric method of directly characterizing the change in association between the two items over time. When the observed variables represent measures of negative affectivity, which they did on our application, this fitted function quantifies the level and pattern of comorbidity more directly than other existing methods.

The second method designed for the purpose of analyzing the first type of clus-

tering mentioned above offers a means of handling more than two variables simultaneously. This approach utilizes conditional and composite likelihood in a clever way so as to eliminate the need to model means or dependencies of secondary interest. Specifically, by analyzing an altered data set composed of within-subject cross products and conditioning on the marginal totals of each variable we get a model free of nuisance parameters, leaving only the pairwise log odds ratio functions to be estimated. Each of these two methods provide a readily interpretable and meaningful description of the dynamic association structure within an individual.

In the second class of problems we look to characterize the associations between family members and their temporal patterns. In this case, we cannot simply take a univariate measure and effectively regress it against time to characterize the longitudinal association structure. This is because longitudinal family data inherently is modulated by multiple timing variables; even cross sectional measurement of family members is subjected to a timing variable– age. Members of a family encounter events at particular chronological times and are affected by developmental processes which depend on age. Therefore, to properly understand the dynamic association structure it is necessary to consider causes modulated by both age and time. All existing work on similar problems only accounts for a single timing index (typically age) in the analysis. In chapter 3 we present an unprecedented methodology which accounts for multiple timing variables in longitudinal family data. Our approach proves to be a challenging endeavor since many data settings correspond to situations where there is not enough difference between age and time gaps to disaggregate the two phenomena.

The supposition underlying the family data problem is that there are some latent time and age indexed contributions to familial correlation, called "environment" for lack of a better word. Members of the family are exposed to these environments and the amount to which they are affected by them are functions of age. Age-specific sensitivity to events occurring in chronological time represents a new approach

to conceptualizing the longitudinal association structure in a family. The principal was also applied to binary data in chapter 4, although binary data does not appear informative enough to support the disaggregation that is possible with continuous data, even under favorable data conditions. In this case, we describe an interesting sub-model that can be readily fit to binary data using composite likelihood, and still addresses the issue of multiple timing variables modulating associations.

In this dissertation we considered a kind of nested clustering– clustered longitudinal data. Supposing there were multiple variables measured on each member of the family, we would be presented with a further level of nesting. An interesting subject of future work would be to extend and combine these models in some sense to study the change between individual, within individual and over time in the association structure.

APPENDICES

APPENDIX A

Gradient of bivariate binary likelihood with respect to $\tilde{\Theta}$

The mapping from Θ to $\tilde{\Theta}$ has Jacobian matrix

$$J = \begin{pmatrix} J_{11} & J_{12} & J_{13} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{A.1}$$

where by differentiation of 2.12 we have

$$
\begin{aligned}
J_{11} &= 1/P_{11}(t;X) + 1/(1 + P_{11}(t;X) - M_1(t;X) - M_2(t;X)) + \\
&\quad 1/(M_1(t;X) - P_{11}(t;X)) + 1/(M_2(t;X) - P_{11}(t;X)) \tag{A.2} \\
J_{12} &= -1/(M_1(t;X) - P_{11}(t;X)) - 1/(1 + P_{11}(t;X) \\
&\quad - M_1(t;X) - M_2(t;X)) \tag{A.3} \\
J_{13} &= -1/(1 + P_{11}(t;X) - M_1(t;X) \\
&\quad - M_2(t;X)) - 1/(M_2(t;X) - P_{11}(t;X)). \tag{A.4}
\end{aligned}
$$

The Jacobian of the the reverse map from $\tilde{\Theta}$ to Θ is given by

$$J^{-1} = \begin{pmatrix} -1/J_{11} & -J_{12}/J_{11} & -J_{13}/J_{11} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{A.5}$$

The components of the score function $\partial \mathbf{L}/\partial \tilde{\Theta}$ can all be easily obtained using the chain rule:

$$\partial \mathbf{L}/\partial \tilde{\Theta} = \sum_u \partial \Theta_u/\partial \tilde{\Theta} \cdot \partial \mathbf{L}/\partial \Theta_u. \tag{A.6}$$

