Statistical methods for surrogacy and hypothesis testing in HIV research

Jiacheng Wu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Ying Qing Chen, Chair

Chongzhi Di

Kwun Chan

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical methods for surrogacy and hypothesis testing in HIV research

Jiacheng Wu

Chair of the Supervisory Committee:
Professor Ying Qing Chen
Department of Biostatistics

This dissertation focuses on the estimation of surrogacy for biomarkers in time-to-event setting and hypothesis testing for mixed-effects models. In the first part, we develop a new measure to evaluate the proportion of treatment effect explained by surrogate markers in the time-to-event setting, with an application to HIV clinical trials. In the second part, we develop a hypothesis testing procedure for random-effects meta-analysis. We consider the exact likelihood ratio tests for two hypotheses with boundary problems, including testing the global null and homogeneity. The proposed method works well regardless of the number of studies. We apply the methodology to assess the association between circumcision and HIV among men who have sex with men. In the third part, we develop a general hypothesis testing framework in mixed-effects models. We are interested in testing the hypothesis that involves both fixed-effects parameters and random-effects variance component and we propose a novel construction of independent score statistics from both components. We consider the linear combinations of two score statistics. We illustrate the power tradeoff between different linear combination weights and propose to choose the weight based on Bayes and Minimax criteria so that power can be balanced well across the alternative space. We apply the methodology to random-effects meta-analysis, set-based genetic association analysis and time-varying treatment effect in survival analysis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my dissertation advisors, Ying Chen and Chongzhi Di, for their tremendous support, guidance and patience throughout my PhD studies. They set a good example of what statisticians they should be to solve real world problems and make an impact on the scientific community. I am grateful to Chongzhi Di for providing me with interesting dissertation projects and guiding me along the way. I am forever indebted to Forrest Crawford, my advisor at Yale, for teaching me how to conduct statistical and applied research and for his kindness and encouragement. I am grateful for the education at UW that trains me to be a principled statistician. I am grateful for my RA supervisors, Susan Shortreed for her support of my graduate studies. I would like to thank faculty members, Ann Duerr, Li Hsu, Gary Chan, Alex Luedtke, and Lurdes Inoue who provided help during my PhD studies.

I am deeply indebted to Gitana Garofalo for her kindness that makes the Biostatistics program wonderful and her help and support. I would like to thank all my friends at the department of Biostatistics and UW that helped me and supported me through this journey. Last but not least, I am grateful to my parents for their unconditional love.

DEDICATION

to my family

Chapter 1

INTRODUCTION

This dissertation consists of research on three topics related to surrogacy measure and hypothesis testing, motivated by HIV/ADIS research. In the first part, we develop a new measure to evaluate the surrogacy of a marker in the time-to-event setting with an application to HIV clinical trials. In the second part, we develop a hypothesis testing procedure for random-effects meta-analysis and meta-regression with an application to assess the association between circumcision and HIV among men who have sex with men. In the third part, we develop a general hypothesis testing framework for testing both fixed-effects parameters and random-effects variance component in mixed-effects models. We apply the methodology to meta-analysis, genetic association studies and time-varying treatment effect in survival analysis. In this chapter, we outline the motivations and provide an overview of the methodological contributions.

1.1 Proportion of treatment explained by surrogate marker for time-to-event data

We usually conduct clinical trials to evaluate the effectiveness of a new treatment. However, sometimes the clinical endpoint is rare requiring a large sample size, and the trial takes a long time to follow up. Instead of depending on clinical endpoints, there has been considerable attention in using surrogate biomarkers to decide the efficacy of a new treatment. Surrogate biomarkers are usually intermediate in the disease development process. They can be measured earlier and are correlated with the clinical endpoint. For example, in HIV trials, CD4 counts and viral loads are widely used as surrogate measures of how a treatment affects patients' health, and they are correlated with the development of AIDS and death.

Using surrogate biomarker can sufficiently reduce the sample size of a clinical trial, shorten the trial time and thus decrease the trial cost.

Prentice (1989) defined surrogacy in terms of hypothesis testing, in which the test to reject the null with no treatment effect using the surrogate marker should also reject the null using the true clinical endpoint. But this definition is too strict and does not quantify the magnitude of using surrogate biomarkers to predict the clinical endpoint based on treatment groups. Researchers are more interested in the proportion of treatment effect explained (PTE) by surrogate biomarker. In regression framework, PTE is defined as the proportion of reduction of regression coefficients for the treatment indicator after adjusting for surrogate markers (Weir and Walley, 2006). But previous work on surrogate markers has focused on binary endpoint and model-based measure of surrogacy.

In chapter 2, we introduce a new measure of PTE in time-to-event setting. This measure captures the time-varying nature of surrogacy and does not depend on parametric or semi-parametric models. We propose a plug-in estimator of this measure and prove that it converges to a Gaussian process. We provide Wald-type confidence interval which has nominal coverage in simulation studies. We apply this new measure to an HIV clinical trial to assess the surrogacy of CD4 count. We also apply it to Asia Cohort Consortium data set, a large-scale observational study, to assess the proportion of added risk of smoking associated with age.

1.2 Exact Likelihood ratio tests for random-effects meta-analysis and meta-regression

Meta-analysis has been widely used across various scientific disciplines to combine results from independent studies and strengthen the evidence about treatment efficacy or outcome exposure relationship. Fixed-effects models and random-effects models are two common approaches in meta-analysis. Fixed-effects models assume that the true effect is the same for all studies. Its estimator is constructed by taking a weighted average of individual study effects with weights being inversely proportional to variances of these effects estimates. On

the other hand, random-effects models allow for between-study heterogeneity by assuming effects from individual studies to be random.

Existing testing procedures for random-effects models focus on testing the average effect. DerSimonian and Laird (1986) first proposed an inference procedure to test the average effect while plugging in a method-of-moment estimator for the heterogeneity parameter (variance component). It gained popularity due to its simplicity and easy implementation (Hardy and Thompson, 1996; Biggerstaff and Tweedie, 1997; Thompson and Sharp, 1999; Hartung and Knapp, 2001a). However, it has been shown that this test always yields less significant p-values than that based on fixed effect models in many scenarios (Han and Eskin, 2011). Thus, it lacks power for discovering new associations even though it is designed to account for heterogeneity. Besides, this approach can have inaccurate confidence interval coverage and Type I error, especially when the number of studies is small or moderate (Brockwell and Gordon, 2001; DerSimonian and Kacker, 2007; Guolo and Varin, 2017), which is often the case in practice due to resource constraints.

Recently there has been more attention in testing the global null hypothesis in the random-effects models that there is no effect in any of the study, which corresponds to test that both average effect and variance component are zero (Han and Eskin, 2011). Existing testing procedures for the global null rely on asymptotic results, but they may not work well as the number of studies is typically small to moderate. Another difficulty arises due to the nonstandard situation where the variance component might be zero, which is on the boundary of its parameter space.

In Chapter 3, we consider exact likelihood ratio tests for two hypotheses with boundary problems, including testing the global null and homogeneity. Based on spectral decomposition, we characterize exact distributions of the likelihood ratio test under the null and alternative hypotheses. This facilitates fast computation of not only the null distribution for p-values but also the power function, which allows comprehensive power comparison between tests based on random-effects and fixed-effects models and provides tools for practitioners in planning and designing their studies. The proposed test performs well regardless of the

number of studies, and can have substantially higher power than tests based on fixed-effects models in the presence of heterogeneity.

1.3 Hypothesis testing for both fixed effects and random effects in mixed-effects models

Mixed effects models are widely used in various settings, including meta-analysis, longitudinal and correlated data analysis, nonparametric smoothing, genetic association and sequencing studies and survival analysis. In many scenarios, it is of interest to study a set of many parameters simultaneously, and the main objective is to test the global null hypothesis that all parameters are zero. Treating them as fixed effects may suffer from low power, especially when the number of parameters is large or there is little information to estimate each parameter reliably. An alternative approach is to treat these parameters as random effects that follow an underlying common distribution, e.g., Gaussian distribution. Under the random effects model, testing global null can be specified as testing that both fixed-effects parameters and random-effects variance components are zero. Examples include testing the existence of effects in meta-analysis, whether a set of single-nucleotide polymorphisms (SNPs) are associated with disease outcomes in genetic variant association studies, and whether there is a time-varying treatment effect in survival analysis.

Except for a few special cases, the asymptotic distribution of the likelihood ratio test is complicated and intractable. Alternatively, score tests have gained popularity for testing the variance component because they only require fitting the model under the null and can reduce the computational complexity (Lin and Breslow, 1996; Lin, 1997; Lin and Zhang, 1999; Verbeke and Molenberghs, 2003; Zhang and Lin, 2003). Under a set of regularity conditions, score statistics for two parameters usually have (asymptotically) multivariate Gaussian distribution, and thus can be combined into a χ^2 test statistic. However, the asymptotic behavior for the score statistic of the variance component is nonstandard and also correlated with the score statistic for fixed-effects parameters.

In Chapter 4, we address these challenges and consider a general mixed-effects framework.

We propose a novel construction of independent score statistics from fixed-effects parameters and random-effects variance components, and we consider the linear combinations of two score statistics. We show that the null and alternative distributions of this linear combination have tractable asymptotic distribution and are easy to approximate by a non-central chi-squared random variable, which facilitates fast power computation. We illustrate the power tradeoff between linear combination weights and propose to choose the weight based on Bayes and Minimax criteria so that power can be balanced well across the alternative space. In comprehensive simulation studies, we compare our method to existing nonlinear methods of combining independent score statistics.

Chapter 2

NONPARAMETRIC ESTIMATION OF PROPORTION OF TREATMENT EXPLAINED BY SURROGATE MARKER FOR TIME-TO-EVENT DATA

2.1 Introduction

The aim of clinical trials is to evaluate the effectiveness of a new treatment. But it is costly and time-consuming to conduct and follow up clinical trials. Instead of depending on clinical endpoints, there has been considerable attention in using surrogate biomarkers to decide the efficacy of a new treatment. For example, in HIV trials, CD4 counts are widely used as a surrogate measure of how a treatment affects patients' health. Using surrogate biomarker can sufficiently reduce the sample size of a clinical trial, shorten the trial time and thus decrease the trial cost.

Prentice (1989) defined surrogacy in terms of hypothesis testing, in which the test to reject the null with no treatment effect using the surrogate marker should also reject the null using the true clinical endpoint. But this definition is too strict and does not allow assessing the strength of surrogate biomarkers to predict clinical endpoint based on treatment groups. Researchers are more interested in the proportion of treatment explained (PTE) by surrogate biomarker. In regression framework, PTE is defined as the ratio of regression coefficients before and after adjusting for surrogate biomarker (Weir and Walley, 2006). For instance, Freedman et al (1992) considered PTE in the binary outcome setting as a ratio of regression coefficients in two logistic regression models. Lin et al (1997) extended PTE to time-to-event data in two Cox proportional hazard models,

$$\begin{aligned} \lambda(t|Z) &= \lambda_{10}(t)e^{\alpha Z}, \\ \lambda(t|Z,X) &= \lambda_{20}(t)e^{\beta Z+\gamma X}, \end{aligned} \tag{2.1}$$

where Z is binary treatment or exposure and X is surrogate marker, and PTE is defined as

$$\phi = 1 - \frac{\beta}{\alpha}. \tag{2.2}$$

However, Cox models in (2.1) cannot be true simultaneously, and the interaction between Z and X is omitted. Another drawback of this definition is that the PTE estimate is variable and sometimes can be below 0 or above 1. Alternatively, Wang and Taylor (2002) proposed F-measure in the binary endpoint setting without assuming regression models and allowing more flexibility in modeling,

$$\phi = \frac{A_1 - A_{10}}{A_1 - A_0},$$

where $A_1 = \text{pr}(T = 1|Z = 1)$, $A_0 = \text{pr}(T = 1|Z = 0)$, $A_{10} = \sum_x \text{pr}(T = 1|Z = 1, X = x)\text{pr}(X = x|Z = 0)$, and T is binary endpoint and X is categorical surrogate marker. The denominator is the treatment effect, and the numerator is the difference of the survival probability in the treated group due to the covariate adjustment in the control group. However, all these PTE measures are model-based and do not vary over time. In time-to-event settings, PTE could be time-varying, and we are interested in when the surrogate biomarker can explain the treatment the most.

In this chapter, we propose a new PTE measure in time-to-event setting and allow PTE to change over time. We show that PTE defined in Lin et al (1997) is a special case of our definition. We estimate PTE nonparametrically and derive its variance estimator. We show in simulation studies that this estimator has nominal confidence interval coverage. We apply this new measure to a HIV clinical trial to assess the surrogacy of CD4+ count. We also apply it to Asia Cohort Consortium dataset, a large-scale observational study, to assess the proportion of added risk of smoking associated with age.

2.2 Methods

2.2.1 Definition

We consider binary treatment $Z = 0, 1$ with 1 indicating the treatment group and categorical surrogate marker with K levels, $X = 1, \ldots, K$. For now, we assume that surrogate marker is

measured at a time point $c_0 > 0$ and we consider the time $t \geq c_0$. Denote the time-to-event by T. Let $S_1(t) = \text{pr}(T \geq t|Z = 1)$ be unadjusted survival probability for the treatment group and $S_0(t) = \text{pr}(T \geq t|Z = 0)$ be the unadjusted survival probability for the control group. Conditional hazard function of Z and X is defined as

$$\lambda(t|Z = z, X = x) = \lim_{\Delta t \to 0} \frac{\text{pr}(T \leq t + \Delta t|T \geq t, Z = z, X = x)}{\Delta t},$$

for $z = 0, 1$ and $x = 1, \ldots, K$. Conditional hazard $\lambda(t|Z = z)$ can be defined in a similar way. Note that the unadjusted survival function for $Z = 0$ and $Z = 1$ can also be written as $S_0(t) = \exp\left\{-\int_0^t \lambda(u|Z = 0)du\right\}$ and $S_1(t) = \exp\left\{-\int_0^t \lambda(u|Z = 1)du\right\}$.

The adjusted hazard function for the treatment group using the covariate distribution in the control group is defined as

$$\lambda_1^*(t) = \sum_{x=1}^{K} \lambda(t|Z = 1, X = x)\text{pr}(X = x|T \geq t, Z = 0),$$

where $\text{pr}(X|T \geq t, Z = 0)$ is the conditional distribution function of X, provided that the failure time T in the control group is greater than or equal to t. Adjusted hazard function measures the instantaneous rate of death in the treatment group if their surrogate marker values are distributed as those who survive at t in the control group. We then define the adjusted survival probability for the treatment group in terms of adjusted hazard rate

$$S_1^*(t) = \exp\left\{-\int_0^t \lambda_1^*(u)du\right\}.$$

Finally, we define the proportion of treatment effect explained by surrogate marker for the time-to-event data as

$$\phi(t) = \frac{g\{S_1(t)\} - g\{S_1^*(t)\}}{g\{S_1(t)\} - g\{S_0(t)\}}, \tag{2.3}$$

where $g(x) = \log\{-\log(x)\}$ is the complementary log-log link function. We can interpret $g\{S_1(t)\} - g\{S_1^*(t)\}$ as the change of survival probability in the complementary log-log scale due to the change in the surrogate marker distribution from treatment group to control group. In addition, $g\{S_1(t)\} - g\{S_0(t)\}$ can be interpreted as the treatment effect in the

complementary log-log scale at time t. Thus, $\phi(t)$ can be interpreted as a measure of the proportion of treatment effect on survival that is explained by surrogate marker. Unlike equation (2.2), this definition does not depend on a statistical model and allows PTE to change over time. It is useful when researchers are interested in the time when surrogate marker explains the treatment effect the most.

2.2.2 Example

Here we show that our definition is the same as the definition in equation (2.2) if two Cox models are assumed. Under a working Cox model of the form $\lambda(t|Z,X) = \lambda_{20}(t)\exp(\beta Z + \gamma X)$, we have

$$\frac{-\log S_1^*(t)}{-\log S_0(t)} = \frac{\int_0^t \sum_{x=1}^K \lambda_{20}(u) e^{\beta+\gamma x} \mathrm{pr}(X=x|T\geq u, Z=0)du}{\int_0^t \sum_{x=1}^K \lambda_{20}(u) e^{\gamma X} \mathrm{pr}(X=x|T\geq u, Z=0)du} = e^{\beta}.$$

Under another working Cox model of the form $\lambda(t|Z) = \lambda_{10}(t)\exp(\alpha Z)$, we have

$$\frac{-\log S_1(t)}{-\log S_0(t)} = \frac{\int_0^t \lambda_{10}(u) e^{\alpha} du}{\int_0^t \lambda_{10}(u) du} = e^{\alpha}.$$

Thus PTE in equation (2.3) can be expressed as

$$\phi(t) = 1 - \frac{g\{S_1^*(t)\} - g\{S_0(t)\}}{g\{S_1(t)\} - g\{S_0(t)\}} = 1 - \frac{\log\left\{\frac{-\log S_1^*(t)}{-\log S_0(t)}\right\}}{\log\left\{\frac{-\log S_1(t)}{-\log S_0(t)}\right\}} = 1 - \frac{\beta}{\alpha}, \tag{2.4}$$

same as the definition in equation (2.2). Note that in this calculation, different working models are imposed in the numerator and denominator in (2.4). But if we use the same model in the numerator and denominator in (2.4), our proposed PTE measure in (2.3) will not reduce to the PTE in (2.2).

2.3 Estimation and Inference

We adopt standard notations to estimate and make inference for the newly defined PTE when the time-to-event outcomes are subject to censoring. Suppose we have n subjects with T_i and C_i denoting the time-to-event and censoring time, respectively, $i = 1, \cdots, n$. The observed

data consists of n independent and identically distributed (iid) $(U_i, \delta_i, Z_i, X_i), i = 1, \cdots, n$, where $U_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, Z_i is the binary variable for treatment and control, and X_i is the categorical variable for surrogate marker. Let $N_i(t) = I(U_i \leq t, \delta_i = 1)$ be the observed counting indicator and $Y_i(t) = I(U_i \geq t)$ be the at-risk indicator. Without loss of generality, sometimes we may drop the subscript i for general terms.

We order the m distinct event time in the treatment group ($Z = 1$): $t_1 < \ldots < t_m$. Let a_{1ki} be the number of events in the treatment group with surrogate marker strata k at time $t_i, i = 1, \ldots, m$. Let n_{1ki} be the number of individuals at risk in treatment group ($Z = 1$) with surrogate marker strata k at time t_i. Let n_{0ki} be the number of individuals at risk in control group ($Z = 0$) with surrogate marker strata k at time t_i. Kaplan-Meier estimators (Kaplan and Meier, 1958) for $S_0(t) = P(T \geq t|Z = 0)$ and $S_1(t) = P(T \geq t|Z = 1)$ are

$$\hat{S}_0(t) = \prod_{i:t_i<t} \left(1 - \frac{\sum_{k=1}^{K} a_{0ki}}{\sum_{k=1}^{K} n_{0ki}}\right), \tag{2.5}$$

$$\hat{S}_1(t) = \prod_{i:t_i<t} \left(1 - \frac{\sum_{k=1}^{K} a_{1ki}}{\sum_{k=1}^{K} n_{1ki}}\right), \tag{2.6}$$

respectively. The conditional failure probability in the treatment group with surrogate marker strata k at time t_i is $q_{1ki} = a_{1ki}/n_{1ki}$. The adjusted conditional probability for treatment group using surrogate marker distribution in the control group is $q^*_{1i} = \sum_{k=1}^{K} w_{ki} q_{1ki}$, where $w_{ki} = n_{0ki}/\sum_{k=1}^{K} n_{0ki}$. The adjusted survival curve $S^*_1(t)$ can be estimated by

$$\hat{S}^*_1(t) = \prod_{i:t_i \leq t} (1 - q^*_{1i}). \tag{2.7}$$

Therefore, we obtain a plug-in estimator for PTE in equation (2.3) using the estimators of survival functions in equation (2.5), (2.6) and (2.7),

$$\hat{\phi}(t) = \frac{g\{\hat{S}_1(t)\} - g\{\hat{S}^*_1(t)\}}{g\{\hat{S}_1(t)\} - g\{\hat{S}_0(t)\}}. \tag{2.8}$$

Without loss of generality, sometimes we may drop the subscript i for general terms. In addition, we also drop the dependency of t in the PTE to simplify notation and write

$$\phi = \frac{g_1 - g^*_1}{g_1 - g_0},$$

where $g_1 = g(S_1)$, $g_0 = g(S_0)$, $g_1^* = g(S_1^*)$ and $g(x) = \log\{-\log(x)\}$. Its estimator is

$$\hat{\phi} = \frac{\hat{g}_1 - \hat{g}_1^*}{\hat{g}_1 - \hat{g}_0},$$

where $\hat{g}_1 = g(\hat{S}_1)$, $\hat{g}_0 = g(\hat{S}_0)$ and $\hat{g}_1^* = g(\hat{S}_1^*)$.

Theorem 1. *Under the conditions:*

1. *There exists a constant $\tau > 0$, such that for $t \in (c_0, \tau)$, $S_1(\tau)S_0(\tau) > 0$ and $1 - \{1 - F(\tau_-)\}\{1-G(\tau_-)\} < 1$ where c_0 is the time to measure the marker, F is the distribution function of time-to-event T, and G is the distribution function of censoring time C.*

2. *Survival probability $S_1(t) \neq S_0(t)$ for $t \in (c_0, \tau)$.*

3. *Random censoring: censoring time C is independent of the failure time T and marker X.*

We have that $n^{1/2}(\hat{\phi}-\phi)$ converges to a zero-mean Gaussian process with covariance function $E\{\psi(t)\psi(t')\}$ and

$$\psi(t) = \frac{g_1 - g_1^*}{(g_1 - g_0)^2}\eta_0 + \frac{g_1^* - g_0}{(g_1 - g_0)^2}\eta_1 + \frac{1}{g_0 - g_1}\eta_2,$$

where

$$\eta_0 = -\frac{1}{\log(S_0)} I(Z=0) \int_0^t \frac{dM(u|Z=0)}{E\{Y(u)I(Z=0)\}}, \quad \eta_1 = -\frac{1}{\log(S_1)} I(Z=1) \int_0^t \frac{dM(u|Z=1)}{E\{Y(u)I(Z=1)\}},$$

$$\eta_2 = -\frac{1}{\log(S_1^*)} \int_0^t \sum_{x=1}^{K} I(Z=1, X=x) \frac{pr(X=x|T \geq u, Z=0)}{E\{Y(u)I(Z=1,X=x)\}} dM(u|Z=1, X=x)$$
$$- \frac{1}{\log(S_1^*)} \int_0^t \sum_{x=1}^{K} Q(u,x)\lambda(u|Z=1, X=x)du, \tag{2.9}$$

$$Q(u,x) = \frac{I(X=x, Z=0, T \geq u)}{pr(T \geq u|Z=0)pr(Z=0)} - \frac{pr(X=x|Z=0, T \geq u)}{pr(T \geq u|Z=0)pr(Z=0)} I(Z=0, T \geq u), \tag{2.10}$$

and $N(u) = I(T \leq C, T \leq u)$ *is observed counting indicator,* $Y(u) = I(T \geq u, C \geq u)$ *is at-risk indicator, and* $dM(u|Z = z) = dN(u) - Y(u)\lambda(u|Z = z)du$ *is counting process martingale for* $z = 0, 1$. *We also define* $dM(u|Z = z, X = x) = dN(u) - Y(u)\lambda(u|Z = z, X = x)du$ *for* $x = 1, \cdots, K$ *and* $z = 0, 1$,

The proof of theorem 1 is in the Appendix A.1. We can use this result to consistently estimate the covariance of $n^{1/2}\left\{\hat{\phi}(t) - \phi(t)\right\}$ by

$$\widehat{\sigma}_{\phi}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i(t)\hat{\psi}_i(s),$$

where

$$\hat{\psi}_i(t) = \frac{g_1 - g_1^*}{(g_1 - g_0)^2}\eta_{0i} + \frac{g_1^* - g_0}{(g_1 - g_0)^2}\eta_{1i} + \frac{1}{g_0 - g_1}\eta_{2i},$$

and

$$\begin{aligned}
\eta_{0i} &= -\frac{1}{\log(S_0)}I(Z_i = 0)\int_0^t \frac{dN_i(u) - Y_i(u)\lambda(u|Z = 0)du}{E\left\{Y(u)I(Z = 0)\right\}},\\
\eta_{1i} &= -\frac{1}{\log(S_1)}I(Z_i = 1)\int_0^t \frac{dN_i(u) - Y_i(u)\lambda(u|Z = 1)du}{E\left\{Y(u)I(Z = 1)\right\}},\\
\eta_{2i} &= -\frac{1}{\log(S_1^*)}\int_0^t \sum_{x=1}^{K} I(Z_i = 1, X_i = x)\frac{\mathrm{pr}(X = x|T \geq u, Z = 0)}{E\left\{Y(u)I(Z = 1, X = x)\right\}}\left\{dN_i(u) - \right.\\
&\quad \left. Y_i(u)\lambda(u|Z = 1, X = x)du\right\} - \frac{1}{\log(S_1^*)}\int_0^t \sum_{x=1}^{K} Q_i(u,x)\lambda(u|Z = 1, X = x)du,\\
Q_i(u,x) &= \frac{I(X_i = x, Z_i = 0, T_i \geq u)}{\mathrm{pr}(T \geq u|Z = 0)\mathrm{pr}(Z = 0)} - \frac{\mathrm{pr}(X = x|Z = 0, T \geq u)}{\mathrm{pr}(T \geq u|Z = 0)\mathrm{pr}(Z = 0)}I(Z_i = 0, T_i \geq u).
\end{aligned} \tag{2.11}$$

2.4 Simulation

We simulate survival data from Cox model with Weibull baseline hazard

$$\lambda(t|Z, X) = \alpha\rho t^{\rho-1}\exp(\beta_1 Z + \beta_2 X + \beta_3 ZX).$$

We consider binary treatment $Z = 0, 1$ and binary surrogate marker $X = 0, 1$ that need not be balanced in exposed and unexposed group, i.e., $\mathrm{pr}(X = 1|Z = 1) = p_1$ and $\mathrm{pr}(X = 1|Z =$

$0) = p_0$. There are equal number of observations in treatment and control group. Here we derive the analytical form of the true PTE. Under Cox model

$$\lambda(t|Z,X) = \alpha\rho t^{\rho-1}\exp(\beta_1 Z + \beta_2 X + \beta_3 ZX),$$

the survival probability is

$$\text{pr}(T \geq t|Z,X) = \exp(-\alpha t^{\rho} e^{\beta_1 Z+\beta_2 X+\beta_3 ZX}).$$

The true survival function for $Z = 0$ and $Z = 1$ is

$$\begin{aligned} S_0(t) &= \text{pr}(T \geq t|Z=0) = (1-p_0)\exp(-\alpha t^{\rho}) + p_0\exp(-\alpha t^{\rho} e^{\beta_2}), \\ S_1(t) &= \text{pr}(T \geq t|Z=1) = (1-p_1)\exp(-\alpha t^{\rho} e^{\beta_1}) + p_1\exp(-\alpha t^{\rho} e^{\beta_1+\beta_2+\beta_3}), \end{aligned} \tag{2.12}$$

respectively. Using Bayes formula

$$\text{pr}(X|T \geq t, Z=0) = \frac{\text{pr}(T \geq t|Z=0,X)\text{pr}(X|Z=0)}{\text{pr}(T \geq t|Z=0)},$$

we obtain

$$\text{pr}(X=0|T \geq t, Z=0) = (1-p_0)\frac{\exp(-\alpha t^{\rho})}{S_0(t)},$$

$$\text{pr}(X=1|T \geq t, Z=0) = p_0\frac{\exp(-\alpha t^{\rho} e^{\beta_2})}{S_0(t)},$$

where $S_0(t) = \text{pr}(T \geq t|Z=0) = (1-p_0)\exp(-\alpha t^{\rho}) + p_0\exp(-\alpha t^{\rho} e^{\beta_2})$. The adjusted hazard function for $Z = 1$ using the covariate distribution of $Z = 0$ is $\lambda_1^*(t) = \sum_{x=0}^{1}\lambda(t|Z=1,X=x)\text{pr}(X=x|T \geq t, Z=0)$. So the adjusted survival function for $Z = 1$ using the covariate distribution of $Z = 0$ is

$$S_1^*(t) = \exp\left[-\int_0^t \frac{\alpha\rho u^{\rho-1}}{S_0(u)}\left\{p_0\exp(\beta_1+\beta_2+\beta_3-\alpha u^{\rho}e^{\beta_2}) + (1-p_0)\exp(\beta_1-\alpha u^{\rho})\right\}du\right]. \tag{2.13}$$

The true PTE is accordingly derived from (2.12) and (2.13).

In this simulation, we consider two time points $t = 2, 7$, and choose $\alpha = 0.1, \rho = 1, \beta_1 = 0.5, \beta_2 = 1.5, \beta_3 = -0.8, p_0 = 0.25$ and $p_1 = 0.75$, in which case surrogate marker has a partial surrogacy with the true PTE approximately 0.69 and 0.57, respectively for $t = 2$ and

	$t = 2$		$t = 7$	
	censor = 35%	censor= 50%	censor= 35%	censor= 50%
N = 1000				
Bias	0.0511	0.0659	0.0164	0.0197
Variance	0.0835	0.1084	0.0286	0.0607
Sampling Variance	0.0674	0.1007	0.0293	0.0630
Coverage	0.956	0.946	0.946	0.949
N = 2000				
Bias	0.0272	0.0153	0.0121	0.0108
Variance	0.0341	0.0375	0.0137	0.0223
Sampling Variance	0.0337	0.0347	0.0136	0.0234
Coverage	0.957	0.951	0.950	0.944

Table 2.1: Simulation result. Data is generated by Cox model with censoring rate 35% and 50% and varying sample size. Bias, variance, sampling variance, and confidence interval coverage are shown at two time points, $t = 2$ and $t = 7$.

