Study Design for Longitudinal and High Dimensional Measures

by

Meihua Wu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2013

Doctoral Committee:

Assistant Professor Brisa N. Sánchez, Chair
Professor Ana V. Diez-Roux
Professor Trivellore E. Raghunathan
Professor Peter X.K. Song

Many Thanks to Family and Friends.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Brisa N. Sánchez for her truly outstanding guidance on my Ph.D. study. I have benefited so much from her deep insights on how to approach various statistical problems. In addition to biostatistics, I have learned a lot of useful techniques for academic writing and presentation skills from her. Also, I wish to express my great appreciation to Dr. Peter X.K. Song for his invaluable suggestions on my dissertation research, his overwhelming enthusiasm on the statistical problems, and his broad knowledge. I am deeply appreciative of their continuous encouragement, support and help in my Ph.D. study.

I wish to thank Dr. Ana V. Diez-Roux and Dr. Trivellore Eachambadi Raghunathan for serving as members of my dissertation committee and providing me very helpful comments on my dissertation. I am very grateful to Dr. Ana V. Diez-Roux for providing me with the deep insights for the application of my dissertation from the viewpoint of an epidemiologist and sharing me with plenty of data sets for the motivating examples. I also deeply appreciate Dr. Trivellore Eachambadi Raghunathan's constructive comments and criticisms of my dissertation.

I would like to express my gratitude to my fellow students Dr. Jian Kang, Dr. Youna Hu and Dr. Nanhua Zhang. I had numerous discussions with them about my my dissertation and they offered many helpful and sincere suggestions.

Finally, my appreciation also goes to the Department of Biostatistics at the University of Michigan which has provided a strongly supportive environment for study and research. I am grateful to all the staff members, fellow students and friends in

the department for their sincere help and support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Study Design for Longitudinal and High Dimensional Measures

by

Meihua Wu

Chair: Brisa N. Sánchez

Study design is the foundation of successful clinical or epidemiological studies. Ever since the seminal work of Fisher (1935), research in this area has blossomed and many innovative concepts and approaches have been developed. Despite extensive literature on study design, new challenges for study design continue to emerge as innovative technologies push the limits of what can be investigated with a clinical or epidemiological study. For instance, tools for ecological momentary assessment of behaviors or biological markers, or high throughput experiment devices such as microarrays open the opportunity to measure complex biological processes over time, or the expression levels of millions of genetics or proteomics biomarkers simultaneously. In this dissertation, we develop novel design methodologies for studies employing these new data collection techniques, namely: 1) studies involving repeated measures of nonlinear profiles in biomarker studies with the objective of estimating features of the profile; 2) studies involving data with underlying functional response with the objective of capturing the mean profile and between subject variability; 3) studies involving high dimensional genetics and proteomics data with the objective of constructing classifiers with high probability of correct classification. Correspondingly, our research is

motivated by three practical applications: 1) salivary cortisol studies for investigating the association between cardiovascular disease and stress; 2) urinary progesterone studies for reproductive health; 3) studies involving high dimensional genetics and proteomics data with the objective of constructing classifiers with high probability of correct classification. This dissertation contributes novel study design methodologies for studies that involve related but distinct data structures. We demonstrate the use of the methods with various examples to enhance the potential of their used across a variety of settings. The new design methodology will thus enable investigators to better evaluate the feasibility and cost-efficiency of the study in the planning stage and ultimately improve the chance of success of studies involving longitudinal and high dimensional data.

CHAPTER I

Introduction

Study design is the foundation of successful clinical or epidemiological studies. Statistical aspects of design include the determination of the sample size, mode of randomization, the sampling units, the timing (or location) of sample collection, among others. The optimal choice of these design features ultimately depends on the goal of the proposed study. For example, if the purpose of the study is to test a scientific hypothesis, then the optimal design should maximize the power of the test given a predefined threshold for Type I errors. If, instead, the estimation of a quantity describing a biological or clinical mechanism is of interest, the optimal design should provide strategies to minimize the variance or mean squared errors of the estimated quantity. Alternatively if the objective is prediction of future outcomes, the design should enable maximum prediction accuracy. In addition to statistical concerns, other factors will also impact the optimal design. For example, the design should minimize the risk and burden on the subjects, as well as the implementation cost. Furthermore, a good design should also quantify the uncertainty in the assumption of the study scenarios and provide adequate analysis for evaluating the robustness of the design.

The field of experimental design was pioneered by the seminal work of *Fisher* (1935). Ever since then, the research in this area has blossomed and many innovative concepts and approaches have been developed, such as response surface methodology

(*Myers et al.*, 2009), Bayesian design (*Chaloner*, 1995), adaptive design (*Chow and Chang*, 2007), model-based design (*Fedorov and Hackl*, 1997). Comprehensive reviews for many fundamental results and effective applications of the design methodologies can be found in *Cochran et al.* (1992); *Montgomery* (2008).

In spite of the extensive literature on study design, new challenges for study design continue to emerge as innovative technologies are always pushing the limits of what can be investigated with a clinical or epidemiological study. High throughput experiment devices, such as microarrays, open the opportunity to measuring the expression levels of millions of genetics or proteomics biomarkers simultaneously. From these data one may be able to construct accurate and reliable classifiers for predicting the risk or prognosis of various diseases. Likewise with ecological momentary assessment tools, repeated measures of a biomarkers or behaviors can be taken on a subject in real time to more accurately measure behaviors or biological processes in real world contexts (*Shiffman et al.*, 2008). These measurement techniques will often give rise to nonlinear functional profiles often with substantial between-individual variation.

In this dissertation, we develop novel design methodologies for the studies employing these new data collection techniques: 1) studies involving repeated measures of nonlinear profiles; 2) studies involving data with underlying functional response with the objective of capturing the mean profile and between subject variability; 3) studies involving high dimensional genetics and proteomics data with the objective of constructing classifiers with high probability of correct classification. Correspondingly, our research is motivated by three practical applications: 1) salivary cortisol studies for investigating the association between cardiovascular disease and stress; 2) urinary progesterone studies for reproductive health; 3) studies for constructing classifiers of long term survival after kidney transplant. The background and motivations for each of the design problems will be discussed as follows.

1.1 Background and Motivations

1.1.1 Design for Studies Involving Repeated Measures of Nonlinear Profiles

Cardiovascular disease remains one of the leading causes of mortality and morbidity in the US, and an area of health where race-ethnic and socioeconomic disparities persist. In addition to known risk factors, such as overweight, psychosocial stress has been proposed as a cause for cardiovascular disease, and as one of the roots of health disparities (*Kaplan and Keil*, 1993; *Diez Roux et al.*, 2001). However, subjective measures of stress, such as self-reported questionnaires, are susceptible to response bias. Instead, a field-friendly biological marker of stress, salivary cortisol, is now collected in many population-based epidemiological studies to assess the relationship between stress and disease (*Adam and Kumari*, 2009). Although there is great promise in using salivary cortisol as an objective measure of stress, the best statistical methods to properly collect and summarize salivary cortisol data in epidemiological studies, and thus take full advantage of the millions of dollars spent yearly in assay costs alone, remain largely unknown.

In Chapter 2, we will develop statistical methods to better quantify stress response when using salivary cortisol. Salivary cortisol exhibits a non-linear diurnal pattern through the length of the day, so-called stress response. The stress response varies from day to day within a given individual, and also exhibits variation between individuals; that is, it exhibits multiple levels of variability (*Smyth et al.*, 1997). To capture the variability, studies often collect salivary cortisol from multiple days on groups of individuals (*Adam et al.*, 2006; *Cohen et al.*, 2006). However, the number of days and samples per day, as well as the time of day when the samples are collected have so far been chosen in an ad hoc manner. We will discuss optimal design strategies for multi-level sampling of salivary cortisol in population studies that max-

imize the precision of summaries of the stress response. Improving the precision of the summaries will ultimately reduce overall study costs and increase the ability to detect associations between stress and disease.

We will use the Stress sub-study of the Multi-Ethnic Study of Atherosclerosis (MESA Stress) as one of the illustrating examples. MESA Stress is among a handful of large scale epidemiological studies using salivary cortisol. MESA Stress examines the role of stress as a contributor to a range of precursors of cardiovascular disease, and collects data on 1000 participants. The study is unique in size, multiethnic composition, and sampling of repeated measures of salivary cortisol during the day and over multiple days. The methods we develop are motivated by this study, which we use to illustrate the statistical challenges present in epidemiology studies of the stress response.

1.1.2 Design for Studies for Measuring the Variability in Longitudinal Process

The mean profile and the temporal variability are two important features of a longitudinal process. While accurate estimation of the mean profile has been the primary focus of existing approaches for longitudinal study designs, few design approaches to characterize and estimate the variability exist. The variability is itself important in a longitudinal study because the variability can be directly associated with a subsequent outcome of interest (*Elliott*, 2007). A practical example is the urinary progesterone, which is a biomarker important for assessing the women reproductive health. The variability of the urinary progesterone across women is believed to be associated with various reproductive characteristics such as demographics (*Windham et al.*, 2002), occupations (*Gold et al.*, 1995), exercise habit (*De Souza et al.*, 2010). Therefore, further analysis of the association between these health predictors/outcomes and the urinary progesterone profiles hinges on our ability to accurately measure the vari-

ability of the progesterone across individuals. In order to the capture the variability of urinary progesterone, it is not uncommon to collect samples everyday during the entire menstrual cycle of the participant (*Waller et al.*, 1998). Such densely sampled data points are helpful in reconstructing the entire progesterone profile. However, such sampling schedule is difficult to implement in large scale studies since the cost associated with the collection and assay of the samples become very high. Therefore, deriving the simplified and optimal sampling schedules to measure the progesterone level is highly desirable.

In Chapter 3, we will develop methods for optimal sampling schedules that increase our ability to capture the mean profile and the variability of the longitudinal process simultaneously. We will explore the various strategies that are most suitable for deriving schedules for measure the mean profile and the between subject variability. The results we obtained can be useful in practice because they will not only enable practitioners to design cost effective studies, but ultimately help elucidate the relationships between longitudinal process and health outcomes.

We will use the urinary metabolite progesterone data from *Brumback and Rice* (1998) to illustrate our new method. The data set were collected as part of early pregnancy loss studies conducted by the Institute for Toxicology and Environmental Health at the University of California, Davis in collaboration with the Reproductive Epidemiology Section of the California Department of Health Services, Berkeley. The data set contained progesterone profiles within menstrual cycles of 51 women with healthy reproductive function. For some women more than one menstrual cycle was recorded and in total 91 menstrual cycles have been measured. The data set also include the conceptive and non-conceptive status for each cycle, which also allows us to assess the performance of the sampling schedules in terms of predicting the health outcomes from the longitudinal samples.

1.1.3 Design for Studies Involving High Dimensional Genetics and Proteomics Measures

Rapid technological development has expanded the use of high throughput data collection in clinical and epidemiology studies. Such studies typically gather measurements on thousands or millions of proteomics or genetics biomarkers from each subject *Elaine R.* (2008); *Schuster* (2008). Sophisticated statistical analysis is conducted, for instance, to identify associations between disease prognosis and the biomarkers. The results of association studies can shed light on important biomarkers that are differentially expressed for various prognosis groups, which then forms the basis of the development of accurate and reliable classifiers for disease prognosis. These classifiers are among the key components for realizing the promise of personalized medicine. For instance, physicians can select the most effective treatment for a patient based on the prediction of the classifiers, given the measurement of the biomarkers from this very patient *Hamburg and Collins* (2010). Novel approaches for deriving such classifiers for high dimensional data have been developed and applied to real world experiments, leading to promising results.

The utility of high throughput data for clinical settings hinges on well developed classifiers. In turn, a good classifier will depend on proficient study design. For instance, determining the sample size needed to construct an accurate classifier is a design question that comes to mind. This aspect of the study design is particularly relevant for high throughput experiments because the implementation cost, while often in decline, is still prohibitively high and makes it almost impractical to conduct a study with more than a few hundred subjects. Therefore, investigators are highly motivated to obtain reliable sample size estimation that enhances cost-efficiency and minimizes participant risk. Furthermore,since there are a variety of technologies emerging to measure different type of features, the investigators are often interested in combining two or more types of the features to enhance the PCC

of the study. These features can be genetics biomarkers, proteomics biomarkers, or clinical covariates. However, the literature has not formally established for the results regarding how much gain in terms of PCC could potentially be realized when we combine these features in the study and under what scenarios we expect to achieve the highest gain. These questions need to be carefully addressed in the design stage before the investigators commit substantial amount of funds and time to collect several types of features.

In Chapter 4, we will develop methods to efficiently and accurately compute sample size for studies involving high dimensional data and clinical covariates. We will calibrate the sample size to construct a classifier that meets the pre-selected probability of correct classification. We also derive the upper and lower bounds of the PCC gain when two different types of the features are collected in the study. Our research will enable the investigators to better evaluate the feasibility and cost-efficiency of the study in the planning stage and ultimately improve the chance of successful development of more accurate classifiers for disease prognosis prediction.

We apply the new methodology to the design of studies aimed at constructing high dimensional classifiers to predict long-term survival after kidney transplant. Although one year kidney graft survival is greater than 90% due to improved immunosuppression, long-term kidney allograft outcome has not changed dramatically over the last decade (*Meier-Kriesche et al.*, 2004). Traditionally, serum creatine levels have been used as the non-invasive surrogate markers to follow renal allograft function. However serum creatine levels are neither sensitive or specific of long-term survival. For example, the positive predictive value of serum creatine for 7 years graft survival is only 59% (*Kaplan et al.*, 2003). So there is a need to identify better biomarkers to construct more accurate classifiers to predict long-term survival after kidney transplant. Recently a new study is being for planned for such purpose. In total 108 proteins will be measured simultaneously by microarray and classifiers for kidney graft sur-

vival will be developed based on the high dimensional microarray data. This study provides the ideal scenario for evaluating the performance of our methods. We will compute the sample size requirement to achieve a reasonable PCC. The robustness of the sample size calculation will also be assessed through various design assumptions.

1.2 Existing Methods

In this section, we will review the existing methods related to the design problems of interest in this dissertation.

1.2.1 Design for Studies Involving Repeated Measures of Nonlinear Profiles

The longitudinal salivary cortisol profile can be modeled by the nonlinear mixed effect model (*Lindstrom and Bates*, 1990). Thus we review methods for longitudinal design in studies with nonlinear parametric profiles with the focus of selecting optimal sampling schedules and determining the number of repeated profiles (e.g. days) per subject. The key ingredients in this setting are choice of model, , incorporation of uncertainty of parameter values at design stage, criterion for comparing designs and its computation. Determination of the relative magnitude of variability across multiple levels of sampling

Parametric nonlinear mixed models assume that the observations y_{ij} for the subject i at time t_{ij} can be modeled by

$$y_{ij} = f(t_{ij}, \theta_{ij}) + \epsilon_{ij}$$

where $f(t, \theta)$ is the known parametric functions for the profile of the response interest. $f(t, \theta)$ can be any nonlinear functions which allows us to accurately capture the diurnal profile of the salivary cortisol throughout the period of one day. $\theta_i \sim N(\theta, \Sigma)$

is the parametric specific to subject i. θ and Σ are the population mean and variance of the subject specific parameters θ_{ij}. The random effect component of the model provides a simple and straightforward framework to account for the between subject variability. Finally $\epsilon_{ij} \sim N(0, \sigma^2)$ are independent measurement errors. Estimation of the nonlinear mixed models has been established in (*Lindstrom and Bates*, 1990; *Pinheiro and Bates*, 1995). Estimation software can be found in both R and SAS.

Besides the models for salivary cortisol profiles, appropriate criterion for comparing various designs is another crucial ingredient for optimal design. The most widely used approach for selecting the optimal design is based on the Fisher's information matrix (*Retout et al.*, 2002). For a given model, and a given set of population parameters θ (e.g. obtained from preliminary studies), we would like to identify the optimal design $T = (t_1, \ldots, t_d)$ that minimize the estimation variance of $\hat{\theta}$ in the full study. By MLE theory, the MLE $\hat{\theta}$ follows normal distribution with mean θ and variance $I^{-1}(\theta)$ where $I(\theta)$ is the Fisher information matrix. When θ is one-dimensional, one could find the optimal design by minimizing $I^{-1}(\theta)$. When the dimension of θ is higher than one, we need to employ an appropriate criterion to summarize the variance of θ into one quantity. *Atkinson et al.* (2007) discuss many such criteria, such as D-optimality ($\det(I(\theta))$) and A-optimality ($\mathrm{tr}(I^{-1}(\theta))$). Among these criteria, D-optimality is most widely used for several reasons: 1) $\det(I(\theta))$ is the reciprocal of the confidence region for the MLE $\hat{\theta}$; 2) $\det(I(\Phi))$ is invariant under reparameterization of θ.

At the design stage, the values of θ are unknown but can be approximated if preliminary data exist. Such estimates will naturally have potentially high degree of uncertainty. Bayesian designs are a way to incorporate uncertainty into the design criterion (*Chaloner*, 1995). A full Bayesian adaptive approach for the optimal design is described by *Stroud et al.* (2001). In this approach, the design problem is formulated as identifying the optimal design of the next subject to measure the parameter, given the data we have already collected. To achieve this goal, the authors suggest that we

could maximize a well-defined utility function under the posterior distribution given the data already observed,

$$U(T|Data) = \int u(y_{\text{next}}, T, \theta) p(y_{\text{next}}|\theta) p(\theta|Data) dy_{\text{next}} d\theta$$

where $u(y, T, \theta)$ is the utility function and is defined as the precision minus cost in *Stroud et al.* (2001). $p(y|\theta)$ is the sampling distribution for data given the parameter. $p(\theta|Data)$ is the posterior distribution given the observed data. y_{next} denotes the observed data from the next patient. This Bayesian approach provides a coherent framework for adaptively updating the optimal design for each new subject and the resulting individualized optimal design is highly efficient. However, we find it difficult to implement this approach in population based epidemiology studies because individualized sampling protocol poses a huge administrative cost. Nevertheless, the Bayesian design perspective provides a useful formulation for incorporating uncertainty regarding preliminary parameter estimates used at the design stage.

Once the criterion or utility function is well defined, we need to find an efficient algorithm to solve the associated optimization problem. It turns out that optimization of the criterion or utility function is itself difficult because the criterion or utility is multivariate function of the sampling schedule $(t_1, \ldots, t_d)$ with many local maxima. People have tried to solve this problem with different approaches, which we will discuss as follows.

One approach that has been widely studied is based on the continuous design for convex criterion (*Fedorov and Hackl*, 1997). Under this formulation, subjects are assigned to groups, each group with a different schedule. The schedule for each group is termed "elementary design". We define a set A, called the candidate design, which consists of the weighted sum of elementary designs. Weighted sum of two elementary design a_1 and a_2 is denoted by $w_1 a_2 + w_2 a_2$ where $w_1, w_2 \geq 0$ and $w_1 + w_2 = 1$. The

notation $w_1a_2 + w_2a_2$ means that for an incoming subject, he/she will be assigned to design a_i with probability w_i. With weighted sum defined this way, a criterion function $f(\cdot)$ is convex if and only if for any design a_1, a_2 and weights w_1, w_2, we have

$$f(w_1a_2 + w_2a_2) \leq w_1f(a_1) + w_2f(a_2)$$

Most of the criterion functions, such as D-optimality and A-optimality, belongs to the convex criterion if appropriate one-to-one transformation is applied. For a convex criterion, the optimal design can be represented as a finite weighted sum of elementary designs and Federov-Wynn algorithm has been developed to identify the optimal design.

The optimal design based on convex functions is highly efficient and is commonly used in the PD/PK experiment, where the sampling protocols are followed exactly. In these settings, subjects are randomly assigned to each of the elementary designs that form the final optimal design.

However, randomization in the elementary designs could lead to implementation difficulties in epidemiology studies. It is not uncommon for a epidemiology study to focus on more than one objective and collect various responses besides salivary cortisol. In this case, the randomization of cortisol sampling protocols could substantially increase the overall cost, if not disrupt the measurement of the other responses. Therefore, we need to carefully weight the benefit and cost when employing this type of design strategy.

A fixed sampling protocol without random assignment is more practical for epidemiology studies. Thus efficient optimization algorithm should be employed to identify a fixed (non-randomized) optimal design. General optimization algorithm, such as Newton-Raphson Simplex, are not very efficient because of the high dimension of the optimization space and the numerous local minima. A wide variety of exchange algorithms have instead been developed for optimal design problem. For example,

Ogungbenro et al. (2005) propose the so called Modified Federov-Wynn algorithm. In this algorithm, we define a pool of admissible sampling times and choose an initial design. Then iteratively we replace one time point in the current design with a time point from the pool of candidate sampling times, such that the value of the criterion function increases after the replacement. The procedure is repeated until we cannot further improve the criterion function. This approach is extremely fast in terms of convergence. However it is not guaranteed that a global optimum is reached with such algorithm, because it can easily get stuck in local optimum. *Choi et al.* (2007) describes stochastic variant by incorporating simulated annealing. In this algorithm, the replacement of sampling times is accepted stochastically with a nonzero probability even if replacement does not result in an increment of the criterion function. This allows the algorithm to jump out of the local optimum and reach the global optimum eventually.

Finally, the design for the studies with multi-level sampling, which is commonly employed in salivary cortisol studies, has also been considered in the design literature. *Raudenbush and Liu* (2000) derived formulas to determine the optimal number of samples per day and number of days for sampling under two-level linear mixed models. A similar design problem characterized by a three-level linear mixed model is investigated by *Roy et al.* (2007).

1.2.2 Existing Methods for Design for Studies for Measuring the Variability in Longitudinal Process

Because of the similarities, the existing literature for measuring the variability of longitudinal process overlaps substantially with that for measuring the mean of a nonlinear profile. Therefore, we will only focus on the difference and avoid restating the common results.

In *Retout et al.* (2002), the longitudinal process is again modeled by the nonlinear

mixed effect model. But instead of computing the Fisher information matrix only for the mean parameter θ, we compute the joint information matrix for both the mean parameter θ and variability parameter Σ. Then we identify the the optimal schedule for maximizing an efficiency criterion, such as D-optimality or A-optimality, using the various optimization technique, as has been described. While the nonlinear mixed model approach has been widely employed, there are various aspects that can be improved. In particular, under this model, the temporal pattern of the variability is derived from the mean profile. So it might not be flexible enough to model the pattern of the between-subject variability over time, which is the prime focus of our design problem.

A more flexible approach for modeling the variability is based on the functional principal component analysis (FPCA (*Rice and Silverman*, 1991; *Silverman*, 1996)). In this case, we assume that the observation y_{ij} for the subject i at time t_{ij} can be modeled by

$$y_{ij} = f_i(t_{ij}) + \epsilon_{ij}$$

where ϵ_{ij} are the independent measurement errors. $f_i(t)$, the subject specific profile is defined nonparametrically. In particular, $f_i(t)$ is assumed to be Gaussian process with mean $E(f_i(t)) = f(t, \theta)$ where $f(t, \theta)$ is the population averaged profile. We can obtain a parsimonious representation of $f_i(t)$ by applying FPCA, i.e. $f_i(t) = f(t, \eta) + \sum_{k=1}^{\infty} \alpha_{ik}\beta_k(t)$ with $\alpha_{ik} = \int g_i(t)\beta_k(t)dt$ is the loading on each functional principal component $\beta_k(t)$ for each subject i, and $\alpha_{ik} \overset{iid}{\sim} N(0, d_k)$. The principal component functions $\beta_k(t)$'s maximize $Var\left(\int f_i(t)\beta_k(t)dt\right)$ and satisfy the orthonormal requirement $\int \beta_k(t) \cdot \beta_{k'}(t)dt = 0$ for $k \neq k'$ and $\int \beta_k(t) \cdot \beta_k(t)d = 1$. Parsimonious representation can be achieved by selecting a finite number of principal components. The major advantage of the FPCA approach is that the FPCA does not place any assumptions on the temporal pattern of the variability of subject specific profile. Estimation of the FPCA has been developed by (*James et al.*, 2000; *Peng*, 2009).

Various penalties that encourage the smoothness of the principal components have also been proposed (*Silverman*, 1996; *Yao et al.*, 2005; *Yao and Lee*, 2006; *Martinez et al.*, 2010).

1.2.3 Existing Methods for Study Design for High Dimensional Genetics and Proteomics Measures

There are three basic challenges in designing the studies to construct classifiers based on the high dimensional genetics or proteomics measures. We will review the existing methods for each for challenges as follows.

The first challenge is to choose an appropriate and reasonable objective for the sample size calculation. *Hwang et al.* (2002) consider the sample size requirement for rejecting the global null hypothesis of no biomarkers being differentially expressed between groups. However, such hypothesis is not directly related to performance of the classifier and the sample size obtained in this way might not be directly relevant to the purpose of classification. *Mukherjee et al.* (2003) propose using the probability of correct classification (PCC) as the benchmark for evaluating the performance of the classifier and calibrate sample size computation to achieve a predefined PCC. Using PCC as the objective is appealing because it is a direct measure of the performance of the classifiers. But it still has some limitations. The achievable range of PCC varies depending on the practical problem and in some cases, the signal is so weak that the PCC is far less than 1 even with infinite sample size. *Dobbin and Simon* (2007) recognize this issue and propose a better strategy. The authors construct the ideal classifier assuming complete knowledge of the data model and compute the associated ideal PCC. The ideal PCC becomes the upper bound for the PCC of any classifier derived from sampled data. Then they recommend calibrating the sample size to the ideal PCC minus a certain tolerance, which avoids the possibility of setting an objective PCC so high that it cannot be reached even with infinite sample size.

The second challenge is the data analysis tools employed to construct the classifiers. Because of the high dimensionality of the biomarkers, the number of biomarkers is often much larger than the total number of the observations (e.g. the $p \gg n$ scenario). In this case, many conventional statistical analysis tools are not applicable. For example, the sample covariance matrix is always singular and cannot be employed in Fisher's linear discriminant because it involves inverse of covariance matrix. Therefore, regularity structures are often imposed in high dimensional analysis. For example, *Dobbin and Simon* (2007) includes a biomarker selection step to reduce the dimension of the data subsequently used to develop the classifiers. While this approach is simple to implement, it has some drawbacks in practical applications. In particular, the biomarker selection depends on the type I error threshold which needs to be determined by simulation which in turns depends on knowledge of the true model parameters. Since these true values of the model parameters are rarely known with confidence in the design stage, the choice of the type I error threshold could be sub-optimal. In order to address these issues, *Donoho and Jin* (2009) propose a rare-weak model as the framework for high dimensional classification. To be more precise, the model assumes that the number of biomarkers that are useful for classification is typically very small relative the total number of the biomarkers (i.e. useful biomarkers are "rare"); the effect size of the useful biomarkers can be very small relative to the noise (i.e. "weak"). This formulation is strikingly different from the traditional statistical framework and it is no surprise that it leads to some interesting and useful results. *Jin* (2009) shows that the linear classification problem can be divided into two classes: infeasible and feasible. In the infeasible class, the effect size is so weak and the biomarkers are so rare that any linear classifiers will asymptotically perform as bad as a random assignment classifier. For the feasible class, however, *Donoho and Jin* (2009) propose a novel linear classifier based on a higher criticism threshold (HCT) such that the PCC will approach 1 asymptotically

as the the number of biomarkers goes to infinity.

The last challenge is the computation of the PCC given the design objective and the classifier for finite samples. This places a substantial computational burden on the direct estimation of the PCC and makes it too slow to be practical. *Mukherjee et al.* (2003) use a learning curve method to extrapolate the PCC, which requires that extensive data is already available in order to estimate the learning curve parameter. *de Valpine et al.* (2009) propose to compute the PCC by combining simulation and approximation of the PCC formula, which speeds up the computation. However, this combination approach seems to be complicated to implement and the accuracy of the approximation is not yet established. For the newly developed HCT classifiers, only the asymptotic PCC has been consider in *Donoho and Jin* (2009) and no efficient and practical method for computing the PCC for the finite samples has been proposed.

1.3 Approaches

In this dissertation, we will address the design problem of interest by integrating prior results and developing novel design methodologies.

In Chapter 2, we will discuss optimal design strategies for multi-level sampling of salivary cortisol in population studies that maximize the precision of summaries of the stress response. We use conditionally linear mixed models to model the cortisol profile. Instead of D-optimality, we use the Bayesian average of the estimation variance of features of the cortisol response as the criterion for identifying the optimal design. We derive a closed form solution for the estimation variance of the features, which partitions variance into variance due to: the sampling schedule, the between-subject variability and the between-day variability. We extend the result of *Raudenbush and Liu* (2000)to the case of nonlinear response profile and use it to decide on the optimal choice for the number of samples per day and the number of the days for sampling. This straightforward formula for the estimation variance also leads to

other benefits, such as simplifying the computation of the Bayesian average of the estimation variance. We identify the optimal design by inhomogeneous Markov chain simulation (*Muller et al.*, 2004), which progressively improves the precision of the Bayesian average and converges the global optimum. We prefer this approach to the FW algorithm because in our case every subject follows the same schedule and because incorporating the Bayesian design within the FW algorithm would carry a high computational burden. We apply these results to identify optimal design for populations similar to those in the MESA Stress study and incorporate considerations of cost of the multi-level approach.

In Chapter 3, we will develop methods that increase our ability to simultaneously capture the mean and variance profile in repeated measure designs. We review the existing methods based on parametric mixed models and show that accommodating a broad range of between subject variability patterns at the design stage is an area that needs improvement. We utilize functional principal component analysis (FPCA) as a method to characterize the variability structure of the longitudinal process, and develop a method for identifying the optimal sampling schedules for the mean profile and between subject variability. We conduct a small scale simulation study to compare designs derived using our proposed approach to those from using PMM-based approaches. We apply these new methods to two design problems: obtaining sampling schedules for salivary cortisol that also consider flexible models for between subject variability, and sampling schedules for urinary progesterone profiles an early biomarker for conceptive status.

In Chapter 4, we will develop new methods to enhance the sample size compuation of the studies involving high dimensional data. We calibrate the sample size to achieve so that the study could achieve a pre-defined PCC. We enhance the existing method by incorporating cross-validation (CV) and high criticism threshold (HCT) into the feature selection of the classifier. Our new approaches are data driven and

the PCC estimates derived from these approaches are achievable in practice if the corresponding classifiers are employed. In terms of computation, we propose a new simulation method based on order statistics that allows us to efficiently compute the PCC based on the HCT method. Furthermore, we derive an inequality for the upper and lower bounds of the achievable PCC when combining two types of features in the study. We evaluate the performance and validity of our proposed method by extensive simulations. Finally we use the prediction of long term survival after kidney transplant to illustrate the application of our new approaches.

This dissertation contributes novel study design methodologies for studies that involve related but distinct data structures. The design objectives for studies involving these data can include minimizing estimation variance, reducing cost, accurate prediction and high probability of correct classification. The new design methodology will enable the investigators to better evaluate the feasibility and cost-efficiency of the study in the planning stage and ultimately improve the chance of success of studies involving longitudinal and high dimensional data.

CHAPTER II

Designing Salivary Cortisol Studies: the Case of Salivary Cortisol in the Multi-Ethnic Study of Atherosclerosis

2.1 Introduction

Cardiovascular disease remains one of the leading causes of mortality and morbidity in the US, and an area of health where race-ethnic and socioeconomic disparities remain unexplained. In addition to known risk factors, psychosocial stress has been proposed as a cause for cardiovascular disease, and as one of the roots of health disparities (*Kaplan and Keil*, 1993; *Diez Roux et al.*, 2001; *Williams et al.*, 1997; *Gallo and Matthews*, 1999). Subjective measures of stress, such as self-reported questionnaires, are susceptible to response bias, hence salivary cortisol, a field-friendly biological marker of stress, is now collected in many population-based epidemiological studies (*Adam and Kumari*, 2009). Salivary cortisol exhibits a non-linear diurnal pattern through the length of the day, so-called stress response (Figure 2.1). The stress response not only varies between individuals, but also exhibits variability from day to day within a given individual (*Smyth et al.*, 1997). To capture said variability, studies often collect salivary cortisol from multiple days on groups of individuals (*Adam et al.*, 2006; *Cohen et al.*, 2006). However, because very little work exists on

cortisol study design (*Kraemer et al.*, 2006). the number of days and samples per day, as well as the time of day when the samples are collected have been chosen in an ad hoc manner.

We aim to develop an approach of optimal design for mutiple-level sampling of the non-linear salivary cortisol profile, to maximize the precision of model-based features of the population average stress response in community based studies. Salivary cortisol is typically summarized into features (*Adam and Kumari*, 2009). Although the features are commonly constructed from raw data (e.g. AUC constructed with a trapezoidal rule given observed cortisol values and time points), we take a model-based approach by modeling the cortisol profile using (non-linear) random effects models (*Laird and Ware*, 1982; *Lindstrom and Bates*, 1990; *Blozis and Cudeck*, 1999). Our approach incorporates the following design components: 1) Determine the sampling schedule for each day; 2) Decide the number of samples per day and the number of days for sampling; 3) Incorporate practical constraints such as the maximum number of samples collected per participant; 4) Incorporate a wide range of plausible values for the model parameters instead of a single point estimate; 5) Incorporate the implementation cost into the consideration of optimal design; and 6) Incorporate feasibility constrains for large scale community-based studies (e.g. we restrict our attention to designs where the sampling schedule is the same for every subject every day). We develop this approach by extending and integrating optimal design theory and methods from several areas.

Optimal design of repeated measurement studies has received the attention of many researchers. The area of pharmacokinetic and pharmacodynamic modeling (PK/PD) has seen many fruitful results (see *Fedorov and Hackl* (1997); *Fedorov* (2010) for review). In particular, *Retout et al.* (2002) consider the optimal sampling times for measuring the pharmocokinetics of a drug in a Phase III clinical trial. The drug concentration is analyzed using nonlinear mixed effect models which

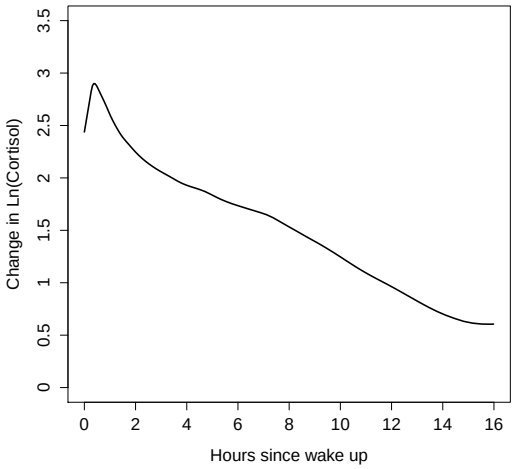


Figure 2.1: LOWESS plot of log(Cortisol) showing 1) a steep climb right after waking up that reaches the peak at approximately 30 minutes after waking up; 2) a fast decline after reaching the peak; 3) a slower rate of decline for the rest of the day.

incorporate the between subject variability. *Retout and Mentré* (2003) generalize the results of *Retout et al.* (2002) to incorporate inter-occasion variability (occasions meaning different days of clinical visits), where inter-occasion variability is assumed to be independent of the between subject variability. *Anisimov et al.* (2007); *Fedorov et al.* (2010) use differential equations to model drug concentration profiles and incorporate within and between subject variability. The Bayesian perspective has also been brought into the optimal design to address two critical issues: incorporate the uncertainty associated with the parameters in the design stage and thus improve the robustness of the optimal design (*Chaloner*, 1995), and provide a general framework to streamline the development of adaptive sampling time of the next subject based on the data from previously enrolled subjects (*Stroud et al.*, 2001). More generally, designs for repeated measures have also been considered (e.g. *Roy et al.* (2007); *Basagaña and Spiegelman* (2010); *Basagaña et al.* (2010)). Specifically, *Raudenbush and Liu* (2000) derived formulas to determine the optimal number of samples per day and number of days for sampling under two-level linear mixed models.

From the computational point of view, two critical decisions in deriving the op-

timal design are the choice of the criterion measuring efficiency of the design, and the numerical algorithms used to solve the optimization problem. The D-optimality criterion, i.e. the determinant of the information matrix, is a popular criterion (*Atkinson et al.*, 2007) since it can be interpreted as the size of the confidence region of the model parameters and it is very useful when the goal is to optimize the variance of the model parameters. However, using the estimation variance of the summary statistics of the profile and mean squared error as the criterion for sample size may also be desirable (*Dawson*, 1998; *Choi et al.*, 2007). Further, hybrid criteria that incorporate the precision as gain and implementation cost as loss to form a utility function, which balances the trade-off between efficiency and overall budget, have also been employed (*Fedorov et al.*, 2002; *Tack and Vandebroek*, 2004; *Bacchetti et al.*, 2008). Nevertheless, most criteria rely on the evaluation of the information matrix, which can be a computationally intensive task itself. *Bazzoli et al.* (2009) compare several linear approximation and simulation based approaches for the evaluation of the information matrix.