For example, for θ_t, the intercept in the log odds ratio trajectory (2.9), we first get

$$
\begin{aligned}
\partial \mathbf{L}/\partial \theta_t &= \partial \mathbf{L}/\partial \mathrm{LOR}(t;X) \cdot \partial \mathrm{LOR}(t;X)/\partial \theta_t + \partial \mathbf{L}/\partial M_1(t;X) \cdot \partial M_1(t;X)/\partial \theta_t + \\
& \quad \partial \mathbf{L}/\partial M_2(t;X) \cdot \partial M_2(t;X)/\partial \theta_t \\
&= \partial \mathbf{L}/\partial \mathrm{LOR}(t;X) \cdot \partial \mathrm{LOR}(t;X)/\partial \theta_t \tag{A.7}
\end{aligned}
$$

by applying the chain rule $\partial \mathbf{L}/\partial \theta_t = \sum_u \partial \Theta_u/\partial \theta_t \cdot \partial \mathbf{L}/\partial \Theta_u$ and using the structure of (2.7)-(2.9). Next we would use the chain rule to get

$$
\begin{aligned}
\partial \mathbf{L}/\partial \mathrm{LOR}(t;X) &= \partial \mathbf{L}/\partial P_{11}(t;X) \cdot \partial P_{11}(t;X)/\partial \mathrm{LOR}(t;X) + \\
& \quad \partial \mathbf{L}/\partial M_1(t;X) \cdot \partial M_1(t;X)/\partial \mathrm{LOR}(t;X) + \\
& \quad \partial \mathbf{L}/\partial M_2(t;X) \cdot \partial M_2(t;X)/\partial \mathrm{LOR}(t;X) \\
&= \partial \mathbf{L}/\partial P_{11}(t;X) \cdot \partial P_{11}(t;X)/\partial \mathrm{LOR}(t;X) \tag{A.8}
\end{aligned}
$$

where $\partial M_1(t;X)/\partial \mathrm{LOR}(t;X)$ and $\partial M_2(t;X)/\partial \mathrm{LOR}(t;X)$ vanish as they are the $2, 1$ and $3, 1$ entries of the Jacobian. Finally, the factor $\partial \mathbf{L}/\partial P_{11}(t;X)$ in (2.8) is obtained

by directly differentiating $\mathbf{L}$ with respect to $P_{11}(t; X)$ as a free variable:

$$
\begin{aligned}
d\mathbf{L}/dP_{11}(t; X) \;=\; & \sum_{ijk} Y_{it11}/P_{11}(t; X_{it}) - Y_{it10}/(M_1(t; X_{it}) - P_{11}(t; X_{it})) - \\
& Y_{it01}/(M_2(t; X_{it}) - P_{11}(t; X_{it})) + \\
& Y_{it00}/(1 + P_{11}(t; X_{it}) - M_1(t; X_{it}) - M_2(t; X_{it})). \quad \text{(A.9)}
\end{aligned}
$$

The factor $\partial P_{11}(t; X)/\partial \mathrm{LOR}(t; X)$ in (A.8) is the $1,1$ entry of the Jacobian.

APPENDIX B

Asymptotic pointwise confidence intervals for parameters in the continuous longitudinal family data model

Let $\hat{\boldsymbol{\theta}}$ be the estimated parameter value based on the converged optimization algorithm, while $\boldsymbol{\theta}^\star$ denotes the true parameter value. Since we are doing maximum likelihood estimation we know that the properly scaled and shifted parameter estimates are asymptotically normal with covariance equal to the inverse fisher information evaluated at the true parameter value, $\mathbf{I}(\boldsymbol{\theta}^\star)$. We know $\mathbf{I}(\boldsymbol{\theta}^\star)$ is defined as the variance of the score function, therefore

$$\mathbf{I}(\hat{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{j=1}^{N} \nabla \mathbf{L}_j(\hat{\boldsymbol{\theta}})' \nabla \mathbf{L}_j(\hat{\boldsymbol{\theta}}) \tag{B.1}$$

is a consistent estimator. That is, the mean of the gradient outer product at the converged point. Although the gradient is too computationally cumbersome to be used to optimization, the expression above only requires a single evaluation. To begin, we need to differentiate $\mathbf{L}_j$ with respect to the mean vector $\boldsymbol{\mu}_j(\boldsymbol{\theta})$ and the covariance matrix $\boldsymbol{\Sigma}_j(\boldsymbol{\theta})$:

