STATISTICAL METHODS FOR ANALYZING RANDOMIZED TRIALS AND BRAIN IMAGING DATA

by

Bingkai Wang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

February 2021

Abstract

My thesis work focuses on developing reliable and innovative statistical methods for improving analyses of biomedical data. I address two types of questions: improving precision of randomized clinical trials and identifying brain networks using brain imaging. For randomized clinical trials, we proved the statistical validity of two commonly used methods to improve precision while being robust to misspecification of models used in the analysis. We demonstrated our results by re-analyzing completed randomized trials and showed that substantial precision gain can be achieved by these two methods. For brain imaging, we proposed a consistent estimator for the brain networks that are common across people. Applied to a motor-task functional magnetic resonance imaging data set, our estimator identifies meaningful brain networks that are consistent with current scientific understandings of motor networks.

Primary Readers: Ramin Mojtabai, Bryan Lau, Xi Luo, Brian Caffo (Co-advisor) and Michael Rosenblum (Advisor).

ABSTRACT

Secondary Readers: Alyssa Moran and Elizabeth Ogburn.

Acknowledgments

I would like to express the deepest appreciation to my primary advisor, Michael Rosenblum, who is an excellent mentor and supportive friend. I would like to sincerely thank my co-advisor, Brian Caffo. I am extremely fortunate to have both of you as my advisors and I have learned precious knowledge and skills from you.

I am deeply appreciative for the time of my thesis committee and their advice for my research.

I would like to thank my collaborators Marcia Irene Canto, Elizabeth Ogburn, Xi (Rossi) Luo, Yi Zhao, Masoumeh Amin-Esmaeili, Ryoko Susukida and Paniz Charkhchi. It is my great pleasure and honor to work with you.

I would like to give special thanks to Marie Diener-West, Yi Zhao, Ning Yang and Dong Xi. I was a teaching assistant of courses instructed by Marie for three years and she helped me be a good team member and taught me how to teach. Yi is not only a fantastic collaborator, but also my great friend. When I visited Cornell University, Ning kindly advised me and generously gave me research suggestions and guidance. Dong was my supervisor when I did an internship at Novartis. He is the one that showed me the world of industrial pharmaceuticals and provided me different and inspiring perspectives of my research.

I would like to thank my peers and friends in Johns Hopkins University and Cornell University.

I would like to thank the Department of Biostatistics and the Johns Hopkins Bloomberg School of Public Health for providing the resources for study, research and well-being.

Finally, I would like to thank my fiancee, Ke Wu, and my parents, Jing Li and Longjun Wang, for their full support during the entire course of my study. This thesis is dedicated for you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

My thesis research lies in two areas: (1) causal inference for clinical trials and (2) statistical modeling for brain imaging. In causal inference for clinical trials, the open question I address is how to perform model-robust inference, i.e. valid inference without requiring a correct model, and how to improve precision when estimating a treatment effect. In the analysis of brain imaging data, I study the identification of brain networks that are common among people. These two areas of my research differ substantially. However, a common component of my research is that the proposed statistical methods improve the existing methods both empirically (via simulation and real data analysis) and theoretically (by asymptotic results, i.e., consistency and asymptotic variance). In addition, both areas have connections to what is classically called "adjustment". In the randomized trial setting, we seek robust and theoretically valid forms of adjustment. In brain imaging we exploit parsimony.

Chapter 2 focuses on ANCOVA in randomized clinical trials. ANCOVA involves fitting a linear regression model for the primary outcome on intercept, treatment allocation and baseline variables. The estimated regression coefficient of the treatment allocation term is the ANCOVA estimator of the average treatment effect (Yang and Tsiatis, 2001). According to two surveys (Pocock et al., 2002; Austin et al., 2010), there is confusion regarding the statistical validity of ANCOVA when the linear model is misspecified. This is an important question, because model misspecification is almost unavoidable in practice. We clarify this issue by proving that the ANCOVA estimator is robust to arbitrary model misspecification and explain how ANCOVA can reduce variance. Our results apply to analyses of phase 2 or 3 trials and can help reduce the required sample size to achieve a desired power. Applied to randomized trials for mild cognitive impairment, schizophrenia, and depression, we demonstrate that ANCOVA brings a 4% to 32% precision gain.

Chapter 3 concerns stratified randomization and covariate adjustment, which are two commonly used methods for improving precision and power in clinical trials. Stratified randomization refers to a randomization procedure that, by design, attempts to create balance across study arms in strata of baseline variables. Covariate adjustment means adjusting for baseline variables in estimation at the end of the trial (or at interim analyses). According to a survey by

Kahan and Morris (2012), many trials do not fully capture the combined precision gain from these two methods, which may lead to increased sample size or prolonged trial duration. We derive consistency and asymptotic normality for a large class of estimators that involves stratified randomization and covariate adjustment. We show that these two methods can lead to substantial gains in precision and power by re-analyzing three trials of substance use disorder treatments, where the variance reduction due to stratified randomization and covariate adjustment ranges from 1% to 36%. Our results are most useful in improving precision of phase 2 or 3 trials and can handle a variety of outcome types, repeated measures outcomes and missing outcome data.

Chapter 4 considers a problem of jointly modeling multiple covariance matrices, assuming a proportion of eigenvectors to be shared across while the rest are individual-specific. This problem is motivated by functional magnetic resonance imaging (fMRI) data, where correlations among brain regions form covariance matrices, and our goal is to identify common brain networks (i.e. shared eigenvectors), which represent correlations in functional brain measures consistent across subjects. We solve this problem by proposing consistent estimators of the shared eigenvectors and the number of the shared eigenvectors. In a data set of motor-task fMRI, our estimator identifies meaningful brain networks that are consistent with current scientific understandings of motor networks during a motor paradigm. In general, our proposed estimator can help better understand brain functions and is also applicable to similar questions in other areas, such as genomics and economics.

Chapter 5 discusses future directions of Chapters 2-4.

Chapter 2

Analysis of Covariance (ANCOVA) in Randomized Trials: More Precision and Valid Confidence Intervals, Without Model Assumptions

The content of this chapter is reproduced from Wang et al. (2019) available at `https://doi.org/10.1111/biom.13062`.

2.1 Introduction

Pocock et al. (2002), in a survey of 50 randomized trial reports, found that 36 used covariate adjustment, but only 12 emphasized adjusted over unadjusted estimators. They stated that "the statistical properties of covariate-adjustment are quite complex and often poorly understood, and there remains confusion as to what is an appropriate statistical strategy." Austin et al. (2010), in a paper titled "A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals", surveyed 114 randomized trial articles and found that only 39 of them presented an adjusted analysis.

We focus on the analysis of covariance (ANCOVA) estimator, referred to as "ANCOVA I" by Yang and Tsiatis (2001). It involves fitting a linear regression model for the primary outcome with intercept and main terms for the treatment assignment and baseline variables. For trials with continuous-valued or change score outcomes, covariate adjustment often involves the ANCOVA estimator. We use the term "covariate adjustment" to refer to the ANCOVA estimator.

Concerns have been raised about the validity of ANCOVA for analyzing randomized trial data when the linear model is misspecified. For example, Kraemer (2015) states "The linear model used for covariate adjusting (e.g., analysis of covariance) assumes ... that there is no interaction between the covariates and the treatment effect." and "Given these risks for bias, ANCOVA should not generally be used for such adjustment." Ludvigsson et al. (2008) and Montalban et al. (2017) both checked whether data is normally distributed before applying ANCOVA and Ludvigsson et al. (2008) state that "ANCOVA involves the assumption of normally distributed response data and homogeneity of variances." These authors are wise to have the general concern about model misspecification, since in many contexts it can lead to biased or uninterpretable analyses. However, when the ANCOVA estimator is used to analyze randomized trial data, it has special robustness properties that obviate the above concerns.

Yang and Tsiatis (2001) proved that the ANCOVA estimator is consistent under arbitrary misspecification of the linear model. We build on this result by proving that the standard error, computed as if the linear model were correct, is also consistent. Therefore, not only estimates but also confidence intervals and hypothesis tests conducted as if the linear model were correct are asymptotically valid even when the linear model is arbitrarily misspecified, e.g., when the true relationships between variables are non-linear, and/or when there is treatment effect heterogeneity. This is important since it is not possible to rule out all types of model misspecification.

We attempt to provide intuition behind the precision gains from covariate adjustment by showing a direct analogy to ordinary least squares linear regression. We prove that the asymptotic variance reduction (i.e., precision gain) due to covariate adjustment equals the fraction of variance in the primary outcome explained by the baseline variables, beyond what is already explained by the main effect of treatment. This holds under arbitrary model misspecification and leads to a simple formula for estimating the variance reduction due to ANCOVA. This variance reduction is important since it equals the reduction in the required sample size to achieve a desired power.

The above results build on key ideas from Tsiatis et al. (2008); Rubin and van der Laan (2008); Moore and van der Laan (2009a); Moore et al. (2011); Rubin and van der Laan (2011); Jiang et al. (2018); Tian et al. (2019). As in their work, our results are asymptotic, i.e., they hold in the limit as sample size grows to infinity while the set of covariates is fixed. A special case of the results of Bugni et al. (2018) coincides with a special case of our result in Section 4.2 (explained in that section).

We present data analyses based on three completed randomized clinical trials for treatment of mild cognitive impairment (MCI) (Petersen et al., 2005), schizophrenia (Jarskog et al., 2013), and depression (Treatment for Adolescents With Depression Study (TADS) Team, 2004), respectively. By analyzing these data sets, we demonstrate how covariate adjustment can reduce variance, have greater added value in large trials in terms of reducing the required sample size to achieve a desired power, and increase power even when by chance there is perfect or near-perfect balance across arms in the baseline variables (due to the ANCOVA estimator's smaller standard error).

In the next section, we describe the three trials. In Section 2.3, we define the unadjusted estimator, ANCOVA estimator, and covariate imbalance. In Section 2.4, we present our main results. Illustrations are provided in Section 2.5, where trial analyses are presented. Some practical recommendations for applying covariate adjustment are given in Section 2.6.

2.2 Three Completed Randomized Clinical Trials

2.2.1 Mild Cognitive Impairment (MCI) Trial

The "Vitamin E and Donepezil for the Treatment of Mild Cognitive Impairment" (MCI) phase 3 randomized trial was completed in 2004 (Petersen et al., 2005). The goal was to estimate the effect of a drug treatment on preventing progression from MCI to Alzheimer's disease. Participants were randomized to three arms: the drug Donepezil, Vitamin E, and placebo control. We compare the Donepezil arm (253 participants, 33% missing outcomes) to the placebo arm (259 participants, 28% missing outcomes). The primary outcome was time to progression to Alzheimer's disease. In order to apply the ANCOVA estimator, which requires a continuous or change score outcome, we instead use the change in Clinical Dementia Rating-sums of boxes score (CDR-SB) between baseline and 18 months. We use the following baseline variables for adjustment: age, gender, Alzheimer's Disease Assessment Scale (ADAS)-cognitive score, Mini–Mental State Examination (MMSE) score, Activities of Daily Living total score, Global Deterioration scale, and CDR-SB.

2.2.2 Metformin for Weight Loss (METS) Trial

The "Metformin for weight loss and metabolic control in overweight outpatients with schizophrenia and schizoaffective disorder" trial, referred to as "METS", is a phase 4 randomized trial completed in 2010 (Jarskog et al., 2013). Participants were randomly assigned to two arms: Metformin (treatment, 75 participants, 15% missing outcomes) and placebo (control, 71 participants, 14% missing outcomes). The primary outcome was weight loss over 16 weeks. We use this outcome and the following baseline variables: age, gender, Clinical Global Impressions (CGI) severity rating score, tobacco use, illicit drug use, alcohol use, weight and body mass index (BMI).

2.2.3 Treatment for Adolescents with Depression Study (TADS)

The "Treatment for Adolescents with Depression Study" (TADS) is a phase 3, four-arm, randomized trial completed in 2003 (Treatment for Adolescents With Depression Study (TADS) Team, 2004). The goal was to evaluate cognitive-behavioral therapy (CBT) and Fluoxetine (FLX), each alone and combined (CMB), for treating major depressive disorder in adolescents (age 12–17). Participants were randomized to four arms: FLX only (109 participants, 15% missing outcomes), CBT only (111 participants, 29% missing outcomes), combined (CMB, 107 participants, 16% missing outcomes), and placebo (112 participants, 20% missing outcomes). The co-primary outcomes were the change in Children's Depression Rating Scale-Revised (CDRS-R) score and improvement of Clinical Global Impressions (CGI) severity rating score at 12 weeks. We focus on the former outcome and adjust for the following baseline variables: age, gender, CDRS-R score, CGI severity rating score, Children's Global Assessment Scale score (CGAS), Reynolds Adolescent Depression Scale total score (RADS), suicide ideation score, current major depressive episode duration, and comorbidity (indicator of any other psychiatric disorder except dysthymia).

2.3 Definitions

2.3.1 Estimators of Average Treatment Effect

We focus on randomized clinical trials where each participant contributes the generic data vector $(\boldsymbol{W}, A, Y)$, where $\boldsymbol{W}$ is a $k \times 1$ column vector of predefined baseline variables, A is the study arm assignment, and Y is the outcome. We assume that Y is continuous or a change score (difference between a score measured at follow-up and baseline). We assume the study arm assignment indicator A is binary ($A = 1$ for treatment and $A = 0$ for control). For trials with more than 1 treatment (e.g., TADS), we consider each treatment arm vs. control comparison separately.

The components of the baseline vector $\boldsymbol{W}$ can be continuous, binary, ordinal and/or categorical. All variables are assumed to be bounded. We assume that the components of $(1, \boldsymbol{W}^t)$ are linearly independent, i.e., no component is a linear combination of the others; otherwise at least one component is redundant and can be dropped from the corresponding design matrix.

For each participant $i = 1, \ldots, n$, we observe the data vector $(\boldsymbol{W}_i, A_i, Y_i)$, which we assume to be an independent, identically distributed draw from the unknown, joint distribution on the generic data vector $(\boldsymbol{W}, A, Y)$. The only assumption that we make about the joint distribution is that the study arm A is randomly assigned with equal probability to treatment or control independent of the baseline variables $\boldsymbol{W}$. This holds by design in trials using simple randomization with equal probability of assignment to each study arm, which is the type of trial design that we consider throughout. The assumptions in this paragraph do not hold in trials that use stratified block randomization or covariate-adaptive randomization, which are discussed in Section 2.6.

The goal is to estimate the population average treatment effect $\Delta = E[Y|A = 1] - E[Y|A = 0]$, i.e., the difference between population means if everyone in the study population had been assigned to treatment versus control. We focus throughout on estimating the average treatment effect Δ, since that is the principal quantity of interest in the primary efficacy analysis of randomized trials (Tsiatis et al., 2008). An estimator of the average treatment effect Δ is called robust to arbitrary model misspecification if it is consistent under the aforementioned assumptions. These assumptions do not put any restrictions on the joint distribution of $\boldsymbol{W}, A, Y$ other than A being randomly assigned with equal probability to each arm independent of the baseline variables $\boldsymbol{W}$. E.g., the baseline variables can be correlated with each other and the treatment can be more/less effective for different subpopulations defined by the baseline variables.

Denote the unadjusted estimator (which ignores baseline variables) of the average treatment effect Δ by

$$\widehat{\Delta}^{unadj} = \frac{\sum_{i=1}^{n} Y_i A_i}{\sum_{i=1}^{n} A_i} - \frac{\sum_{i=1}^{n} Y_i(1-A_i)}{\sum_{i=1}^{n}(1-A_i)}.$$

The unadjusted estimator is consistent, i.e., converges to Δ as the sample size goes to infinity.

The ANCOVA estimator of the average treatment effect Δ adjusts for chance imbalance between study arms in $\boldsymbol{W}$. It is computed by fitting the following linear regression model

$$E[Y|A,\boldsymbol{W}] = \beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{W}}^t \boldsymbol{W}, \tag{2.1}$$

using ordinary least squares (OLS). Denote the estimated coefficients by $\widehat{\beta}_0, \widehat{\beta}_A, \widehat{\boldsymbol{\beta}}_{\boldsymbol{W}}$. The ANCOVA estimator $\widehat{\Delta}^{ancova}$ of the average treatment effect Δ is the estimated coefficient $\widehat{\beta}_A$.

According to Huitema (2011), the ANCOVA model assumes: (i) a linear relationship between the outcome and the other variables, i.e., $Y = \beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{W}}^t \boldsymbol{W} + \varepsilon$, where ε is the error term, and (ii) the distribution of the error ε is normal with mean 0 conditional on A and $\boldsymbol{W}$. These assumptions may fail to hold if there is an interaction between treatment and covariate (Kraemer, 2015), if there are unmeasured prognostic covariates that are correlated with $\boldsymbol{W}$ (Austin et al., 2010), or if the outcome is non-linearly related to the covariates. Fortuitously, the key statistical properties of ANCOVA (consistency of the point estimate and standard error) hold under any of these types of model misspecification.

Yang and Tsiatis (2001) proved that the ANCOVA estimator is consistent for Δ, i.e., $\widehat{\beta}_A$ converges to Δ in probability, even under arbitrary misspecification of the linear model (2.1). Furthermore, the ANCOVA estimator is asymptotically normal and we denote its asymptotic variance as $Var^*(\widehat{\Delta}^{ancova})$, i.e., $n^{1/2}(\widehat{\Delta}^{ancova} - \Delta)$ converges to a normal distribution with mean 0 and variance $Var^*(\widehat{\Delta}^{ancova})$. Yang and Tsiatis (2001) also proved that when the probability of being randomized to each study arm is equal (as assumed here), the ANCOVA estimator has asymptotic variance at most that of the unadjusted estimator; if any baseline variable is correlated with the outcome, then ANCOVA is strictly more precise.

We use the ANCOVA estimator $\widehat{\Delta}^{ancova} = \widehat{\beta}_A$ to estimate the average (also called marginal) treatment effect $\Delta = E[Y|A=1] - E[Y|A=0]$, which is not assumed to be constant across strata of $\boldsymbol{W}$. We emphasize this to avoid confusion, since the conventional interpretation of the estimated coefficient $\widehat{\beta}_A$ is the conditional treatment effect. That interpretation does not apply when the model is misspecified. For example, when the treatment effect differs within strata of $\boldsymbol{W}$, then the conditional treatment effect is not a single number but instead is a function mapping each stratum of $\boldsymbol{W}$ to the corresponding effect. Though it is of independent interest to estimate the conditional treatment effect, this is often much more challenging and requires more assumptions than estimating the marginal treatment effect Δ (since it involves estimating a function rather than a single number). The reason for considering baseline variables at all when estimating the marginal treatment effect Δ is that this can improve precision and power by accounting for chance imbalances across study arms (Yang and Tsiatis, 2001).

The imbalance $\boldsymbol{I}$ between study arms in the baseline variables $\boldsymbol{W}$, called chance imbalance or covariate imbalance, is the difference between sample means of $\boldsymbol{W}$ comparing treatment versus control arms: $\boldsymbol{I} = \sum_{i=1}^{n} A_i \boldsymbol{W}_i \,/ \sum_{i=1}^{n} A_i - \sum_{i=1}^{n}(1-A_i)\boldsymbol{W}_i \,/ \sum_{i=1}^{n}(1-A_i)$. Although A is independent of $\boldsymbol{W}$ by design, in any realization the baseline variables can be imbalanced.

2.3.2 ANCOVA Variance Decomposition and Definition of $R^2_{Y-\Delta A\sim \boldsymbol{W}}$

We review properties of OLS regression and define a quantity ($R^2_{Y-\Delta A\sim \boldsymbol{W}}$) that plays a key role in our main results in Section 2.4. All results below hold under arbitrary model misspecification.

Consider regressing a generic response variable Z on a covariate vector $\boldsymbol{X}$ using the linear model $E[Z|\boldsymbol{X}] = \beta_0 + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}$. We assume that all variables are bounded and the components of $(1, \boldsymbol{X}^t)$ are linearly independent. If the model is misspecified, i.e., if for every possible $\beta_0, \boldsymbol{\beta}_{\boldsymbol{X}}$ we have $E[Z|\boldsymbol{X}] \neq \beta_0 + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}$, then the OLS estimator $\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}$ (based on independent, identically distributed vectors $(\boldsymbol{X}_i, Z_i) : i = 1, \ldots, n$) still converges to a limit, denoted $\underline{\beta}_0, \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}$. The variance of Z decomposes as $Var(Z) = Var(\underline{\beta}_0 + \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}) + Var(Z - \underline{\beta}_0 - \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X})$, where $\underline{\beta}_0 + \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}$ is the predicted response and $Z - \underline{\beta}_0 - \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}$ is the residual. In other words, the response variance is the sum of the prediction variance and residual variance. The fraction of the variance of Z explained by covariates $\boldsymbol{X}$, denoted $R^2_{Z \sim \boldsymbol{X}}$, is defined as $1 - Var(Z - \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X})/Var(Z)$ (where we omit the intercept $\underline{\beta}_0$ here and below since it does not impact the variance).