$t = 7$. Table 2.1 shows the bias, variance, sampling variance, and confidence interval coverage over 1000 replications. As the sample size increases, the bias and variance decrease. The increase in censoring rate increases the bias and variance. The proposed variance estimator is similar to the sampling variance, and the 95% confidence interval has approximate 95% coverage in repeated experiments.

2.5 Data analysis

2.5.1 Application to clinical trial

We study a clinical trial that aims to evaluate the effectiveness of antiretroviral therapy (ART) on preventing sexual transmission of HIV-1 in HIV-1 serodiscordant couples. It was

a phase three, randomized, controlled, multi-center trial. There were 1763 enrolled couples with one individual HIV-infected and the other individual not HIV-infected. They were assigned at random in a 1:1 ratio to two arms. Treatment arm initiated ART for the HIV-infected individual upon enrollment while control arm received HIV primary care without initiation of ART until the HIV-infected individual had two consecutive measurements of a CD4+ count below 250 cells/m^3 or develops an AIDS-defining illness. The study lasted for 78 months. The primary endpoint is the incident HIV infection occurring in the partners of HIV-infected cases. Date of the randomization is used as time origin. We use CD4+ count at week 12 after randomization as the surrogate marker. We are interested to estimate the proportion of treatment effect explained by week 12 CD4+ count. Very few individuals have event before week 12, and for individuals that do not have week 12 CD4+ count, the CD4+ count closest to week 12 is used. We dichotomize CD4+ count into 7 groups: 0-200, 201-250, 251-300, 301-350, 351-400, 401-450, greater than 450. Figure 2.1 shows the adjusted and unadjusted survival curve and the corresponding PTE. Adjusted survival curve lies in between unadjusted control and treatment survival curve. We truncate the PTE before 500 days because it is unstable at the beginning of the study. PTE varies with time, and horizontal line is around 0.34 calculated based on equation (2.2) from two Cox models. Week 12 CD4+ count explains the treatment effect the most at around 1,500 days.

2.5.2 Application to observational study

We use the Asia Cohort Consortium, a large-scale observational study that recruited approximately 1 million people in 19 cohorts in Asia (Zheng et al, 2011). Baseline data on BMI, age, sex, smoking status and follow-up data on death status are collected for all participants. All the cohorts have at least 5 years of follow-up time except for Taiwan Cardiovascular Disease Risk Factors Two-Township Study (CVDFACTS) cohort. We focus on the proportion of added risk of smoking associated with baseline age. We dichotomize the baseline age into 4 groups: less than or equal to 50, 51-60, 61-70, greater than 70. We study three cohorts: Singapore Chinese Health Study (SCHS), Taiwan Cardiovascular Disease Risk Factor Two-

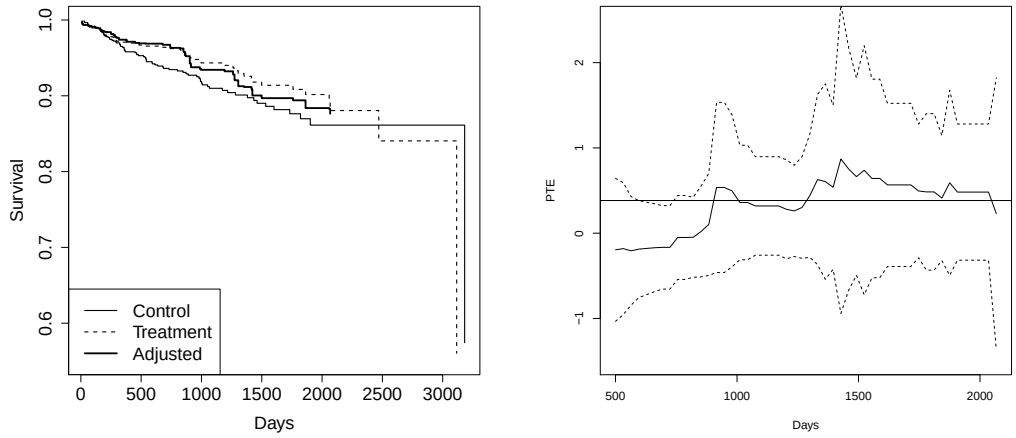


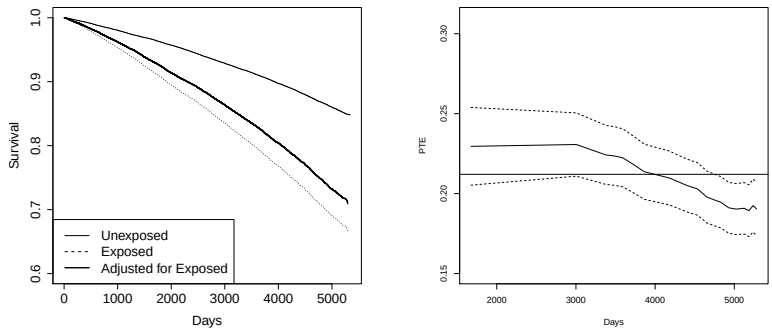
Figure 2.1: Application to HIV clinical trial. Left panel is the adjusted and unadjusted survival curve. Right panel is PTE with 95% pointwise confidence interval. In the right, the horizontal line is calculated based on equation (2.2) from two Cox models.

Township Study (CVDFACTS), Bangladesh Health Effects of Arsenic Longitudinal Study (HEALS). Figure 2.2 shows the adjusted and unadjusted survival curve in the left and the corresponding PTE with 95% pointwise confidence interval in the right. Due to the large sample size, the survival curve is smooth, and the adjusted survival curve lies in between the two unadjusted curves. The static PTEs calculated based on two Cox models in equation (2.2) lie in the middle between of the time-varying PTE. In the Singapore dataset, PTE decreases with time. In Taiwan dataset, PTE first increases and then decreases with time. In Bangladesh dataset, PTE exhibits a decreasing pattern.

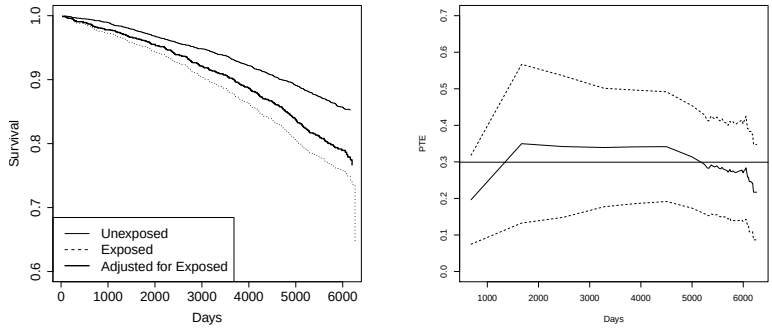
2.6 Discussion

In this chapter, we proposed a new measure of PTE in time-to-event setting based on unadjusted and adjusted survival functions and allow PTE to be time-varying. We showed that PTE defined in Lin et al (1997) is a special case of our definition. We proposed a plug-in estimator for this new measure and used a consistent variance estimator for inference. Sim-

Singapore



Taiwan



Bangladesh

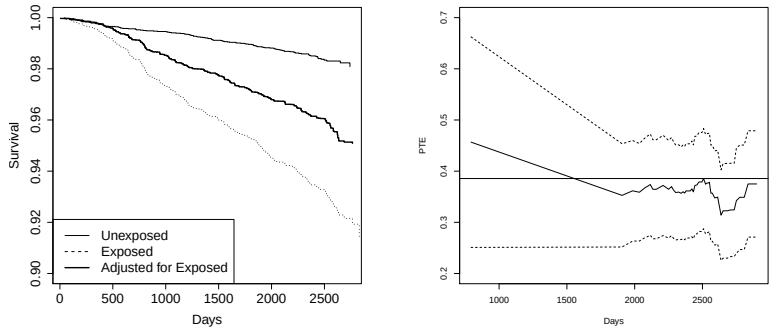


Figure 2.2: Application for Asia Cohort Consortium data. Three cohorts from Singapore, Taiwan and Bangladesh are shown. Left panel is the survival curve and right panel is the PTE with 95% pointwise confidence interval. Horizontal line in the right is calculated based on equation (2.2) from two Cox models.

ulation study indicated that the proposed estimator converges to the truth as sample size grows and the Wald-type confidence interval has nominal coverage. We applied the method in a clinical trial and an observational study.

There are theoretical challenges to derive the variance of the proposed plug-in estimator. The covariance between S_1^* and S_0 may not be zero since the marker distribution in the control group is used to calculate the adjusted survival function in the treatment group. In addition, the covariance of failure time between different strata may not be zero. In light of these difficulties, we adopt counting process martingale techniques to derive the variance of the plug-in estimator.

We only considered the surrogate marker measured at a single time point c_0. It is, however, easy to adapt our framework to allow time-varying surrogate marker X_t with $t \geq c_0$. The estimation and inference strategies remain the same by substituting X with X_t. In addition, we only deal with categorical marker, and future work will focus on extending the adjustment to continuous marker. The adjusted survival function is based on the adjusted hazard rate, and another way to define adjusted survival function is

$$\sum_{x=1}^{K} \text{pr}(T \geq t | Z = 1, X = x) \text{pr}(X = x | Z = 0).$$

This approach is akin to the F-measure in Wang and Taylor (2002), and we leave it as a future research direction. One drawback of using PTE to measure surrogacy is that its interpretation is hard when it is below 0 or above 1.

Note that the estimand proposed in this chapter is not causal but rather an empirical measure that can be easily calculated from the data. By contrast, Frangakis and Rubin (2002) proposed the principal stratification framework to evaluate the effectiveness of surrogate marker. Several methods to evaluate surrogacy have been proposed in this causal framework (Gilbert and Hudgens, 2008; Li et al, 2010; Wolfson and Gilbert, 2010; Huang and Gilbert, 2011; Zigler and Belin, 2012; Huang et al, 2013; Luedtke and Wu, 2017). However, the estimand in principal stratification framework is not identifiable without additional assumptions.

Chapter 3

EXACT LIKELIHOOD RATIO TESTS FOR META-ANALYSIS AND META-REGRESSION BASED ON RANDOM-EFFECTS MODELS

3.1 Introduction

Meta-analysis has been widely used across various scientific disciplines to combine results from independent studies. Fixed-effects models and random-effects models are two common approaches in meta-analysis. Fixed-effects models assume that the true effect is the same for all studies, while random-effects models allow for between-study heterogeneity. We first introduce notation. Consider the setting that there are p independent studies with sample sizes n_i, $i = 1, 2, \ldots, p$. Let β_i denote the true effect of interest from the ith study, X_i and σ_i denote its estimate and corresponding standard error, respectively. If each individual study is sufficiently large, one can assume that central limit theorem applies to parameter estimator X_i and that the summary statistics follows Gaussian distributions, $X_i \sim N(\beta_i, \sigma_i^2)$. Fixed-effects models assume a common effect across studies, i.e., $\beta_1 = \beta_2 = \ldots = \beta_p = \mu$, with the null hypothesis of no effects represented as $\mu = 0$. A random-effects model assumes that the true effects follow a Gaussian distribution with unknown mean μ and variance τ^2, i.e.,

$$M_r : \; X_i = \beta_i + \epsilon_i, \quad \beta_i \sim N(\mu, \tau^2), \quad \epsilon_i \sim N(0, \sigma_i^2), \quad i = 1, \ldots, p. \tag{3.1}$$

Under this model, the global null of no effects is represented as $H_0 : \mu = 0, \tau^2 = 0$, and the null hypothesis of homogeneity is represented as $H_{0,\tau^2} : \tau^2 = 0$.

In the literature, there were several works aiming at testing the average effect, i.e., $H_0' : \mu = 0$, under random-effects models. This is a different hypothesis from the global null H_0. For example, DerSimonian and Laird (1986) first proposed an inference procedure to test

H_0' while plugging in a method-of-moment estimator for the heterogeneity parameter τ^2. It gained popularity due to its simplicity and easy implementation (Hardy and Thompson, 1996; Biggerstaff and Tweedie, 1997; Thompson and Sharp, 1999; Hartung and Knapp, 2001a). However, it has been shown that this test always yields less significant p-values than that based on fixed effect models in many scenarios Han and Eskin (2011). Thus, it lacks power for discovering new associations even though it is designed to account for heterogeneity. Besides, this approach can have inaccurate confidence interval coverage and Type I error, especially when the number of studies is small or moderate (Brockwell and Gordon, 2001; DerSimonian and Kacker, 2007; Guolo and Varin, 2017), which is often the case in practice due to resource constraint.

Han and Eskin (2011) considered the likelihood ratio test (LRT) for the global null hypothesis $H_0 : \mu = 0, \tau^2 = 0$, which can be interpreted as no effects in each of the studies. They argued that testing the global null is scientifically relevant and potentially more powerful, as it is able to discover associations that occur in some studies where fixed effect-based tests may miss in the presence of heterogeneity. However, they did not derive the exact distribution of the test statistics, though the asymptotic distribution is available. To calculate p-values, they proposed simulation-based tabulated values for some specific and simplified scenarios (e.g., common sample sizes across studies). In this chapter, we develop the exact distribution of the LRT both under the null and alternative hypotheses, providing a testing procedure that works well regardless of the number of studies.

Testing the global null poses statistical challenges since the between-study variability is on the boundary of the parameter space under the global null and the standard regularity condition for LRT requires that the true parameter is in the interior of the parameter space. It is worth mentioning that Zeng and Lin (2015) and Kosmidis et al (2017) both considered asymptotic distribution of the maximum likelihood estimator (MLE) and (penalized) LRT. Their results require that the true value of the between-study heterogeneity be nonzero and thus do not apply when the true value is zero as is the case under the global null H_0 and homogeneity H_{0,τ^2}.

This chapter focuses on *exact* likelihood ratio test for the global null and homogeneity under the random-effects meta-analysis model, addressing difficulties arising from limited number of studies and the *nonstandard* situation that the variance component is on the boundary of its parameter space under the null. In Section 3.2, we characterized the exact distribution of the LRT both under the null and alternative hypotheses based on spectral decomposition, which allows very fast calculation of p-values and power function comprehensively across all types of alternatives. In Section 3.3, we conducted power comparisons between proposed tests and fixed-effects testing procedures, offering some insight for practitioners. In Section 3.4, we applied the methodology to a meta analysis to access the association between circumcision and HIV and other sexually transmitted infections among men who have sex with men. In Section 3.5, we extended the results to meta-regression.

3.2 Likelihood ratio tests for random-effects meta-analysis

3.2.1 Testing the global null based on random-effects models

We first consider the LRT for the global null hypothesis $H_0 : \mu = 0, \tau^2 = 0$. Denote by $Y_i = X_i/\sigma_i$ and the model M_r in (3.1) is equivalent to

$$Y_i = \sigma_i^{-1}\beta_i + \epsilon_i',\ \beta_i \sim N(\mu, \tau^2),\ \epsilon_i' \sim N(0,1).$$

It is easy to see that $Y_i \sim N\left(\mu/\sigma_i, 1 + \tau^2/\sigma_i^2\right)$ for $i = 1, \ldots, p$, or in vector form, $Y \sim N_p(\mu Z, V_\tau)$, where $Z = (1/\sigma_1, \ldots, 1/\sigma_p)^T = \Sigma^{-1/2}1$, $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $V_\tau = I_p + \tau^2\Sigma^{-1}$. We are interested in testing the hypothesis $H_0 : \mu = 0, \tau^2 = 0$ versus the alternative hypothesis $H_a : \mu \neq 0$ or $\tau^2 > 0$.

Twice the log likelihood function is (up to a constant that does not depend on the parameters)

$$2\log L(\mu, \tau^2) = -(Y - \mu Z)^T V_\tau^{-1}(Y - \mu Z) - \log|V_\tau|. \tag{3.2}$$

Under the alternative hypothesis H_a, fixing τ^2 and solving the score equation for μ gives

$$\hat{\mu} = (Z^T V_\tau^{-1} Z)^{-1} Z^T V_\tau^{-1} Y.$$

Plugging it back to (3.2), we have the profile log likelihood

$$2\log L(\hat{\mu}, \tau^2) = -Y^T P_\tau^T V_\tau^{-1} P_\tau Y - \log|V_\tau|,$$

where $P_\tau = I_p - Z(Z^T V_\tau^{-1} Z)^{-1} Z^T V_\tau^{-1}$. Under the null hypothesis H_0, we have $2\log L(0,0) = -Y^T Y$. The LRT statistic for testing H_0 is

$$LRT_p = \sup_{\tau^2} \left\{ Y^T P_0 Y - Y^T P_\tau^T V_\tau^{-1} P_\tau Y - \log|V_\tau| \right\} + Y^T (I_p - P_0) Y, \tag{3.3}$$

where $P_0 = I_p - Z(Z^T Z)^{-1} Z^T$ is a projection matrix of rank $(p-1)$. The following theorem characterizes the exact distribution of the LRT statistic under H_0.

Theorem 2. *Under the random-effects meta-analysis model M_r in* (3.1)*, the exact distribution of the LRT statistic under H_0 can be characterized as*

$$LRT_p \overset{d}{=} \sup_{\tau^2} \left\{ \sum_{i=1}^{p-1} \frac{\tau^2 \xi_i}{1+\tau^2 \xi_i} w_i^2 - \sum_{i=1}^{p} \log\left(1 + \frac{\tau^2}{\sigma_i^2}\right) \right\} + u^2,$$

where $(w_1, \ldots, w_{p-1}, u)$ are independent standard Gaussian random variables and $(\xi_1, \ldots, \xi_{p-1})$ are $(p-1)$ nonzero eigenvalues of the matrix $\Sigma^{-1} P_0$.

The proof of Theorem 2 is in Appendix A.2. Theorem 2 characterizes the exact distribution of the LRT for any study size p and within-study variability $(\sigma_1^2, \ldots, \sigma_p^2)$. The LRT_p can be decomposed nicely into two terms, with the first term corresponding to the variance component, which has a complex distribution, the second term corresponding to the fixed effect, which behaves like a standard χ_1^2 distribution. The first term is the supremum of a stochastic function involving $(p-1)$ independent χ_1^2 random variables, and generally does not have a closed form solution except for the special case of common variance across studies $\sigma_1^2 = \ldots = \sigma_p^2$.

Although the analytic form of this distribution is intractable, it can be simulated using a very fast algorithm that involves basic arithmetic operations on standard normal random variables. The supremum over τ^2 can be performed with a grid search, e.g., a set of 200 equally spaced grid points over $[-12, 12]$ in logarithm scale. It is also of interest to study the power of the exact LRT, which is characterized as follows.

Theorem 3. *Under the random-effects meta-analysis model M_r in* (3.1)*, the exact distribution of the likelihood ratio test statistic under alternative $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$ can be represented as follows,*

$$LRT_p \overset{d}{=} \sup_{\tau^2} \left\{ \sum_{i=1}^{p-1} (1+\tau_a^2\xi_i) \frac{\tau^2\xi_i}{1+\tau^2\xi_i} w_i^2 - \sum_{i=1}^{p} \log\left(1+\frac{\tau^2}{\sigma_i^2}\right) \right\} + \left[\mu_a U^T Z + \left(1+\tau_a^2 U^T\Sigma^{-1}U\right)^{1/2} \left\{ w^T b + u(1-b^T b)^{1/2} \right\} \right]^2,$$

where $(w_1, \ldots, w_{p-1}, u)$ are independent standard Gaussian random variables, $w = (w_1, \ldots, w_{p-1})^T$, $(\xi_1, \ldots, \xi_{p-1})$ are $(p-1)$ nonzero eigenvalues of $\Sigma^{-1}P_0$, and U is $p \times 1$ matrix that can be constructed by eigenvalue decomposition of $I_p - P_0$ such that $U^TU = 1$ and $UU^T = I_p - P_0$. Furthermore, $W = AB$ is a $p \times (p-1)$ matrix, where the $p \times (p-1)$ matrix A satisfies $A^TA = I_{p-1}$ and $AA^T = P_0$, and the $(p-1) \times (p-1)$ matrix B satisfies $A^T\Sigma^{-1}A = B\,diag(\xi_i)B^T$. Thus W satisfies $W^TW = I_{p-1}$, $WW^T = P_0$ and $W^T\Sigma^{-1}W = diag(\xi_i)$. Finally, $b = \tau_a^2(1+\tau_a^2U^T\Sigma^{-1}U)^{-1/2} diag\{(1+\tau_a^2\xi_i)^{-1/2}\} W^T\Sigma^{-1}U$.

The proof of Theorem 3 is in Appendix A.3. Similar to the case under the null, although the exact distribution of the LRT under the alternative has no closed form, it can be simulated using a very fast algorithm that involves simple arithmetic operations on p independent standard normal random variables. Based on Theorem 2 and 3, we can in fact calculate the power function for the LRT for the whole alternative space, given a specific meta-analysis scenario with certain study size p and precision per study $(\sigma_1^2, \ldots, \sigma_p^2)$. This allows us to conduct a comprehensive power calculation for the LRT in section 3.3.

We also consider the special case with constant standard errors across studies, i.e., $\sigma_i^2 = \sigma^2$ for $i = 1, 2, \ldots, p$, under which the exact distribution of the LRT can be simplified.

Corollary 1. *When $\sigma_i^2 = \sigma^2$ for $i = 1, 2, \ldots, p$, the LRT_p under $H_0 : \mu = 0, \tau^2 = 0$ can be simplified as*

$$LRT_p \overset{d}{=} \left(R - p - p\log\frac{R}{p} \right) I\left(R > p\right) + u^2,$$

where $R = \sum_{i=1}^{p-1} w_i^2 \sim \chi_{p-1}^2$ and u^2 is a χ_1^2 random variable independent of R. LRT_p under

$H_a : \mu = \mu_a, \tau^2 = \tau_a^2$ can be simplified as

$$LRT_p \stackrel{d}{=} \left(R_{\tau_a} - p - p \log \frac{R_{\tau_a}}{p} \right) I(R_{\tau_a} > p) + \left\{ \frac{p^{1/2}\mu_a}{\sigma} + u \left(1 + \frac{\tau_a^2}{\sigma^2} \right)^{1/2} \right\}^2$$

where $R_{\tau_a} = (1 + \tau_a^2/\sigma^2)\, R = (1 + \tau_a^2/\sigma^2) \sum_{i=1}^{p-1} w_i^2$ follows a scaled χ^2_{p-1} random variable.

The proof of Corollary 1 is provided in Appendix A.4. Han and Eskin (2011) provided simulation-based tabulated values for calculating p-values under the assumption on common σ_i^2 (or sample size). We now provide the exact distribution of the LRT both under the null and under the alternative in this special setting, which is very easy and fast to simulate as it only depends on two independent χ^2 random variables.

The asymptotic distribution of the LRT_p in (3.3) is not standard χ^2_2 due to the null value of τ^2 being on the boundary of its parameter space. It has a $0.5\chi^2_1 + 0.5\chi^2_2$ asymptotic distribution under the null hypothesis based on the results in Self and Liang (1987). Figure 3.1 shows the QQ-plots of the $0.5\chi^2_1 + 0.5\chi^2_2$ mixture versus the finite sample distribution of LRT_p in (3.3) with $p = 5, 10, 50, 100$ and common σ_i^2. The quantiles are generated by 500,000 simulations. The $0.5\chi^2_1 + 0.5\chi^2_2$ mixture is a conservative approximation for the finite sample distribution. For example, the 99% quantile is 8.29 for $0.5\chi^2_1 + 0.5\chi^2_2$ distribution and 7.27 for the finite sample distribution with $p = 5$. Figure 3.1 also illustrates the probability mass that unconstrained MLE $\hat{\tau}^2$ equals 0 as a function of p. Unless p is very large, this probability is substantially larger than 0.5 as indicated by the asymptotic theory.

Based on Corollary 1, one can easily verify the asymptotic distribution of the LRT_p when p is large under H_0. It follows that $\bar{R} = R/p \to 1$ in probability and $p^{1/2}(\bar{R} - 1) \to N(0, 2)$ in distribution. Thus $pr(R > p) \to 0.5$. Notice that under $R > p$,

$$\begin{aligned} R - p - p \log \left(\frac{R}{p} \right) &= R - p - p \left(\frac{R-p}{p} - \frac{1}{2} \left(\frac{R-p}{p} \right)^2 \right) + o_p(1) \\ &= \frac{1}{2} \left[\sqrt{p}(\bar{R} - 1) \right]^2 + o_p(1) \to \chi^2_1, \end{aligned}$$

in distribution. Therefore $LRT_p \to 0.5\chi^2_1 + 0.5\chi^2_2$ in distribution. However, the number of studies p is small to medium in typical meta-analysis settings, so we always recommend using

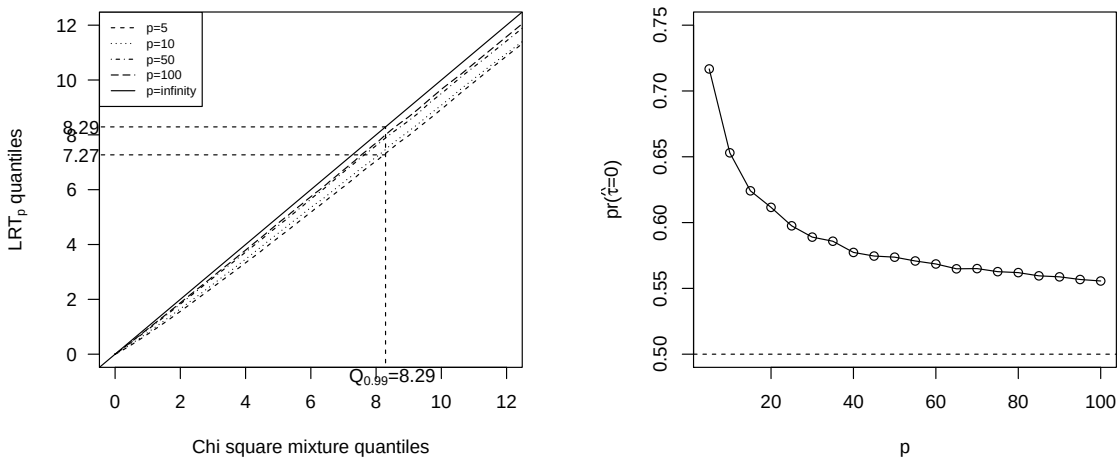


Figure 3.1: Left: QQ-plots for comparing a $0.5\chi^2_1 + 0.5\chi^2_2$ mixture versus the finite sample distribution of LRT_p with $p = 5, 10, 50, 100$ and common σ_i^2. Right: The probability that the unconstrained MLE $\hat{\tau}^2$ equals 0 as a function of p.

the exact LRT instead of asymptotics in practice.

3.2.2 Testing homogeneity based on random-effects model

In this subsection, we consider the problem of testing for homogeneity of effects across studies. Under the random-effects meta-analysis model, the null hypothesis of homogeneity can be represented as $H_{0,\tau^2} : \tau^2 = 0$. We consider the likelihood ratio test (LRT) statistics

$$LRT_{p,\tau^2} = \sup_{\tau^2} \left\{ Y^T P_0 Y - Y^T P_\tau^T V_\tau^{-1} P_\tau Y - \log |V_\tau| \right\},$$

and restricted likelihood ratio test (RLRT) statistics

$$RLRT_{p,\tau^2} = \sup_{\tau^2} \left\{ Y^T P_0 Y - Y^T P_\tau^T V_\tau^{-1} P_\tau Y - \log |V_\tau| - \log |Z^T V_\tau^{-1} Z| + \log |Z^T Z| \right\}.$$

Similar to the proof of Theorem 2 and 3, we can show that under $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$ the exact distributions of LRT_{p,τ^2} and $RLRT_{p,\tau^2}$ are

$$LRT_{p,\tau^2} \stackrel{d}{=} \sup_{\tau^2} \left\{ \sum_{i=1}^{p-1} \frac{(1+\tau_a^2\xi_i)\tau^2\xi_i}{1+\tau^2\xi_i} w_i^2 - \sum_{i=1}^{p} \log\left(1+\frac{\tau^2}{\sigma_i^2}\right) \right\},$$

$$RLRT_{p,\tau^2} \stackrel{d}{=} \sup_{\tau^2} \left\{ \sum_{i=1}^{p-1} \frac{(1+\tau_a^2\xi_i)\tau^2\xi_i}{1+\tau^2\xi_i} w_i^2 - \sum_{i=1}^{p} \log\left(1+\frac{\tau^2}{\sigma_i^2}\right) - \log\left(\sum_{i=1}^{p} \frac{1}{\sigma_i^2+\tau^2}\right) + \log\left(\sum_{i=1}^{p} \frac{1}{\sigma_i^2}\right) \right\},$$

which depend on the value of τ_a^2, but not on μ_a. Their null distributions are a special case when $\tau_a^2 = 0$.

3.3 Simulations

In this section, we conduct comprehensive power comparison of the exact likelihood ratio tests based on random-effects models versus fixed-effects testing procedures for testing the global null hypothesis. We also compare the power of the exact likelihood ratio tests based on random-effects models versus Cochran's Q statistic for testing homogeneity.

3.3.1 Testing the global null

The fixed-effects model assumes a common effect across studies. It can be represented as

$$M_f : \; X_i = \mu + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2), \quad i = 1, \ldots, p.$$

It is straightforward to show that the likelihood ratio test statistic (equivalent to Wald and score statistics) for testing $H_{0f} : \mu = 0$ takes the form of

$$T_f = \frac{\left(\sum_{i=1}^{p} \sigma_i^{-2} X_i\right)^2}{\sum_{i=1}^{p} \sigma_i^{-2}}.$$

Under H_{0f}, T_f follows a standard χ_1^2 distribution. Under $H_a : \mu \sim N(\mu_a, \tau_a^2)$, T_f follows a scaled non-central chi-square distribution, $c\chi_1^2(k)$, where $c = 1 + \tau_a^2 \sum_i \sigma_i^{-4} / \sum_i \sigma_i^{-2}$ and $k = \mu_a^2 \left(\sum_i \sigma_i^{-2}\right)^2 / \left(\sum_i \sigma_i^{-2} + \tau_a^2 \sum_i \sigma_i^{-4}\right)$. The power function of T_f takes the form $1 - F_{\chi_1^2(k)}\left\{F_{\chi_1^2}^{-1}(1-\alpha)/c\right\}$.