$$\frac{\partial \mathbf{L}_j}{\partial \boldsymbol{\mu}_j(\boldsymbol{\theta})} = -\frac{1}{2} \left(\frac{\partial \left(\boldsymbol{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}) \right)' \boldsymbol{\Sigma}_j(\boldsymbol{\theta})^{-1} \left(\boldsymbol{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}) \right)}{\partial \boldsymbol{\mu}_j(\boldsymbol{\theta})} \right)$$

$$= -\frac{1}{2} \left(-2 \boldsymbol{\Sigma}_j(\boldsymbol{\theta})^{-1} \left(\boldsymbol{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}) \right) \right)$$

$$= \boldsymbol{\Sigma}_j(\boldsymbol{\theta})^{-1} \left(\boldsymbol{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}) \right) \tag{B.2}$$

and

$$\frac{\partial \mathbf{L}_j}{\partial \boldsymbol{\Sigma}_j(\boldsymbol{\theta})} = -\frac{1}{2} \left(\frac{\partial \log(|\boldsymbol{\Sigma}_j(\boldsymbol{\theta})|)}{\partial \boldsymbol{\Sigma}_j(\boldsymbol{\theta})} + \frac{\partial \left(\boldsymbol{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}) \right)' \boldsymbol{\Sigma}_j(\boldsymbol{\theta})^{-1} \left(\boldsymbol{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}) \right)}{\partial \boldsymbol{\Sigma}_j(\boldsymbol{\theta})} \right)$$

$$= -\frac{1}{2} \left(\boldsymbol{\Sigma}_j(\boldsymbol{\theta})^{-1} - \boldsymbol{\Sigma}_j(\boldsymbol{\theta})^{-1} \left(\boldsymbol{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}) \right) \left(\boldsymbol{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\theta}) \right)' \boldsymbol{\Sigma}_j(\boldsymbol{\theta})^{-1} \right). \tag{B.3}$$

Then by using the chain rule the gradient with respect to the covariance driving parameters in $\boldsymbol{\theta}$ can be acquired through

$$\frac{\partial \mathbf{L}_j}{\partial \boldsymbol{\theta}_k} = \text{Tr} \left(\left(\frac{\partial \mathbf{L}_j}{\partial \boldsymbol{\Sigma}_j(\boldsymbol{\theta})} \right)' \mathbf{K}_{\boldsymbol{\theta}} \right) \tag{B.4}$$

where $\mathbf{K}_{\boldsymbol{\theta}} = \frac{\partial \boldsymbol{\Sigma}_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k}$. The elements of $\mathbf{K}_{\boldsymbol{\theta}}$ depend on which functional parameter $\boldsymbol{\theta}_k$ is part of. If $\boldsymbol{\theta}_k$ is underlying ψ_1, then

$$[\mathbf{K}_{\boldsymbol{\theta}}]_{n,m} = 2 r_1(t_n, t_m) \cdot \psi_1(a) \cdot \frac{\partial \psi_1(a)}{\partial \boldsymbol{\theta}_k}. \tag{B.5}$$

Similarly if $\boldsymbol{\theta}_k$ is underlying ψ_2, φ, ξ, then

$$[\mathbf{K}_{\boldsymbol{\theta}}]_{n,m} = 2 r_2(a_n, a_m) \cdot \psi_2(a) \cdot \frac{\partial \psi_2(a)}{\partial \boldsymbol{\theta}_k} \tag{B.6}$$

$$[\mathbf{K}_{\boldsymbol{\theta}}]_{n,m} = 2 \mathcal{I}(i_n = i_m) \cdot \varphi(a) \cdot \frac{\partial \varphi(a)}{\partial \boldsymbol{\theta}_k} \tag{B.7}$$

$$[\mathbf{K_\theta}]_{n,m} = 2\mathcal{I}(n=m) \cdot \xi(a) \cdot \frac{\partial \xi(a)}{\partial \boldsymbol{\theta}_k} \tag{B.8}$$

is the form of $\mathbf{K_\theta}$, respectively. If $\boldsymbol{\theta}_k$ is one of the parameters underlying $r_1(\cdot, \cdot)$ or $r_2(\cdot, \cdot)$ then the structure of $\mathbf{K_\theta}$ is

$$[\mathbf{K_\theta}]_{n,m} = \psi_1(a_n)\psi_1(a_m)\frac{\partial r_1(t_n, t_m)}{\partial \boldsymbol{\theta}_k} \tag{B.9}$$

or

$$[\mathbf{K_\theta}]_{n,m} = \psi_2(a_n)\psi_2(a_m)\frac{\partial r_2(a_n, a_m)}{\partial \boldsymbol{\theta}_k} \tag{B.10}$$