We apply the above variance decomposition to the linear regression model (2.1) that is used in computing the ANCOVA estimator. Let $(\underline{\beta}_A, \underline{\boldsymbol{\beta}}_{\boldsymbol{W}})$ denote the limit in probability of the OLS estimator $(\widehat{\beta}_A, \widehat{\boldsymbol{\beta}}_{\boldsymbol{W}})$ for the linear model (2.1) as sample size n goes to infinity. Our interest is in the variance in the outcome Y explained by baseline variables $\boldsymbol{W}$, beyond what is already explained by treatment A. Therefore, we set the response to be $Z = Y - \underline{\beta}_A A$ and regressor to be $\boldsymbol{X} = \boldsymbol{W}$. The following variance decomposition, analogous to the decomposition of $Var(Z)$ above, is proved in the Supporting Information:

$$Var(Y - \underline{\beta}_A A) = Var(\underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W}) + Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W}). \tag{2.2}$$

The corresponding fraction of the variance in the outcome Y explained by the baseline variables $\boldsymbol{W}$, beyond what is already explained by (the main effect of) treatment A, is denoted by

$$R^2_{Y-\Delta A\sim\boldsymbol{W}} = 1 - Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}^t_{\boldsymbol{W}}\boldsymbol{W})/Var(Y - \underline{\beta}_A A). \tag{2.3}$$

The subscript in $R^2_{Y-\Delta A\sim\boldsymbol{W}}$ is to indicate that this R-squared represents the fraction of variance of $Y-\Delta A$ explained by $\boldsymbol{W}$, where we made the substitution $\underline{\beta}_A = \Delta$ on the right side of (2.3), which holds by the consistency result of Yang and Tsiatis (2001).

The importance of $R^2_{Y-\Delta A\sim\boldsymbol{W}}$ is that, as we show below, it is identical to the asymptotic variance reduction (equivalently, the sample size reduction) comparing the ANCOVA estimator to the unadjusted estimator, and that this holds under arbitrary misspecification of (2.1). This result builds on fundamental ideas from Rubin and van der Laan (2008); Moore and van der Laan (2009a) as described below.

2.4 $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ and the Relationship Among Unadjusted Estimator, ANCOVA Estimator, and Covariate Imbalance, Under Model Misspecification

All results below hold under arbitrary model misspecification. Our first result, in Section 2.4.1, is an equivalence between the ordinary least squares variance decomposition (2.2) and a variance decomposition relating the unadjusted estimator, ANCOVA estimator, and covariate imbalance. Second, in Section 2.4.2, we show that the variance estimator for ANCOVA computed by standard statistical software is consistent. Our third result, in Section 2.4.3, is a simple formula for the variance reduction (equivalently, the sample size reduction) due to covariate adjustment. These results build on ideas from prior work as described below.

2.4.1 Connecting OLS Regression to the Relationship Among Unadjusted Estimator, ANCOVA Estimator, and Covariate Imbalance

Jiang et al. (2018) proved the following relationship among the unadjusted estimator $\widehat{\Delta}^{unadj}$, ANCOVA estimator $\widehat{\Delta}^{ancova}$, and chance imbalance $\boldsymbol{I}$:

$$\widehat{\Delta}^{unadj} \approx \boldsymbol{\beta}^{t}_{\boldsymbol{W}}\boldsymbol{I} + \widehat{\Delta}^{ancova}. \tag{2.4}$$

(Formally, the difference between the left and right sides of the above display, after multiplying by $n^{1/2}$, converges to 0 in probability.) They also showed the following variance decomposition:

$$Var^{*}(\widehat{\Delta}^{unadj}) = Var^{*}(\boldsymbol{\beta}^{t}_{\boldsymbol{W}}\boldsymbol{I}) + Var^{*}(\widehat{\Delta}^{ancova}), \tag{2.5}$$

where Var^{*} denotes asymptotic (i.e., large sample) variance.

We show that the above variance decomposition among the unadjusted estimator, chance imbalance, and ANCOVA estimator is identical to the variance decomposition (2.2) for OLS, under arbitrary model misspecification. Specifically, we prove in the Supporting Information that each term in (2.5) equals 4 times the corresponding term in (2.2), i.e., $Var^{*}(\widehat{\Delta}^{unadj}) = 4Var(Y - \beta_{A}A)$,

$Var^*(\underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{I}) = 4Var(\underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W})$, and $Var^*(\widehat{\Delta}^{ancova}) = 4Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W})$. This is summarized in Figure 2.1, where the first row is the variance decomposition in OLS, the second row is the variance decomposition of $\widehat{\Delta}^{unadj}$ from Jiang et al. (2018), and our contribution is to connect them by proving equality of quantities in the same column. When model (2.3.1) is misspecified, all equalities in Figure 2.1 still hold. These relationships are used to prove the results in Sections 2.4.2 and 2.4.3.

2.4.2 Robustness of the ANCOVA Variance Estimator to Arbitrary Model Misspecification

Consider the ANCOVA model-based variance estimator for $\widehat{\Delta}^{ancova}$ that is output by standard statistical software such as 'summary.lm' in R or 'proc reg' in SAS, which we denote by $\widehat{Var}(\widehat{\Delta}^{ancova})$. The formula for $\widehat{Var}(\widehat{\Delta}^{ancova})$ is

$$\widehat{Var}(\widehat{\Delta}^{ancova}) = \frac{\widehat{Var}(Y - \widehat{\beta}_0 - \widehat{\beta}_A A - \widehat{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W})}{(n-1)[\widehat{Var}(A) - \widehat{Cov}(\boldsymbol{W}, A)^t \widehat{Var}(\boldsymbol{W})^{-1} \widehat{Cov}(\boldsymbol{W}, A)]} \tag{2.6}$$

where on the right side $\widehat{Var}, \widehat{Cov}$ are the sample variance and sample covariance, respectively, where degrees of freedom are taken into account. (See the Supporting Information for precise definitions of these.) The following theorem shows that the above variance estimator is robust to arbitrary model misspecification. (See the Supporting Information for proof.)

Theorem 1. *Given the assumptions in Section 2.3.1, which do not assume that the linear model (2.1) is correctly specified, n times the estimated variance $\widehat{Var}(\widehat{\Delta}^{ancova})$ converges in probability to the true asymptotic variance $Var^*(\widehat{\Delta}^{ancova})$ of the ANCOVA estimator $\widehat{\Delta}^{ancova}$.*

The above theorem implies that confidence intervals and Wald-type hypothesis tests conducted as if the linear model were correct are asymptotically valid even when the linear model is arbitrarily misspecified. The $1-\alpha$ confidence interval for the coefficient on the A term in (2.1) that is output by the aforementioned, standard linear regression software is

$$\left(\widehat{\Delta}^{ancova} - t_{n-p,\alpha/2}\sqrt{\widehat{Var}(\widehat{\Delta}^{ancova})}, \widehat{\Delta}^{ancova} + t_{n-p,\alpha/2}\sqrt{\widehat{Var}(\widehat{\Delta}^{ancova})}\right), \qquad (2.7)$$

where $t_{n-p,\alpha/2}$ is the $\alpha/2$-quantile of the t-distribution with $n-p$ degrees of freedom where p is the number of coefficients in the linear model (2.1). For large n and fixed p, the quantile $t_{n-p,\alpha/2}$ is approximately the $\alpha/2$-quantile of the standard normal distribution. It follows from the above theorem that the above display is an asymptotically valid confidence interval for the average treatment effect Δ, under arbitrary model misspecification.

Bugni et al. (2018) focused on trials using covariate-adaptive randomization, but their results also have implications for simple randomization as considered here. In particular, the special case of the above theorem where $\boldsymbol{W}$ is a single, categorical variable follows from Theorem 4.3 and Remark 4.6 of Bugni et al. (2018).

2.4.3 $R^2_{Y-\Delta A\sim\boldsymbol{W}}$ Equals Precision Gain (and Sample Size Reduction) Due to Adjustment, Even Under Arbitrary Model Misspecification

Borm et al. (2007) and Rubin and van der Laan (2008) connect the R-squared from regressing Y on $\boldsymbol{W}$ to the variance reduction due to ANCOVA, while Moore and van der Laan (2009a) and Moore et al. (2011) make a similar connection in the context of binary outcomes and estimators based on logistic regression models. Each of the aforementioned approaches requires conditions (such as the linear model being correctly specified or that $\Delta = 0$) or requires additional factors to connect the R-squared to the variance reduction due to covariate adjustment. (See the Supporting Information for more details.) Building on key ideas from their approaches, we prove that the R-squared $R^2_{Y-\Delta A\sim\boldsymbol{W}}$ equals the variance reduction due to ANCOVA without requiring these conditions or extra factors; this R-squared (which differs from the prior work above by incorporating A) is robust to arbitrary model misspecification.

It follows from the relationships in Figure 2.1 that the fraction $R^2_{Y-\Delta A\sim\boldsymbol{W}}$ of the variance in the outcome Y explained by the baseline variables $\boldsymbol{W}$, beyond what is explained by the treatment A, equals the asymptotic variance reduction due to ANCOVA, i.e.,

$$R^2_{Y-\Delta A \sim \boldsymbol{W}} = 1 - \frac{Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}{}^t \boldsymbol{W})}{Var(Y - \underline{\beta}_A A)} = 1 - \frac{Var^*(\widehat{\Delta}^{ancova})}{Var^*(\widehat{\Delta}^{unadj})}. \tag{2.8}$$

The first equality is the definition of $R^2_{Y-\Delta A \sim \boldsymbol{W}}$, and the second shows that $R^2_{Y-\Delta A \sim \boldsymbol{W}}$ equals the variance reduction due to ANCOVA (expression on the right). The rightmost expression, by definition, equals one minus the asymptotic relative efficiency (also called Pitman efficiency) comparing the unadjusted to the ANCOVA estimator.

In practice, $R^2_{Y-\Delta A \sim \boldsymbol{W}}$ can be estimated by $\widehat{R}^2_{Y-\Delta A \sim \boldsymbol{W}} = 1 - \widehat{Var}(\widehat{\Delta}^{ancova})/\widehat{Var}(\widehat{\Delta}^{unadj})$, where $\widehat{Var}(\widehat{\Delta}^{ancova})$ is the variance of the ANCOVA estimator output by standard statistical software as in (2.6), and $\widehat{Var}(\widehat{\Delta}^{unadj})$ is the variance of the unadjusted estimator estimated analogously (by regressing Y on A and an intercept).

The variance reduction (2.8) due to ANCOVA is important since it equals the fractional sample size reduction that can be achieved through covariate adjustment when holding the desired power fixed, asymptotically. A variance reduction of $p\%$ means that the sample size required to achieve a desired power is also reduced by $p\%$. Therefore, $\widehat{R}^2_{Y-\Delta A \sim \boldsymbol{W}}$ can be used to estimate the benefits of covariate adjustment in terms of sample size reduction. The $\widehat{R}^2_{Y-\Delta A\sim\boldsymbol{W}}$ values from our data sets range from 4% to 32%, which can be translated into 4% to 32% sample size reductions.

$$\begin{array}{ccccc}
Var(Y-\underline{\beta}_A A) & = & \overbrace{Var(\underline{\beta}_{\boldsymbol{W}}^t \boldsymbol{W})}^{\text{Variance in } Y \text{ explained by } \boldsymbol{W}} & + & \overbrace{Var(Y-\underline{\beta}_A A-\underline{\beta}_{\boldsymbol{W}}^t \boldsymbol{W})}^{\text{Residual variance after adjusting for } \boldsymbol{W}} \\
\| & & \| & & \| \\
\frac{1}{4}Var^*(\hat{\Delta}^{unadj}) & = & \underbrace{\frac{1}{4}Var^*(\underline{\beta}_{\boldsymbol{W}}^t \boldsymbol{I})}_{\text{Variance in the unadjusted estimator explained by imbalance in } \boldsymbol{W}} & + & \underbrace{\frac{1}{4}Var^*(\hat{\Delta}^{ancova})}_{\text{Residual variance after adjusting for chance imbalance in } \boldsymbol{W}}
\end{array}$$

Figure 2.1: Variance decomposition equivalence between linear regression and estimators of average treatment effect. The variance decomposition in the first row is a result of OLS linear regression. The second row gives the asymptotic variance decomposition of the unadjusted estimator, which is a minor extension of key results from Jiang et al., 2018; Tian et al., 2019. Our contribution is to connect the two variance decompositions by showing their equivalence, i.e., quantities in the same column are equal, under arbitrary model misspecification.

2.5 Clinical Trial Applications

Our data analyses for each application (MCI, METS, TADS) are summarized in Table 2.1 and described below. All baseline variables were standardized and missing baseline values were imputed by the median for continuous variables and the mode for binary and categorical variables. All participants with missing outcomes were removed from the analysis, for simplicity; in practice, missing outcome data would be handled as described in Section 2.6. Point estimates and standard errors are rounded to the nearest 0.01. "Confidence Interval" is abbreviated as "CI".

Table 2.1: Summary of clinical trial data analyses: unadjusted estimator for average treatment effect, adjusted estimator (ANCOVA) for average treatment effect, 95% confidence intervals (CI), and estimated variance reduction due to adjustment. Negative (positive) estimates are in the direction of clinical benefit (harm).

Trial Name	Unadjusted Estimator (95% CI)	ANCOVA Estimator (95% CI)	Variance Reduction ($\widehat{R}^2_{Y-\Delta A \sim \boldsymbol{W}}$)
MCI	-0.19(-0.49, 0.11)	-0.18(-0.45, 0.08)	25%
METS	-3.66(-6.83, -0.49)	-3.60(-6.71, -0.50)	4%
TADS(FLX)	-1.44(-6.02, 3.15)	-4.36(-8.14, -0.58)	32%
TADS(CBT)	2.22(-1.93, 6.38)	0.50(-3.20, 4.20)	21%
TADS(CMB)	-6.64(-10.97, -2.32)	-7.65(-11.28, -4.03)	30%

For the MCI trial, the unadjusted treatment effect estimate was $\widehat{\Delta}^{unadj} = -0.19$ CDR-SB points with standard error 0.15 and 95% CI $(-0.49, 0.11)$, and the ANCOVA estimate was $\widehat{\Delta}^{ancova} = -0.18$ CDR-SB points with standard error 0.13 and 95% CI $(-0.45, 0.08)$. Compared to the unadjusted estimator, the ANCOVA estimator has a 14% narrower confidence interval and 25% smaller variance, indicating that researchers planning to perform an adjusted analysis could achieve the same precision as the unadjusted analysis with approximately 25% fewer participants.

For the METS trial, the unadjusted treatment effect estimate is $\widehat{\Delta}^{unadj} =$

-3.66 kg of weight change with standard error 1.62 and 95% CI $(-6.83, -0.49)$, and the ANCOVA estimate is $\widehat{\Delta}^{ancova} = -3.60$ with standard error 1.58 and 95% CI $(-6.71, -0.50)$. Adjustment resulted in a 4% variance reduction.

For the TADS trial, as shown in Table 2.1, covariate adjustment results in substantial variance reduction for all three treatment arms. This stands out for the Fluoxetine arm, where we estimated that covariate adjustment reduced asymptotic variance by 32%. The ANCOVA estimator, unlike the unadjusted estimator, leads to a statistically significant treatment effect -4.36 CDRS-R points (p-value 0.01); the 95% CI of the ANCOVA estimator $(-8.14, -0.58)$ excludes zero, but that of the unadjusted estimator $(-6.02, 3.15)$ does not.

2.6 Practical Recommendations

Consider the case where the primary outcome Y is a change score (difference between final score and baseline score). In some cases, adjusting for the baseline score alone brings substantial variance reduction. For example, for TADS(FLX) and TADS(CMB), adjusting for only the baseline CDRS-R score gives a similar variance reduction as adjusting for all of the baseline covariates. In other cases, the baseline score can have negligible impact while the other covariates provide substantial variance reduction. E.g., in the MCI trial the baseline score provided approximately 0% variance reduction while the other covariates led to an estimated 25% variance reduction. It is fine to adjust for correlated baseline variables as long as each adds some new prognostic information for the primary outcome.

When the trial has missing outcomes, under the assumption of missing at random (that the outcome distribution is the same for those with missing outcomes as for those with observed outcomes, conditional on treatment assignment and baseline covariates), the unadjusted estimator may no longer be consistent. This can happen if, e.g., participants who benefit more from treatment are more likely to drop out than those who benefit less. The ANCOVA estimator remains consistent under missing at random if the ANCOVA model is correctly specified. To add robustness to model misspecification, one can use a propensity score model for missing outcomes (modeling the probability of missingness given treatment assignment and covariates with, e.g., a logistic regression model) as the inverse weight when fitting the ANCOVA model among those with observed outcomes. According to Robins et al. (2007), this estimator due to Marshall Joffe is doubly-robust, i.e., consistent as long as one of the two models (propensity score model or ANCOVA model) is correctly specified, under the missing at random assumption. For the three trial examples in this paper, the ANCOVA estimator (which does not incorporate information from participants with missing outcomes) and the aforementioned doubly-robust extension (which incorporates information from all participants) gave similar estimates and confidence intervals. See the Supporting Information for details.

For large trials, e.g., with total sample size at least 500, adjusting for prognostic baseline variables (if there are any) is highly recommended since it reduces the required sample size to achieve a desired power (EMA, 2015). This is counter to the (false) intuition that in large trials there is little to gain from covariate adjustment since randomization will likely leave little imbalance to adjust for. Adjustment can still be useful at large sample sizes, and arguably can be more useful since it leads to greater absolute reductions in the required sample size. For example, the METS trial involved 146 participants and our estimate of $R^2_{Y-\Delta A \sim \boldsymbol{W}}$ is 4%. If this were the true value of $R^2_{Y-\Delta A \sim \boldsymbol{W}}$, it would mean about 6 fewer participants required to achieve the same power as the unadjusted estimator; if this trial were 10 times larger, i.e., 1460 participants, then covariate adjustment would lead to a sample size reduction of approximately 60 participants.

Even when a randomized trial ends up having negligible imbalance, covariate adjustment can still increase power when the hypothesis test is based on dividing the estimator by its standard error and rejecting the null hypothesis when this ratio exceeds a threshold. This results from the fact that power is related to the variance through the standard error in the denominator of the test statistic. For example, even though the MCI trial is well balanced, there is still an estimated 25% variance reduction from adjustment. When there are prognostic baseline variables, we recommend that covariate adjustment be preplanned as the primary efficacy analysis.

If the outcome is binary, count, ordinal or time-to-event, then covariate adjustment can be done using estimators of, e.g., Moore and van der Laan (2009a), Lu and Tsiatis (2011), Howard et al. (2012) and Díaz et al. (2018). However, robust variance estimators typically must be used, e.g., when constructing confidence intervals or conducting hypothesis tests. The sandwich estimator could be used as described by Tsiatis et al. (2008); alternatively, the nonparametric bootstrap could be used. Because of these results for other outcome types, it was surprising that when using ANCOVA it is unnecessary to use these robust variance estimators (since as proved in Section 2.4.2 the standard, model-based variance estimator for ANCOVA is already robust to arbitrary model misspecification).

We assumed equal randomization probabilities to the two trial arms. If unequal probabilities are used, then a robust variance estimator is needed for ANCOVA.

We also assumed that the data vector for each participant is an independent, identically distributed draw from an unknown distribution. This assumption does not hold if stratified randomization or covariate-adaptive randomization is used. For stratified randomization and some types of covariate-adaptive randomization, Bugni et al. (2018) showed that if the covariates in the ANCOVA model are the indicators of the strata used in the randomization procedure, then the ANCOVA estimator is consistent; furthermore, its model-based variance estimator is consistent if the limiting probability of assignment to each arm is $1/2$ within each stratum. The previous sentence holds regardless of whether the true data generating distribution satisfies any of the ANCOVA model assumptions. In general, adjusting for stratification variables is recommended when using stratified randomization or covariate-adaptive randomization (Lachin et al., 1988; Kahan and Morris, 2012; EMA, 2015). It is an open question, to the best of our knowledge, as to what happens when more variables than the stratification indicators are included in the ANCOVA model under such randomization schemes, in terms of consistency of the ANCOVA estimator and how to compute its asymptotic variance under arbitrary model misspecification.

How to best pick the set of covariates to use in an adjusted estimator is a challenging problem. The methods of Moore and van der Laan (2009a) and Moore et al. (2011) use cross-validation, Bloniarz et al. (2016) and Tian et al. (2019) use LASSO, and Wager et al. (2016) use a combination of regression and cross-validation. All aspects of the covariate adjustment method need to be prespecified in the study protocol (FDA and EMA, 1998).

Chapter 3

Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment

The content of this chapter is reproduced from Wang et al. (2019) available at `https://arxiv.org/abs/1910.13954` and has been submitted to *Journal of American Statistical Association* for consideration.

CHAPTER 3. MODEL-ROBUST INFERENCE UNDER STRATIFIED RANDOMIZATION AND COVARIATE ADJUSTMENT

3.1 Introduction

A joint guidance document from the U.S. Food and Drug Administration and the European Medicines Agency (FDA and EMA, 1998) states that "Pretrial deliberations should identify those covariates and factors expected to have an important influence on the primary variable(s), and should consider how to account for these in the analysis to improve precision and to compensate for any lack of balance between treatment groups." More recent regulatory guidance documents also encourage consideration of baseline variables in order to improve precision in randomized trials (EMA, 2015; FDA, 2019, 2020). Though there is a rich statistical literature on model-robust methods to adjust for baseline variables in randomized trials that use simple randomization, less is known for trials that use other forms of randomization. This is a practical concern since, as discussed below, many clinical trials use other forms of randomization.

"Covariate-adaptive randomization" refers to randomization procedures that take baseline variables into account when assigning participants to study arms. The goal is to achieve better balance across study arms in preselected strata of the baseline variables compared to simple randomization (which ignores baseline variables). E.g., balance on disease severity, a genetic marker, or another variable thought to be correlated with the primary outcome could be sought.

CHAPTER 3. MODEL-ROBUST INFERENCE UNDER STRATIFIED RANDOMIZATION AND COVARIATE ADJUSTMENT

The simplest and most commonly used type of covariate-adaptive randomization is stratified permuted block randomization (Zelen, 1974), referred to as "stratified randomization" throughout, for conciseness.