Optimization algorithms play a key role, not only from the view of the number of criterion evaluations, but also from the point of view of convergence. Although general optimization algorithms could be used, they often fail to take advantages of the special properties of problem, such as the convexity of the criteria (*Duffull et al.*, 2002). Therefore, there is a demand for algorithms dedicated for the optimal design. For studies that allow different subjects to have different sampling schedules, the Federov-Wynn (FW) algorithm (*Fedorov and Hackl*, 1997; *Retout et al.*, 2007) is an appropriate choice since it automatically divides the subjects into an optimal number of groups and provides the optimal sampling schedules for each group. *Ogungbenro et al.* (2005) propose an exchange algorithm which, compared to the FW algorithm, could potentially reduce computational burden by avoiding the enumeration of candidate schedules.

Our approach for the cortisol design builds upon previous work, by integrating prior results and extending approaches to our particular problem. We use conditionally linear mixed models to model the cortisol profile. Instead of D-optimality, we use the Bayesian average of the estimation variance of features of the cortisol response as the criterion for identifying the optimal design (Section 2). We derive a closed form solution for the estimation variance of the features, which partitions variance into variance due to: the sampling schedule, the between-subject variability and the between-day variability (Section 3). We extend the result of *Raudenbush and Liu* (2000) to the case of nonlinear response profile and use it to decide on the optimal choice for the number of samples per day and the number of the days for sampling (Section 4). This straightforward formula for the estimation variance also lead to other benefits, such as simplifying the computation of the Bayesian average of the estimation variance. We identify the optimal design by inhomogeneous Markov chain simulation (*Muller et al.*, 2004), which progressively improves the precision of the Bayesian average and converges the global optimum. We prefer this approach to the FW algorithm because in our case every subject follows the same schedule and because incorporating the Bayesian design within the FW algorithm would carry a high computational burden. We apply these results to identify optimal design for the populations similar to those in MESA Stress study and incorporate considerations of cost of the multi-level approach (Section 5). We end the paper with a discussion of our findings (Section 6).

2.2 Statistical Models for Salivary Cortisol

We discuss two commonly used parametric models for salivary cortisol and how commonly used features of the curve can be written in terms of model parameters. Because the models are special cases of nonlinear mixed models (*Lindstrom and Bates*, 1990), we review this more general class of models first.

2.2.1 Models for the Cortisol Profile

Suppose on day $j = 1, \ldots, m$, subject $i = 1, \ldots, n$ has d cortisol observations $y_{ij} = (y_{ij1}, y_{ij2}, \ldots, y_{ijd})$ taken at time points $T = \{t_{ij1}, t_{ij2}, \ldots, t_{ijd}\}$ after wake up, then the underlying assumption is

$$y_{ijk} = f(t_{ijk}, \theta_{ij}) + \epsilon_{ijk} \tag{2.1}$$

where $f(t, \theta)$ is the continuous response profile which depends on the subject and day-specific parameters θ_{ij}. We assume the error ϵ_{ijk} for sample $k = 1, \ldots, m$ on day j follows $MVN(0, \Sigma^{\epsilon})$ distribution.

To incorporate subject and day variability into the model, we decompose θ_{ij} into three parts: the fixed population parameter θ, a random subject effect θ_i, and a nested random day effect: $\theta_{i(j)}$

$$\theta_{ij} = \theta + \theta_i + \theta_{i(j)} \tag{2.2}$$

We assume both random effect vectors follow multivariate normal distributions:

$$\theta_i \overset{iid}{\sim} MVN(0, \Sigma^s) \qquad \theta_{i(j)}|\theta_i \overset{iid}{\sim} MVN(0, \Sigma^d) \tag{2.3}$$

where the variability in the subject and day curves can be quantified indirectly through the model variability parameters Σ^s and Σ^d respectively. We assume subject level effects θ_i are independent across subjects, and the day level effects $\theta_{i(j)}$ are independent across days conditional on θ_i.

Two parametric models used to describe the salivary cortisol profile fit within the nonlinear mixed effects model (2.1). Because the cortisol profile over the course of the day can be approximately divided into three periods, a piecewise linear mixed

model has been employed to describe the curve (*Hajat et al.*, 2010):

$$f(t;\theta) = \theta_0 + \theta_1 t + \theta_2 (t - k_1)_+ + \theta_3 (t - k_2)_+ \tag{2.4}$$

where k_1 and k_2 are pre-specified knots, often set at 30 minutes and 2 hours, and $(x)_+ = x$ when $x > 0$ and $(x)_+ = 0$ if $x \leq 0$. The piecewise linear mixed model is straightforward to interpret. The intercept parameter θ_0 represents the cortisol level at wake up. The slopes for the three periods are θ_1, $\theta_1 + \theta_2$, $\theta_1 + \theta_2 + \theta_3$, respectively. Another model for cortisol profile is inspired by a simplified one-compartment pharmacokinetic model (*Stroud et al.*, 2004):

$$f(t;\theta) = \theta_0 + \theta_1 t + \theta_2 t \exp(-\theta_3 t) \tag{2.5}$$

which depends on parameters describing the baseline cortisol level at wake up time (θ_0), a linear term for the average change over time (θ_1), an amplitude parameter (θ_2) approximately describing the height of the morning rise, and a reactivity parameter (θ_3) that describes how fast the morning peak rises and declines and when it occurs.

2.2.2 Cortisol Features

Through a survey of the existing studies on salivary cortisol, we identified various cortisol features of interest to stress researchers. These features are wake up level, cortisol awakening response (CAR), evening decline, area under the cortisol curve (AUC), and prediction at a certain time t. Table 2.1 provides the descriptions, formulas, and references for each of the features found thorough our literature review. Such features can be constructed as functions of model parameters. For instance, if we denote the area under the curve by G, then, G can be written as $G(\theta) = \int f(t,\theta)dt$; the average slope between t_1 and t_2 is $G(\theta) = \frac{1}{t_2 - t_1}(f(t_2,\theta) - f(t_1,\theta))$. For a study with the aim of accurately measuring G, the objective of the optimal design will nat-

Features	Descriptions	Formula	References
Baseline	salivary cortisol level at wake up	$f(0,\theta)$	*Kumari et al.* (2009)
CAR	Difference between the peak (often at 0.5h) and the baseline	$f(0.5,\theta) - f(0,\theta)$	*Pruessner et al.* (1997)
Evening Decline	The average slope in the evening period (10h-16h)	$\frac{1}{16-10}(f(16,\theta) - f(10,\theta))$	*Adam* (2006)
AUC	The area under the cortisol curve in the period between 0h and 16h	$\int_0^{16} f(t,\theta)dt$	*Badrick et al.* (2007)
Prediction	Prediction of the cortisol level at time t	$f(t,\theta)$	*Powell et al.* (2002)

Table 2.1: Cortisol Features

urally be minimizing $Var(\hat{G})$, the variance of the estimated feature. Next we derive the asymptotic variances of $\hat{\theta}$, and subsequently derive $Var(\hat{G})$.

2.3 Asymptotic Variances

The estimation of model (2.1), based on maximum likelihood estimation (MLE), has been established (e.g. *Lindstrom and Bates* (1990)) and can be conducted conveniently with widely available software, such as R and SAS. Here we focus on the variance of model parameters and the features of the curve.

2.3.1 Variance of the MLE

Denote the MLE estimate of the population parameter θ by $\hat{\theta}$. Then we can evaluate the precision of estimation by considering $Var(\hat{\theta})$, which is asymptotically the inverse of the information matrix $I(\theta)$. Unfortunately, there is no closed form solution for $I(\theta)$ for general nonlinear function $f(t,\theta)$. Nevertheless, by Taylor expansion,

Retout and Mentré (2003) proposed a approximation of $I(\theta)$:

$$I(\theta) = \tilde{X}(T,\theta)' D^{-1}(T,\theta)\tilde{X}(T,\theta) + A(T,\theta) \tag{2.6}$$

To fully appreciate (2.6), we need to introduce some notation as follows. Let $X(t,\theta) = df(t,\theta)/d\theta$ be the gradient of $f(t,\theta)$, then form a design matrix $X(T,\theta) = (X(t_1,\theta), \ldots, X(t_d,\theta))'$ by evaluating the gradient at each time point in a given day. Define $\tilde{X}(T,\theta) = (X(T,\theta)', \ldots, X(T,\theta)')'$ to be m stacked copies $X(T,\theta)$ (m days), let $\bar{X}(T,\theta)$ be a block diagonal matrix with m copies of $X(T,\theta)$ on the diagonal, and finally $\Sigma = \begin{pmatrix} \Sigma^s + \Sigma^d & \Sigma^s & \cdots & \Sigma^s \\ \Sigma^s & \Sigma^s + \Sigma^d & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma^s \\ \Sigma^s & \cdots & \Sigma^s & \Sigma^s + \Sigma^d \end{pmatrix}$. Under such notation, we have $D(T,\theta) = \bar{X}(T,\theta)\Sigma\bar{X}(T,\theta)' + \sigma^2 I$ and $A(T,\theta) = -\frac{1}{2}(\frac{\partial vec(D)}{\partial \theta^t})^t \frac{\partial vec(D^{-1})}{\partial \theta^t}$. The formula for $I(\theta)$ is relatively complex to understand and evaluate. As a result, we consider some special cases of $f(t,\theta)$ that makes the evaluation of $I(\theta)$ more accessible.

The simplest case is when $f(t,\theta)$ is linear in θ, i.e., we can write $f(t,\theta)$ as $f(t,\theta) = X(t)\theta$. In this case, $D(T,\theta)$ will not depend on θ and consequently $A(T,\theta) = 0$. Therefore, $I(\theta)$ reduces to $\tilde{X}(T)' D^{-1}(T)\tilde{X}(T)$ for linear models.

In spite of the simplicity of the information matrix under the linear assumption, practical data analysis may demand more sophisticated models, as in the case of cortisol. The conditionally linear mixed model (*Blozis and Cudeck*, 1999) is one class of models that meet such demand yet is still mathematically tractable. Under CLMM, θ is partitioned into two parts: $\theta = (\eta, \phi)$ where η is the linear parameter and ϕ is the nonlinear parameter, and it is assumed that the between subject and between day variability only affect the linear parameter η. With such assumptions, we can write $f(t,\theta_{ij}) = H(t,\phi)\eta_{ij}$ and the gradient becomes $X(t,\theta) = (H(t,\phi), \frac{\partial H(t,\phi)}{\partial \phi})$. By incorporating the fixed nonlinear parameter ϕ, CLMM is able to accommodate a wider

array of biological models than the linear model. At the same time, it is still relative easy to work with. In particular, we can show that approximation formula (2.6) is the exact closed form solution for the information matrix under CLMM (Appendix A). In this way, when we employ CLMM for characterize our data, we are protected against any potential biased associated with the Taylor expansion approximation.

For CLMM, we will show in Appendix B that the ijth element of the matrix $A(T,\theta)$ can be simplified as $A_{ij} = \frac{1}{2}\text{tr}(\frac{\partial Q}{\partial \theta_i}Q^{-1}\frac{\partial Q}{\partial \theta_j}Q^{-1} + (m-1)\frac{\partial P}{\partial \theta_i}P^{-1}\frac{\partial P}{\partial \theta_j}P^{-1})$ where $P = X(T,\theta)\Sigma^d X(T,\theta)' + \sigma^2$ and $Q = P + X(T,\theta)\Sigma^s X(T,\theta)'$. In this case, a sufficient condition for $A(T,\theta) = 0$ is $\frac{\partial Q}{\partial \theta_i} = 0$ and $\frac{\partial P}{\partial \theta_i} = 0$ for all i. Because model (2.5) fits into the framework of CLMM with $\eta = (\theta_0, \theta_1, \theta_2)$, $\phi = \theta_3$, $H(t,\phi) = (1,\, t,\, t \cdot \exp(-\phi t))$ and the parameter estimates from Table 2.2 satisfy the conditions for $A(T,\theta) = 0$, we have $I(\theta) = \tilde{X}(T,\theta)'D^{-1}(T,\theta)\tilde{X}(T,\theta)$ for the nonlinear model (2.5) for salivary cortisol.

2.3.2 Variance of Features

Suppose a study is conducted to measure a feature G derived from the profile $f(t,\theta)$. We assume this quantity G can be written as the function of the parameter θ, i.e. $G = G(\theta)$, as described in Section 2.2.2 and Table 1. For a study with the aim of accurately measuring G, the objective of the optimal design will naturally be minimizing the variance of the estimated feature, $Var(\hat{G})$. By the Delta method, we can relate $Var(\hat{G})$ with $Var(\hat{\theta})$ by

$$Var(\hat{G}) = \nabla G' \cdot Var(\hat{\theta}) \cdot \nabla G = \nabla G' \cdot \frac{1}{n} I(\theta)^{-1} \cdot \nabla G \tag{2.7}$$

where ∇G denotes $dG/d\theta$, the gradient of G. In the optimal design literature, designs minimizing (2.7) are called c-optimal designs (*Atkinson et al.*, 2007).

Because (2.7) involves variance components at multiple levels, it is critical to

have a panoramic understanding of how the variance structure ultimately impacts the variance of the features. In this manner, we can answer many design questions that naturally arise in practice. For example, whether sampling more days for each individual is worth the extra cost, whether the variance components have a substantial impact on the choice of the daily sampling schedules, among others. To address these questions, we study the mathematical properties of (2.7) by examining a simple closed from solution for $I^{-1}(\theta)$. When $A(T,\theta) = 0$ and $I(\theta) = \tilde{X}(T,\theta)'D^{-1}(T,\theta)\tilde{X}(T,\theta)$, it can be shown (*Roy et al.*, 2007) that

$$I(\theta)^{-1} = \Sigma^s + \frac{1}{m}(\Sigma^d + \sigma^2(X'(T,\theta)X(T,\theta))^{-1}) \tag{2.8}$$

In Appendix C, we provide a proof for (2.8) which can be extended to an arbitrary number of levels of variabilities. Equation (2.8) is useful in that the between subject variability Σ^s, the between day variability Σ^d and the design matrix $X(T)$ are completely separated from each other. Therefore, we can write the variance of estimated feature $\hat{G}$ (2.7) as

$$Var(\hat{G}(n,m,d,T)) = \frac{1}{n}\nabla G'\Sigma^s\nabla G + \frac{1}{n\cdot m}\nabla G'\Sigma^d\nabla G + \frac{\sigma^2}{n\cdot m}\nabla G'(X'(T,\theta)X(T,\theta))^{-1}\nabla G \tag{2.9}$$

That is, the estimation variance of $\hat{G}$ can be decomposed into three components, each of them is associated with a clear interpretation: $\nabla G'\Sigma^s\nabla G$ and $\nabla G'\Sigma^d\nabla G$ are the between subject variability and the between day variability for feature G; $\sigma^2\nabla G'(X'(T,\theta)X(T,\theta))^{-1}\nabla G$ is the estimation variance attributed to the schedule T alone. By (2.9), it is clear that the estimation variance is inflated by the between subject variability and between day variability, each with scale $\frac{1}{n}$ and $\frac{1}{nm}$, respectively.

2.4 The Optimal Design in a Longitudinal Study with Repeated Measures

The design for a longitudinal study with repeated measures consists of four components: (n, m, d, T) where n is the total number of subjects; m is the number of days of data collection; and d samples per day are collected according to sampling schedule $T = (t_1, \ldots, t_d)$. We denote by $\hat{\theta}(n, m, d, T)$ the maximum likelihood estimate of θ for a design (n, m, d, T). The asymptotic variance of $\hat{\theta}(n, m, d, T)$ is $\frac{1}{n} I(\theta)^{-1}$, with $I(\theta)$ given in (2.6). The structure of the variance decomposition for the feature $\hat{\theta}$, (2.9), allows us to derive several important results regarding the optimal design (n, m, d, T), as more fully described in what follows.

2.4.1 The Daily Sampling Schedule

Since $Var(\hat{G})$ depends on sampling schedule T only through the term $\nabla G'(X'(T, \theta)X(T, \theta))^{-1}\nabla G$, the optimal schedule T for minimizing $Var(\hat{G})$ can be obtained by minimizing $\nabla G'(X'(T, \theta)X(T, \theta))^{-1}\nabla G$, without any knowledge of the between subject or within subject variance components Σ^s, Σ^d, σ^2, or the number of days m. This result is relevant in two ways. Firstly, a reliable estimate for Σ^s or Σ^d is typically not available in the design stage, especially for those designs relying on small-scale pilot studies. Thanks to this result, the design practitioners can be sure that optimal schedules T are robust against any inaccuracy or misspecification of the variance components Σ^s or Σ^d. Secondly, this result allows us to implement a sequential greedy algorithm for identifying the optimal design: for a given number of samples per day d, we can first identify the optimal time points T, then find the optimal number of days m given T, and finally the optimal number of subjects n. This strategy reduces the computation complexity since we do not need to search the whole design space (n, m, d, T).

It is easy to see that the above result still applies when we want to minimize the weighted average of the variance of several quantities, i.e. $\sum w_k Var(\hat{G}_k)$. One application of this extension is that we can identify the optimal schedules that will minimize the prediction errors over a series of predefined time point $\{t_k\}_{k=1}^K$ by minimizing $\sum_{k=1}^K X(t_k)(X'(T)X(T))^{-1}X(t_k)$, which again does not depend on Σ^s, Σ^d, σ^2 or m.

2.4.2 The Number of Sampling Days and Subjects

In the ideal situation, one can enroll as many subjects and instruct them to collect samples as many days as possible. In reality, ethical and budgetary constraints prevent excessive sampling. Thus it becomes imperative to determine the optimal number of days for sampling and the number of subjects to enroll within such constraints. There are two types of constraints that we consider important and their impacts on the optimal design are discussed as follows.

We first consider the situation where the total number of samples are capped. In this case, we compare the efficiency of two competing study designs $E_1 = (\alpha n, m, d, T)$ and $E_2 = (n, \alpha m, d, T)$ where $\alpha > 1$. These two designs share the same total number of samples αnmd, but $E_1 = (\alpha n, m, d, T)$ enrolls more subjects with fewer samples per day whereas $E_2 = (n, \alpha m, d, T)$ collects more days of data from fewer subjects. Then we can show that the variance of the features will always be lower under design E_1 compared to E_2:

$$Var(\hat{G}(\alpha n, m, T)) \leq Var(\hat{G}(n, \alpha m, T)) \tag{2.10}$$

The equality holds only when $\nabla G' \Sigma^s \nabla G = 0$ (An example satisfying this condition is $G = \phi$ in CLMM). The inequality suggests that when between subject variability is not zero, enrolling more independent subjects to the study is a more effective way

to improve precision than expanding the number days by the same proportion.

In other situations, the budget, instead of the total number of samples, becomes the major constraint for conducting the study. This could happen when the cost to enroll a subject is much higher than the cost of keeping the subject in the study. In this situation, designs with more independent subjects and fewer days might not be cost effective, even if they are statistically more efficient. Therefore, it becomes necessary to determine the optimal design within a fixed budget. We can conceptualize the overall cost for a design (n, m, d, T) as $c_1 n + c_2 nm + c_3 nmd$, where c_1 is the cost for the initial enrollment for one subject; the c_2 is the daily cost for each subject to stay in the study; c_3 is the cost to collect or to analyze each sample. Then for a fixed budget, the optimal number of days of sampling is

$$m_0 = \frac{c_1 \nabla G'(\Sigma^d + \sigma^2 (X'(T,\theta)X(T,\theta))^{-1}) \nabla G}{(c_2 + dc_3) \nabla G' \Sigma^s \nabla G} \tag{2.11}$$

This formula is similar to the results in *Raudenbush and Liu* (2000), which deals with the similar problem under the linear model assumption. Given the choice of m_0, the optimal number of subjects is $n_0 = \frac{C}{c_1 + m_0 c_2 + m_0 d c_3}$. Unlike the optimal sampling schedule T, the number of days m_0 and the number of subjects n_0 depends not only on θ but also on Σ^s and Σ^d. As a result, we consider the robustness of optimal design carefully, which is the focus of the next section.

2.4.3 Robust Design Using Bayesian Approach

From the above derivation, we can see the optimal design (n, m, d, T) as a whole is dependent on the parameter estimate $(\theta, \Sigma^s, \Sigma^d, \sigma)$. So we reach a paradox: We design a study to estimate parameters, but the optimal study design depends on the true values of parameters in question. In order to address this issue, we need to identify a design that is optimal for a wide array of values within the scientifically plausible

parameter space. An optimal design with such property is referred to as robust design (*Atkinson et al.*, 2007). In the study design literature, minimax and Bayesian are two most widely used approaches to identify robust designs. The objective of the minimax approach is to find the design that minimizes the maximum loss quantified by estimation variance, i.e. $\max_{(\theta,\Sigma^s,\Sigma^d,\sigma^2)} Var(\hat{G}(n,m,d,T))$. For the Bayesian approach, a prior distribution $p(\theta,\Sigma^s,\Sigma^d,\sigma^2)$ for $(\theta,\Sigma^s,\Sigma^d,\sigma^2)$ is assumed, and then the design that minimizes expected variance $\int Var(\hat{G}(n,m,d,T))dp(\theta,\Sigma^s,\Sigma^d,\sigma^2)$ is found. *Atkinson et al.* (2007) suggests that the minimax approach might be less desirable because the maximum often occurs on the boundary of the plausible parameter space. So it could lead to a potential loss of efficiency since too much focus is placed on boundary parameters. We opt for the Bayesian approach to incorporate the knowledge regarding model parameters estimated from MESA Stress.

Assuming the variability parameters $(\Sigma^s,\Sigma^d,\sigma^2)$ are independent of the mean parameter θ in the prior distribution, i.e., $p(\theta,\Sigma^s,\Sigma^d,\sigma^2) = p(\theta)p(\Sigma^s,\Sigma^d,\sigma^2)$, the Bayesian average is

$$\int Var(\hat{G})dp(\theta,\Sigma^s,\Sigma^d,\sigma^2) = \mathrm{tr}\left\{K(\theta)\left(\frac{E(\Sigma^s)}{n}+\frac{E(\Sigma^d)}{n\cdot m}\right)\right\} + \frac{E(\sigma^2)}{n\cdot m}E\left\{\nabla G'(\theta)(X'(T,\theta)X(T,\theta))^{-1}\nabla G(\theta)\right\}$$

where $K(\theta) = E(\nabla G(\theta)\nabla G(\theta)')$. This formula is very similar to (2.9). Therefore, in the Bayesian design, we can still employ a greedy search for the optimal sampling schedule T by minimizing $E(X'(T,\theta)X(T,\theta))^{-1}$, independent of the variability parameters $(\Sigma^s,\Sigma^d,\sigma^2)$; and find the best number of days to collect samples using a formula similar to (2.11) with Σ^s, Σ^d and $\nabla G'(\theta)(X'(T,\theta)X(T,\theta))^{-1}\nabla G(\theta)$ replaced by their expectations. Further, since $(\Sigma^s,\Sigma^d,\sigma^2)$ affect $\int Var(\hat{G})dp(\theta,\Sigma^s,\Sigma^d,\sigma^2)$ only through their prior expectation, no full distribution function for $(\Sigma^s,\Sigma^d,\sigma^2)$ is actually needed to carry out a Bayesian design. This substantially reduces the computational

burden for the Bayesian design.

2.4.4 Computation of the Bayesian Design

The most practical computational algorithm for the Bayesian design depends on the model being used (e.g., (2.4) vs (2.5)). For the piecewise linear model (2.4), ∇G and $X(t)$ are not functions of θ. So the Bayesian design is the same as a naive design ignoring possible misspecification of θ. In this case, we identify T by minimizing $\nabla G'(X'(T)X(T))^{-1}\nabla G$ as is prescribed in Section 2.4.1. Since the set of the candidate schedules is finite, the most straightforward way to solve the optimization problem is to enumerate all candidate schedules and find the minimum. This approach will guarantee a global minimum is found. Alternatively, we can use more sophisticated algorithms, such as Federal-Wynn algorithm (*Retout et al.*, 2007) and exchange algorithm *Ogungbenro et al.* (2005). We defer the comparison of these approaches in the discussion section.

For the nonlinear model (2.5), ∇G and $X(t,\theta)$ are functions of θ. In this case, we need to minimize $E(\nabla G'(\theta)X'(\theta,T)X(\theta,T))^{-1}\nabla G\theta))$. The expectation is taken over the prior distribution of θ, which is multivariate normal with the mean and variance parameters estimated from prior data (eg, the MESA Stress data set Table 2.2). Since there is no closed form solution for the expectation, we employ Monte Carlo methods to evaluate $E(\nabla G'(\theta)X'(\theta,T)X(\theta,T))^{-1}\nabla G\theta))$. Monte Carlo sampling is typically computationally intensive. So the methods used for the linear model (2.4), such as enumeration and the exchange algorithm, are no longer applicable. Instead, we employ the inhomogeneous Markov chain simulation (*Muller et al.*, 2004) as a more efficient approach to solve the optimization problem. This simulation approach performs random walk over the space of candidate schedules in order to locate the optimal schedule. Compared to enumeration or the exchange algorithm, it can improve computation efficiency because it will gradually increase of the precision of evaluat-

ing the expectation and does not waste time to compute the expectation for those schedules unlikely to be optimal.

2.5 Design of Salivary Cortisol Studies

In this section, we derive the optimal design for the salivary cortisol studies by using the results discussed in Sections 2.3 - 2.4. We will use the MESA Stress study as an illustrating example.

2.5.1 Preliminary Parameter Estimates: MESA Stress Study

MESA is a population-based, multi-site, large scale epidemiological study which focuses on identifying predictors of subclinical cardiovascular disease (CVD) (*Bild et al.*, 2002). MESA Stress is an ancillary study to MESA, which examines the role of stress as a contributor to a range of precursors of cardiovascular disease. A key component of the study is the inclusion of repeated measures of salivary cortisol each day over multiple days in order to characterize daily profiles of the salivary cortisol as well as within-individual variability in the stress response.

Approximately 1000 participants, free of clinical CVD, collected salivary cortisol samples using cotton swabs over three weekdays at six pre-specified times as follows: wake-up time, 30 minutes after wake up, 10 am, before lunch, 6 pm, and before bed. Each participant recorded the time of data collection on a daily card, but a track-cap device (a container with a cap that registers the time at which it is opened) was used to monitor the times at which the data were actually collected. We use data from 850 participants for whom the self reported and track-cap time differed by less than 15 minutes, and at least 15 samples over the course of three days were collected. We restricted the sample in this manner to avoid any possible bias in estimating parameters needed for study design (e.g. variance components). The final analytical sample consisted of 47.8% males and 52.0% were 65 years or older. Table 2.2 gives

		Piecewise Linear Mixed Model*		Nonlinear Mixed Model*	
		Parameter	S.E	Parameter	S.E
Mean					
	θ_0	2.328	0.0331	2.266	0.0252
	θ_1	1.026	0.0553	-0.116	0.00273
	θ_2	-1.587	0.0651	3.138	0.3582
	θ_3	0.449	0.0245		
	ϕ			1.951	0.1834
	σ^2	0.500^2		0.535^2	
Between Subject Variance**					
	$Var(\theta_0)$	0.470^2		0.347^2	
	$Var(\theta_1)$	0.254^2		0.036^2	
	$Var(\theta_2)$				
	$Var(\theta_3)$	0.273^3			
Between Day Variance**					
	$Var(\theta_0)$	0.081^2		0.023^2	
	$Var(\theta_1)$			0.022^2	
	$Var(\theta_2)$				
	$Var(\theta_3)$	0.028^2			

*Model parameters are not comparable across models.
**Covariance are not shown in the table. If the variance is estimated to be zero, it will be indicated by blanks.

Table 2.2: Parameter Estimates from MESA Stress

parameter estimates for the piecewise linear model (2.4) and the nonlinear model (2.5). The parameter estimates are consistent with the profile depicted in Figure 2.1 and the results from *Hajat et al.* (2010). Since the choice of number of days and samples per day depends on the variance parameters, note that the between day variance is less than 37% of the between subject variance across all parameters. For example, the between subject variance for the intercept in the piecewise linear mixed model is 0.47^2 and between day variance is 0.081^2.

2.5.2 Optimal designs for salivary cortisol features

We consider designs for cortisol features that can be written in terms of θ as summarized in Table 2.1, and, in addition, we also consider two weighted compound objectives (Section 2.4.1). One of them is the average prediction errors over time points 0h, 0.5h, 1h,1.5h, 2h, 3h, 4h, 5h, 16h. The set of time points forms a dense grid near the beginning of the day where the salivary cortisol is expected to change dramatically. Since the prediction errors over these time points are directly comparable to each other, we set all the weights to 1. The other compound objective is the weighted sum of the variance for measuring baseline, CAR, evening decline and AUC. Because the variances for measuring these cortisol features differ dramatically in terms of magnitude, an appropriate weighting scheme is required to ensure no one feature over powers the rest. We choose the weights (0.01, 0.004, 0.986, 0.0001), which are the normalized versions of the reciprocal of the optimized values for the objective function for each feature G, i.e., $\frac{1}{\nabla G'(X'(T,\theta)X(T,\theta))^{-1}\nabla G}$.

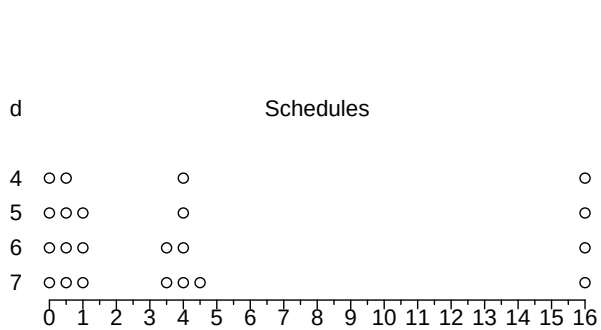
With the cortisol features and underlying model clearly defined, we set out to identify the optimal design for measuring these cortisol features. Since not all design are applicable in a large-scale population epidemiology study, we restrict our attention to a subset of candidate designs (n, m, d, T) that meet certain requirements. Because the linear model and the nonlinear model both contain 4 mean parameters, we need to collect at least 4 samples per day to ensure the information matrix is not singular. At the same time, we cap the total number of samples per day at 7 so as to avoid unnecessary burden on the participants. Hence the possible number of samples per day is $d = 4, 5, 6, 7$. Further, we require the sampling times in the schedule T be chosen from $\frac{1}{2}$ hour intervals from wake-up time: 0h, 0.5h, ..., 16h. In this way the participants may be more likely to follow the sampling protocol. We will consider $m = 1, 2, 3, 4, 5$ days of sampling for each subject. Finally, the total number of subjects n will be determined by the study budget after m, d, and T are chosen. For

exposition, we focus on the design based on nonlinear model (2.5). The design based the piecewise linear model (2.4) shares many similarities and is presented in Appendix D.

We first determine the sampling schedule T using the algorithms described in Section 2.4.4. The optimal schedules for each cortisol features are shown in left column of Figure 2.2. For the number of samples per day d ranging from 4 to 7, the sampling times are represented by circles over the time line. The sampling schedules aiming to capture different cortisol features are different, although some patterns across features are apparent. For example, regardless the features in question, the time points 0, 0.5, 16 are almost always included in the optimal schedules. The three time points are important because cortisol levels at these time points largely define the overall shape of cortisol profile throughout the day: The profile begins a rapid climb from baseline $t = 0$ and reaches the peak at $t = 0.5$ and gradually decline until the bedtime $t = 16$. Once these key time points are included, additional time points are assigned to locations that boost the efficiency of estimating the features under investigation. An exception to this rule is that not all schedules for the evening decline require wake-up or 0.5h samples.

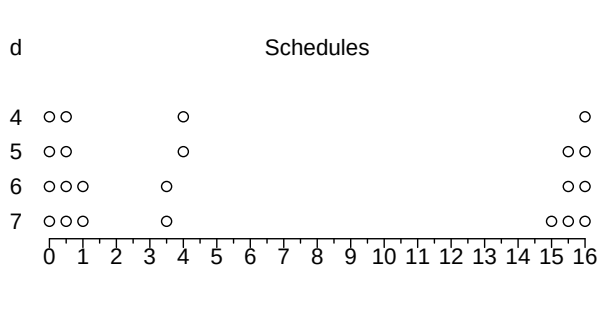
The variance for estimating the features under different optimal schedules are listed in the middle column of Figure 2.2. The variances are computed for a hypothetical study with one participant collecting samples for one to five days. To get the variance for a study with n independent participants, we only need to divide the variance values by n. In general, the variance decreases as we expand the days of sampling. However, the reduction in variance differs significantly from feature to feature. For example, using the 6-sample-per-day schedule under the nonlinear model, the variance to measure baseline, evening decline and AUC is reduced by 13%, 14% and 6.3% if we collect samples over 3 days instead of 2. On the other hand, the variance to measure CAR is reduced substantially by 33%. This disparity in efficiency

gain illustrates the results in Section 2.4.2. By definition, CAR is the difference of the cortisol level between 0.5h and 0h. Since the time points are so close to each other, the majority of the subject level deviation is canceled out when we take the difference. As a result, the between subject variability for CAR ($\nabla G' \Sigma^s \nabla G$ =3.3e-4) is close to zero. Then, by (2.10), expanding the days of sampling is almost as good as enrolling more independent subject into the study in terms of increasing efficiency for CAR. For other cortisol features, however, the between subject variability $\nabla G' \Sigma^s \nabla G$ are relatively high compared to between subject variability $\nabla G' \Sigma^d \nabla G$ or the estimation errors $\sigma^2 \nabla G' (X'X)^{-1} \nabla G$. As a result, increasing the days for sampling does not increase the effective sample size in the same proportion. In summary, the efficiency gain by expanding the number of days will vary depend on features under investigation and should be analyzed on a case-by-case basis.