We consider the alternative hypothesis $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$ with μ_a and τ_a ranging from 0 to 2. We compare the LRT for testing the global null from fixed-effects and random-effects models. We use Theorem 2 to simulate $100,000$ samples from the null distribution of LRT_p in (3.3) to calculate the critical value, and then we use Theorem 3 to simulate $100,000$ samples from the alternative distribution to calculate the power of random-effects LRT.

We vary the number of studies $p = 5, 10, 20, 50$. We randomly generate study-specific variances σ_i^2 to follow an inverse gamma distribution with mean 1 and standard deviation equal to 0 ($\sigma_i^2 = 1$ for $i = 1, \ldots, p$), 0.5 and 2. Figure 3.2 shows the power heat map under different study sizes for the common study variance ($\sigma_i^2 = 1$ for $i = 1, \ldots, p$). When τ_a is small, the power for fixed-effects and random-effects model are similar. However, when τ_a is large, indicating a large heterogeneity between studies, the random-effects model has larger power than the fixed-effects model. The results for σ_i^2 with standard deviation 0.5 and 2 are shown in Figure 3.3 and 3.4, and they are similar to Figure 3.2.

3.3.2 Comparing with the test for the average effect

There are many procedures for testing the average effect $H_0' : \mu = 0$ under random-effects models. Guolo and Varin (2017) reviewed some testing procedures for H_0' and found out that the Hartung and Knapp method (Hartung and Knapp, 2001a,b; Knapp and Hartung, 2003), the signed profile log-likelihood ratio statistic and the Skovgaard's statistic (Guolo, 2012) can preserve type one error, regardless of the number study size. Although $H_0' : \mu = 0$ is a different hypothesis than the global null hypothesis of $H_0 : \mu = 0, \tau^2 = 0$, we can still compare the testing procedures for H_0' with our proposed LRT for H_0 by treating H_0' as a working model/hypothesis. We conduct comprehensive power comparisons between the proposed random-effects global LRT and the three testing procedures for H_0'. We vary the study size p to be 5, 10 and 20, and choose the study variability σ_i^2 randomly from exponential(1) distribution. Figure 3.5 shows the power difference between the random-effects global LRT and the other three testing procedures for $H_0' : \mu = 0$ over a range of alternatives $\mu = \mu_a$ and $\tau = \tau_a$. The random-effects global LRT has higher power than the other three testing

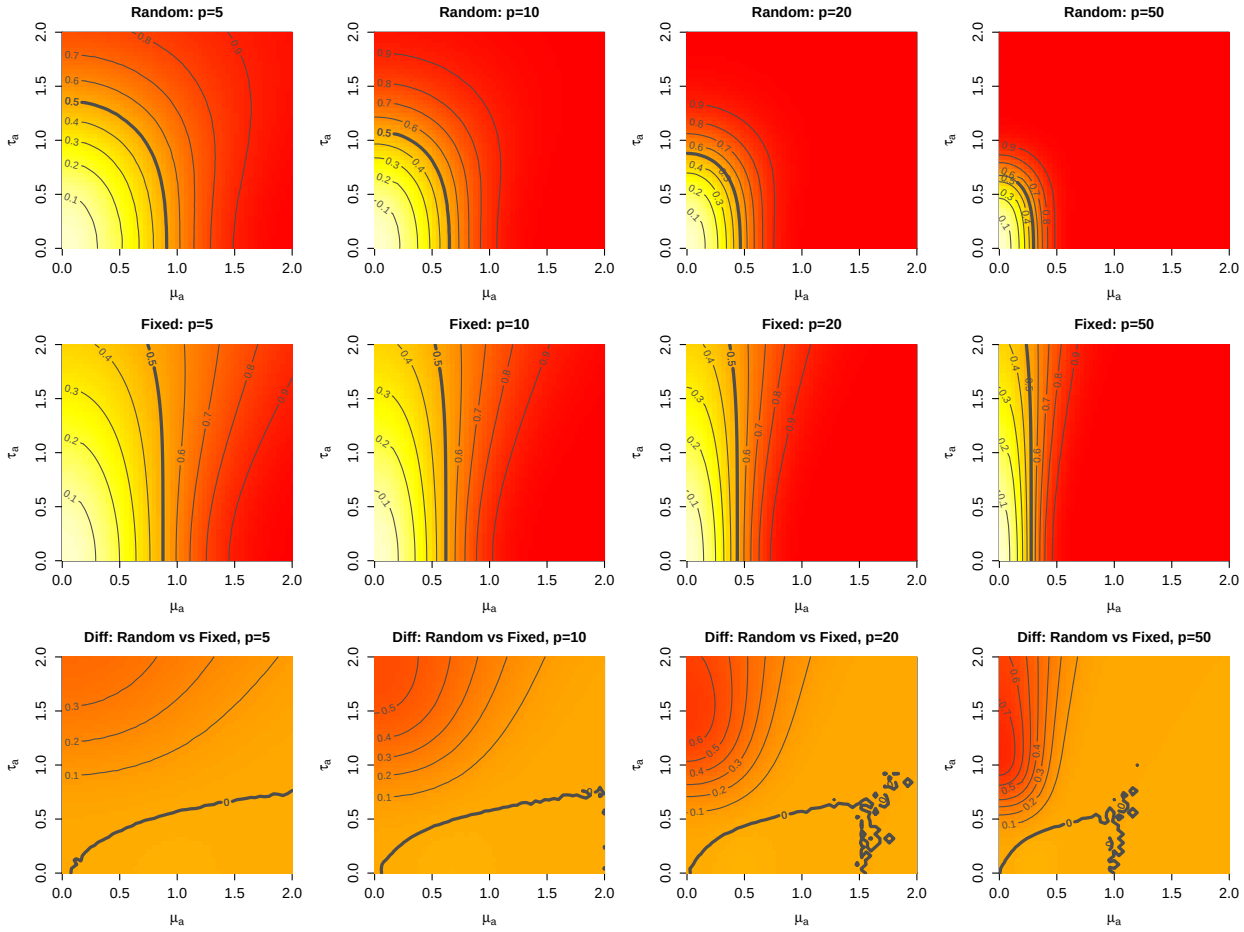


Figure 3.2: Power functions for tests based random-effects models (first row) and fixed-effects models (second row). The third row shows the power difference between the two testing procedures.

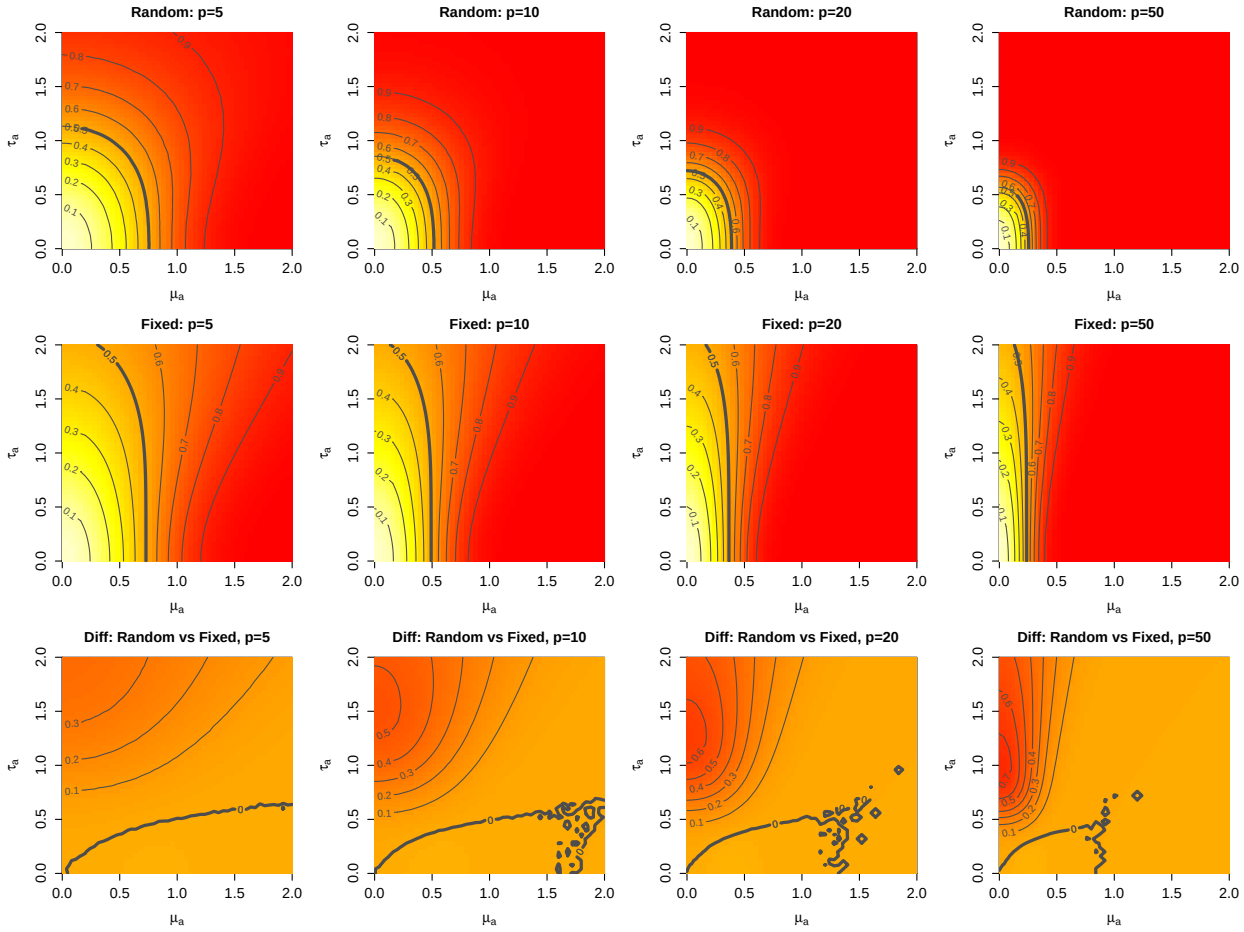


Figure 3.3: Power comparison for the random-effects model and fixed-effects model over a wide range of alternatives, where σ_i^2 is from the inverse gamma distribution with mean of 1 and standard deviation of 0.5. For each plot, the x-axis and y-axis are the mean and standard deviation of the alternative hypothesis. The third row shows the power difference between the two models.

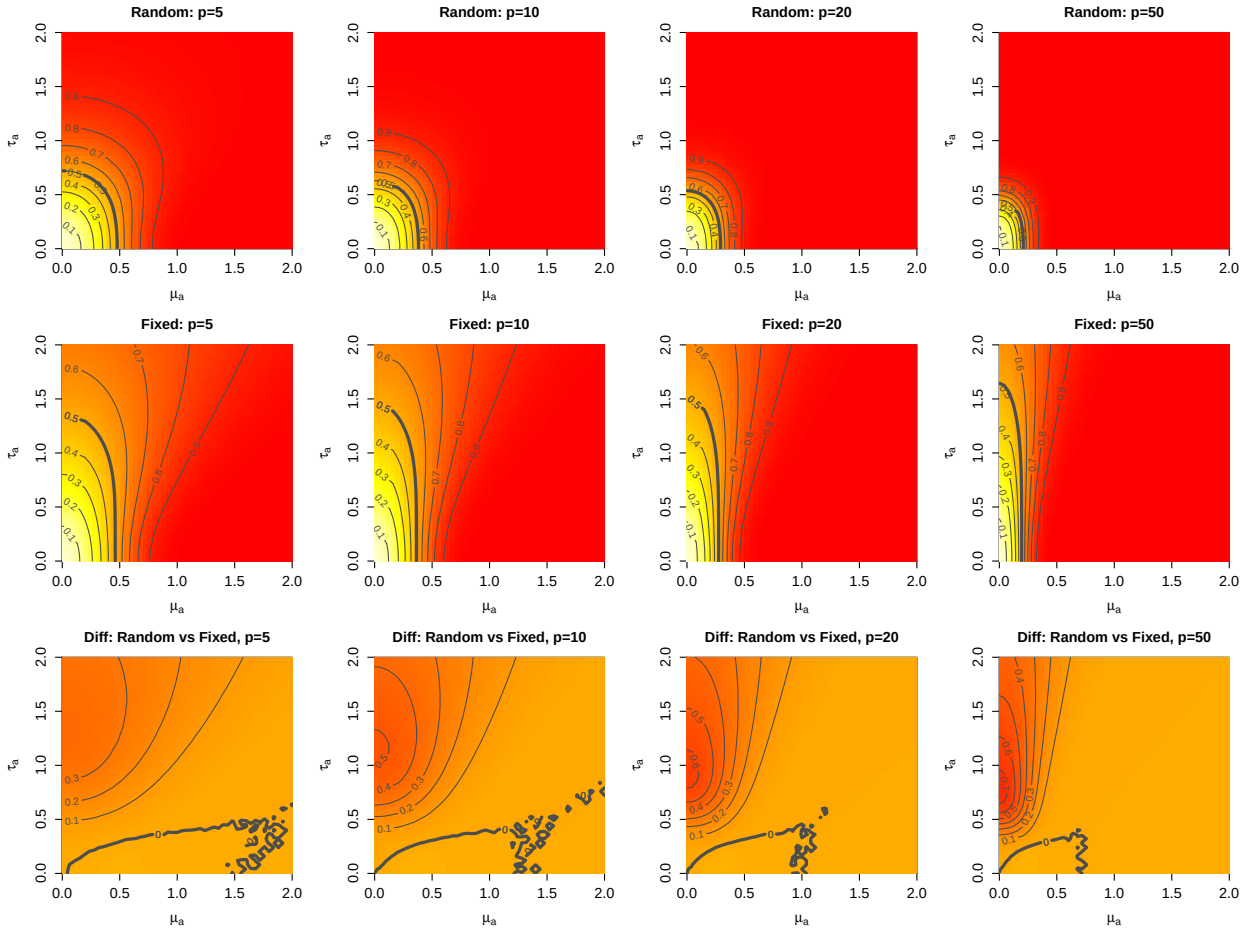


Figure 3.4: Power comparison for the random-effects model and fixed-effects model over a wide range of alternatives, where σ_i^2 is from the inverse gamma distribution with mean of 1 and standard deviation of 2. For each plot, the x-axis and y-axis are the mean and standard deviation of the alternative hypothesis. The third row shows the power difference between the two models.

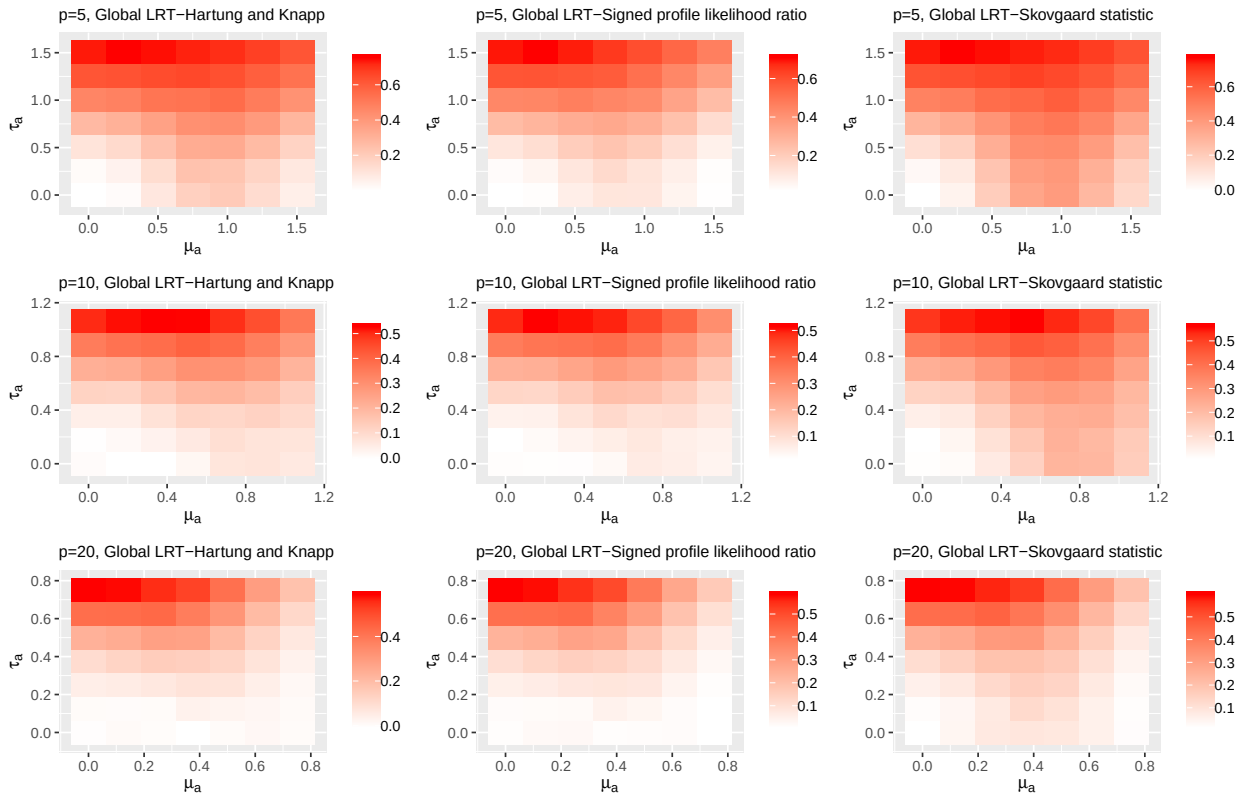


Figure 3.5: Power differences between the random-effects global LRT and the other testing procedures for $H_0' : \mu = 0$ over a wide range of alternatives, where σ_i^2 is from exponential(1) distribution.

procedures across all alternatives. Since the other three testing procedures are testing a different hypothesis $H_0' : \mu = 0$ compared to $H_0 : \mu = 0, \tau^2 = 0$, it is expected that the global LRT will have higher power. But it is interesting to see that when there is no heterogeneity, i.e. $\tau_a^2 = 0$, the global LRT still have higher power. Figure 3.6 show that the fixed-effects global test also has higher power than the other three testing procedures in all scenarios. We also used the same study variance $\sigma_i^2 = 1, i = 1, \ldots, p$, and the results are similar.

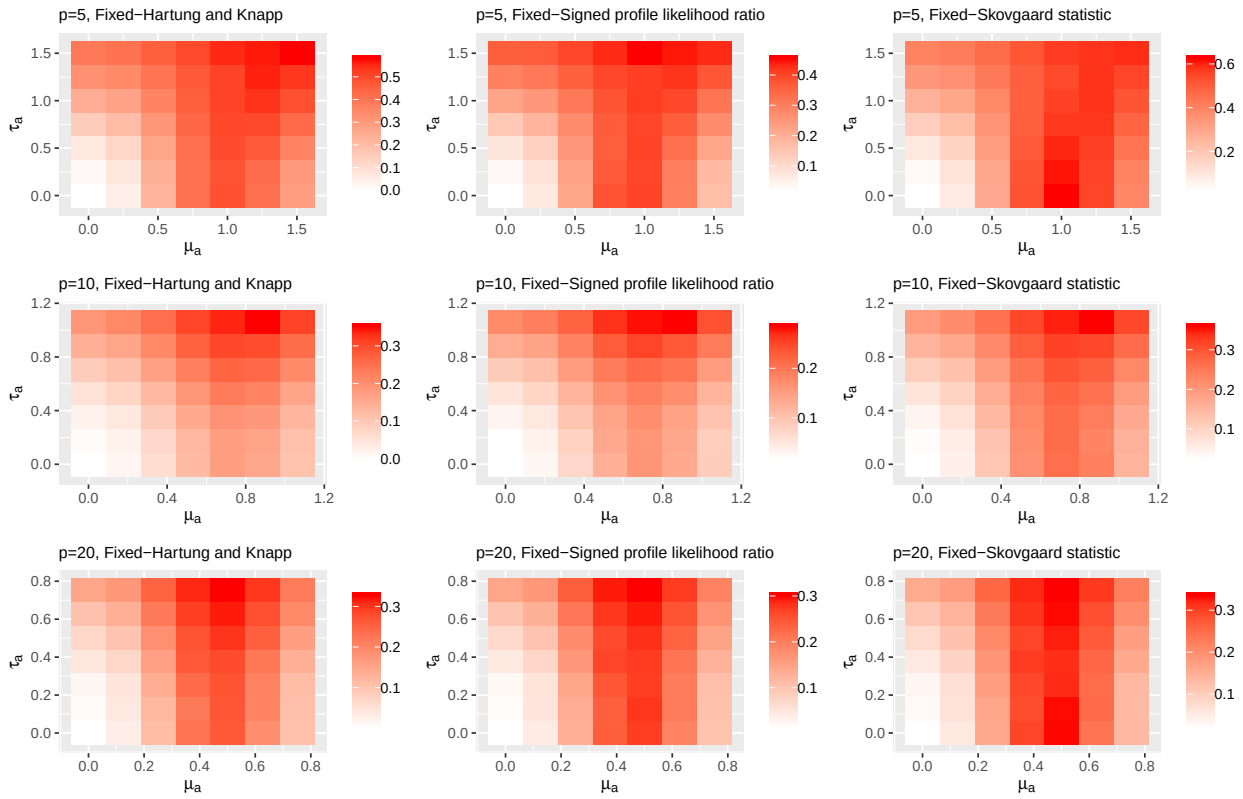


Figure 3.6: Power differences between the fixed-effects global test and the other testing procedures for $H_0' : \mu = 0$ over a wide range of alternatives, where σ_i^2 is from exponential(1) distribution.

3.3.3 Testing homogeneity

A popular test for testing homogeneity is Cochran's Q statistic (Cochran, 1937). Under the model $X_i \sim N(\beta_i, \sigma_i^2), i = 1, \ldots, p$, we test the null $\beta_1 = \ldots = \beta_p$ with Q statistic

$$Q = \sum_{i=1}^{p} \sigma_i^{-2}(X_i - \bar{X})^2,$$

where $\bar{X} = \sum_{i=1}^{p} w_i X_i$ and $w_i = \sigma_i^{-2} / \sum_{j=1}^{p} \sigma_j^{-2}$. Let $w = (w_1, \ldots, w_p)^T$, $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, and $A = (I_p - w1^T)\Sigma^{-1}(I_p - 1w^T)$, where 1 denotes a vector of 1's with length p. Then Cochran's Q statistic can be expressed as $Q = X^T AX$.

Under the null $\beta_1 = \ldots = \beta_p$, Q follows a χ^2_{p-1} distribution. Under $H_a : X_i \sim N(\mu_a, \sigma_i^2 + \tau_a^2)$, Q follows a weighted sum of χ^2_1 distribution $\sum_i \lambda_i \chi^2_1$, where λ_i's are eigenvalues of the matrix $A(\Sigma + \tau_a^2 I_p)$. We can approximate this distribution by a non-central chi square distribution $\chi^2_k(\delta)$ based on the method in Liu et al (2009).

We vary the study size $p = 5, 10, 20, 50$ and choose the study variability to be the same $\sigma_i^2 = 1$, or to follow exponential distribution with rate parameter 1, or an evenly spaced sequence between 0 and 1 (increment is $1/(p+1)$ excluding 0). Figure 3.7 shows the power comparison for Cochran's Q statistic and the LRT. When $\sigma_i^2 = 1$, $i = 1, \ldots, p$, the two tests have similar power. When σ_i^2, $i = 1, \ldots, p$, come from exponential distribution or a sequence between 0 and 1, LRT has higher power than Cochran's Q test when τ_a is small and moderate but for large τ_a, Cochran's Q test has higher power. The proposed LRT has advantages to detect heterogeneity when the study size is moderate or large ($p \geq 20$). But when the p is small, the difference between LRT and Cochran's Q is small. We also calculated the power of restricted likelihood ratio test (RLRT), and it is very similar to the power of LRT, thus omitted here.

3.4 Application

Men who have sex with men (MSM) are at high risk for HIV infection, particularly in low-income countries. Male circumcision was shown to reduce the risk of female-to-male HIV

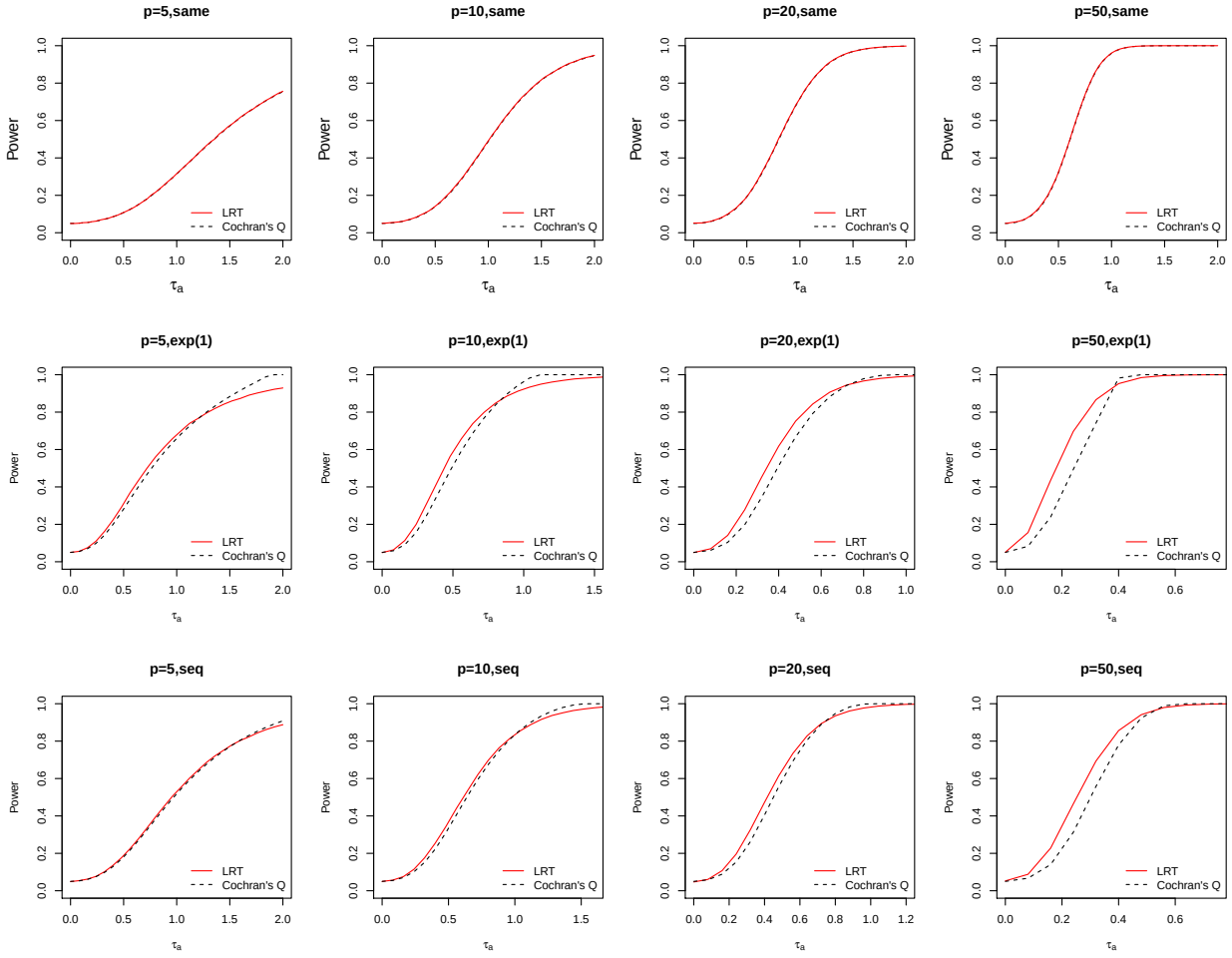


Figure 3.7: Power comparison between Cochran's Q statistic and LRT.

transmissions among heterosexual men from three randomized controlled trials conducted in Africa (Auvert et al, 2005; Bailey et al, 2007; Gray et al, 2007). However, the effect of circumcision to protect MSM is not clear. According to two meta-analyses in 2008 (Millett et al, 2008)and 2011 (Wiysonge et al, 2011), non-significant associations between HIV infection and circumcision was found in more than 20 observational studies. From a meta-analysis in 2018 (Sharma et al, 2018), circumcision was found significantly associated with the reduction of HIV infection in 18 observational studies. Understanding the association of circumcision and HIV infection is important for conducting evidence-based HIV prevention strategies.

Recently Yuan et al (2019) reviewed 45 published results before March 8, 2018 to study the association between circumcision and HIV and other sexually transmitted infections (STIs) among men who have sex with men (MSM). Log odds ratios and the corresponding standard errors are used for the strength of association. We conduct the global LRT and homogeneity LRT by generating 10^6 samples to approximate their null distribution. For the global null hypothesis, fixed-effects global test has a p-value of 0.865, while the proposed random-effects global test has a p-value of $< 10^{-6}$. The reason that they give very different results is because one study with null effect has very small standard error, thus higher weights, which makes the fixed-effects global test fail to reject the null. This suggests that fixed-effects global test can be very sensitive to the standard errors and potentially lose power. Cochran's Q statistic has a p-value of $< 10^{-16}$ and the global LRT has a p-value $< 10^{-6}$, indicating a significant amount of heterogeneity. Using the random-effects approach in DerSimonian and Laird (1986) (DL), we estimated the log odds ratio to be -0.258 with a 95% CI of $(-0.395, -0.120)$ and a p-value of 0.0002. It suggests that circumcision has an effect on preventing HIV transmission among MSM.