respectively. The score function with respect to $\boldsymbol{\theta}_k$ is then obtained by combining (B.4) with the proper expression for $\mathbf{K_\theta}$, described by (B.5)-(B.10). Similarly, the score with respect a mean driving parameter, $\boldsymbol{\theta}_d$, can be obtained through

$$\frac{\partial \mathbf{L}_j}{\partial \boldsymbol{\theta}_d} = \mathrm{Tr}\left(\left(\frac{\partial \mathbf{L}_j}{\partial \boldsymbol{\mu}_j(\boldsymbol{\theta})}\right)' \mathbf{D_\theta}\right) \tag{B.11}$$

where

$$[\mathbf{D_\theta}]_{n,m} = \frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{\theta}_d}. \tag{B.12}$$

Using the above identities coupled with (B.2) to evaluate the score function at the converged point yields a consistent estimator of the fisher information, and thus a consistent estimator of the covariance matrix of the parameter estimates in $\mathbf{I}(\hat{\boldsymbol{\theta}})$.

APPENDIX C

Properties of the unconditional covariance of η_j

To prove the unconditional covariance formula, suppose $Z(t)$ has mean zero $\forall t$ and that

$$E\left(Z(t)Z(s)|\delta\right) = \exp(-\delta|t-s|^\kappa) \tag{C.1}$$

First it is convenient to define $\tau = e^{-\delta}$ and rewrite the conditional covariance as

$$E\left(Z(t)Z(s)|\tau\right) = \exp(\log(\tau)|t-s|^\kappa) = \tau^{|t-s|^\kappa} \tag{C.2}$$

By smoothing, the unconditional covariance, $E\left(Z(t)Z(s)\right)$, is nothing but an expectation against the distribution of τ, which is assumed to be beta distribution with parameters α_1, α_2. Thus,

$$E\left(Z(t)Z(s)\right) = E_\tau\left(E\left(Z(t)Z(s)|\tau\right)\right) = E_\tau\left(\tau^{|t-s|^\kappa}\right)$$

$$= \frac{\Gamma(\alpha_1+\alpha_2+|t-s|^\kappa)\Gamma(\alpha_1+|t-s|^\kappa)}{\Gamma(\alpha_1+\alpha_2+|t-s|^\kappa)\Gamma(\alpha_1+|t-s|^\kappa)}\int_0^1 \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\tau^{|t-s|^\kappa+\alpha_1-1}(1-\tau)^{\alpha_2-1}d\tau$$

$$= \frac{\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_1+|t-s|^\kappa)}{\Gamma(\alpha_1+\alpha_2+|t-s|^\kappa)\Gamma(\alpha_1)}\int_0^1 \frac{\Gamma(\alpha_1+\alpha_2+|t-s|^\kappa)}{\Gamma(\alpha_1+|t-s|^\kappa)\Gamma(\alpha_2)}\tau^{|t-s|^\kappa+\alpha_1-1}(1-\tau)^{\alpha_2-1}d\tau$$

$$= \frac{\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_1+|t-s|^\kappa)}{\Gamma(\alpha_1+\alpha_2+|t-s|^\kappa)\Gamma(\alpha_1)} \tag{C.3}$$

The last line follows because the integral in the previous line is that of a Beta($\alpha_1 + |t-s|^\kappa, \alpha_2$) density, and the identity is proven.

To prove that the exchangeable correlation is a limiting case, we first prove that For any positive constant k,

$$\lim_{z\downarrow 0}\frac{\Gamma(kz)}{\Gamma(z)} = \frac{1}{k} \tag{C.4}$$

It is convenient to use the following representation of $\Gamma(z)$, which is valid for any $z > 0$:

$$\Gamma(z) = \frac{1}{z}\prod_{n=1}^{\infty}\frac{\left(1+\frac{1}{n}\right)^z}{1+\frac{z}{n}} \tag{C.5}$$

We can use this to represent the ratio $\Gamma(kz)/\Gamma(z)$. After some algebra, we have

$$\frac{\Gamma(kz)}{\Gamma(z)} = \frac{1}{k}\prod_{n=1}^{\infty}\left(1+\frac{1}{n}\right)^{z(k-1)}\frac{\left(1+\frac{z}{n}\right)}{\left(1+\frac{kz}{n}\right)}. \tag{C.6}$$