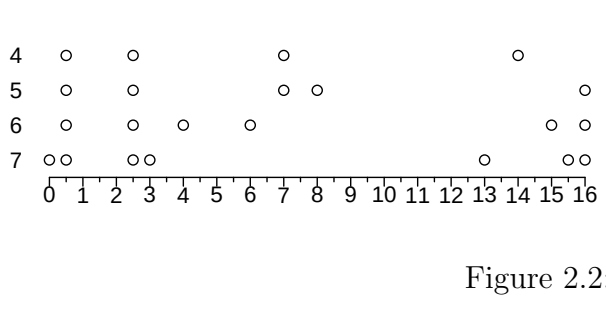
d	Variance 1 Day	2 Day	3 Day	4 Day	5 Day	Cost Adjusted Variance 10^−2 1 Day	2 Day	3 Day	4 Day	5 Day
4	0.402	0.261	0.214	0.191	0.177	0.667	0.54	0.531	0.551	0.583
5	0.31	0.215	0.183	0.168	0.158	0.527	0.462	0.477	0.512	0.554
6	0.268	0.194	0.169	0.158	0.150	0.467	0.432	0.459	0.508	0.556
7	0.246	0.184	0.162	0.152	0.146	0.438	**0.424**	0.46	0.513	0.568

(a) Baseline



d	Variance 10^−3 1 Day	2 Day	3 Day	4 Day	5 Day	Cost Adjusted Variance 10^−5 1 Day	2 Day	3 Day	4 Day	5 Day
4	4	2.66	2.22	1.99	1.86	6.64	5.51	5.5	5.76	6.14
5	3.5	2.41	2.05	1.87	1.76	5.95	5.19	5.33	5.7	6.16
6	3.23	2.28	1.96	1.80	1.71	5.62	5.08	5.33	5.79	6.32
7	3.05	2.19	1.9	1.76	1.67	5.42	**5.06**	5.39	5.92	6.51

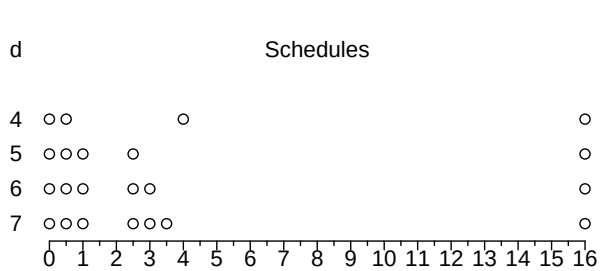
(b) Evening Decline



d	Variance 1 Day	2 Day	3 Day	4 Day	5 Day	Cost Adjusted Variance 1 Day	2 Day	3 Day	4 Day	5 Day
4	82.7	67.6	62.6	60.1	58.6	1.37	1.4	1.55	1.74	1.93
5	76.6	64.6	60.6	58.6	57.3	1.30	1.39	1.58	1.79	2.01
6	73.5	63.1	59.5	57.8	56.7	1.28	1.41	1.62	1.86	2.1
7	71.4	62	58.8	57.2	56.3	**1.27**	1.43	1.67	1.93	2.20

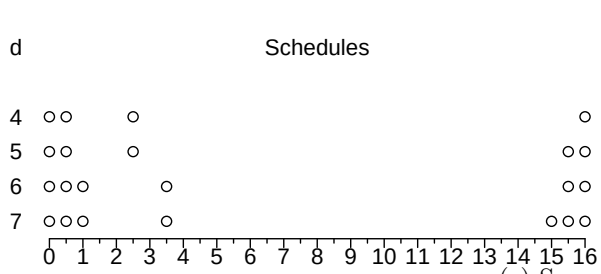
(c) AUC

Figure 2.2: Optimal Design under the Nonlinear Model



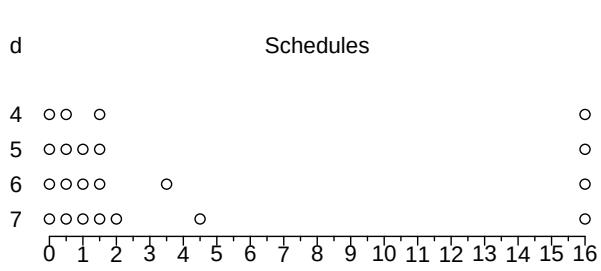
	Variance					Cost Adjusted Variance 10^−2				
d	1 Day	2 Day	3 Day	4 Day	5 Day	1 Day	2 Day	3 Day	4 Day	5 Day
4	0.562	0.281	0.188	0.141	0.113	0.933	0.582	0.465	0.407	0.372
5	0.427	0.215	0.142	0.106	0.0853	0.726	0.462	0.368	0.323	0.299
6	0.373	0.185	0.122	0.0914	0.0745	0.648	0.412	0.333	0.294	0.276
7	0.335	0.172	0.112	0.0854	0.0683	0.596	0.398	0.318	0.288	**0.267**

(d) CAR



	Variance 10^−3					Cost Adjusted Variance 10^−5				
d	1 Day	2 Day	3 Day	4 Day	5 Day	1 Day	2 Day	3 Day	4 Day	5 Day
4	3.94	2.62	2.18	1.96	1.83	6.54	5.42	5.4	5.66	6.03
5	3.45	2.38	2.02	1.84	1.73	5.87	5.11	5.25	5.61	6.06
6	3.18	2.24	1.93	1.77	1.68	5.54	5	5.24	5.69	6.21
7	3.01	2.15	1.87	1.73	1.64	5.35	**4.97**	5.3	5.82	6.4

(e) Sum of Variances of Cortisol Features



	Variance					Cost Adjusted Variance 10^−2				
d	1 Day	2 Day	3 Day	4 Day	5 Day	1 Day	2 Day	3 Day	4 Day	5 Day
4	3.72	2.60	2.23	2.05	1.94	6.18	5.39	5.54	5.93	6.41
5	3.4	2.44	2.12	1.97	1.87	5.78	5.24	5.52	6.01	6.56
6	3.10	2.30	2.03	1.90	1.82	5.4	**5.14**	5.53	6.11	6.73
7	2.95	2.22	1.98	1.86	1.79	5.25	5.14	5.63	6.27	6.98

(f) Prediction

Figure 2.2: Optimal Design under the Nonlinear Model (Continued)

2.5.3 Cost Considerations

We also consider the cost factor in order to determine the optimal number of days for sampling. Our goal is to find the design (n, m, d, T) that minimizes the estimation variance for a fixed budget. We obtain the itemized cost structure of MESA Stress study as follows: initial enrollment (including the Trackcap Device) $c_1 = \$125$, daily compensation $c_2 = \$25$, cost per assay sample $c_3 = \$4$. So the cost per subject is $125 + 25m + 4md$ and a total of $n = \frac{C}{125+25m+4md}$ subjects can be enrolled if the overall budget of the study is C. By substituting the numbers of (n, m, d, T) into the variance formula for $\hat{G}$, we can compute the cost-adjusted variance, i.e. the overall estimation variance when the total budget is fixed at C. While the optimal choice of (m, d, T) is the same regardless of the budget C, we set C to be \$10000 for purpose of exposition. The cost adjusted variances for different choice of (m, d, T) are shown in the right column of Figure 2.2. The cost-adjusted variance highlighted in bold is the smallest for the feature of interest and it corresponds to the optimal choice of (m, d, T). After adjusting for the cost, baseline, evening decline and AUC can be more accurately measured if we sample subjects for just one or two days and use the resources to enroll more subjects. On the other hand, the opposite strategy should be used if the objective is to measure CAR.

In applying the results from Section 4 to MESA Stress study, we have provided recommendations for the optimal design for measuring salivary cortisol profiles given the mean parameters, variability parameters and the costs. Of course, the assumptions on the parameters reflects the characteristics of the population targeted by MESA Stress. For example, a substantial proportion of the participants in MESA Stress are senior citizens. In this population, the between day variability of many cortisol features is tiny. If the study targets a different population, the model parameters may be different and the optimal design might change accordingly. The same argument also applies to the itemized cost structure since different studies are likely to employ

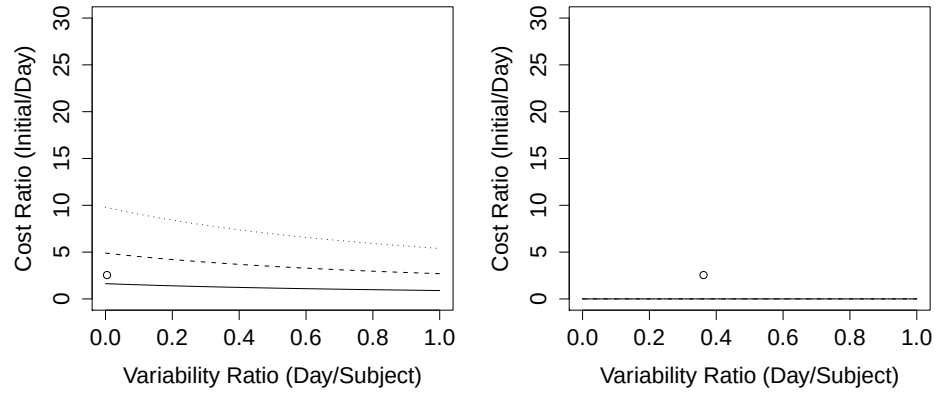
different field techniques to enroll subjects and collect samples. In this context, we find it beneficiary to investigate the optimal design beyond the assumption of MESA Stress on the population and cost structures.

Next we will focus on the optimal number of days. By (2.9), the optimal number of days depends on the Initial : Daily cost ratio as well as the Between-Day : Between-Subject variability ratio. These two ratios are the y-axis and x-axis in Figure 2.3. Each point in the figures represents a particular study setting with corresponding cost ratio and variability ratios. In particular, the dot in each figure indicates the cost ratio and variability ratio for the cortisol features in MESA Stress. The curves divide the spaces into several regions in which a particular number of days of sampling is preferred: long dashes, $- - -$, separate the region 1 day and region 2 days; short dashes, - - -, separate 2 days vs. 3 days; and the dotted line distinguishes 3 days vs. 4 days. These figures are a convenient tool for the investigators to analyze the sensitivity of the optimal design for various population characteristics and cost structures. For example, in the figure for AUC under nonlinear model, the dot corresponding to MESA Stress is far away from the boundary of the one-day region where the dot resides. So even if there is slight changes to the population characteristics or cost structures in MESA Stress, we are still assured that the one-day design is more efficient than other designs.

2.6 Discussion

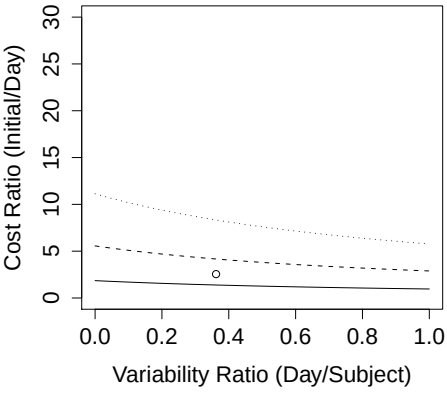
The scientific questions that investigators seek to address with longitudinal data are becoming increasingly complex. One example is understanding the nonlinear biological stress response as measured by salivary cortisol. While a longitudinal perspective to quantify the dynamic profile helps us better understand the complex biological process, it also posts many challenges from the viewpoint of study design.

In the case of salivary cortisol, we considered sampling not only individuals, but

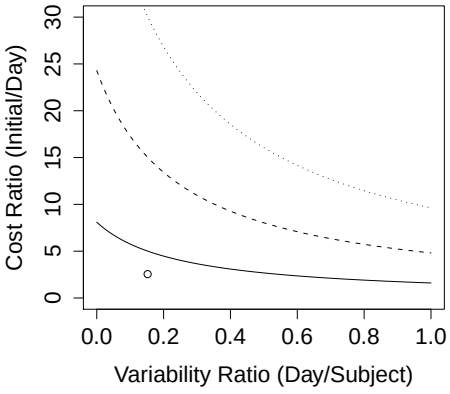


(a) Baseline

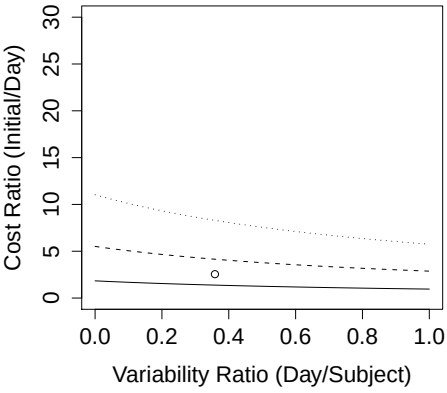
(b) CAR



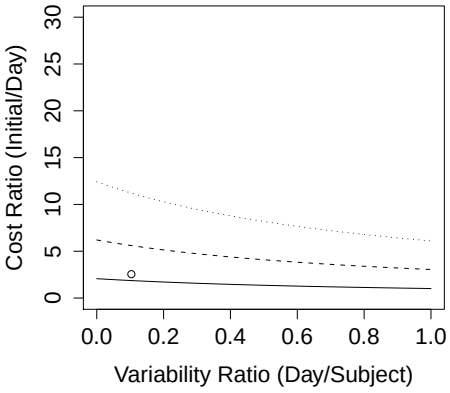
(c) Evening Decline



(d) AUC



(e) Sum of Variances of Cortisol Features



(f) Prediction

Figure 2.3: Cost Ratio vs. Variability Ratio (Nonlinear Model)

also days within individuals, and times within days, with the objective of minimizing estimation variance of profile features. We analyzed the impacts of the multi level variabilities on the study design by deriving a simplified version of the variance formula. Such formula is surprisingly neat in that the terms involving the between subject variability, between day variability, and the daily sampling schedules are completely separated from each other. Hence, we derived several useful properties: 1) the optimal daily sampling schedule depends only on the shape of the response profile; 2) we obtained the optimal number of days for sampling under different cost structures for the study; 3) we showed that the optimal Bayesian design is only affected by the expectation of the variability parameters, but not their full prior distribution. These properties are useful in practice since they substantially reduce the computation burden of Bayesian designs.

We utilized a piecewise linear model (2.4) and nonlinear model (2.5) in our discussion. Both models have been used in the cortisol literature and naturally they have some advantages and disadvantages. The nonlinear model (2.5) provides smooth prediction curves for the cortisol profile throughout the day, and since it is derived from pharmocokinetics, its parameters have a meaningful biological interpretation. On the other hand, the piecewise linear mixed model is much simpler and the corresponding fitting algorithm is faster and more stable numerically. Furthermore, the variance estimates in the piecewise linear model do not depend on the population parameter θ, which is considered to be an advantage for study design since we typically do not know the exact values of θ in the planning stage (see Appendix D).

We note that the proposed designs require very few samples to be taken between 6 hr and 16 hr (bedtime) since wake up. This is due to the fact that both models reduce to a straight line in this period. In this case, the proposed sampling protocols are highly efficient since they do not waste any valuable sampling points in the region which has very little impact on determining the slope of a straight line. On the other

hand, it maybe unwise to totally ignore the possibility that the cortisol profile in this period calls for a more sophisticated model than a decline at constant rate. Current studies have not collected many samples at the later part of the day, hence this region of the salivary cortisol profile is less understood. Thus including one or two sampling points between 6 hr since wake up and bedtime would increase the robustness of the design.

We assume that cortisol sampling is carried out at the exact times prescribed in the protocol. In many practical applications, the participants might find it difficult to follow sampling instructions exactly, causing some deviations from the prescribed sampling times. Usually the deviations will not affect the optimality of the sampling protocol as long as the deviations are small. Furthermore, *Vrijens and Goetghebeur* (1999) suggest that some deviations will help us gain a more complete and detailed picture of the cortisol profiles and potentially improve the estimation of the population parameters.

We also made assumptions regarding the cortisol model. First, in order to derive schedules, we assumed measurement errors in cortisol to follow an iid normal distribution. Residual analysis of MESA Stress data shows no strong indication of deviations from normality or autocorrelations. Nevertheless we can relax this assumption to accommodate multivariate normal distribution with correlated covariance structure, such as first order autoregression. We also assumed there are no random effects associated with the nonlinear parameters. The fitted model from the MESA Stress data indicates that the variability of the nonlinear parameter is virtually zero. As a result we do not consider random effects for the nonlinear parameter ϕ. Doing so greatly simplifies the variance calculations, particularly in the case where we need to model two levels of randomness, i.e. between subject and between day. On the other hand, if the variability of ϕ is not negligible, a potential approach is to linearize the nonlinear term containing ϕ by first order linearization. The appropriateness of

the linearization is discussed in *Bazzoli et al.* (2009). Another possibility is to derive variance formulas using numerical derivatives. This would increase the computational expense to find the optimal sampling design.

There exist several methods for minimizing the variance contribution due to a given schedule, $\nabla G'(X'(T,\theta)X(T,\theta))^{-1}\nabla G$, such as the Fedorov-Wynn algorithm (*Retout et al.*, 2007) and the Modified Fedorov Exchange Algorithm (MFEA) (*Ogungbenro et al.*, 2005). The F-W algorithm divides the population into several groups and provides optimal schedules T for each group. This approach is appealing because a multi group approach may increase the efficiency. For large scales population studies, however, salivary cortisol is among a wide array of biological responses collected and randomization of the schedules could substantially increase the administrative cost. We also tested MFEA in our application of cortisol study design. Since it is a deterministic procedure, we found MFEA to be sensitive to the initial guess of the optimal design and often got stuck in local optimums. Considering these potential issues, we use enumeration for the linear model since it guarantee to obtain the global optimum and the computation is tolerable. We employ inhomogeneous Markov chain simulation for the Bayesian design of the nonlinear model because it is efficient and the inherent randomness allows us to jump out of local optimum.

We can further extend these results in several directions. Throughout this paper, we used parametric models in describing the cortisol profile. Future work can involve non-parametric models which place less restriction on the data than the parametric models. Another extension is to consider the sampling protocols for detecting differences between groups. For example, cortisol profiles may differ by age, gender and ethnic groups. From the statistical point of view, the same method for selecting the efficient sampling protocol can be applied to all groups since optimal sampling protocols are typically robust to small changes in parameters. On the other hand, if the differences in the cortisol profiles are big, new methods might be needed to

simultaneously select schedules for each group, as well as optimizing other criteria like power or sample size.

In summary, our understanding of the structures of the optimal design for features of nonlinear response profiles can be enhanced if we are able to obtain a simple closed form solution for their variances. Decomposing variance formulas into between/within subject variability estimation error can help identify the optimal design more efficiently as well as simplify the Bayesian computation.

CHAPTER III

A Semi-parametric Approach to Select Optimal Sampling Schedules for Measuring the Mean Profile and Variability in Longitudinal Studies

3.1 Introduction

Longitudinal studies with repeated measures are widely employed in medical and epidemiological settings and provide the unique opportunity to investigate an outcome's mean profile and its variability over time. These studies deepen our understanding of the outcome and enhance our ability to identify predictors of change. However, longitudinal studies have to be carefully designed to achieve their potential.

Longitudinal study design is generally complex, requiring the number of subjects and samples per subject and the spacing between measurements (i.e., the sampling schedules), while meeting budgetary and logistical constraints. In one of our motivating examples, investigators want to identify the best timing to collect stress biomarkers, such as salivary cortisol, to estimate the daily nonlinear stress profile and to identify its variability across the population. Relative to methods for sample size and power calculations for repeated measure studies (e.g. *Dawson* (1998); *Raudenbush and Liu* (2000); *Basagaña and Spiegelman* (2010)), methods to determine the sampling schedule have received less attention.

Researchers have discussed approaches to select optimal sampling schedules based on parametric nonlinear mixed models (PMM) for longitudinal data (*Fedorov and Hackl*, 1997; *Stroud et al.*, 2001). These approaches have been used primarily in pharmacodynamics and pharmacokinetics studies and are advantageous from a design and analysis standpoint when mechanistic models for the longitudinal process exist, e.g. kinetic models describing drug clearance rates. However, mechanistic models are not always available and PMM can be too restrictive in modeling the temporal pattern of the variability (see Section 2).

Capturing the variability is an important aspect in a longitudinal study, for at least three reasons: (a) incorporating correct models for the variance structure can improve the quality of inference (*Carroll*, 2003); (b) the variability of a longitudinal predictor can be directly associated with a subsequent outcome of interest (*Elliott*, 2007); (c) understanding the temporal pattern of the variability across individuals can help us identify the optimal sampling windows to assess between individual differences. Adequately capturing the variability of the process will likely to improve efficiency and expand the types of questions that can be addressed.

We propose a unified framework for identifying the optimal sampling schedules to accurately estimate the mean profile and between subject variability. We assume identical sampling schedules for all subjects since it is most practical in large scale studies. We review existing methods based on parametric mixed models in Section 3.2, and show that parsimonious approaches to accommodate a broad range of between subject variability patterns at the design stage is an area that needs improvement. In Section 3.3 we utilize functional principal component analysis (FPCA) to characterize the variability structure of the longitudinal process, and develop a method for identifying the optimal sampling schedules for the mean profile and between subject variability. Section 3.4 contains a small scale simulation study to compare our approach and existing methods. We apply these methods to two design problems in

Section 3.5, and conclude with a discussion in Section 6.

3.2 Optimal Schedules by Parametric Mixed Model

Suppose subject $i = 1, \ldots, N$ will collect samples according to the sampling schedule $T = (t_1, \ldots, t_n)$. We assume the observations $y_i(t_j)$ of subject i at time t_j follow

$$y_i(t_j) = f(t; \eta_i) + \epsilon_{ij} \quad \eta_i \overset{iid}{\sim} N(\eta, \Sigma) \tag{3.1}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ are independent and identically distributed (iid) measurement errors; f is a known nonlinear parametric function; and η_i is a subject-specific vector of length p.

The design goal is to find schedule $T^* = (t_1, \ldots, t_n)$ that minimizes the estimation variance of (η, Σ). Letting $l(\eta, \Sigma)$ be the log-likelihood, the asymptotic variance of the MLE $(\hat{\eta}, \hat{\Sigma})$ is the inverse of the information matrix $I(\eta, \Sigma) = -E \begin{pmatrix} \frac{\partial^2 l}{\partial \eta^t \partial \eta} & \frac{\partial^2 l}{\partial \eta^t \partial \Sigma} \\ \frac{\partial^2 l}{\partial \Sigma^t \partial \eta} & \frac{\partial^2 l}{\partial \Sigma^t \partial \Sigma} \end{pmatrix}$, where by a slight abuse of notation, Σ on the right hand side denotes the vector of the upper triangular elements of the covariance matrix Σ in (3.1). Various optimization criteria based on the information matrix $I(\eta, \Sigma)$ have been developed to select the optimal sampling schedules (*Atkinson et al.*, 2007). D-optimality, i.e. maximizing $\det(I(\eta, \Sigma))$, has many desirable properties: 1) it is the reciprocal of the size of the confidence region for the MLE $(\hat{\eta}, \hat{\Sigma})$; 2) it is invariant under reparameterization of (η, Σ); 3) it is convex, allowing the use special algorithms to find optimal schedules (*Retout et al.*, 2002; *Ogungbenro et al.*, 2005).

This parametric mixed model (PMM) approach is intuitive, but three aspects can be improved. First, the PMM approach might not always be flexible enough to characterize the temporal pattern of the variability, given by $Var(f(t, \eta_i))$ at time t, because both the distribution of η_i and the functional form of $f(t, \eta)$ affect

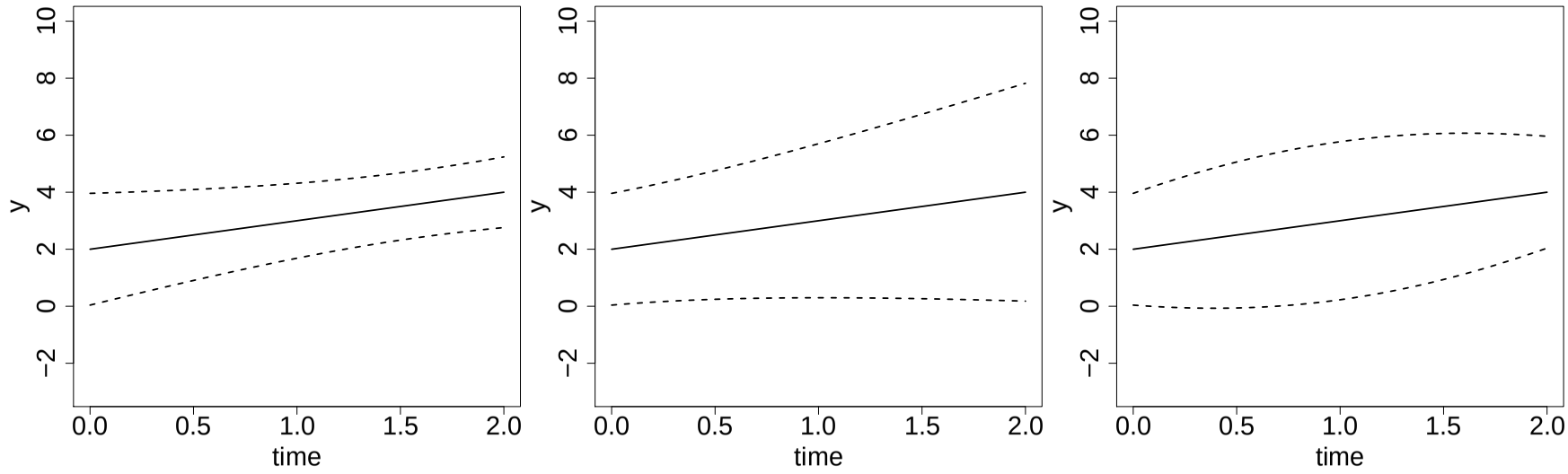
$Var(f(t, \eta_i))$. Figure 1 shows examples where the variance pattern cannot always be captured by a parametric mixed model even though all figures share the same model for the mean. Second, the PMM approach may not adapt to the situation where a mechanistic model for the mean profile is not available. Although spline models, i.e. $f(t, \eta) = \sum_{l=1}^{L} \eta_l B_l(t)$, have been employed to model nonlinear time trends (*Green and Silverman*, 1994) they do not easily lend themselves to deriving meaningful sampling schedules. Due to the linearity of η in a spline-based model for $f(t, \eta)$, the information matrix corresponding to the mean profile, $-E(\frac{\partial^2 l}{\partial \eta^t \partial \eta})$, does not involve η. Therefore, maximizing $\det(-E(\frac{\partial^2 l}{\partial \eta^t \partial \eta}))$ based on one set of basis functions always leads to exactly the same optimal schedule, no matter what longitudinal process is. Last, the information matrix $I(\eta, \Sigma)$ is difficult to evaluate. The major obstacle is that no closed form solution exists for the integral with respect to the individual effect η_i when $f(t, \eta)$ is a general nonlinear function of η. Approximations have not established error rates, and nuermical methods or can be time consuming if we need to compute $I(\eta, \Sigma)$ repeatedly (*Bazzoli et al.*, 2009). So while it is very useful, the PMM approach is not universally applicable.

3.3 Optimal Schedules by Functional Principal Component Analysis

3.3.1 Modeling Strategy

In order to relax the constraints in the PMM approach, we consider modeling the mean profile and variability structures separately as follows:

$$y_i(t_{ij}) = f(t_{ij}; \eta) + g_i(t_{ij}) + \epsilon_{ij} \tag{3.2}$$



(a) (b) (c)

Figure 3.1: Various temporal pattern of the longitudinal processes.
The solid line represent the mean profile, which is assumed to be $f(t;\eta_0,\eta_1)=\eta_0+\eta_1 t$ and (η_0,η_1) is fixed parameter vector. The dash lines represent the variability pattern.
In Figure 3.1b and 3.1a, we assume the between subject variability is derived from the mean profile using random effect of the parameters, i.e. the profile for subject i is $f(t;\eta_{i0},\eta_{i1})$ with $(\eta_{i0},\eta_{i1})\sim N((\eta_0,\eta_1),\Sigma)$. Σ is a 2×2 covariance matrix with elements $\tau_{kl},\ k=1,2;\ l=1,2$. Then the variability $Var(f(t;\eta_{i0},\eta_{i1}))=\eta_0^2\tau_{11}+2\eta_0\eta_1\tau_{12}\cdot t+\eta_1^2\tau_{22}\cdot t^2$ is a quadratic function of t. In this case, the variability in the middle of the time interval is always lower than the variability at least one end point.
In Figure 3.1c, we consider a more flexible structure for the between subject variability, i.e. the profile for subject i is $f(t;\eta_0,\eta_1)+g_i(t)$ where $g_i(t)$ is general random process unrelated to the mean profile $f(t;\eta)$. The variability $Var(f(t;\eta_0,\eta_1)+g_i(t))=Var(g_i(t))$ could demonstrate any pattern depending on the property of $g_i(t)$. In particular, if the variability is higher in the middle of the time interval, the variability structure cannot be characterized by the random effect model employed in Figure 3.1b and 3.1a.

where, similar to PMM, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ are measurement errors; $f(t; \eta)$ is a known parametric function for the mean profile, and η is the fixed population mean parameter as in Section 3.2, which does not depend on individual characteristics covariates. In contrast to model (3.1), where between individual variation is modeled by random effects η_i, we model the subject specific deviation by nonparametric random effect $g_i(t)$. We assume $g_i(t)$'s are *iid* Gaussian processes with $E(g_i(t)) = 0$ and place no restriction on the temporal covariance structure of $g_i(t)$. Subjects with different covariates will have different $g_i(t)$, i.e., i's profile $f(t; \eta) + g_i(t)$ implicitly depends on covariates. This is advantagous because if some covariates exert strong effects on the subject specific profiles, we will observe higher between subject variability that is attributable to those covariates. Because optimal schedule for capturing the variability will automatically include the time intervals where the longitudinal profile exhibits high variability, this model will enable us to collect data in a way that boosts the statistical power when regressing the longitudinal profile on influential covariates.

We use functional principal component analysis (FPCA) (*Rice and Silverman*, 1991) to characterize $g_i(t)$. The FPCA framework consist of finding smooth principal component functions $\beta_k(t)$, $k = 1, 2, 3, \ldots$ that maximize $Var\left(\int g_i(t)\beta_k(t)dt\right)$ with the orthonormal restrictions $\int \beta_k(t) \cdot \beta_{k'}(t)dt = 0$ for $k \neq k'$ and $\int \beta_k(t) \cdot \beta_k(t)d = 1$. Each functional principal coponent account for variance $d_k = Var(\int g_i(t)\beta_k(t)dt)$; we assume d_k are in decreasing order. Under this framework, we have $f_i(t) = f(t, \eta) + \sum_{k=1}^{\infty} \alpha_{ik}\beta_k(t)$, where $\alpha_{ik} = \int g_i(t)\beta_k(t)dt \overset{iid}{\sim} N(0, d_k)$ is the loading on the kth component $\beta_k(t)$ for subject i. Since d_k are in descending order, α_{ik} is typically negligible for large k. Hence we only need the first few principal components $\beta_k(t)$, $k = 1, \ldots, r$, to capture the majority of the variability. This approximation leads to the reduced rank model introduced by *James et al.* (2000)

$$y_i(t) = f(t, \eta) + \sum_{k=1}^{r} \alpha_{ik}\beta_k(t) + \epsilon \tag{3.3}$$

James et al. (2000) further developed an EM algorithm to obtain the maximum likelihood estimate of the parameters since the reduced rank model has only finite number of components. With FPCA, the variability of $f_i(t)$ can be summarized by a few independent statistics d_k, which are much easier to handle than the $p \times p$ covariance matrix Σ in (3.1).

3.3.2 Design Approach

Let S be the set of the admissible sampling times. Theoretically, our method does not place any restriction on S. However, investigators will probably prefer to limit S to sampling times that can be implemented in practice. Let $T = (t_1, \ldots, t_m)$ denote a candidate schedule where $t_j,\ j = 1, \ldots, m$ is chosen from S and $t_j \neq t_{j'}$ if $j \neq j'$. We denote the subset of candidate schedules by S_c. Our design goal is to select a sampling schedule $T^* = (t_1, \ldots, t_m) \in S_c$ such that it accurately estimates both mean and variability. So we would like to minimize the estimation variance of η and $D = diag(d_1, \ldots, d_r)$. We follow the information matrix approach for the mean parameter η and extend it to incorporate the variance parameter D. Based on the reduced rank model (3.3), we derive the information matrix $I(\eta, D; T)$. Unlike the parametric approach, there is a closed form solution for $I(\eta, D; T)$ which can be written as a block diagonal matrix, with blocks $I(\eta; T)$ and $I(D; T)$. Denoting ∇ as the gradient, we have $I(\eta; T) = \nabla f(T, \eta) \cdot A \cdot \nabla f(T, \eta)$ where $A = B(T)DB(T)' + \sigma^2 I$ and $B(t) = (\beta_1(t), \ldots \beta_r(t))$. The kl^{th} element of $I(D; T)$ can be written as $I(D; T)_{kl} = \frac{1}{2} tr(A^{-1}\beta_k(T)\beta_k(T)'A^{-1}\beta_l(T)\beta_l(T)')$.

From the pool of all candidate schedules S_c, we identify the optimal sampling schedule T^* by maximizing an appropriate criterion. Two scenarios are particularly interesting: when we have a good mechanistic model for the mean profile and the variance, then we can derive an optimal schedule T^* for both mean and variability by maximizing the D-optimal criterion (*Atkinson et al.*, 2007): $T^* =$

$\text{argmax}_{T \in S_c} \det(I(\eta, D; T))$. If a mechanistic model is absent and splines are employed to model mean profile, then we can still perform FPCA to reconstruct the variance components and derive optimal schedule to capture the variance. Since the variance is only part of the model, we employ the D_s-optimal criterion (*Atkinson et al.*, 2007) to obtain T^*: $T^* = \text{argmax}_{T \in S_c} \det(I(D; T))$.

3.3.3 Implementations

Our design approach can be implemented in three steps: 1) prepare preliminary data for estimating the FPCA; 2) estimate FPCA model from preliminary data, including the selection of appropriate number of components and smoothing parameter via cross validations; 3) maximize the optimization criterion to find T^*. We sketch the implementations of each step below and provide more detailed discussions including brief review of FPCA estimation in the Web Appendices.

Preparation: Densely sampled preliminary data over the entire time interval of interest is needed to reconstruct the variability pattern, which can be achieved by densely sampling profiles for individuals or combining data from individuals if they each have different (sparse) sampling points. If a particular time interval is not covered by the preliminary data, no interesting variance structure inside such interval can be identified by our method, or any other statistical method, simply due to lack of information. As a result, the optimal schedule derived from the FPCA method is unlikely to include any sampling time from such intervals.

Estimation: Various estimation procedures have been developed for FPCA, starting from the seminal work of *Rice and Silverman* (1991) based on the eigenfunctions of the covariance kernel $Cov(f_i(t), f_i(s))$. *James et al.* (2000) developed an EM algorithm for the reduced rank model. In considering the joint modeling of pairs of sparse functional data, *Zhou et al.* (2008) suggests a smoothness penalty, $\lambda \int \beta_k''(t)^2 dt$, on the functional principal components $\beta_k(t)$, which reduce the mean

squared error in estimating $\beta_k(t)$ especially when the data are sparse and irregular. For fixed values of λ and r, we use the smoothing penalty from *Zhou et al.* (2008) and derive an EM algorithm to estimate $\beta_k(t)$ and D. Furthermore, at each iteration of EM algorithm, we employ a singular value decomposition (SVD) to reparameterize the principal components. The SVD enforces orthonormality of components, so that the assumed diagonal form of D holds in every iteration; and improves convergence speed. Implementation details of this estimation method are discussed in the Appendix E.

Selection of (λ, r): A popular way to select r and λ is by examining the k-fold cross validation score $s(r, \lambda)$, which is the average of log likelihoods of testing data based on the model estimated from training data. Since higher $s(r, \lambda)$ suggests a better model, the appropriate smoothing parameter for a FPCA model with r components is chosen as $\lambda_r^* = \text{argmax}_\lambda s(r, \lambda)$. The appropriate number of principal components r, however, is less straight forward. Because the model with r components is nested within the model with $r + 1$ components, $s(r) = s(r, \lambda_r^*)$ almost always increase as r increases. Therefore maximizing the CV score $s(r)$ does not lead to a parsimonious model. We employ a rule based approach inspired by the "scree plot" (*Johnson and Wichern*, 2007). For pre-specified threshold $b\%$, we select $r^* = \min\{r | \frac{s(r+1)-s(r)}{s(r)} \leq b\%\}$, i.e. the smallest r such that the improvement in CV score by adding one more component is less than the threshold for negligible improvement. Details about the computation of $s(r, \lambda)$, the scree plot and the rule based approach are discussed, with examples, in Appendix F.

Optimization: The algorithms for identifying the optimal sampling schedules that maximizes $\det(I(\eta, D; T))$ or $\det(I(D; T))$ are also worth consideration. Both criteria are complex functions of the sampling schedule T, and may have many local maximum, which makes it difficult to use conventional optimization algorithms, such as Simplex, to find the global maximum. Instead, if there are only a few time points

to choose from, enumeration or a grid search works well. However, if there are many possible choices for the time points, sophisticated optimization methods are needed. We implement a Metropolis-Hastings algorithm (*Metropolis et al.*, 1953; *Hastings*, 1970) that introduces randomness into the optimization and is guaranteed to reach global maximum if the Markov chain has converged. The main idea is that since $\det(I(\eta, D; T))$ and $\det(I(D; T))$ are positive, we can treat them as the probability function (less a constant) of a multivariate distribution of $t_1, \cdots, t_n$. Then the maximum of the criterion function corresponds to the mode of the probability distribution, which the Markov chain will visit with probability one and can be identified easily. More details about the implementations of the algorithm can be found in Appendix G.

3.4 Simulation Study

Simulation Setup. We use simulations to compare the performance of FPCA and parametric mixed model (PMM) in the context of selecting optimal schedules. In order to highlight the similarities and differences of the two approaches, we consider two simulation scenarios:

- Sim A: The between subject variability is induced from random effect in the mean profile.
- Sim B: The between subject variability is not related to the mean profile.