The association between circumcision and HIV may differ by the study characteristics. However, it is not possible to conduct multivariate meta-regression because only a small number of studies reported complete study characteristics. Therefore, we conduct subgroup meta-analyses in Table 3.1, where we show the p-values for random-effects global LRT, fixed-

effects global test, random-effects homogeneity LRT, and Cochran's Q test for homogeneity. For example, for 20 high-income countries, the fixed-effects global test has a p-value of 0.062 and the random-effects global test has a p-value of 0.072. It suggests that there is no heterogeneity and no effect of circumcision on HIV prevention. Furthermore, we find that the homogeneity LRT has a p-value of 0.09 and Cochran's Q test has a p-value 0.037. The DL approach estimates the log odds ratio of studies from high-income countries to be -0.008 with a 95% CI of $(-0.107, 0.090)$ and a p-value of 0.87. These results agree with the global test. On the other hand, for 23 low-income countries, the fixed-effects global test has a p-value of 4.8×10^{-13}, while the random-effects global LRT has a p-value of $< 10^{-6}$. Thus we have enough evidence to reject the global null. Both LRT for homogeneity and Cochran's Q test suggest that there is a significant amount of heterogeneity in low-income countries with p-values $< 10^{-6}$ and 1.1×10^{-11}, respectively. The DL approach estimates the log odds ratio to be -0.538 with a 95% CI of $(-0.891, -0.184)$ and a p-value of 0.003, suggesting that despite heterogeneity among low-income countries, there is still statistical evidence that circumcision is protective for HIV transmission.

3.5 Extension to meta-regression

It is straightforward to extend previous results to the random-effects meta-regression model,

$$M_s : X_i = \tilde{Z}_i^T \mu + \beta_i + \epsilon_i, \quad \beta_i \sim N(0, \tau^2), \quad \epsilon_i \sim N(0, \sigma_i^2), \quad i = 1, \ldots, p, \tag{3.4}$$

where $\tilde{Z}_i = (\tilde{z}_{i1}, \ldots, \tilde{z}_{iq})^T$ is the study-specific covariates, q is the number of covariates, and μ is q-dimensional vector. When $\tilde{Z}_i$ is a column of ones, the meta-regression model simplifies to the meta-analysis model. Denote by $Y_i = X_i/\sigma_i$ and $Z_i = \tilde{Z}_i/\sigma_i$, so model M_s in (3.4) is equivalent to

$$Y_i = Z_i \mu + \beta_i' + \epsilon_i', \ \beta_i' \sim N(0, \tau^2/\sigma_i^2), \ \epsilon_i' \sim N(0, 1),$$

		N	Global LRT	Fixed	Homogeneity LRT	Cochran's Q
Income	High	20	0.072	0.062	0.09	0.037
	Low	23	<E-6	4.8E-13	<E-6	1.1E-11
LGBT rights	Support	26	<E-6	0.6840	<E-6	<E-16
	Neither/oppose	18	1E-5	0.024	1E-4	0.0018
WHO region	Americas	16	0.0783	0.0656	0.08	0.0036
	Europe	4	0.9140	0.8890	1	0.5730
	Western Pacific	12	0.0098	0.7170	7E-4	0.0272
	Southeast Asia	5	1E-5	6.3E-5	0.0012	9E-4
	Africa	5	<E-6	<E-16	0.0011	0.0027
Sexual behavior	Insertive anal sex	8	<E-6	2.7E-12	2E-4	0.0014
	Receptive anal sex	6	0.033	0.030	0.11	0.13
Mean or median age	>29 years	22	8E-5	0.5860	<E-6	3.2E-6
	<=29 years	21	<E-6	9.5E-13	<E-6	7.7E-7
Study design	Cohort	9	0.3740	0.3990	0.123	0.3770
	Case-control	2	0.2670	0.2980	0.068	0.0671
	Cross-sectional	34	<E-6	1.0000	<E-6	<E-16
Recruitment setting	Clinic-based	9	0.1030	0.0549	0.022	0.0076
	Non-clinic-based	34	<E-6	0.0003	<E-6	<E-16
HIV assessment	Laboratory test	39	<E-6	0.6680	<E-6	<E-16
	Self-reported	6	0.0015	0.1530	2E-4	0.0006
Circumcision assessment	Genital examination	10	0.0629	0.0573	0.018	0.5740
	Self-reported	29	<E-6	0.0244	<E-6	<E-16
Circumcision	<= 34%	24	<E-6	<E-16	4.6E-6	2.8E-11
	> 34%	19	9E-5	0.1510	<E-6	1.5E-6
Consistent condom use	<= 38%	6	0.0200	0.0146	1	0.5190
	> 38%	4	0.0004	0.0031	0.0035	0.0171
HIV-testing history	<= 53%	9	<E-6	7.9E-12	<E-6	1.4E-6
	> 53%	9	0.5700	0.5310	0.097	0.2330
Total		45	<E-6	0.8650	<E-6	<E-16

Table 3.1: Subgroup analysis to assess the association between circumcision and HIV among men who have sex with men. We show the p-values for the random-effects global LRT, fixed-effects global test, the random-effects homogeneity LRT, and fixed-effects Cochran's Q test.

or in vector form $Y \sim N(Z\mu, V_\tau)$ with $Z = (Z_1, \ldots, Z_p)^T$ and $V_\tau = I_p + \tau^2\Sigma^{-1}$. We are interested in testing the null hypothesis:

$$H_{0r} : \mu_{k+1} = \mu_{k+1}^0, \ldots, \mu_q = \mu_q^0, \tau^2 = 0,$$

versus the alternative:

$$H_{ar} : \mu_{k+1} \neq \mu_{k+1}^0 \text{ or } \ldots \text{ or } \mu_q \neq \mu_q^0 \text{ or } \tau^2 \neq 0.$$

We partition the fixed-effects parameters $\mu = (\mu_1^T|\mu_2^T)^T$, where under the null hypothesis H_{0r}, $\mu_2 = \mu_2^0$ are known. We also partition the corresponding design matrix into $Z = (Z_1|Z_2)$. Similar to the previous arguments, we can show that likelihood ratio test for H_{0r} versus H_{ar} has the form

$$LRT_p' = \sup_{\tau^2} \left\{-Y^T P_\tau^T V_\tau^{-1} P_\tau Y - \log|V_\tau|\right\} + (Y - Z_2\mu_2^0)^T S_1 (Y - Z_2\mu_2^0)$$

where $P_\tau = I_p - Z(Z^T V_\tau^{-1} Z)^{-1} Z^T V_\tau^{-1}$ and $S_1 = I_p - Z_1(Z_1^T Z_1)^{-1} Z_1^T$.

Theorem 4. *Under the random-effects meta-analysis model M_s in* (3.4)*, the exact distribution of the LRT statistic under H_{0r} can be characterized as*

$$LRT_p' \overset{d}{=} \sup_{\tau^2} \left\{ \sum_{i=1}^{p-q} \frac{\tau^2\xi_i}{1+\tau^2\xi_i} w_i^2 - \sum_{i=1}^{p} \log\left(1 + \frac{\tau^2}{\sigma_i^2}\right) \right\} + \sum_{i=1}^{q-k} u_i^2 + (Z_2\mu_2^0)^T Z_2\mu_2^0,$$

where $(w_1, \ldots, w_{p-q}, u_1, \ldots, u_{q-k})$ are $(p-k)$ independent standard Gaussian random variables, $(\xi_1, \ldots, \xi_{p-q})$ are the $(p-q)$ nonzero eigenvalues of $\Sigma^{-1}P_0$ and $P_0 = I_p - Z(Z^T Z)^{-1} Z^T$.

3.6 Discussion

In this chapter, we proposed likelihood ratio test for the global null and homogeneity under random-effects meta-analysis model, and compared the power of the two tests to the fixed-effects global test and Cochran's Q test, respectively. Our proposed global LRT has higher power than the fixed-effects global test when there is heterogeneity. In addition, our proposed homogeneity LRT has advantages and disadvantages compared to Cochran's Q test,

depending on the amount of heterogeneity and study size. We furter extended the LRT of the global null to the meta-regression setting.

For the random-effects meta-analysis, both the average effect and heterogeneity are of interest. We can apply two step procedures to differentiate three types of hypothesis: the average effect $\mu = 0$, the homogeneity $\tau^2 = 0$, and the global null $H_0 : \mu = 0, \tau^2 = 0$. The proposed random-effects global LRT can be used as a first step to test the global null to see if there is an average effect or heterogeneity. If the global null is rejected, then apply the homogeneity LRT to see if there is heterogeneity or apply the procedures in Guolo and Varin (2017) to see if the average effect is zero. Note that conventional meta-analysis focuses on testing the average effect $H_0' : \mu = 0$. In our simulations, we find that the conventional meta-analysis testing procedures for H_0' lack power to detect association compared to the tests aimed at the global null (the random-effects global LRT and the fixed-effects global test) even under the scenario where heterogeneity is absent.

Traditional meta-analysis focuses on combining different studies to assess the association between exposure and outcome. Recently Bhattacharjee et al (2012) proposed to combine the association between a single exposure with multiple related outcomes in the analysis of genetic association studies of heterogeneous traits. For example, one can combine the association between one variant with different types of cancer to investigate the association between this variant and cancer in general. Since the effects of individual variants on different types of cancer can be heterogeneous, a random-effects meta-analysis model is a natural choice. Global LRT developed in this chapter can be applied in this setting to detect this association. It would be interesting to compare the subset-based approach in the fixed-effects framework in Bhattacharjee et al (2012) with the global LRT in this chapter. We leave this for future work.

Chapter 4

HYPOTHESIS TESTING FOR BOTH FIXED EFFECTS AND RANDOM EFFECTS IN MIXED-EFFECTS MODELS

4.1 Introduction

Mixed-effects models (Breslow and Clayton, 1993) are widely used in various settings, including meta-analysis, longitudinal and correlated data analysis, nonparametric smoothing, genetic association and sequencing studies, and survival analysis. In many scenarios, it is of interest to study a set of many parameters simultaneously and the main objective is to test whether all parameters are zero. Treating them as fixed effects may suffer from low power, especially when the number of parameters is large or there is little information to estimate each parameter reliably. An alternative approach is to treat these parameters as random effects that follow an underlying common distribution, e.g., Gaussian distribution. Under the random effects model, testing global null can be specified as testing a small number of parameters, e.g., mean and variance of random effects.

Consider a likelihood function (or a pseudo/partial/composite likelihood) denoted as $L(\alpha, \beta; y)$, where y is data, α is a m dimensional nuisance parameter and β is a p dimensional parameter of interest. The main objective is to test the global null hypothesis $H_0 : \beta = 0$. Instead of testing these p fixed parameters, one can use a mixed-effects model to combine information and potentially improve power, $\beta = Z\mu + \delta$, where Z is a $p \times q$ design matrix, μ is a q dimensional fixed-effects parameter and δ is a p dimensional random effect. Typically we assume δ follows a multivariate Gaussian distribution with mean 0 and covariance matrix $\tau^2\Omega$, where Ω is known and τ^2 is a random-effects variance component. Testing the hypothesis that β is zero is equivalent to testing $H_0 : \mu = 0, \tau^2 = 0$. We now provide some motivating examples that are special cases of this general statistical problem.

4.1.1 Example 1: random-effects meta-analysis

Meta-analysis aims to aggregate information from multiple studies to assess the association between an outcome and an exposure variable. Suppose we have n independent studies. Let β_i be the true effect of interest from the ith study, X_i and σ_i be its estimate and corresponding standard error, respectively, $i = 1, \ldots, n$. A random-effects meta-analysis model assumes

$$X_i = \beta_i + \epsilon_i, \quad \beta_i \sim N(\mu, \tau^2), \quad \epsilon_i \sim N(0, \sigma_i^2), \quad i = 1, \ldots, n, \tag{4.1}$$

where the true effect β_i follows a Gaussian distribution with mean μ and variance τ^2. Under this model, the null hypothesis of no effects among all studies can be represented as $H_0 : \mu = 0, \tau^2 = 0$ (Han and Eskin, 2011; Wu et al, 2020), in which μ is a fixed-effects parameter and τ^2 is a random-effects variance component.

4.1.2 Example 2: set-based genetic variant association analysis

In genetic association studies, testing the association between individual genetic variants (e.g., single-nucleotide polymorphisms, SNPs) and disease outcomes generally has low power when the genetic variants are rare. In order to improve power, one can aggregate variants by a prior defined set (e.g., transcripts, genes, pathways), and test the association for the set of variants instead of individual variants (Lee et al, 2012; Sun et al, 2013; Lin et al, 2013; Su et al, 2017, 2018; He et al, 2018). Let Y_i, $i = 1, \ldots, n$ be the disease outcome (continuous or discrete), $X_i \in \mathbf{R}^m$ be potential confounders, $G_i \in \mathbf{R}^p$ be genotypes of p variants aggregated by a prior defined set, where the genotypes are coded as $0, 1, 2$, representing the number of minor alleles. The model is

$$g\{E(Y_i)\} = X_i^T \alpha + G_i^T \beta, \tag{4.2}$$

where g is a link function, e.g. $g(x) = x$ for continuous trait and $g(x) = \exp(x)/(1 + \exp(x))$ for binary trait. To leverage information across p variants, Sun et al (2013) modeled $\beta = (\beta_1, \ldots, \beta_p)^T$ as $\beta_j = Z_j^T \mu + \delta_j$, where $Z_j \in \mathbf{R}^q$ are characteristics of jth variant, $\mu \in \mathbf{R}^q$ are fixed-effects regression coefficients and δ_j is a random effect that follows a Gaussian

distribution with mean 0 and variance τ^2. Here μ is a fixed-effects parameter and τ^2 is a random-effects variance component. Under the random-effects model, testing the null hypothesis of no association between disease trait and sets of genetic variants is equivalent to testing $H_0 : \mu = 0, \tau^2 = 0$.

4.1.3 Example 3: time-varying treatment effect in extended Cox model

Another example arises in randomized clinical trials where the outcome is time-to-event and treatment effects might vary over time. Saegusa et al (2014) considered an extended Cox proportional hazard model

$$\lambda(t|X, Z) = \lambda_0(t) \exp\{X^T\alpha + Z\beta(t)\}, \tag{4.3}$$

where $\lambda(t|X, Z)$ is the hazard function for the covariate X and the treatment indicator Z, and $\lambda_0(t)$ is the baseline hazard function. In addition, α is regression coefficients for X and $\beta(t)$ is the time-varying coefficient for treatment Z. The null hypothesis that treatment has no effect for survival at any time can be represented as $H_0 : \beta(t) = 0$ for any $t \geq 0$. To flexibly model the time-varying treatment effect, we can use B-splines or smoothing splines to represent $\beta(t)$ as

$$\beta(t) = \mu + \sum_{i=1}^{p} \delta_i G_i(t), \tag{4.4}$$

where $G_i(t), i = 1, \ldots, p$ are basis functions and δ_i is the corresponding coefficient. Since we introduce extra p parameters $(\delta_1, \ldots, \delta_p)$, in order to avoid overfitting, we penalize the smoothness of $\beta(t)$ by

$$\int \{\beta'(t)\}^2 \, dt = \int \left\{ \sum_i \delta_i G_i'(t) \right\}^2 dt = \delta^T \Sigma \delta,$$

where Σ is a $p \times p$ matrix with (i, j)th element to be $\int G_i'(t) G_j'(t) dt$. By exploiting the connection between penalized spline and random-effects model (Saegusa et al, 2014), we can treat $\delta \sim N(0, \tau^2 \Sigma^{-1})$. Here μ is a fixed-effects parameter and τ^2 is a random-effects variance component. The null hypothesis of no time-varying treatment effect can be represented as

$H_0 : \mu = 0, \tau^2 = 0.$

The goal of this chapter is to test the global null hypothesis that the set of parameters are all zeros, which is equivalent to testing both fixed-effects parameters and random-effects variance components under the random-effects model. However, testing variance components is nonstandard, because their true value is on the boundary of its parameter space under the null. Except for a few special cases, the asymptotic distribution of the likelihood ratio test is complicated and intractable. Alternatively, score tests have gained popularity for testing the variance component because they only require fitting the model under the null and can reduce the computational complexity (Lin and Breslow, 1996; Lin, 1997; Lin and Zhang, 1999; Verbeke and Molenberghs, 2003; Zhang and Lin, 2003). Under a set of regularity conditions, score statistics for two parameters usually have (asymptotically) multivariate Gaussian distribution, and thus can be combined into a χ^2 test statistic. However, the asymptotic behavior for the score statistic of the variance component is nonstandard and also correlated with the score statistic for fixed-effects parameters.

Several methods have been proposed for combining score statistics. Fisher's (Fisher, 1925) and Tippett's (Tippett et al, 1931) procedures are classical methods to combine independent tests based on p-values. In the context of genetic variant association studies, there are a flurry of methods that utilize linear combination of scores statistics for fixed-effects parameters parameters and variance components. Lee et al (2012) and Su et al (2017) considered the test statistic to be the minimum of the p-values from a grid of linear combinations. But the score statistics proposed in Lee et al (2012) are correlated, leading to a restricted linear combination space while the score statistics proposed in Su et al (2017) are independent. He et al (2018) chose the test statistic to be the maximum of the standardized linear combination of two independent score statistics. Su et al (2017) and Su et al (2018) considered adaptively weighted linear combination where the weights come from a transformation of p-values from two independent tests. Even though some of these methods use linear combinations of score statistics, they are data-adaptive and non-linear methods in the sense that the weight

depends on the data and their null distributions involve numerical integration or grid search.

To test the global null hypothesis, we propose novel conmbinations of two asymptotic independent score statistics for fixed-effects parameters and random-effects variance components. The independence property is desirable for the linear combination of score statistics $S_\rho = \rho S_\mu + (1-\rho)S_{\tau^2}$, where S_μ and S_{τ^2} are score statistics for fixed-effects parameters and random-effects variance components, respectively, and $0 \leq \rho \leq 1$. We consider the alternative hypothesis $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$. For a fixed ρ, the null and alternative distributions of S_ρ are mixture of chi-squared distributions, which is easy to approximate (Liu et al, 2009). Thus it is convenient to calculate the analytical power of each linear combination given an alternative hypothesis. It is important to notice that for different linear combinations, there will be trade-offs of power across the alternative space. For example, S_{τ^2}, corresponding to $\rho = 0$, is only powerful when the true $\tau^2 \neq 0$ and is not powerful when the true $\tau^2 = 0$. On the other hand, S_μ, corresponding to $\rho = 1$, is only powerful when the true $\mu \neq 0$. An ideal choice of ρ would balance the power well across all types of alternatives. To this end, we propose Bayes and Minimax decision rules to find an omnibus linear combination that performs reasonably well in terms of power across the alternative space.

Note that our procedure is different than previous proposals in many aspects. First, our procedure is not data-adaptive in the sense that the weight is not a function of the data. Once the optimal weight is chosen by Bayes or Minimax decision rules, it is treated as fixed. Second, our procedure directly compares power from different linear combinations and tries to balance the trade-off. The optimal weight is chosen such that power is relatively high across all alternatives. In practice, researchers' prior information about plausible alternatives can be incorporated into our testing procedures to directly maximize power. Third, the null distribution is equivalent to the sum of weighted chi-square distributions and does not require numerical integration or grid search. Thus the computational cost is relatively low.

The organization of this chapter is as follows. In section 4.2, we present a hypothesis testing framework for generalized linear mixed-effects models. Next we show three examples: meta-analysis in section 4.3, set-based genetic variant association test in section 4.4 and time-

varying treatment effect for extended Cox model in section 4.5. In section 4.6, we review procedures to combine independent score statistics. In section 4.7, we conduct comprehensive simulations of power comparisons for the three examples. In section 4.9, we conclude with some suggestions for practice.

4.2 General framework for generalized linear mixed-effects model

For simplicity of presentation, we present the main results in the context of generalized linear mixed-effects model. It is easy to extend these results to any type of mixed-effects model, i.e. example 3 of the spline-based extended Cox model.

Let $(Y_i, X_i, G_i), i = 1, \ldots, n$, be independent and identically distributed data, where for ith subject, Y_i is the outcome, $X_i \in \mathbf{R}^m$ are potential confounders, and $G_i \in \mathbf{R}^p$ are covariates of interest. We consider a generalized linear model

$$g\{E(Y_i|X_i, G_i)\} = X_i^T\alpha + G_i^T\beta,$$

where g is the link function, $\alpha \in \mathbf{R}^m$ and $\beta \in \mathbf{R}^p$ are regression coefficients for X_i and G_i, respectively. We are interested in testing whether $\beta = 0$. We can combine information by modeling $\beta_j, j = 1, \ldots, p$, as

$$\beta_j = Z_j^T\mu + \delta_j,$$

where $Z_j \in \mathbf{R}^q$ is the design matrix for jth coefficient, $\mu \in \mathbf{R}^q$ is the coefficient, and δ_j is the effect unexplained by Z_j. We assume that $\delta_j, j = 1, \ldots, p$, follows a Gaussian distribution of mean 0 and variance τ^2. Let $Y = (Y_1, \ldots, Y_n)^T$, $X = (X_1^T, \ldots, X_n^T)^T$, $G = (G_1^T, \ldots, G_n^T)^T$, $Z = (Z_1^T, \ldots, Z_p^T)^T$ and $\delta = (\delta_1, \ldots, \delta_p)^T$. Let $n \times q$ matrix GZ be the product of G and Z. Then we can write the model in matrix form

$$g\{E(Y|X, G)\} = X\alpha + (GZ)\mu + G\delta.$$

Testing $\beta = 0$ is equivalent to test $H_0 : \mu = 0, \tau^2 = 0$. Denote the marginal log likelihood by $\ell(\alpha, \mu, \tau^2)$. The score statistic for μ under H_0 is

$$U_\mu = \left.\frac{\partial \ell}{\partial \mu}\right|_{\mu=0,\tau^2=0} = (GZ)^T(Y - \tilde{\pi}),$$

where $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_n)^T$, $\tilde{\pi}_i = g^{-1}\{X_i^T\tilde{\alpha}\}$ and $\tilde{\alpha}$ is the MLE of α under H_0. Naively, if we define the score statistic for τ^2 in a similar way as $\partial\ell/\partial\mu|_{\mu=0,\tau^2=0}$, it will be correlated with U_μ. Linear combinations of two correlated statistics will restrict the search space and potentially lose power. Therefore, we modify the score for τ^2 such that it is calculated under $\tau^2 = 0$ without the restriction on $\mu = 0$, and the resulting score for τ^2 will be independent of U_μ. The score for τ^2 under $\tau^2 = 0$ is

$$S_{\tau^2} = \frac{\partial\ell}{\partial\tau^2}\bigg|_{\tau^2=0} = (Y-\hat{\pi})^T GG^T(Y-\hat{\pi}),$$

where $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)^T$, $\hat{\pi}_i = g^{-1}\{X_i^T\hat{\alpha} + (G_i^T Z)\hat{\pi}\}$, and $(\hat{\alpha}, \hat{\pi})$ is MLE of (α, π) under $\tau^2 = 0$. In Appendix A.6, we provide another approach of constructing score statistic for τ^2 under H_0 and we prove the asymptotic equivalence between the two approaches.

Using Taylor expansion, we have the approximations $Y - \tilde{\pi} \approx P_1(Y-\pi)$ and $Y - \hat{\pi} \approx P_2(Y-\pi)$, where $P_1 = I - DX(X^TDX)^{-1}X^T$, $P_2 = I - DM(M^TDM)^{-1}M^T$, and $M = (X, GZ)$. In addition, $\pi = (\pi_1, \ldots, \pi_n)^T$ is the mean of Y under H_0 with $\pi_i = g^{-1}(X_i^T\alpha)$, and D is the covariance matrix of Y under H_0. For example, $D = \sigma^2 I_n$ for continuous outcome, and $D = \mathrm{diag}(d_1, \ldots, d_n)$ for binary outcome where $d_i = \pi_i(1-\pi_i)$. Denote by $U_\mu = (GZ)^T P_1(Y-\pi)$ and $U_{\tau^2} = G^T P_2(Y-\pi)$. Thus the score for μ is asymptotically equivalent to

$$S_\mu \approx (Y-\pi)^T K_1 (Y-\pi), \tag{4.5}$$

where $K_1 = P_1^T(GZ)\mathrm{cov}(U_\mu)^{-1}(GZ)^T P_1$ and $\mathrm{cov}(U_\mu) = (GZ)^T\left\{D - DX(X^TDX)^{-1}X^TD\right\}(GZ)$. Under H_0, asymptotically S_μ follows a χ^2_q distribution. The score for τ^2 is asymptotically equivalent to

$$S_{\tau^2} \approx (Y-\pi)^T K_2 (Y-\pi), \tag{4.6}$$

where $K_2 = P_2^T GG^T P_2$. Under H_0, asymptotically S_{τ^2} follows a mixture of chi-squared distribution $\sum_i \lambda_i \chi^2_{1,i}$, where λ_i is the eigenvalues of $G^T\left\{D - DM(M^TDM)^{-1}M^TD\right\}G$.

We consider the linear combination of two score statistics $S_\rho = \rho S_\mu + (1-\rho)S_{\tau^2}$, where $0 \le \rho \le 1$. Thus $S_\rho \approx (Y-\pi)^T K_\rho (Y-\pi)$, where $K_\rho = \rho K_1 + (1-\rho)K_2$. Since K_1 and K_2 may

have different magnitude, we normalize them such that $\text{tr}(K_1) = \text{tr}(K_2)$ before combining them. In practice we can estimate D by $\tilde{D} = \text{diag}\{\tilde{d}_1, \ldots, \tilde{d}_n\}$ where $\tilde{d}_i = \sum_{i=1}^{n}(Y_i - \tilde{\pi}_i)^2/n$ for continuous trait and $\tilde{d}_i = \tilde{\pi}_i(1 - \tilde{\pi}_i)$ for binary trait, $i = 1, \ldots, n$. $\tilde{\pi}_i = g^{-1}(X_i^T\tilde{\alpha})$ is the estimated mean outcome under the null and $\tilde{\alpha}$ is the MLE of α under H_0. Note that we also substitute D by $\tilde{D}$ in the definition of K_ρ.

For fixed ρ, we can characterize the distribution of S_ρ in the following proposition.

Proposition 1. *(a) Under H_0, S_μ and S_{τ^2} are (asymptotically) independent.*

(b) Under H_0, S_ρ asymptotically follows a mixture of central chi-squared distribution $\sum_{i=1}^{n}\lambda_i\chi^2_{1i}$, where λ_i are eigenvalues of $D^{1/2}K_\rho D^{1/2}$ for $i = 1, \ldots, n$.

(c) Denote by π_a and D_a the mean and covariance matrix of Y under the alternative hypothesis $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$, respectively. Under H_a, $S_\rho \sim \sum_{i=1}^{n}\eta_i\chi^2_{1,i}(b_i^2)$ asymptotically, where η_i is the eigenvalue of $\tilde{K}_\rho = D_a^{1/2}K_\rho D_a^{1/2}$ and $b = (b_1, \ldots, b_n)^T = U^T D_a^{-1/2}(\pi_a - \pi)$ by the spectral decomposition of $\tilde{K}_\rho = U\Lambda U^T$.

The proof of proposition 1 is in Appendix A.7. In addition, we prove the independence of scores in the general setting in Appendix A.10. We can approximate the distribution of S_ρ by a scaled non-central chi-squared distribution $c\chi^2_l(k)$ based on the moment-matching method in Liu et al (2009), in which the unknown parameters are determined by the first four cumulants of the quadratic form. The details of the approximation are in Appendix A.8. The approximation allows fast and comprehensive power calculations for a fixed ρ. Denote by $F_{0\rho}$ and $F_{a\rho}$ the cumulative distribution function of S_ρ under H_0 and H_a, respectively. The rejection region at α level is $S_\rho > F_{0\rho}^{-1}(1-\alpha)$. Thus the power for rejecting the alternative $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$ is $G(\mu_a, \tau_a^2; \rho) = 1 - F_{a\rho}\left\{F_{0\rho}^{-1}(1-\alpha); \mu_a, \tau_a^2, \rho\right\}$. Note that the analytical power formula is available if we can write the alternative mean μ_a and covariance matrix D_a explicitly.

4.2.1 Optimal linear combinations of score statistics

Let $G(\mu_a, \tau_a^2; \rho)$ be the power based on S_ρ for testing $H_0 : \mu = 0, \tau^2 = 0$ versus $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$. If the alternative (μ_a, τ_a^2) is known in advance, the optimal weight should achieve the maximum power, namely,

$$\rho_a = \underset{0 \le \rho \le 1}{\operatorname{argmax}}\, G(\mu_a, \tau_a^2; \rho). \tag{4.7}$$

We call this power the oracle power since it is obtained as if the alternative is known. However, in practice, the alternative hypothesis is generally not known.

To illustrate that different linear combinations will have different power behavior in the alternative space, we show the power heat map from Example 2 in Figure 4.1. If we choose a small ρ favoring S_{τ^2} over S_μ, we would perform well in terms of power in regions where μ_a is small and τ_a^2 is large but not in regions where μ_a is large and τ_a^2 is small. On the other hand, If we choose a large ρ favoring S_μ over S_{τ^2}, we would have high power in regions where μ_a is large and τ_a^2 is small but not in regions where μ_a is small and τ_a^2 is large. We can see a trade-off of power for different ρ, and a good choice of ρ would balance the power well in all the alternatives. Thus we propose two methods to find an omnibus linear combination that performs reasonably well in terms of power across all the alternatives, via Bayes and minimax decision rules. First we define the risk function.

Definition 1 (Risk)**.** *The risk function is defined as the power loss of S_ρ compared to the oracle procedure S_{ρ_a} for each fixed alternative (μ_a, τ_a^2),*

$$R(\mu_a, \tau_a^2; \rho) = G(\mu_a, \tau_a^2; \rho_a) - G(\mu_a, \tau_a^2; \rho), \tag{4.8}$$

where under a specific alternative (μ_a, τ_a^2), ρ_a is the optimal weight, and $G(\mu_a, \tau_a^2; \rho_a)$ is the maximum achievable power within the linear combination family.