Compared with simple randomization, covariate-adaptive randomization can be advantageous in minimizing imbalance and improving efficiency (Efron, 1971; Pocock and Simon, 1975; Wei, 1978). Due to these benefits, covariate-adaptive randomization has become a popular approach in clinical trials. According to a survey by Lin et al. (2015), 183 out of their sample of 224 randomized clinical trials published in 2014 in leading medical journals used some form of covariate-adaptive randomization. Stratified randomization was implemented by 70% of trials in this survey. Another method for covariate-adaptive randomization is the biased-coin design by Efron (1971), which we call "biased-coin randomization" throughout. Other examples include Wei's urn design (Wei, 1978) and rerandomization (Morgan and Rubin, 2012). We only consider the following two types of covariate-adaptive randomization: stratified randomization and biased-coin randomization.

Concerns have been raised regarding how to perform valid statistical analyses at the end of trials that use covariate-adaptive randomization. Adjusting for stratification variables is recommended (Lachin et al., 1988; Kahan and Morris, 2012; EMA, 2015). However, this recommendation is not reliably carried out. Kahan and Morris (2012) sampled 65 published trials from major medical journals from March to May 2010 and found that 41 implemented covariate-adaptive randomization (among which 29 used stratified randomization), but only 14 adjusted in the primary analysis for the variables used in the randomization procedure. Furthermore, many results on how to conduct the primary efficacy analysis in trials that use stratified randomization require one to assume a correctly specified regression model, e.g., Shao et al. (2010); Shao and Yu (2013); Ma et al. (2015, 2018); Yang et al. (2020). Our focus is on model-robust estimators, i.e., estimators that do not require such an assumption when there is no missing data or when outcome data are missing completely at random; when censoring depends on baseline variables, additional assumptions are generally required.

Yang and Tsiatis (2001) showed that the analysis of covariance (ANCOVA) estimator is consistent and asymptotically normal under simple randomization, and that this holds under arbitrary misspecification of the linear regression model used to construct the estimator. Analogous results for the ANCOVA estimator were shown by Bugni et al. (2018) under a variety of covariate-adaptive randomization procedures that include stratified and biased-coin randomization; however, their results only allow adjustment for the variables used in the randomization procedure. The proofs of our results build on key ideas from their work as described below. The results of Li and Ding (2020) and Liu and Yang (2020) for the ANCOVA estimator are robust to arbitrary misspecification of the linear regression model; however, they use the randomization inference framework while many clinical trials are analyzed using the superpopulation inference framework (as done here); see Robins (2002) for a comparison of these frameworks. All of the results in this paragraph are for the ANCOVA estimator, and so do not apply to logistic regression models for binary outcomes nor to commonly used models for time-to-event outcomes. Ye and Shao (2020) derived asymptotic distributions for log-rank and score tests in survival analysis under covariate-adaptive randomization; however, estimation was not addressed.

For trials using stratified or biased-coin randomization, to the best of our knowledge, it was an open problem to determine (in the commonly used superpopulation inference framework and without making parametric model assumptions) the large sample properties of estimators that involve any of the following features: binary or time-to-event outcomes, adjustment for baseline variables in addition to those in the randomization procedure, and missing data under the missing at random assumption. This is the problem that we address, and we think that each of the above features can be important in the analysis of clinical trials. For example, binary and time-to-event outcomes are commonly used in clinical trials. According to a survey by Austin et al. (2010) on trials published in leading medical journals in 2007, 74 out of 114 trials involved binary or time-to-event outcomes. As we show in our data analyses, the addition of baseline variables beyond those used for stratified randomization can lead to substantial precision gains. Handling missing data is also important in order to avoid bias in treatment effect estimation.

Under regularity conditions, we prove that a large class of estimators is consistent and asymptotically normally distributed in randomized trials that use stratified or biased-coin randomization, and we give a formula for computing their asymptotic variance. This class of estimators consists of all M-estimators that are consistent under simple randomization. Examples are listed in Section 3.4. We prove analogous results for the Kaplan-Meier (K-M) estimator (Kaplan and Meier, 1958) of the survival function. Underlying these results is our general technique for characterizing the large sample behavior of asymptotically linear estimators under stratified or biased-coin randomization, described in Section 3.7.

Our theorems imply that under standard regularity conditions, whenever an estimator in our class is consistent and asymptotically normally distributed under simple randomization, then it is consistent and asymptotically normally distributed under stratified (or biased-coin) randomization. Also, its influence function is the same regardless of whether data is generated under simple, stratified or biased-coin randomization. This can be advantageous since for many estimators used to analyze randomized trials, their influence functions have already been derived under simple randomization. An estimator's influence function can be input into our formula (3.5) to produce a consistent variance estimator under stratified and biased-coin randomization.

As in the aforementioned work, we assume that the randomization procedure and analysis method have been completely specified before the trial starts, as is typically required by regulators (FDA and EMA, 1998; EMA, 2015; FDA, 2019, 2020).

In the next section, we describe three trial examples to which we apply our methods. In Section 3.3, we describe our setup, notation and assumptions. We present our main results in Section 3.4. In Section 3.5, we give example estimators for continuous and binary outcomes to which our general results apply. In Section 3.6, we present asymptotic results for the Kaplan-Meier estimator for time-to-event outcomes. Trial applications are provided in Section 3.7. Practical recommendations and future directions are discussed in Section 3.8.

3.2 Three completed trials that used stratified randomization

In some cases, the outcomes in our analyses differ from the primary outcomes in the corresponding trials. This is because we wanted similar outcomes across trials for illustration.

3.2.1 Buprenorphine tapering and illicit opioid use (NIDA-CTN-0003)

The trial of “Buprenorphine tapering schedule and illicit opioid use” in the National Drug Abuse Treatment Clinical Trials Network (NIDA-CTN-0003), is a phase-3 randomized trial completed in 2005 (Ling et al., 2009). The goal was to compare the effects of a short or long taper schedule after buprenorphine stabilization of patients with opioid use disorder. Patients were randomized into two arms: 28-day taper (control, 259 patients, 36% missing outcomes) and 7-day taper (treatment, 252 patients, 21% missing outcomes), stratified by maintenance dose (3 levels) measured at randomization. The outcome of interest is a binary indicator of whether a participant’s urine tested at the end of the study is opioid-free. In addition to the stratification variable, we adjust for the following baseline variables: sex, opioid urine toxicology results, the Adjective Rating Scale for Withdrawal (ARSW), the Clinical Opiate Withdrawal Scale (COWS) and the Visual Analog Scale (VAS).

3.2.2 Prescription opioid addiction treatment (NIDA-CTN-0030)

The “Prescription Opioid Addiction Treatment Study” (NIDA-CTN-0030) is a phase-3 randomized trial completed in 2013 (Weiss et al., 2011). The goal was to determine whether adding individual drug counseling to the prescription of buprenorphine/naloxone would improve outcomes for patients with prescription opioid use disorder. Though this study adopted a 2-phase adaptive design, we focus on the first phase, in which patients were randomized into standard medical management (control, 330 patients, 10% missing outcomes) or standard medical management plus drug counseling (treatment, 335 patients, 13% missing outcomes). Randomization was stratified by the presence or absence of (i) a history of heroin use and (ii) current chronic pain, resulting in 4 strata. The outcome of interest is the proportion of negative urine laboratory results among all tests (treated as a continuous outcome between 0 and 1). Among all 5 urine laboratory tests during the first 4 weeks of phase I, if a patient missed two consecutive visits, then the outcome is regarded as missing. We included the following baseline variables in the analysis: randomization stratum, age, sex and urine laboratory results.

3.2.3 Internet-delivered treatment for substance abuse (NIDA-CTN-0044)

The phase-3 randomized trial “Internet-delivered treatment for substance abuse” (NIDA-CTN-0044) was completed in 2012 (Campbell et al., 2014). The goal was to evaluate the effectiveness of a web-delivered behavioral intervention, Therapeutic Education System (TES), in the treatment of substance abuse. Participants were randomly assigned to two arms: treatment as usual (control, 252 participants, 19% missing outcomes) and treatment as usual plus TES (treatment, 255 participants, 18% missing outcomes).

Randomization was stratified by site, patient’s primary substance of abuse (stimulant or non-stimulant) and abstinence status at baseline. Unfortunately, the available data set for this trial did not include the site variable. Our analyses and claims in Section 3.7 assume that the only randomization strata are the patient’s primary substance of abuse and abstinence status at baseline (4 levels overall). Our theorems imply that ignoring one or more randomization stratum variables leads to conservative variance estimates when using our variance formulas, as explained in Section 3.8.

After randomization, each participant was followed for 12 weeks with 2 urine laboratory tests per week. The outcome of interest is the proportion of negative urine lab results among all tests (treated as a continuous outcome between 0 and 1). If a participant missed visits of more than 6 weeks, the outcome is regarded as missing. We adjust for randomization stratum and the following additional baseline variables: age, sex and urine laboratory result.

We also analyze a second outcome: time to abstinence, defined as the time to first two consecutive negative urine tests during the study. Censoring time is defined as the first missing visit. We used the data from the first 6 weeks of follow-up in our data analysis of this time-to-event outcome, during which 99% of the events occurred.

3.3 Definitions and assumptions

3.3.1 Data generating distributions

We focus on two-arm randomized trials that use simple, stratified or biased-coin randomization. Let n denote the sample size. For each participant $i = 1, \dots, n$, let Y_i denote the primary outcome, M_i denote whether Y_i is observed ($M_i = 1$) or missing ($M_i = 0$), A_i denote study arm assignment ($A_i = 1$ if assigned to treatment and $A_i = 0$ if assigned to control), and $\boldsymbol{X}_i$ denote a vector of baseline covariates. This notation is for real-valued outcomes, e.g. continuous or binary outcomes. Modified definitions, assumptions, and results for time-to-event outcomes are in Section 3.6.

We use the Neyman-Rubin potential outcomes framework (Neyman et al., 1990), which assumes the existence of potential outcomes $Y_i(0)$ and $Y_i(1)$ for each participant i. These represent the outcome that would be observed under assignment to study arm 0 or 1, respectively. Though using potential outcomes introduces additional notation, it is needed in order to rigorously define the data generating distributions under the different randomization procedures that we consider. We make the following consistency assumption linking the observed outcome Y_i to the potential outcomes: $Y_i = Y_i(A_i) = Y_i(1)A_i + Y_i(0)(1 - A_i)$ for each participant i. Also, let $M_i(a)$ be the indicator of whether participant i would have a non-missing outcome if they get assigned to study arm $a \in \{0, 1\}$. We assume, analogous to the consistency assumption above, that $M_i = M_i(A_i) = M_i(1)A_i + M_i(0)(1 - A_i)$.

For each participant i, we define the full data vector (including potential outcomes, some of which are not observed) $\boldsymbol{W}_i = (Y_i(1), Y_i(0), M_i(1), M_i(0), \boldsymbol{X}_i)$ and the observed data vector $\mathbf{O}_i = (A_i, \boldsymbol{X}_i, Y_iM_i, M_i)$. The reason that the product Y_iM_i appears in the observed data vector $\mathbf{O}_i$ is to encode that whenever the outcome is missing ($M_i = 0$), the outcome value Y_i is not available in $\mathbf{O}_i$ (since $Y_iM_i = 0$). The data available to the analyst at the end of the trial are $\mathbf{O}_1, \ldots, \mathbf{O}_n$.

We make the following assumptions on the distribution of $\{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n\}$:

Assumption 1.

(i) $\boldsymbol{W}_i, i = 1, \ldots, n$ are independent, identically distributed samples from an unknown joint distribution P on $\boldsymbol{W} = (Y(1), Y(0), M(1), M(0), \boldsymbol{X})$.

(ii) Missing at random: $M(a) \perp\!\!\!\perp Y(a) | \boldsymbol{X}$ for each arm $a \in \{0, 1\}$, where $\perp\!\!\!\perp$ denotes independence.

Throughout, we use E to denote the expectation with respect to distribution P.

3.3.2 Randomization procedures: simple, stratified, and biased-coin

First consider simple randomization, which assigns study arms $A_1, \ldots, A_n$ by independent Bernoulli draws each with fixed probability π of being 1, e.g., using a random number generator. By design, the draws are independent of each other and of all participant characteristics measured before randomization or not impacted by randomization. Therefore, we have that $(A_1, \ldots, A_n)$ is independent of $(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)$, and that the observed data $\mathrm{O}_1, \ldots, \mathrm{O}_n$ are independent, identically distributed.

Next consider stratified or biased-coin randomization, where treatment allocation depends on predefined baseline strata, such as gender, age, site, disease severity, or combinations of these. We refer to the baseline strata that are used in the randomization procedure as “randomization strata”. The baseline stratum of participant i is denoted by the single, categorical variable S_i taking K possible values. For example, if randomization strata are defined by 4 sites and a binary indicator of high disease severity, then S has $K = 8$ possible values. Let S_i denote the stratification variable for participant i and let $\mathcal{S} = \{1, \ldots, K\}$ denote the set of all K randomization strata. The goal of stratified or biased-coin randomization is to achieve balance in each stratum; that is, the proportion of participants assigned to the treatment arm is targeted to the prespecified proportion $\pi \in (0, 1)$, e.g. $\pi = 0.5$. Throughout, the stratification variable S is encoded in the baseline covariate vector $\boldsymbol{X}$ using $K - 1$ dummy variables that make up the first $K - 1$ components of $\boldsymbol{X}$ (which can include additional baseline variables).

Stratified randomization uses permuted blocks to assign treatment. For each randomization stratum, a randomly permuted block with fraction π 1's (representing treatment) and $(1 - \pi)$ 0's (representing control) is used for sequential allocation. When a block is exhausted, a new block is used.

Biased-coin randomization can be applied when $\pi = 0.5$ and it allocates participants sequentially by the following rule for $k = 1, \ldots, n$:

$$P(A_k = 1 | S_1, \ldots, S_k, A_1, \ldots, A_{k-1}) = \begin{cases} 0.5, \text{ if } \sum_{i=1}^{k-1}(A_i - 0.5)I\{S_i = S_k\} = 0 \\ \lambda, \text{ if } \sum_{i=1}^{k-1}(A_i - 0.5)I\{S_i = S_k\} < 0 \\ 1-\lambda, \text{ if } \sum_{i=1}^{k-1}(A_i - 0.5)I\{S_i = S_k\} > 0 \end{cases}$$

where $\lambda \in (0.5, 1]$, $I\{Z\}$ is the indicator function that has value 1 if Z is true and 0 otherwise, and by convention the first participant is assigned with probability 0.5 to each arm. Our results for biased-coin randomization assume that $\pi = 0.5$.

When comparing the three types of randomization procedures (simple, stratified, or biased-coin), we assume that all use the same value of π. For the stratified randomization and biased-coin designs, it follows by construction (and was shown by Bugni et al., 2018) that the study arm assignments $(A_1, \ldots, A_n)$ are conditionally independent of the participant baseline variables and potential outcomes $(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)$ given the randomization strata $(S_1, \ldots, S_n)$. Intuitively, this is because the study arm assignment procedure only has access to the participants' randomization strata. Under stratified or biased-coin randomization, the observed data vectors $\mathbf{O}_1, \ldots, \mathbf{O}_n$ are not independent.

Under any of the three randomization procedures, the observed data vectors $\mathbf{O}_1, \ldots, \mathbf{O}_n$ are identically distributed; that is, the distribution of $\mathbf{O}_1$ is the same as that of $\mathbf{O}_2$, etc. Let P^* denote this distribution, i.e., the distribution of a generic, observed data vector $\mathbf{O} = (A, \boldsymbol{X}, YM, M)$. This distribution is the same for each of the three randomization procedures, and is that induced by first drawing a single realization $\boldsymbol{W} = (Y(1), Y(0), M(1), M(0), \boldsymbol{X})$ from the distribution P (see Assumption 1), then drawing A as an independent Bernoulli draw with probability π of being 1, and lastly applying the consistency assumptions $Y = Y(1)A + Y(0)(1-A)$ and $M = M(1)A + M(0)(1-A)$ to construct Y, the (non)-missingness indicator M, and their product YM. The corresponding expectation with respect to P^* is denoted E^*, which is used below. The claims in this paragraph are proved in the Supplementary Material.

3.3.3 Targets of inference (estimands) and estimators

For continuous and binary outcomes, our goal is to estimate a population parameter Δ^*, which is a contrast between the marginal distributions of $Y(1)$ and $Y(0)$. For example, Δ^* can be defined as the population average treatment effect $E[Y(1)] - E[Y(0)]$.

We consider M-estimators of Δ^* (van der Vaart, 1998, Ch. 5). Let $\boldsymbol{\theta} = (\Delta, \boldsymbol{\beta}^t)^t$ denote a column vector of $p+1$ parameters where $\Delta \in \mathbb{R}$ is the parameter of interest and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a column vector of p nuisance parameters. We define the M-estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\Delta}, \widehat{\boldsymbol{\beta}}^t)^t$ to be the solution to the following estimating equations:

$$\sum_{i=1}^{n} \boldsymbol{\psi}(A_i, \boldsymbol{X}_i, Y_i, M_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{3.1}$$

where $\boldsymbol{\psi}$ is a column vector (with $p+1$ components) of known functions. We define $\widehat{\Delta}$ to be the estimator of Δ^*. We assume that $\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})$ does not depend on the outcome Y when $M = 0$ (since then Y is missing). Many estimators used in clinical trials, including all estimators defined in Section 3.5.1, can be expressed as solutions to estimating equations (3.1) for an appropriately chosen estimating function $\boldsymbol{\psi}$.

For time-to-event outcomes, the K-M estimator of the survival curve is commonly used. Since it is not an M-estimator, our general result (Theorem 2) for M-estimators below does not apply. We separately prove analogous results for the K-M estimator; see Section 3.6.

We assume regularity conditions similar to the classical conditions that are used for proving consistency and asymptotic linearity of M-estimators for independent, identically distributed data, as given in Section 5.3 of van der Vaart (1998). One of the conditions is that the expectation of the estimating equations

$$E^*[\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})] = \mathbf{0}, \tag{3.2}$$

has a unique solution in θ, which is denoted as $\underline{\boldsymbol{\theta}} = (\underline{\Delta}, \underline{\boldsymbol{\beta}}^t)^t$. The other regularity conditions are given in the Supplementary Material.

We assume that the estimating equations ψ were chosen to ensure that the property $\Delta^* = \underline{\Delta}$ holds. This property is generally needed to show consistency of the M-estimator $\widehat{\Delta}$ for Δ^* under simple randomization, and has previously been proved for all of the estimators in Section 3.5.1. In general, whether the property $\Delta^* = \underline{\Delta}$ holds does not depend on the randomization procedure (simple, stratified, or biased-coin randomization); this is because the property depends only on ψ, P and P^*.

Results in Section 5.3 of van der Vaart (1998) imply that under simple randomization, given Assumption 1 and the regularity conditions in the Supplementary Material, $\widehat{\Delta}$ converges in probability to $\underline{\Delta}$ and is asymptotically normally distributed with asymptotic variance that we denote by $\widetilde{V}$. We focus on determining what happens under stratified or biased-coin randomization, where our main result (Section 3.4) is that consistency and asymptotic normality still hold but the asymptotic variance may be smaller (and a consistent variance estimator is given).

3.4 Main result for M-estimators

Consider the setup in Section 3.3.3, where the M-estimator $\widehat{\Delta}$ is defined. The proof of the following theorem (and all results in the paper) is given in the

Supplementary Material:

Theorem 2. *Assume the regularity conditions in the Supplementary Material, $\Delta^* = \underline{\Delta}$, and Assumption 1. Then under simple, stratified, or biased-coin randomization, we have consistency, i.e., $\widehat{\Delta} \to \Delta^*$ in probability, and asymptotic linearity, i.e.,*

$$\sqrt{n}(\widehat{\Delta} - \Delta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(A_i, \boldsymbol{X}_i, Y_i, M_i) + o_p(1), \tag{3.3}$$

where the influence function $IF(A, \boldsymbol{X}, Y, M)$ is the first entry of $-\boldsymbol{B}^{-1}\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \underline{\boldsymbol{\theta}})$ for $\boldsymbol{B} = E^*\left[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\underline{\boldsymbol{\theta}}}\right]$.

For stratified and biased-coin randomization, $\sqrt{n}(\widehat{\Delta} - \Delta^*) \xrightarrow{d} N(0, V)$ for

$$V = \widetilde{V} - \frac{1}{\pi(1-\pi)} E^*\left[E^*\left\{(A-\pi) IF(A, \boldsymbol{X}, Y, M)|S\right\}^2\right], \tag{3.4}$$

where $\widetilde{V} = E^*\{IF(A, \boldsymbol{X}, Y, M)^2\}$ is the asymptotic variance under simple randomization. The asymptotic variance V can be consistently estimated by formula (3.5) below.

Theorem 2 implies that whenever an M-estimator $\widehat{\Delta}$ is consistent and asymptotically normally distributed under simple randomization, then it is consistent and asymptotically normally distributed under stratified (or biased-coin) randomization with equal or smaller asymptotic variance. Also, its influence function is the same regardless of whether data is generated under simple, stratified, or biased-coin randomization. This can be advantageous since for many estimators used to analyze randomized trials, their influence functions have already been derived under simple randomization; the last display in Theorem 2 gives a formula for calculating the asymptotic variances for these estimators under the other two randomization procedures, in terms of the influence function.