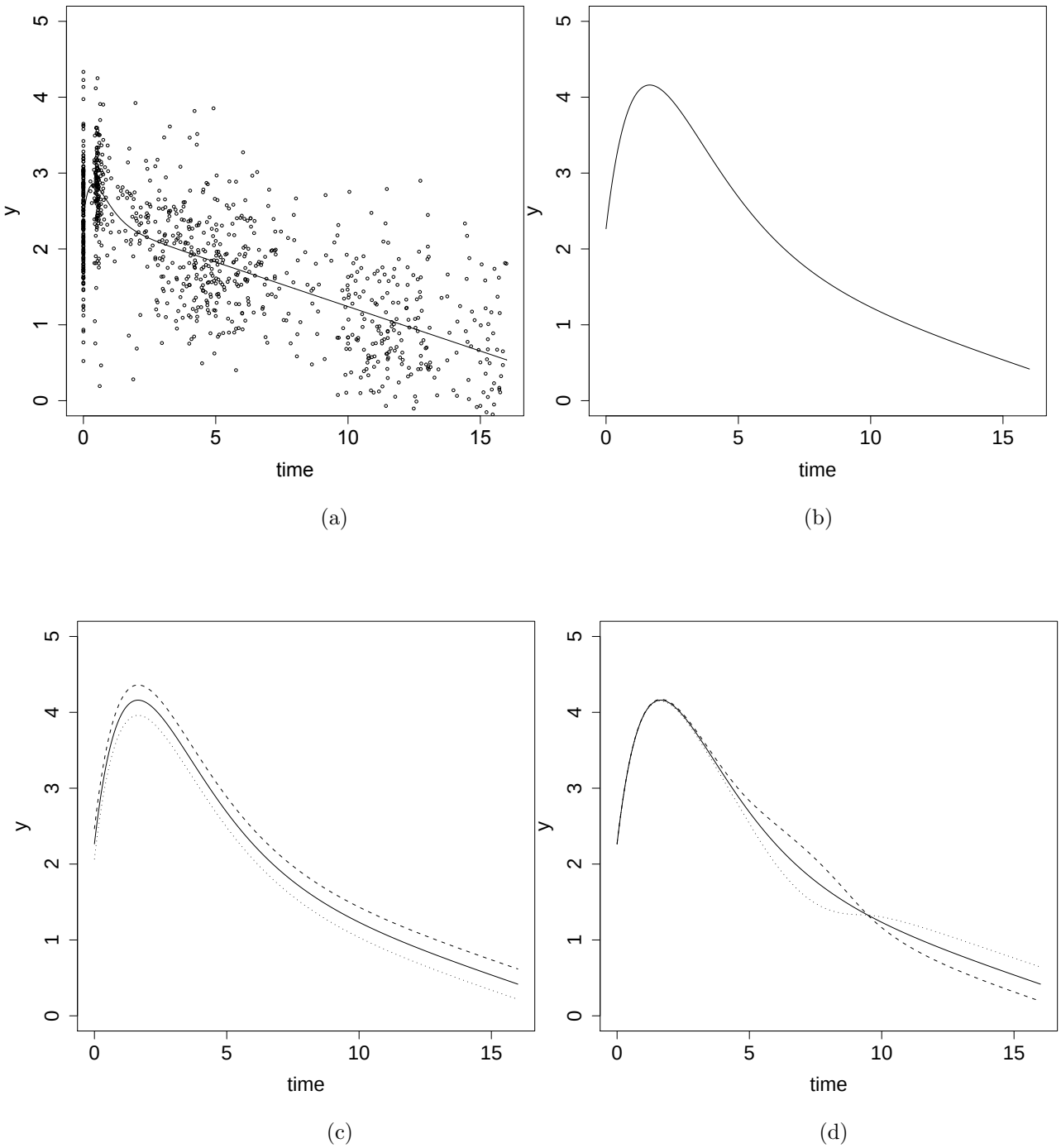
In both Sim A and Sim B, we employ $f(t; \eta) = \eta^0 + \eta^1 t + \eta^2 t \exp(-\eta^3 t)$ as the mean profile, which is motivated by the salivary cortisol profile in the MESA Stress study (Section 5.1). Figure 3.2a presents a scatter plot of a random sub-sample of the MESA Stress data and the estimated mean from this model. Figure 3.2b provides a graphical presentation of $f(t, \eta)$ used in the simulation from 0hr to 16hr. In both simulations we assume the set of the candidate sampling times $S = \{0, 0.5, \ldots, 16\}$

and the candidate schedules S_c consists of schedules with 7 sampling times chosen from S.

In Sim A, we assume the between subject variability is induced by random effects of the parametric mean profile $f(t, \eta)$ as in (3.1) where $\eta_i = (\eta_i^0, \eta_i^1, \eta_i^2, \eta_i^3)^t$ is the subject specific parameter. We assume $(\eta_i^0, \eta_i^1, \eta_i^2)$ follow multivariate normal distribution with mean (η^0, η^1, η^2) and variance $\Sigma = \text{diag}(0.347^2, 0.036^2, 0.3^2)$. We keep $\eta_i^3 = \eta^3$ as fixed such that the formula for the information matrix is exact. Assuming the true model and true parameter is known, the ideal sampling schedule T^A_{ideal} maximizing $\det(I_{true}(\eta, \Sigma; T))$ where $I_{true}(\cdot, T)$ is the information matrix under the tru model. In this scenario, T^A_{ideal} =(0.0, 0.5, 1.5, 2.0, 4.0, 15.5, 16.0) where 1.5 and 2 are close to the peak; 4 is near the inflection point of the curve; 0, 0.5, 15.5 and 16 are near the end points of the time interval.

In Sim B, we assume the between subject variability is not related to the parametric mean profile $f(t, \eta)$. To generate the between subject variability, we employ the reduced rank model (3.3) with $r = 2$ components: $y_i(t_{ij}) = f(t, \eta) + \sum_{k=1}^{2} \alpha_{ik}\beta_k(t_{ij}) + \epsilon_{ij}$ where $\alpha_{i1} \sim N(0, 0.8^2)$ and $\alpha_{i2} \sim N(0, 0.7^2)$ are independent unobserved loadings on the components $\beta_1(t)$ and $\beta_2(t)$. A graphical representation of the components is given by Figures 3.2c Figure 3.2d. $\beta_1(t)$ acts as a random intercept; as a result our ability to estimate the variability associated with $\beta_1(t)$ will not depend on the placement of the sampling times. On the other hand, $\beta_2(t)$ alters the decline of $f_i(t)$ after the peak in a nonlinear fashion with a maximum deviation occurring near $t = 7$. Assuming the true model and true parameters are known, the ideal sampling schedule maximizing $\det(I_{true}(\eta, D; T)$ is T^B_{ideal} =(0.0, 1.0, 1.5, 4.0, 7.0, 15.5, 16.0). In particular, the 4th sampling time $t = 7$ is included to capture the variability of $\beta_2(t)$. The rest of the sampling times are included to capture the important landmarks of the mean profile.

In order to make the simulation as realistic as possible, we pretend we only have access to (simulated) preliminary data instead of the true model. For each of the



(a) (b)

(c) (d)

Figure 3.2: Mean and Variability Structure in the Simulations.
(a) Scatter Plot of a Random Sample of the MESA Stress Data. The timing of the measures is indexed as time since waking up.
(b) Mean Profile for Sim A and Sim B: $f(t;\eta) = \eta^0 + \eta^1 t + \eta^2 t \exp(-\eta^3 t)$ where $\eta^0 = 2.23$, $\eta^1 = -0.12$, $\eta^2 = 3.14$, $\eta^3 = 0.55$.
(c) The mean profile $f(t, \boldsymbol{\eta})$ plus (dash) and minus (dot) $\beta_1(t)$
(d) The mean profile $f(t, \boldsymbol{\eta})$ plus (dash) and minus (dot) $\beta_2(t)$

scenarios, we simulate 1000 data sets with 200 subjects, representing "preliminary data". Each of the 200 subjects was assumed to take 9 samples at t_{ij} $j = 1, \ldots 9$ and t_{ij} randomly chosen from $\{0, 0.5, \ldots, 16\}$. Random noise $\tau_{ij} \sim N(0, 0.1^2)$ is added to t_{ij} to simulate noncompliance of the subjects. Given the times for each subject, outcome data were generated according to the true model from Sim A or Sim B (above). Regardless of the data generating model, we fit PMM and FPCA to the preliminary data. nlme() function in R is used to fit PMM (Appendix H) and the smoothing FPCA outlined in Section 3.3 and described in the Appendix E was used to fit FPCA. The smoothing parameter λ and the number of principal components r were selected by cross validation as described in Section 3.3. In particular, we use the rule based approach to select r and the threshold for negligible improvement is set to $b\% = 1\%$. We identify optimal schedules by the Metropolis-Hastings algorithm described in Section 3.3. The optimization criterion is $\det(I_{est}(\eta, \Sigma; T))$ or $\det(I_{est}(\eta, D; T))$ using model parameters estimated from the preliminary data.

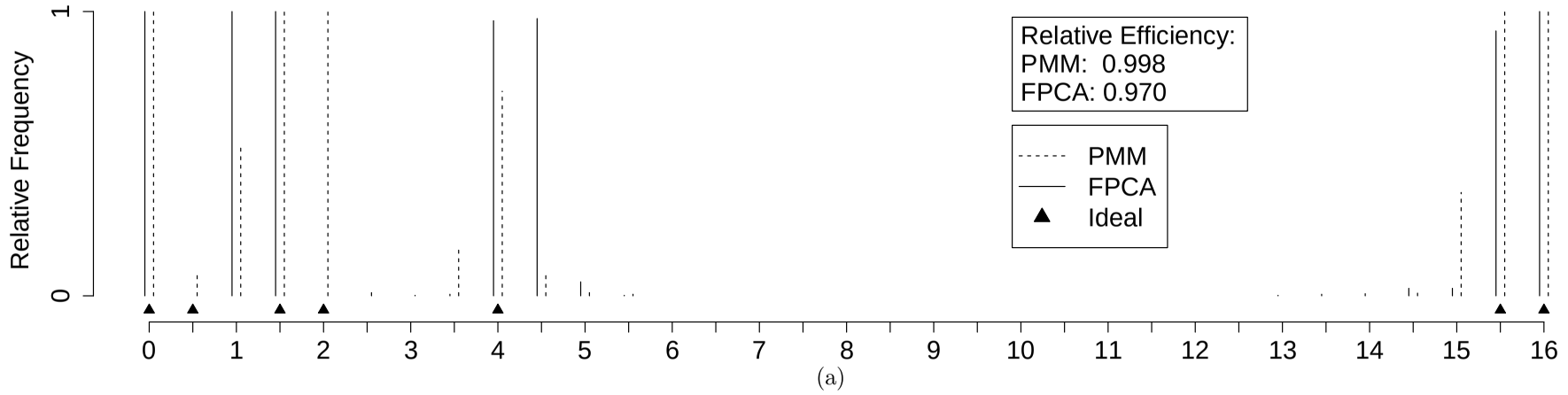
We evaluate the performance of the sampling schedules selected by PMM and FPCA in two ways. First, we investigate whether these schedules are in agreement with ideal schedules in each scenario by tabulating the relative frequency of each sample time selected. Second, we use the efficiency relative to the ideal schedules as a numerical benchmark. Relative efficiency (*Atkinson et al.*, 2007) is defined as $\frac{\det(I_{true}(\cdot;T_{candidate}))^{1/p}}{\det(I_{true}(\cdot;T^Q_{ideal}))^{1/p}}$, where $Q = A$ or B represents the two simulation scenarios; $T_{candidate}$ and T^Q_{ideal} are candidate and ideal schedules respectively; p is the number of parameters under the data generating model. Because the ideal schedule maximizes the optimization criterion under the true model, the relative efficiency is ≤ 1, with higher values indicating better schedules.

Simulation Results. The FPCA algorithm converged for all preliminary data sets in Sim A and Sim B. However, the nlme() function for fitting PMM only converged for 858 preliminary data sets in Sim A and 528 in Sim B. We restrict our comparison

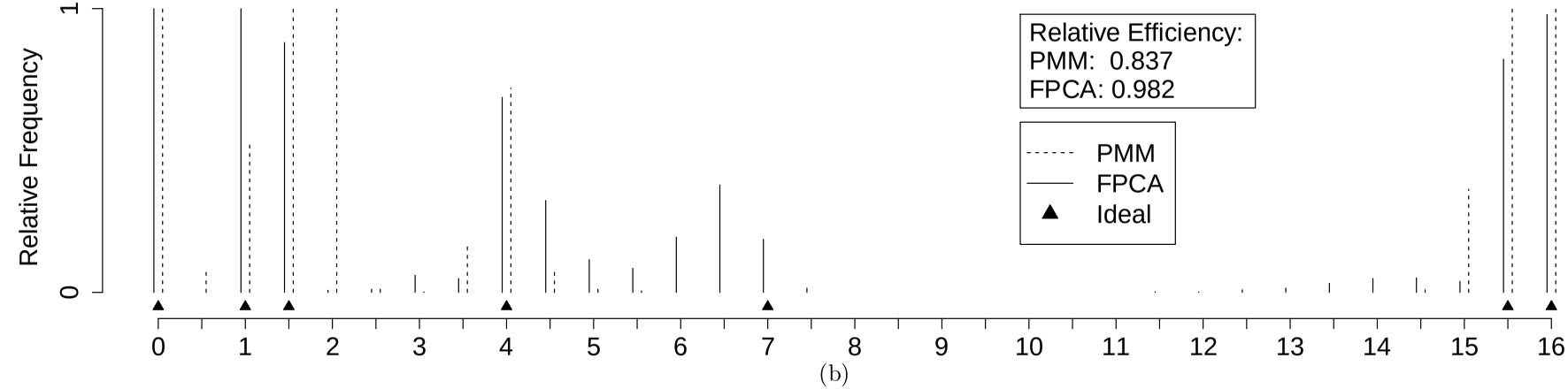
to the simulations where both methods converged.

For Sim A (Figure 3.3a), both FPCA and PMM provide schedules that are similar to the ideal schedule. There is perfect agreement between the two methods at the end points 0 and 16 as well as the peak time 1.5. Furthermore, at least one of 3.5 or 4, near the inflection point of the profile (Figure 3.2b), appears in all the optimal schedules. Both methods assign exactly one sampling time within the period (0, 1.5) where the curve peaks, although they have slightly different preferences: FPCA always picks 1 while PMM places more weight on 0.5. Of all the seven sampling times, there is only one consistent discrepancy between the two methods. FPCA always selects 4.5 while PMM always selects 2. However, 4.5 and 2 are adjacent to 4 and 1.5, both of which are already selected by the two methods. The average relative efficiency is 0.998 for PMM and 0.970 for FPCA. Both methods can provide reasonably good schedules, although the PMM approach has a slight advantage.

For Sim B (Figure 3.3b), the ideal schedule includes a sampling point at time $t = 7$, which is the best location to capture the variability associated with the second component $\beta_2(t)$. The optimal schedules by FPCA place at least one sample in the interval of $[6,\ 8]$ with probability 0.70. By comparison, the PMM approach never includes any sampling time in the $[6,\ 8]$ time interval. For the rest of the sampling times, FPCA and PMM have good agreement, primarily influenced by the need to capture the mean structure. The relative efficiency is 0.837 for PMM and 0.982 for FPCA. The efficiecy gain of FPCA is primiarily due to FPCA's capability in including the sampling times critical for detecting the variance components. Sim B demonstrates the FPCA approach is more applicable than the PMM in the scenarios where the temporal variability pattern is not induced by the random effects on the mean profile.



(a)



(b)

Figure 3.3: Relative frequency for sampling time is selected as part of the optimal schedule: (a) Results for Simulation A; (b) Results for Simulation B

FPCA ——

PMM - - - -

▲ denotes the ideal sampling times for each simulation.

Relative efficiency is compared against the ideal sampling schedules. The higher and closer to 1, the better the efficiency.

3.5 Application

3.5.1 Design for Salivary Cortisol Studies

The use of salivary cortisol as a biomarker of stress, which exhibits a nonlinear diurnal pattern through the length of the day (Figure 3.2a), is increasingly common in epidemiology studies (*Adam and Kumari*, 2009). Salivary cortisol can be assayed objectively using standardized techniques, and it is less likely to suffer from biases due to individual's interpretations of stress questionnaires. We compare the PMM and FPCA methods in the design for salivary cortisol studies.

As preliminary data for the design, we use data from the Stress ancillary study of the Multi-Ethnic Study of Atherosclerosis (MESA Stress). MESA Stress (*Hajat et al.*, 2010) is an epidemiological study that examines the role of stress as a contributor to a range of precursors of cardiovascular disease (CVD). We use data from 800 individuals who collected 6 samples for 3 days. Figure 3.2a presents a scatter plot of a random sub-sample of this data.

The design goal is to identify the sampling schedule that maximizes the precision to measure the mean profile and the variability. We consider sampling times $S = \{0, 0.5, 1, \ldots, 16\}$, i.e. the sampling period begins at wake up time, and ends with bed time, $t = 16$h (similar to the MESA Stress study). The set of candidate schedules S_c consists of the schedules with 6 sampling times from S.

Preliminary analysis suggest $f(t;\eta) = \eta_0 + \eta_1 t + \eta_2 t \exp(-\eta_3 t)$ is a suitable mean profile for salivary cortisol (*Stroud et al.*, 2004). We use this mean profile in both PMM and FPCA approaches. For PMM, we fit model (3.1) and use $\det(I(\eta, \Sigma; T))$ as the optimization criterion. For FPCA, we fit the reduced rank model (3.3) and use $\det(I(\eta, D; T))$ as the optimization criterion. The values of $r = 3$ and $\lambda = 2000$ were determined by 10-fold cross validation. Because there are only $\binom{33}{6}$ candidate schedules, we enumerate all of them to identify the optimal schedule.

The 10 schedules with the highest criterion values obtained from each method are given in Table 1. All schedules include four common sampling times: 0, 0.5, 1 and 16 hour after wake up. Wake up and 16 hours are the beginning and the end of the time period under investigation, and hence are naturally included. Since salivary cortisol curves exhibit high curvature between 0 and 1 hours since wake up, 0.5 and 1 are also included in the sampling schedules. The difference between the two methods is revealed in the time period between 2 and 16. PMM will place the remaining two sample times close to both ends of this time period and none in between. This is because in the 2-16 hrs time period, the exponential term $\eta_2 t \exp(-\eta_3 t)$ is essentially zero and the mean profile is dominated by the linear term $\eta_0 + \eta_1 t$. Since the linear term forces the variability to be higher at the end points (Section 2), more sampling times are placed at both end points. However, FPCA detects a different temporal pattern for the variability and places the sampling times at around 4 and 10 hours after wake up. We believe the FPCA schedule is better because it not only covers covers almost all the sampling regions of the PMM schedules, but also includes new sampling region that is not included in PMM schedules. This new sampling region can be discovered mainly because the FPCA approach places less restriction on the variance structure.

3.5.2 Urinary Progesterone Study

Urinary progesterone is an important biomarker of reproductive health. (*De Souza et al.*, 2010). In order to capture the variations of urinary progesterone, it is not uncommon to collect samples everyday during the entire menstrual cycle of the participant (*Waller et al.*, 1998). Densely sampled data points are helpful in reconstructing the entire progesterone profile, but are difficult and costly to implement in large scale studies. Thus a design question naturally arises: which simplified sampling schedule, with fewer sampling times, adequately captures the variation of the progesterone

Time1	Time2	Time3	Time4	Time5	Time6
0	0.5	1	3.0	15.5	16
0	0.5	1	2.5	15.5	16
0	0.5	1	3.5	15.5	16
0	0.5	1	3.0	15.0	16
0	0.5	1	2.5	15.0	16
0	0.5	1	3.5	15.0	16
0	0.5	1	3.0	14.5	16
0	0.5	1	2.5	14.5	16
0	0.5	1	4	15.5	16
0	0.5	1	3.0	15.0	16

(a) Parametric Mixed Model

Time1	Time2	Time3	Time4	Time5	Time6
0	0.5	1	4	10.5	16
0	0.5	1	4	10.0	16
0	0.5	1	4.5	10.5	16
0	0.5	1	4.5	10.0	16
0	0.5	1	4	9.5	16
0	0.5	1	4	11	16
0	0.5	1	4.5	9.5	16
0	0.5	1	4.5	11	16
0	0.5	1	4	9.0	16
0	0.5	1	4.5	9.0	16

(b) FPCA

Table 3.1: The ten best sampling schedules chosen by parametric mixed model and FPCA based approaches

profile? We answer this question using the FPCA approach.

As preliminary data, we use the urinary metabolite progesterone data from *Brumback and Rice* (1998), which were collected as part of early pregnancy loss studies. The data set contained progesterone profiles of 91 menstrual cycles from 51 women with healthy reproductive function. For illustration, we randomly select only one cycle from each of the women who contributed data of multiple cycles to ensure independent data across cycles. As is standard practice in endocrinological research, progesterone profiles were aligned by the day of ovulation (day=0) and then truncated at each end to present curves of equal length (24 days).

Figure 3.4a shows a scatter plot of the progesterone measures in the data set, with a local polynomial smooth summarizing the overall trend. Although the trend appears to resemble an horizontally stretched 'S' shape, common parametric functions, such as the logistic function, fail to provide a satisfactory fit. *Brumback and Rice* (1998) suggested that a nonparametric spline model is a better choice for the population mean profile.

In the absence of a mechanistic model for the mean, we consider study designs that maximize the efficiency of measuring the variability of the progesterone profile. We first center the data at each time point with the mean estimated by a local polynomial smoother. Then we perform FPCA on the centered data to obtain the functional principal components. The number of components $r = 3$ and the smoothing parameter $\lambda = 1000$ were determined by 10-fold cross validation with $b = 1\%$ as the threshold. The three principal components are shown in Figure 3.4b-3.4d. The first component accounts for 53.2% of the variance and can be interpreted as a constant cycle level deviation from the mean. The second and third component accounts for 36.6% and 1.4% of overall variance. These two components capture the local deviations at various times within a cycle.

Since three components are sufficient to characterize the progesterone data, sam-

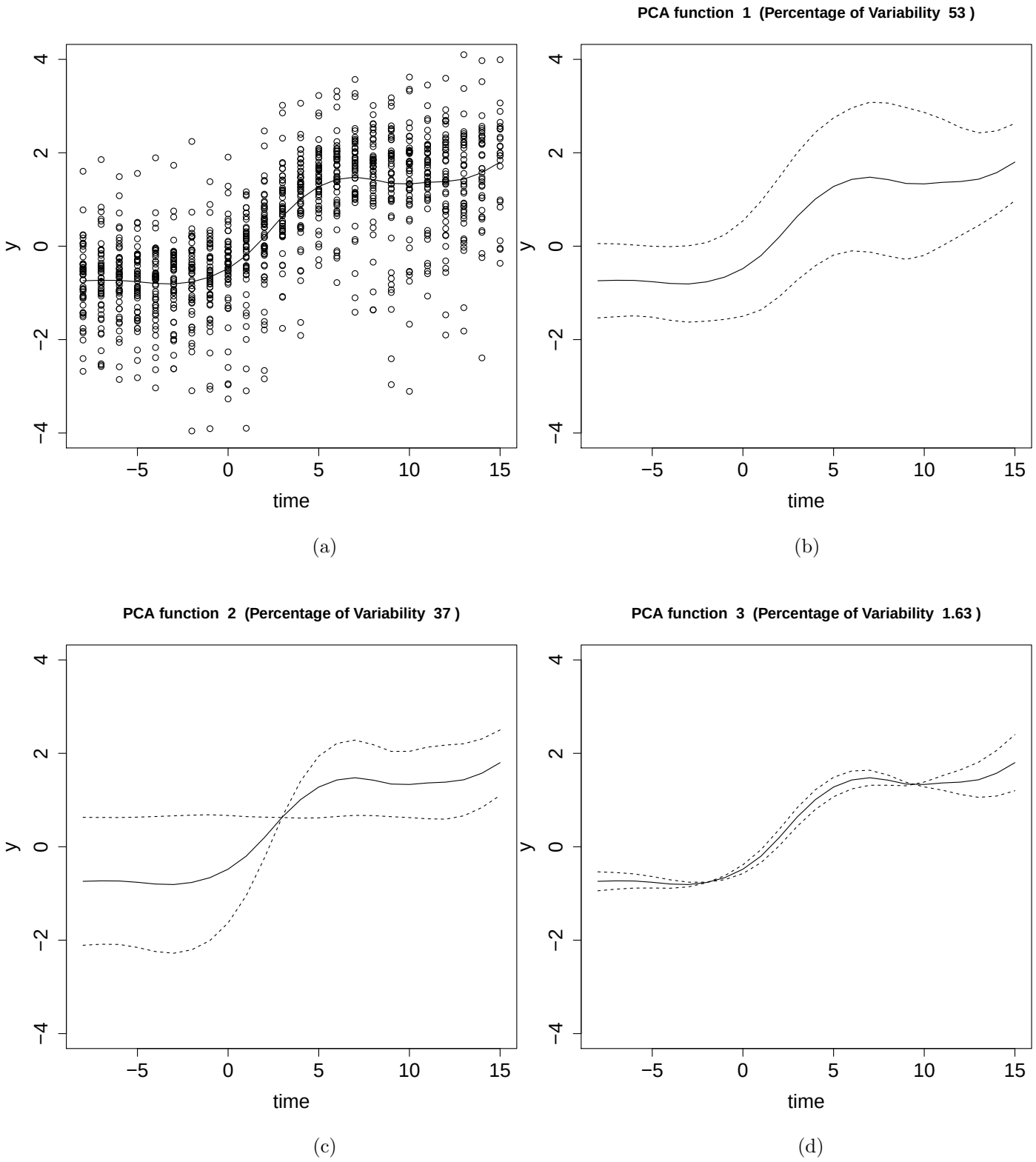


Figure 3.4: (a) Scatter Plot of Progesterone Data and Local Polynomial Smoother; (b)-(c) Functional principal components of the urinary progesterone data. The solid line indicates the mean profile. The dash lines represent the mean profile +/- one standard deviation of the functional principal components.

pling schedules with at least $m = 3$ sampling times will adequately estimate the variance $D = diag(d_1, d_2, d_3)$ of the principal components. We consider $S = \{-8, -7, \ldots, 15\}$ as the admissible sampling times and the set candidate schedules S_c consists of schedules with 3 sampling times from S.

Table 3.2 lists the 10 best schedules with 3 samples each and the associated values of $\det(I(D;T))$. These schedules offer similar level of efficiency when it comes to measuring the variability across the cycles (similar $\det(I(D;T))$. In addition, the sampling times of these schedules also exhibit a clear pattern: $t = 15$ is considered to be important by all schedules; the rest of the sampling times are clustered in two intervals $[-8,\ -4]$ and $[4,\ 6]$. The choice of the sampling times can be intuitively understood if we refer to principal components (Figure 3.4b-3.4d). The deviation from the mean of the second and third components are relatively higher at $[-8,\ -4]$ and $[4,\ 6]$ and $t = 15$. Collecting samples in these locations maximizes our ability to identity the variability associated with these components. The first component deviates from the mean relatively uniformly during the cycle; hence it has little impact on the choice of the sampling times even if it accounts for a substantial proportion of the overall variability.

Furthermore, we evaluate the performance of the optimal schedules derived from FPCA in the context of detecting the association between progesterone and a covariate: conceptive status. We consider leave-one-out prediction rate of the schedules for predicting the conceptive status. Prediction rates were computed by a logistic regression model with conceptive status as the outcome and progesterone levels measured at the days indicated by each of the derived schedules as the predictors. The leave-one-out prediction rate of the ten best sampling schedules obtained by FPCA are between 92%-94% (the fifth column of Table 3.2), which suggests that progesterone values collected with these schedules are highly predictive of conceptive status. Furthermore, we rank all candidate schedules in S_c by the leave-one-out prediction rate.

Time1	Time2	Time3	det($I(D)$) (10^{-8})	Leave One Out Prediction Rate	Ranking among All Candidate Schedules
-8	5	15	6.95	92%	94%
-8	6	15	6.92	94%	96%
-7	5	15	9.87	90%	89%
-7	6	15	6.85	90%	89%
-8	4	15	6.83	92%	94%
-6	5	15	6.81	90%	89%
-6	6	15	6.79	94%	96%
-5	6	15	6.78	92%	94%
-4	6	15	6.78	92%	94%
-5	6	15	6.78	94%	96%

Table 3.2: The ten best sampling schedules for urinary progesterone chosen by the FPCA approach
The leave-one-out prediction rate is for the prediction of conceptive status.
The ranking in the last column is based on the leave-one-out prediction rate.

The sampling schedules we obtained perform better than 89%-96% of all candidate schedules (the last column of Table 3.2). Sampling urinary progesterone at the times selected by our FPCA based approach are among the best times to study the association between conceptive status and progesterone levels. It should be noted that when we use FPCA to derive these optimal schedules we do not make use of any information regarding the conceptive status. With the help of FPCA we reconstructed the temporal pattern of the variability, and subsequently identified time points critical for detecting the association between the longitudinal profiles and covariates.

3.6 Discussion

We propose an approach to longitudinal study design that optimizes estimation of the mean and between-subject variability. Our approach employs a semiparametric model that characterizes the mean profile and the variability separately. We use functional principal component analysis (FPCA) to derive a parsimonious and flexible representation of the temporal pattern of the variability. Smoothing of the princi-

pal components is incorporated to enhance estimation stability given the potentially unbalanced and irregular sampling of the preliminary data. The population mean can be modeled with known parametric functions or with splines. Following the existing literature, we employ D-optimality for the mean and variability parameters as the optimization criterion. Simulations suggest that if between subject variability is induced by random effects in a parametric mean profile, the FPCA approach and existing methods based on parametric mixed model (PMM) lead to equally competent sampling schedules. For more flexible variability structures, however, the FPCA approach is superior. We apply the new approach to two real world examples. In the first example we show that the FPCA approach identifies new sampling regions for measuring the mean and the variability of salivary cortisol profiles that are not discovered by the PMM approach. In the second example, the mean profile of progesterone in menstrual cycles is modeled by splines and the PMM approach is no longer applicable. We employ the FPCA approach to identify optimal schedules for capturing the variability of the profiles and show that schedules we obtained are also highly predictive of conceptive status. In both simulations and real world examples, we show that the FPCA approach is more robust flexible for capturing the variability of the process.

Under the framework of FPCA, $Var(f_i(t))$ is modeled by the functional principal components $\beta_k(t)$'s and the variance of the principal components $D = \text{diag}(d_1, \ldots, d_r)$. While both terms are involved in the optimization criterion $\det(I(D))$, we only focus on the estimation of D in forming the criterion. The rationale behind such decision is that the individual deviation of subject i is captured only by the subject-specific loadings α_{ik} but not the functional principal components $\beta_k(t)$ which are shared by all subjects. Therefore we focus on subject specific loadings α_{ik} when our goal is to maximize our ability to detect individual deviations. At the population level, the distribution of α_{ik} is determined by the variance d_k; thus focusing on the loadings im-

plies focusing on the estimation of d_k. Because the number of components is expected to be low, only a few samples would need to be measured in future studies.

FPCA is a versatile modeling tool and its application could be seen in many areas of statistics. For example, *Fedorov and Hackl* (1997) uses FPCA to model correlated data and consider the design that minimizes prediction errors. But to our knowledge, our proposed method is the first to employ FPCA to derive optimal sampling schedules for the estimation of the between subject variability.

Besides parametric mixed model, mean and variance models can be employed for longitudinal data (*Davidian et al.*, 1988). Unlike the parametric mixed model, the mean and variance model only characterizes the marginal distribution of the data. As a result, it provides no insight into the between-individual variability. Therefore, we do not consider the mean and variance model in our comparison.

There are some limitations to our approach. The FPCA approach requires preliminary data with dense sampling. In the cortisol example, each individual collects very few samples in the preliminary data, but dense sampling can be achieved by including more individuals. In the progesterone example, each woman already has samples every day during the menstrual cycles so we only need a small number of women for the preliminary data. Furthermore, we need to check the normal assumptions of the principal component loadings α_{ik}. We observed that the loadings for the first few components exhibited small deviations from normality at the tail distribution. Because we are most interested in the temporal pattern of the majority of variability from the first few components, these small deviations might not affect our results significantly. Nevertheless, we recommend checking the distributions of the estimated loadings.

The framework of determining optimal schedules based on FPCA can be extended in many ways depending on the design problems. One direction will be to obtain the sampling schedules for estimating $\beta_k(t)$, which is modeled by smoothing spline. One

can consider a surrogate for the smoothing spline, for which the optimal schedule is known, such as a piecewise linear spline (the optimal schedule is simply the set of the knots). Another extension could be the sampling schedules for a study incorporating multilevel sampling. In the second example, repeated cycles from the same woman were excluded. To incorporate such data, one could employ a hierarchical FPCA model and derive the information matrix for the grand mean and variability parameters at multiple levels. Furthermore, our method currently consider only one optimal schedule in the design due to practical constraint of large scale epidemiology studies. Nevertheless, it is also interesting to consider multiple optimal schedules in the design since it could potentially improve the efficiency of the design (*Mentré et al.*, 1997). Lastly, deriving theoretical results of the FPCA approach, such as its asymptotic properties and alternative methods to select the number of principal components, would be of interest.

CHAPTER IV

Design for Studies Involving High Dimensional Features and Other Covariates

4.1 Introduction

Rapid technological advances and bench scientific findings have given rise to the use of high throughput data in clinical studies, where measurements on hundreds or even millions of biomarkers are gathered from each subject (*Elaine R.*, 2008; *Schuster*, 2008). Screening analysis of association between biomarkers and disease status enables researchers to identify promising biomarkers that are differentially expressed across disease groups and are useful to develop classifiers for disease prognosis. Classifiers are one of the key components for realizing the promise of personalized medicine. For instance, clinicians can design an effective treatment plan for a patient utilizing the prediction of validated classifiers, given the measurement of biomarkers from this very patient (*Hamburg and Collins*, 2010). Several approaches to deriving classifiers from high dimensional biomarkers have been developed in the literature, and when applied to real world experiments, some promising results have been reported (*Clarke et al.*, 2008; *Simon*, 2008; *Wang et al.*, 2008). The utility of high throughput data in clinical practice hinges on the availability of well developed and validated classifiers, which in turn requires proficient study design to generate training and validation

datasets. In particular, determining the sample size needed for the deviation of a powerful classifier is a central design question.

Sample size determination has been essential in the design of clinical studies. Every clinical study needs it to begin with. As a matter of fact, this task is traditionally carried out in the context of hypothesis testing, in which type I error (or size of a test) and type II error (or power of a test) are treated as target objectives for the sake of achieving certain optimality. However, these concepts of errors no long exist in the setting of classification analysis, and thus a new statistical framework of design needs to be established. The primary objective of this paper is to provide systematic development of design and sample size calculation for classification analysis involving high-dimensional features. Our software produced from this work will furnish a timely and useful toolbox to clinical studies.

This research work has been primarily motivated from our collaborative project "NEPTUNE" Nephrotic Syndrome Study Network administrated by multiple principal clinical investigators in the Division of Nephrology at University of Michigan School of Medicine. The NEPTUNE consortium, funded with 10.25 million dollars by NIH, University of Michigan and NephCure Foundation, is an unprecedented research endeavor to conquer this aggressive rare renal disease. It's the largest ever committed to collectively by nephrologists in USA and Canada to study molecular mechanisms for rare renal diseases, including Membranous Nephropathy (MN) and Focal or Segmental Glomerulosclerosis (FSGS)/Minimal Change Disease (MCD). The consortium is designed to address multiple scientific goals, one of which is to extract important tissue-based mRNA biomarkers to classify patients into different risk groups and predict their clinical outcomes such as remission status (i.e. remission versus no remission). One of the co-authors (Song) was in charge of designing the NEPTUNE consortium (*Gadegbeku et al.*) using a somewhat ad-hoc approach, and a comprehensive generalization of that original method is very appealing, because such

design methodology is needed in practice. This motivated us to extend and rigorously justify the design framework in general settings. For example, a new study was recently proposed by a clinician at University of Michigan Kidney Transplantation Center, which aimed to predict the graft survival of a patient after kidney transplant, which is an important outcome for measuring treatment effectiveness. In this study, the long term prognosis outcome after kidney transplant is divided into two groups: stable and rejected. Overall the study being conceived will be carried out in two parts (Figure 1).

In Part I, a commercial array allows us to collect 108 proteomics biomarkers from each patient in both stable and rejected groups. Then classifiers for graft survival will be constructed using the microarray data. The investigators suggest that among these proteins, according to literature, approximately 10 proteins are likely to be informative for predicting graft survival. For each informative biomarker, the difference in mean expression between stable group and rejected group should be at least 0.8 of the standard deviation, whereas for the non-informative biomarkers, the difference in mean expression is 0. In Part II, additionally, the investigators will consider adding a new source of clinical predictors such as measures of lab tests, and/or patients' demographic characteristics into classifiers in the hope of improving classification performance.