Definition 2 (Bayes rule)**.** *If we assume that the alternative parameter (μ_a, τ_a^2) follows a prior distribution $f(\mu_a, \tau_a^2)$, Bayes decision rule will minimize the average power loss,*

$$\rho_b = \underset{0 \le \rho \le 1}{argmin} \int R(\mu_a, \tau_a^2; \rho) f(\mu_a, \tau_a^2) d\mu_a d\tau_a^2. \tag{4.9}$$

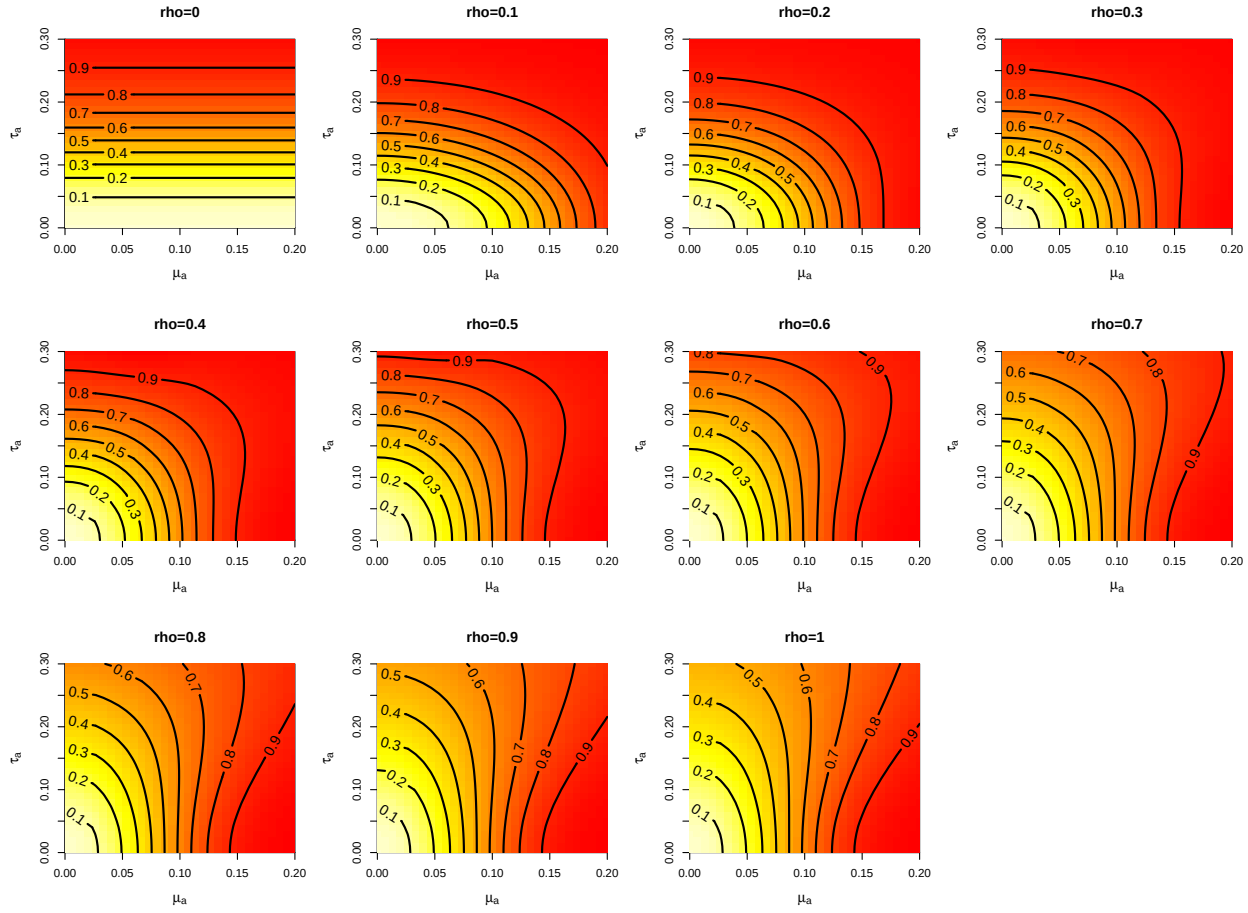


Figure 4.1: Genetic association test with continuous trait: power heat map and contours of a specific linear combination $S_\rho = \rho S_\mu + (1-\rho)S_{\tau^2}$ with $\rho = 0, 0.1, \ldots, 0.9, 1$. Each panel shows the power function across a wide range of alternatives, $\mu_a \in [0, 0.2]$ and $\tau_a \in [0, 0.3]$, for fixed rho.

It is easy to see that Bayes rule is equivalent to maximize the average power,

$$\rho_b = \underset{0\le\rho\le1}{\operatorname{argmax}} \int G(\mu_a, \tau_a^2; \rho) f(\mu_a, \tau_a^2) d\mu_a d\tau_a^2.$$

Definition 3 (Minimax rule)**.** *Minimax decision rule will minimize the worst case scenario power loss,*

$$\rho_m = \underset{0\le\rho\le1}{argmin} \max_{(\mu_a,\tau_a^2)} \left\{ R(\mu_a, \tau_a^2; \rho) \right\}. \tag{4.10}$$

Researchers' prior information about the plausible alternatives can be incorporated into our testing procedures to directly maximize power through either Bayes or Minimax procedure. One can also define a similar minimax procedure but using relative power instead of power difference

$$\rho'_m = \underset{0\le\rho\le1}{\operatorname{argmin}} \max_{(\mu_a,\tau_a^2)} \left\{ 1 - \frac{G(\mu_a, \tau_a^2; \rho)}{G(\mu_a, \tau_a^2; \rho_a)} \right\}.$$

Throughout this chapter we would use the definition in equation (4.10) for minimax procedure. Note that there are many other procedures to combine independent tests shown in section 4.6, but they are not motivated by maximizing power for alternative hypothesis. Next we show three examples of combining score statistics.

4.3 Example 1: random-effects meta-analysis

Let $Y_i = X_i/\sigma_i$ and the model in equation (4.1) is equivalent to $Y \sim N_n(\mu Z, V_{\tau^2})$, where $Z = \Sigma^{-1/2}1$, $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ and $V_{\tau^2} = I_n + \tau^2\Sigma^{-1}$. Under $H_0 : \mu = 0, \tau^2 = 0$, $Y \sim N_n(0, I_n)$. The log likelihood function is $\ell(\mu, \tau^2) = -\frac{1}{2}(Y - \mu Z)^T V_{\tau^2}^{-1}(Y - \mu Z) - \frac{1}{2}\log|V_{\tau^2}|$. The score for μ under $H_0 : \mu = 0, \tau^2 = 0$ is

$$\left.\frac{\partial \ell}{\partial \mu}\right|_{\mu=0,\tau^2=0} = -Z^T Y,$$

or equivalently

$$S_\mu = Y^T Z(Z^T Z)^{-1} Z^T Y = Y^T K_1 Y, \tag{4.11}$$

where $K_1 = Z(Z^T Z)^{-1} Z$. We modify the score for τ^2 such that it is evaluated under $\tau^2 = 0$ without the restriction of $\mu = 0$. The reason is that the resulting score for τ^2 will be

independent of S_μ. The score for τ^2 under $\tau^2 = 0$ is

$$S_{\tau^2} = \left.\frac{\partial \ell}{\partial \tau^2}\right|_{\tau^2=0} = (Y - Z\tilde{\mu})^T \Sigma^{-1} (Y - Z\tilde{\mu}) = Y^T K_2 Y. \tag{4.12}$$

where $\tilde{\mu} = (Z^T Z)^{-1} Z^T Y$ is the MLE of μ under $\tau^2 = 0$, and $K_2 = (I_n - K_1)\Sigma^{-1}(I_n - K_1)$. Note that S_μ and S_{τ^2} are independent under H_0. Linear combination of scores can be written as $S_\rho = \rho S_\mu + (1-\rho) S_{\tau^2} = Y^T K_\rho Y$, where $K_\rho = \rho K_1 + (1-\rho) K_2$. Under H_0, $Y \sim N_n(0, I_n)$, thus $S_\rho \sim \sum_{i=1}^n \lambda_i \chi^2_{1,i}$ where λ_i is the eigenvalues of K_ρ. Under H_a : $\mu = \mu_a, \tau^2 = \tau_a^2$, we have $Y \sim N_n(\mu_a Z, V_{\tau_a})$ where $V_{\tau_a} = \text{diag}(1 + \tau_a^2/\sigma_i^2) = I_n + \tau_a^2 \Sigma^{-1}$. We can write $S_\rho = Y^T V_{\tau^2}^{-1/2} V_{\tau^2}^{1/2} K_\rho V_{\tau^2}^{1/2} V_{\tau^2}^{-1/2} Y = W^T \tilde{K}_\rho W$, where $\tilde{K}_\rho = V_{\tau^2}^{1/2} K_\rho V_{\tau^2}^{1/2}$ and $W = V_{\tau^2}^{-1/2} Y \sim N_n(\mu_a V_{\tau_a}^{-1/2} Z, I_n)$. From the eigenvalue decomposition of $\tilde{K}_\rho = U \Lambda U^T$, we have that under H_a, $S_\rho \sim \sum_{i=1}^n \eta_i \chi^2_{1,i}(b_i^2)$, where η_i is diagonal elements of Λ and $b = (b_1, \ldots, b_n)^T = \mu_a U^T V_{\tau_a}^{-1/2} Z$.

We can calculate the power under $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$ using a grid of μ_a and τ_a, and calculate the optimal weight from Bayes and Minimax decision rules. In section 4.7.1, we show the simulation results.

4.4 Example 2: set-based genetic variant association analysis

The general framework in section 4.2 is the same as the situation in the set-based genetic association testing, thus the results can be directly applied. In this section, we present a more detailed discussion. From equation (4.2) and $\beta_j = Z_j^T \mu + \delta_j$, we have the model

$$g\{E(Y_i)\} = X_i^T \alpha + (G_i^T Z)\mu + G_i^T \delta, \tag{4.13}$$

where $Z = (Z_1, \ldots, Z_p)^T$ is a $p \times q$ matrix of variant characteristics and $\delta = (\delta_1, \ldots, \delta_p)^T$ are the random effects. The representation in equation (4.13) encompasses several well-known models in set-based association test as special cases. If we set $\delta = 0$ and Z to be a $p \times 1$ vector of 1, (4.13) reduces to the model for burden test (Morgenthaler and Thilly, 2007; Li and Leal, 2008; Madsen and Browning, 2009). If we set $\mu = 0$, (4.13) reduces to the model of sequence kernel association test (SKAT) with linear kernel (Wu et al, 2011). Note that

there is no random intercept in equation (4.13), unlike the case in longitudinal studies where random intercepts are commonly used. The reason is that we are interested in modeling the coefficients β_j, $j = 1, \ldots, p$.

It is possible to include weights in the testing framework by assuming the individual variant effects δ_j to follow a distribution of mean 0 and variance $w_j\tau^2$, where w_j is a non-negative weight, $j = 1, \ldots, p$. Under this assumption, the score statistic for τ^2 is

$$wS_{\tau^2} = (Y - \hat{\pi})^T GWG^T(Y - \hat{\pi}),$$

where $W = \text{diag}(w_1, \ldots, w_p)$. Usually we can choose the weight w_j to be a decreasing function of the observed minor alleles frequency (MAF) f_j, $j = 1, \ldots, p$. For example, we can choose the weight based on the beta function, $\sqrt{w_j} = f_j^{a-1}(1 - f_j)^{b-1}$ where $a \geq 0$ and $b \geq 0$.

Following the same logic in section 4.2, we can construct score statistics as in (4.5) and (4.6) and consider the linear combination $S_\rho \approx (Y - \pi)^T K_\rho (Y - \pi)$ where $K_\rho = \rho K_1 + (1 - \rho)K_2$. From the results in proposition 1, under H_0, S_ρ follows a mixture of chi-squared distributions, $\sum_{i=1}^n \lambda_i \chi^2_{1,i}$, where λ_i are eigenvalues of $D^{1/2}K_\rho D^{1/2}$. For the alternative distribution, we distinguish fixed alternative and random alternative. In section 4.7.2, we show the simulation results.

4.4.1 Fixed alternative

Under fixed alternative $H_a : \beta = \beta_a$, the mean of Y is $\pi_a = g^{-1}(X\alpha + G\beta_a)$, and the covariance matrix of Y is $D_a = \sigma^2 I_n$ for continuous trait and $D_a = \text{diag}(v_1, \ldots, v_n)$ for binary trait where $v_i = \pi_{ai}(1 - \pi_{ai})$. Note that S_μ and S_{τ^2} are independent under H_0 but not necessarily independent under H_a. Then under H_a, S_ρ follows a mixture of non-central chi-squared distribution $\sum_{i=1}^n \eta_i \chi^2_{1i}(b_i^2)$, where η_i are eigenvalues from the eigenvalue decomposition $D_a^{1/2} K_\rho D_a^{1/2} = U\Lambda U^T$ and $b = (b_1, \ldots, b_n)^T = U^T D_a^{-1/2}(\pi_a - \pi)$.

4.4.2 Random alternative

Consider the random alternative $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$, i.e. $\beta \sim N(\mu_a, \tau_a^2)$. For continuous trait, under H_a, $Y \sim N(\pi_a, D_a)$ where $\pi_a = X\alpha + (GZ)\mu_a$, $D_a = \tau_a^2 GG^T + \sigma^2 I_n$. Thus, under H_a, $S_\rho = Y^T K_\rho Y \sim \sum_{i=1}^n \eta_i \chi_{1,i}^2(b_i^2)$ where η_i is the eigenvalues from the eigenvalue decomposition $D_a^{1/2} K_\rho D_a^{1/2} = U\Lambda U^T$ and $b = (b_1, \ldots, b_n)^T = U^T D_a^{-1/2}(GZ)\mu_a$. For binary trait, the mean π_a and the covariance matrix D_a do not have analytical forms. Therefore, it is difficult to obtain analytical power under random alternative for binary trait.

4.5 Example 3: time-varying treatment effect in extended Cox model

Although this example does not belong to the generalized linear mixed-effects model discussed in section 4.2, the idea of combining scores statistics can be generalized to any type of mixed-effects models.

In this section, we adapt the framework of testing time-varying treatment effect under extended Cox model in Saegusa et al (2014). Let $Y = \min(T, C)$ be the minimum of time to event T and censoring time C, $\Delta = I(T \leq C)$ be the indicator for a failure event, Z be the treatment indicator and $X \in \mathbf{R}^m$ be the covariate vectors. The observed data consists of n independent and identically distributed copies $\{(Y_i, \Delta_i, X_i, Z_i)\}$, $i = 1, \ldots, n$. Let $t_1 < t_2 < \ldots < t_r$ be the ordered distinct observed failure times. Consider the extended Cox model in equation (4.3) and the spline representation in equation (4.4). Since partial likelihood is evaluated at the observed failure times $t_1, \ldots, t_r$ only, we can define $\beta = (\beta(t_1), \ldots, \beta(t_r))^T = 1\mu + G\delta$, where $1 \in \mathbf{R}^r$ is a vector of 1, $\delta = (\delta_1, \ldots, \delta_p)^T$ and G is $r \times p$ matrix with its (i, j) element to be $G_i(t_j)$. Under this model, the null can be expressed as $H_0 : \mu = 0, \delta = 0$. To avoid overfitting, we consider penalized partial log likelihood is

$$\ell_p\left(\alpha, \mu, \delta, \tau^2\right) = \ell\left(\alpha, \mu, \delta\right) - \frac{1}{2\tau^2}\delta^T \Sigma \delta,$$

where τ^2 is a tuning parameter that controls the smoothness of $\beta(\cdot)$, Σ is a $p \times p$ matrix with

(i,j)th element to be $\int G_i'(t)G_j'(t)dt$, and ℓ is the log partial likelihood

$$\ell\left(\alpha,\mu,\delta\right)=\sum_{i=1}^{n}\Delta_i\left[X_i^T\alpha+Z_i\beta(Y_i)-\log\left\{\sum_{j=1}^{n}\exp\left\{X_j^T\alpha+Z_j\beta(Y_j)\right\}I(Y_j\geq Y_i)\right\}\right].$$

For statistical inference, we can utilize the connection between penalized spline and random effects model. By treating δ as random effects with mean 0 and covariance matrix $\tau^2\Sigma^{-1}$, we have the model

$$\lambda(t|X,Z)=\lambda_0(t)\exp\left[X^T\alpha+Z\left\{\mu+\sum_{i=1}^{p}\delta_iG_i(t)\right\}\right],\quad \delta\sim N(0,\tau^2\Sigma^{-1}). \tag{4.14}$$

The resulting marginal partial log likelihood is

$$\ell_m(\alpha,\mu,\tau^2)=\log\left[\int_\delta\exp\left\{\ell(\alpha,\mu,\delta)\right\}\exp\left(-\frac{1}{2\tau^2}\delta^T\Sigma\delta\right)d\delta\right],$$

which can be used for inference. The null hypothesis can be expressed as $H_0:\mu=0,\tau^2=0$. The score for μ under H_0 is $1^T\tilde{U}$, where $\tilde{U}=\partial\ell(\tilde{\alpha},\mu=0,\delta=0)/\partial\beta$, and $\tilde{\alpha}$ is the maximum partial likelihood estimate of $\ell(\alpha,\mu=0,\delta=0)$, the Cox model without treatment. The kth element of $\tilde{U}$ is

$$\sum_{i=1}^{n}I(Y_i=t_k)\Delta_i\left\{Z_i-\frac{\sum_j Z_j\exp(X_j^T\hat{\alpha})I(Y_j\geq t_k)}{\sum_j\exp(X_j^T\hat{\alpha})I(Y_j\geq t_k)}\right\}.$$

The covariance matrix of $\tilde{U}$ is V. The standardized score statistic for μ is

$$S_\mu=\tilde{U}^T1(1^TV1)^{-1}1^T\tilde{U}. \tag{4.15}$$

The score statistic for τ^2 under H_0 is $\tilde{U}^TG\Sigma^{-1}G^T\tilde{U}$. However these two scores are not independent. We can modify the score statistic for τ^2 by projecting $\tilde{U}$ onto $1^T\tilde{U}$ and the modified score is

$$S_{\tau^2}=\tilde{U}^TW^TG\Sigma^{-1}G^TW\tilde{U} \tag{4.16}$$

where $W=I-V1(1^TV1)^{-1}1^T$ and W is constructed such that $W\tilde{U}$ and $1^T\tilde{U}$ are asymptotically independent, implying the asymptotic independence of S_μ and S_{τ^2} under H_0. The score from this construction is asymptotically equivalent to the score of τ^2 under

the working model $\tau^2 = 0$ without the restriction of $\mu = 0$, namely, $\hat{U}^T G\Sigma^{-1} G^T \hat{U}$ where $\hat{U} = \partial\ell(\hat{\alpha}, \mu = \hat{\mu}, \delta = 0)/\partial\beta$, and $\hat{\alpha}$ and $\hat{\mu}$ are the maximum partial likelihood estimate of $\ell(\alpha, \mu, \delta = 0)$.

Consider the linear combination of scores $S_\rho = \rho S_\mu + (1-\rho) S_{\tau^2} = \tilde{U}^T K_\rho \tilde{U}$, where $K_\rho = \rho K_1 + (1-\rho) K_2$, $K_1 = 1(1^T V 1)^{-1} 1^T$ and $K_2 = W^T G \Sigma^{-1} G^T W$. Under H_0, we have $\tilde{U} \sim N_n(0, V)$ and $S_\rho \sim \sum_{i=1}^n \lambda_i \chi^2_{1,i}$ asymptotically, where λ_i is the eigenvalues of $V^{1/2} K_\rho V^{1/2}$. Note that if $\rho = 1$ this test is the usual log-rank test, and if $\rho = 0$ this test is to test the proportional hazard.

For the alternative hypothesis and power, we consider fixed and random alternative shapes of $\beta(t)$. For the fixed alternative, we consider specific shape of $\beta(t)$ from previous literature.

4.6 Existing methods to combine independent tests

In this section, we survey existing methods to combine score statistics. These methods are non-linear combination methods. Except for the Fisher and Tippet combinations, these procedures were proposed in the set-based genetic association testing context (Example 2) in recent years. The Fisher and Tippet procedures for combining independent tests have been around for a long time, and were also used in testing time-varying treatment effects for time-to-event data (Saegusa et al, 2014).

Note that under the null, S_μ and S_{τ^2} are independent by construction. Throughout this section, we adopt the following notation. Let $F_{0\mu}$ and $f_{0\mu}$ be the cumulative distribution function and density function of S_μ under H_0, respectively. Let $F_{0\tau^2}$ be the cumulative distribution function of S_{τ^2} under H_0.

4.6.1 Tippett's procedure

Let $P_\mu = 1 - F_{0\mu}(S_\mu)$ and $P_{\tau^2} = 1 - F_{0\tau^2}(S_{\tau^2})$ be the p-values for S_μ and S_{τ^2}. Tippett's procedure (Tippett et al, 1931) used the minimum of the p-values as the test statistic, $T = \min(P_\mu, P_{\tau^2})$. The distribution function of T under H_0 is $P_0(T < t) = 1 - P_0(p_\mu \geq$

$t, p_{\tau^2} \geq t) = 1-(1-t)^2$, where we use the fact that p-values are uniformly distributed under the null. Thus at significance level α, the rejection region is $T < 1-(1-\alpha)^{1/2}$.

4.6.2 Fisher's procedure

Fisher's procedure (Fisher, 1925) used the test statistic $T = -2\log(P_\mu) - 2\log(P_{\tau^2})$. Under H_0, $T \sim \chi^2_4$. At significance level α, the rejection region is $T > F^{-1}_{\chi^2_4}(1-\alpha)$.

4.6.3 Linear combination that maximizes standardized test statistics (MaxStat)

Under H_0, $S_\mu \sim \chi^2_q$ and $S_{\tau^2} \sim \sum_{i=1}^n \lambda_i \chi^2_{1,i}$. Consider the linear combination $S(\rho) = \rho S_\mu + (1-\rho)S_{\tau^2}$. He et al (2018) chose to maximize the standardized linear combination

$$\frac{S(\rho)-\mu(\rho)}{\sigma(\rho)},$$

where $\mu(\rho) = \rho q + (1-\rho)\Lambda_1$, $\sigma(\rho) = \sqrt{2\rho^2 q + 2(1-\rho)^2\Lambda_2}$, $\Lambda_1 = \sum_{i=1}^{p-q}\lambda_i$ and $\Lambda_2 = \sum_{i=1}^{p-q}\lambda_i^2$. Some calculation yields that the maximum is achieved at

$$\rho^* = \underset{0\leq\rho\leq 1}{\operatorname{argmax}} \left\{ \frac{S(\rho)-\mu(\rho)}{\sigma(\rho)} \right\} = \begin{cases} \frac{\Lambda_2(S_\mu - q)}{q(S_{\tau^2}-\Lambda_1)+\Lambda_2(S_\mu - q)} & \text{if } S_\mu > q \text{ and } S_{\tau^2} > \Lambda_1 \\ 0 & \text{if } (S_\mu - q)/\sqrt{2q} \leq (S_{\tau^2}-\Lambda_1)/\sqrt{2\Lambda_2} \\ 1 & \text{if } (S_\mu - q)/\sqrt{2q} > (S_{\tau^2}-\Lambda_1)/\sqrt{2\Lambda_2} \end{cases}.$$

The test statistic is

$$T = \frac{S(\rho^*)-\mu(\rho^*)}{\sigma(\rho^*)} = \begin{cases} \sqrt{\frac{(S_\mu - q)^2}{2q} + \frac{(S_{\tau^2}-\Lambda_1)^2}{2\Lambda_2}} & \text{if } S_\mu > q \text{ and } S_{\tau^2} > \Lambda_1 \\ \max\left\{(S_\mu - q)/\sqrt{2q}, (S_{\tau^2}-\Lambda_1)/\sqrt{2\Lambda_2}\right\} & \text{otherwise} \end{cases}.$$

Note that ρ^* is a function of S_μ and S_{τ^2}. This procedure is data adaptive in the sense that the optimal weight ρ^* is a function of the data. Denote the cumulative distribution function of T by F_{0T} under H_0. Since S_μ and S_{τ^2} are independent, He et al (2018) proved that the cumulative distribution of T follows

$$F_{0T}(t) = P_0(T \leq t) = \begin{cases} F_{0\tau^2}(t_{\tau^2})F_{0\mu}(q) + \int_q^t F_{0\tau^2}(\delta(u))f_{0\mu}(u)du, & \text{if } t > 0 \\ F_{0\tau^2}(t_{\tau^2})F_{0\mu}(t_\mu), & \text{if } t \leq 0, \end{cases}$$

where $t_\mu = t\sqrt{2q} + q$, $t_{\tau^2} = t\sqrt{2\Lambda_2} + \Lambda_1$, and $\delta(u) = \sqrt{t^2 - \frac{(u-q)^2}{2q}}\sqrt{2\Lambda_2} + \Lambda_1$. Thus at significance level α, the rejection region is $T > F_{0T}^{-1}(1-\alpha)$.

4.6.4 Linear combination that minimizes p-values (MinPvalue)

Consider the linear combination of scores $S_\rho = \rho S_\mu + (1-\rho)S_{\tau^2}$. Denote its p-value by $P_\rho = 1 - F_{S_\rho}(S_\rho)$, where F_{S_ρ} is the cumulative distribution function of S_ρ. Lee et al (2012) and Su et al (2017) chose a test statistic that is the minimum of p-values,

$$T = \inf_{0\le\rho\le 1} P_\rho.$$

Note that the two score statistics considered in Lee et al (2012) are asymptotically correlated but the two scores considered in this chapter are independent asymptotically. Denote the observed value of T by $t = \inf_{0\le\rho\le 1} p_\rho$, where $p_\rho = 1 - F_{S_\rho}(s_\rho)$ and s_ρ is the observed value of S_ρ. Choose a grid of $\rho, \rho_1 < \ldots < \rho_b$, to evaluate T and t. We can calculate the cumulative distribution of T under H_0,

$$\begin{aligned}
P_0(T \le t) &= P_0\left(\inf_\rho \left(1 - F_{S_\rho}(S_\rho)\right) \le t\right) \\
&= P_0\left(1 - \sup_\rho F_{S_\rho}(S_\rho) \le t\right) \\
&= 1 - P_0\left(\sup_\rho F_{S_\rho}(S_\rho) \le 1 - t\right) \\
&= 1 - P_0\left(S_{\rho_1} \le F_{S_{\rho_1}}^{-1}(1-t), \ldots, S_{\rho_b} \le F_{S_{\rho_b}}^{-1}(1-t)\right).
\end{aligned}$$

Let $q_\rho(1-t) = F_{S_\rho}^{-1}(1-t)$. Since $S_\rho = \rho S_\mu + (1-\rho)S_{\tau^2}$, we simplify

$$P_0(T \le t) = 1 - E_0\left[P_0\left(S_{\tau^2} < \min_\rho\left\{\frac{q_\rho(1-t) - \rho S_\mu}{1-\rho}\right\}\right)\middle| S_\mu\right].$$

Let $\delta(x;t) = \min_\rho\left\{\frac{q_\rho(1-t)-\rho x}{1-\rho}\right\}$. Denote by F_{0T} the cumulative distribution function of T under H_0,

$$F_{0T}(t) = P_0(T \le t) = 1 - \int_0^\infty F_{0\tau^2}(\delta(x;t)) f_{0\mu}(x)dx.$$

Thus at significance level α, the rejection region is $T < F_{0T}^{-1}(\alpha)$.

4.6.5 Adaptively weighted linear combinations (Adaptive)

Su et al (2017) and Su et al (2018) used the adaptively weighted linear combination $T = \rho_\mu Z_\mu + \rho_{\tau^2} Z_{\tau^2}$, where $Z_\mu = -2\log(P_\mu)$ and $Z_{\tau^2} = -2\log(P_{\tau^2})$ are log transformed p-values. The weights ρ_μ and ρ_τ are determined by

$$\rho_\mu = \frac{Z_\mu}{\sqrt{Z_\mu^2 + Z_{\tau^2}^2}}, \quad \rho_{\tau^2} = \frac{Z_{\tau^2}}{\sqrt{Z_\mu^2 + Z_{\tau^2}^2}}.$$

Therefore the test statistic can be simplified as $T = Z_\mu^2 + Z_{\tau^2}^2$. Under the null, $Z_\mu \sim \chi_2^2$ and $Z_{\tau^2} \sim \chi_2^2$. Notice that $P_0(Z_\mu^2 < x) = 1 - \exp\left(-\frac{1}{2}\sqrt{x}\right)$. Thus the cumulative distribution of T under H_0 is

$$F_{0T}(t) = P_0(T < t) = P_0(Z_\mu^2 + Z_{\tau^2}^2 < t) = \int_0^t \left[1 - \exp\left(-\frac{1}{2}\sqrt{x}\right)\right] \frac{1}{4\sqrt{x}} \exp\left(-\frac{1}{2}\sqrt{x}\right) dx.$$

Thus at significance level α, the rejection region is $T > F_{0T}^{-1}(1-\alpha)$.

4.7 Simulation studies

Although analytical power formula for minimax procedure is available, analytical power formula for other existing combination methods is not available. Thus we use simulation based methods for power comparison.

4.7.1 Example 1: random-effects meta-analysis

We choose $n = 10$ and the study variabilities $\sigma_i^2, i = 1, \ldots, n$ are generated from exponential(1) distribution. We consider the alternative $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$, where μ_a consists of 30 equally spaced points between 0 and 1, and τ_a consists of 30 equally spaced points between 0 and 2. To calculate power, we simulate data from $X_i \sim N(\mu_a, \tau_a^2 + \sigma_i^2)$, and conduct tests with α level of 0.05. We repeat this process for 10^4 times to obtain an average empirical power under a particular (μ_a, τ_a^2). We compare S_μ, S_{τ^2}, Minimax and Bayes decision rules with uniform prior on the grid. We also conducted testing procedures from section 4.6 but we did not show the results because these testing procedures are not widely used

in meta-analysis. The minimax linear combination is $0.32S_{\mu} + 0.68S_{\tau^2}$. The Bayes linear combination is $0.24S_{\mu} + 0.76S_{\tau^2}$. Note that we normalize the score matrix K_1, K_2 such that $\text{tr}(K_1) = \text{tr}(K_2)$ before the linear combination of the score statistics.