For any n

$$\lim_{z\downarrow 0}\frac{\left(1+\frac{z}{n}\right)}{\left(1+\frac{kz}{n}\right)} = 1 \tag{C.7}$$

Since both the numerator and denominator are continuous and bounded for $z > 0$

and each converge to 1. Similarly, for any n,

$$\lim_{z \downarrow 0} \left(1 + \frac{1}{n}\right)^{z(k-1)} = 1 \tag{C.8}$$

So,

$$\lim_{z \downarrow 0} \prod_{n=1}^{\infty} \left(1 + \frac{1}{n}\right)^{z(k-1)} \cdot \frac{(1 + \frac{z}{n})}{(1 + \frac{kz}{n})} = 1 \tag{C.9}$$

which proves the result. Now, by the constraint that $\rho = \alpha_1/(\alpha_1 + \alpha_2)$, it follows that $\alpha_1 + \alpha_2 = \alpha_1/\rho$. Substituting, the unconditional covariance can be written only as a function of α_1.

$$E\left(Z(t)Z(s)\right) = \frac{\Gamma(\alpha_1/\rho)\Gamma(\alpha_1 + |t-s|^\kappa)}{\Gamma(\alpha_1/\rho + |t-s|^\kappa)\Gamma(\alpha_1)} = \frac{\Gamma(\alpha_1/\rho)}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_1 + |t-s|^\kappa)}{\Gamma(\alpha_1/\rho + |t-s|^\kappa)} \tag{C.10}$$

Clearly,

$$\lim_{\alpha_1 \downarrow 0} \frac{\Gamma(\alpha_1 + |t-s|^\kappa)}{\Gamma(\alpha_1/\rho + |t-s|^\kappa)} = 1 \tag{C.11}$$

So,

$$\lim_{\alpha_1 \downarrow 0} E(Z(t)Z(s)) = \lim_{\alpha_1 \downarrow 0} \frac{\Gamma(\alpha_1/\rho)}{\Gamma(\alpha_1)} \tag{C.12}$$

Using (C.4) with $k = 1/\rho$ and $z = \alpha_1$, the result is proven.

To prove the final property we again use the constraint that $\alpha_2 = (1 - \rho)\alpha_1/\rho$, and the variance of τ as a function of α_1 can be written as

$$v(\alpha_1) = \frac{\rho(1 - \rho)}{\alpha_1 + \rho} \tag{C.13}$$

Clearly, $\lim_{\alpha_1 \to \infty} v(\alpha_1) = 0$, at which point the distribution of τ has mean ρ and no variance. Thus, $\tau = \rho$ with probability 1, and the result is proven.

BIBLIOGRAPHY

BIBLIOGRAPHY

Achenbach, T. and Edelbrock, C. (1983). *Manual for Child Behavior Checklist and Revised Child Behavior Profile*. Burlington: University of Vermont, Department of Psychiatry.

Almasy, L., Amos, C. I., Bailey-Wilson, J. E., Cantor, R. M., Jaquish, C. E., Martinez, M., Neuman, R. J., Olson, J. M., Palmer, L. J., Rich, S. S., Spencer, M. A., and MacCluer, J. W. (2003). Genetic analysis workshop 13: Analysis of longitudinal family data for complex diseases and related risk factors. *BMC Genetics*, 4:S1.

Bentler, P. M. and Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annu. Rev. Psychol.*, 47:563–592.

Besag, J. (1974). Spatial interactions and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B*, 36:192–236.

Bock, R. (1960). Components of variance analysis as a structural and discriminal analysis for psychological tests. *British Journal of Statistical Psychology*, 13:151–163.

Bock, R. and Bargmann, R. (1966). Analysis of covariance structures. *Psychometrika*, 31:507–533.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, 26(2):211–252.

Browne, M. W. (1984). Asymptotically distribution free methods for the analysis of covariance structures. *British Journal of Mathematical Statistics in Psychology*, 37:62–83.

Burnham, K. and Anderson, D. (1998). *Model selection and inference: a practical information-theoretic approach*. Springer-Verlag, New York.

Carey, V., Zeger, S. L., and Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526.

Corte, C. and Zucker, R. A. (2008). Self-concept disturbances: Cognitive vulnerability for early drinking and early drunkenness in adolescents at high risk for alcohol problems. *Addictive Behaviors*, 33(10):1282–1290.