In this paper, we focus on the development and implementation of classification design methodologies in clinical studies involving high-dimensional features, with the number of the features larger than the sample size. In Section 2, we begin with a brief review on the existing methods related to our design problem and then discuss our strategies to approach the problem. Section 3 presents the core material of this paper, including (i) the framework of classification design with high-dimensional features; (ii) implementation of cross-validation scheme to evaluate PCC; (iii) a new simulation method utilizing order statistics to efficiently evaluate PCC of high criti-

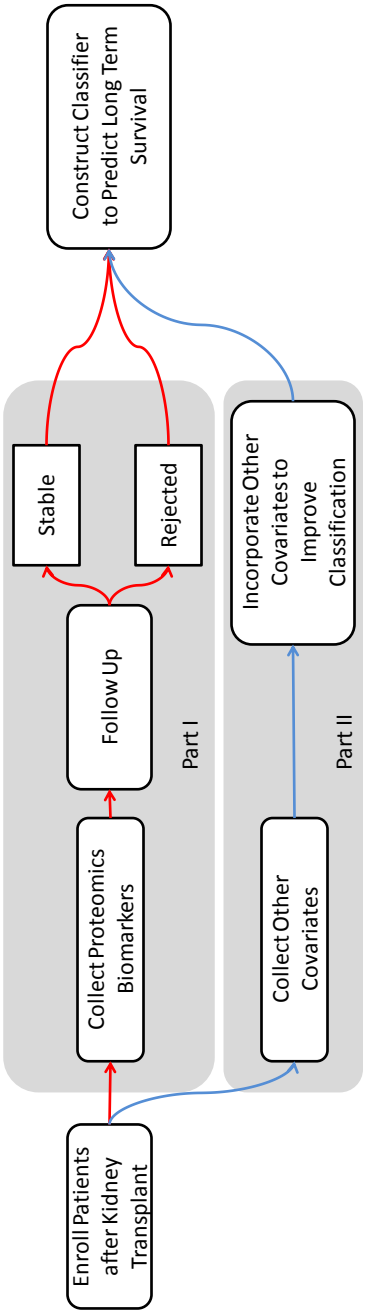


Figure 4.1: A Study for Predicting Long Term Survival after Kidney Transplant

cism thresholding classifier; (iv) an inequality to establish the upper and lower bounds of achievable PCC when two sources of biomarkers are added simultaneously to enhance classifier in the study. In Section 4, we conduct simulation studies to compare the relative efficiency of three design strategies in terms of potential bias in estimated PCC and required sample size. In Section 5 we illustrate the use of these design methods in a practical example of predicting graft survival after kidney transplant. In Section 6, we discuss R software implementation of these new methods. Finally, we conclude this paper with discussion and recommendation for high-dimensional classification design.

4.2 Roadmap and Strategies

We approach the design question by three integral steps: (i) choosing a reasonable objective to calibrate the sample size; (ii) building an efficient classifier for high-dimensional features; (iii) improving classification performance through incorporating new sources of features. We will discuss each of these steps in detail in the rest of the paper.

In the first step, a reasonable objective is key to develop a proficient design for sample size calculation. As mentioned above, conventional Type I error and power are used in hypothesis testing to determine sample size. Such approach has been considered in designs for high-dimensional classification, based on a global null hypothesis of no biomarkers being differentially expressed between groups (e.g. *Hwang et al.* (2002)). However, such hypothesis setup does not directly measure the performance of the classifier. In our views, a more desirable objective function for the study design of high-dimensional classification analysis is the probability of correct classification (PCC) (*Mukherjee et al.*, 2003) since PCC enables us to directly measure the performance of classifier. *de Valpine et al.* (2009) proposed a hybrid method based on simulation and approximation of the PCC formula, which is shown to speed up the

computation of PCC.

To quantify optimality in design, the upper bound of PCC needs to be determined. *Dobbin and Simon* (2007) proposed an interesting procedure to construct an ideal classifier with the availability of the full knowledge about the underlying data model, under which the PCC of the ideal classifier is computed. This optimistic PCC corresponds to the PCC of a study with infinite sample size and therefore gives an upper bound of PCC. In other words, for a practical design with finite sample size, the actual sample size will be calibrated against the such best PCC minus a certain tolerance specified by clinicians.

The second step is to choose an appropriate classification method. For the so-called $p \gg n$ scenarios, namely the number of biomarkers is much larger than the total number of observations, many conventional methods, such as Fisher linear determinant analysis, are not directly applicable. One way to overcome such difficulty is to utilize variable selection strategies in the building of classifiers. The resulting classifier will only use biomarkers whose p-values, obtained by, for example, two-sample comparison test, are lower than a predefined threshold level. Thus, the PCC of this classifier heavily depends on a choice of the threshold for feature selection. *Dobbin and Simon* (2007) proposed a design method to evaluate the best PCC, in which they assume that there exists the optimal threshold maximizing the PCC. Unfortunately, this assumption will not hold in real-world design settings because there exists uncertainty in connection to the procedure of the determining threshold values. Dobbin and Simon's strategy will be apparently higher than a PCC obtained in actual analysis stage, and thus results in an underestimated sample size. Therefore, it is desirable to incorporate the actual thresholding procedure into the design framework so that consistency between design and analysis is enforced, which ensures the needed sample size for the study.

One popular thresholding procedure is cross-validation (CV). In spite of being con-

ceptually simple, the CV approach is computationally demanding. Recently, many advance feature selection methods have been proposed, including regularization approaches such as lasso (**?**) and elastic net (**?**), and specific approaches developed for biomarker selection such as **?** and **?**. In this paper, we are particularly interested in higher criticism thresholding (HCT) method proposed by *Donoho and Jin* (2009) because of the nice properties such as well justified asymptotic theory concerning feasibility and power of classification. HCT determines the threshold value by maximizing the deviation from the null empirical process indexed by ordered p-values. HCT uses the rare-and-weak model framework, where classification problems asymptotically fall into one of two classes: (i) the infeasible class, referring to a scenario where signals are so weak and rare that no linear classifier exists to outperform a naive random classifier; and (ii) the feasible class, referring to the case where the PCC of HCT approaches to 1 asymptotically as the number of biomarkers goes to infinity.

In the third step, we consider evaluating the benefit of adding additional sources of features to classification analysis. This is an important practical problem, as many practitioners believe that using more features in classification can improve the performance limitlessly. Typical features include clinical predictors such as measures of lab test, and/or patients' demographic characteristics. Some researchers have investigated the benefits of feature augmentation in terms of the receiver operator curve (ROC) (**?????**). However, the current literature has not provided any rigorous answers concerning the amount of potential PCC gain resulted from feature augmentation, and under which scenarios we expect to achieve the highest PCC gain. These questions are critical and need to be carefully addressed in the classification design analysis before substantial amounts of funds and time are committed to data collection.

4.3 Methods

4.3.1 Setup

We begin with a brief review of a general framework for the design of studies for classification proposed by *Dobbin and Simon* (2007). Suppose the study population can be divided into two groups: Group +1 and Group -1. In our motivating example, Group +1 consists of subjects who have a stable graft condition after the kidney transplant and Group -1 consists of the subjects who reject the transplant. To plan a future study, we presumably collect training data $D = \{(\boldsymbol{x}_i, y_i)|i = 1, \dots n\}$ from n subjects; $y_i = \{+1, -1\}$ is the group label for subject i; the prevalences of Group +1 and Group -1 are $P(y_i = 1) = \pi_1$ and $P(y_i = 1) = 1 - \pi_1$, respectively. Let $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^t \in R^p$ be a high-dimensional vector of features for subject i. For instance, $\boldsymbol{x}_i$ could be the vector of 108 proteomics biomarkers in the study of predicting graft survival after kidney transplant.

In the step of data processing, we standardize all features to ensure that their standard deviation is 1. Furthermore, we center features around their means and align the signs of features with group labels, such that $E(\boldsymbol{x}_i|y_i = 1) = \boldsymbol{\mu}$ and $E(\boldsymbol{x}_i|y_i = -1) = -\boldsymbol{\mu}$ with $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^t$ and $\mu_j \geq 0$ for $j = 1, \dots, p$. Assume that all features are conditionally independent of each other and they follow the multivariate normal distribution within each group:

$$\boldsymbol{x}_i|y_i \sim \begin{cases} N(+\boldsymbol{\mu}, I) & y_i = +1 \\ N(-\boldsymbol{\mu}, I) & y_i = -1 \end{cases}.$$

The effect size vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^t$ presents signal strengths of features for the purpose of classification. Higher values of μ_j suggest better separation between two groups in feature j and consequently feature j is more likely to be important for classification. Here notice that the dimension of $\boldsymbol{\mu}$ is very high, ranging from hundreds

to millions depending on an actual problem. To overcome the curse of dimensionality, sparsity is a common assumption. That is, it is believed that there are only a small number of biomarkers that are differentially expressed between groups, i.e. their effect sizes μ_j are not zero. These features are considered essential to construct a classifier and the number of these features is m ($\ll p$). The other $p-m$ features, however, are noise and not differentially expressed, with $\mu_j = 0$. For the ease of exposition, we may reorder the features such that the first m features are important, and the resulting $\boldsymbol{\mu}$ has the form $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_m, \underbrace{0, \cdots, 0}_{p-m \text{ copies}})^t$. In general, the values in $(\mu_1, \cdots, \mu_m)^t$ are unknown at the design stage. In this paper, we assume we can obtain a lower bound μ_0 for $(\mu_1, \cdots, \mu_m)^t$ based on prior research results or a certain hypothesis and thus and replace $\boldsymbol{\mu}$ by $\boldsymbol{\mu} = (\underbrace{\mu_0, \cdots, \mu_0}_{m \text{ copies}}, \underbrace{0, \cdots, 0}_{p-m \text{ copies}})^t$. Clearly such simplification will lead to a conservative estimate of the sample size, which is acceptable in practice when no good pilot studies are available to estimate $\mu_1, \cdots, \mu_m$ satisfactory.

A linear classifier is then employed for the classification problem, and their performance is evaluated. Constructing a linear classifier is equivalent to using training data D to derive a certain weighting scheme G that allocates weights $\boldsymbol{w} = (w_1, \cdots, w_p) \in R^p$, denoted by $\boldsymbol{w} = G(D)$. The classification rule for a new subject is: (i) if $\boldsymbol{w} \cdot \boldsymbol{x}_i \geq 0$, subject i is assigned to Group 1; (ii) if $\boldsymbol{w} \cdot \boldsymbol{x}_i < 0$, subject i is assigned to Group -1. Here $\boldsymbol{a} \cdot \boldsymbol{b}$ denotes the inner product of two vectors.

4.3.2 Objective Function

The performance of a given classifier may be evaluated by the probability of correct classification (PCC). We employ PCC in this paper as the target objective function for sample size determination. In the setting of two groups, the PCC is expressed as

$$\begin{aligned} PCC &= P(\text{Subject } i \text{ is classified to Group 1} | y_i = 1) \times P(y_i = 1) \\ &\quad + P(\text{Subject } i \text{ is classified to Group -1} | y_i = -1) \times P(y_i = -1). \end{aligned}$$

Notice that the probability of Subject i being classified to Group 1 given $y_i = 1$ is the sensitivity while the probability of Subject i being classified to Group -1 given $y_i = -1$ is the specificity. The PCC is a weighted average of these two important operating characteristics of a classifier. Given the weight $\boldsymbol{w}$, it is easy to derive the PCC of a linear classifier as $PCC(\boldsymbol{w}; \boldsymbol{\mu}, n) = \Phi(\frac{\boldsymbol{w}\cdot\boldsymbol{\mu}}{\sqrt{\boldsymbol{w}\cdot\boldsymbol{w}}})$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

In the best scenario, all important features are included in the classifier. This leads to the oracle classifier and the associated PCC is $PCC_{oracle} = \Phi(\frac{m\mu_0}{\sqrt{m}}) = \Phi(\sqrt{m}\mu_0)$. In the worst scenario, none of the important features are included in the classifier, leading to a random classifier and the resulting PCC is $PCC_{random} = 0.5$. Clearly, a practically achievable PCC lies in between the best and worst scenarios, and its exact value depends on how much information can be extracted from the training data. Thus, we can set a PCC target as $PCC_{target} \in (PCC_{random},\ PCC_{oracle})$.

Let $PCC(n)$ be the PCC of a study with n subjects. Then the sample size requirement is defined as the smallest n such that $PCC(n) \geq PCC_{target}$. If the inverse function of $PCC(n)$ can be derived, then the sample size could be easily determined by $n \geq PCC^{-1}(PCC_{target})$. However, a closed form solution for $PCC^{-1}(\cdot)$ rarely exists. In this case, we can employ the numerical algorithm to evaluate it and obtain the sample size. For instance, the binary search algorithm can useful. We find n_1 and n_2 such that $PCC(n_1) < PCC_{target}$ and $PCC(n_2) > PCC_{target}$. Since $PCC(n)$ is monotone increasing with respect to n, the required sample size n should fall inside the interval of (n_1, n_2) and be efficiently identified by the binary search method.

4.3.3 Feature Selection

In general the weighting scheme G can assign non-zero weights to all available features. However, using all available features in the classifier can reduce PCC if most of them are not important, raising the noise of level in the classifier. Instead,

performing regularized feature selection allows us to include only important features in the classifier and consequently enhance its performance. This is an especially appealing approach where the sample size n is much less than the total number of features, p.

Feature selection is primarily driven by pairwise associations between the features x_{ij} and the group assignment y_i, in which features having strongest associations will be selected. To proceed, let $Z = (z_1, \ldots, z_p)$ be the vector of z-scores derived from a potential training data D:

$$z_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i x_{ij} \quad j = 1, \ldots, p.$$

Then $z_j \sim N(\sqrt{n}\mu_0, 1)$ for important feature $j = 1, \ldots, m$ and $z_j \sim N(0, 1)$ otherwise. In other words, the z-scores for those $p - m$ unimportant features appear to be clustered around 0 while the z-scores for those m important features are away from zero. Therefore, a natural strategy for feature selection is to choose an appropriate threshold λ such that we only include features satisfying $|z_j| \geq \lambda$ $j = 1, \ldots p$. With a given threshold λ, we can incorporate this parameter of feature selection into the definition of weighting scheme:

$$w_j = \begin{cases} 1, & \text{if } z_j > \lambda; \\ -1, & \text{if } z_j < -\lambda; \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

We denote the weighting scheme (4.1) with threshold λ by G_λ. In actual data analysis, we can employ various thresholding procedures $H(D)$ to determine threshold λ and the corresponding weight scheme becomes $G_{H(D)}$. Once again, it is worth emphasizing that the thresholding procedure $H(D)$ should depend on the training data D only but not the true effect size vector $\boldsymbol{\mu}$, since $\boldsymbol{\mu}$ is never fully known in practical applications.

In the following we will introduce two procedures that may be employed to determine λ.

Higher Criticism Threshold

The higher criticism threshold (HCT), proposed by *Donoho and Jin* (2009), provides a data-driven approach to select λ in a high-dimensional classification analysis. Using the distribution of p-values obtained from univariate association screening test for individual features with the group assignment, HCT allows us determine a threshold λ for weighting scheme (4.1). Since only important features are used for classification, HCT can substantially improve the PCC.

Now we present the detail of HCT procedure, which is denoted by $HCT(D)$ with respect to training data D. Recall that z_j is the z-score obtained in an association test for feature j with group assignment y_i. Then the resulting two-sided p-value is $\pi_j = 2 \times \{1 - \Phi(|z_j|)\}$. For an unimportant feature j, $z_j \sim N(0,1)$ and the corresponding $\pi_j \sim Uniform(0,1)$. In contrast, for an important feature j, $z_j \sim N(\sqrt{n}\mu_0, 1)$. Therefore the p-value π_j for the important features does not follow $Uniform(0,1)$ and tends to be extremely small compared to p-values for the unimportant features. As a result, we only need to focus on the smallest $\lceil p\alpha_0 \rceil$ p-values ($\lceil x \rceil$ denote the smallest integer larger than x and we typically choose $\alpha_0 = 10\%$) so as to identify the important features. Denote these smallest p-values in an increasing order by $(\pi_{(1)}, \cdots, \pi_{(\lceil p\alpha_0 \rceil)})$. *Donoho and Jin* (2009) showed that the l-th ordered p-value with $l = \text{argmax}_{k=1,\ldots,\lceil p\alpha_0 \rceil} \sqrt{p} \frac{k/p - \pi_{(k)}}{\sqrt{k/p(1-k/p)}}$ provides an appropriate cutoff for feature selection. So the features whose p-values are less than $\pi_{(l)}$ are considered being important for classification. Equivalently, we may convert the p-value cutoff $\pi_{(l)}$ to the z-score threshold $HCT(D) = |\Phi^{-1}(\frac{\pi_{(l)}}{2})|$. Then the resulting PCC is $E_D \left\{ PCC(G_{HCT(D)}(D); \boldsymbol{\mu}, n) \right\}$ where the expectation is taken over the sampling distribution of training data D.

The HCT brings new insight to the asymptotic properties of linear classifiers under the so-called rare and weak model (*Donoho and Jin*, 2009). This model is of specific interest in the theory of high-dimensional classification, referring to the mechanism in that important features become rarer and their effect size become weaker when the total number of features increases, namely $p \to \infty$. In the rare and weak model suggested in (*Donoho and Jin*, 2009), the number of important features m increases with p through the relation: $m = p^{1-\beta}$ where the parameter $\beta \in (0, 1)$ essentially controls the sparsity; β closer to 1 implies a smaller number of important features. The effect size μ_0 vanishes to zero along p according to $\mu_0 = \frac{\sqrt{2r \log p}}{\sqrt{n}}$ where $r \in (0, 1)$ controls the signal strength. Moreover, we see that the sample size n is in the order of $c \times (\log(p))^\gamma$ with c and $\gamma > 0$. Within this assumed framework, *Donoho and Jin* (2009) proves that the PCC of any linear classifiers is characterized only by (β, r) through a certain function $\rho(\beta)$: (i) when $r > \rho(\beta)$, the classification analysis is asymptotically feasible and the PCC of the HCT classifier approaches to 1 as $p \to \infty$; and (ii) when $r < \rho(\beta)$, the classification analysis is asymptotically unfeasible and the PCC of any linear classifier approaches to 0.5 as $p \to \infty$. The result of asymptotic feasibility is very critical for guiding the design of classification analysis. It allows us to make a timely decision on the feasibility of a study under the planning stage. Simply verifying inequality $r > \rho(\beta)$ can help investigators to avoid wasting their investment of time and effort to data collection for a failed study. The above asymptotic results of feasibility will be illustrated through simulation in Appendix A using our R software package.

Cross Validation Threshold

Cross validation (CV) is a popular data-driven method to choose tuning parameters in the statistical literature. Since the value of λ can also be regarded as a tuning parameter, it can be determined by CV.

We denote the CV thresholding procedure by $CV(D)$ where D is training data. The main idea behind $CV(D)$ to choose an appropriate λ is to maximize the apparent PCC, denoted by $\tilde{PCC}(D, s)$, which is function of both a threshold s and the training data D. The apparent PCC can be computed as the following steps:

Step 1: Randomly divide a training data D into k equal-sized subsets, $D^1, \cdots, D^k$.

Step 2: For each $q = 1, \ldots, k$, treat D^q as a CV testing set and the rest of the data D^{-q} as a CV training set, and then use (4.1) to obtain the weighting $w(D^{-q}, s)$ from the CV training set D^{-q} and a given s value where z-scores are calculated from D^{-q} only.

Step 3: Calculate the apparent PCC on a CV testing set D^q as

$$\begin{aligned} & \tilde{PCC}(D^q, w(D^{-q}, \lambda)) \\ = & \sum_{(\boldsymbol{x}_i, y_i) \in D^q} \{I(\boldsymbol{x}_i \cdot w(D^{-q}, \lambda) \geq 0) I(y_i = 1) + I(\boldsymbol{x}_i \cdot w(D^{-q}, \lambda) < 0) I(y_i = -1)\} \end{aligned}$$

Step 4: Repeat Step 2 and Step 3 for all $q = 1, \ldots, K$ and then calculate the overall apparent PCC $\tilde{PCC}(D, s) = \frac{1}{k} \sum_{q=1}^{k} \tilde{PCC}(D^q, w(D^{-q}, \lambda))$.

Step 5: Finally select λ to maximize the overall apparent PCC: $CV(D) = \text{argmax}_{s>0} \tilde{PCC}(D, s)$ for a sequence of dense grid points.

For a study employing the CV thresholding procedure for feature selection, the expected PCC is $E_D \left\{ PCC(G_{CV(D)}(D); \boldsymbol{\mu}) \right\}$ where the expectation is taken over sampling distribution the training data D.

4.4 Implementation

Now we present numerical methods to implement the calculation of PCC. *Dobbin and Simon* (2007) suggested that the PCC of a study is given by $\max_{s \in (0,1)} E_D \{ PCC(G_s(D); \boldsymbol{\mu}, n) \}$. Clearly their approach implicitly assumes that the optimal threshold maximizing is chosen in the future classification analysis after data collection is completed. However, this assumption does not seem to be realis-

tic. This is because the optimal threshold depends on the effect size vector $\boldsymbol{\mu}$ but $\boldsymbol{\mu}$ is never fully known in an actual study. So in practice, the thresholding procedure $H(D)$ can only produce a threshold from a given training data D, which mostly is not optimal. Therefore, Dobbin and Simon's approach to calculate the PCC ignores the sampling uncertainty arising from the thresholding procedure in an actual application. As a result, this strategy tends to overestimate the PCC, and subsequently the sample size requirement will be underestimated.

In order to address this issue, we need to incorporate the thresholding procedure $H(D)$ into the PCC estimation, so that it is matched seamlessly to actual data analysis. In the following we present details pertaining to implementation of PCC estimation employing HCT and CV as the thresholding procedure, respectively.

4.4.1 Implementation of Cross Validation Threshold

We propose the following algorithm to evaluate the expected PCC, $E_D\left\{PCC(G_{CV(D)}(D); \boldsymbol{\mu}, n)\right\}$, when the CV thresholding employed for feature selection. The algorithm is based on Monte Carlo simulation method to estimate the expectation with respect to sampling distribution of the training data D. At the kth iteration of the simulation, we execute:

(a) Generate a training data D with sample size n based on a design assumption on $\boldsymbol{\mu}$;

(b) Employ CV on the simulated D to determine $\lambda = CV(D)$;

(c) Compute the z-scores z from the training data D;

(d) Use (4.1) to derive weighting scheme w based on the z-scores z and threshold λ;

(e) Acquire $PCC_{(k)} = \Phi(\frac{\boldsymbol{w}\cdot\boldsymbol{\mu}}{\sqrt{\boldsymbol{w}\cdot\boldsymbol{w}}})$.

We repeat the above steps 1-5 for N iterations, generating $PCC_{(1)}$, ..., $PCC_{(N)}$. Then the Monte Carlo estimate of $E_D\left\{PCC(G_{CV(D)}(D);\boldsymbol{\mu},n)\right\}$ is given by $\frac{1}{N}\sum_{k=1}^{N} PCC_{(k)}$. The above algorithm has been implemented in an R packages; see more details in Section 7 Software.

4.4.2 Implementation of Higher Criticism Threshold

Since $HCT(D)$ is a data-driven thresholding procedure, Monte Carlo simulation is in principle applied to evaluate $E_D\left\{PCC(G_{CV(D)}(D);\mu,n)\right\}$. The complication for the implementation arises from the fact that $HCT(D)$ depends on the training data D exclusively through the $\lceil p\alpha_0\rceil$ smallest p-values. In this paper, we propose instead of simulating training data D, a computational fast algorithm that directly simulates the $\lceil p\alpha_0\rceil$ smallest p-values from the distribution of ordered statistics. The algorithm is described as follows:

(a) Simulate z-scores for m important features from $z_j \sim N(\sqrt{n}\mu_0, 1)$, $j = 1, \ldots, m$;

(b) Convert the above z-scores to two-sided p-values by $\pi_j = 2 \times (1 - \Phi(|z_j|))$, $j = 1, \ldots, m$;

(c) Simulate a random variable $u \sim Beta(\lceil p\alpha_0\rceil, p - m + 1 - \lceil p\alpha_0\rceil)$;

(d) Simulate variables $v_1, \ldots, v_{\lceil p\alpha_0\rceil - 1}$ independently from $Uniform(0, u)$;

(e) Sort vector $(z_1, \ldots, z_m, v_1, \ldots, v_{\lceil p\alpha_0\rceil-1}, u)^t$ in an increasing order.

In Appendix B, we show that the $\lceil p\alpha_0\rceil$ smallest values given by the vector $(\pi_1, \ldots, \pi_m, v_1, \ldots, v_{\lceil p\alpha_0\rceil-1}, u)^t$ has the same joint distribution as the $\lceil p\alpha_0\rceil$ smallest p-values $(\pi_{(1)}, \ldots, \pi_{(\lceil p\alpha_0\rceil)})^t$ derived from training data D. As a by-product, the above algorithm also supplies the z-score $z_1, \ldots, z_m$ for m important features.

Using this new algorithm to generate the $\lceil p\alpha_0\rceil$ smallest p-values gives rise to a clear computational benefit over simulating the entire vector of p-values based on

training data D. In the latter case, np random variables need to be generated in order to simulate training data set as opposed to the former case where only $\lceil p\alpha_0 \rceil + m$ variables are generated. Thus the computational efficiency ratio is $\frac{np}{\lceil p\alpha_0 \rceil + m}$, which is approximately equal to $\frac{n}{\alpha_0}$ because m is much smaller than p. For the instance of $n = 100$ individuals in a study and the smallest $\alpha_0 = 10\%$ of p-values being considered, the algorithm is roughly 1000 times more efficient than generating the p-values from training data using routine Monte Carlo algorithm similar to that given in Section 4.1.

To incorporate this algorithm to evaluate the expected PCC, $E_D\left\{PCC(G_{HCT(D)}(D); \boldsymbol{\mu})\right\}$, as follows: At the kth iteration, we perform the following steps:

(a) Generate the $\lceil p\alpha_0 \rceil$ smallest p-values $(\pi_{(1)}, \ldots, \pi_{(\lceil p\alpha_0 \rceil)})^t$ and the corresponding z-scores for m important features, with given sample size n and the inputs of μ_0 and m;

(b) Determine the threshold $\lambda = HCT(D)$;

(c) Use (4.1) to calculate the weight $w_1, \ldots, w_m$ of m important feature using their z-scores $z_1, \ldots, z_m$ and λ .

(d) Calculate $PCC_{(k)} = \Phi(\frac{\mu_0 \sum_{j=1}^{m} w_j}{\sqrt{\#w}})$, where $\#w$ denotes the number of elements in the vector $(\pi_{(1)}, \ldots, \pi_{(\lceil p\alpha_0 \rceil)})^t$ whose values are smaller than $2 \times (1 - \Phi(\lambda))$.

We repeat steps 1-4 for N iterations, generating $PCC_{(1)}$, ..., $PCC_{(N)}$. Then the Monte Carlo estimate of $E_D\left\{PCC(G_{HCT(D)}(D); \boldsymbol{\mu}, n)\right\}$ is given by $\frac{1}{N}\sum_{k=1}^{N} PCC_{(k)}$. The above algorithm has been implemented in an R packages; see more details in Section 7 Software.

4.5 Augmenting with New Sources of Features

In practice, investigators typically collect multiple sources of features and intend to use all of them in classification analysis. The key question concerning the augmentation of features is to investigate the potential PCC gain resulted from adding new sources of features to the study. For simplicity, let us focus on two sources of features (e.g. molecular biomarkers and clinical variables). Denote the features already in the study by Type A and the new set of features to be added by Type B. Let p_A and p_B be the respective dimension of Type A and Type B features. For subject i in the study, we collect measurements $\boldsymbol{x}_i^A \in R^{p_A}$ and $\boldsymbol{x}_i^B \in R^{p_B}$. Again $y_i = +1, -1$ is the group assignment for subject i. The prevalence for Group +1 and Group -1 is $P(y_i = 1) = \pi_1$ and $P(y_i = -1) = 1 - \pi_1$. After centering these features as discussed in Section 3, we assume the following joint conditional distribution for features $\boldsymbol{x}_i^A$ and $\boldsymbol{x}_i^B$:

$$\begin{pmatrix} \boldsymbol{x}_i^A \\ \boldsymbol{x}_i^B \end{pmatrix} | y_i \sim N \left(y_i \begin{pmatrix} \boldsymbol{\mu}^A \\ \boldsymbol{\mu}^B \end{pmatrix}, \begin{pmatrix} \Sigma^A & 0 \\ 0 & \Sigma^B \end{pmatrix} \right)$$

where $\boldsymbol{\mu}^A$ and $\boldsymbol{\mu}^B$ are the respective effect size vectors, and Σ^A and Σ^B are the respective variance matrices. It is worth noting that we do not place any restriction on the effect size vectors and the variance matrices. For instance, the important features do not have to be sparse and correlation among the features of the same type is allowed. For mathematical convenience, we assume that the correlation across two sources of features is zero. The results given in this section are applicable to a wide range of scenarios, including but not limited to the rare-and-weak model we introduce in Section 3.3.

To study PCC gain, we need to consider the PCC of linear classifiers in three cases. including 1) Type A features only; 2) Type B features only; and 3) Type A features augmented with Type B features. We denote the respective weights for the Type A

and Type B features are w_A and w_B, and here again we do not place any assumptions on these weights. For instance, the weights can be derived from any thresholding procedures for feature selection. It is known that the PCC of a linear classifier can be written as $Q_{\pi_1}(t) = \Phi(t - \frac{1}{2}\log(\frac{1-\pi_1}{\pi_1})\frac{1}{t}) \times \pi_1 + \Phi(t + \frac{1}{2}\log(\frac{1-\pi_1}{\pi_1})\frac{1}{t}) \times (1 - \pi_1)$ with $t = \frac{\boldsymbol{\mu}\cdot\boldsymbol{w}}{\sqrt{\boldsymbol{w}^t\Sigma\boldsymbol{w}}}$ and $\boldsymbol{\mu}$ is the effect size vector, Σ is the variance and $\boldsymbol{w}$ is the weight. Correspondingly, conditioning on the weights w_A and w_B, the PCC of the classifier using Type A features only is $PCC_A = Q_{\pi_1}(\frac{\boldsymbol{\mu}^A\cdot\boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t\Sigma^A\boldsymbol{w}_A}})$; the PCC of the classifier using Type B features only is $PCC_B = Q_{\pi_1}(\frac{\boldsymbol{\mu}^B\cdot\boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_B^t\Sigma^B\boldsymbol{w}_B}})$; and finally, the PCC of the classifier using Type A and Type B features simultaneously is $PCC_{AB} = Q_{\pi_1}(\frac{\boldsymbol{\mu}^A\cdot\boldsymbol{w}_A+\boldsymbol{\mu}^B\cdot\boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t\Sigma^A\boldsymbol{w}_A+\boldsymbol{w}_B^t\Sigma^B\boldsymbol{w}_B}})$.

In Appendix C, we will prove the following theorem regarding an inequality for three PCC's:

$$\min(PCC_A, PCC_B) \leq PCC_{AB} \leq Q_{p_1}(\sqrt{2}\cdot Q_{\pi_1}^{-1}(\max(PCC_A, PCC_B))) \qquad (4.2)$$

The first equality will hold when $\frac{\boldsymbol{w}_k^t\Sigma^k\boldsymbol{w}_k}{\boldsymbol{w}_j^t\Sigma^j\boldsymbol{w}_j} \to 0$ where $(j,k) = \begin{cases} (A,B) & \text{if } PCC_A < PCC_B \\ (B,A) & \text{if } PCC_A \geq PCC_B \end{cases}$. The second equality is reached when $PCC_A = PCC_B$ and the weights are scaled such that $w_A\Sigma^A w_A = w_B\Sigma^B w_B$. The inequality provides an upper bound of the PCC of the classifier including both two types of features. The upper bound is a function of the PCC of the classifier including only the more predictive type of features, i.e. $\max(PCC_A, PCC_B)$. If either PCC_A or PCC_B approaches to 1, the upper bound will also approach to 1. We compare the PCC_A and PCC_B and the upper bound numerically in the simulation section 6.2.

4.6 Simulation Experiments

We conduct two major simulation studies to investigate the performance of the proposed design methods. In Simulation A, we focus on the PCC estimated by the methods under finite sample scenarios mimicking practical studies. In Simulation B, we illustrate numerically the upper bounds of the PCC when combining two types of features based on the inequality (4.2) in Section 5.

4.6.1 Simulation A

We set up the simulation study as follows: Consider there are $m = 10$ important features that are differentially expressed in two groups of patients. The minimal effect size of important features is $\mu_0 = 0.4$ and the two groups are equally proportioned with $p_1 = p_2 = 0.5$. Varying the total sample size from $n = 20$ to $n = 300$, we compute the PCC using the DS, CV and HCT thresholding procedures.

The relations between the expected PCC and the sample size are displayed in Panel (a) and (b) of Figure 4.2 with $p = 500$ and $p = 10^5$ respectively. It is easy to see that the PCC increases along with the sample size n. This is not surprising because larger data set typically provides more information regarding which features are important to be included in classification.

The estimated PCC differ across various classification methods. The PCC estimated by the DS method is always the highest. This is primarily because it has made use of the true effect size vector $\boldsymbol{\mu}$ to choose the threshold λ. However, due to its reliance on $\boldsymbol{\mu}$, the DS method has no counterpart in actual data analysis. As a result, the PCC the DS method provides is overly optimistic about the actual ability in feature selection. On the other hand, the PCC estimated by the CV and the HCT method do not rely on information on $\boldsymbol{\mu}$. These PCC estimates truly reflect the achievable performance of the corresponding classifiers in the real world applications. Similarly, the receiver operating characteristic (ROC) curves of these three methods

PCC_{target}	DS	DSCV	HCT
0.70	34	42	40
0.75	46	52	58
0.80	62	72	82
0.85	90	106	138

(a) Sim A1

PCC_{target}	DS	DSCV	HCT
0.70	88	92	108
0.75	104	108	130
0.80	126	130	158
0.85	164	190	230

(b) Sim A2

Table 4.1: Sample Size Requirement

(Panel (c) and (d)) also suggests that the DS method provides is overly optimistic about the actual ability in feature selection reflected by the CV and the HCT method.

It is important to point out that the difference in PCC estimated by these methods have strong implication on the sample size calculation. Table 4.1 lists the sample size requirement suggested by three aforementioned design methods. Since the DS method often predicts overly optimistic PCC, it recommends much smaller sample size than the CV method or the HCT method. A study could be severely "under powered" if it is designed solely based on the DS method. Furthermore, because the gap of the PCC curves become stable as the PCC increases, such discrepancy of the PCC estimated across these methods gradually translates into a substantial difference in sample size. These observations highlight the importance of calculating the sample size based on PCC estimates that are achievable by statistical methods at the data analysis stage.

4.6.2 Simulation B

Now we illustrate the feature augmentation. Inequality (4.2) provides an upper bound of the PCC when we supplement new type of features to the study relative to the PCC's of each type of features individually, i.e. $\max(PCC_A, PCC_B)$. We can

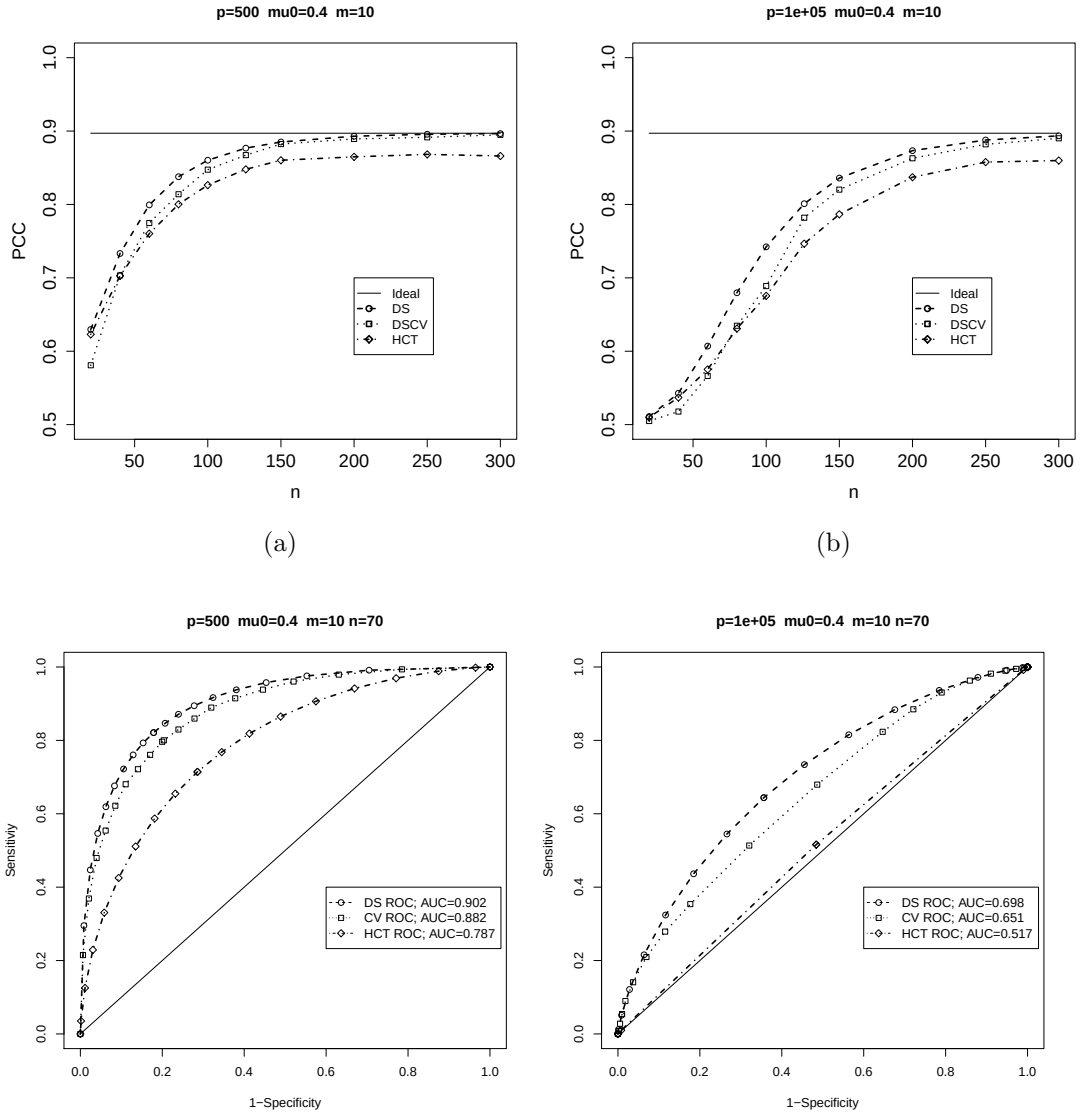

p=500 mu0=0.4 m=10
PCC
n
Ideal
DS
DSCV
HCT
(a)
p=1e+05 mu0=0.4 m=10
PCC
n
Ideal
DS
DSCV
HCT
(b)
p=500 mu0=0.4 m=10 n=70
Sensitivity
1−Specificity
DS ROC; AUC=0.902
CV ROC; AUC=0.882
HCT ROC; AUC=0.787
(c)
p=1e+05 mu0=0.4 m=10 n=70
Sensitivity
1−Specificity
DS ROC; AUC=0.698
CV ROC; AUC=0.651
HCT ROC; AUC=0.517
(d)

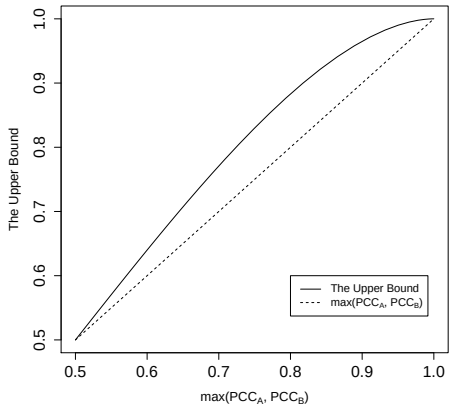
Figure 4.2

visualize in Figure 4.3 the upper bound (the left column) and the improvement in PCC (the right column) for Group 1 prevalence (p_1) ranging from 0.5 to 0.7. If both PCC_A and PCC_B are small, then $\max(PCC_A, PCC_B)$ will also be small. In this case, Figure 4.3 suggests that the upper bound of the PCC of classifiers that incorporate both Type A and Type B features is only slightly higher than $\max(PCC_A, PCC_B)$. In other words, if either type of features is not predictive of the group assignment, supplementing new type of features with the type of features already in the study will not greatly improve the PCC. If both types of features are of medium quality, i.e. both PCC_A and PCC_B are in the medium range around 0.8, then we could obtain the highest gain (approximately 10% or less) in PCC by incorporating the new type of features. Finally, if either PCC_A or PCC_B is very high, then the augmentation with new features will not provide much more information for classification than using only one type of the features. Therefore, the PCC cannot be significantly improved in this case.