Figure 4.2 shows the absolute power of four procedures and their power loss with an oracle procedure, Max, that is the maximum power within the linear combination family if the alternative (μ_a, τ_a^2) is known in advance. We can see that S_{μ} suffers from significant power loss in regions with large τ_a and S_{τ^2} suffers from significant power loss in regions with large μ_a. Minimax and Bayes procedures can balance the power very well. Their powers are similar although Bayes procedure puts less weight on μ compared to Minimax procedure, resulting in a slightly larger power loss for large μ_a.

We also conducted the likelihood ratio test (LRT) for the global null in Wu et al (2020). In Figure 4.3 we show the pairwise differences between different testing procedures. We find that LRT performs better than Minimax, Fisher, Tippett and Maxstat in almost all alternatives except in some regions, which may be caused by numerical accuracy. However, the oracle procedure combining score statistics (Max) has higher power than LRT. This indicates that knowing the alternatives for combining score statistics can have power gain compared to LRT although in practice alternatives are not known in advance.

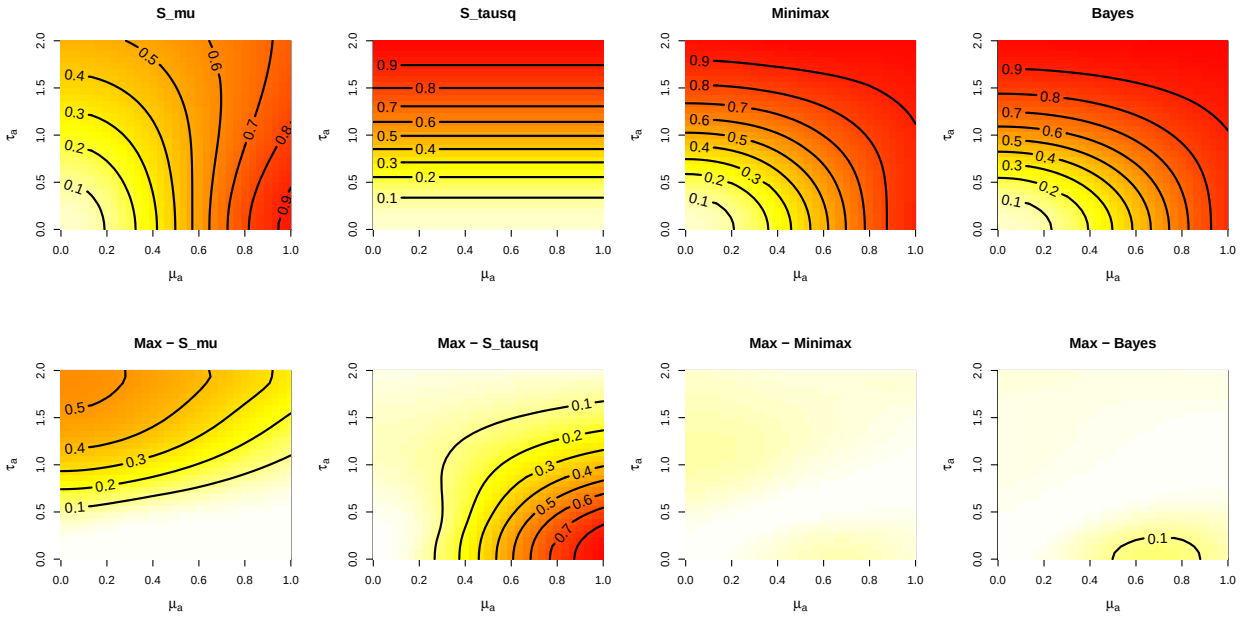


Figure 4.2: Meta-analysis: power heat map of S_μ, S_{τ^2}, Minimax and Bayes procedures. "Max" is the maximum achievable power within the linear combination family if the alternative is known.

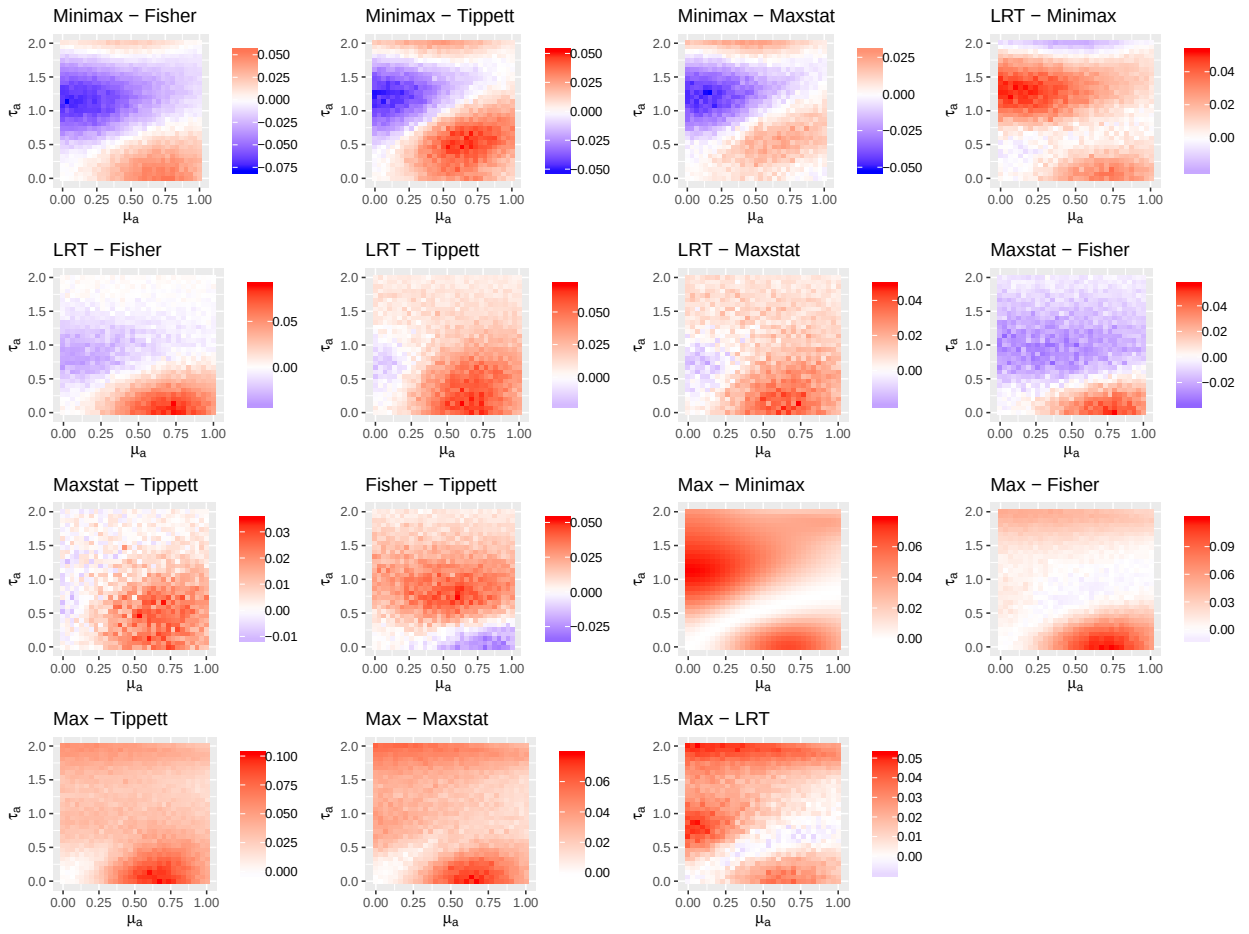


Figure 4.3: Meta-analysis: pairwise power differences among Minimax, Fisher, Tippett, Maxstat, LRT, and Max procedures.

4.7.2 Example 2: set-based genetic variant association analysis

Continuous trait

The simulation setting is similar to Lin and Tang (2011) and Sun et al (2013). We generate $p = 10$ variants in a region under the Hardy-Weinberg equilibrium with minor allele frequency (MAFs), $f_j = 0.005j$ for $j = 1, \ldots, 10$. We generate two continuous covariates X_1 and X_2 from standard normal distribution. For each individual, the continuous phenotype response

variable is generated by

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \sum_{j=1}^{10} \beta_j G_j + \epsilon, \tag{4.17}$$

where $\alpha_0 = -1, \alpha_1 = 0.5, \alpha_2 = -1$, $\beta_j, j = 1, \cdots, 10$, are regression coefficients for genetic variants and ϵ follows standard normal distribution. We consider the alternative that β_j is from $N(\mu_a, \tau_a^2)$ where μ_a consists of 18 equally spaced points between 0 and 0.2 and τ_a consists of 18 equally spaced points between 0 and 0.3. We evaluate the type I error and power under the significance level 0.05 and choose the sample size $n = 500$.

To calculate power, we first simulate $\beta_j, j = 1, \cdots, 10$ from $N(\mu_a, \tau_a^2)$. For given $\beta_a = (\beta_1, \ldots, \beta_{10})^T$, we simulate data from equation (4.17) and conduct test from the minimax procedure in equation (4.10), Tippett procedure in section 4.6.1, Fisher procedure in section 4.6.2, maximum test statistics (MaxStat) in section 4.6.3, minimum p-values (MinPvalue) in section 4.6.4, and adaptive procedure in section 4.6.5. We repeat this process for 10^4 times to obtain an average empirical power under a particular (μ_a, τ_a^2). The minimax linear combination of score statistics is $0.32S_\mu + 0.68S_{\tau^2}$. The Bayes procedure with uniform prior is $0.26S_\mu + 0.74S_{\tau^2}$ but we do not show its results since they are similar to the results of Minimax. Note that we normalize the score matrix K_1, K_2 such that $\mathrm{tr}(K_1) = \mathrm{tr}(K_2)$ before the linear combination of the score statistics.

Figure 4.4 shows the empirical power loss of different testing procedures compared to the oracle procedure Max, that is the optimal power within the linear combination family for the alternative (μ_a, τ_a^2). We can see that Minimax and MaxStat procedures can balance the power very well and their power loss with the oracle procedure is relatively small. Fisher's procedure performs well in regions with large heterogeneity and small mean effect, but suffers from a significant power loss in regions with large mean effect and small heterogeneity. Tippett's procedure has moderate power loss in almost every parts of the alternative region. MinPvalue and Adaptive have very similar power and they behave well in regions with large τ_a and small μ_a but not in regions with small τ_a and large μ_a. We also show the empirical pairwise power differences in Figure 4.5.

Figure 4.6 shows the power of the linear combination as a function of ρ for selected alternatives. We can see a clear trade-off of power for different linear combinations. As ρ increases from 0 to 1, the power of S_ρ increases first, then hits its maximum and finally decreases. We can generally achieve high power across different alternatives for ρ between 0.2 and 0.4. In Appendix A.9, we demonstrate that the analytical power formula from proposition 1 is accurate compared to the empirical power for continuous trait.

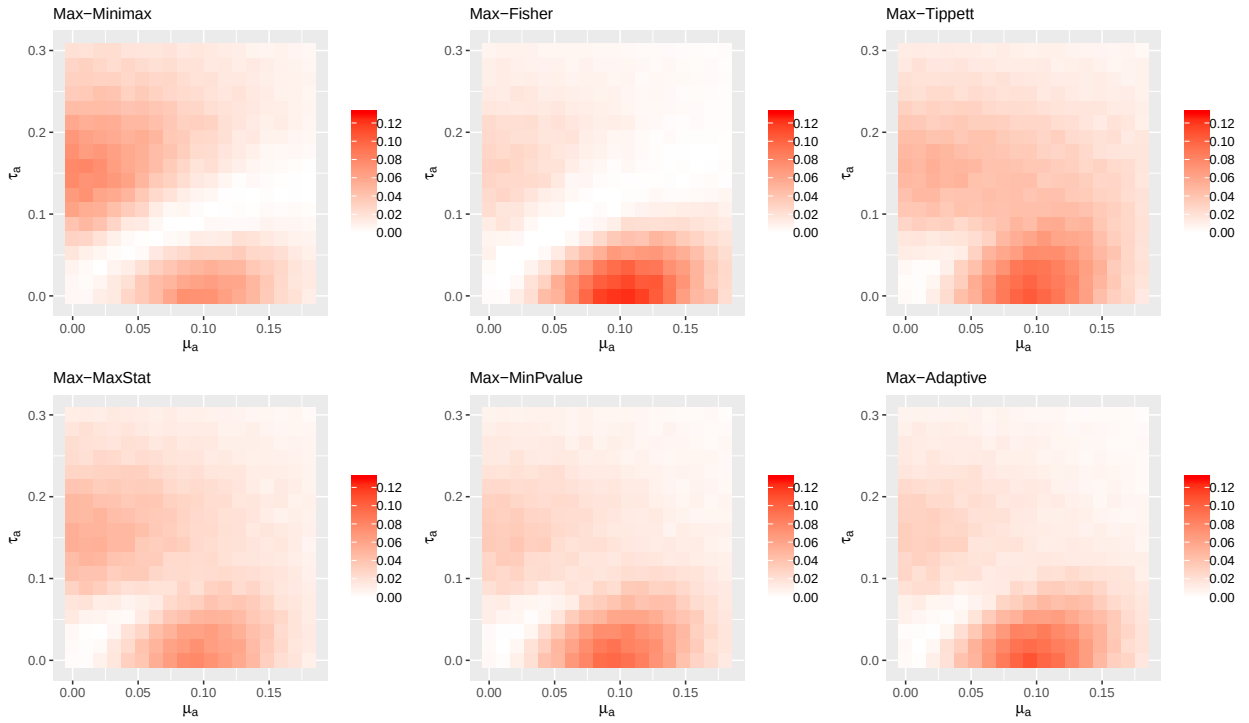


Figure 4.4: Genetic association test with continuous trait: power loss of various combination procedures compared to the oracle procedure "Max."

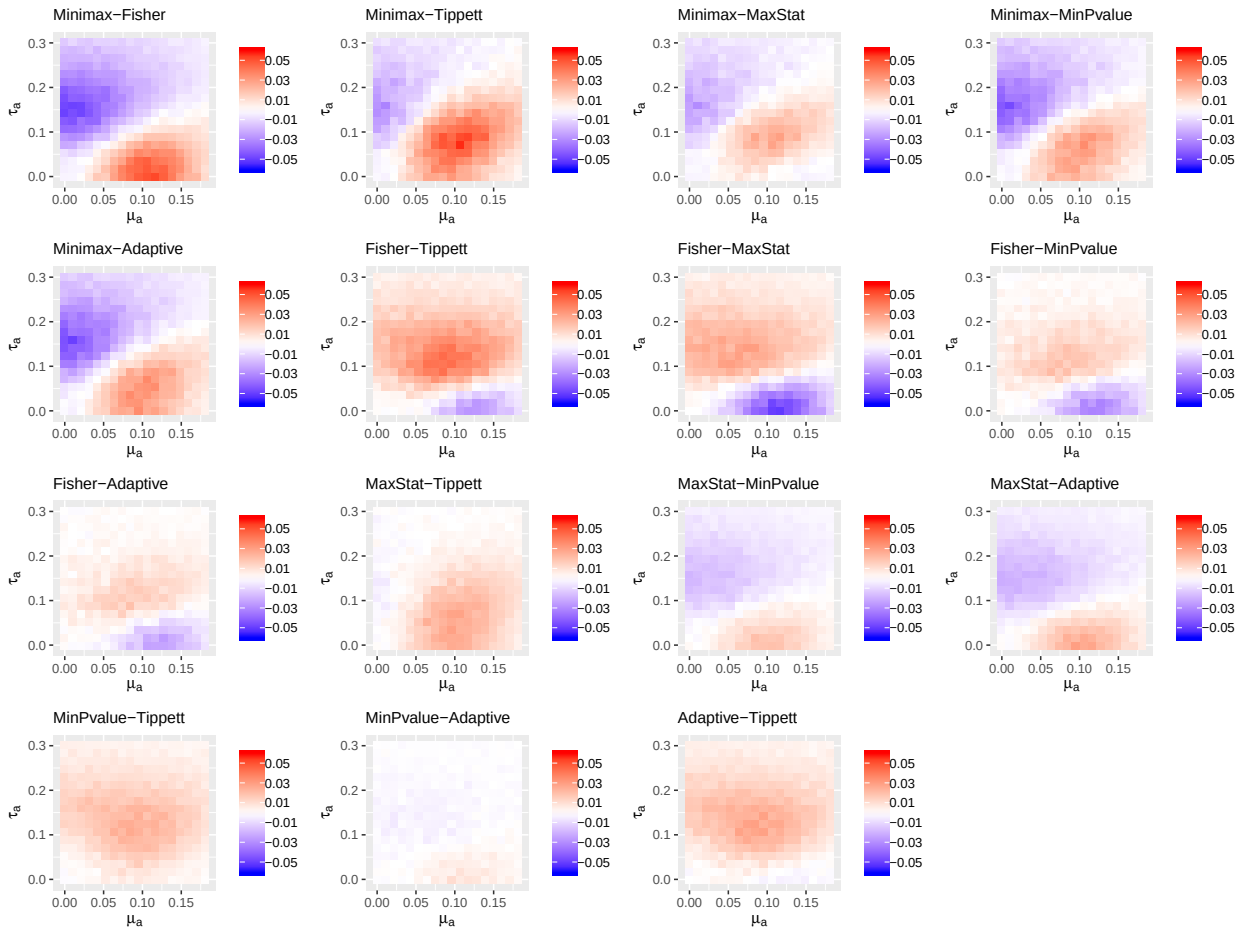


Figure 4.5: Genetic association test with continuous trait: pairwise power differences among Minimax, Fisher, Tippett, Maxstat, MinPvalue, Adaptive procedures.

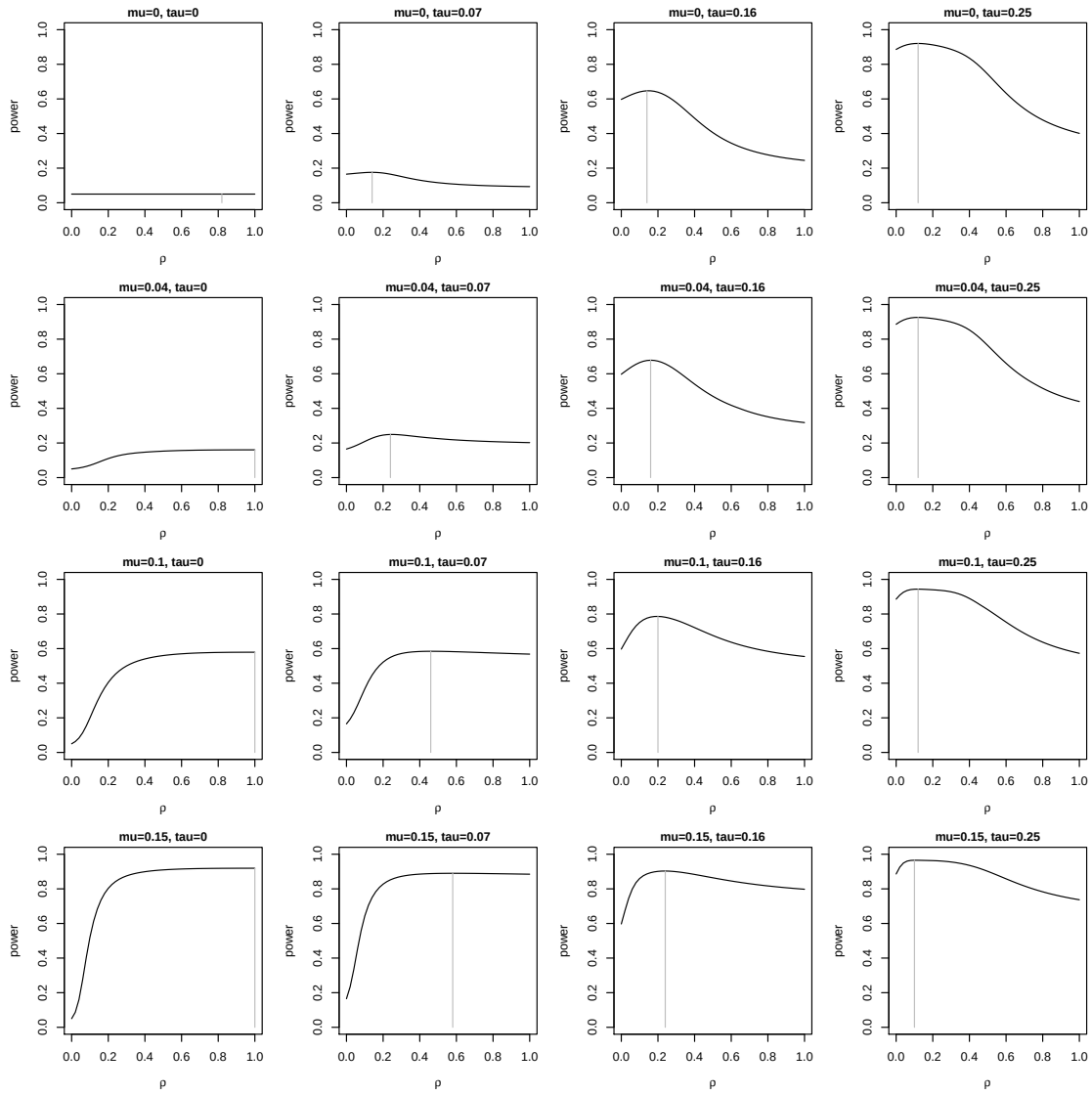


Figure 4.6: Genetic association test with continuous trait: power curves of $S_\rho = \rho S_\mu + (1-\rho)S_{\tau^2}$ as a function of ρ for a specific alternative (μ_a, τ_a).

Binary trait

Similar to the setting in continuous trait, we generate binary phenotype response by

$$\text{logit}\left\{P(Y=1|X_1, X_2, G)\right\} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \sum_{j=1}^{10} \beta_j G_j, \tag{4.18}$$

where $\text{logit}(x) = \log\{x/(1-x)\}$ and $\alpha_0 = -1, \alpha_1 = 0.5, \alpha_2 = -1$. We choose sample size to be $n = 500$ and significance level at 0.05. We evaluate power by generating β_j from $N(\mu_a, \tau_a^2)$ similar to the continuous case. The minimax optimal linear combination is $0.3S_\mu + 0.7S_{\tau^2}$. The Bayes optimal linear combination using uniform prior is $0.24S_\mu + 0.76S_{\tau^2}$. Figure 4.7 shows the empirical power loss from the oracle procedure Max, and the results are similar to the continuous case in Figure 4.4. Figure 4.8 shows the empirical power pairwise difference between Minimax, Fisher, Tippett, Maxstat, MinPvalue, Adaptive and Max procedures. It is similar to the continuous case in Figure 4.5.

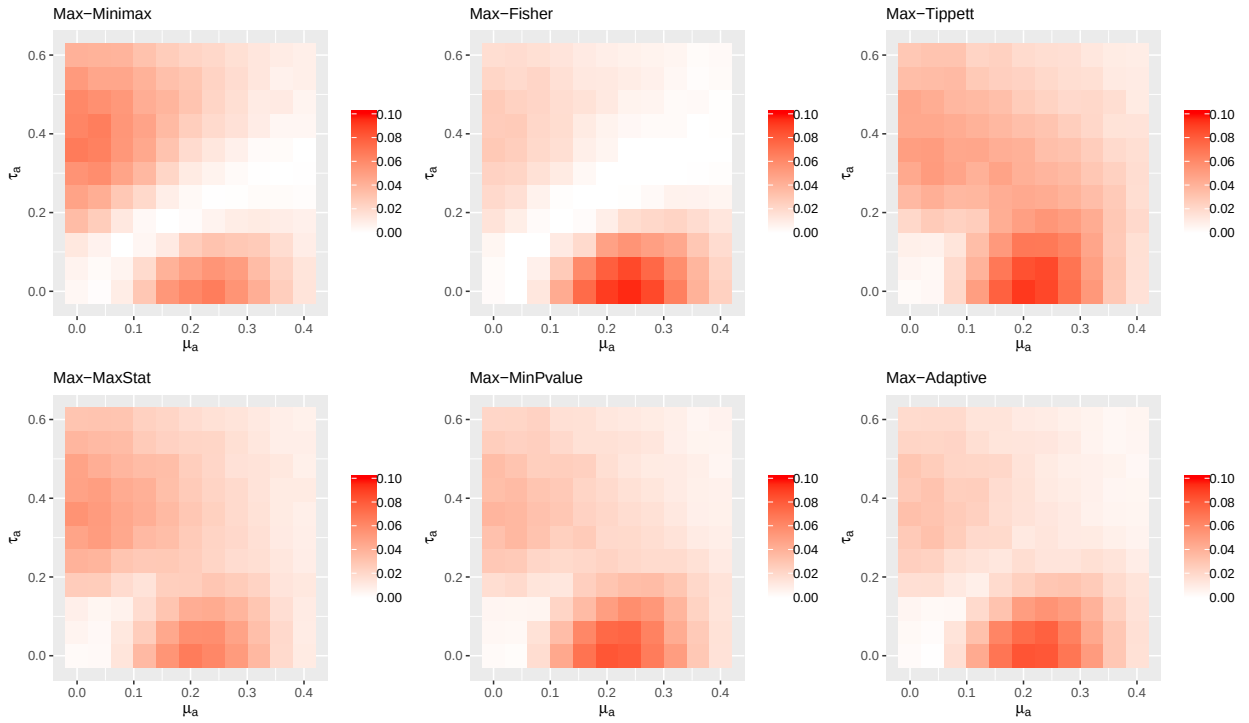


Figure 4.7: Genetic association test with binary trait: power loss of various combination procedures compared to the oracle procedure "Max."

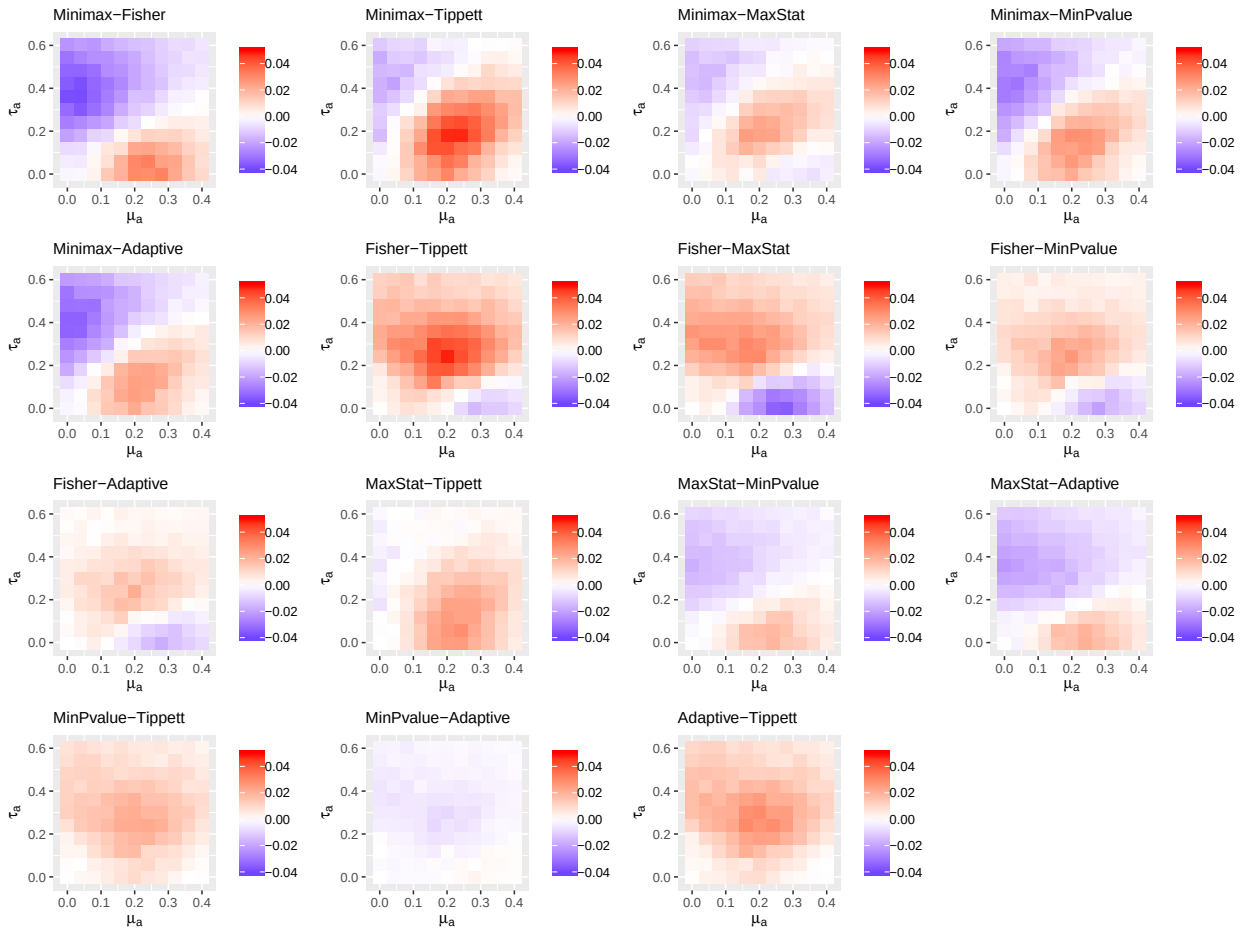


Figure 4.8: Genetic association test with binary trait: pairwise power differences among Minimax, Fisher, Tippett, Maxstat, MinPvalue, Adaptive procedures.