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729:737.

de Leeuw, J. (2006). Principal components analysis of binary data by iterated singular value decomposition. *Computational statistics and data analysis*, 50:21–39.

Falconer, D. S. (1981). *Introduction to Quantitative Genetics, 2nd Ed.* Longman, London.

Fisher, R. A. (1919). The genesis of twins. *Genetics*, 4:489–499.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80:141–151.

Gao, F., Wahba, G., Klein, R., and Klein, B. (2001). Smoothing spline anova for multivariate bernoulli observations, with application to ophthalmology data. *Journal of the American Statistical Association*, 96:127–147.

Gauderman, W. J. and Conti, D. V. (2005). Commentary: Models for longitudinal family data. *International Journal of Epidemiology*, 34:1077–1079.

Gauderman, W. J., Macgregor, S., Briollais, L., Scurrah, K., Tobin, M., Park, T., Wang, D., Rao, S., John, S., and Bull, S. (2003). Longitudinal data analysis in pedigree studies. *Genetic Epidemiology*, 25:S18–28.

Gauderman, W. J. and Morrison, J. L. (2000). Evidence for age-specific genetic relative risks in lung cancer. *American Journal of Epidemiology*, 151(1):41–49.

Gilmour, A., Anderson, R., and Rae, A. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika*, 72:593–599.

Guo, W. (2002). Functional mixed effects models. *Biometrics*, 58:121–128.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54:99–108.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Heagerty, P. and Lele, S. (1998). A composite likelihood approach to binary data in space. *Journal of the American Statistical Association*, 93:1099–1111.

Hopper, J., Hannah, M., and Mathews, J. (1984). Genetic analysis workshop ii: Pedigree analysis of a binary trait without assuming an underlying liability. *Genetic Epidemiology*, 1:183–188.

Jaffrezic, F. and Pletcher, S. D. (2000). Statistical models for estimating the genetic basis of repeated measures and other function-valued traits. *Genetics*, 156:913–922.

Jaffrezic, F., Thompson, R., and Pletcher, S. D. (2004). Multivariate character process models for the analysis of two or more correlated function-valued traits. *Genetics*, 168:477–487.

Joreskög, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4):443–477.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association*, 96:1272–1281.

Kempthorne, O. (1969). *An Introduction to Genetic Statistics*. John Wiley.

Kingsolver, J. G., Gomulkiewicz, R., and Carter, P. A. (2001). Variation, selection, and evolution of function valued traits. *Genetica*, 112-113:87–104.

Kirkpatrick, M. and Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite dimensional characters. *Journal of Mathematical Biology*, 27:429–450.

Kirkpatrick, M., Lofsvold, D., and Bulmer, M. (1990). Analysis of inheritance, selection and evolution of growth trajectories. *Genetics*, 124:979–993.

Kirkpatrick, M. and Meyer, K. (2004). Direct estimation of genetic principal components: Simplified analysis of complex phenotypes. *Genetics*, 168:2295–2306.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, 33:133–158.

Laird, N. M. (1991). Topics in likelihood-based methods for longitudinal data analysis. *Statistica Sinica*, 1(1):33–50.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Lande, R. and Arnold, S. J. (1983). The measurement of selection on correlated characters. *Evolution*, 37(6):1210–1226.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society Series B*, 54(1):3–40.

Lindenberger, U., Nagel, I. E., Chicherio, C., Li, S. C., Heekeren, H. R., and Backman, L. (2008). Age-related decline in brain resources modulates genetic effects on cognitive functioning. *Frontiers in Neuroscience*, 2(2):234–244.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:221–239.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, 78(1):153–160.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley Sons.

Macgregor, S., Knott, S. A., White, I., and Visscher, P. M. (2003). Longitudinal variance-components analysis of the framingham heart study data. *BMC Genetics*, 4:S22.

McCulloch, C. E. (2003). *Generalized linear mixed models.* Institute of Mathematical Statistics.

Meyer, K. (1998). Estimating covariance functions for longitudinal data using a random regression model. *Genetics Selection Evolution*, 30:221–240.

Meyer, K. and Hill, W. (1997). Estimation of genetic and phenotypic covariance functions for longutindal or 'repeated' records by restricted maximum likelihood. *Livestock Production Science*, 47:185–200.

Molenberghs, G. and Ritter, L. L. (1996). Methods for analyzing multivariate binary data, with association between outcomes of interest. *Biometrics*, 52:1121–1133.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data.* Springer, New York.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B*, 68(2):179–199.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49:115–132.