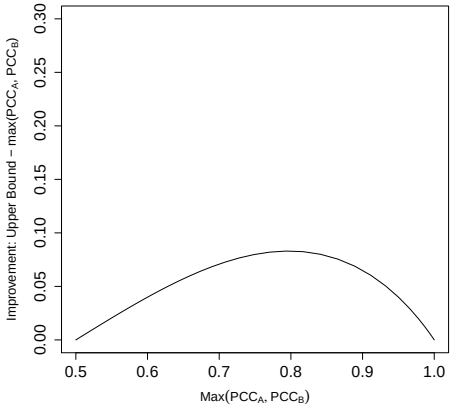
4.7 Application

Through our R package, we now apply the methods developed in this paper to design our motivating study, that is a new study proposed by a clinician at University of Michigan Kidney Transplantation Center, which aims to predict the graft survival of a patient after kidney transplant using proteomics biomarkers. In this study, patients are to be classified into two groups: stable transplant and rejected transplant. Overall the planning of the study will be carried out in two steps (Figure 1).

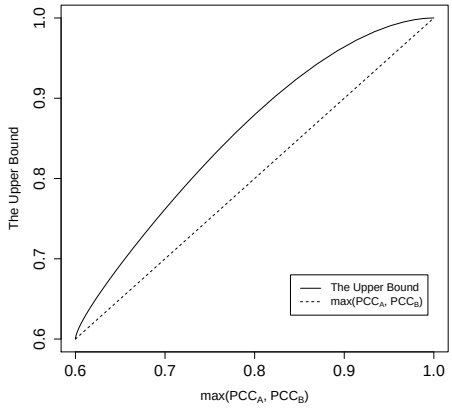
In Step I, $p = 108$ proteomics biomarkers will be collected using microarray for each patient in both stable and rejected group, and classifiers for graft survival will be developed based on the microarray data. The investigator suggested that among these proteins, approximately $m = 10$ of them are informative for predicting graft survival, and for each informative biomarker, the difference in mean between the



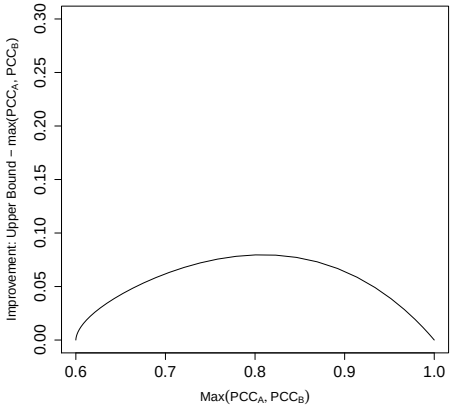
(a) $p_1 = 0.5$



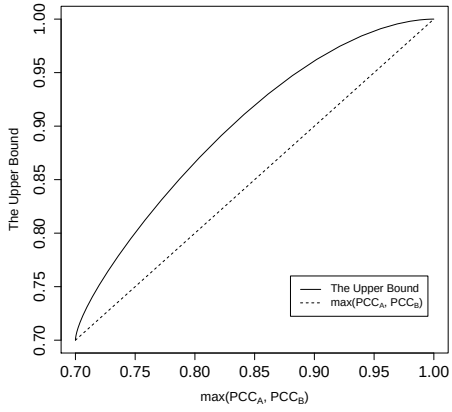
(b) $p_1 = 0.5$



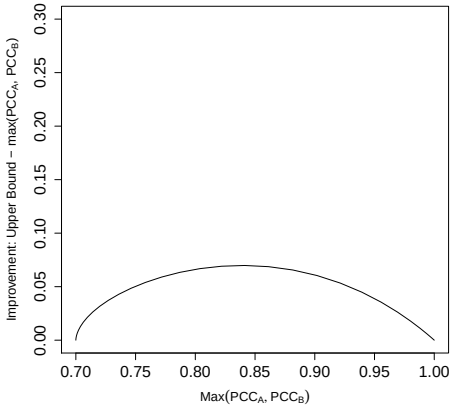
(c) $p_1 = 0.6$



(d) $p_1 = 0.6$



(e) $p_1 = 0.7$



(f) $p_1 = 0.7$

Figure 4.3: The Upper Bounds of PCC with Two Types of Features and its Improvement over a Single Type of Features.

stable group and the rejected group should be at least 0.8 of the standard deviation and for the non-informative biomarkers, the difference in mean is 0.

Given these inputs from the investigators, we check the feasibility of the proposed classification analysis. Under the rare-and-weak model, the strength parameter $r = \frac{\mu_0^2 n}{2 \log p} = 5.13$ and the sparsity parameter $\beta = 1 - \frac{\log m}{\log p} = 0.51$. Referring to the definition of $\rho(\beta)$ in *Donoho and Jin* (2009), we confirm $r > \rho(\beta)$. Therefore, the classification analysis is feasible and at least there is at least one classifier (e.g. HCT) whose PCC approaches to 1 as $p \to \infty$.

Then we set an appropriate target for the PCC. The PCC of the ideal classifier is $PCC_{ideal} = \Phi(\frac{m\mu_0}{\sqrt{m}}) = 0.90$, and the PCC of the random classifier is $PCC_{random} = 0.50$ since both groups are of equal proportion in our design. Thus any values between $PCC_{random} = 0.50$ and $PCC_{ideal} = 0.90$ could be a legitimate PCC_{target}.

Panel (a) of Figure 4.4 which shows the sample size n needed to achieve PCC estimated by various methods for this proposed study. In general, as the sample size becomes larger, the PCC increases from the lower bound 0.50 to the upper bound 0.90. Panel (b) of Figure 4.4 displays the sample size requirement for various PCC targets. A large sample size is required when a higher PCC target is set for the classifiers in the study. In addition, the DS method clearly underestimates the sample size requirement when compared to the other methods.

In Step II, the investigator will consider adding clinical predictors such as lab test measures, and/or patient demographic characteristics into the classifier to hopefully improve classification performance. Let us consider an interesting scenario where the study currently being planned with the proteomics biomarker alone can only achieve a low target PCC, say 0.7, and the investigator considers to improve the PCC by incorporating a new type of features, such as patients' renal functional measures, including proteinuria, GFR, hematuria, ALB, CHOL, CHEAT, C3, C4, etc. In Panel (c) of Figure 4.4, we present the achievable region of the PCC with both types of

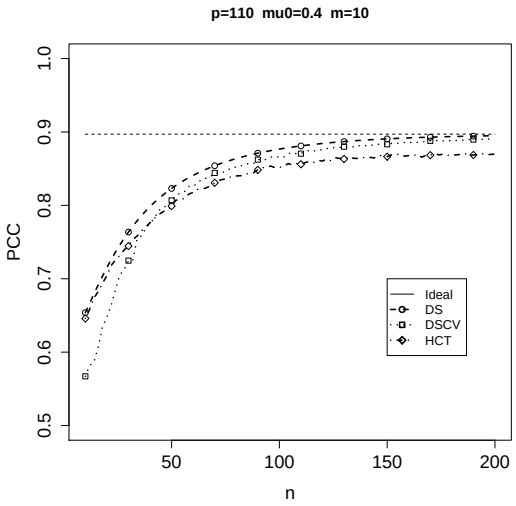
features (PCC_{AB}, in the notation of Inequality (4.2)) for various value of the PCC with the new features (PCC_B). We keep the PCC with the proteomics markers (PCC_A) fixed at 0.7 since it is already planned in the study. When the new features are not as informative as the proteomics markers ($PCC_B \leq PCC_A$), combining the two leads to limited improvement of the classifier at best and might actually degrade the classifier due to the noise introduced by the new features. On the other hand, if the new features are more informative ($PCC_B \geq PCC_A$), incorporating them in the classifier is likely to substantially enhance the PCC, potentially exceeding 90% PCC.

4.8 Software: HDDesign

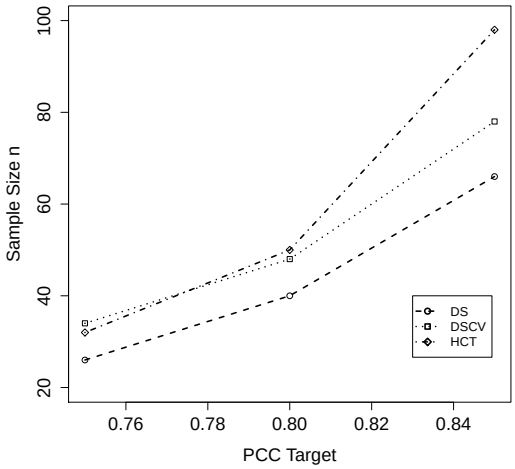
We have implemented the proposed methodologies in this paper in a user-friendly R package named "HDDesign". This package includes functions i) to determine feasibility of a classification analysis; ii) to compute the upper bound of a PCC for any linear classifiers, using the Dobbin and Simon (DS) , CV, or HCT method; iii) to estimate the PCC of design method given the input of design parameters; and iv) to determine the sample size requirement to achieve a pre-specified target PCC for a design method. There are various design methods available in the package to choose. For more details concerning the usage of the functions in the package, please refer to the manual of the package.

4.9 Discussion

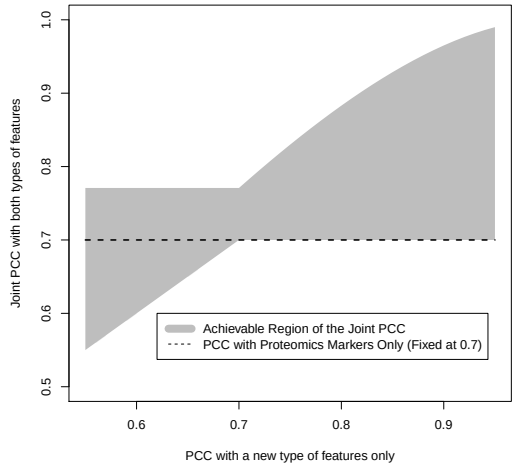
In this paper, we develop and compare designs for classification analysis in the presence of high-dimensional features using PCC as the objective for sampling size estimation. We discuss the PCC estimation method by *Dobbin and Simon* (2007), and extend this approach by implementing both cross-validation method and the HCT method to compute the actual PCC that would be obtained in an practical



(a) PCC vs. Sample Size



(b) Sample Size Needed to Achieve PCC Target



(c) Achievable Region of the Joint PCC with Proteomics Makers and a New Type of Features in the classifier.

Figure 4.4: Application: Study Design for Predicting Survival after Kidney Transplant

application. We propose a new simulation algorithm based on order statistics that allows us to efficiently compute the expected PCC of the HCT classifier. We derive an inequality for the upper and lower bounds of the achievable PCC when supplementing new types of features into the study. We employ simulations and a a practical example to compare the relative efficiency of the three designs and evaluate potential bias in terms of the PCC and the corresponding sample size. More importantly, we supply an R package that implements all the proposed methods, which will fill in the software need of design for classification analysis with high-dimensional data.

For future work, our work on the design of high dimensional classification studies can be improved in the following directions. Classification of more than two groups of patients could be considered since it commonly appears in many clinical studies. We would also be interested in investigating how various correlation structures of features affect classification performance. An interesting scenario is non-zero correlations among the features. For example, if two features are positively correlated, they will provide similar information and thus they are less effective for classification when compared to two independent features. When the positive correlation is ignored, we risk overestimating the PCC. Similarly, the opposite statement is true for negative correlations. Potentially, we could adjust for the correlation if we plugins the values of correlations into the formula for calculating the PCC. However, reliable estimate for the correlations are very difficult to obtain given the enormous pairs of correlations ($\frac{(p-1)p}{2}$ for p features) and the limited preliminary data available in the study planning stage. Without such reliable estimate, the plugin approach might not significantly improve, or even deteriorate the accuracy of PCC calculation. So for simplifying design inputs, many methods, including our own, assume features are independent. We believe a more promising approach to incorporate correlation in the PCC calculation is to derive the bounds of the PCC assuming the plausible range of the correlation values. This approach will be more robust than the plugin method.

CHAPTER V

Conclusion

In this dissertation, we developed novel design methodologies for studies employing 1) repeated measures of nonlinear profiles; 2) functional responses; 3) high dimensional genetics and proteomics data.

In Chapter 2, we focus on studies involving repeated measures of nonlinear profiles. The profiles were characterized by the conditionally mixed model. We considered sampling not only individuals, but also days within individuals, and times within days. The objective was minimizing estimation variance of profile features. We analyzed the impacts of the multi level variabilities on the study design by deriving a simplified version of the variance formula. Such formula is surprisingly neat in that the terms involving the between subject variability, between day variability, and the daily sampling schedules are completely separated from each other. Hence, we derived several useful properties: 1) the optimal daily sampling schedule depends only on the shape of the response profile; 2) we obtained the optimal number of days for sampling under different cost structures for the study; 3) we showed that the optimal Bayesian design is only affected by the expectation of the variability parameters, but not their full prior distribution. These properties are useful in practice since they substantially reduce the computation burden of Bayesian designs. We apply the new methods to the salivary cortisol studies for investigating the association between cardiovascular

disease and stress.

In Chapter 3, we considered studies involving data with underlying functional response with the objective of capturing the mean profile and between subject variability. We propose an approach to longitudinal study design that optimizes estimation of the mean and between-subject variability. Our approach employs a semiparametric model that characterizes the mean profile and the variability separately. We used functional principal component analysis (FPCA) to derive a parsimonious and flexible representation of the temporal pattern of the variability. Smoothing of the principal components was incorporated to enhance estimation stability given the potentially unbalanced and irregular sampling of the preliminary data. The population mean can be modeled with known parametric functions or with splines. Following the existing literature, we employed D-optimality for the mean and variability parameters as the optimization criterion. Simulations suggest that if between subject variability is induced by random effects in a parametric mean profile, the FPCA approach and existing methods based on parametric mixed model (PMM) lead to equally competent sampling schedules. For more flexible variability structures, however, the FPCA approach is superior. We apply the new approach to two real world examples. In the first example we show that the FPCA approach identifies new sampling regions for measuring the mean and the variability of salivary cortisol profiles that are not discovered by the PMM approach. In the second example, the mean profile of progesterone in menstrual cycles is modeled by splines and the PMM approach is no longer applicable. We employ the FPCA approach to identify optimal schedules for capturing the variability of the profiles and show that schedules we obtained are also highly predictive of conceptive status. In both simulations and real world examples, we show that the FPCA approach is more robust flexible for capturing the variability of the process.

In Chapter 4, we explore designs for studies involving high dimensional genetics

and proteomics data with the objective of constructing classifiers with high probability of correct classification (PCC). We begin with a review of a statistical framework proposed by *Dobbin and Simon* (2007), which incorporates feature selection in the estimation of the PCC for high dimensional features. We recognize that the thresholding step in this approach requires the complete knowledge of the effect size, which is not available in practical applications. As a result, this approach overestimates the PCC achievable by practical data analysis tools and consequently underestimates the sample size requirement necessary to obtain the PCC target. In order to address this issue, we develop two design approaches based on cross-validation (CV) and high criticism threshold (HCT), respectively. These approaches are data driven and do not demand any information regarding the true effect size. Therefore, the PCC estimates derived from these approaches are achievable in practice if the corresponding classifiers are employed. In terms of computation, we propose a new simulation method based on order statistics that allows us to efficiently compute the PCC based on the HCT method. Furthermore, since there are a variety of technologies emerging to measure different type of the high dimensional features, the investigators are often interested in supplementing new type of the features into the study to enhance the PCC of the study. So we derive an inequality for the upper and lower bounds of the achievable PCC when adding new types of feature in the study. We evaluate the performance and validity of our proposed method by simulations. Finally we use the prediction of long term survival after kidney transplant to illustrate the application of our new approaches.

Overall this dissertation contributes three novel methodologies for study designs involving three distinct data structures. The methods were evaluated by simulations and illustrated with real data from diverse application scenarios. Furthermore these methodologies can be improved to meet the need of more complex design problems.

For measuring the profile features as described in Chapter 2, it will be beneficial

to consider non-parametric models which place less restriction on the data than the parametric models we used. Another extension is to consider the sampling protocols for detecting differences between groups. For example, cortisol profiles may differ by age, gender and ethnic groups. From the statistical point of view, the same method for selecting the efficient sampling protocol can be applied to all groups since optimal sampling protocols are typically robust to small changes in parameters. On the other hand, if the differences in the cortisol profiles are big, new methods might be needed to simultaneously select schedules for each group, as well as optimizing other criteria like power or sample size.

We can also extend the framework of determining optimal schedules based on FPCA as described in Chapter 3 in many ways depending on the design problems. One direction will be to obtain the sampling schedules for estimating $\beta_k(t)$, which is modeled by smoothing spline. One can consider a surrogate for the smoothing spline, for which the optimal schedule is known, such as a piecewise linear spline (the optimal schedule is simply the set of the knots). Another extension could be the sampling schedules for a study incorporating multilevel sampling. In the second example, repeated cycles from the same woman were excluded. To incorporate such data, one could employ a hierarchical FPCA model and derive the information matrix for the grand mean and variability parameters at multiple levels. Furthermore, our method currently consider only one optimal schedule in the design due to practical constraint of large scale epidemiology studies. Nevertheless, it is also interesting to consider multiple optimal schedules in the design since it could potentially improve the efficiency of the design (*Mentré et al.*, 1997). Lastly, deriving theoretical results of the FPCA approach, such as its asymptotic properties and alternative methods to select the number of principal components, would be of interest.

Building on top of these methods, we can explore more complex design for studies involving longitudinal measurement. For instance, the investigator might be inter-

ested in generating the optimal design for measuring a profile feature (Chapter 2) as well as capturing the between subject variability (Chapter 3). One approach to this problem is to define a utility function that combines both the profile feature objective and the between subject variability objective. However, these two objectives are of different scales: variance for the measuring the profile feature; D-optimality for capturing the variability. Therefore it is difficult to define an appropriate utility function that is readily interpretable. Alternatively, we can identify the design that maximize our ability to capture the variability among those designs that are capable of measuring the profile feature to the pre-defined precision. This approach is often referred to as the minimax approach, which will provide more interpretable results in our case. We can extend this approach by developing efficient algorithm to identify the minimax design specific for our problem and evaluating other theoretic properties.

Finally, our work on the design of high dimensional classification studies can be improved in the following directions. Classification of more than two groups of patients could be considered since it commonly appears in many clinical studies. We would also be interested in investigating how various correlation structures of features affect classification performance. An interesting scenario is non-zero correlations among the features. For example, if two features are positively correlated, they will provide similar information and thus they are less effective for classification when compared to two independent features. Therefore we will risk overestimating the PCC when positive correlation is ignored. The opposite statement will be true for negative correlations. Potentially, we could adjust for the correlation if we plugins the values of correlations into the formula for calculating the PCC. However, reliable estimate for the correlations are very difficult to obtain given the enormous pairs of correlations ($\frac{(p-1)p}{2}$ for p features) and the limited preliminary data available in the study planning stage. Without such reliable estimate, the plugin approach might not significantly improve, or even deteriorate the accuracy of PCC calculation. So for simplifying de-

sign inputs, many methods, including our own, assume features are independent. We believe a more promising approach to incorporate correlation in the PCC calculation is to derive the bounds of the PCC assuming the plausible range of the correlation values. This approach will be more robust than the plugin method.

APPENDICES

APPENDIX A

Derivation of the Information Matrix for Conditionally Linear Mixed Model

In this section, we will derive the closed form solution of the information matrix for conditionally linear mixed model. Under the framework of CLMM, we have

$$y_{ijk} = f(t;\theta_{ij}) + \epsilon_{ijk} = H(t,\phi)\eta_{ij} + \epsilon_{ijk} \tag{A.1}$$

where $H(t,\phi)$ is a row vector that depends on the measurement time t and a nonlinear population parameter ϕ. η_{ij} is a linear parameter vector. $\epsilon_{ij} \sim N(0,\Sigma^{\epsilon})$ is the measurement error. To incorporate the between subject and between day variability, we assume

$$\begin{aligned} \eta_{ij}|\eta_i &\sim N(\eta_i,\Sigma^d) \\ \eta_i &\sim N(\eta,\Sigma^s). \end{aligned} \tag{A.2}$$

Consider a study in which we have n subjects and each subject takes their cortisol sample for m days according to the schedule $T = (t_1, t_2, \ldots, t_d)$, i.e., the schedule is the same for all days. The corresponding cortisol measurements are $y_i =$

$(y_{i11}, \ldots y_{i1d}, y_{i21} \ldots y_{imd})$. We introduce some more notations: $H(T,\phi) = (H(t_1,\phi), \ldots, H(t_d,\phi))'$ is the design matrix, $\tilde{H}(T,\phi) = (H(T,\phi)', \ldots, H(T,\phi)')'$ is m copies of $H(T,\phi)$ stacked together, $\bar{H}(T,\phi)$ is a block diagonal matrix with $H(T,\phi)$ on the diagonal.

Let $\Theta_i = (\eta_{i1}^t, \ldots \eta_{im}^t)^t$ be the vector collecting linear random effects of all days from the same subject. According to (A.2), the distribution of Θ_i is given by

$$\Theta_i \sim MVN(\Theta, \Sigma) \qquad \Sigma = \begin{pmatrix} \Sigma^s + \Sigma^d & \Sigma^s & \cdots & \Sigma^s \\ \Sigma^s & \Sigma^s + \Sigma^d & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma^s \\ \Sigma^s & \cdots & \Sigma^s & \Sigma^s + \Sigma^d \end{pmatrix} \tag{A.3}$$

where $\Theta = (\eta^t, \ldots, \eta^t)^t$ is d copies of the population mean η and Σ is a block matrix.

The joint distribution of cortisol measurement for subject i over the d days is therefore

$$y_i \sim MVN(H(T,\phi)\Theta,\ D(T,\phi)) \qquad D(T,\phi) = \bar{H}(T,\phi)\Sigma\bar{H}(T,\phi)^t + \Sigma^\epsilon \tag{A.4}$$

Consequently the log-likelihood contribution for subject i (less a constant) is

$$l_i = -\frac{1}{2}\log|D(T,\phi)| - \frac{1}{2}(y_i - \tilde{H}(T,\phi)\eta)^t D^{-1}(T,\phi)(y_i - \tilde{H}(T,\phi)\eta) \tag{A.5}$$

Then by large sample theory, the MLE of $\theta = (\eta, \phi)$ will follow a multivariate normal distribution with mean (η, ϕ) and variance $I^{-1}(\eta,\phi)/n$, where

$$I(\eta,\phi) = -E\begin{pmatrix} \frac{\partial^2 l_i}{\partial\eta^t\partial\eta} & \frac{\partial^2 l_i}{\partial\phi\partial\eta} \\ \frac{\partial^2 l_i}{\partial\phi\partial\eta^t} & \frac{\partial^2 l_i}{\partial\phi\partial\phi} \end{pmatrix} \tag{A.6}$$

is the expected information matrix with components

$$E\left(\frac{\partial^2 l_i}{\partial\eta^t\partial\eta}\right) = -\tilde{H}(T,\phi)^t\cdot D^{-1}(T,\phi)\cdot\tilde{H}(T,\phi) \quad (A.7)$$

$$E\left(\frac{\partial^2 l_i}{\partial\phi^t\partial\eta}\right) = -\tilde{H}(T,\phi)^t\cdot D^{-1}(T,\phi)\cdot\frac{\partial\tilde{H}(T,\phi)}{\partial\phi^t}\cdot\eta \quad (A.8)$$

$$\begin{aligned} E\left(\frac{\partial^2 l_i}{\partial\phi\partial\phi}\right) &= -\left(\frac{\partial\tilde{H}(T,\phi)}{\partial\phi}\cdot\eta\right)^t\cdot D^{-1}(T,\phi)\cdot\left(\frac{\partial\tilde{H}(T,\phi)}{\partial\phi}\cdot\eta\right) \\ &\quad -\frac{1}{2}\mathrm{tr}\left(D(T,\phi)\cdot\frac{\partial D(T,\phi)}{\partial\phi}\cdot D(T,\phi)\cdot\frac{\partial D(T,\phi)}{\partial\phi}\right) \end{aligned} \quad (A.9)$$

Now if we define $X(t,\theta) = (H(t,\phi)', \frac{\partial H(t,\phi)}{\partial\phi})'$ and $X(T,\theta)$, $\tilde{X}(T,\theta)$, $\bar{X}(T,\theta)$ we can combine (A.7)-(A.9) as

$$I(\theta) = \tilde{X}(T,\theta)'D^{-1}(T,\theta)\tilde{X}(T,\theta) + F(T,\theta) \quad (A.10)$$

Therefore the Taylor expansion approximation of $I(\theta)$ for nonlinear mixed models becomes exact under CLLM.

APPENDIX B

Derivation of $A(T, \theta)$

In this section, we will compute the element of $A(T, \theta)$. To simplify the notation, we denote $X(t, \theta)$ as X and $D(T, \theta)$ as D. So D can be written as

$$D = \begin{pmatrix} F+B & B & \cdots & B \\ B & F+B & \ddots & \vdots \\ \vdots & \ddots & \ddots & B \\ B & \cdots & B & F+B \end{pmatrix}$$

where $F = X\Sigma^d X' + \sigma^2 I$ and $B = X\Sigma^s X'$. Then it is easy to show that the inverse of D can be written as

$$D^{-1} = \begin{pmatrix} U+V & V & \cdots & V \\ V & U+V & \ddots & \vdots \\ \vdots & \ddots & \ddots & V \\ V & \cdots & V & U+V \end{pmatrix}$$

where $U = F^{-1}$ and $V = -(F + mB)^{-1}BF^{-1}$. Further we have

$$\begin{aligned} V &= -(F + mB)^{-1}\frac{1}{m}(mB + F - F)F^{-1} \\ &= \frac{1}{m}((F + mB)^{-1} - F^{-1}) \end{aligned}$$

Then the diagonal block of $\frac{\partial D}{\partial \theta_i}\frac{\partial D^{-1}}{\partial \theta_j}$ is

$$\begin{aligned} \frac{\partial D}{\partial \theta_i}\frac{\partial D^{-1}}{\partial \theta_j} &= \frac{\partial (F + B)}{\partial \theta_i}\frac{\partial (U + V)}{\partial \theta_j} + (m - 1)\frac{\partial B}{\partial \theta_i}\frac{\partial V}{\partial \theta_j} \\ &= -\frac{1}{m}(\frac{\partial Q}{\partial \theta_i}Q^{-1}\frac{\partial Q}{\partial \theta_j}Q^{-1} + (m - 1)\frac{\partial P}{\partial \theta_i}P^{-1}\frac{\partial P}{\partial \theta_j}P^{-1}) \end{aligned}$$

where $P = X(T, \theta)\Sigma^d X(T, \theta)' + \sigma^2$ and $Q = P + X(T, \theta)\Sigma^s X(T, \theta)'$. Then the ij element of $A = -\frac{1}{2}(\frac{\partial vec(D)}{\partial \theta^t})^t \frac{\partial vec(D^{-1})}{\partial \theta^t}$ is

$$\begin{aligned} A_{ij} &= \mathrm{tr}\left(-\frac{1}{2}\frac{\partial D}{\partial \theta_i}\frac{\partial D^{-1}}{\partial \theta_j}\right) \\ &= \frac{1}{2}\mathrm{tr}\left(\frac{\partial Q}{\partial \theta_i}Q^{-1}\frac{\partial Q}{\partial \theta_j}Q^{-1} + (m - 1)\frac{\partial P}{\partial \theta_i}P^{-1}\frac{\partial P}{\partial \theta_j}P^{-1}\right) \end{aligned}$$

APPENDIX C

Derivation of the Inverse of the Information Matrix

In this section, we will prove that for $I(\theta) = nX^t D^{-1} X$ where $D = X\Sigma X^t + \sigma^2 I$ where

$$\Sigma = \begin{pmatrix} \Sigma^s + \Sigma^d & \Sigma^s & \cdots & \Sigma^s \\ \Sigma^s & \Sigma^s + \Sigma^d & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma^s \\ \Sigma^s & \cdots & \Sigma^s & \Sigma^s + \Sigma^d \end{pmatrix}$$

then $I^{-1}(\theta) = \frac{1}{n}\Sigma^s + \frac{1}{nm}(\Sigma^d + \sigma^2 (X^t X)^{-1})$

In the paper, we have shown that $I(\theta)$ is the information matrix for θ for the following hierarchical model:

$$\begin{aligned} Y_{ij} &\sim N(X\theta_{ij}, \sigma^2 I) \\ \theta_{ij} &\sim N(\theta_i, \Sigma_d) \\ \theta_i &\sim N(\theta, \Sigma_s) \end{aligned}$$

where $j = 1, \ldots, m$ and $i = 1, \ldots, n$. Then by MLE theory, we can compute the asymptotic variance of $\hat{\theta}$ by $I^{-1}(\theta)$. At the same time, since everything is assumed

normally distributed, we can directly compute the variance of $\hat{\theta}$. We can write the complete likelihood of the model as (less a multiplicative constant that does not involve $\theta_{ij}, \theta_i, \theta$)

$$\prod_i^n \int \prod_j^m \int \exp(-\frac{1}{2}((Y_{ij} - X\theta_{ij})' \frac{1}{\sigma^2}(Y_{ij} - X\theta_{ij}) + (\theta_{ij} - \theta_i)'\Sigma_d^{-1}(\theta_{ij} - \theta_i) + (\theta_i - \theta)\Sigma_s^{-1}(\theta_i - \theta)))d\theta_{ij}d\theta_i$$

Integrate with respect to θ_{ij}, we have

$$\prod_i^n \int \exp(-\frac{1}{2}(-\sum_j^m (X'Y_{ij} + \Sigma_d^{-1}\theta_i)'(\frac{1}{\sigma^2}X'X + \Sigma_d^{-1})^{-1}(X'Y_{ij} + \Sigma_d^{-1}\theta_i) + \theta_i'\Sigma_d^{-1}\theta_i + (\theta_i - \theta)\Sigma_s^{-1}(\theta_i - \theta)))d\theta_i \quad \text{(C.1)}$$

Using the fact that $\Sigma_d^{-1} - \Sigma_d^{-1}(\frac{1}{\sigma^2}X'X + \Sigma_d^{-1})\Sigma_d^{-1} = (\Sigma_d + \sigma^2(X'X)^{-1})^{-1}$, the above becomes

$$\prod_i^n \int \exp(-\frac{1}{2}(m \cdot \theta_i'(\Sigma_d + \sigma^2(X'X)^{-1})^{-1}\theta_i - 2\theta_i'(\frac{1}{\sigma^2}X'X + \Sigma_d^{-1})^{-1}\sum_j^m X'Y_{ij} + (\theta_i - \theta)\Sigma_s^{-1}(\theta_i - \theta)))d\theta_i$$

Then integrate with respect to θ_i

$$\exp(-\frac{1}{2}(-n(\sum_i^n (H + \Sigma_s^{-1}\theta)'(\Sigma_s^{-1} + m(\Sigma_d + \sigma^2(X'X)^{-1})^{-1})^{-1}(H + \Sigma_s^{-1}\theta) + n\theta'\Sigma_s^{-1}\theta)))$$

where $H = (\frac{1}{\sigma^2}X'X + \Sigma_d^{-1})^{-1}\sum_j^m X'Y_{ij}$. The quadratic term involving θ inside the exponential function is

$$-\frac{1}{2}\theta' n(\Sigma_s^{-1} - \Sigma_s^{-1}(\Sigma_s^{-1} + m(\Sigma_d + \sigma^2(X'X)^{-1})^{-1})^{-1}\Sigma_s^{-1})\theta$$

So the variance of $\hat{\theta}$ is $(n(\Sigma_s^{-1} - \Sigma_s^{-1}(\Sigma_s^{-1} + m(\Sigma_d + \sigma^2(X'X)^{-1})^{-1})^{-1}\Sigma_s^{-1}))^{-1} = \frac{1}{n}(\Sigma_s + \frac{1}{m}(\Sigma_d + \sigma^2(X'X)^{-1}))$.