For the unadjusted estimator $\widehat{\Delta} = \sum_{i=1}^{n} Y_i A_i / \sum_{i=1}^{n} A_i - \sum_{i=1}^{n} Y_i(1-A_i) / \sum_{i=1}^{n}(1-A_i)$, our Theorem 2 is equivalent to Theorem 4.1 of Bugni et al. (2018) under stratified or biased-coin randomization. In the special case of continuous outcomes, if the ANCOVA estimator is used with $X = S$, then Theorem 2 is equivalent to the result in section 4.2 of Bugni et al. (2018) under stratified or biased-coin randomization, though their results also handle other types of covariate-adaptive randomization.

Theorem 2 above extends the results of Bugni et al. (2018) to handle the class of M-estimators, that is, estimators calculated by solving estimating equations (3.1). This includes, for example, the ANCOVA estimator that adjusts for baseline covariates in addition to those used in the randomization procedure (Example 1 of Section 3.5.1 below), the standardized logistic regression estimator for binary outcomes (Example 2 of Section 3.5.1), and the DR-WLS estimator (Example 3 of Section 3.5.1). This class of estimators also includes the inverse-probability-weighted estimator (IPW, Robins et al., 1994), the augmented inverse probability weighted estimator (AIPW, Robins et al., 1994; Scharfstein et al., 1999), the Mixed-effects Model for Repeated Measures estimator (MMRM, Mallinckrodt et al., 2003; Siddiqui et al., 2009; EMA, 2019), and targeted maximum likelihood estimators (TMLE) that converge in 1-step (van der Laan and Gruber, 2012), among others. Thus, Theorem 2 covers estimators that handle various outcome types, repeated measures outcomes, missing outcome data, and covariate adjustment. Our proof relies on key ideas from Lemmas B.1 and B.3 in the Supplement of Bugni et al. (2018).

We prove consistency of the following estimator for the asymptotic variance V, which is the following empirical counterpart of the right side of (3.4):

$$\widehat{V} = \tilde{V}_n - \frac{1}{\pi(1-\pi)} E_n\left[E_n\{(A-\pi)IF(A, \boldsymbol{X}, Y, M)|S\}^2\right], \tag{3.5}$$

where $\tilde{V}_n$ is the sandwich variance estimator of $\widehat{\Delta}$ (Section 3.2 of Tsiatis, 2007), defined as the first-row first-column entry of

$$\frac{1}{n}\left\{E_n\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right]\right\}^{-1}\left\{E_n\left[\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \widehat{\boldsymbol{\theta}})\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \widehat{\boldsymbol{\theta}})^t\right]\right\}$$
$$\left\{E_n\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right]\right\}^{-1,t},$$

and E_n denotes expectation with respect to the empirical distribution of the observed data $\mathbf{O}_1, \dots, \mathbf{O}_n$.

3.5 Example estimators for continuous and binary outcomes

3.5.1 ANCOVA, standardized logistic regression, and DR-WLS

We give several examples of estimators that Theorem 2 applies to. For estimators defined in Examples 1-3, the parameter of interest, i.e. Δ^*, is the average treatment effect defined as $E[Y(1)] - E[Y(0)]$, and we denote $\boldsymbol{Z} = (1, A, \boldsymbol{X}^t)^t$. In Examples 1 and 2, we assume no missing data and we do not assume that the working models, i.e., the linear regression model in Example 1 and the logistic regression model in Example 2, are correctly specified.

Example 1. For continuous outcomes, the ANCOVA estimator $\widehat{\Delta}_{ancova}$ for Δ^* involves first fitting a linear regression working model $E[Y|A, \boldsymbol{X}] = \beta_0 + \Delta A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}$ using ordinary least squares and then letting $\widehat{\Delta}_{ancova}$ be the estimate of Δ. The ANCOVA estimator can be equivalently calculated by solving estimating equations (3.1) letting

$$\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta}) = \{Y - (\beta_0 + \Delta A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X})\}\boldsymbol{Z}.$$

Example 2. For binary outcomes, the standardized logistic regression estimator $\widehat{\Delta}_{logistic}$ is calculated by first fitting a working model: $P(Y = 1|A, \boldsymbol{X}) =$ expit$(\beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X})$, where expit$(x) = 1/(1 + e^{-x})$, and getting the maximum likelihood estimates $(\widehat{\beta}_0, \widehat{\beta}_A, \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t)^t$. Then define $\widehat{\Delta}_{logistic} = \frac{1}{n}\sum_{i=1}^n\{$expit$(\widehat{\beta}_0 + \widehat{\beta}_A + \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}_i) -$ expit$(\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}_i)\}$. Equivalently, the estimator $\widehat{\Delta}_{logistic}$ is the solution to estimating equations (3.1) letting

$$\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta}) = \begin{pmatrix} \text{expit}(\beta_0 + \beta_A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}) - \text{expit}(\beta_0 + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}) - \Delta \\ \{Y - \text{expit}(\beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X})\}\boldsymbol{Z} \end{pmatrix}.$$

This estimator is mentioned as potentially useful in COVID-19 treatment and prevention trials in a recent FDA guidance (FDA, 2020).

Example 3. When some outcomes are missing, then one can estimate Δ^* by the DR-WLS estimator, which generalizes the estimators in Examples 1 and 2. However, due to missing data this estimator requires additional assumptions (as is true for all estimators) described below. The DR-WLS estimator can be used with binary or continuous outcomes. The estimator is calculated by first fitting the logistic regression working model:

$$P(M=1|A,\boldsymbol{X}) = \text{expit}(\alpha_0 + \alpha_A A + \boldsymbol{\alpha}_{\boldsymbol{X}}^t \boldsymbol{X}) \tag{3.6}$$

and getting the maximum likelihood estimates $(\widehat{\alpha}_0, \widehat{\alpha}_A, \widehat{\boldsymbol{\alpha}}_{\boldsymbol{X}}^t)^t$ of parameters $(\alpha_0, \alpha_A, \boldsymbol{\alpha}_{\boldsymbol{X}}^t)^t$. Next, fit the following working model for the outcome given study arm and baseline variables (from the generalized linear model family):

$$E[Y|A,\boldsymbol{X}] = g^{-1}(\beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}), \tag{3.7}$$

with weights $1/\text{expit}(\widehat{\alpha}_0 + \widehat{\alpha}_A A_i + \widehat{\boldsymbol{\alpha}}_{\boldsymbol{X}}^t \boldsymbol{X}_i)$ using only the data with $M_i = 1$. Here the inverse link function is $g^{-1}(x) = x$ for continuous outcomes and $g^{-1}(x) = \text{expit}(x)$ for binary outcomes. Third, the DR-WLS estimator is

$$\widehat{\Delta}_{DR-WLS} = \frac{1}{n}\sum_{i=1}^{n}\{g^{-1}(\widehat{\beta}_0 + \widehat{\beta}_A + \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}) - g^{-1}(\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X})\}.$$

The DR-WLS estimator can be expressed as the solution to estimating equations (given in the Supplementary Material) of the general form (3.1). For the DR-WLS estimator, we assume that at least one of the two working models (3.6) and (3.7) is correctly specified, and $\inf_{(a,\mathrm{x})\in(\mathcal{A},\mathcal{X})} P(M=1|a,\mathrm{x}) > 0$, where $(\mathcal{A},\mathcal{X})$ is the support of $(A,\boldsymbol{X})$.

The ANCOVA estimator and the standardized logistic regression estimator are special cases of the DR-WLS estimator. If there are no missing data, which means $M_i = 1$ for $i = 1, \ldots, n$, and the regression weights used to fit (3.7) are constant, then $\widehat{\Delta}_{DR-WLS}$ reduces to $\widehat{\Delta}_{ancova}$ for continuous outcomes and to $\widehat{\Delta}_{logistic}$ for binary outcomes. The DR-WLS estimator can be generalized to allow the addition of interaction terms in the model (3.7).

3.5.2 Asymptotic Results for Estimators in Examples 1-3

Under simple randomization and assuming that $\Delta^* = \underline{\Delta}$, consistency and asymptotic normality for the estimators in Examples 1-3 have been proved by Yang and Tsiatis (2001); Scharfstein et al. (1999); Robins et al. (2007), respectively. Under stratified or biased-coin randomization, Theorem 2 applies to these estimators since each is an M-estimator. In particular, under the conditions in the theorem, each of the three estimators is consistent and asymptotically normal with asymptotic variance that is consistently estimated by (3.5).

Under the additional conditions (a)-(c) listed in the corollary below, for each estimator in Examples 1-3, its asymptotic variance is the same regardless of whether simple, stratified, or biased-coin randomization is used; also, the asymptotic variance is consistently estimated by the sandwich variance estimator $\widetilde{V}_n$. Under such conditions, the estimators and their corresponding sandwich variance estimators can be used to perform hypothesis tests and construct confidence intervals that are asymptotically correct.

Recall that we assume throughout that S is encoded by dummy variables in $\boldsymbol{X}$.

Corollary 1. *Assume that* $\Delta^* = \underline{\Delta}$*, the regularity conditions in the Supplementary Material, and Assumption 1. Consider the ANCOVA estimator or the standardized logistic regression estimator. If any of the conditions (a)-(c) below holds, then under simple, stratified, or biased-coin randomization, the estimator is consistent and asymptotically normally distributed with asymptotic variance* $V = \widetilde{V}$*; furthermore, the sandwich variance estimator is consistent. Conditions:*

(a) $\pi = 0.5$*;*

(b) the outcome regression model (3.7) includes indicators for the randomization strata and also treatment-by-randomization-strata interaction terms;

(c) the outcome regression model (3.7) is correctly specified.

For the special case of the ANCOVA estimator with $\boldsymbol{X} = S$, Corollary 1 with condition (a) or (b) was proved by Bugni et al. (2018). The claims in Corollary 1 also hold for the the DR-WLS estimator if at least one of the two working models (3.6) and (3.7) is correctly specified and $\inf_{(a,\mathrm{x})\in(\mathcal{A},\mathcal{X})} P(M = 1|a, \mathrm{x}) > 0$, where $(\mathcal{A}, \mathcal{X})$ is the support of $(A, \boldsymbol{X})$.

3.6 Estimators involving time-to-event outcomes

3.6.1 Notation and Assumptions

For time to event outcomes, we use slightly modified notation and assumptions compared to above. We assume that the outcome is right-censored. Let Y_i denote the failure time and M_i denote the censoring time. Other variables including $A_i, \boldsymbol{X}_i$ and the potential outcomes $Y_i(a), M_i(a)$ for $a = 0, 1$ are defined analogously as in Section 3.3. For each participant $i \in \{1, \ldots, n\}$, we observe $(A_i, \boldsymbol{X}_i, U_i, \delta_i)$, where $U_i = \min\{Y_i, M_i\}$ and $\delta_i = I\{Y_i \leq M_i\}$. We further define a restriction time τ such that the time window $t \in [0, \tau]$ is of interest. We define P^* and E^* analogously as in Section 3.3.2, except here they represent the distribution and expectation, respectively, for a single observed data vector $(A, \boldsymbol{X}, U, \delta)$.

The following assumption is made in place of Assumption 1:

Assumption 1'.

(i) $\boldsymbol{W}_i, i = 1, \ldots, n$ are independent, identically distributed samples from an unknown joint distribution P on $\boldsymbol{W} = (Y(1), Y(0), M(1), M(0), \boldsymbol{X})$.

(ii) Censoring completely at random: $M(a) \perp\!\!\!\perp Y(a)$ for each arm $a \in \{0, 1\}$.

(iii) $P(\min\{Y(a), M(a)\} > \tau) > 0$ for each $a = 0, 1$.

Compared with Assumption 1, Assumption 1'(i) is the same as Assumption 1(i), and Assumption 1'(ii) assumes censoring completely at random instead of missing at random. This modification of the assumption on missing data is because we consider the K-M estimator and its consistency generally requires Assumption 1'(ii). Assumption 1'(iii) is often made in survival analysis, which states that there is a positive probability that both the failure time and censoring time occur after τ (under each study arm assignment).

3.6.2 Kaplan-Meier estimator under simple, stratified, and biased-coin randomization

One commonly-used method for survival analysis is the K-M estimator. The goal is to estimate the survival curve $\{S_0^{(a)}(t) : t \in [0, \tau]\}$ for each $a = 0, 1$, where $S_0^{(a)}(t) = P(Y(a) > t)$. This represents the survival curve if everyone in the study population were assigned to study arm a. The K-M estimator is defined as

$$\widehat{S}_n^{(a)}(t) = \prod_{j:T_j \le t} \left(1 - \frac{\sum_{i=1}^n \delta_i I\{A_i = a\} I\{U_i = T_j\}}{\sum_{i=1}^n I\{A_i = a\} I\{U_i \ge T_j\}}\right),$$

where $\{T_j, j = 1, \ldots, m_n\}$ is the list of unique observed failure times.

While the K-M estimator does not adjust for any baseline variable, its variance under simple randomization is typically different than under stratified or biased-coin randomization, and this is not accounted for by standard methods for estimating its variance. Specifically, the standard method for variance estimation will typically overestimate the K-M variance under stratified or biased-coin randomization, leading to wasted power. Our variance estimator below avoids this problem. Since the K-M estimator estimates a survival function rather than a real number or a vector, our Theorem 2 on M-estimators does not apply. The following theorem gives the asymptotic distribution of the K-M estimator under our three different types of randomization. It involves the influence function $IF^{(a)}(A_i, U_i, \delta_i; t)$ for the K-M estimator under simple randomization (Kosorok, 2008, Section 4.2), which is also given in the Supplementary Material.

Theorem 3. *Given Assumption 1', under simple, stratified, or biased-coin randomization, we have for each* $t \in [0, \tau]$ *that*

$$\sqrt{n}(\widehat{S}_n^{(a)}(t) - S_0^{(a)}(t)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF^{(a)}(A_i, U_i, \delta_i; t) + o_{p^*}(1), \tag{3.8}$$

where $o_{p^*}(1)$ *represents a sequence of random variables converging to* 0 *in probability uniformly over* $t \in [0, \tau]$.

For stratified and biased-coin randomization, the process $\{\sqrt{n}(\widehat{S}_n^{(a)}(t) - S_0^{(a)}(t)) : t \in [0, \tau]\}$ *converges weakly to a mean* 0, *tight Gaussian process with covariance function* $V^{(a)}(t, t')$ *defined in the Supplementary Mate-*

rial, which has the following property: for any $t \leq \tau$*,*

$$V^{(a)}(t,t) = \widetilde{V}^{(a)}(t,t) - \frac{1}{\pi(1-\pi)} E^* \left[E^* \left\{ (A-\pi) IF^{(a)}(A,U,\delta;t) | S \right\}^2 \right], \tag{3.9}$$

where $\widetilde{V}^{(a)}(t,t)$ *is the asymptotic variance under simple randomization.* $V^{(a)}(t,t)$ *can be consistently estimated as described in the Supplementary Material.*

Analogous to Theorem 2, Theorem 3 implies that the influence function of the K-M estimator is the same under simple, stratified, and biased-coin randomization. The above theorem implies that under stratified or biased-coin randomization, the K-M estimator is consistent and asymptotically normally distributed with equal or smaller asymptotic variance than under simple randomization. The asymptotic covariance function of the K-M estimator under stratified or biased-coin randomization is given in Appendix C of the Supplementary Material. It can be used to construct pointwise confidence intervals and a simultaneous confidence band.

The challenge in proving Theorem 3 is that the traditional tool for deriving asymptotic normality in survival analysis, i.e., martingale central limit theorems such as Theorem II.5.1 of Andersen et al. (2012) or Theorem 5.1.1 of Fleming and Harrington (2011), is not applicable here because of the dependence among data points introduced by stratified or biased-coin randomization. To overcome the above difficulty, in the proof of Theorem 3 we first developed a central limit theorem for sums of random functions under stratified randomization (Lemma 5 in the Supplementary Material) based on the empirical process results of Shorack and Wellner (2009) combined with generalizations of the techniques from Bugni et al. (2018). We then proved Theorem 3 by generalizing the arguments in our proof of Theorem 2 to handle random functions. We conjecture that, using our central limit theorem, Theorem 3 can be generalized to apply to other estimators of survival functions, such as the covariate-adjusted estimators proposed by Lu and Tsiatis (2011); Zhang (2015), which may improve precision even further.

3.6.3 Other estimators for time-to-event outcomes

Another parameter of interest is the restricted mean survival time, defined as $\Delta^* = E[\min\{Y(1), \tau\} - \min\{Y(0), \tau\}]$. One covariate adjusted estimator of the restricted mean survival time is the augmented inverse probability weighted (AIPW) estimator of Moore and van der Laan (2009b). This estimator is an M-estimator, to which our Theorem 2 applies. When the survival probability at a given time point is the parameter of interest, one can use the K-M estimator or the method from Moore and van der Laan (2009b).

3.7 Clinical trial applications

3.7.1 Binary and continuous outcomes

Table 3.1 summarizes our data analyses involving binary and continuous outcomes. The outcome is binary for NIDA-CTN-0003 and is continuous for NIDA-CTN-0030 and NIDA-CTN-0044. In all cases, the target of inference is the average treatment effect defined as $E[Y(1)] - E[Y(0)]$.

All missing baseline values were imputed by the median for continuous variables and mode for binary or categorical variables. The only estimator in Table 3.1 that adjusts for missing outcomes is the DR-WLS estimator; all other estimators omit data from the participants with missing outcomes. Negative (positive) estimates are in the direction of clinical benefit (harm). For all estimators presented in Table 3.1, the 95% confidence interval (CI) is constructed using the normal approximation with variance calculated from formula (3.5).

For NIDA-CTN-0003, the outcome is binary and "adjusted estimator" in Table 1 refers to the standardized logistic regression estimator. The unadjusted point estimate is -0.104 with 95% CI $(-0.204, -0.004)$. If randomization strata and additional baseline variables are adjusted for (as in the row "Adjusted estimator ($\boldsymbol{X}$)" in Table 1), the point estimate is unchanged but the 95% CI $(-0.184, -0.024)$ is substantially smaller. The corresponding variance reduction due to covariate adjustment, defined as one minus the variance ratio of

Table 3.1: Summary of clinical trial data analyses with each cell giving the point estimate and 95% CI of an estimator. Each row is for a different estimator. "Adjusted estimator" refers to the standardized logistic estimator for the trial with binary outcome (column 2) and to the ANCOVA estimator for the trials with continuous outcomes (columns 3 and 4). The variable in parentheses after the estimator name indicates which variables (if any) are adjusted for, with S denoting the randomization strata only and $\boldsymbol{X}$ denoting the randomization strata and additional baseline variables.

		Clinical Trial:	
	NIDA-CTN-0003	NIDA-CTN-0030	NIDA-CTN-0044
Unadjusted estimator	-0.104(-0.204, -0.004)	0.015(-0.023, 0.052)	-0.093(-0.149, -0.038)
Adjusted estimator (S)	-0.110(-0.209, -0.009)	0.015(-0.022, 0.052)	-0.089(-0.145, -0.033)
Adjusted estimator ($\boldsymbol{X}$)	-0.104(-0.184, -0.024)	0.012(-0.022, 0.046)	-0.087(-0.142, -0.032)
DR-WLS estimator ($\boldsymbol{X}$)	-0.099(-0.180, -0.019)	0.012(-0.022, 0.045)	-0.091(-0.148, -0.035)

"Adjusted estimator ($\boldsymbol{X}$)" to the unadjusted estimator, is 36%. This is equivalent to needing 36% fewer participants to achieve the same power as a trial that uses the unadjusted estimator, asymptotically.

NIDA-CTN-0030 and NIDA-CTN-0044 had continuous-valued outcomes and "Adjusted estimator" in Table 1 refers to the ANCOVA estimator. Covariate adjustment brings 17% and 3% variance reduction for NIDA-CTN-0030 and NIDA-CTN-0044, respectively, compared to the unadjusted estimator. In all cases, the variance reduction from covariate adjustment is larger when the baseline variables are more strongly prognostic for (i.e., more strongly correlated with) the outcome.

In all three trials, the variance reduction due to adjusting for baseline variables beyond S, defined by one minus the variance ratio of "adjusted estimator ($\boldsymbol{X}$)" and "adjusted estimator (S)", is the same (to the nearest percent) as the corresponding variance reduction comparing "adjusted estimator ($\boldsymbol{X}$)" to the unadjusted estimator. This is expected for the ANCOVA estimator since Bugni et al. (2018) showed that "adjusted estimator (S)" and the unadjusted estimator are asymptotically equivalent when the randomization probability $\pi = 0.5$. Also, in all three trials, the "DR-WLS estimator ($\boldsymbol{X}$)", which handles missing outcomes under the missing at random assumption, has a similar point estimate and 95% CI compared to "adjusted estimator ($\boldsymbol{X}$)", which omits missing outcomes. We recommend using the DR-WLS estimator in practice since it is consistent under weaker assumptions than the other estimators considered here.

We next compare the estimated variance (and resulting confidence intervals) based on the sandwich variance estimator versus the variance estimator (3.5). For the unadjusted estimator, using the sandwich variance estimator instead of formula (3.5) may lead to conservative variance estimates, as implied by Theorem 2. For example, for NIDA-CTN-0044, the 95% CI of the unadjusted estimator constructed by formula (3.5) is $(-0.149, -0.038)$, while the 95% CI calculated using the sandwich variance formula is $(-0.162, -0.025)$, which is 23% wider. The former 95% CI is asymptotically correct assuming outcomes are missing completely at random, an assumption that is generally needed for the the unadjusted estimator to be consistent. Furthermore, the variance of the unadjusted estimator calculated by formula (3.5) (which is consistent) is

34% smaller than the variance calculated by the sandwich variance estimator (which is conservative). In contrast, for the adjusted estimator or the DR-WLS estimator, since all three trials have randomization probabillity $\pi = 0.5$, the sandwich variance estimator is not conservative; this follows from Corollary 1.