Muthén, B., Kaplan, D., and Hollis, M. (1987). On the structural edquation modeling with data that are not completely missing at random. *Psychometrika*, 52:431–462.

Muthén, L. K. and Muthén, B. O. (2004). *Mplus User's Guide.* Muthén and Muthén, Los Angeles, CA.

Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Nunez-Anton, V. (1998). Longitudinal data analysis: non-stationary error structures and antedependence models. *Applied Stochastic Models Data Analysis*, 13:279–287.

Nunez-Anton, V. and Zimmerman, D. (2000). Modeling non-stationary longitudinal data. *Biometrics*, 56:699–705.

Ochi, Y. and Prentice, R. L. (1985). Likelihood inference in a correlated probit regression model. *Biometrika*, 71:531–543.

Pawitan, Y., Reilly, M., Nilsson, E., Cnattingius, S., and Lichtenstein, P. (2005). Estimation of genetic and environmental factors for binary traits using family data. *Statistics in Medicine*, 24:1613–1618.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. on the inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society Series A*, 195:79–150.

Pendergast, J. F., Gange, S. T., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review*, 64(1):89–118.

Pinheiro, J. and Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, 4:11–23.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag.

Pletcher, S. D. and Geyer, C. J. (1999). The genetic analysis of age-dependent traits: Modeling the character process. *Genetics*, 153:825–833.

Pletcher, S. D. and Jaffrezic, F. (2002). Generalized character process models: Estimating the genetic basis of traits that cannot be observed and that change with age or environmental conditions. *Biometrics*, 58:157–162.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44:1033–1048.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47:825–839.

Rabe-Hesketh, S. and Skrondal, A. (2001). Parameterization of multivariate random effects models for categorical data. *Biometrics*, 57:1256–1264.

Ramsey, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer, New York.

Rao, C. R. (1971). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Assocation*, 67(337):112–115.

Rasmussen, C. E. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Riska, B., Atchley, W. R., and Rutledge, J. (1985). A genetic analysis of targeted growth in mice. *Genetics*, 107:79–101.

Satorra, A. and Saris, W. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50:83–90.

Scheiner, S. M. (1993). Genetics and evolution of phenotypic plasticity. *Annual Review of Ecology and Systematics*, 24:35–68.

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Trans. American Mathematical Society*, 44(3):522–536.

Searle, S. R. (1968). Another look at henderson's methods of estimating variance components. *Biometrics*, 24(4):749–787.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Shedden, K. and Zucker, R. A. (2008). Regularized finite mixture models for probability trajectories. *Psychometrika*, 73.

Skrondal, A. and Rabe-Hesketh, S. (2005). *Entry for the Encyclopedia of Statistics in Behavioral Science.* John Wiley.

Soler, J. M. and Blangero, J. (2003). Longitudinal familial analysis of blood pressure involving parametric (co)variance functions. *BMC Genetics*, 4:S87.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40:961–971.

Tatar, M., Promislow, D. E. L., Khazaeli, A. A., and Curtsinger, J. W. (1996). Age-specific patterns of genetic variables in drosophilia melanogaster. ii. fecundity and its genetic covariance with age-specific mortality. *Genetics*, 143:849–858.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92:519–528.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *Annals of Statistics*, 23:1865–1895.

Wang, Y. (1997). Odds ratio estimation in bernoulli smoothing spline anova models. *Statistician*, 48:49–56.

Windle, M. and Lerner, R. M. (1986). Reassessing the dimensions of temperamental individuality across the life span: The revised dimensions of temperament survey. *Journal of Adolescent Research*, 1(2):213–230.

Yang, Q., Chazaro, I., Cui, J., Guo, C.-Y., Demissie, S., Larson, M., Atwood, L. D., Cupples, A., and DeStephano, A. L. (2003). Genetic analyses of longitudinal phenotype data: a comparison of univariate methods and a multivariate approach. *BMC Genetics*, 4:S29.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77:642–648.

Zhao, L. P., Prentice, R. L., and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society Series B*, 54(3):805–811.

Zucker, R. A., Ellis, D. A., Bingham, C. R., Fitzgerald, H. E., and Sanford, K. P. (1996). Other evidence for at least two alcoholisms: Life course variation in antisociality and heterogeneity of alcoholic outcome. *Development and Psychopathology*, 8:831–848.