It is easy to generalize the above results to the arbitrary level of variability. We can consider the hierarchy model:

$$\begin{aligned} Y_{i_1\cdots i_p} &\sim N(X\theta_{i_1\cdots i_p}, \sigma^2 I) \\ \theta_{i_1\cdots i_p} &\sim N(\theta_{i_2\cdots i_p}, \Sigma_1) \\ &\vdots \\ \theta_{i_p} &\sim N(\theta, \Sigma_p) \end{aligned}$$

Here $i_j = 1..n_j$ for $j = 1..p$. Then we set $A_1 = \sigma^2(X'X)^{-1}$, $A_{j+1} = \frac{1}{n_j}(\Sigma_j + A_j)$ for $j = 1..p-1$ and $H_{i_1\cdots i_p} = X'Y_{i_1\cdots i_p}$, $H_{i_{j+1}\cdots i_p} = \sum_{i_j=1}^{n_j}(\Sigma_j^{-1} + A_j^{-1})^{-1}H_{i_j\cdots i_p}$ for $j = 1..p-1$. Then we will show that after we have integrated out all $\theta_{i_q\cdots i_p}$ $q = 1..p-1$, the terms inside the exponential function involving is $\theta_{i_{q+1}\cdots i_p}$ is (less $\frac{1}{2}$)

$$\begin{aligned} - & \sum_{i_q=1}^{n_q}((H_{i_q\cdots i_p} + \Sigma_q^{-1}\theta_{i_{q+1}\cdots i_p})'(\Sigma_q^{-1} + A_q^{-1})^{-1}(H_{i_q\cdots i_p} + \Sigma_q^{-1}\theta_{i_{q+1}\cdots i_p})) \\ + & \; n_q\theta'_{i_{q+1}\cdots i_p}\Sigma_q^{-1}\theta_{i_{q+1}\cdots i_p} \\ + & \; (\theta_{i_{q+1}\cdots i_p} - \theta_{i_{q+2}\cdots i_p})'\Sigma_{q+1}^{-1}(\theta_{i_{q+1}\cdots i_p} - \theta_{i_{q+2}\cdots i_p}) \end{aligned}$$

We prove this statement by induction:

1. By referring to (C.1), the statement is true for $q = 1$.

2. Assume the statement is true for $q = k$, Then we need to show the expression after we have integrated out $\theta_{i_{k+1}\cdots i_p}$. By induction assumption, the expression involving $\theta_{i_{k+1}\cdots i_p}$ after integrated out $\theta_{i_k\cdots i_p}$ is

$$
\begin{aligned}
&- \sum_{i_k=1}^{n_k}((H_{i_k\cdots i_p}+\Sigma_k^{-1}\theta_{i_{k+1}\cdots i_p})'(\Sigma_k^{-1}+A_k^{-1})^{-1}(H_{i_k\cdots i_p}+\Sigma_k^{-1}\theta_{i_{k+1}\cdots i_p}))\\
&+ n_k\theta'_{i_{k+1}\cdots i_p}\Sigma_k^{-1}\theta_{i_{k+1}\cdots i_p}\\
&+ (\theta_{i_{k+1}\cdots i_p}-\theta_{i_{k+2}\cdots i_p})'\Sigma_{k+1}^{-1}(\theta_{i_{k+1}\cdots i_p}-\theta_{i_{k+2}\cdots i_p})
\end{aligned}
$$

which can be written as

$$
\begin{aligned}
&\theta'_{i_{k+1}\cdots i_p}(\Sigma_{k+1}^{-1}+n_k(\Sigma_k^{-1}-\Sigma_k^{-1}(\Sigma_k^{-1}+A_k^{-1})^{-1}\Sigma_k^{-1}))\theta_{i_{k+1}\cdots i_p}\\
&- 2\theta'_{i_{k+1}\cdots i_p}(\sum_{i_k=1}^{n_k}(\Sigma_k^{-1}+A_k^{-1})^{-1}H_{i_k\cdots i_p}+\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p})\\
&+ \theta'_{i_{k+2}\cdots i_p}\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p}
\end{aligned}
$$

substitute the definition of A_{j+1} and H_{j+1}, we have

$$\theta'_{i_{k+1}\cdots i_p}(\Sigma_{k+1}^{-1}+A_{j+1}^{-1})\theta_{i_{k+1}\cdots i_p}-2\theta'_{i_{k+1}\cdots i_p}(H_{i_{k+1}\cdots i_p}+\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p})+\theta'_{i_{k+2}\cdots i_p}\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p}$$

Then if we integrate out $\theta_{i_{k+1}\cdots i_p}$ we have

$$(H_{i_{k+1}\cdots i_p}+\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p})'(\Sigma_{k+1}^{-1}+A_{j+1}^{-1})^{-1}(H_{i_{k+1}\cdots i_p}+\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p})+\theta'_{i_{k+2}\cdots i_p}\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p}$$

We collect all the terms involving $\theta_{i_{k+2}\cdots i_p}$,

$$
\begin{aligned}
&- \sum_{i_{k+1}=1}^{n_{k+1}}(H_{i_{k+1}\cdots i_p}+\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p})'(\Sigma_{k+1}^{-1}+A_{j+1}^{-1})^{-1}(H_{i_{k+1}\cdots i_p}+\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p})\\
&+ n_{k+1}\theta'_{i_{k+2}\cdots i_p}\Sigma_{k+1}^{-1}\theta_{i_{k+2}\cdots i_p}\\
&+ (\theta_{i_{k+2}\cdots i_p}-\theta_{i_{k+3}\cdots i_p})'\Sigma_{k+2}^{-1}(\theta_{i_{k+2}\cdots i_p}-\theta_{i_{k+3}\cdots i_p})
\end{aligned}
$$

In this way, we have shown the statement is also true for $q=k+1$.

Then after we integrate out all random components, we are left with

$$\exp(-\frac{1}{2}(\sum_{i_k=1}^{n_p}((H_{i_p}+\Sigma_p^{-1}\theta)'(\Sigma_p^{-1}+A_p^{-1})^{-1}(H_{i_p}+\Sigma_p^{-1}\theta))+n_p\theta'\Sigma_p^{-1}\theta))$$

The quadratic term involving θ is

$$\theta' n_p(\Sigma_p^{-1}-\Sigma_p^{-1}(\Sigma_p^{-1}+A_p^{-1})^{-1}\Sigma_p^{-1})\theta$$

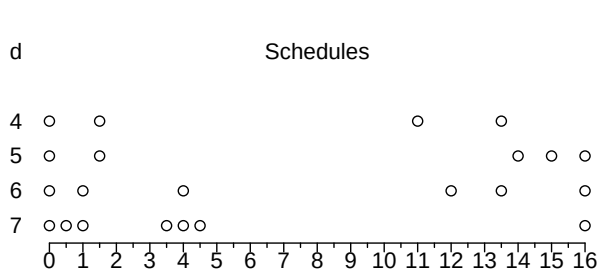
So the $Var(\hat{\theta}) = (n_p(\Sigma_p^{-1}-\Sigma_p^{-1}(\Sigma_p^{-1}+A_p^{-1})^{-1}\Sigma_p^{-1}))^{-1} = \frac{1}{n_p}(\Sigma_p+A_p) = \frac{1}{n_p}(\Sigma_p + \frac{1}{n_{p-1}}(\Sigma_{p-1}+\cdots+\frac{1}{n_1}(\Sigma_1+\sigma^2(X'X)^{-1}))\cdots)$.

APPENDIX D

Optimal Design and Cost Analysis Based on Piecewise Linear Model

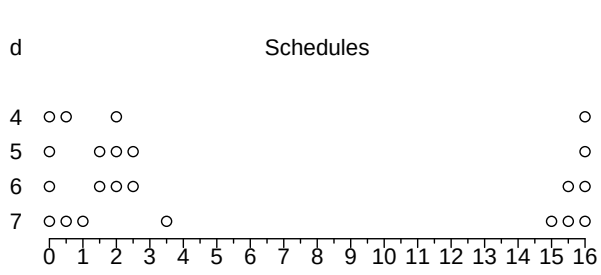
In this section, we will present the figures for the design under the piecewise linear model. The interpretations of these figures are the same as those for the nonlinear model, which have been discussed in detail in the paper. The designs under two models share many similarities with only one major exception: more samples per day is more cost effective for measuring features under nonlinear model but not for the piecewise linear model.

We also notice that more samples per day is more cost effective for measuring features under nonlinear model but not for the piecewise linear model.



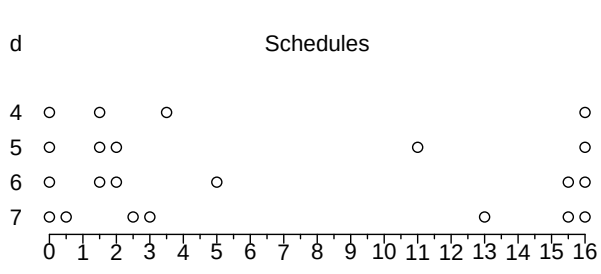
	Variance					Cost Adjusted Variance 10^−2				
d	1 Day	2 Day	3 Day	4 Day	5 Day	1 Day	2 Day	3 Day	4 Day	5 Day
4	0.477	0.349	0.306	0.285	0.272	0.793	**0.723**	0.76	0.824	0.898
5	0.477	0.349	0.306	0.285	0.272	0.812	0.75	0.797	0.87	0.953
6	0.477	0.349	0.306	0.285	0.272	0.83	0.779	0.833	0.915	1.01
7	0.477	0.349	0.306	0.285	0.272	0.85	0.807	0.87	0.96	1.06

(a) Baseline



	Variance					Cost Adjusted Variance 10^−2				
d	1 Day	2 Day	3 Day	4 Day	5 Day	1 Day	2 Day	3 Day	4 Day	5 Day
4	0.142	0.141	0.140	0.14	0.140	**0.236**	0.291	0.348	0.404	0.461
5	0.142	0.140	0.14	0.140	0.140	0.241	0.302	0.364	0.426	0.489
6	0.141	0.14	0.140	0.140	0.139	0.246	0.312	0.38	0.448	0.516
7	0.142	0.141	0.14	0.140	0.140	0.253	0.325	0.398	0.471	0.545

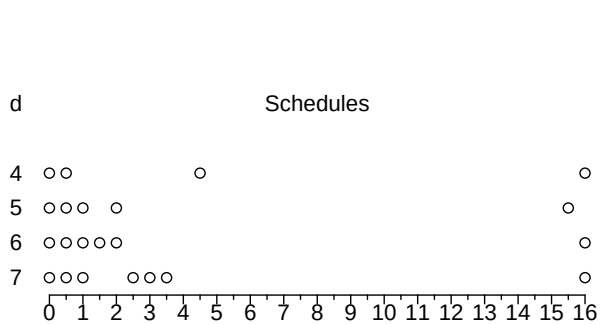
(b) Evening Decline



	Variance					Cost Adjusted Variance				
d	1 Day	2 Day	3 Day	4 Day	5 Day	1 Day	2 Day	3 Day	4 Day	5 Day
4	1861	1845	1840	1837	1836	**30.9**	38.2	45.6	53.1	60.6
5	1854	1842	1838	1836	1834	31.5	39.6	47.8	56	64.2
6	1851	1840	1837	1835	1834	32.2	41	50	58.9	67.8
7	1850	1840	1836	1835	1834	32.9	42.5	52.2	61.8	71.5

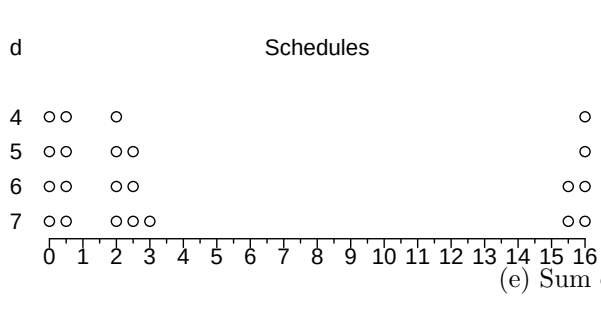
(c) AUC

Figure D.1: Optimal Design under the Piecewise Linear Model



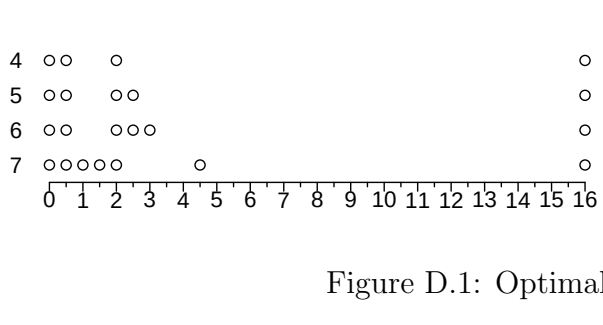
	Variance					Cost Adjusted Variance 10^−2				
d	1 Day	2 Day	3 Day	4 Day	5 Day	1 Day	2 Day	3 Day	4 Day	5 Day
4	0.516	0.266	0.183	0.141	0.116	0.857	0.551	0.453	0.408	0.383
5	0.445	0.230	0.159	0.123	0.102	0.756	0.495	0.413	0.376	**0.356**
6	0.441	0.229	0.158	0.122	0.101	0.768	0.51	0.429	0.393	0.374
7	0.441	0.229	0.158	0.122	0.101	0.786	0.529	0.448	0.413	0.395

(d) CAR



	Variance					Cost Adjusted Variance 10^−2				
d	1 Day	2 Day	3 Day	4 Day	5 Day	1 Day	2 Day	3 Day	4 Day	5 Day
4	39.9	38.9	38.6	38.4	38.3	**66.3**	80.5	95.6	111	126
5	39.5	38.7	38.4	38.3	38.2	67.1	83.1	99.8	117	134
6	39.4	38.6	38.4	38.2	38.2	68.5	86.1	104	123	141
7	39.2	38.5	38.3	38.2	38.1	69.8	89	109	129	149

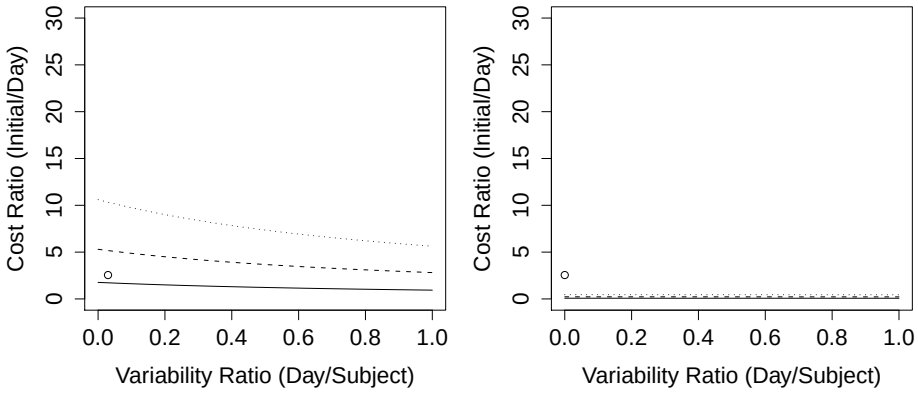
(e) Sum of Variances of Cortisol Features



	Variance					Cost Adjusted Variance 10^−2				
d	1 Day	2 Day	3 Day	4 Day	5 Day	1 Day	2 Day	3 Day	4 Day	5 Day
4	39.9	38.9	38.6	38.4	38.3	**66.3**	80.5	95.6	111	126
5	39.5	38.7	38.4	38.3	38.2	67.1	83.1	99.8	117	134
6	39.3	38.6	38.4	38.2	38.2	68.4	86	104	123	141
7	39.3	38.6	38.3	38.2	38.1	69.9	89.1	109	129	149

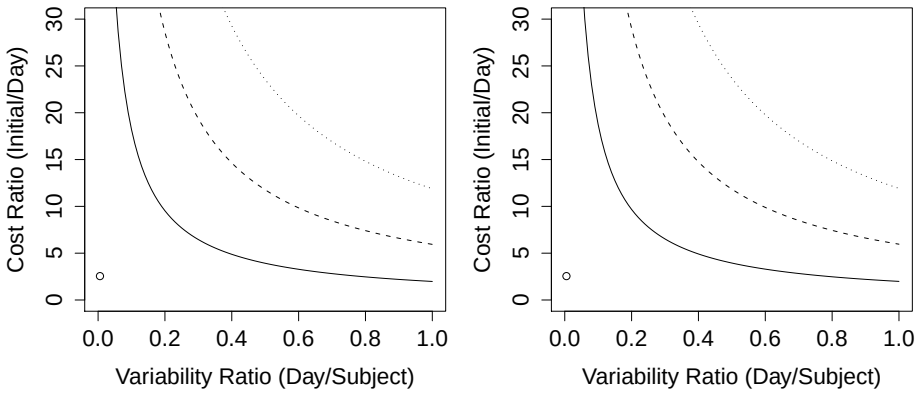
(f) Prediction

Figure D.1: Optimal Design under the Piecewise Linear Model (Continued)



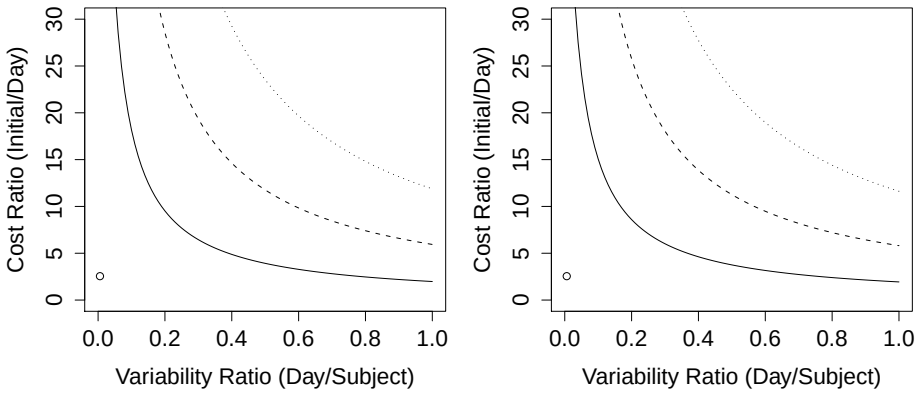
(a) Baseline

(b) CAR



(c) Evening Decline

(d) AUC



(e) Sum of Variances of Cortisol Features

(f) Prediction

Figure D.2: Cost Ratio vs. Variability Ratio (Piecewise Linear Model)

APPENDIX E

Estimation of FPCA

A.1 Previous Literature

There are generally two major approaches for estimating FPCA, depending on whether the covariance kernel $Cov(f_i(t), f_i(s))$ is used.

When *Rice and Silverman* (1991) first introduced FPCA, they provided an estimation method for the principal components by identifying $\beta_k(t)$'s as the eigenfunctions of the covariance kernel $Cov(f_i(t), f_i(s))$. This method works well for functional data for which $y_i(t)$ is regularly and densely sampled and $Cov(f_i(t), f_i(s))$ could be accurately estimated. For longitudinal data, however, $y_i(t)$ is often irregularly and sparsely sampled. In this case, the variability in estimating $Cov(f_i(t), f_i(s))$ is relatively high, which results in poor estimation of $\beta_j(t)$. In order to address this issue, *Yao et al.* (2005) proposed using the local smoothing technique to improve the estimation of $Cov(f_i(t), f_i(s))$ and subsequently $\beta_k(t)$. The asymptotic properties of this method were later established by *Hall et al.* (2006). Nevertheless, *Peng* (2009) points out that this method still has some limitations: the estimated covariance kernel is not necessarily positive semi-definite and the estimated residual variance $\hat{\sigma}^2$ could be negative. Given the difficulty in estimating $Cov(f_i(t), f_i(s))$, methods that avoid the

use of $Cov(f_i(t), f_i(s))$ have been considered. *James et al.* (2000) introduce the the reduced rank model and develop an EM algorithm for obtaining $\beta_k(t)$ without estimating $Cov(f_i(t), f_i(s))$. *Peng* (2009) recognized the EM algorithm in *James et al.* (2000) does not take advantage of the orthonormality of $\beta_k(t)$'s. Thus they propose a Newton Raphson algorithm that maximizes the incomplete likelihood over the Stiefel manifold which consists of orthonormal vectors. Simulations suggest that the Newton Raphson algorithm is more accurate than EM algorithm and the local smoothing approach. However, the Newton Raphson algorithm they proposed is challenging to implement.

Despite their differences, all the methods discussed above assume $\beta_k(t)$ is modeled by a linear combinations of spline bases, i.e. $\beta_k(t) = \sum_{l=1}^{q} b_k(t)\theta_{jl}$. This assumption leads an important question: what is the optimal number of spline bases? While many authors have suggested using cross validation type statistics to find the best number of spline bases, the sparse nature of longitudinal data often restricts the number to be quite small. In this case, the estimated $\hat{\beta}_k(t)$ tends to be "wiggly" and unstable. A better solution is to employ sufficiently many bases and impose some regularity conditions on $\beta_k(t)$ at the same time. *Silverman* (1996) propose a smoothed version of their original FPCA by defining a new inner product and subsequently a new norm for $\beta_j(t)$: $< \beta_k, \beta_{k'} >_\lambda = \int \beta_k(t)\beta_{k'}(t)dt + \lambda \int \beta_k''(t)\beta_{k'}''(t)dt$ and $||\beta_k||_\lambda^2 =< \beta_k, \beta_k >_\lambda$. Here λ is the smoothing parameter. Although this approach fits into the original FPCA framework nicely, the orthogonormality of $\beta_k(t)$ under $<, >_\lambda$ becomes difficult to interpret since it now depends on the smoothing parameter λ. As an extension to *Yao et al.* (2005), *Yao and Lee* (2006) proposed a penalized spline models for FPCA. However, this approach only performs smoothing for the mean of the longitudinal process but not the principal components $\beta_k(t)$. In considering the joint modeling of pairs of sparse functional data, *Zhou et al.* (2008) suggests a smoothness penalty, $\lambda \int \beta_k''(t)^2 dt$, on the functional principal components $\beta_k(t)$, which reduce the mean

squared error in estimating $\beta_k(t)$ especially when the data are sparse and irregular.

A.2 Our Approach

Given the rich literature on FPCA, we want to find an estimation method that fits our needs. We prefer the EM approach for the reduced rank model because it can handle the sparsity of the longitudinal data well, and automatically computes the loadings on each principal component for every subject. We would also like to employ penalized spline to reduce the artifacts due to the limited number of spline bases. Lastly, we want to take advantage of the orthonormality of $\beta_k(t)$ to improve the estimation efficiency. In the following paragraphs, we will describe a modification of *James et al.* (2000) that incorporates the smoothing penalty *Zhou et al.* (2008):

For subject i, let the sampling times be $T_i = (t_{i1}, \dots t_{in_i})$ and the observations be $Y_i = (y_i(t_{i1}), \dots y_i(t_{in_i}))'$. We model the principal components by linear combinations of spline basis $b_l(t)$: $\beta_k(t) = \sum_{l=1}^{q} b_l(t)\theta_{lk}$. Let B_i be a basis matrix such that $(B_i)_{jl} = b_l(t_{ij})$; $\theta_k = (\theta_{1k}, \dots, \theta_{qk})'$; Θ be the coefficient matrix such that $\Theta = (\theta_1, \dots \theta_r)$ and finally $\alpha_i = (\alpha_{i1}, \dots \alpha_{ir})'$. Then the reduced rank model can be written in the matrix form:

$$Y_i = f(T, \eta) + B_i\Theta\alpha_i + \epsilon_i \tag{E.1}$$

where $\alpha_i \sim N(0, D)$, $\epsilon_i \sim N(0, \sigma^2 I)$. For the purpose of identification, we impose the orthonormal constraints: $\Theta'\Theta = I$, $\int b_l(t)b_{l'}(t)dt = \delta_{ll'}$. Then we have the marginal distribution of Y_i:

$$Y_i \sim N(f(T, \eta), \Sigma_i) \qquad \Sigma_i = B_i\Theta D\Theta' B_i' + \sigma^2 I$$

and the observed log likelihood given data

$$\sum_{i=1}^{n} -\frac{n_i}{2}\log 2\pi - \frac{1}{2}\log\Sigma_i - \frac{1}{2\sigma^2}(Y_i - f(T,\eta))'\Sigma_i^{-1}(Y_i - f(T,\eta))$$

It is difficult to find the MLE for the observed likelihood given its complexity. Instead, we employ the EM algorithm and work with the complete data likelihood with α_i assumed to be known:

$$\sum_{i=1}^{n} -\frac{n_i}{2}\log 2\pi - \frac{n_i}{2}\log\sigma^2 - \frac{1}{2}\log D - \frac{1}{2\sigma^2}(Y_i - f(T,\eta) - B_i\Theta\alpha_i)'(Y_i - f(T,\eta) - B_i\Theta\alpha_i) - \frac{1}{2}\alpha_i' D^{-1}\alpha_i$$

We could encourage the smoothness of $\beta_k(t)$ by introducing a second derivative penalty $\lambda\int \beta_k''(t)^2 dt$ with λ being a positive smoothing parameter. To write the penalty in terms of θ_k, we let H be a matrix such $H_{ll'} = \int b_l(t)b_{l'}(t)dt$. Now we have $\lambda\int \beta_k''(t)^2 dt = \lambda\theta_k' H\theta_k$. Then we could obtain a smoothed version of FPCA by maximizing the penalized log likelihood

$$\begin{aligned} Q &= \sum_{i=1}^{n}\{-\frac{n_i}{2}\log 2\pi - \frac{n_i}{2}\log\sigma^2 - \frac{1}{2}\log D \\ &\quad -\frac{1}{2\sigma^2}(Y_i - f(T,\eta) - B_i\Theta\alpha_i)'(Y_i - f(T,\eta) - B_i\Theta\alpha_i) - \frac{1}{2}\alpha_i' D^{-1}\alpha_i\} \\ &\quad -\sum_{k=1}^{r}\lambda\theta_k' H\theta_k \end{aligned}$$

(The penalty term on θ_k is equivalent to incorporating a prior $\theta_k \sim N(0, \frac{1}{2\lambda}H^{-1})$ in Bayesian statistics.)

Green (1990) show that EM algorithm is also applicable to the penalized log likelihood when the penalty term does not involve latent variables. Hence we can employ EM algorithm for maximizing Q. The E-step and M-step of the EM algorithm are as follows. For the E-step, we denote $E(\alpha_i|Y_i,\eta,\Theta,D,\sigma^2)$ by $\hat{\alpha}_i$ and $E(\alpha_i\alpha_i'|Y_i,\eta,\Theta,D,\sigma^2)$ by $\alpha_i\alpha_i'$. We have

$$\begin{aligned} \hat{\alpha}_i &= (\sigma^2 D^{-1} + \Theta' B_i' B_i\Theta)^{-1}\Theta' B_i'(Y_i - f(T,\eta)) \\ \hat{\alpha_i\alpha_i'} &= \hat{\alpha}_i\hat{\alpha}_i' + (D^{-1} + \Theta' B_i' B_i\Theta/\sigma^2)^{-1} \end{aligned}$$

For the M-step, we could maximize Q iteratively over $\eta, \Theta, D, \sigma^2$. The formula for η,

D, and σ^2 are identical to those in *James et al.* (2000) and they are omitted here.

The interesting question is how to maximize Q while preserving the orthonormality of the columns of Θ. Our solution to this problem is a reparameterization based on singular value decomposition (SVD). In this setting, the optimization is carried out column by column of Θ. Suppose θ_k is being considered and the rest of the columns, denoted by $\Theta_{(k)}$ are kept fixed. Then θ_k should be orthogonal to the column space of $\Theta_{(k)}$. We perform a SVD on $\Theta_{(k)}$ and we have

$$\Theta_{(k)} = U \cdot S \cdot V'$$

where U is $q \times q$ orthogonal matrix, S is diagonal matrix and V is $(r-1) \times (r-1)$ orthogonal matrix. Let $(u_1, \ldots u_q)$ be the columns of U. Then $(u_r, \ldots u_q)$ spans the space orthogonal to the column space of $\Theta_{(k)}$. Then we could parameterize θ_k as

$$\theta_k = \sum_{l=r}^{q} u_l p_l = \tilde{U} P$$

where $\tilde{U} = (u_r, \ldots u_q)$ and $P = (p_r, \ldots, p_q)'$. Under this parameterization, θ_k is always orthogonal to the column space of $\Theta_{(k)}$ and there is no restriction on P. Now by focusing on the terms in Q that involve θ_k, we only need to minimize:

$$\sum_{i=1}^{N} (Y_i - f(T,\eta) - B_i \Theta \alpha_i)'(Y_i - f(T,\eta) - B_i \Theta \alpha_i) + \tilde{\lambda} \theta_k' H \theta_k$$

where $\tilde{\lambda} = 2\sigma^2 \lambda$. We can rewrite it as

$$\begin{aligned} & \sum_{i=1}^{N} ((Y_i - f(T,\eta) - B_i \Theta_{(k)} \alpha_{i(k)}) - B_i \theta_k)'((Y_i - f(T,\eta) - B_i \Theta_{(k)} \alpha_{i(k)}) - B_i \theta_k) + \tilde{\lambda} \theta_k' H \theta_k \\ = & \sum_{i=1}^{N} ((Y_i - f(T,\eta) - B_i \Theta_{(k)} \alpha_{i(k)}) - B_i \tilde{U} P)'((Y_i - f(T,\eta) - B_i \Theta_{(k)} \alpha_{i(k)}) - B_i \tilde{U} P) \\ & + \tilde{\lambda} P' \tilde{U}' H \tilde{U} P \end{aligned}$$

The closed form solution for P to minimize the above equation is

$$\hat{P} = \left(\sum_{i=1}^{N} \hat{\alpha}_{ik}^2 \tilde{U}' B_i' B_i \tilde{U} + \tilde{\lambda} \tilde{U}' H \tilde{U} \right)^{-1} \sum_{i=1}^{N} \tilde{U}' B_i' (\hat{\alpha}_{ik} (Y_i - f(T, \eta)) - \sum_{l \neq k} \alpha_{ik} \hat{\alpha}_{il} B_i \theta_l)$$

Then we have $\hat{\theta}_k = \tilde{U} \hat{P}$ and we can normalize $\hat{\theta}_k$ by dividing it by its norm. We could repeat the same procedures for $k = 1, \ldots, r$ iteratively until convergence is reached.

The reparameterization based on SVD allows us to optimize Q under the constraint that the columns of Θ are orthonormal. This procedure works when the number of columns, i.e., the number of principal components is strictly larger than the number of spline bases, which is generally true. This method is much simpler to understand and implement than working with Stiefel manifold. This method could potentially be extended to other optimization problems with orthonormal restrictions.

APPENDIX F

Selecting the Number of Components and the Smoothing Parameter

In this section, we will discuss using cross validation score $s(r, \lambda)$ to select the smoothing parameter λ and the number of components r in the FPCA model. We will use the simulation examples (Sim A and Sim B) to illustrating the approach.

The cross validation score $s(r, \lambda)$ is computed by k-fold cross validation. We randomly and evenly split the preliminary data in the simulation into k groups. We treat one group as testing data set and the rest $k - 1$ groups as the training data set. We use training data to estimate a FPCA model for the given r and λ. Then we compute the log likelihood of the testing data given the estimated model. We repeat the process k times so that every group becomes the testing data exactly once. Then $s(r, \lambda)$ is computed as the average of log likelihoods across the k testing sets. Because of the normal assumption, the $s(r, \lambda)$ is equivalent to negative of the mean square error, less some constant. In general, the higher the $s(r, \lambda)$, the better the model fit. In the simulations and other two real data applications, we specify $k = 10$.

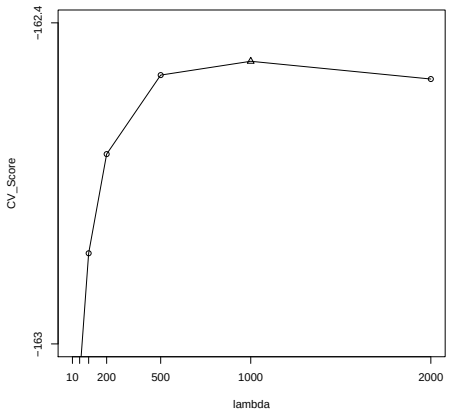
In the simulation, we consider smoothing parameter λ ranges from 10 to 2000. The appropriate λ for a FPCA model with r components is chosen as $\lambda_r^* = \text{argmax}_\lambda s(r, \lambda)$. Figure F.1 plots $s(r, \lambda)$ vs. λ for $r = 1, 2, 3$ for a sample of the simulated dataset from

Sim A and Sim B. The highest $s(r, \lambda)$ in each case is marked by a triangle and the corresponding λ is chosen as λ_r^*.

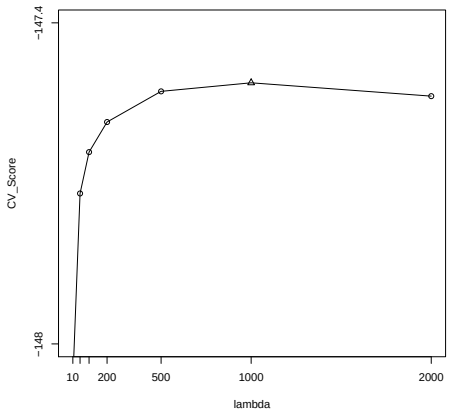
Let $s(r) = s(r, \lambda_r)$, i.e. the CV score for a model with r principal components and the appropriate smoothing parameter λ_r. Because the model with r components is nested within the model with $r + 1$ components, $s(r) = s(r, \lambda_r)$ almost always increases as r increases. Therefore maximizing the CV score $s(r)$ does not always lead to a parsimonious model in practice. In this case, we use a "scree plot" as a tool to visually select the appropriate number of components (*Johnson and Wichern*, 2007). In a scree plot, the CV score $s(r)$ is plotted against r and we are interested in the elbow point r^* where the improvement of $s(r)$ after r^* is relatively much smaller than those before r^*. In other words, the model fit will not significantly improve if we already have at least r^* principal components in the model. Figure F.2a and F.2b present scree plots for data sets simulated for Sim A and Sim B. In general, $s(r)$ increases substantially from $r = 1$ to $r = 2$ but there is virtually no improvement in $s(r)$ from $r = 2$ to $r = 3$. So it appears that $r = 2$ principal components are sufficient for the data in the simulation replicates shown in Figure F.2a and F.2b.

While the scree plot is simple to understand, it bears some subjective influence from the analyst and cannot be implemented in a simulation scenario where analyst intervention is absent. Inspired by the scree plot, we propose an objective rule based approach. We set a threshold b for negligible improvement and select the appropriate r for the data as $r^* = \min\{r | \frac{s(r+1)-s(r)}{s(r)} \leq b\}$, i.e. the smallest r such that the improvement in CV score by adding one more component is less than the threshold for negligible improvement. The rule based approach can be clearly defined and carried out in the simulation without outside intervention. To specify the threshold of negligible improvement, we could consult with the investigators or refer to the scree plot. For example, in Figure F.2a and F.2b, we notice from the scree plots that improvement in CV score is generally larger than 1% from $r = 1$ to $r = 2$ and less

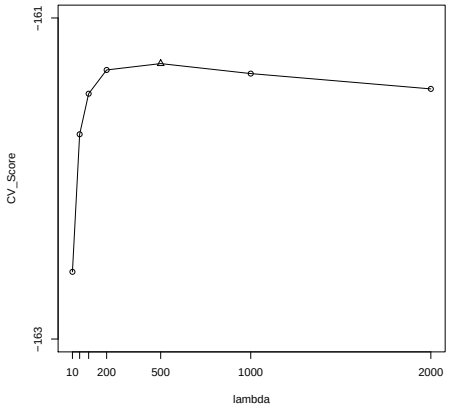
than 1% from $r = 2$ to $r = 3$. Therefore we set the threshold to be $b = 1\%$ for negligible improvement in this simulation.



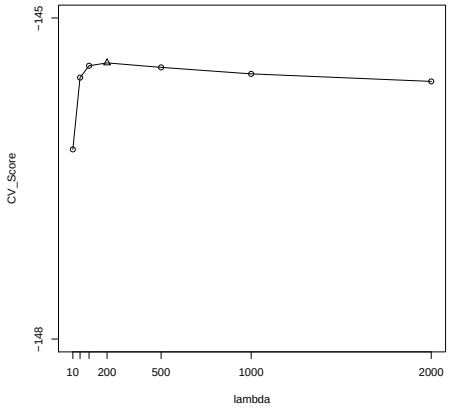
(a) Sim A: r=1



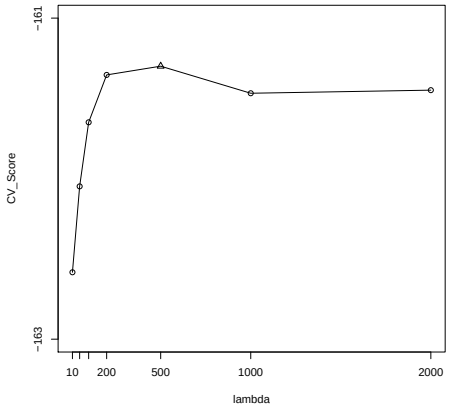
(b) Sim B: r=1



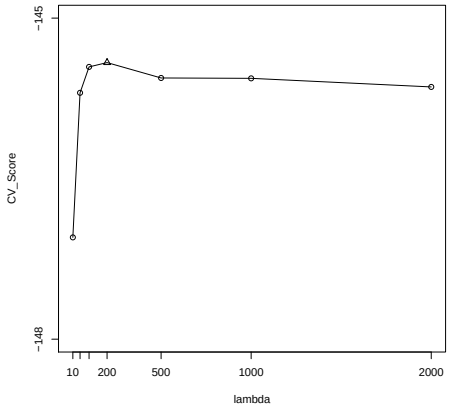
(c) Sim A: r=2



(d) Sim B: r=2

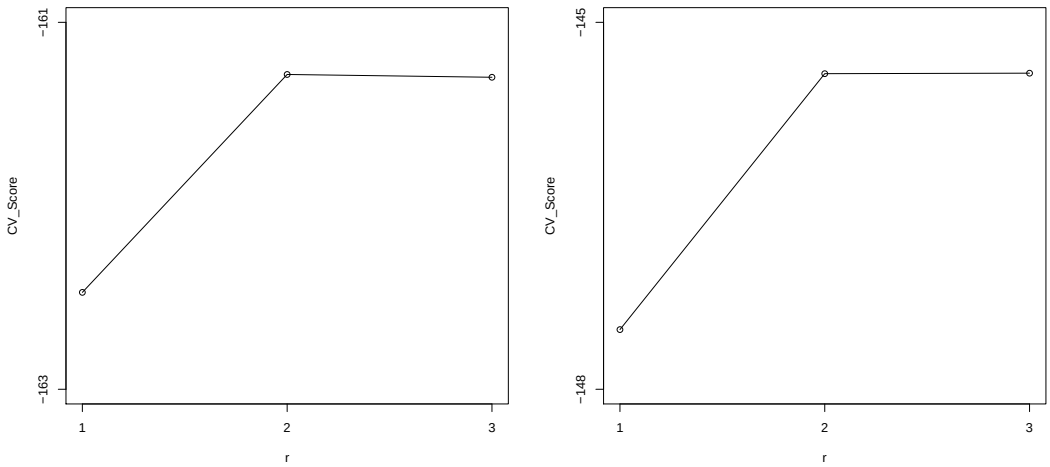


(e) Sim A: r=3



(f) Sim B: r=3

Figure F.1: CV score vs. λ for various choice of r. The highest CV score $s(r, \lambda)$ is marked by the triangles.