3.7.2 Time-to-event outcome

Figure 3.1 presents the K-M estimator for time-to-abstinence in the treatment group as defined in Section 3.2.3 for study NIDA-CTN-0044. We estimated the variance of the K-M estimator in two different ways: one ignored the stratification variable and was the estimated variance returned by the "survfit" function in R; the other used our proposed variance formula that takes the stratification into account. For each of the two variance estimators, we constructed corresponding point-wise confidence intervals for the K-M estimator.

While Figure 3.1 shows that confidence intervals based on different variance estimators are very close to each other, there are variance reductions due to accounting for stratification, which can be translated into sample size reduction needed to achieve the desired power. The variance reduction ranges from 1% to 12% as we consider the survival function at different time points. Among all time points, the first time point (one week after randomization) has the greatest variance reduction. The variance formula (3.9) from Theorem 3 accounts for the improved precision due to stratified randomization (unlike stan-

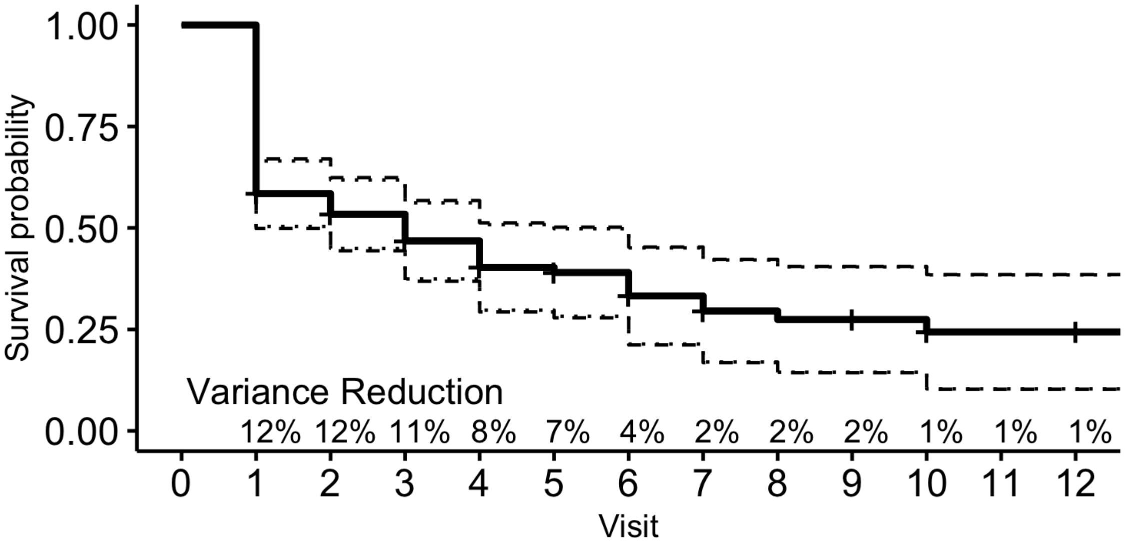

Figure 3.1: The K-M estimator of survival function for NIDA-CTN-0044 treatment group. The solid line is the estimated survival function. Dashed and dotted lines, respectively, represent confidence intervals using the standard method and confidence intervals accounting for randomization strata using (3.9); the dashed and dotted lines are very similar and almost coincide. “Variance Reduction” and the associated percentages represent the variance reduction due to accounting for stratified randomization using (3.9).

dard methods that ignore stratification variables); this can be used to construct more powerful hypothesis tests based on the K-M estimator divided by its standard error. The corresponding figure and results for the control group are given in the Supplementary Material and are qualitatively similar to those described above for the treatment group.

3.8 Discussion

The primary efficacy analysis in confirmatory randomized trials is typically based on a treatment effect estimator that is asymptotically linear under simple randomization; i.e., for an appropriately chosen influence function IF, the estimator has the form (3.3) when estimating a scalar/vector or (3.8) when estimating a function such as a survival curve. All estimators in this paper have this property, and we proved for each estimator covered by Theorems 2-3 that under stratified and biased-coin randomization, it is asymptotically linear with the same influence function as under simple randomization. We then gave formulas (3.4) and (3.9) for the asymptotic variance under stratified (and biased-coin) randomization in terms of the influence function.

Though our theorems cover a variety of estimators used to analyze randomized trials, they do not handle every estimator. However, our results point to a general approach for deriving the asymptotic behavior under stratified and biased-coin randomization of any estimator that is known to be asymptotically linear under simple randomization. The approach is to (a) conjecture that under stratified and biased-coin randomization it is asymptotically linear with the same influence function as under simple randomization; (b) prove this, which may need to be tailored to the estimator, e.g., using techniques as shown in the Supplementary Material for M-estimators and the K-M estimator; (c) apply results from the Supplementary Material (Proposition 1 or Lemma 5) to show that the asymptotic variance is given by (3.4) for scalar/vector parameters or by (3.9) for functions. An area of future research is to apply this approach to the estimators of Lu and Tsiatis (2011); Zhang (2015) that use covariate adjustment to improve precision of the K-M estimator.

Our asymptotic results, just as many asymptotic results under the commonly used superpopulation inference framework for randomized trials, assume that the number of randomization strata is fixed and the number of participants in each stratum goes to infinity. This may be a reasonable approximation when no stratum has a small number of participants. In our data examples, the smallest stratum has 49 participants. An area of future research is to consider cases where some randomization strata have few participants.

In our data analyses of NIDA-CTN-0044, the stratification variable "site" was not available in our data set. It was therefore neither used in the estimators nor in the corresponding variance estimates. The variance formulas (3.4) and (3.9) in this case are asymptotically conservative. This is because the outer expectation in the rightmost terms of these formulas are unchanged or decreased if S is replaced by a coarsening of S (defined as merging several randomization strata together in a preplanned way, in the analysis); this follows from the conditional Jensen's inequality. This result may be useful more generally, e.g., when some strata are so small compared to the sample size that stratum-specific evaluation of the empirical means E_n in (3.5) and the corresponding estimator for (3.9) cannot be reliably done. In such cases a pre-planned, coarsened stratum indicator could be used and the resulting hypothesis test would still control Type I error, asymptotically.

Stratified randomization is related to stratified sampling designs, also called "two-phase sampling" (Sen, 1988; Breslow and Wellner, 2007; Bai et al., 2013). To the best of our knowledge, asymptotic results for these designs do not directly apply to our problem; a key difference is that asymptotic results for stratified sampling designs often involve finite population inference (commonly used in survey sampling), while here we use superpopulation inference (commonly used in analyzing randomized trials).

We provide R functions to calculate the variance for estimators including those in Examples 1-3 and the K-M estimator.

Chapter 4

Semiparametric Partial Common Principal Component Analysis for Covariance Matrices

The content of this chapter is reproduced from Wang et al. (2020) available at `https://doi.org/10.1111/biom.13369`.

4.1 Introduction

Common principal component analysis (CPCA) is an approach that simultaneously models multiple covariance matrices. It extends the idea of principal component analysis by assuming all covariance matrices share the same set of eigenvectors. Since it was first introduced by Flury (1984), CPCA has been extensively applied in various fields including statistics (Gu, 2016; Pepler et al., 2016), finance (Goyal et al., 2008; Xu et al., 2019), and computer science (Ye et al., 2012; Hadjipantelis et al., 2015).

Extensions of CPCA have been investigated from multiple angles. Flury (1987) proposed partial common principal component analysis (PCPCA), where only a proportion of the eigenvectors was assumed to be shared across covariance matrices and the rest to be individual-specific. Another direction relaxed the Gaussianity assumption in CPCA, resulting in asymptotic theory for non-Gaussian distributions (Boik, 2002; Hallin et al., 2010). Other extensions include Bayesian approaches (Hoff, 2009), algorithm acceleration (Browne and McNicholas, 2014) and modifications for high-dimensional data (Franks and Hoff, 2019). Among these extensions of CPCA, PCPCA continues to be appealing, as it relaxes the assumption of a completely common eigenspace across matrices while partially preserving the straightforward interpretation of common eigenvectors, i.e., eigenvectors shared across matrices. Related work on this topic includes Krzanowski (1984), Schott (1999), Boik (2002), Lock et al. (2013) and Pepler et al. (2016).

In spite of these extensions, some questions related to PCPCA remain unanswered. Given the number of common eigenvectors, how one can identify the common eigenvectors from a pool of eigenvectors requires further investigation. Flury (1987) assumed that “some order of the common components is defined”. Some of the literature assumed that the common eigenvectors are those associated with the largest eigenvalues across all covariance matrices (Schott, 1999; Crainiceanu et al., 2011). However, common eigenvectors may be associated with small eigenvalues, or the corresponding eigenvalue of a common eigenvector ranks differently across matrices. This question becomes more challenging if the number of common eigenvectors is unknown or the data is not Gaussian distributed. Regarding these points, Pepler et al. (2016) developed a non-parametric method to select the common eigenvectors in the special case of two covariance matrices. With multiple asymmetric matrices as the response, Lock et al. (2013) proposed a linear model to identify latent factors that explain the joint and individual data variation (JIVE) and Zhou et al. (2016) generalized JIVE to a common and individual feature extraction (CIFE) framework. None of these methods, however, studied the asymptotic properties.

In this paper, we propose a semiparametric PCPCA approach, which can consistently estimate the common eigenvectors, without making any assumptions on the ranks of eigenvalues that are associated with common eigenvectors. Our method builds on an idea from Krzanowski (1984), where a semiparametric approach was proposed in the context of CPCA. We extend this idea to semiparametric PCPCA and provide asymptotic results for our methods as the number of matrices, or the number of samples to estimate each matrix, goes to infinity (both with fixed dimension). If the number of samples goes to infinity, our results do not require the data to be Gaussian distributed. When the number of common eigenvectors is unknown, we develop a sequential testing procedure, which effectively controls the type I error for Gaussian distributed data. As shown in the simulation study, our method outperforms existing methods in estimating the common eigenvectors in a variety of scenarios.

In the next section, we introduce PCPCA. In Section 4.3, we present our proposed semiparametric method to identify the common eigenvectors. We evaluate the performance of our proposed method through simulation studies in Section 4.4. An application to an fMRI data set is provided in Section 4.5. Section 4.6 summarises this paper and discusses future directions.

4.2 Model and assumptions

We consider a data set, $\{\mathbf{y}_{it}\}$, for $t \in \{1, \ldots, T\}$ and $i \in \{1, \ldots, n\}$, where $\mathbf{y}_{it} \in \mathbb{R}^p$ are independent and identically distributed random samples from a p-dimensional distribution with mean zero and covariance matrix Σ_i. In our application example, $\mathbf{y}_{it}$ is a sample of brain fMRI measurements of p regions from subject i at time point t. We assume that Σ_i satisfies the following partial common principal component (PCPC) model:

$$\mathbf{\Sigma}_i = \sum_{j=1}^{k} \lambda_{ij} \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j^\top + \sum_{l=1}^{p-k} \lambda_{i(l+k)} \mathbf{r}_{il} \mathbf{r}_{il}^\top, \tag{4.1}$$

where $\{\lambda_{ij}\}_{j=1}^{p}$ are the eigenvalues of covariance matrix $\mathbf{\Sigma}_i$ and k is the largest integer such that formulation 4.1 holds. The $\boldsymbol{\gamma}_j$, for $j = 1, \ldots, k$, are the unit-length common eigenvectors across subjects. Let $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k) \in \mathbb{R}^{p \times k}$ ($k \leq p$) be the orthonormal matrix of the common eigenvectors. The $\mathbf{r}_{il}$, for $l = 1, \ldots, p-k$, are unit-length individual-specific eigenvectors of subject i. Let $\mathbf{R}_i = (\mathbf{r}_{i1}, \ldots, \mathbf{r}_{i,p-k}) \in \mathbb{R}^{p \times (p-k)}$ be the orthonormal matrix of the individual-specific eigenvectors. We assume that $\mathbf{R}_i$ is orthogonal to $\mathbf{\Gamma}$, i.e., $\mathbf{\Gamma}^\top \mathbf{R}_i = \mathbf{0}$. Let $\mathbf{\Lambda}_i = \mathrm{diag}\{\lambda_{i1}, \ldots, \lambda_{ik}\}$ and $\mathbf{\Psi}_i = \sum_{l=1}^{p-k} \lambda_{i(l+k)} \mathbf{r}_{il} \mathbf{r}_{il}^\top$. Then, the PCPC model (4.1) can be reformulated as:

$$\mathbf{\Sigma}_i = \mathbf{\Gamma} \mathbf{\Lambda}_i \mathbf{\Gamma}^\top + \mathbf{\Psi}_i. \tag{4.2}$$

First proposed by Flury (1987), the PCPC model has an interpretation analogous to CPCA. A CPC, defined by $\boldsymbol{\gamma}_j \boldsymbol{\gamma}_j^\top$ for $j = 1, \ldots, k$, is shared across all matrices. We emphasize that our definition of CPC is different from Flury (1984) (or the principal component in PCA) where a CPC is defined as $\boldsymbol{\gamma}_j^\top \mathbf{y}_{it}$, since we focus on the shared covariance structure across matrices instead of individual-specific eigenvalues. For example, in our application, a CPC represents a functional brain network in the sense that it represents correlations in functional brain measures consistent across subjects. The corresponding diagonal entry of Λ_i is interpreted as the variation of the CPC in subject i. On the other hand, Ψ_i is the individual-specific model component, which varies across subjects. A toy example of the PCPC model with $p = 4$ and $k = 2$ is shown in Figure 4.1.

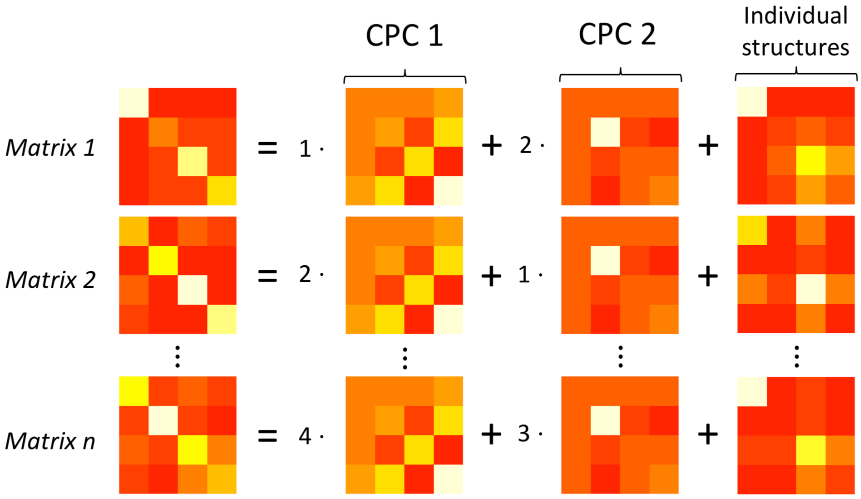

Figure 4.1: An example of the PCPC model. Each covariance matrix consists of two CPCs and an individual structure. Each CPC has rank 1 and norm 1.

Our goal is to both find k and estimate Γ consistently, as either $n \to \infty$ or $T \to \infty$, with p fixed. When estimating Γ, existing methods, such as Schott (1999) and Crainiceanu et al. (2011), assumed that CPCs are associated with the largest eigenvalues across all covariance matrices. This is a restrictive assumption, since the corresponding eigenvalue of a CPC may rank consistently low or differently across matrices. For instance, our toy example in Figure 4.1 shows that CPCs are associated with small eigenvalues in matrices 1 and 2, but with large eigenvalues in matrix n. In addition, in many scientific applications, there is no priori reason to assume that the variation explained by the common components dominates the variation explained by the individual-specific components. For this reason, we do not make any assumptions regarding the rank of CPC-related eigenvalues.

The PCPC model shares some common features with existing partial information decomposition methods, but there exist major differences. Crainiceanu et al. (2011) provided a population value decomposition (PVD) model where common eigenvectors are extracted from concatenated individual eigenvectors. This procedure presumes that common inter-subject components are associated with the largest individual eigenvalues. Moreover, the approach does not consider group level diagonalization as a goal. Lock et al. (2013) introduced the JIVE model, which decomposes information from multiple data sources into common components and individual components. Zhou et al. (2016) generalized the JIVE model by a CIFE framework, which has the same objective function as JIVE. Unlike the PCPC model, the common components identified by JIVE and CIFE are not unique, which can make them hard to interpret in practice. Furthermore, PVD, JIVE and CIFE are empirical methods with no asymptotic guarantees.

More recently, Wang et al. (2019) proposed a common reducing subspace model, which assumes $\mathbf{\Sigma}_i = \mathbf{\Gamma}\mathbf{\Omega}_0\mathbf{\Gamma}^\top + \widetilde{\mathbf{\Gamma}}\mathbf{\Omega}_i\widetilde{\mathbf{\Gamma}}^\top$, where $(\mathbf{\Gamma}, \widetilde{\mathbf{\Gamma}}) \in \mathbb{R}^{p\times p}$ forms an eigenbasis and $\mathbf{\Omega}_0 \in \mathbb{R}^{k\times k}, \mathbf{\Omega}_i \in \mathbb{R}^{(p-k)\times(p-k)}, i = 1, \ldots, n$ are positive definite matrices. This model can be reformulated as $\mathbf{\Sigma}_i = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^\top + \mathbf{\Psi}_i$, where $\mathbf{\Lambda} \in \mathbb{R}^{k\times k}$ is a positive definite diagonal matrix shared across i and $\mathbf{\Psi}_i \in \mathbb{R}^{p\times p}$ is a positive semi-definite matrix orthogonal to $\mathbf{\Gamma}$ with rank $p-k$. Compared with the PCPC model (4.1), this model requires $\mathbf{\Lambda}_i \equiv \mathbf{\Lambda}$, and is hence a special case of the PCPC model.

To achieve the identifiability of $\mathbf{\Gamma}$ and the consistency of our proposed estimator, for the PCPC model (4.1), we impose the following assumptions for the asymptotics when $n \to \infty$.

Assumption A (for $n \to \infty$):

1. T and p are fixed with $T > 0$ and $p > 1$.

2. Each $(\lambda_{i1}, \ldots, \lambda_{ik})$, $i \in \{1, \ldots, n\}$, is an independent and identically distributed random sample from a distribution with finite mean $(\lambda_1^*, \ldots, \lambda_k^*)$ and finite variance. Furthermore, elements of $(\lambda_{i1}, \ldots, \lambda_{ik})$ are independent of each other.

3. Each $\mathbf{\Psi}_i$, $i \in \{1, \ldots, n\}$, is an independent and identically distributed random sample from a distribution with finite mean $\mathbf{\Psi}^*$ and finite second-order moment. Both $\mathbf{\Psi}_i$ and $\mathbf{\Psi}^*$ are symmetric positive semi-definite matrices with rank $p-k$ and are orthogonal to $\mathbf{\Gamma}$.

4. The matrix $\mathbf{\Gamma}\mathbf{\Lambda}^*\mathbf{\Gamma}^\top + \mathbf{\Psi}^*$ has distinct eigenvalues, where $\mathbf{\Lambda}^* = \mathrm{diag}\{\lambda_1^*, \ldots, \lambda_k^*\}$.

5. For each $i \in \{1, \ldots, n\}$, y_{it} is normally distributed given Σ_i.

To the best of our knowledge, we are the first to provide asymptotic results as the number of matrices goes to infinity. Different from the literature where n is fixed (Flury, 1987; Boik, 2002; Pepler et al., 2016; Wang et al., 2019), Assumptions A (2) and (3) assume $(\lambda_{i1}, \ldots, \lambda_{ik})$ and $\boldsymbol{\Psi}_i$ are random variables instead of fixed parameters, since otherwise, the number of parameters would explode as n increases. Assumption A (4) is required for identifiability of $\boldsymbol{\Gamma}$ in matrix perturbation theory. The Gaussian assumption in Assumption A (5) is made for convenience and is stronger than required for our results. For the proof, we only need the fourth-order moment of y_{it} to be the same as the fourth-order moment of a Gaussian distribution, with mean 0 and covariance Σ_i.

In some cases, n is small but T is large. For example, in fMRI data analysis, the number of subjects may be small, but subjects may have long fMRI scans. In other measures with rapid sampling, such as electroencephalograms, this is frequently the case. For such data sets, we prove a similar asymptotic theory as $T \to \infty$ with n and p fixed. This asymptotic theory requires the following assumptions.

Assumption B (for $T \to \infty$):

1. n and p are fixed with $n > 1$ and $p > 1$.

2. For each $i \in \{1, \ldots, n\}$, $(\lambda_{i1}, \ldots, \lambda_{ik})$ and $\boldsymbol{\Psi}_i$ are fixed.

3. The eigenvalues of $\sum_{i=1}^{n} \Sigma_i/n$ are distinct.

4. The fourth-order moment of $\mathbf{y}_{it}$ is bounded for $i = 1, \ldots, n$.

Assumption B (2) implies that the asymptotics are conditional on $\Sigma_i, i = 1, \ldots, n$. The reason to pursue conditional asymptotics is that n is fixed and inference on these specific n distributions is of interest. Unlike Assumption A where a Gaussian distribution is assumed, Assumption B is semiparametric, since it does not put constraints on higher-order moments, except that the fourth-order moment is bounded. Compared with existing asymptotic results for $T \to \infty$, Assumption B is weaker. Flury (1987) and Schott (1999) both assumed that $\{\mathbf{y}_{it}\}$ are normally distributed. Boik (2002) provided asymptotic results for non-normal data, but modeled eigenvalues as known smooth functions of parameters. Though Pepler et al. (2016) and Hallin et al. (2010) relaxed Assumptions B (3) and (4), the former work only focused on the case when $n = 2$ and the latter was for CPCA.