4.7.3 Example 3: time-varying treatment effect in extended Cox model

We generate survival data from extended Cox model $\lambda(t) = \exp\{\alpha_1 X_1 + \alpha_2 X_2 + \beta(t)Z\}$ with sample size $n = 100$, where baseline hazard function is constant, X_1, X_2 are from standard normal distribution and $\alpha_1 = 0.5, \alpha_2 = -1$. The censoring distribution is uniform distribution on $[0, c]$, where c is chosen such that around 30% subjects are censored. We use smoothing splines for the basis functions.

Scenario	$\beta(t)$	Minimax	Fisher	Tippett	MaxStat	MinPvalue	Adaptive
1	0	0.051	0.054	0.051	0.052	0.051	0.052
2	$\log(1.5)$	0.325	0.301	0.303	0.308	0.304	0.307
3	$0.8t$	0.536	0.556	0.499	0.516	0.524	0.530
4	$-0.5t(t-2.6)$	0.375	0.378	0.335	0.349	0.350	0.359
5	$0.25\exp(0.8t)$	0.457	0.444	0.410	0.423	0.427	0.431
6	$0.7\exp(-t)$	0.480	0.447	0.435	0.445	0.460	0.449
7	$\frac{0.6\exp(3.5t)}{1+\exp(3.5t)}$	0.406	0.381	0.373	0.381	0.382	0.384
8	$0.5\log(0.75t)$	0.753	0.757	0.679	0.705	0.706	0.728
9	$0.7\log\frac{2}{1+5t}$	0.509	0.533	0.506	0.514	0.522	0.525
10	$1.2I(t \geq 1)$	0.375	0.405	0.370	0.379	0.386	0.392
11	$0.8\cos(2\pi t/2.7)$	0.560	0.575	0.504	0.526	0.538	0.547

Table 4.1: Power comparison among various combination procedures for 11 fixed alternatives under the extended Cox model.

We conduct simulations with various alternative shapes of $\beta(t)$ considered in Saegusa et al (2014) and they are shown in Figure 4.9. From these 11 alternatives, the minimax and bayes procedure using uniform prior leads to the same linear combination $0.4S_{\mu} + 0.6S_{\tau^2}$. Table 4.1 illustrates the power from Minimax, Fisher, Tippett, MaxStat, MinPvalue and Adaptive procedures. We see that in scenario 2, 5, 6, 7, Minimax has higher power than other procedures. In scenario 4, 8, Minimax has similar power to Fisher's procedure. In scenario 3, 9, 19, 11, Minimax has lower power than Fisher's procedures.

Figure 4.10 shows the power of linear combination as a function of ρ among the 11 scenarios. We can see a clear trade-off of power for different linear combinations and ρ between 0.3 and 0.4 generally gives high power.

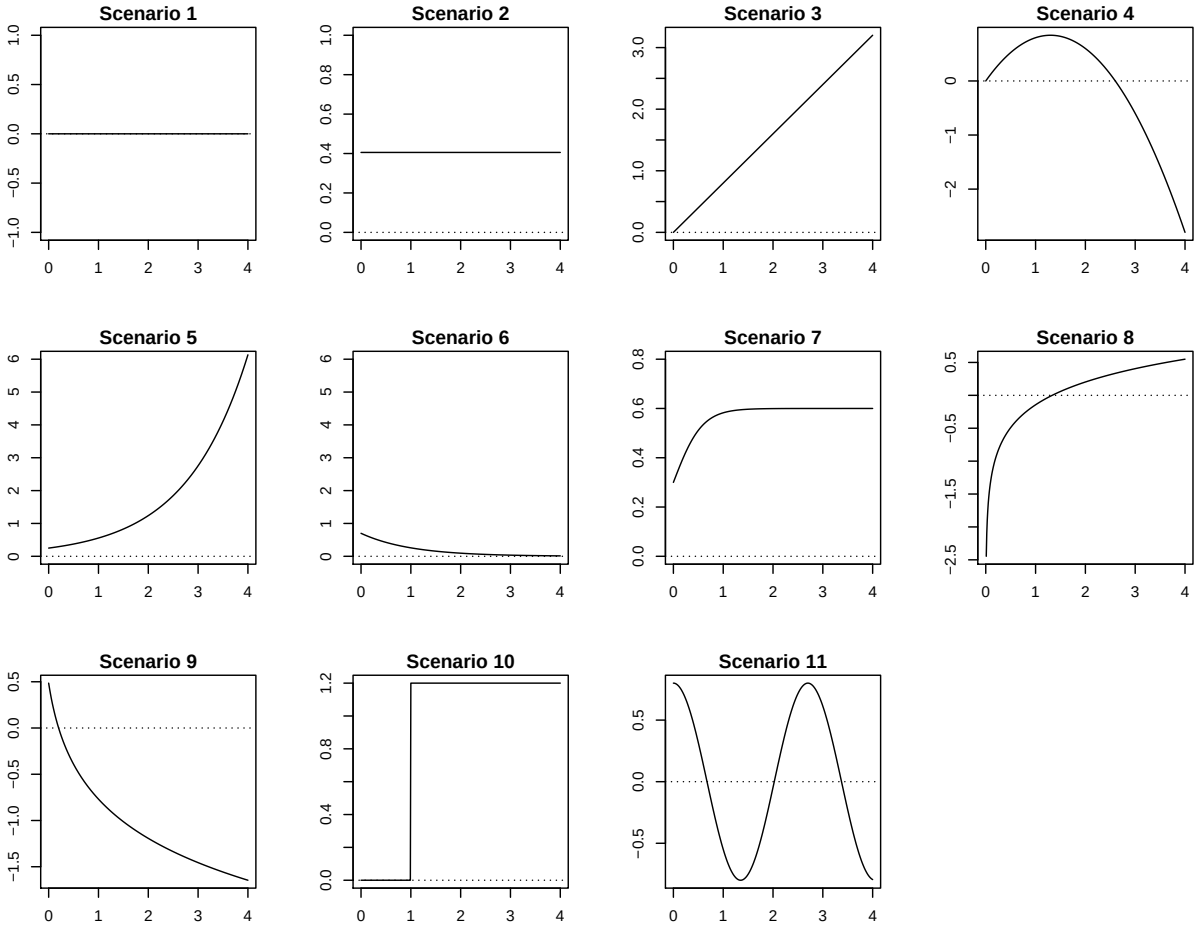


Figure 4.9: Shapes of hazard ratio function $\beta(t)$ used in simulations for the extended Cox model.

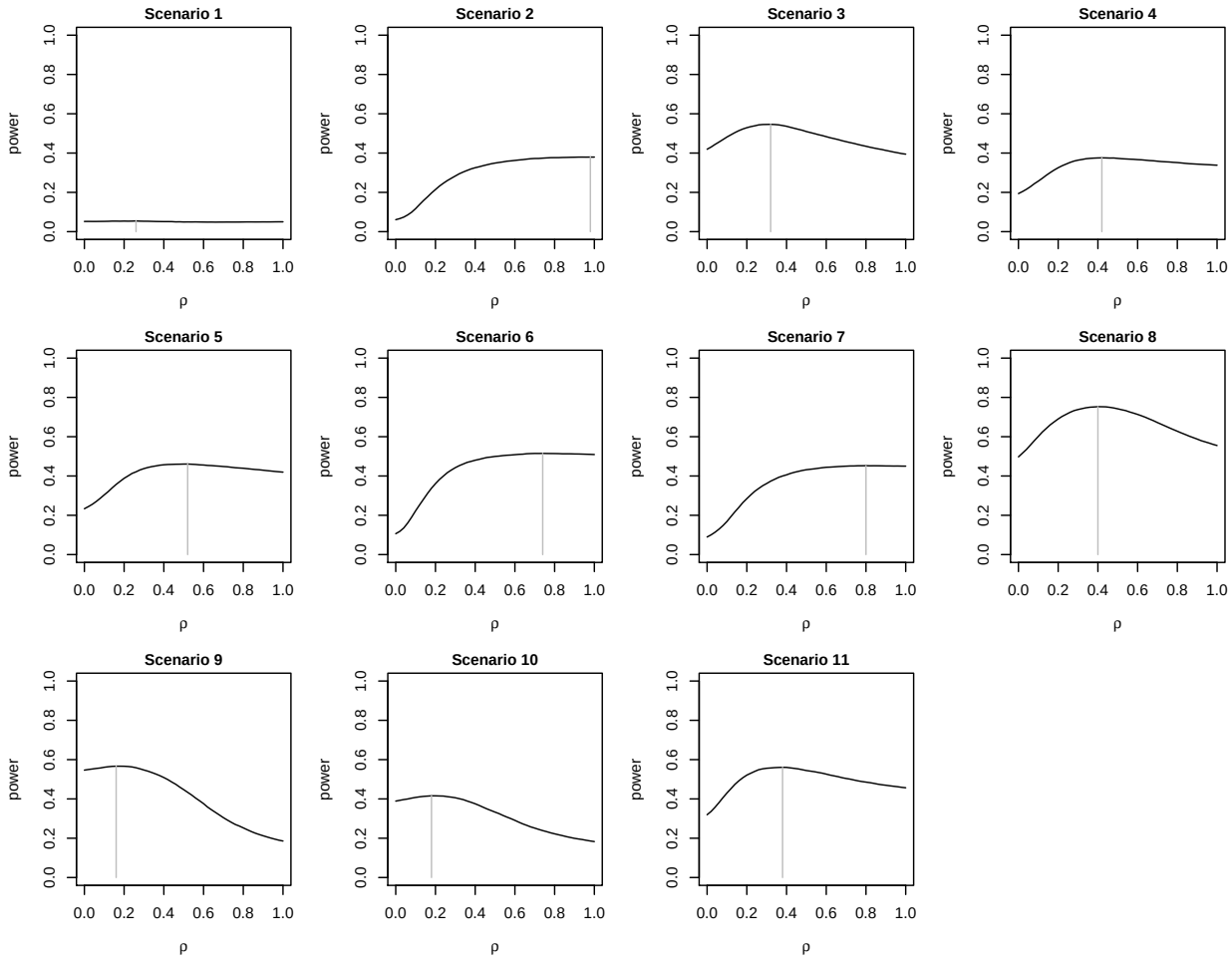


Figure 4.10: Extended Cox model with fixed alternatives: power curves of $S_\rho = \rho S_\mu + (1 - \rho)S_{\tau^2}$ as a function of ρ for various scenarios.

4.8 Application

4.8.1 Dallas Heart Study

Dallas heart study is designed to evaluate the social and biological factors contributing to the ethnic differences in the cardiovascular health (Victor et al, 2004). Researchers collected variables like gender, ethnic group, age and triglyceride level and sequenced three candidate genes (ANGPTL3, ANGPTL4, ANGPTL5). In this analysis, we followed the data analysis strategy in Sun et al (2013). We included 3409 individuals and focused on rare variants where the observed MAF $f_j < 0.05$. There are 1500 men and 1909 women. There are 1762 black, 988 white, 586 hispanic and 73 other ethnic groups. There are 85, 90 and 96 variants in gene ANGPTL3, ANGPTL4, ANGPTL5, respectively. A rare variant is functional if it is missense, nonsense or frameshift, and there are 36, 31 and 27 functional variants in gene ANGPTL3, ANGPTL4, ANGPTL5, respectively. We analyze the data based on functional variants only or all variants.

We are interested in the association between log-transformed triglyceride levels with each candidate gene after adjusting for gender, race and age. The results from weighted and unweighted versions of the test are shown in Table 4.2 and Table 4.3, respectively. The weights are set by $w_j = \left\{f_j^{a-1}(1-f_j)^{b-1}\right\}^2$, where $a = 1$ and $b = 25$. We specify the matrix Z as a vector of $\sqrt{w_j}$. Minimax linear combination is $0.32S_\mu + 0.68S_{\tau^2}$ based on the simulations in section 4.7.2. We use $\alpha = 0.05$ as significance level.

The results of the weighted tests and unweighted tests are similar. Our discussion will focus on the weighted tests in Table 4.2. When we include only functional variants, gene ANGPTL3 and ANGPTL4 are significant based on all the combination procedures. Gene ANGPTL3 has significant mean effect (p-value of S_μ: 0.006) but non-significant heterogeneity (p-value of S_{τ^2}: 0.382). For gene ANGPTL3, Minimax procedure has smaller p-values than other combination procedures. Gene ANGPTL4 has significant mean effect (p-value of S_μ: 0.031) and significant heterogeneity (p-value of S_{τ^2}: 0.001). For gene ANGPTL4, Minimax has larger p-values than other combination procedures. It agrees with the simulation results

that minimax has higher power when the mean effect is large and there is not so much heterogeneity. None of the tests can find association of log-transformed triglyceride and functional variants in gene ANGPTL5.

When we include all variants, gene ANGPTL3 has non-significant mean effect (p-value of S_μ: 0.119) and non-significant heterogeneity (p-value of S_{τ^2}: 0.067). Gene ANGPTL4 has non-significant mean effect (p-value of S_μ: 0.0.865) but significant heterogeneity (p-value of S_{τ^2}: 0.001). Gene ANGPTL3 is significant based on Fisher's procedure but not the other combination procedures. Gene ANGPTL4 is significant based on all the combination methods except for Minimax. None of the tests can find association of log-transformed triglyceride and all variants in gene ANGPTL5.

	S_μ	S_{τ^2}	Minimax	Fisher	Tippett	MaxStat	MinPvalue	Adaptive
				Functional				
ANGPTL3	0.006	0.382	0.009	0.016	0.012	0.010	0.013	0.013
ANGPTL4	0.031	0.001	0.001	0.000	0.002	0.002	0.001	0.001
ANGPTL5	0.248	0.630	0.419	0.446	0.434	0.396	0.430	0.455
				All				
ANGPTL3	0.119	0.067	0.069	0.046	0.129	0.102	0.094	0.077
ANGPTL4	0.865	0.001	0.075	0.007	0.002	0.008	0.003	0.003
ANGPTL5	0.669	0.903	1.000	0.909	0.891	0.220	0.884	0.904

Table 4.2: P-values from the weighted tests for the Dallas Heart Study. Results from functional variants only or all variants are shown.

	S_μ	S_{τ^2}	Minimax	Fisher	Tippett	MaxStat	MinPvalue	Adaptive
				Functional				
ANGPTL3	0.006	0.232	0.016	0.010	0.012	0.012	0.012	0.012
ANGPTL4	0.031	0.004	0.001	0.001	0.009	0.006	0.003	0.004
ANGPTL5	0.248	0.372	0.343	0.312	0.434	0.450	0.430	0.375
				All				
ANGPTL3	0.119	0.029	0.039	0.023	0.057	0.055	0.046	0.040
ANGPTL4	0.865	0.003	0.043	0.018	0.006	0.011	0.008	0.007
ANGPTL5	0.669	0.665	0.805	0.806	0.888	0.369	0.884	0.839

Table 4.3: P-values from unweighted tests for the Dallas Heart Study. Results from functional variants only or all variants are shown.

4.8.2 HIVNET 012

The HIVNET 012 study was a randomized clinical trial to study the effectiveness of single-dose nevirapine (NVP) versus short-course zidovudine (AZT) to prevent mother-to-child HIV transmissions among HIV infected pregnant women in Uganda (Jackson et al, 2003). The trial initially had a placebo arm but it was dropped after a study in Thailand showed that short-course AZT can reduce mother-to-child transmissions of HIV. Infants who were born without HIV could still be susceptible for HIV infection after birth due to breastfeeding. In the NVP arm, mothers received a single dose of NVP at the onset of labor and their children received a single does of NVP within 72 hours of birth. In the AZT arm, mothers received AZT at the onset of labor, followed by a dose every 3 hours during labor, and their children received an oral dose twice daily for the first 7 days of life. The primary endpoints of this study were HIV infection and HIV-free survival of infants at 6–8 weeks, 14–16 weeks, and 18 months of age. It was found that NVP was associated with a 41% reduction of mother-to-child HIV transmissions through the age of 18 months compared to AZT. Thus, NVP was a

simpler medication and was more effective in preventing HIV transmissions.

In this application, we focus on the 18-months survival of infants as our outcome. In our analysis, we exclude the second twins or more. There are 310 women in the NVP arm and 306 women in the AZT arm. We show the estimated Kaplan-Meier survival function and its log negative log transformation in Figure 4.11. We see that NVP has better survival than AZT but the log-rank test has a p-value of 0.14, suggesting that the difference is not significant. However, the non-significant result may due to the sample size or the violation of the proportional hazard. It is known that log-rank test will lose power to detect survival differences if the proportional hazard does not hold. The log negative log transformation in Figure 4.11 suggests that hazards for two groups are not proportional. Grambsch-Therneau test (Grambsch and Therneau, 1994) for proportional hazard has a p-value of 0.02, suggesting that the hazard ratio is not constant.

In light of the potential non-proportional hazard between two arms, we use the testing procedure in section 4.5 to test the time-varying treatment effect. In Table 4.4, we show the p-values of different testing procedures. Minimax procedure uses the linear combination $0.4S_{\mu}+0.6S_{\tau^2}$ obtained from the simulation studies in section 4.7.3. We show the unadjusted results and adjusted results with viral load, CD4 and birth weight as covariates. Both unadjusted and adjusted log-rank tests suggest the treatment effect is not significant (p-value: 0.145 and 0.066 respectively). Both unadjusted and adjusted tests for proportionality suggest time-varying hazard ratio (p-value: 0.02 and 0.026 respectively). All the combination methods suggest a significant benefit of NVP compared to AZT. We also conducted weighted log rank tests with weights coming from G^{ρ} family (Harrington and Fleming, 1982). With a usual choice of ρ between 0 and 1, we didn't find statistical significance. We also tried $\rho = 2, 3, 4, 5, 6$ and the results were not significant. Not until $\rho = 7$ is considered, the test becomes significant. It indicates that this approach is sensitive to the choice of ρ.

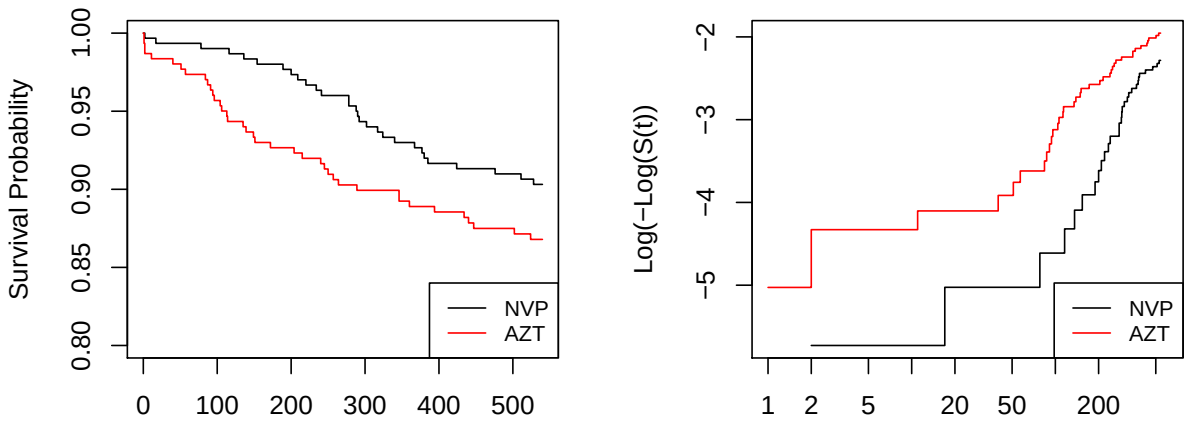


Figure 4.11: HIVNET 012 Study: estimated Kaplan-Meier survival probability (left) and log negative log transformation of the survival probability (right) for two treatment groups.

	Unadjusted	Adjusted
Log-rank test S_{μ}	0.145	0.066
Proportional hazard S_{τ^2}	0.020	0.026
Minimax	0.025	0.016
Fisher	0.020	0.012
Tippett	0.039	0.050
MaxStat	0.039	0.037
MinPvalue	0.036	0.030
Adaptive	0.031	0.026

Table 4.4: P-values for effects of treatment on infants' 18-month survival in the HIVNET 012 Study.

4.9 Discussion

In this chapter, we studied linear combination of score statistics for mixed-effects models with three examples in meta analysis, genetic variant association test and survival analysis. We proposed Minimax and Bayes decision rules to determine the weight with the aim of balancing power under a range of alternatives. We conducted comprehensive power comparisons between the proposed Minimax and Bayes procedure with other non-linear combination approaches.

We find that there does not exist one test that is universally more powerful than other tests across all types of alternatives. The power difference between different methods is not very large except for the Tippett combination. Fisher's procedure has significant power loss when μ_a is large and τ_a is small. MaxSta, MinPvalue and Adaptive procedure seem to balance the power well but they require numerical integration or grid search to calculate power. Minimax procedure can balance the power well across alternatives and is fast in computation once the weight is determined by power. If prior knowledge on the plausibility of the alternative regions exists, one identify the weight using Bayes or Minimax procedures.

There are many procedures for combining different test statistics besides the approaches described in this chapter. For example, Liu and Xie (2019) considered Cauchy combination of p-values under arbitrary dependency structures. However, the goal of this chapter is to combine score statistics from the perspective of power. A testing procedure will perform well in certain regions of the alternative space but not so well in other regions. This trade-off motivates us to use decision theory to find the optimal linear combination such that power is relatively high for all types of the alternative.

Mixed-effects model is a powerful approach to impose structure and combine information for estimation and inference. It has been widely used in the set-based genetic variant association test which motivated the development of this chapter. But the idea of using linear combination of score statistics is less known outside this area. We hope that this research can motivate hypothesis testing of mixed-effects models in other areas of application.

BIBLIOGRAPHY

Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, Puren A (2005) Randomized, controlled intervention trial of male circumcision for reduction of hiv infection risk: the anrs 1265 trial. PLoS medicine 2(11):e298

Bailey RC, Moses S, Parker CB, Agot K, Maclean I, Krieger JN, Williams CF, Campbell RT, Ndinya-Achola JO (2007) Male circumcision for hiv prevention in young men in kisumu, kenya: a randomised controlled trial. The lancet 369(9562):643–656

Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P, Yeager M, Chung CC, Chanock SJ, Chatterjee N, et al (2012) A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. The American Journal of Human Genetics 90(5):821–835

Biggerstaff B, Tweedie R (1997) Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Statistics in Medicine 16(7):753–768

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

Brockwell SE, Gordon IR (2001) A comparison of statistical methods for meta-analysis. Statistics in Medicine 20(6):825–840

Cochran WG (1937) Problems arising in the analysis of a series of similar experiments. Supplement to the Journal of the Royal Statistical Society 4(1):102–118

Crainiceanu CM, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66(1):165–185

DerSimonian R, Kacker R (2007) Random-effects model for meta-analysis of clinical trials: an update. Contemporary clinical trials 28(2):105–114

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Controlled clinical trials 7(3):177–188

Fisher R (1925) Statistical methods for research workers. Edinburgh Oliver & Boyd

Fleming TR, Harrington DP (1991) Counting processes and survival analysis, vol 169. John Wiley & Sons

Frangakis CE, Rubin DB (2002) Principal stratification in causal inference. Biometrics 58(1):21–29

Freedman LS, Graubard BI, Schatzkin A (1992) Statistical validation of intermediate endpoints for chronic diseases. Statistics in Medicine 11(2):167–178

Gilbert PB, Hudgens MG (2008) Evaluating candidate principal surrogate endpoints. Biometrics 64(4):1146–1154

Grambsch PM, Therneau TM (1994) Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81(3):515–526

Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F, Kiwanuka N, Moulton LH, Chaudhary MA, Chen MZ, et al (2007) Male circumcision for hiv prevention in men in rakai, uganda: a randomised trial. The lancet 369(9562):657–666

Guolo A (2012) Higher-order likelihood inference in meta-analysis and meta-regression. Statistics in Medicine 31(4):313–327

Guolo A, Varin C (2017) Random-effects meta-analysis: the number of studies matters. Statistical Methods in Medical Research 26(3):1500–1518

Han B, Eskin E (2011) Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. The American Journal of Human Genetics 88(5):586–598

Hardy RJ, Thompson SG (1996) A likelihood approach to meta-analysis with random effects. Statistics in Medicine 15(6):619–629

Harrington DP, Fleming TR (1982) A class of rank test procedures for censored survival data. Biometrika 69(3):553–566

Hartung J, Knapp G (2001a) On tests of the overall treatment effect in meta-analysis with normally distributed responses. Statistics in medicine 20(12):1771–1782

Hartung J, Knapp G (2001b) A refined method for the meta-analysis of controlled clinical trials with binary outcome. Statistics in medicine 20(24):3875–3889

He Q, Liu Y, Peters U, Hsu L (2018) Multivariate association analysis with somatic mutation data. Biometrics 74(1):176–184

Huang Y, Gilbert PB (2011) Comparing biomarkers as principal surrogate endpoints. Biometrics 67(4):1442–1451

Huang Y, Gilbert PB, Wolfson J (2013) Design and estimation for evaluating principal surrogate markers in vaccine trials. Biometrics 69(2):301–309

Jackson JB, Musoke P, Fleming T, Guay LA, Bagenda D, Allen M, Nakabiito C, Sherman J, Bakaki P, Owor M, et al (2003) Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of hiv-1 in kampala, uganda: 18-month follow-up of the hivnet 012 randomised trial. The Lancet 362(9387):859–868

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

Knapp G, Hartung J (2003) Improved tests for a random effects meta-regression with a single covariate. Statistics in medicine 22(17):2693–2710

Kosmidis I, Guolo A, Varin C (2017) Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. Biometrika 104(2):489–496

Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762–775

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. The American Journal of Human Genetics 83(3):311–321

Li Y, Taylor JM, Elliott MR (2010) A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. Biometrics 66(2):523–531

Lin D, Fleming T, De Gruttola V, et al (1997) Estimating the proportion of treatment effect explained by a surrogate marker. Statistics in Medicine 16(13):1515–1527

Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. The American Journal of Human Genetics 89(3):354–367

Lin X (1997) Variance component testing in generalised linear models with random effects. Biometrika 84(2):309–326

Lin X, Breslow NE (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association 91(435):1007–1016

Lin X, Zhang D (1999) Inference in generalized additive mixed modelsby using smoothing splines. Journal of the royal statistical society: Series b (statistical methodology) 61(2):381–400

Lin X, Lee S, Christiani DC, Lin X (2013) Test for interactions between a genetic marker set and environment in generalized linear models. Biostatistics 14(4):667–681

Liu H, Tang Y, Zhang HH (2009) A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Computational Statistics & Data Analysis 53(4):853–856

Liu Y, Xie J (2019) Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. Journal of the American Statistical Association pp 1–18

Luedtke AR, Wu J (2017) Efficient principally stratified treatment effect estimation in crossover studies with absorbent binary endpoints. arXiv preprint arXiv:171205835

Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS genetics 5(2):e1000,384

Millett GA, Flores SA, Marks G, Reed JB, Herbst JH (2008) Circumcision status and risk of hiv and sexually transmitted infections among men who have sex with men: a meta-analysis. Jama 300(14):1674–1684

Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 615(1-2):28–56

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58(3):545–554

Prentice RL (1989) Surrogate endpoints in clinical trials: definition and operational criteria. Statistics in Medicine 8(4):431–440

Saegusa T, Di C, Chen YQ (2014) Hypothesis testing for an extended cox model with time-varying coefficients. Biometrics 70(3):619–628

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

Sharma SC, Raison N, Khan S, Shabbir M, Dasgupta P, Ahmed K (2018) Male circumcision for the prevention of human immunodeficiency virus (hiv) acquisition: a meta-analysis. BJU international 121(4):515–526

Su YR, Di CZ, Hsu L (2017) A unified powerful set-based test for sequencing data analysis of gxe interactions. Biostatistics 18(1):119–131

Su YR, Di C, Bien S, Huang L, Dong X, Abecasis G, Berndt S, Bezieau S, Brenner H, Caan B, et al (2018) A mixed-effects model for powerful association tests in integrative functional genomics. The American Journal of Human Genetics 102(5):904–919

Sun J, Zheng Y, Hsu L (2013) A unified mixed-effects model for rare-variant association in sequencing studies. Genetic epidemiology 37(4):334–344

Thompson SG, Sharp SJ (1999) Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in Medicine 18(20):2693–2708

Tippett LHC, et al (1931) The methods of statistics. The Methods of Statistics

Verbeke G, Molenberghs G (2003) The use of score tests for inference on variance components. Biometrics 59(2):254–262

Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, Leonard D, Basit M, Cooper RS, Iannacchione VG, Visscher WA, et al (2004) The dallas heart study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. The American journal of cardiology 93(12):1473–1480

Wang Y, Taylor JM (2002) A measure of the proportion of treatment effect explained by a surrogate marker. Biometrics 58(4):803–812

Weir CJ, Walley RJ (2006) Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Statistics in Medicine 25(2):183–203

Wiysonge CS, Kongnyuy EJ, Shey M, Muula AS, Navti OB, Akl EA, Lo YR (2011) Male circumcision for prevention of homosexual acquisition of hiv in men. Cochrane Database of Systematic Reviews (6)

Wolfson J, Gilbert P (2010) Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. Biometrics 66(4):1153–1161

Wu J, Di C, Hsu L, Chen Y (2020) Likelihood ratio tests for meta-analysis and meta-regression based on random-effects models, submitted

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics 89(1):82–93

Yuan T, Fitzpatrick T, Ko NY, Cai Y, Chen Y, Zhao J, Li L, Xu J, Gu J, Li J, Hao C, Yang Z, Cai W, Cheng CY, Hao Y, Luo Z, Zhang K, Wu G, Meng X, Grulich A, Zou H (2019) Circumcision to prevent hiv and other sexually transmitted infections in men who have sex with men: a systematic review and meta- analysis of global data. The Lancet HIV

Zeng D, Lin D (2015) On random-effects meta-analysis. Biometrika 102(2):281–294

Zhang D, Lin X (2003) Hypothesis testing in semiparametric additive mixed models. Biostatistics 4(1):57–74

Zheng W, McLerran DF, Rolland B, Zhang X, Inoue M, Matsuo K, He J, Gupta PC, Ramadas K, Tsugane S, et al (2011) Association between body-mass index and risk of death in more than 1 million Asians. New England Journal of Medicine 364(8):719–729