(a) Replicate 657 in Sim A

(b) Replicate 981 in Sim B

Figure F.2: Scree plots for selected replicates in the simulations

APPENDIX G

Algorithm for Identifying the Optimal Schedule

Let $f(T) = f(t_1, \cdots, t_n) > 0$ be the objective function with $T = (t_1, \cdots, t_n)$ being the sampling schedule. The maximization is with respect to the sampling schedule T and the values of t_i $i = 1, \ldots, n$ are taken (without replacement) from the set of feasible sampling times S. If S contains n_S elements, there are $\binom{n_S}{n}$ sampling schedules in the pool of all candidate schedules S_c, so this number can be extremely large even if n_S and n are only of moderate size, which makes almost impossible to enumerate all candidate sampling schedules for the purpose of maximization, particularly in a simulation setting. Therefore, in the following we describe a more efficient maximization algorithm based on the Metropolis-Hastings algorithm (*Metropolis et al.*, 1953; *Hastings*, 1970).

Algorithm: Select an initial sampling schedule T_0 from the pool of candidate schedules S_c. Also set up two storage arrays x_f and x_T of length M where M is the total number iterations to be run. Let T_{k-1} denote the sampling schedules selected at the $k-1$ iteration.

During the kth iteration:

(a) Let G_{k-1}, denote the set of feasible sampling times not included in the sampling schedule T_{k-1}(i.e. the complement of T_{k-1} with respect to S.

(b) Randomly select one sampling time from T_{k-1} and replace it with a sampling time randomly selected from G_{k-1} to create a new sampling schedule T_{temp}.

(c) Compute $f(T_{k-1})$ and $f(T_{temp})$.

(d) If $f(T_{temp}) > f(T_{k-1})$ then we let $T_k = T_{temp}$.

(e) Otherwise, we generate a random number $q \sim Binomial(\alpha)$. If $q = 1$, then $T_k = T_{temp}$, otherwise $T_k = T_{k-1}$.

(f) Store the sampling schedule T_k in $x_T[k]$ and objective value $f(T_k)$ in $x_f[k]$.

Repeat the iteration for M times. Then we identify the maximum objective values from x_f and corresponding sampling schedule from x_T.

In this algorithm, we are treating the $f(T)$ as the probability function (less a constant) of a multivariate distribution of $t_1, \cdots, t_n$. The distribution can be simulated with the Metropolis-Hastings algorithm and a uniform proposal distribution. Since the mode of the distribution is identical to the maximum of the objective function $f(\cdot)$ less a constant, we are guaranteed to reach the maximum if the Metropolis-Hastings algorithm has run long enough to converge. So the optimal objective function and the corresponding optimal schedule will appear in x_f and x_T with probability 1. Many methods for checking the convergence of Markov chains have been developed and *Cowles and Carlin* (1996) provides a comprehensive review for these methods.

APPENDIX H

Estimating Parametric Mixed Model with the R package nlme

The nlme package is used to estimate linear and nonlinear mixed effect models. We will provide an example regarding how we use nlme to estimate the nonlinear mixed effect models in the simulation.

The mathematical model can be summarized as follows:

$y_{ij} = \eta^{0i} + \eta^{1i} t_j + \eta^2 t \cdot \exp(-\eta^3 t) + \epsilon_{ij}$

where $\epsilon_{ij} \sim N(0, \sigma^2)$; the random effects are $\begin{pmatrix} \eta^{0i} \\ \eta^{1i} \end{pmatrix} \sim N(\begin{pmatrix} \eta^0 \\ \eta^1 \end{pmatrix}, \Sigma)$. The fixed effect parameters are $\eta = (\eta^0, \eta^1, \eta^2, \eta^3)$.

The code to estimate the model can be summarized as:

```
fit=nlme(y~eta0+eta1*time+eta2*time*exp(-eta3*time),
data=data,
fixed=eta0+eta1+eta2+eta3~1,
random=eta0+eta1~1—ID,
start=c(2.264, -0.1152, 1.1464, 0.6682) )
```

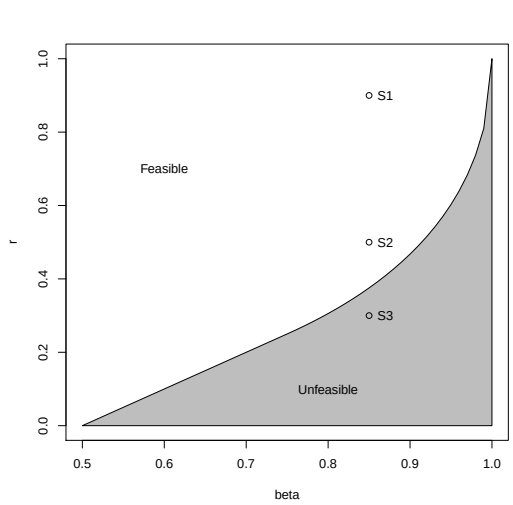
The meaning of the arguments are as follow:

- The 1st argument is the nonlinear model formula.
- The 2nd argument is data set for estimation.
- The 3rd argument is to specify the fixe effect.
- The 4th argument is to specify random effect with group variable being ID.
- The 5th argument is to specify the starting values for the fixed effect parameters.

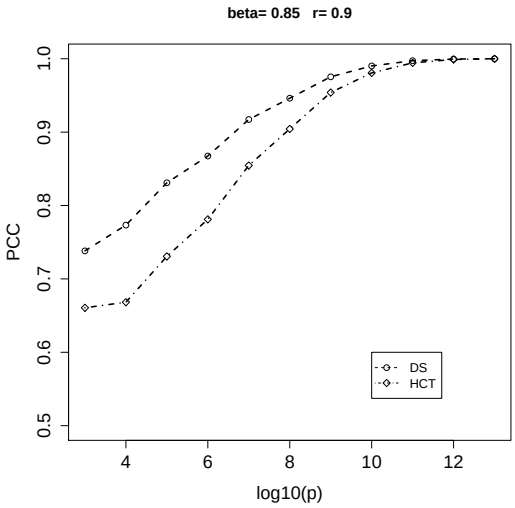
APPENDIX I

Simulation to Verify the Asymptotic Properties of Higher Criticism Threshold Classifier

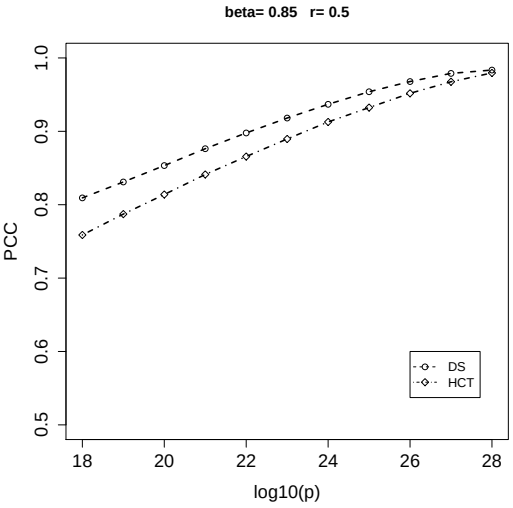
These simulation scenarios investigate the asymptotic feasibility of the classification problem under the rare-and-weak model. The values of strength parameter r and sparsity parameter β for scenario S1, S2 and S3 are shown in the phase diagram (Figure I.1a). S1 and S2 are in the feasible region while S3 is not. The sample size n is related to p by $n = c \times (\log(p))^{\gamma}$ with $c = 5$ and $\gamma = 1$. The PCC estimated by the DS method and the HCT method are computed. The PCC by CV is omitted because it requires too much resource and time to compute for very large p. Figure I.1b-I.1d suggest that for both DS and HCT method, when the total number of features $p \to \infty$, the PCC in S1 and S2 will go to 1 while the PCC in S3 will goes to 0.5. These results are consistent with the theory for the asymptotic feasibility.



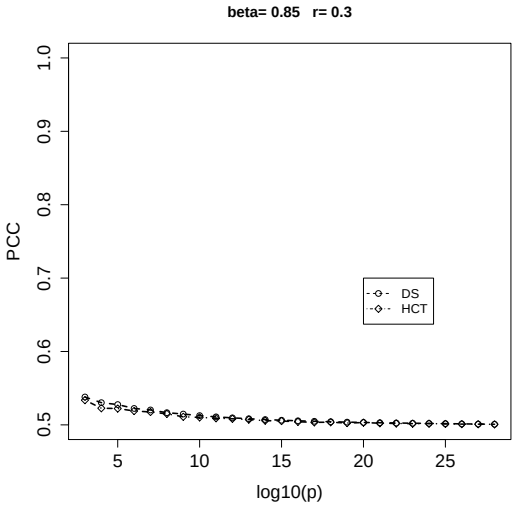
(a) Values of r and β in Simulation S1 S2 S3



(b) PCC vs. Sample Size for Simulation S1 Shown in the Phase Diagram



(c) PCC vs. Sample Size for Simulation S2



(d) PCC vs. Sample Size for Simulation S3

Figure I.1: Simulations for Evaluating Asymptotic Feasibility

APPENDIX J

Derivation of the Order-Statistics-Based Algorithm for Sampling P-values

The efficient algorithm to simulate the $k = \lceil p\alpha_0 \rceil$ smallest p-values among all the features is broken down into three steps:

I Simulate $(\pi_1, \ldots, \pi_m)$

II Simulate the k smallest values from the vector $(\pi_{m+1}, \ldots, \pi_p)$ and denote them by $(v_1, \ldots, v_k)$.

III Identify the k smallest values from the vector $(\pi_1, \ldots, \pi_m, v_1, \ldots, v_k)$.

Step I is relatively straight forward: for $j = 1, \ldots, m$, we generate $z_j \sim N(\sqrt{n}\mu_0, 1)$ and then compute $\pi_j = 2 \times (1 - \Phi(|z_j|))$.

Step II, however, is more involved. Since for $j = m+1, \ldots, p$, $z_j \sim N(0,1)$ and $\pi_j = 2 \times (1 - \Phi(|z_j|))$, we have $\pi_j \sim Uniform(0,1)$. So $(v_1, \ldots, v_k)$ are jointly distributed as the 1st to kth ordered statistics of the $p - m$ independent uniform random variable in $(0,1)$. The joint density function of $(v_1, \ldots, v_k)$ is

$$f(v_1, \ldots, v_k) = \frac{(1 - v_k)^{p-m-k}(p-m)!}{(p-m-k)!} \times I(0 \leq v_1 \leq, \ldots \leq v_k \leq 1)$$

From the join density, we can derive that v_k is marginally distributed as $Beta(k, p - m + 1 - k)$ with marginal density function

$$f(v_k) = \frac{v_k^{k-1}(1 - v_k)^{p-m-k}(p - m)!}{(k - 1)!(p - m - k)!} \times I(0 \leq v_k \leq 1)$$

Then the joint distribution of $(v_1, \ldots, v_{k-1})$ conditional on v_k is

$$f(v_1, \ldots, v_{k-1}|v_k) = \frac{(k - 1)!}{v_k^{k-1}} \times I(0 \leq v_1 \leq, \ldots \leq v_k \leq 1)$$

In other words, conditional on v_k, $(v_1, \ldots, v_{k-1})$ is the ordered statistics of $k - 1$ uniform variables in $(0, v_k)$. Such observation allows us to simulate $(v_1, \ldots, v_{k-1}, v_k)$ in an efficient manner:

(a) Simulate $v_k \sim Beta(k, p - m + 1 - k)$.

(b) Simulate $\tilde{v}_1, \ldots, \tilde{v}_{k-1}$ independently from $Uniform(0, v_k)$

(c) Sort $\tilde{v}_1, \ldots, \tilde{v}_{k-1}$ in increasing order and we have $v_1, \ldots, v_{k-1}$.

Once we have simulated both $(v_1, \ldots, v_k)$ and $(\pi_1, \ldots, \pi_m)$, we proceed with Step III by identifying the smallest k values in $(v_1, \ldots, v_k, \pi_1, \ldots, \pi_m)$, which are the same as the k smallest values from the vector $(\pi_1, \ldots, \pi_p)$.

APPENDIX K

Proof of the Inequality for the Upper and Lower Bound of the PCC When Combinging Two Types of Features

In the section, we will provide a proof for the inequality for the upper and lower bound of the PCC when we combine two types of features.

Lemma K.1. *$Q_{\pi_1}(t)$ is continuous and strictly increasing for $t \in (0, \infty)$.*

Proof. Continuity is obvious because all of the functions involved in $Q_{\pi_1}(t)$ are continuous. So we focus on the monotonicity. Let $G_{\pi_1}(t,k) = \Phi(t-k)\cdot\pi_1 + \Phi(t+k)\cdot(1-\pi_1)$. Then

$$\frac{\partial G_{\pi_1}(t,k)}{\partial k} = -\phi(t-k)\cdot\pi_1 + \phi(t+k)\cdot(1-\pi_1)$$

$\frac{\partial G_{\pi_1}(t,k)}{\partial k} > 0$ for $k < \frac{1}{2t}\log(\frac{1-\pi_1}{\pi_1})$; $\frac{\partial G_{\pi_1}(t,k)}{\partial k} = 0$ for $k = \frac{1}{2t}\log(\frac{1-\pi_1}{\pi_1})$; $\frac{\partial G_{\pi_1}(t,k)}{\partial k} < 0$ for $k > \frac{1}{2t}\log(\frac{1-\pi_1}{\pi_1})$. So $G_{\pi_1}(t,k)$ reaches maximum when $k = \frac{1}{2t}\log(\frac{1-\pi_1}{\pi_1})$. For

$0 < t_1 < t_2 < \infty$,

$$\begin{aligned} Q_{\pi_1}(t_1) &= \Phi(t_1 - \frac{1}{2}\log(\frac{1-\pi_1}{\pi_1})\frac{1}{t_1}) \cdot \pi_1 + \Phi(t_1 + \frac{1}{2}\log(\frac{1-\pi_1}{\pi_1})\frac{1}{t_1}) \cdot (1-\pi_1) \\ &< \Phi(t_2 - \frac{1}{2}\log(\frac{1-\pi_1}{\pi_1})\frac{1}{t_1}) \cdot \pi_1 + \Phi(t_2 + \frac{1}{2}\log(\frac{1-\pi_1}{\pi_1})\frac{1}{t_1}) \cdot (1-\pi_1) \\ &\leq \Phi(t_2 - \frac{1}{2}\log(\frac{1-\pi_1}{\pi_1})\frac{1}{t_2}) \cdot \pi_1 + \Phi(t_2 + \frac{1}{2}\log(\frac{1-\pi_1}{\pi_1})\frac{1}{t_2}) \cdot (1-\pi_1) \\ &= Q_{\pi_1}(t_2) \end{aligned}$$

The first inequality is due to the fact that $\Phi(\cdot)$ is strictly increasing. The second inequality is due to the fact that $G_{\pi_1}(t,k)$ reaches maximum when $k = \frac{1}{2t}\log(\frac{1-\pi_1}{\pi_1})$.

□

Theorem K.2. *Under the assumptions previously discussed, we have the following inequality*

$$\min(PCC_A, PCC_B) \leq PCC_{AB} \leq Q_{p_1}(\sqrt{2} \cdot Q_{\pi_1}^{-1}(\max(PCC_A, PCC_B)))$$

The first equality will hold when $\frac{w_k \Sigma^k w_k}{w_j \Sigma^j w_j} \to 0$ *where* $(j,k) = \begin{cases} (A,B) & \text{if } PCC_A < PCC_B \\ (B,A) & \text{if } PCC_A \geq PCC_B \end{cases}$. *The second equality is reached when* $PCC_A = PCC_B$ *and the weights are scaled such that* $w_A \Sigma^A w_A = w_B \Sigma^B w_B$.

Proof. We begin with the special case when we have$PCC_A = PCC_B$. Because $Q_{\pi_1}(\cdot)$ is strictly increasing, we have $\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}} = \frac{\boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}}$. Then because of the symmetry, we could fix the values of $\boldsymbol{\mu}^A$ and $\boldsymbol{w}_A$ and let $\boldsymbol{\mu}^B$ and $\boldsymbol{w}_B$ vary. Let $s = \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}} = \frac{\boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}}$. Then we denote $\theta = \sqrt{\boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}$ and we have $\boldsymbol{\mu}^B \cdot \boldsymbol{w}_B = s\theta$. So we have

$$\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}} = \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + s\theta}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \theta^2}}$$

The derivative with respect to θ is

$$\frac{s \times (\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A) - \theta \times (\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A)}{(\theta^2 + \boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A)^{3/2}}$$

The derivative is positive when $\theta < \frac{s \times (\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A)}{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}$ and negative when $\theta > \frac{s \times (\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A)}{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}$. In order words, $\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + s\theta}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \theta^2}}$ reaches maximum when $\theta = \theta_M = \frac{s \times (\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A)}{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}$ and minimum when $\theta \to 0$ or $\theta \to \infty$.

Maximum: Since $s = \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}}$ then $\theta_M = \frac{s \times (\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A)}{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A} = \sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}$, which means that in order to obtain the maximum, the values of $\boldsymbol{\mu}^B$ and $\boldsymbol{w}_B$ will satisfy: $\sqrt{\boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B} = \theta_M = \sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}$ and $\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A = s\theta_M = \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B$. In this case,

$$\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}} \leq \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}} = \sqrt{2} \times \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}}$$

Because $Q_{\pi_1}(\cdot)$ is an increasing function, we have $PCC_{AB} \leq Q_{\pi_1}(\sqrt{2} \times Q_{\pi_1}^{-1}(\max(PCC_A, PCC_B)))$.

Minimum: When $\theta \to 0$, $\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + s\theta}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \theta^2}} \to \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}}$. When $\theta \to \infty$, $\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + s\theta}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \theta^2}} \to s = \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}}$. Therefore, we have

$$\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}} \leq \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}}$$

Because $Q_{\pi_1}(\cdot)$ is an increasing function, we have $\min(PCC_A, PCC_B) \leq PCC_{AB}$.

Next we employ the result of the special case to prove the result for the general case where $PCC_A \neq PCC_B$. Without of loss of generality, we assume $PCC_A \geq PCC_B$, i.e. $\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}} \geq \frac{\boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}}$, because $Q_{\pi_1}(\cdot)$ is continuous strictly increasing. We choose a number $\eta > 1$ such that $\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}} = \frac{\eta \times \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}}$. Let $\boldsymbol{\mu}^{\tilde{B}} = \eta \times \boldsymbol{\mu}^B$ and

apply the previous results, we have

$$\begin{aligned}\frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}} &< \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \eta \times \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}} \\ &= \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^{\tilde{B}} \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}} \leq \sqrt{2} \times \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}}\end{aligned}$$

The first inequality holds because $\tau_2 \cdot w_2 > 0$. The last inequality holds because of the result for the case of $PCC_A = PCC_{\tilde{B}}$.

The situation for the minimum is similar. We choose a positive number $0 < \gamma < 1$ such that $\frac{\gamma \times \boldsymbol{\mu}^A \cdot \boldsymbol{w}_A}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A}} = \frac{\boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}}$ and let $\boldsymbol{\mu}^{\tilde{A}} = \gamma \times \boldsymbol{\mu}^A$. Then we have

$$\begin{aligned}\frac{\boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}} &\leq \frac{\boldsymbol{\mu}^{\tilde{A}} \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}} \\ &= \frac{\gamma \times \boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}} < \frac{\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B}{\sqrt{\boldsymbol{w}_A^t \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^t \Sigma^B \boldsymbol{w}_B}}\end{aligned}$$

The first inequality holds because of the result for the case of $PCC_{\tilde{A}} = PCC_B$. The last inequality holds because $\tau_1 \cdot w_1 > 0$.

□

BIBLIOGRAPHY

BIBLIOGRAPHY

Adam, E. K. (2006), Transactions among adolescent trait and state emotion and diurnal and momentary cortisol activity in naturalistic settings, *Psychoneuroendocrinology*, *31*(5), 664–679.

Adam, E. K., and M. Kumari (2009), Assessing salivary cortisol in large-scale, epidemiological research, *Psychoneuroendocrinology*, *34*(10), 1423–1436.

Adam, E. K., L. C. Hawkley, B. M. Kudielka, and J. T. Cacioppo (2006), Day-to-day dynamics of experience–cortisol associations in a population-based sample of older adults, *Proceedings of the National Academy of Sciences of the United States of America*, *103*(45), 17,058–17,063.

Anisimov, V., V. Fedorov, and S. Leonov (2007), Optimal design of pharmacokinetic studies described by stochastic differential equations, in *mODa 8 - Advances in Model-Oriented Design and Analysis*, pp. 9–16.

Atkinson, A. C., A. N. Donev, and R. Tobias (2007), *Optimum experimental designs, with SAS*, Oxford University Press.

Bacchetti, P., C. E. McCulloch, and M. R. Segal (2008), Simple, defensible sample sizes based on cost efficiency, *Biometrics*, *64*(2), 577–594.

Badrick, E., C. Kirschbaum, and M. Kumari (2007), The relationship between smoking status and cortisol secretion, *J Clin Endocrinol Metab*, *92*(3), 819–824.

Basagaña, X., and D. Spiegelman (2010), Power and sample size calculations for longitudinal studies comparing rates of change with a time-varying exposure, *Statistics in Medicine*, *29*(2), 181–192.

Basagaña, X., X. Liao, and D. Spiegelman (2010), Power and sample size calculations for longitudinal studies estimating a main effect of a time-varying exposure, *Statistical Methods in Medical Research*.

Bazzoli, C., S. Retout, and F. Mentre (2009), Fisher information matrix for nonlinear mixed effects multiple response models: Evaluation of the appropriateness of the first order linearization using a pharmacokinetic/pharmacodynamic model, *Statistics in Medicine*, *28*(14), 1940–1956.

Bild, D. E., et al. (2002), Multi-ethnic study of atherosclerosis: objectives and design, *American Journal of Epidemiology*, *156*(9), 871–881.

Blozis, S. A., and R. Cudeck (1999), Conditionally linear Mixed-Effects models with latent variable covariates, *Journal of Educational and Behavioral Statistics*, *24*(3), 245–270.

Brumback, B. A., and J. A. Rice (1998), Smoothing spline models for the analysis of nested and crossed samples of curves, *Journal of the American Statistical Association*, *93*(443), 961–976.

Carroll, R. J. (2003), Variances are not always nuisance parameters, *Biometrics*, *59*(2), 211–220.

Chaloner, K. (1995), Bayesian experimental design: A review, *Statistical Science*, *10*(3), 273–304.

Choi, L., B. Caffo, and C. Rohde (2007), Optimal sampling times in bioequivalence studies using a simulated annealing algorithm, *Statistics and Computing*, *17*(4), 337–347.

Chow, S., and M. Chang (2007), *Adaptive design methods in clinical trials*, CRC Press.

Clarke, R., H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang (2008), The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, *Nature reviews. Cancer*, *8*(1), 37–49.

Cochran, W. G., W. G. Cochran, and G. M. Cox (1992), *Experimental designs*, Wiley.

Cohen, S., W. J. Doyle, and A. Baum (2006), Socioeconomic status is associated with stress hormones, *Psychosomatic Medicine*, *68*(3), 414–420.

Cowles, M. K., and B. P. Carlin (1996), Markov chain monte carlo convergence diagnostics: A comparative review, *Journal of the American Statistical Association*, *91*, 883–904.

Davidian, M., R. J. Carroll, and W. SMITH (1988), Variance functions and the minimum detectable concentration in assays, *Biometrika*, *75*(3), 549 –556.

Dawson, J. D. (1998), Sample size calculations based on slopes and other summary statistics, *Biometrics*, *54*(1), 323–330.

De Souza, M., R. Toombs, J. Scheid, E. O'Donnell, S. West, and N. Williams (2010), High prevalence of subtle and severe menstrual disturbances in exercising women: confirmation using daily hormone measures, *Human Reproduction*, *25*(2), 491 –503.

de Valpine, P., H. Bitter, M. P. S. Brown, and J. Heller (2009), A simulation–approximation approach to sample size planning for high-dimensional classification studies, *Biostatistics*.

Diez Roux, A. V., et al. (2001), Neighborhood of residence and incidence of coronary heart disease, *The New England Journal of Medicine*, *345*(2), 99–106.

Dobbin, K. K., and R. M. Simon (2007), Sample size planning for developing classifiers using high-dimensional DNA microarray data, *Biostatistics*, *8*(1), 101 –117.

Donoho, D., and J. Jin (2009), Feature selection by higher criticism thresholding achieves the optimal phase diagram, *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *367*(1906), 4449–4470.

Duffull, S. B., S. Retout, and F. Mentré (2002), The use of simulated annealing for finding optimal population designs, *Computer Methods and Programs in Biomedicine*, *69*(1), 25–35.

Elaine R., M. (2008), The impact of next-generation sequencing technology on genetics, *Trends in Genetics*, *24*(3), 133–141.

Elliott, M. R. (2007), Identifying latent clusters of variability in longitudinal data, *Biostatistics*, *8*(4), 756 –771.

Fedorov, V. (2010), Optimal experimental design, *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*, 581–589.

Fedorov, V. V., and P. Hackl (1997), *Model-oriented design of experiments*, Springer.

Fedorov, V. V., R. C. Gagnon, and S. L. Leonov (2002), Design of experiments with unknown parameters in variance, *Applied Stochastic Models in Business and Industry*, *18*(3), 207–218.

Fedorov, V. V., S. L. Leonov, and V. A. Vasiliev (2010), Pharmacokinetic studies described by stochastic differential equations: Optimal design for systems with positive trajectories, in *mODa 9 - Advances in Model Oriented Design and Analysis*, edited by A. Giovagnoli, A. C. Atkinson, B. Torsney, and C. May, pp. 73–80, Physica-Verlag HD, Heidelberg.

Fisher, R. (1935), *The Design of Experiments*.

Gadegbeku, C. A., et al. (), Design of the nephrotic syndrome study network (neptune): A multi-disciplinary approach to understanding primary glomerular nephropathy.

Gallo, L. C., and K. A. Matthews (1999), Do negative emotions mediate the association between socioeconomic status and health?, *Annals of the New York Academy of Sciences*, *896*, 226–245.

Gold, E. B., B. Eskenazi, B. L. Lasley, S. J. Samuels, M. O'Neill Rasor, J. W. Overstreet, and M. B. Schenker (1995), Epidemiologic methods for prospective assessment of menstrual cycle and reproductive characteristics in female semiconductor workers, *American Journal of Industrial Medicine*, *28*(6), 783–797.

Green, P. J. (1990), On use of the EM for penalized likelihood estimation, *Journal of the Royal Statistical Society. Series B (Methodological)*, *52*(3), 443–452.

Green, P. J., and B. W. Silverman (1994), *Nonparametric regression and generalized linear models: a roughness penalty approach*, CRC Press.

Hajat, A., et al. (2010), Socioeconomic and race/ethnic differences in daily salivary cortisol profiles: the multi-ethnic study of atherosclerosis, *Psychoneuroendocrinology*, *35*(6), 932–943.

Hall, P., H. Muller, and J. Wang (2006), Properties of principal component methods for functional and longitudinal data analysis, *The Annals of Statistics*, *34*(3), 1493–1517.

Hamburg, M. A., and F. S. Collins (2010), The path to personalized medicine, *New England Journal of Medicine*, *363*(4), 301–304.

Hastings, W. K. (1970), Monte carlo sampling methods using markov chains and their applications, *Biometrika*, *57*(1), 97–109.

Hwang, D., W. A. Schmitt, G. Stephanopoulos, and G. Stephanopoulos (2002), Determination of minimum sample size and discriminatory expression patterns in microarray data, *Bioinformatics (Oxford, England)*, *18*(9), 1184–1193.

James, G. M., T. J. Hastie, and C. A. Sugar (2000), Principal component models for sparse functional data, *Biometrika*, *87*(3), 587–602.

Jin, J. (2009), Impossibility of successful classification when useful features are rare and weak, *Proceedings of the National Academy of Sciences*, *106*(22), 8859 –8864.

Johnson, R. A., and D. W. Wichern (2007), *Applied Multivariate Statistical Analysis*, Pearson Prentice Hall.

Kaplan, B., J. Schold, and H. Meier-Kriesche (2003), Poor predictive value of serum creatinine for renal allograft loss, *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, *3*(12), 1560–1565.

Kaplan, G. A., and J. E. Keil (1993), Socioeconomic factors and cardiovascular disease: a review of the literature, *Circulation*, *88*(4 Pt 1), 1973–1998.

Kraemer, H. C., J. Giese-Davis, M. Yutsis, R. O'Hara, E. Neri, D. Gallagher-Thompson, C. B. Taylor, and D. Spiegel (2006), Design decisions to optimize reliability of daytime cortisol slopes in an older population, *American Journal of Geriatric Psychiatry*, *14*(4), 325–333.

Kumari, M., E. Badrick, T. Chandola, E. K. Adam, M. Stafford, M. G. Marmot, C. Kirschbaum, and M. Kivimaki (2009), Cortisol secretion and fatigue: associations in a community based cohort, *Psychoneuroendocrinology*, *34*(10), 1476–1485.

Laird, N. M., and J. H. Ware (1982), Random-Effects models for longitudinal data, *Biometrics*, *38*(4), 963–974.

Lindstrom, M. J., and D. M. Bates (1990), Nonlinear mixed effects models for repeated measures data, *Biometrics*, *46*(3), 673–687.

Martinez, J. G., F. Liang, L. Zhou, and R. J. Carroll (2010), Longitudinal functional principal component modeling via stochastic approximation monte carlo, *The Canadian Journal of Statistics*, *38*(2), 256–270.

Meier-Kriesche, H., J. D. Schold, T. R. Srinivas, and B. Kaplan (2004), Lack of improvement in renal allograft survival despite a marked decrease in acute rejection rates over the most recent era, *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, *4*(3), 378–383.

Mentré, F., A. Mallet, and D. Baccar (1997), Optimal design in Random-Effects regression models, *Biometrika*, *84*(2), 429–442, doi:10.1093/biomet/84.2.429.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953), Equation of state calculations by fast computing machines, *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Montgomery, D. C. (2008), *Design and Analysis of Experiments*, John Wiley and Sons.

Mukherjee, S., P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, and J. P. Mesirov (2003), Estimating dataset size requirements for classifying DNA microarray data, *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *10*(2), 119–142.

Muller, P., B. Sanso, and M. De Iorio (2004), Optimal bayesian design by inhomogeneous markov chain simulation, *Journal of the American Statistical Association*, *99*(467), 788–798.

Myers, R. H., D. C. Montgomery, and C. M. Anderson-Cook (2009), *Response surface methodology: process and product optimization using designed experiments*, John Wiley and Sons.

Ogungbenro, K., G. Graham, I. Gueorguieva, and L. Aarons (2005), The use of a modified fedorov exchange algorithm to optimise sampling times for population pharmacokinetic experiments, *Computer Methods and Programs in Biomedicine*, *80*(2), 115–125.

Peng, J. (2009), A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data, *Journal of Computational and Graphical Statistics*, *18*(4), 995–1015.

Pinheiro, J. C., and D. M. Bates (1995), Approximations to the Log-Likelihood function in the nonlinear Mixed-Effects model, *Journal of Computational and Graphical Statistics*, *4*(1), 12–35.

Powell, L. H., et al. (2002), Physiologic markers of chronic stress in premenopausal, Middle-Aged women, *Psychosom Med*, *64*(3), 502–509.

Pruessner, J. C., J. Gaab, D. H. Hellhammer, D. Lintz, N. Schommer, and C. Kirschbaum (1997), Increasing correlations between personality traits and cortisol stress responses obtained by data aggregation, *Psychoneuroendocrinology*, *22*(8), 615–625.

Raudenbush, S. W., and X. Liu (2000), Statistical power and optimal design for multisite randomized trials, *Psychological Methods*, *5*(2), 199–213.

Retout, S., and F. Mentré (2003), Further developments of the fisher information matrix in nonlinear mixed effects models with evaluation in population pharmacokinetics, *Journal of Biopharmaceutical Statistics*, *13*(2), 209.

Retout, S., F. Mentre, and R. Bruno (2002), Fisher information matrix for non-linear mixed-effects models: evaluation and application for optimal design of enoxaparin population pharmacokinetics, *Statistics in Medicine*, *21*(18), 2623–2639.

Retout, S., E. Comets, A. Samson, and F. Mentre (2007), Design in nonlinear mixed effects models: Optimization using the Fedorov-Wynn algorithm and power of the wald test for binary covariates, *Statistics in Medicine*, *26*(28), 5162–5179.

Rice, J. A., and B. W. Silverman (1991), Estimating the mean and covariance structure nonparametrically when the data are curves, *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(1), 233–243.

Roy, A., D. K. Bhaumik, S. Aryal, and R. D. Gibbons (2007), Sample size determination for hierarchical longitudinal designs with differential attrition rates, *Biometrics*, *63*(3), 699–707.

Schuster, S. C. (2008), Next-generation sequencing transforms today's biology, *Nat Meth*, *5*(1), 16–18.

Shiffman, S., A. A. Stone, and M. R. Hufford (2008), Ecological momentary assessment, *Annual Review of Clinical Psychology*, *4*, 1–32.

Silverman, B. W. (1996), Smoothed functional principal components analysis by choice of norm, *The Annals of Statistics*, *24*(1), 1–24.

Simon, R. (2008), Development and validation of biomarker classifiers for treatment selection, *Journal of statistical planning and inference*, *138*(2), 308–320.

Smyth, J. M., M. C. Ockenfels, A. A. Gorin, D. Catley, L. S. Porter, C. Kirschbaum, D. H. Hellhammer, and A. A. Stone (1997), Individual differences in the diurnal cycle of cortisol, *Psychoneuroendocrinology*, *22*(2), 89–105.

Stroud, J. R., P. Müller, and G. L. Rosner (2001), Optimal sampling times in population pharmacokinetic studies, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *50*(3), 345–359.

Stroud, L. R., G. D. Papandonatos, D. E. Williamson, and R. E. Dahl (2004), Applying a nonlinear regression model to characterize cortisol responses to corticotropin-releasing hormone challenge, *Annals of the New York Academy of Sciences*, *1032*, 264–266.

Tack, L., and M. Vandebroek (2004), Budget constrained run orders in optimum design, *Journal of Statistical Planning and Inference*, *124*(1), 231–249.

Vrijens, B., and E. Goetghebeur (1999), The impact of compliance in pharmacokinetic studies, *Statistical Methods in Medical Research*, *8*(3), 247–262.

Waller, K., S. H. Swan, G. C. Windham, L. Fenster, E. P. Elkin, and B. L. Lasley (1998), Use of urine biomarkers to evaluate menstrual function in healthy premenopausal women, *American Journal of Epidemiology*, *147*(11), 1071 –1080.

Wang, Y., D. J. Miller, and R. Clarke (2008), Approaches to working in high-dimensional data spaces: gene expression microarrays, *British Journal of Cancer*, *98*(6), 1023–1028.

Williams, D. R., Y. Yu, J. S. Jackson, and N. B. Anderson (1997), Racial differences in physical and mental health.

Windham, G. C., E. Elkin, L. Fenster, K. Waller, M. Anderson, P. R. Mitchell, B. Lasley, and S. H. Swan (2002), Ovarian hormones in premenopausal women: Variation by demographic, reproductive and menstrual cycle characteristics, *Epidemiology*, *13*(6), 675–684.

Yao, F., and T. C. M. Lee (2006), Penalized spline models for functional principal component analysis, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 3–25.

Yao, F., H. Muller, and J. Wang (2005), Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association*, *100*(470), 577–590.

Zhou, L., J. Z. Huang, and R. J. Carroll (2008), Joint modelling of paired sparse functional data using principal components, *Biometrika*, *95*(3), 601 –619.