In addition to Assumption A or Assumption B, we also assume that $\{\mathbf{y}_{it}\}, t = 1, \ldots, T$ are independent of each other. However, in many real-world applications, such as our fMRI data example, $\{\mathbf{y}_{it}\}$ can be temporarily correlated. We consider a generic constraint on the temporal correlation that, for $t' < t$, $E[\mathbf{y}_{it}\mathbf{y}_{it'}^\top] = \boldsymbol{D}_{t-t'}$ where $\boldsymbol{D}_{t-t'}$ is a diagonal matrix and $\boldsymbol{D}_{t-t'} = \mathbf{0}$ if $t > t' + c$ for some constant c. This assumed constraint is satisfied for many time series models, including Bickel and Gel (2011) and Guo et al. (2016), and approximately satisfied under the auto-regressive model. Under this assumption, our theoretical results for $n \to \infty$ still hold. Alternatively, when $T \to \infty$, under the same assumption, one can adopt an auto-regressive moving-average (ARMA) model for pre-whitening $\{\mathrm{y}_{it}\}$ to remove temporal dependence, which is commonly used in fMRI data analysis (Lindquist et al., 2008; Olszowy et al., 2019).

4.3 Estimation

In this section, we introduce our estimation procedure under two scenarios: (1) the number of CPCs is known and (2) the number of CPCs is unknown. When the number of CPCs is known, we prove that our proposed estimator of the common eigenvectors is consistent. When the number of CPCs is unknown, the estimation procedure has two steps: we first use a sequential testing procedure to estimate the number of CPCs, and then calculate our proposed estimator using the estimated number of CPCs.

4.3.1 The number of CPCs is known

When the number of CPCs is known, we propose to estimate Γ in two steps: first getting p CPC candidates for Γ, denoted as $\widehat{\Gamma}_{\text{candi}} \in \mathbb{R}^{p \times p}$, and then selecting k columns from $\widehat{\Gamma}_{\text{candi}}$ as $\widehat{\Gamma} \in \mathbb{R}^{p \times k}$.

In the first step, $\widehat{\Gamma}_{\text{candi}}$ is calculated as the eigenvectors of $\overline{\mathrm{S}} = \sum_{i=1}^{n} \mathrm{S}_i / n$, where $\mathrm{S}_i = \sum_{t=1}^{T} \mathbf{y}_{it}\mathbf{y}_{it}^\top/T$ is the sample covariance matrix of subject i. The columns of $\widehat{\mathbf{\Gamma}}_{\text{candi}}$ are ordered in a way that the corresponding eigenvalue of each eigenvector is decreasing. We first define the consistency of an eigenvector estimator.

> **Definition 1.** Let $\{\mathrm{x}_s : s = 1, 2, \dots\}$ denote a series of random vectors in $\mathbb{R}^p$ with ℓ_2-norm 1; that is $\|\mathrm{x}_s\|_2 = 1$ for all s. Let x be a vector in $\mathbb{R}^p$ such that $\|\mathrm{x}\|_2 = 1$. As $s \to \infty$, x_s is consistent to x if $|\langle \mathrm{x}_s, \mathrm{x}\rangle| \xrightarrow{P} 1$, where $\langle \cdot, \cdot \rangle$ is the inner product defined in $\mathbb{R}^p$ and $\xrightarrow{P}$ denotes convergence in probability.

Under Definition 1, the following theorem shows that k out of p columns of $\widehat{\mathbf{\Gamma}}_{\text{candi}}$ are consistent estimators of the columns of $\mathbf{\Gamma}$ as n or T goes to infinity, which is a direct generalization of spectral properties of $\overline{\mathrm{S}}$.

Theorem 4. *Assume the PCPC model* (4.1) *holds.*

1. *Under Assumption A, for any column γ_j of $\mathbf{\Gamma}$ ($j = 1, \dots, k$), there exists a column of $\widehat{\mathbf{\Gamma}}_{\text{candi}}$ that is consistent to γ_j as $n \to \infty$. Explicitly, let $\mathrm{e}_l \in \mathbb{R}^p$ denote a p-dimensional vector with the l-th entry one and rest zero, then, there exists $l(j) \in \{1, \dots, p\}$, such that*

$$|\langle \boldsymbol{\gamma}_j, \widehat{\mathbf{\Gamma}}_{\text{candi}}\mathbf{e}_{l(j)}\rangle| \xrightarrow{P} 1.$$

2. *Under Assumption B, for any column γ_j of $\mathbf{\Gamma}$ ($j = 1, \dots, k$), there exists a*

column of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$ *that is consistent to* γ_j *as* $T \to \infty$.

Theorem 4 implies that, by properly ordering the columns of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$, we can achieve that the j-th column of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$ converges in probability to γ_j for $j = 1, \ldots, k$. To find this ordering, we define a deviation from commonality metric for each column of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$:

$$\operatorname{Dev}(\{\mathbf{y}_{it}\}, \widehat{\boldsymbol{\Gamma}}_{\text{candi}}, j) = \frac{1}{n(p-1)} \sum_{l=1, l\neq j}^{p} \frac{\sum_{i=1}^{n} (\widehat{\gamma}_j^\top \mathbf{S}_i \widehat{\gamma}_l)^2}{(\widehat{\gamma}_j^\top \bar{\mathbf{S}} \widehat{\gamma}_j)(\widehat{\gamma}_l^\top \bar{\mathbf{S}} \widehat{\gamma}_l)}, \tag{4.3}$$

where $\widehat{\gamma}_j$ is the j-th column of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$.

For the deviation from commonality metric, we expect it to be small if $\widehat{\gamma}_j \widehat{\gamma}_j^\top$ is close to a true CPC and large otherwise. For illustration, we assume T is large enough such that the sample estimates can be replaced by their population targets; that is, $\widehat{\gamma}_j \widehat{\gamma}_j^\top = \gamma_{\tilde{j}} \gamma_{\tilde{j}}^\top$ for some $\tilde{j} \in \{1, \ldots, k\}$ and $\mathbf{S}_i = \boldsymbol{\Sigma}_i$. Then the PCPC model (4.1) implies that $\widehat{\gamma}_j^\top \mathbf{S}_i \widehat{\gamma}_l = 0$ for $l \neq j$ and hence $\operatorname{Dev}(\{\mathbf{y}_{it}\}, \widehat{\boldsymbol{\Gamma}}_{\text{candi}}, j) = 0$. If $\widehat{\gamma}_j \widehat{\gamma}_j^\top$ and $\widehat{\gamma}_l \widehat{\gamma}_l^\top$ are not close to any CPC, then $\widehat{\gamma}_j^\top \mathbf{S}_i \widehat{\gamma}_l = \widehat{\gamma}_j^\top \boldsymbol{\Psi}_i \widehat{\gamma}_l \neq 0$ for some i and hence $\operatorname{Dev}(\{\mathbf{y}_{it}\}, \widehat{\boldsymbol{\Gamma}}_{\text{candi}}, j) > 0$. In general, on the right-hand side of Equation (4.3), the numerator captures the sum of the squared (j, l) element in $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}^\top \mathbf{S}_i \widehat{\boldsymbol{\Gamma}}_{\text{candi}}$ and the denominator is a normalizing term that eliminates the effect of magnitude difference in the eigenvalues. The following theorem shows that $\operatorname{Dev}(\{\mathbf{y}_{it}\}, \widehat{\boldsymbol{\Gamma}}_{\text{candi}}, j)$ can be used to order the columns of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$ and estimate $\boldsymbol{\Gamma}$.

Theorem 5. *For all $j \in \{1, \ldots, k\}$, let $\widehat{\gamma}_{j_n}$ be the estimate of γ_j in $\widehat{\mathbf{\Gamma}}_{\text{candi}}$, where $j_n = \arg\max_{l \in \{1,\ldots,p\}} |\langle \widehat{\gamma}_l, \gamma_j \rangle|$. Assume the PCPC model* (4.1) *holds.*

1. *Under Assumption A, as $n \to \infty$, we have*

$$\operatorname{Dev}(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j_n) \xrightarrow{P} \frac{1}{T}.$$

In addition, let $L_n = \{1, \ldots, p\} \setminus \{1_n, \ldots, k_n\}$, then there exists a positive constant C independent of n, such that, as $n \to \infty$,

$$\min_{l \in L_n} \operatorname{Dev}(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{candi}, l) \xrightarrow{P} \frac{1}{T} + C.$$

2. *Under Assumption B, as $T \to \infty$, we have*

$$\operatorname{Dev}(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j_n) \xrightarrow{a.s.} 0 \quad \textit{and} \quad \min_{l \in L_n} \operatorname{Dev}(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, l) \xrightarrow{a.s.} \widetilde{C},$$

where $\widetilde{C}$ is a positive constant independent of T and $\xrightarrow{a.s.}$ denotes convergence almost surely.

Given Theorem 5, in practice, we can rank the columns of $\widehat{\mathbf{\Gamma}}_{\text{candi}}$ in increasing order of $\operatorname{Dev}(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j)$ for $j = 1, \ldots, p$ and select the first k columns as $\widehat{\mathbf{\Gamma}}$. If $\widehat{\gamma}_j$ is selected, we call $\widehat{\gamma}_j$ a common eigenvector estimate and $\widehat{\gamma}_j \widehat{\gamma}_j^\top$ a CPC estimate.

In Theorem 5, the asymptotic results of $n \to \infty$ and $T \to \infty$ are different, which results from the different assumptions made in the two cases. When $n \to \infty$, the deviation from commonality metric is related to the fourth-order moment of $\mathbf{y}_{it}$, which yields a positive probability limit for a CPC estimate. When $T \to \infty$, we have $\widehat{\gamma}_j^\top \mathbf{S}_i \widehat{\gamma}_l \xrightarrow{a.s.} 0$ for $l \neq j$ if and only if $\widehat{\gamma}_j \widehat{\gamma}_j^\top$ is a CPC estimate, making the deviation from commonality metric converge to 0 only for a CPC estimate. Despite these differences, CPC estimates in both cases have the least deviation from commonality metric among all columns of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$ asymptotically, which is essential for identifying CPC estimates from $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$.

When n and T are small, some columns of $\widehat{\boldsymbol{\Gamma}}$ may have a large deviation from commonality metric and are not “close” to any CPC. This bias, however, will disappear as $n \to \infty$ or $T \to \infty$, as guaranteed by Theorems 4 and 5. While our theorems hold for all $k \in \{0, 1, \ldots, p-2, p\}$, the convergence rate can be faster for larger k. Under Assumption A, when k is large, $\{\mathbf{y}_{it}\}$ for different i share more in common, which reduces the variability of the eigenvectors of $\overline{\mathbf{S}}$. Under Assumption B, as k increases, the number of parameters in the PCPC model (1) decreases, and the effective sample size to estimate each parameter increases. We leave the study of the convergence rate as a function of p and k to future research. The estimating procedure is summarized in Algorithm 1.

With the number of CPCs given, we generalize the results of Flury (1987) in three directions. First, Algorithm 1 can consistently estimate CPCs as $n \to \infty$, a case Flury (1987) did not cover. Second, when $T \to \infty$, Theorems 4 and 5

Algorithm 1 An algorithm to estimate CPCs in model (4.1) when k is known.

```
Input: A Data set $\{\mathbf{y}_{it}\}$, $t = 1, \dots, T$, $i = 1, \dots, n$, and $k \in \{1, \dots, p-2, p\}$.
  1. Calculate the sample covariance matrix $\mathbf{S}_i = \sum_{t=1}^{T} \mathbf{y}_{it}\mathbf{y}_{it}^{\top}/T$ for each $i$.
  2. Perform eigendecomposition on $\overline{\mathbf{S}} = \sum_{i=1}^{n} \mathbf{S}_i/n$ and obtain the estimated
     eigenvectors denoted as $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$.
  3. Reorder the columns of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$ such that $\mathrm{Dev}(\{\mathbf{y}_{it}\}, \widehat{\boldsymbol{\Gamma}}_{\text{candi}}, j)$ is increasing
     in $j$, and let $\widehat{\boldsymbol{\Gamma}}$ be the first $k$ columns of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$.
Output: A $p \times k$ orthonormal matrix $\widehat{\boldsymbol{\Gamma}}$.
```

relax the Gaussian assumption made by Flury (1987). Third, Theorems 4 and 5 guarantee the identification of the CPCs without making assumptions on the ranks of CPC-related eigenvalues.

Theorems 1 and 2 allow that $p > T$ when $n \to \infty$. When implementing Algorithm 1, the only condition is that $\overline{\mathbf{S}}$ is positive definite, which is generally true if $p < nT$. However, a large p may substantially increase the computational complexity and affect finite-sample accuracy, as discussed in Sections 3.3 and 4, respectively.

4.3.2 The number of CPCs is unknown

Based on the idea of Schott (1999), we use a sequential hypothesis testing approach to find k. For $j = 0, 1, \dots, p-2$, we sequentially perform the following testings

$$\mathrm{H}_{0,j} : k = j \quad \leftrightarrow \quad \mathrm{H}_{1,j} : k \geq j + 1.$$

Starting from $j = 0$, if $\mathrm{H}_{0,j}$ is rejected, then we proceed to test $\mathrm{H}_{0,j+1}$; otherwise we estimate $\widehat{k} = j$. Before the first test, we order the columns of $\widehat{\mathbf{\Gamma}}_{\text{candi}}$ such that $\mathrm{Dev}\left(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j\right)$ is increasing in j. When testing the j-th hypothesis, we simulate the distribution of $\mathrm{Dev}\left(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j+1\right)$ under $\mathrm{H}_{0,j}$, denoted as $\widehat{F}_{j+1}$, and reject $\mathrm{H}_{0,j}$ if $\mathrm{Dev}\left(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j+1\right)$ is smaller than the α-quantile of $\widehat{F}_{j+1}$. The logic of this rejection rule is that $\mathrm{Dev}\left(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j+1\right)$ is small under $\mathrm{H}_{1,j}$, but the α-quantile of $\widehat{F}_{j+1}$ is generally large, since $\widehat{\gamma}_{j+1}$ is not a common eigenvector under the null hypothesis. Adjusting for multiple testing is unnecessary here, since the family-wise type I error is $\mathbb{P}(\widehat{k} \geq k_0 + 1 | k = k_0) = \alpha$, if the truth is $k = k_0$.

Given $\mathbf{\Gamma}$ and $\{\lambda_{i1}, \ldots, \lambda_{ip}\}_{i=1}^{n}$ defined in the PCPC model (4.1), we calculate $\widehat{F}_{j+1}$ by repeating the following steps for m times. In practice, we can approximate $\mathbf{\Gamma}$ using $\widehat{\mathbf{\Gamma}}$ output in Algorithm 1, estimate $\{\lambda_{i1}, \ldots, \lambda_{ij}\}$ by diagonal entries of $\widehat{\mathbf{\Gamma}}\mathbf{S}_i\widehat{\mathbf{\Gamma}}$ and estimate $\{\lambda_{i(j+1)}, \ldots, \lambda_{ip}\}$ by the non-zero eigenvalues of $\mathbf{S}_i - \widehat{\mathbf{\Gamma}}\,\mathrm{diag}\{\lambda_{i1}, \ldots, \lambda_{ij}\}\widehat{\mathbf{\Gamma}}^\top$. We emphasize that, different from Algorithm 1, where p can be larger than T, the above approximations are valid when $p \leq T$.

1. For each $i = 1, \ldots, n$, independently and uniformly generate $\mathbf{R}_i^{(\text{sim})}$ from the sample space $\{\mathbf{R}_i^{(\text{sim})} \in \mathbb{R}^{p\times(p-j)} : \mathbf{R}_i^{(\text{sim})\top}\mathbf{R}_i^{(\text{sim})} = \boldsymbol{I}_{n-j}, \mathbf{R}_i^{(\text{sim})\top}\mathbf{\Gamma} = \mathbf{0}\}$.

2. Construct $\mathbf{\Sigma}_i^{(\text{sim})} = (\mathbf{\Gamma}, \mathbf{R}_i^{(\text{sim})})\,\mathrm{diag}\{\lambda_{i1}, \ldots, \lambda_{ip}\}(\mathbf{\Gamma}, \mathbf{R}_i^{(\text{sim})})^\top$. Generate $\mathbf{y}_{it}^{(\text{sim})}$, $t = 1, \ldots, T$, from multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}_i^{(\text{sim})}$.

3. Given the data set $\{\mathbf{y}_{it}^{(\text{sim})}\}$, calculate $\widehat{\mathbf{\Gamma}}_{\text{candi}}^{(\text{sim})}$ as described in Algorithm 1 and output $\mathrm{Dev}\left(\{\mathbf{y}_{it}^{(\text{sim})}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}^{(\text{sim})}, j+1\right)$.

Then $\widehat{F}_{j+1} = \sum_{l=1}^{m} \delta_l / m$, where δ_l denotes a point mass at $\mathrm{Dev}\left(\{\mathbf{y}_{it}^{(\text{sim})}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}^{(\text{sim})}, j+1\right)$ output by the l-th simulation. The following theorem shows that the type I error rate for each test is bounded by α under regularity assumptions.

Theorem 6. *Assume the PCPC model (4.1) holds, $\{\mathbf{y}_{it}|\Sigma_i\}$ follows a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance Σ_i, and $\mathbf{R}_i$ follows a uniform distribution on its sample space defined in Section 4.2. Then under $\mathrm{H}_{0,j}$, as $m \to \infty$, $\widehat{F}_{j+1}$ converges in distribution to the true distribution of $\mathrm{Dev}\left(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j+1\right)$ given $\mathbf{\Gamma}$ and $\{\lambda_{i1}, \ldots, \lambda_{ip}\}, i = 1, \ldots, n$.*

A key assumption in Theorem 3 is that data are Gaussian distributed. When this assumption does not hold, the sequential testing procedure tends to be conservative, i.e. $\widehat{k} < k$, since the $\widehat{F}_{j+1}$ is likely to underestimate the mean and deviation of the distribution of $\mathrm{Dev}\left(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j+1\right)$. In practice, an ad hoc solution is to let $\widehat{k}$ be the smallest j such that $E_m[\widehat{F}_j] - \sqrt{Var_m(\widehat{F}_j)}$ is smaller than $\mathrm{Dev}\left(\{\mathbf{y}_{it}\}, \widehat{\mathbf{\Gamma}}_{\text{candi}}, j\right)$, where E_m and Var_m represent sample average and variance respectively. This solution shares the central idea of gap statistics in Tibshirani et al. (2001), which is used to determine the number of clusters in clustering. The procedure of finding k and estimating $\mathbf{\Gamma}$ is described in Algorithm 2.

Another method to find k is to use the hierarchy of partial chi-squared statistics proposed by Flury (1987, 1988). A nice summary of these statistics can be found in Pepler et al. (2016). The relevant application of this hierarchy in PCPCA is testing $k = k_1 \nleftrightarrow k = k_2$. However, this approach has two limitations. First, a set of common eigenvector estimates must be prespecified to implement the test, which is unknown under our setting since CPCs can rank differently among matrices. Second, the chi-squared test is valid only as $T \to \infty$, which is a case where our approach also applies. Hence, we do not consider this method to find k in the simulation studies and data application.

4.3.3 Computational complexity

Given parameters k, m, n, T and p, the computational complexity is $O\{np^2(p+T)\}$ for Algorithm 1 and $O\{4\widehat{k}mnp^2(p+T)\}$ for Algorithm 2, where $\widehat{k}$ is the estimate of k by Algorithm 2. It is straightforward to see that the dimension of matrices p drives the computational complexity at a rate of p^3, if T is not too large. Furthermore, finding k can dramatically increase the run time if k and m are large.

As a benchmark for actual run time, we set $k = p = 20, m = n = T = 100$ and ran both algorithms on an Intel I5-8259U 2.3GHz processor in R software for 10 times. On average, Algorithm 1 took 0.04 seconds and Algorithm 2 took 453.01 seconds, where the difference is the run time due to the iterations for

Algorithm 2 A two-step algorithm to estimate CPCs in model (4.1) when k is unknown.

Input: A Data set $\{\mathbf{y}_{it}\}$, $t = 1, \dots, T$, $i = 1, \dots, n$.

Step 1: Get candidates $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$ for $\boldsymbol{\Gamma}$.

1. Calculate the sample covariance matrix $\mathrm{S}_i = \sum_{t=1}^{T} \mathbf{y}_{it}\mathbf{y}_{it}^{\top}/T$ for each i.
2. Perform eigendecomposition on $\overline{\mathrm{S}} = \sum_{i=1}^{n} \mathrm{S}_i/n$ and obtain the estimated eigenvectors $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$.
3. Reorder the columns of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$ such that $\mathrm{Dev}\left(\{\mathbf{y}_{it}\}, \widehat{\boldsymbol{\Gamma}}_{\text{candi}}, j\right)$ is increasing in j.

Step 2: Identify $\widehat{\boldsymbol{\Gamma}}$ from $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$.

1. Initialize $\widehat{k} = 0$.
2. Test the hypothesis $\mathrm{H}_{0,\widehat{k}} : k = \widehat{k} \leftrightarrow \mathrm{H}_{1,\widehat{k}} : k \geq \widehat{k} + 1$ by a simulation test described in Section 4.3.2 with significance level $\alpha = 0.05$ and 1,000 simulations.
3. Based on the testing result: if $\mathrm{H}_{0,\widehat{k}}$ is not rejected, return $\widehat{k}$ and $\widehat{\boldsymbol{\Gamma}}$ as the first $\widehat{k}$ columns of $\widehat{\boldsymbol{\Gamma}}_{\text{candi}}$; if $\widehat{k} = p - 2$ and $\mathrm{H}_{0,\widehat{k}}$ is rejected, return $\widehat{k} = p$ and $\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}}_{\text{candi}}$; otherwise, increase $\widehat{k}$ by 1 and repeat **Step 2** (2).

Output: $\widehat{k} \in \{0, 1, \dots, p-2, p\}$ and a $p \times \widehat{k}$ orthonormal matrix $\widehat{\boldsymbol{\Gamma}}$.

estimating k. In comparison, Flury's algorithm (Flury and Gautschi, 1986) for CPCA, which assumes k is known, took 5.23 seconds under the same setting, which is roughly 100 times slower than Algorithm 1. In practice, one could reduce the run time of Algorithm 2 by parallel programming and improving code efficiency.

4.4 Simulation study

In this section, we perform three simulation studies. The first confirms the asymptotic results given by Theorem 5. The second tests the performance of Algorithms 1 and 2 under various settings. The last compares our proposed method with existing approaches under different scenarios.

4.4.1 Design and data generating mechanism

In the first simulation, we let $p = 20$ and $k = 10$. Define $\lambda_j = e^{0.5(p-j)}$ for $j = 1, \ldots, p$, and assume that $\{\mathrm{y}_{it}\}$ follows a multivariate Gaussian distribution and CPCs rank randomly in each covariance matrix. For the study of the asymptotics as $n \to \infty$, we set $T = 50$ and $n = 50, 100, 500, 1000$; and for the study of the asymptotics as $T \to \infty$, we set $n = 50$ and $T = 50, 100, 500, 1000$. For each combination of n and T, we simulate data and compare the distribution of the k-th smallest deviation from commonality metric, which is the largest metric of CPC estimates, with the distribution of the $(k+1)$-th smallest deviation from commonality metric, which is the smallest metric of non-CPC estimates.

The second simulation is the same as the first one, except that we consider combinations of different settings: (1) $n = T = 15, 30, 100$, (2) $k = 1, 10, 20$ and (3) $\{\mathbf{y}_{it}\}$ follows a multivariate Gaussian distribution versus Gamma distribution. For each combination, we simulate data for 1000 times, run Algorithm 1 to get $\widehat{\mathbf{\Gamma}}$, and run Algorithm 2 to get $\widehat{k}$ for each simulated data set. To measure the performance of Algorithm 1, we define $\sum_{j=1}^{k} \max |\boldsymbol{\gamma}_j^\top \widehat{\mathbf{\Gamma}}|/k$ as the accuracy metric of $\widehat{\mathbf{\Gamma}}$. This metric lies in $[0, 1]$ with larger values indicating better accuracy. To evaluate the sequential testing procedure, we report $\widehat{k}$ and compare it with the true k.

The last simulation compares our proposed method (with or without k known, i.e., Algorithm 1 or Algorithm 2) with Flury's method (Flury, 1987) and the PVD method (Crainiceanu et al., 2011) through 4 scenarios below. By "Flury's method", we mean first running the algorithm given by Flury and Gautschi (1986) to estimate $\widehat{\mathbf{\Gamma}}_{candi}$ in CPCA and then selecting k columns associated with the largest eigenvalues of $\overline{\mathbf{S}}$. Although Flury (1987, 1988) proposed a method to estimate k in PCPCA, we do not implement it here, because the order of CPCs is unknown, as discussed in Section 4.3.2. For the PVD, we use the default setting; that is, first calculating the top k eigenvectors of $\mathbf{S}_i$ (denoted as $\mathbf{U}_i$) and then estimating $\mathbf{\Gamma}$ as the top k eigenvectors of $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_n)$. There are other partial information decomposition methods, such as JIVE and CIFE described in Section 4.2, but they do not have unique CPC estimates, which makes the comparison with these methods via simulation infeasible.

Scenario 1: $\{\mathbf{y}_{it}\}$ follows a Gaussian distribution with large n and T. CPCs are associated with the largest eigenvalues in each covariance matrix.

Scenario 2: $\{\mathbf{y}_{it}\}$ follows a Gaussian distribution with large n and T. The CPC-associated eigenvalues rank randomly in each covariance matrix.

Scenario 3: $\{\mathbf{y}_{it}\}$ follows a Gamma distribution with large n and T. CPCs are associated with the largest eigenvalues in each covariance matrix.

Scenario 4: $\{\mathbf{y}_{it}\}$ follows a Gaussian distribution with small n and T. CPCs are associated with the largest eigenvalues in each covariance matrix.

Scenario 1 serves as the reference case, where the underlying assumptions of all 4 methods are satisfied. Different from Scenario 1, Scenario 2 has randomly ranked CPC-associated eigenvalues, Scenario 3 has Gamma data generating distribution and Scenario 4 has small sample size. For each of the 4 scenarios, we consider two cases: $p = 20$ with $k = 10$ and $p = 100$ with $k = 20$, which represent small-scale and large-scale problem, respectively. When $p = 20$, we set $n = T = 100$ for Scenarios 1-3 and $n = T = 30$ for Scenario 4 and define $\lambda_j = e^{0.5(p-j)}$ for $j = 1, \ldots, p$. When $p = 100$, we set $n = T = 1000$ for Scenarios 1-3 and $n = T = 150$ for Scenario 4 and define $\lambda_j = e^{0.1(p-j)}$ for $j = 1, \ldots, p$. Similar to simulation 2, we use $\sum_{j=1}^{k} \max |\boldsymbol{\gamma}_j^\top \widehat{\boldsymbol{\Gamma}}|/k$ as the accuracy metric of $\widehat{\boldsymbol{\Gamma}}$.

For all simulations, if $\{\mathbf{y}_{it}\}$ follows a Gaussian distribution and CPC-associated eigenvalues rank randomly, we simulate the data as follows for 1000 replications. Given p, n, T, k and $\{\lambda_j\}_{j=1}^{p}$, we sample one $\mathbf{\Gamma}$ from the space $\{\mathbf{\Gamma} : \mathbf{\Gamma}^\top\mathbf{\Gamma} = \boldsymbol{I}_k\}$ as the common eigenvectors, and randomly partition $\{\lambda_j\}_{j=1}^{p}$ into two parts: one with k elements as the eigenvalues corresponding to common eigenvectors (denoted as $\{\lambda_j^*\}_{j=1}^{k}$) and the other one consisting of $p-k$ elements (denoted as $\{\lambda_j^*\}_{j=k+1}^{p}$). For $i = 1, \ldots, n$, we independently sample λ_{ij} from a chi-squared distribution with degrees of freedom λ_j^* and construct $\mathbf{\Psi}_i = \mathbf{U}_i\mathbf{D}_i\mathbf{U}_i^\top$, where $\mathbf{U}_i$ is an independent sample from the space $\{\mathbf{U} : \mathbf{U}^\top\mathbf{U} = \boldsymbol{I}_{n-k}, \mathbf{U}^\top\mathbf{\Gamma} = \mathbf{0}_{(n-k)\times k}\}$ and $\mathbf{D}_i = \operatorname{diag}\{\lambda_{i(k+1)}, \ldots, \lambda_{ip}\}$. Then we construct $\mathbf{\Sigma}_i = \mathbf{\Gamma}\operatorname{diag}\{\lambda_{i1}, \ldots, \lambda_{ik}\}\mathbf{\Gamma}^\top + \mathbf{\Psi}_i$ and $\{\mathbf{y}_{it}, t = 1, \ldots, T\}$ are independently sampled from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i)$. If $\{y_{it}\}$ are not Gaussian distributed, we modify the above procedure by letting $\mathbf{y}_{it} = \mathbf{\Sigma}_i^{-\frac{1}{2}}\tilde{\mathbf{y}}_{it}$, where $\{\tilde{\mathbf{y}}_{it}, t = 1, \ldots, T\}$ are independently sampled from a multivariate-Gamma distribution with mean $\mathbf{0}$, variance $\boldsymbol{I}_p$ and skewness $10\boldsymbol{I}_p$. If CPC-associated eigenvalues are the largest k eigenvalues across matrices, we set $\{\lambda_j^*\}_{j=1}^{k}$ to be the largest k numbers in $\{\lambda_j\}_{j=1}^{p}$.

4.4.2 Simulation results

Simulation results are summarized in Figure 4.2 and Tables 4.1 and 4.2 for simulations 1, 2 and 3 respectively.

Figure 4.2 shows that the deviation from commonality metric converges to its limit when T is fixed and $n \to \infty$, and when n is fixed and $T \to \infty$. This confirms the results of Theorem 5 and indicates that this metric can be used to distinguish CPC estimates and non-CPC estimates when either n or T is large.

Table 4.1 displays the performance of Algorithms 1 and 2 under the different simulation settings. When data are Gaussian distributed, both algorithms have high accuracy whenever k, n, T are small, medium or large. As n and T increases, the performance of both algorithms improves. When the sample size is small, Algorithm 1 still has a high accuracy in estimating $\boldsymbol{\Gamma}$, even under the non-Gaussian distribution setting. In particular, when $n = T = 15 < p$, Algorithm 1 remains valid and has good accuracy, which demonstrates an advantage with small data. Since implementing Algorithm 2 requires $p \leq T$, $\widehat{k}$ is not estimated when $n = T = 15$. As k increases, the accuracy of Algorithm 1 slightly increases, which confirms our discussion in Section 4.3.1. Under the Gamma data generating distribution, the algorithm to find k likely underestimates k when k is large. The reason for this is twofold. First, this algorithm is conservative for non-Gaussian data (as discussed in Section 4.3.2); second, when k is large, the number of null hypotheses to reject is large, which reduces the overall power. As a result, we recommend using Algorithm 2 for Gaussian distributed data. If k is large, one may not find all CPCs, but the identified ones are accurate.

Table 4.2 gives the comparison of our proposed method with k known or unknown to Flury's method and PVD. In the first scenario, all methods perform well, as expected. In the other scenarios, our proposed method performs as good as or better than Flury's method and PVD, even when the true number of CPCs is unknown. In Scenario 2, since both Flury's method and PVD assume CPCs are associated with the largest eigenvalues for each matrix, their accuracy is much lower than our proposed method. In Scenario 3, all four methods have modest accuracy, but our proposed method with k unknown, Flury's method and PVD have lower accuracy due to the non-Gaussian distribution. In contrast, our proposed method with k known remains highly accurate, since it is semiparametric. In Scenario 4, the size of data is limited compared to the dimension of matrices, resulting in small accuracy drops of all methods. However, our proposed method still has the least accuracy drop among all methods. In all scenarios, our proposed method outperforms Flury's method and PVD, even when the true number of CPCs is unknown.

Table 4.1: The accuracy of Algorithm 1 and the sequential testing procedure under different settings with $p = 20$.

		Average accuracy of $\widehat{\boldsymbol{\Gamma}}$			Average $\widehat{k}$		
	Distribution	$k=1$	$k=10$	$k=20$	$k=1$	$k=10$	$k=20$
$n=T=15$	Gaussian	0.81	0.91	0.93	-	-	-
	Gamma	0.67	0.71	0.83	-	-	-
$n=T=30$	Gaussian	0.95	0.97	0.95	1.20	9.68	18.92
	Gamma	0.70	0.86	0.92	0.65	4.04	4.02
$n=T=100$	Gaussian	0.98	0.99	0.95	1.26	10.02	20.00
	Gamma	0.96	0.98	0.95	1.01	9.50	13.73

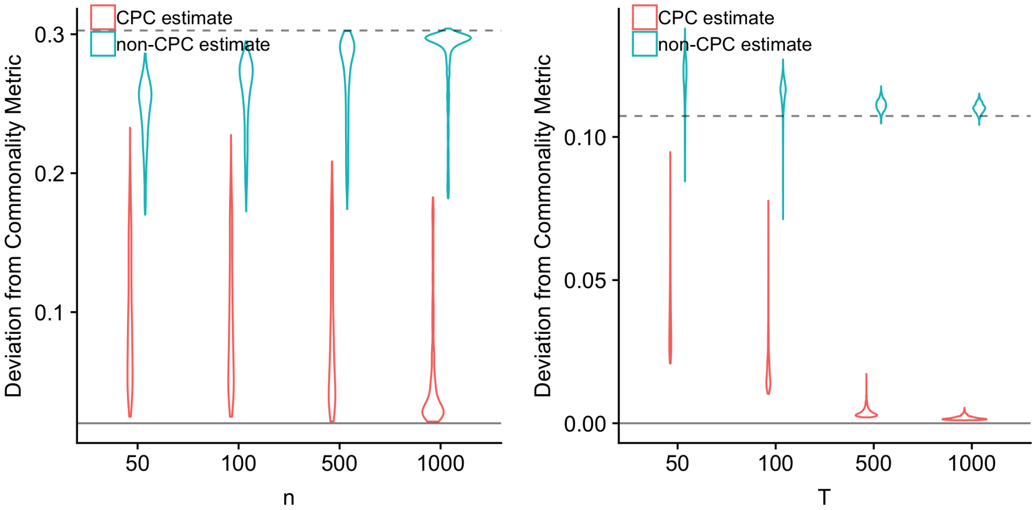

Figure 4.2: Distribution of the "Deviation from commonality" metric (4.3) as n (left panel) or T (right panel) goes to infinity for the last CPC estimate and the first non-CPC estimate. The solid line is the probability limit for the CPC estimate and the dashed line is the probability limit for the non-CPC estimate calculated from Theorem 5. The left panel demonstrates that the metric converges in probability to its limit, while the right panel shows that the metric converges almost surely.

Table 4.2: The accuracy of methods in estimating CPC under different scenarios. Semi-1: the proposed semiparametric method with k known. Semi-2: the proposed semiparametric method with k unknown. Flury: the Flury's method. PVD: population value decomposition.

		Semi-1	Semi-2	Flury	PVD
Scenario 1	$p = 20$	1.00	1.00	1.00	0.99
	$p = 100$	1.00	1.00	1.00	0.99
Scenario 2	$p = 20$	0.99	0.96	0.26	0.50
	$p = 100$	0.99	0.93	0.24	0.20
Scenario 3	$p = 20$	0.98	0.95	0.88	0.94
	$p = 100$	1.00	0.95	0.91	0.95
Scenario 4	$p = 20$	0.99	0.99	0.99	0.95
	$p = 100$	0.99	0.99	0.97	0.96

4.5 Task fMRI data example

We apply the proposed semiparametric PCPC method to the Human Connectome Project (HCP) motor-task fMRI data. The HCP project studies the brain connectome, both structural and functional, of healthy adults. The data set includes $n = 136$ healthy young adults from the most recent S1200 release. Adapted from the experimental design in Buckner et al. (2011) and Thomas Yeo et al. (2011), the task fMRI consists of ten task blocks including two tongue movement blocks, four hand movement blocks (two left and two right) and four foot movement blocks (two left and two right), as well as three 15-second fixation blocks. In each movement task block, a three-second visual cue was first presented followed by a 12-second movement. Participants were instructed to follow the visual cue to either move their tongue, or tap their left/right fingers, or squeeze their left/right toes to map the corresponding motor areas. The tasks were randomly intermixed. Once the ordering was fixed, the task onsets are nearly consistent across participants. We used the fMRI data collected and minimally pre-processed by the Washington University-University of Minnesota HCP Consortium (Van Essen et al., 2012). The HCP pre-processing steps include distortion correction, image alignment, volume segmentation, Montreal Neurological Institute (MNI) space registration, and creating various masks and maps for analysis, and these steps were described in detail in Glasser et al. (2013). Given the pre-processed data, we extracted the time courses ($T = 284$ time points and repetition time $= 0.72$ seconds) from 264 brain regions (Power et al., 2011), which are averaged signals over voxels within the 5 mm radius ball from the region center. Motion correction was conducted to remove the effect due to head movement during the scan by regressing on the 12 motion parameters and taking the residuals as the motion-corrected data (Lindquist et al., 2008). Furthermore, we fit an ARMA(1,1) model (Lindquist et al., 2008) for each brain region to remove temporal correlation. A figure showing the average of the inter-subject and region estimated autocorrelation function before and after removing temporal correlation is given in the Supporting Information. This plot suggests that autocorrealtions were mitigated by ARMA(1,1) filtering. Since Flury's method and PVD are not able to identify the CPCs associated with small eigenvalues and require prespecified k, we present the result from the proposed semiparametric method only. Results of Flury's method and PVD letting $k = p$ are given in the Supporting Information, which differ from the results of the semiparametric method.

We first focused on functional brain regions in the sensorimotor network (Power et al., 2011) with $p = 35$, which is directly related to the task design. According to the Doornik-Hansen test for multivariate normality by Doornik and Hansen (2008), there is no sufficient evidence to reject the null hypothesis that data are normally distributed (p-value 0.12). Among 35 CPC candidates, Algorithm 2 identified 30 as CPCs, which explain 80% of the total variance of the average covariance matrix. Figure 1 in the Supporting Information summarises the results of sequential testings. To explore the relationship between the identified CPCs and the motor tasks, we plotted the average time course of each CPC estimate (i.e., $\sum_{i=1}^{n}\gamma_j^\top \mathbf{y}_{it}/n$ for $j=1,\ldots,30$) and compared it with task time bins. We also visualized brain regions with loading magnitude greater than 0.15 in a brain map. As a result, at least ten of the identified CPCs are related to tasks (no statistical test is performed) and a list of identified brain networks is provided in Table 4.3. Figure 4.3(A) presents an example of task-related CPC (CPC 18). In Figure 4.3(A), the average time course suggests a brain network of right hand movement and left foot movement, which is confirmed by the brain map. In this component, brain areas associated with motor control of the right hand yield high negative loadings (blue regions on the left hemisphere of the brain in Figure 4.3(A)); and regions associated with motor control of the left foot yield high positive loadings (red regions on the right hemisphere of the brain in Figure 4.3(A)). The lateral separation of the brain in terms of the loading sign suggests that during these motor tasks, the associated left and right hemispheres are functionally negatively correlated. Figure 4.3(B) presents an example of the CPC that is not related to the tasks. Even though the time course does not show a clear pattern, this CPC is concentrated in a region of the brain, which is modularized as a tongue region by Power et al. (2011). For the five components that are not identified as CPCs, two appear task-related and three do not. Since Algorithm 2 can be conservative when the true number of CPCs is large and the distribution is non-Gaussian based on simulation, some of these five CPC candidates may be CPCs. For all 35 CPC candidates, the average time course and the brain maps are provided in the Supporting Information.

Besides the analysis of the sensorimotor network, we ran Algorithm 2 on all brain regions ($p = 264$) and the Doornik-Hansen test shows that it is marginally significant for the null hypothesis that the whole-brain data are normally distributed (p-value 0.47). 190 CPC estimates were identified, which explain 50% of the total variance of the average covariance matrix. Among the 190 CPC estimates, 66 are associated with the default mode network, 15 are associated with the visual network, 12 are associated with the sensorimotor network and, 5 are associated with the frontoparietal network. Here we classify a CPC estimate as associated with a brain network if, among the regions with loadings greater than 0.1 in this CPC estimate, at least 25% come from the corresponding network. To compare the results from the sensorimotor-network analysis and whole-brain analysis, we extracted loadings corresponding to the sensorimotor network for each common eigenvector estimate in the whole-brain analysis. Among 30 CPC estimates of sensorimotor-network analysis, 12 are highly correlated (absolute value of inner product larger than 0.7) with some CPC candidates of whole-brain analysis, suggesting that the brain networks encoded by these CPCs retain when taking into account regions outside of the sensorimotor network.

Table 4.3: Task-related brain networks identified by Algorithm 2.

CPC No.	Variance explained	Brain network
3	2.0%	Hands, feet
4	2.2%	Feet
5	2.1%	Right hand
10	2.3%	Right foot
11	2.7%	Feet
15	2.0%	Hands
16	2.2%	Left hand
18	2.1%	Right hand, left foot
23	1.8%	Tongue, feet
24	2.1%	Hands, feet

4.6 Discussion

In this paper, we propose a semiparametric PCPC model and provide algorithms to identify CPCs with or without knowing the true number of CPCs. Furthermore, we prove the asymptotic consistency of our proposed estimators, even when the data generating distribution is non-Gaussian. In simulation studies, our estimator consistently outperforms Flury's method and PVD and shows high accuracy if the number of CPCs is known. Applied to fMRI data, our method identifies meaningful brain networks that match the current find-

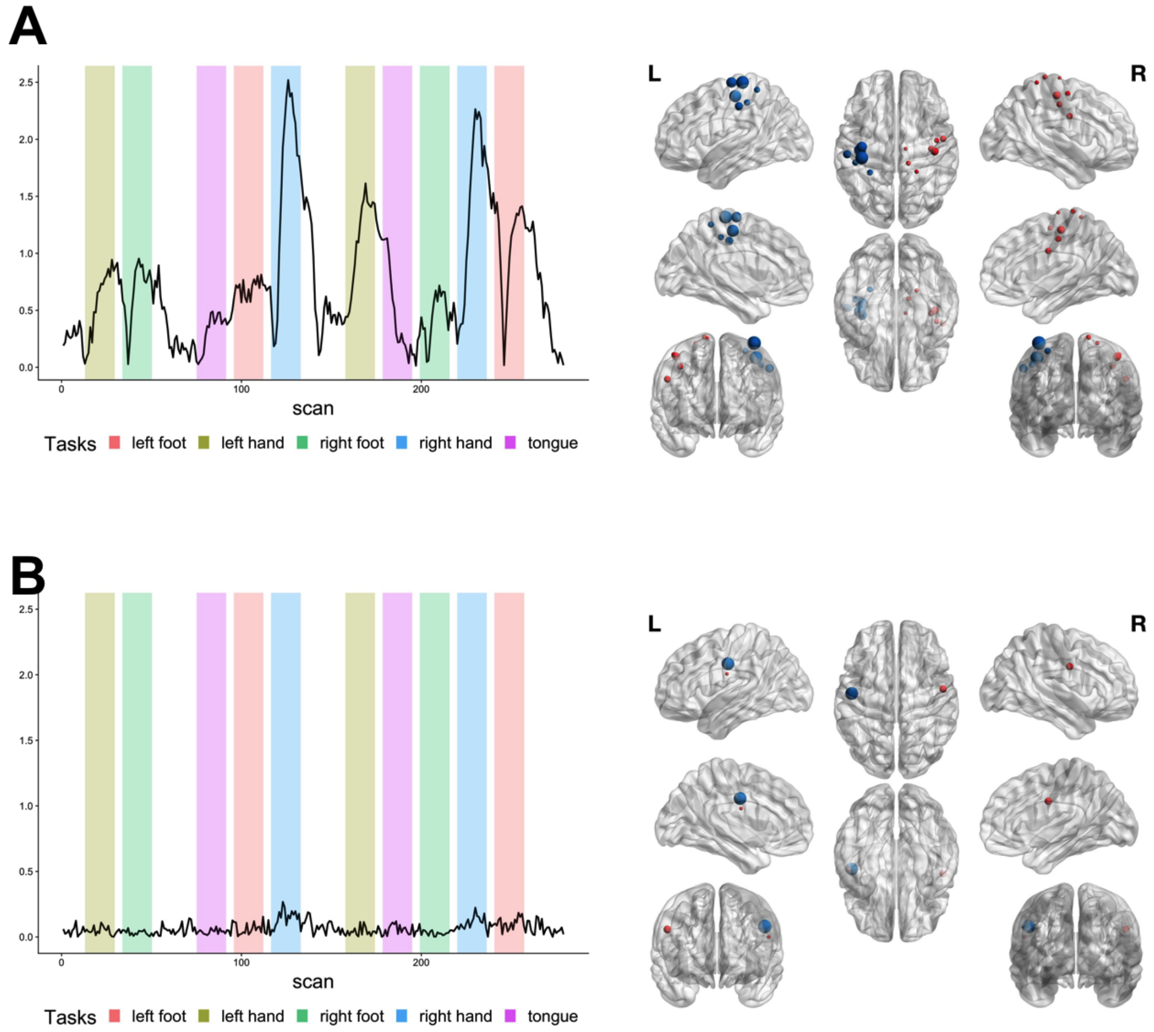

Figure 4.3: Average time course (left panel) and brain regions (right panel) of CPC 18 (upper panel) and CPC 9 (lower panel). In the left panel, each bin represents the time period of a task. In the right panel, each node is a brain region, with size standing for the absolute loading and color representing the sign of the loading (blue for negative and red for positive). Brain regions with absolute loading smaller than 0.1 are not shown in the figure.

ings.

In PCPCA, a CPC may not be associated with the largest eigenvalues across all covariance matrices. For this reason, our proposed method allows for an arbitrary association between CPC and eigenvalues, which makes the model more flexible. One challenge resulting from this flexibility is to find k, the number of CPCs, since the signal of CPCs can be weak or inseparable from non-common principal components. Our proposed algorithm for finding k performs well under Gaussian distribution, but can be conservative if the underlying distribution is non-Gaussian or k is large. Furthermore, sequential hypothesis testing usually requires huge computational resources and can be slow for high-dimensional matrices. Hence, an efficient and robust method for finding k will be one future direction.

Our proposed method, as well as the literature, assumes p, the dimension of covariance matrices, is fixed. One exciting future direction could be finding solutions to handle data with large p but small n and T.

Chapter 5

Discussion

For randomized clinical trials using stratified or biased-coin randomization and having time-to-event outcomes, an open question is how to derive the asymptotic distribution of covariate-adjusted estimators, such as estimators of Zhang (2015) and Lu and Tsiatis (2011) for survival functions. Our conjecture is that results of Chapter 3 can be generalized to handle this question and stratified randomization will lead to a variance reduction with the same formula (3.9) as the K-M estimator.

When identifying shared eigenvectors, our proposed estimator in Chapter 4 assumes that each observed data vector is independent of each other. In brain imaging, however, there is auto-correlation among different brain scans of the same subject. In our data example presented in Section 4.5, we use a time series model to remove such correlation, but how to best handle dependency of data vectors remain future directions.

Appendix A

Supporting information to Chapter 2

Supporting Information, available at `https://doi.org/10.1111/biom.13062`, includes the following: (a) definitions of the sample variance and covariance used in Section 2.4.2; (b) simulations to generate visualizations of the relationship among the imbalance, unadjusted estimator, and ANCOVA estimator; (c) proofs of theoretical results; (d) relationship among different types of R-squared; (e) data analyses accounting for missing data; (f) link to the code for data analysis; and (g) information monitoring with covariate adjustment.

Appendix B

Supplementary Material to Chapter 3

Supplementary Material, available at https://arxiv.org/abs/1910.13954, includes the following: (a) regularity conditions to Theorem 2; (b) consistent estimators of the asymptotic variance V defined in Theorem 2; (c) asymptotic variance $V^{(a)}(t,t')$, which is described in Theorem 3, and a consistent estimator for $V^{(a)}(t,t)$; (d) proofs of Theorem 1, Corollary 1 and Theorem 2 and (e) the results of our data analysis for the K-M estimator for the control group of NIDA-CTN-0044.

Appendix C

Supporting information to Chapter 4

Proofs of Theorems 4, 5, 6, and additional results of data application referenced in Section 3.7 are available with this paper at the Biometrics website on Wiley Online Library `https://doi.org/10.1111/biom.13369`. The R code and data to reproduce the simulations and data application are available at `https://github.com/BingkaiWang/Semi-parametric-PCPCA`.

Bibliography

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (2012). *Statistical Models Based on Counting Processes*. Springer Science & Business Media.

Austin, P., A. Manca, M. Zwarenstein, D. Juurlink, and M. Stanbrook (2010). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology 63*(2), 142 – 153.

Bai, X., A. A. Tsiatis, and S. M. O'Brien (2013). Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics 69*(4), 830–839.

Bickel, P. J. and Y. R. Gel (2011). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(5), 711–728.

Bloniarz, A., H. Liu, C. Zhang, J. Sekhon, and B. Yu (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proc Natl Acad Sci USA 113*(27), 7383–7390.

Boik, R. J. (2002). Spectral models for covariance matrices. *Biometrika 89*(1), 159–182.

Borm, G., J. Fransen, and W. Lemmens (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. *J Clin Epidemiol 60*(12), 1234 – 1238.

Breslow, N. E. and J. A. Wellner (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics 34*(1), 86–102.

Browne, R. P. and P. D. McNicholas (2014). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification 8*(2), 217–226.

Buckner, R. L., F. M. Krienen, A. Castellanos, J. C. Diaz, and B. T. T. Yeo (2011). The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology 106*(5), 2322–2345.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association 113*(524), 1784–1796.

Campbell, A. N., E. V. Nunes, A. G. Matthews, M. Stitzer, G. M. Miele, D. Polsky, E. Turrigiano, S. Walters, E. A. McClure, T. L. Kyle, A. Wahle, P. Van Veldhuisen, B. Goldman, D. Babcock, P. Q. Stabile, T. Winhusen, and U. E. Ghitza (2014). Internet-delivered treatment for substance abuse: A multisite randomized controlled trial. *American Journal of Psychiatry 171*(6), 683–690.

Crainiceanu, C. M., B. S. Caffo, S. Luo, V. M. Zipunnikov, and N. M. Punjabi (2011). Population value decomposition, a framework for the analysis of image populations. *Journal of the American Statistical Association 106*(495), 775–790.

Díaz, I., E. Colantuoni, , D. F. Hanley, and M. Rosenblum (2018). Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime Data Analysis. https://doi.org/10.1007/s10985-018-9428-5.*

Doornik, J. A. and H. Hansen (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics 70*(s1), 927–939.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika 58*(3), 403–417.

EMA (2015). European Medicines Agency Guideline on Adjustment for Baseline Covariates in Clinical Trials. Reference number EMA/CHMP/295050/2013. Committee for Medicinal Products for Human Use (CHMP).

EMA (2019). Missing data in confirmatory clinical trials. Revision 1 - Adopted guideline. CPMP/EWP/1776/99 Rev. 1.

FDA (2019). Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biologics with Continuous Outcomes. Draft Guidance for Industry. https://www.fda.gov/media/123801/download.

FDA (2020). COVID-19: Developing Drugs and Biological Products for Treatment or Prevention. Guidance for Industry. https://www.fda.gov/media/137926/download.

FDA and EMA (1998). E9 statistical principles for clinical trials. *U.S. Food and Drug Administration: CDER/CBER. European Medicines Agency: CPMP/ICH/363/96*.

Fleming, T. R. and D. P. Harrington (2011). *Counting Processes and Survival Analysis*, Volume 169. John Wiley & Sons.

Flury, B. (1988). *Common principal components and related multivariate models*. John Wiley & Sons, Inc.

Flury, B. and W. Gautschi (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing 7*(1), 169–184.

Flury, B. K. (1987). Two generalizations of the common principal component model. *Biometrika 74*(1), 59–69.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association 79*(388), 892–898.

Franks, A. M. and P. Hoff (2019). Shared subspace models for multi-group covariance estimation. *Journal of Machine Learning Research 20*(171), 1–37.

Glasser, M. F., S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage 80*, 105 – 124.

Goyal, A., C. Pérignon, and C. Villa (2008). How common are common return factors across the NYSE and Nasdaq? *Journal of Financial Economics 90*(3), 252 – 271.

Gu, F. (2016). Analysis of correlation matrices using scale-invariant common principal component models and a hierarchy of relationships between correlation matrices. *Structural Equation Modeling: A Multidisciplinary Journal 23*(6), 819–826.

Guo, S., Y. Wang, and Q. Yao (2016). High-dimensional and banded vector autoregressions. *Biometrika 103*(4), 889–903.

Hadjipantelis, P. Z., J. A. D. Aston, H. G. Müller, and J. P. Evans (2015). Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of Mandarin Chinese. *Journal of the American Statistical Association 110*(510), 545–559.

Hallin, M., D. Paindaveine, and T. Verdebout (2010). Optimal rank-based testing for principal components. *The Annals of Statistics 38*(6), 3245–3299.

Hoff, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(5), 971–992.

Howard, G., J. Waller, J. Voeks, V. Howard, E. Jauch, K. Lees, F. Nichols, V. Rahlfs, and D. Hess (2012). A simple, assumption-free, and clinically interpretable approach for analysis of modified Rankin outcomes. *Stroke 43*(3), 664–669.

Huitema, B. (2011). *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, 2nd Edition*. WILEY.

Jarskog, L., R. Hamer, D. Catellier, D. Stewart, L. LaVange, N. Ray, L. Golden,

J. Lieberman, and T. Stroup (2013). Metformin for weight loss and metabolic control in overweight outpatients with schizophrenia and schizoaffective disorder. *American Journal of Psychiatry 170*(9), 1032–1040.

Jiang, F., L. Tian, H. Fu, T. Hasegawa, and L. J. Wei (2018). Robust alternatives to ANCOVA for estimating the treatment effect via a randomized comparative study. *Journal of the American Statistical Association 0*, 1–37.

Kahan, B. C. and T. P. Morris (2012). Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine 31*(4), 328–340.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer.

Kraemer, H. (2015). A source of false findings in published research studies: Adjusting for covariates. *JAMA Psychiatry 72*(10), 961–962.

Krzanowski, W. J. (1984). Principal component analysis in the presence of group structure. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 33*(2), 164–168.

Lachin, J., J. Matts, and L. Wei (1988). Randomization in clinical trials: Conclusions and recommendations. *Controlled Clinical Trials 9*(4), 365 – 374.

Li, X. and P. Ding (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(1), 241–268.

Lin, Y., M. Zhu, and Z. Su (2015). The pursuit of balance: An overview of covariate-adaptive randomization techniques in clinical trials. *Contemporary Clinical Trials 45*, 21 – 25. 10th Anniversary Special Issue.

Lindquist, M. A. et al. (2008). The statistical analysis of fMRI data. *Statistical Science 23*(4), 439–464.

Ling, W., M. Hillhouse, C. Domier, G. Doraimani, J. Hunter, C. Thomas, J. Jenkins, A. Hasson, J. Annon, A. Saxon, J. Selzer, J. Boverman, and R. Bilangi (2009). Buprenorphine tapering schedule and illicit opioid use. *Addiction 104*(2), 256–265.

Liu, H. and Y. Yang (2020). Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*.

Lock, E. F., K. A. Hoadley, J. S. Marron, and A. B. Nobel (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics 7*(1), 523–542.

Lu, X. and A. A. Tsiatis (2011). Semiparametric estimation of treatment effect with time-lagged response in the presence of informative censoring. *Lifetime Data Analysis 17*(4), 566–593.

Ludvigsson, J., M. Faresjö, M. Hjorth, S. Axelsson, M. Chŕamy, M. Pihl, O. Vaarala, G. Forsander, S. Ivarsson, C. Johansson, A. Lindh, N. Nilsson, J. Åman, E. Örtqvist, P. Zerhouni, and R. Casas (2008). GAD treatment and insulin secretion in recent-onset type 1 diabetes. *New England Journal of Medicine 359*(18), 1909–1920.

Ma, W., F. Hu, and L. Zhang (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association 110*(510), 669–680.

Ma, W., Y. Qin, Y. Li, and F. Hu (2018). Statistical inference of covariate-adjusted randomized experiments. *arXiv https://arxiv.org/abs/1807.09678*.

Mallinckrodt, C. H., W. S. Clark, R. J. Carroll, and G. Molenberghs (2003). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics 13*(2), 179–190. PMID: 12729388.

Montalban, X., S. Hauser, L. Kappos, D. Arnold, A. Bar-Or, G. Comi, J. de Seze, G. Giovannoni, H. Hartung, B. Hemmer, F. Lublin, K. Rammohan, K. Selmaj,

A. Traboulsee, A. Sauter, D. Masterman, P. Fontoura, S. Belachew, H. Garren, N. Mairon, P. Chin, and J. Wolinsky (2017). Ocrelizumab versus placebo in primary progressive multiple sclerosis. *N Engl J Med 376*(3), 209–220.

Moore, K., R. Neugebauer, T. Valappil, and M. van der Laan (2011). Robust extraction of covariate information to improve estimation efficiency in randomized trials. *Stat Med 30*(19), 2389–2408.

Moore, K. and M. van der Laan (2009a). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine 28*(1), 39–64.

Moore, K. L. and M. J. van der Laan (2009b). Increasing power in randomized trials with right censored outcomes through covariate adjustment. *Journal of Biopharmaceutical Statistics 19*(6), 1099–1131. PMID: 20183467.

Morgan, K. L. and D. B. Rubin (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics 40*(2), 1263–1282.

Neyman, J. S., D. M. Dabrowska, and T. Speed (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 465–472.

Olszowy, W., J. Aston, C. Rua, and G. B. Williams (2019). Accurate autocorrelation modeling substantially improves fMRI reliability. *Nature communications 10*(1), 1–11.

Pepler, P. T., D. W. Uys, and D. G. Nel (2016). A comparison of some methods for the selection of a common eigenvector model for the covariance matrices of two groups. *Communications in Statistics - Simulation and Computation 45*(8), 2917–2936.

Petersen, R., R. Thomas, M. Grundman, D. Bennett, R. Doody, S. Ferris, D. Galasko, S. Jin, J. Kaye, A. Levey, E. Pfeiffer, M. Sano, C. van Dyck, and L. Thal (2005). Vitamin E and Donepezil for the Treatment of Mild Cognitive Impairment. *New England Journal of Medicine 352*(23), 2379–2388. PMID: 15829527.

Pocock, S., S. Assmann, L. Enos, and L. Kasten (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med 21*(19), 2917–2930.

Pocock, S. J. and R. Simon (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics 31*(1), 103–115.

Power, J. D., A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, et al. (2011). Functional network organization of the human brain. *Neuron 72*(4), 665–678.

Robins, J. M. (2002). Covariance adjustment in randomized experiments and observational studies: Comment. *Statistical Science 17*(3), 309–321.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*(427), 846–866.

Robins, J. M., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science 22*(4), 544–559.

Rubin, D. and M. van der Laan (2008). Covariate adjustment for the intention-to-treat parameter with empirical efficiency maximization. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 229*, https://biostats.bepress.com/ucbbiostat/paper229.

Rubin, D. and M. van der Laan (2011). Targeted ancova estimator in rcts. In M. van der Laan and S. Rose (Eds.), *Targeted Learning: Causal Inference for Observational and Experimental Data*, Chapter 12, pp. 201–215. New York, NY: Springer.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Schott, J. (1999). Partial common principal component subspaces. *Biometrika 86*(4), 899–908.

Sen, P. K. (1988). Asymptotics in finite population sampling. In *Sampling*, Volume 6 of *Handbook of Statistics*, pp. 291 – 331. Elsevier.

Shao, J. and X. Yu (2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics 69*(4), 960–969.

Shao, J., X. Yu, and B. Zhong (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika 97*(2), 347–360.

Shorack, G. R. and J. A. Wellner (2009). *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics.

Siddiqui, O., H. M. J. Hung, and R. O'Neill (2009). MMRM vs. LOCF: A Comprehensive Comparison Based on Simulation Study and 25 NDA Datasets. *Journal of Biopharmaceutical Statistics 19*(2), 227–246. PMID: 19212876.

Thomas Yeo, B. T., F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology 106*(3), 1125–1165.

Tian, L., F. Jiang, T. Hasegawa, H. Uno, M. Pfeffer, and L. Wei (2019). Moving beyond the conventional stratified analysis to estimate an overall treatment efficacy with the data from a comparative randomized clinical study. *Statistics in Medicine 38*(6), 917–932.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 411–423.

Treatment for Adolescents With Depression Study (TADS) Team (2004). Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for adolescents with depression study (TADS) randomized controlled trial. *JAMA 292*(7), 807–820.

Tsiatis, A. (2007). *Semiparametric theory and missing data.* Springer Science & Business Media.

Tsiatis, A., M. Davidian, M. Zhang, and X. Lu (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Stat Med 27*(23), 4658–4677.

van der Laan, M. J. and S. Gruber (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics 8*(1), Article 9.

van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Van Essen, D. C., K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, et al. (2012). The human connectome project: a data acquisition perspective. *Neuroimage 62*(4), 2222–2231.

Wager, S., W. Du, J. Taylor, and R. Tibshirani (2016). High-dimensional regression adjustments in randomized experiments. *Proc Natl Acad Sci USA 113*(45), 12673–12678.

Wang, B., X. Luo, Y. Zhao, and B. Caffo (2020). Semiparametric partial common principal component analysis for covariance matrices. *Biometrics*.

Wang, B., E. L. Ogburn, and M. Rosenblum (2019). Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. *Biometrics 75*(4), 1391–1400.

Wang, B., R. Susukida, R. Mojtabai, M. Amin-Esmaeili, and M. Rosenblum (2019). Model-robust inference for clinical trials that improve precision by stratified randomization and adjustment for additional baseline variables. *arXiv https://arxiv.org/abs/1910.13954*.

Wang, W., X. Zhang, and L. Li (2019). Common reducing subspace model and network alternation analysis. *Biometrics 75*(4), 1109–1120.

Wei, L. J. (1978). The adaptive biased coin design for sequential experiments. *The Annals of Statistics 6*(1), 92–100.

Weiss, R. D., J. S. Potter, D. A. Fiellin, M. Byrne, H. S. Connery, W. Dickinson, J. Gardin, M. L. Griffin, M. N. Gourevitch, D. L. Haller, A. L. Hasson, Z. Huang, P. Jacobs, A. S. Kosinski, R. Lindblad, E. F. McCance-Katz, S. E. Provost, J. Selzer, E. C. Somoza, S. C. Sonne, and W. Ling (2011). Adjunctive Counseling During Brief and Extended Buprenorphine-Naloxone Treatment for Prescription Opioid Dependence: A 2-Phase Randomized Controlled Trial. *JAMA Psychiatry 68*(12), 1238–1246.

Xu, X., C. Y. H. Chen, and W. K. Härdle (2019). Dynamic credit default swap curves in a network topology. *Quantitative Finance 19*(10), 1705–1726.

Yang, L., W. Ma, Y. Qin, and F. Hu (2020). Testing for treatment effect in covariate-adaptive randomized clinical trials with generalized linear models and omitted covariates. *arXiv https://arxiv.org/abs/2009.04136*.

Yang, L. and A. Tsiatis (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician 55*(4), 314–321.

Ye, J., S. Dobson, and S. McKeever (2012). Situation identification techniques in pervasive computing: A review. *Pervasive and Mobile Computing 8*(1), 36 – 66.

Ye, T. and J. Shao (2020). Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(5), 1301–1323.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases 27*(7), 365 – 375.

Zhang, M. (2015). Robust methods to improve efficiency and reduce bias in estimating survival curves in randomized clinical trials. *Lifetime Data Analysis 21*(1), 119–137.

Zhou, G., A. Cichocki, Y. Zhang, and D. P. Mandic (2016). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems 27*(11), 2426–2439.