Zigler CM, Belin TR (2012) A Bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. Biometrics 68(3):922–932

Appendix A

PROOFS AND TECHNICAL DETAILS

A.1 Proof of Theorem 1

Proof. We can write

$$
\begin{aligned}
&n^{1/2}(\hat{\phi}-\phi)\\
&= n^{1/2}\left(\frac{\hat{g}_1-\hat{g}_1^*}{\hat{g}_1-\hat{g}_0}-\frac{g_1-g_1^*}{g_1-g_0}\right)\\
&= n^{1/2}\left(\frac{\hat{g}_1}{\hat{g}_1-\hat{g}_0}-\frac{g_1}{g_1-g_0}-\frac{\hat{g}_1^*}{\hat{g}_1-\hat{g}_0}+\frac{g_1^*}{g_1-g_0}\right)\\
&= \frac{n^{1/2}}{(\hat{g}_1-\hat{g}_0)(g_1-g_0)}\left\{\hat{g}_1(g_1-g_0)-g_1(\hat{g}_1-\hat{g}_0)-\hat{g}_1^*(g_1-g_0)+g_1^*(\hat{g}_1-\hat{g}_0)\right\}\\
&= \frac{1}{(\hat{g}_1-\hat{g}_0)(g_1-g_0)}\left\{(g_1-g_1^*)n^{1/2}(\hat{g}_0-g_0)+(g_1^*-g_0)n^{1/2}(\hat{g}_1-g_1)+(g_0-g_1)n^{1/2}(\hat{g}_1^*-g_1^*)\right\}.
\end{aligned}
$$

Under random censoring assumption (iii), Kaplan-Meier estimator $\hat{S}_0(t)$ and $\hat{S}_1(t)$ satisfies respectively (Fleming and Harrington, 1991)

$$
n^{1/2}\{\hat{S}_0(t)-S_0(t)\} = n^{1/2}(\mathbb{P}_n-\mathbb{P})\left[-S_0(t)I(Z=0)\int_0^t \frac{dM(u|Z=0)}{E\left\{I(Z=0)Y(u)\right\}}\right]+o_p(1),
$$

$$
n^{1/2}\{\hat{S}_1(t)-S_1(t)\} = n^{1/2}(\mathbb{P}_n-\mathbb{P})\left[-S_1(t)I(Z=1)\int_0^t \frac{dM(u|Z=1)}{E\left\{I(Z=1)Y(u)\right\}}\right]+o_p(1).
$$

Next we derive the convergence of $\hat{S}_1^*$. Denote by $\hat{\Lambda}_1^*(t) = -\log \hat{S}_1^*(t)$ and $\Lambda_1^*(t) = -\log S_1^*(t)$. We have

$$
\hat{\Lambda}_1^*(t) = \sum_{i=1}^n \left\{\int_0^t \sum_{x=1}^K \widehat{\mathrm{pr}}(X=x|T\geq u, Z=0)\frac{dN_i(u)I(Z_i=1,X_i=x)}{\sum_{j=1}^n Y_j(u)I(Z_j=1,X_j=x)}\right\},
$$

and

$$
\Lambda_1^*(t) = \int_0^t \sum_{x=1}^K \mathrm{pr}(X=x|T\geq u, Z=0)\lambda(u|Z=1,X=x)du,
$$

where $\widehat{\mathrm{pr}}(X=x|T\geq u, Z=0) = \frac{\widehat{\mathrm{pr}}(X=x,Z=0,T\geq u)}{\widehat{\mathrm{pr}}(Z=0,T\geq u)} = \frac{n^{-1}\sum_{i=1}^n I(X_i=x,Z_i=0,T_i\geq u)}{n^{-1}\sum_{i=1}^n I(Z_i=0,T_i\geq u)}$. It is easy to show that

$$
\begin{aligned}
\hat{\Lambda}_1^*(t)-\Lambda_1^*(t) = &\sum_{i=1}^n \int_0^t \sum_{x=1}^K \frac{\widehat{\mathrm{pr}}(X=x|T\geq u, Z=0)}{\sum_{j=1}^n Y_j(u)I(Z_j=1,X_j=x)}dM_i(u|Z_i=1,X_i=x)I(Z_i=1,X_i=x)\\
&+\int_0^t \sum_{x=1}^K \left\{\widehat{\mathrm{pr}}(X=x|T\geq u, Z=0)-\mathrm{pr}(X=x|T\geq u, Z=0)\right\}\lambda(u|Z=1,X=x)du.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&n^{1/2}\left\{\widehat{\mathrm{pr}}(X=x|T\geq u,Z=0)-\mathrm{pr}(X=x|T\geq u,Z=0\}\right)\\
&=\frac{1}{\widehat{\mathrm{pr}}(Z=0,T\geq u)}n^{1/2}\left\{\widehat{\mathrm{pr}}(X=x,Z=0,T\geq u)-\mathrm{pr}(X=x,Z=0,T\geq u)\right\}\\
&+\mathrm{pr}(X=x,Z=0,T\geq u)n^{1/2}\left\{\frac{1}{\widehat{\mathrm{pr}}(Z=0,T\geq u)}-\frac{1}{\mathrm{pr}(Z=0,T\geq u)}\right\}.
\end{aligned}
$$

Thus by Slutsky's theorem and functional delta method,

$$
\begin{aligned}
n^{1/2}\{\hat{S}_1^*(t)-S_1^*(t)\}=n^{1/2}(\mathbb{P}_n-\mathbb{P})\Bigg[&-S_1^*(t)\int_0^t\sum_{x=1}^K I(Z=1,X=x)\frac{\mathrm{pr}(X=x|T\geq u,Z=0)}{E\left\{Y(u)I(Z=1,X=x)\right\}}\\
&dM(u|Z=1,X=x)-S_1^*(t)\int_0^t\sum_{x=1}^K Q(u,x)\lambda(u|Z=1,X=x)du\Bigg]+o_p(1),
\end{aligned}
$$

where $Q(u,x)$ is given in (2.10). By functional delta method, we have

$$
n^{1/2}(\hat{g}_0-g_0)=n^{1/2}(\mathbb{P}_n-\mathbb{P})\left[\frac{-1}{\log(S_0)}I(Z=0)\int_0^t\frac{dM(u|Z=0)}{E\left\{Y(u)I(Z=0)\right\}}\right]+o_p(1),
$$

$$
n^{1/2}(\hat{g}_1-g_1)=n^{1/2}(\mathbb{P}_n-\mathbb{P})\left[\frac{-1}{\log(S_1)}I(Z=1)\int_0^t\frac{dM(u|Z=1)}{E\left\{Y(u)I(Z=1)\right\}}\right]+o_p(1),
$$

and

$$
\begin{aligned}
n^{1/2}(\hat{g}_1^*-g_1^*)=n^{1/2}(\mathbb{P}_n-\mathbb{P})\Bigg[&\frac{-1}{\log(S_1^*)}\int_0^t\sum_{x=1}^K I(Z=1,X=x)\frac{\mathrm{pr}(X=x|T\geq u,Z=0)}{E\left\{Y(u)I(Z=1,X=x)\right\}}\\
&dM(u|Z=1,X=x)-\frac{1}{\log(S_1^*)}\int_0^t\sum_{x=1}^K Q(u,x)\lambda(u|Z=1,X=x)du\Bigg]+o_p(1).
\end{aligned}
$$

We combine the above results to obtain

$$
n^{1/2}(\hat{\phi}-\phi)=n^{1/2}(\mathbb{P}_n-\mathbb{P})\left\{\frac{g_1-g_1^*}{(g_1-g_0)^2}\eta_0+\frac{g_1^*-g_0}{(g_1-g_0)^2}\eta_1+\frac{1}{g_0-g_1}\eta_2\right\}+o_p(1),\qquad\text{(A.1)}
$$

where η_0,η_1,η_2 are given in equation (2.9). □

A.2 Proof of Theorem 2

Proof. First, we notice that $\log|V_\tau| = \sum_{i=1}^{p}\log(1+\tau^2/\sigma_i^2)$ and $1+\tau^2\xi_i, i=1,\ldots,p$ are the $p-1$ nonzero eigenvalues of $V_\tau P_0 = P_0+\tau^2\Sigma^{-1}P_0$. From Patterson and Thompson (1971) and Crainiceanu and Ruppert (2004), there exists a $p\times(p-1)$ matrix W such that $W^TW = I_{p-1}$, $WW^T = P_0$, $W^TV_\tau W = \mathrm{diag}\{(1+\tau^2\xi_i)\}$, and $Y^TP_\tau^TV_\tau^{-1}P_\tau Y = Y^TW\mathrm{diag}\{(1+\tau^2\xi_i)^{-1}\}W^TY$.

Let $w = W^TY$ and under H_0 we have $w \sim N_{p-1}(0, I_{p-1})$. Notice that $Y^TP_0Y = Y^TWW^TY = \sum_{i=1}^{p-1}w_i^2$. It follows that

$$LRT_p \overset{d}{=} \sup_{\tau^2}\left\{\sum_{i=1}^{p-1}w_i^2 - \sum_{i=1}^{p-1}\frac{w_i^2}{1+\tau^2\xi_i} - \sum_{i=1}^{p}\log\left(1+\frac{\tau^2}{\sigma_i^2}\right)\right\} + Y^T(I_p-P_0)Y.$$

For the projection matrix $I_p - P_0 = Z(Z^TZ)^{-1}Z^T$ with rank 1, there exists an $p\times 1$ matrix U such that $U^TU = 1$ and $UU^T = I_p - P_0$. Define $u = U^TY$ and under H_0, $u \sim N(0,1)$ and $Y^T(I_p-P_0)Y = u^2$. Also note that $\mathrm{cov}(u,w) = U^TW$. Let $A = U^TW$ then $UAW^T = (I_p-P_0)P_0 = 0$. It follows that $U^TUAW^TW = 0$ and $A = 0$. Therefore u and w are uncorrelated normal random variables, and therefore independent. Thus $Y^T(I_p-P_0)Y$ and Y^TP_0Y are independent random variables. Putting it together it follows that

$$LRT_p \overset{d}{=} \sup_{\tau^2}\left\{\sum_{i=1}^{p-1}\frac{\tau^2\xi_i}{1+\tau^2\xi_i}w_i^2 - \sum_{i=1}^{p}\log\left(1+\frac{\tau^2}{\sigma_i^2}\right)\right\} + u^2.$$

□

A.3 Proof of Theorem 3

Proof. Following the proof in section A.2, under the alternative hypothesis: $H_a : \mu = \mu_a, \tau^2 = \tau_a^2$, we have $Y \sim N_p(\mu_a Z, V_{\tau_a})$ and $W^TY \sim N_{p-1}(\mu_a W^TZ, W^TV_{\tau_a}W)$. Since $P_0Z = 0$, then $WW^TZ = 0$. Multiplying W^T leads to $W^TWW^TZ = 0$ and $W^TZ = 0$. Since $W^TV_\tau W = \mathrm{diag}\{(1+\tau^2\xi_i)\}$ then $W^TV_{\tau_a}W = \mathrm{diag}\{(1+\tau_a^2\xi_i)\}$. Thus $W^TY \sim N_{p-1}(0, \mathrm{diag}\{(1+\tau_a^2\xi_i)\})$. Denote by $w = \mathrm{diag}\{(1+\tau_a^2\xi_i)^{-1/2}\}W^TY$ and under H_a, $w \sim N_{p-1}(0, I_{p-1})$. Notice that

$$Y^TP_\tau^TV_\tau^{-1}P_\tau Y \overset{d}{=} \sum_{i=1}^{p-1}\frac{1+\tau_a^2\xi_i}{1+\tau^2\xi_i}w_i^2, \;\; Y^TP_0Y \overset{d}{=} \sum_{i=1}^{p-1}(1+\tau_a^2\xi_i)w_i^2.$$

Consider $v = U^T Y$ and under H_a, $v \sim N_1(\mu_a U^T Z, 1 + \tau_a^2 U^T \Sigma^{-1} U)$. Notice that $Y^T(I_p - P_0)Y = v^2$. Let $t = (v - \mu_a U^T Z)(1 + \tau_a^2 U^T \Sigma^{-1} U)^{-1/2}$ and t is standard normal random variable. Calculate $\text{cov}(w, v) = \tau_a^2 \text{diag}\{(1 + \tau_a^2 \xi_i)^{-1/2}\} W^T \Sigma^{-1} U$ and $\text{cov}(w, t) = \tau_a^2 (1 + \tau_a^2 U^T \Sigma^{-1} U)^{-1/2} \text{diag}\{(1 + \tau_a^2 \xi_i)^{-1/2}\} W^T \Sigma^{-1} U$, where we denote the last covariance as b. We have

$$\begin{pmatrix} w \\ t \end{pmatrix} \sim N_p \left(0, \begin{pmatrix} I_{p-1} & b \\ b^T & 1 \end{pmatrix} \right),$$

and we can write $t = w^T b + u(1 - b^T b)^{1/2}$ where u is an independent standard normal random variable and $(w^T, u)^T \sim N_p(0, I_p)$. Then v can be expressed as

$$v = \mu_a U^T Z + \left(1 + \tau_a^2 U^T \Sigma^{-1} U\right)^{1/2} \left\{ w^T b + u(1 - b^T b)^{1/2} \right\}.$$

This leads us to the desired conclusion. □

A.4 Proof of Corollary 1

Proof. Notice that the nonzero eigenvalues of $\Sigma^{-1} P_0$ are simply $\xi_i = 1/\sigma^2$, $i = 1, \ldots, p-1$. Let

$$f_w(\tau^2) = \frac{\tau^2}{\sigma^2 + \tau^2} \sum_{i=1}^{p-1} w_i^2 - p \log \left(1 + \frac{\tau^2}{\sigma^2} \right),$$

be the function we want to maximize over $\tau^2 \geq 0$. Taking the derivative with respect to τ^2 gives

$$\frac{\partial f_w}{\partial \tau^2} = \frac{p}{(\sigma^2 + \tau^2)^2} \left(\frac{\sigma^2}{p} \sum_{i=1}^{p-1} w_i^2 - \sigma^2 - \tau^2 \right).$$

The maximum is achieved at

$$\hat{\tau}^2 = \sigma^2 \left(\frac{1}{p} \sum_{i=1}^{p-1} w_i^2 - 1 \right)_+,$$

where $z_+ = \max(z, 0)$. Then under H_0,

$$\text{LRT}_p \stackrel{d}{=} f_w(\hat{\tau}^2) + u^2 = \left(R - p - p \log \frac{R}{p} \right) I\left(R > p\right) + u^2,$$

where $R = \sum_{i=1}^{p-1} w_i^2$.

Under H_a, the function to maximize is

$$f_w(\tau^2) = \left(1 + \frac{\tau_a^2}{\sigma^2}\right) \frac{\tau^2}{\sigma^2 + \tau^2} \sum_{i=1}^{p-1} w_i^2 - p \log\left(1 + \frac{\tau^2}{\sigma^2}\right).$$

Taking derivatives yields

$$\frac{\partial f_w}{\partial \tau^2} = \frac{p}{(\sigma^2 + \tau^2)^2} \left(\frac{\sigma^2 + \tau_a^2}{p} \sum_{i=1}^{p-1} w_i^2 - \sigma^2 - \tau^2 \right).$$

The maximum is achieved at

$$\hat{\tau}^2 = \left(\frac{\sigma^2 + \tau_a^2}{p} \sum_{i=1}^{p-1} w_i^2 - \sigma^2 \right)_+.$$

Thus

$$f_w(\hat{\tau}^2) = \left(R_{\tau_a} - p - p \log \frac{R_{\tau_a}}{p} \right) I(R_{\tau_a} > p),$$

where $R_{\tau_a} = (1 + \tau_a^2/\sigma^2) \sum_{i=1}^{p-1} w_i^2$. Notice that in this special case, $P_\tau = P_0 = I_p - \frac{1}{p} 1 1^T$, $U = \frac{1}{\sqrt{p}} \mathbf{1}$ and $W^T \mathbf{1} = 0$ so $b = 0$. Thus under H_a

$$\mathrm{LRT}_p \overset{d}{=} \left(R_{\tau_a} - p - p \log \frac{R_{\tau_a}}{p} \right) I(R_{\tau_a} > p) + \left\{ \frac{p^{1/2} \mu_a}{\sigma} + u \left(1 + \frac{\tau_a^2}{\sigma^2}\right)^{1/2} \right\}^2.$$

□

A.5 Proof of Theorem 4

Proof. We partition the fixed effect parameters $\mu = (\mu_1^T | \mu_2^T)^T$ where under the null hypothesis H_0, $\mu_2 = \mu_2^0$ are known. We also partition the design matrix into $Z = (Z_1 | Z_2)$. It is easy to show that the LRT statistic is

$$\begin{aligned} LRT_p = \sup_{\tau^2} & \left\{ Y^T P_0 Y - Y^T P_\tau^T V_\tau^{-1} P_\tau Y - \log |V_\tau| \right\} + (Y - Z_2 \mu_2^0)^T S_1 (Y - Z_2 \mu_2^0) \\ & - Y^T P_0 Y, \end{aligned} \tag{A.2}$$

where $P_\tau = I_p - Z(Z^T V_\tau^{-1} Z)^{-1} Z^T V_\tau^{-1}$, $P_0 = I_p - S_Z$, $S_Z = Z(Z^T Z)^{-1} Z^T$, $S_1 = I_p - S_{Z_1}$ and $S_{Z_1} = Z_1 (Z_1^T Z_1)^{-1} Z_1^T$. Notice that $(I_p - S_Z) Z_2 = 0$, so

$$Y^T P_0 Y = (Y - Z_2 \mu_2^0)^T (I_p - S_Z)(Y - Z_2 \mu_2^0) - (Z_2 \mu_2^0)^T Z_2 \mu_2^0.$$

Denote by $V = Y - Z_2\mu_2^0$ and equation (A.2) becomes

$$LRT_p = \sup_{\tau^2} \left\{ Y^T P_0 Y - Y^T P_\tau^T V_\tau^{-1} P_\tau Y - \log|V_\tau| \right\} + V^T (S_Z - S_{Z_1}) V + (Z_2\mu_2^0)^T Z_2 \mu_2^0.$$

Let $\xi_i, i = 1, \ldots, p-q$, be the $p-q$ non-zero eigenvalues of the matrix $\Sigma^{-1}P_0$, so $1 + \tau^2\xi_i$ are the $p-q$ nonzero eigenvalues of $V_\tau P_0 = P_0 + \tau^2\Sigma^{-1}P_0$. From Patterson and Thompson (1971) and Crainiceanu and Ruppert (2004), there exists a $p \times (p-q)$ matrix W such that $W^TW = I_{p-q}$, $WW^T = P_0$, $W^TV_\tau W = \text{diag}\{(1+\tau^2\xi_i)\}$ and

$$Y^T P_\tau^T V_\tau^{-1} P_\tau Y = Y^T W \text{diag}\{(1+\tau^2\xi_i)^{-1}\} W^T Y.$$

Let $w = W^TY$ and under H_{0r}, we have $E(w) = W^TZ_1\mu_1 + W^TZ_2\mu_2^0$ and $\text{cov}(w) = I_{p-q}$. Denote by $A = W^TZ$ and notice that $WA = P_0Z = 0$. It follows that $W^TWA = 0$ and $A = 0$. Thus we have $w \sim N_{p-q}(0, I_{p-q})$. Since $S_Z - S_{Z_1}$ is a projection matrix, there exists a $q \times (q-k)$ matrix U such that $UU^T = S_Z - S_{Z_1}$ and $U^TU = I_{q-k}$. Let $u = U^TV$ and under H_{0r}, we have $E(u) = U^TZ_1\mu_1$ and $\text{cov}(u) = I_{q-k}$. Denote by $B = U^TZ_1$ and notice that $UB = (S_Z - S_{Z_1})Z_1 = 0$. It follows that $U^TUB = 0$ and $B = 0$. Thus we have $u \sim N_{q-k}(0, I_{q-k})$. Also, notice that $\text{cov}(u, w) = U^TW$ and denote by $C = U^TW$, so we have $UCW^T = (S_Z - S_{Z_1})P_0 = 0$. Thus $U^TUCW^TW = 0$ and $C = 0$. It follows that $(u^T, w^T) \sim N_{p-k}(0, I_{p-k})$ and all entries are independent standard normal random variables. To conclude,

$$LRT_p \overset{d}{=} \sup_{\tau^2} \left\{ \sum_{i=1}^{p-q} \frac{\tau^2\xi_i}{1+\tau^2\xi_i} w_i^2 - \sum_{i=1}^{p} \log\left(1 + \frac{\tau^2}{\sigma_i^2}\right) \right\} + \sum_{i=1}^{q-k} u_i^2 + (Z_2\mu_2^0)^T Z_2 \mu_2^0.$$

□

A.6 Score projection

The score statistic for μ under $H_0 : \mu = 0, \tau^2 = 0$ is

$$U_\mu = (GZ)^T(Y - \tilde{\pi}),$$

where $\tilde{\pi}$ is the estimated mean of Y under H_0. The score statistic for τ^2 under $H_0 : \mu = 0, \tau^2 = 0$ is

$$S'_{\tau^2} = (Y - \tilde{\pi})^T GG^T(Y - \tilde{\pi}) = U'^T_{\tau^2} U'_{\tau^2},$$

where $U'_{\tau^2} = G^T(Y - \tilde{\pi}) \approx G^T P_1(Y - \pi)$ using the approximation $Y - \tilde{\pi} \approx P_1(Y - \pi)$, where $P_1 = I - DX(X^T DX)^{-1}X^T$. However, U_μ and U'_{τ^2} are not independent. We can remove the projection of U'_{τ^2} on U_μ so that the modified score statistic for τ^2 is asymptotically independent of the score statistic for μ. Note that the projection of a random vector A onto a random vector B is $C = \text{cov}(A, B)\text{var}(B)^{-1}B$. So the projection of $A = G^T P_1(Y - \pi)$ onto $B = (GZ)^T P_1(Y - \pi)$ is given by

$$\begin{aligned} C &= \text{cov}\left\{G^T P_1(Y - \pi), (GZ)^T P_1(Y - \pi)\right\} \text{var}\left\{(GZ)^T P_1(Y - \pi)\right\}^{-1} (GZ)^T P_1(Y - \pi) \\ &= G^T P_1 DP_1^T GZ \left\{(GZ)^T P_1 DP_1^T (GZ)\right\}^{-1} (GZ)^T P_1(Y - \pi). \end{aligned}$$

Thus the modified score for τ^2

$$A - C = G^T \left[I - P_1 DP_1^T GZ \left\{(GZ)^T P_1 DP_1^T (GZ)\right\}^{-1} (GZ)^T\right] P_1(Y - \pi).$$

is asymptotically independent of A.

Proposition 2. *The modified score for τ^2, $A - C$, is the same as the score of τ^2 under $\tau^2 = 0$ without the restriction of $\mu = 0$, namely, $U_{\tau^2} = G^T(Y - \hat{\pi}) \approx G^T P_2(Y - \pi)$, where $\hat{\pi}$ if the mean of Y under $\tau^2 = 0$, $P_2 = I - DM(M^T DM)^{-1}M^T$ and $M = (X, GZ)$. Another word to say is that*

$$A - C = G^T P_2(Y - \pi).$$

Proof. We need to show that

$$P1 - P_1 DP_1^T GZ \left\{(GZ)^T P_1 DP_1^T (GZ)\right\}^{-1} (GZ)^T P_1 = P_2. \tag{A.3}$$

Denote by $A = (X^T DX)^{-1}$, $B = X^T D(GZ)$, $C = (GZ)^T DX$ and $E = (GZ)^T D(GZ)$. The left side of the equation (A.3) can be simplified to

$$\begin{aligned}\text{left} = I &- DXA^{-1}X - D(GZ)\left\{E - CA^{-1}B\right\}^{-1}(GZ)^T + DXA^{-1}B\left\{E - CA^{-1}B\right\}^{-1}(GZ)^T \\ &+ D(GZ)\left\{E - CA^{-1}B\right\}^{-1}CA^{-1}X^T - DXA^{-1}B\left\{E - CA^{-1}B\right\}^{-1}CA^{-1}X^T.\end{aligned}$$

According to block matrix inversion formula, the right side is

$$\begin{aligned}\text{right} &= I - D\begin{pmatrix} X & GZ \end{pmatrix}\begin{pmatrix} A & B \\ C & E \end{pmatrix}^{-1}\begin{pmatrix} X^T \\ (GZ)^T \end{pmatrix} \\ &= I - DXF_{11}X^T - D(GZ)F_{21}X^T - DXF_{12}(GZ)^T - D(GZ)F_{22}(GZ)^T.\end{aligned}$$

where

$$\begin{aligned}F_{11} &= A^{-1} + A^{-1}B\left\{E - CA^{-1}B\right\}^{-1}CA^{-1}, \\ F_{12} &= -A^{-1}B\left\{E - CA^{-1}B\right\}^{-1}, \\ F_{21} &= -\left\{E - CA^{-1}B\right\}^{-1}CA^{-1}, \\ F_{22} &= \left\{E - CA^{-1}B\right\}^{-1}.\end{aligned}$$

After some algebra, we see that the right side is equal to the left side. This completes the proof. □

A.7 Proof of Proposition 1

(a) We want to show that U_μ and U_{τ^2} are asymptotically independent under H_0. We only need to show that $\text{cov}(U_\mu, U_{\tau^2}) = (GZ)^T P_1 D P_2^T G$ is zero. Notice that $M^T P_2 D = 0$, indicating that $X^T P_2 D = 0$ and $(GZ)^T P_2 D = 0$. Also we have $P_2 D = D P_2^T$ so $X^T D P_2^T = 0$ and $(GZ)^T D P_2^T = 0$. Calculate

$$(GZ)^T P_1 D P_2^T G = (GZ)^T (I - DX(X^T DX)^{-1}X^T) D P_2^T G = (GZ)^T D P_2^T G = 0.$$

Thus, S_μ and S_{τ^2} are asymptotically independent under H_0.

(b) Notice that $S_\rho = (Y-\pi)^T D^{-1/2} D^{1/2} K_\rho D^{1/2} D^{-1/2} (Y-\pi)$ and $D^{-1/2}(Y-\pi) \sim N_n(0, I_n)$ asymptotically under H_0. Thus, under H_0 $S_\rho \sim \sum_{i=1}^n \lambda_i \chi^2_{1i}$ where λ_i is the eigenvalues of $D^{1/2} K_\rho D^{1/2}$ for $i = 1, \ldots, n$.

(c) We calculate

$$\begin{aligned} S_\rho &= (Y - \pi_a + \pi_a - \pi)^T D_a^{-1/2} D_a^{1/2} K_\rho D_a^{1/2} D_a^{-1/2} (Y - \pi_a + \pi_a - \pi) \\ &= \left[Z + D_a^{-1/2}(\pi_a - \pi)\right]^T \tilde{K}_\rho \left[Z + D_a^{-1/2}(\pi_a - \pi)\right], \end{aligned}$$

where $Z = D_a^{-1/2}(Y - \pi_a) \sim N(0, I_n)$ asymptotically under H_a and $\tilde{K}_\rho = D_a^{1/2} K_\rho D_a^{1/2}$. By the spectral decomposition of $\tilde{K}_\rho = U \Lambda U^T$, $S_\rho \sim \sum_{i=1}^n \eta_i \chi^2_{1,i}(b_i^2)$ asymptotically under H_a, where η_i is the diagonal of $\Lambda, i = 1, \ldots, n$, and $b = (b_1, \ldots, b_n)^T = U^T D_a^{-1/2}(\pi_a - \pi)$.

A.8 Approximation of mixture of chi-squared distribution

Consider the mixture of chi-squared distribution $Q(X) = \sum_{i=1}^n \lambda_i \chi^2_{1i}(\delta_i)$, where $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$ and $\delta_i \geq 0$. Let $c_k = \sum_{i=1}^n \lambda_i^k + k \sum_{i=1}^n \lambda_i^k \delta_i$ be the kth moment of $Q(X)$. The mean, standard deviation, skewness and kurtosis of $Q(X)$ are given by

$$\mu_Q = c_1, \quad \sigma_Q = \sqrt{2c_2}, \quad \beta_1 = \sqrt{8} s_1, \quad \beta_2 = 12 s_2,$$

where $s_1 = c_3^2 / c_2^{3/2}$ and $s_2 = c_4 / c_2^2$. Liu et al (2009) proposed using a non-central $\chi^2_l(\delta)$ distribution to approximate $Q(X)$.

$$\begin{aligned} P(Q(X) > t) &= P\left(\frac{Q(X) - \mu_Q}{\sigma_Q} > t^*\right) \\ &\approx P\left(\frac{\chi^2_l(\delta) - \mu_X}{\sigma_X} > t^*\right) \\ &= P(\chi^2_l(\delta) > t^* \sigma_X + \mu_X), \end{aligned}$$

where $t^* = (t - \mu_Q)/\sigma_Q$, $\mu_X = \mathrm{E}(\chi^2_l(\delta)) = l + \delta$, $\sigma_X = \sqrt{\mathrm{var}(\chi^2_l(\delta))} = \sqrt{2a}$ and $a = \sqrt{l + 2\delta}$. The parameters l and δ are determined such that the skewness of $Q(X)$ and $\chi^2_l(\delta)$ are equal

and the difference between the kurtosis of $Q(X)$ and $\chi^2_l(\delta)$ is minimized. The solution is given by the following. If $s_1^2 > s_2$,

$$a = 1/\left(s_1 - \sqrt{s_1^2 - s_2}\right), \quad \delta = s_1 a^3 - a^2, \quad l = a^2 - 2\delta.$$

If $s_1^2 \leq s_2$,

$$\delta = 0, \quad l = c_2^3/c_3^2.$$

A.9 Accuracy of the power formula

From Figure A.1, the analytical power based on proposition 1 is very similar to the empirical power, therefore the analytical power formula is accurate.

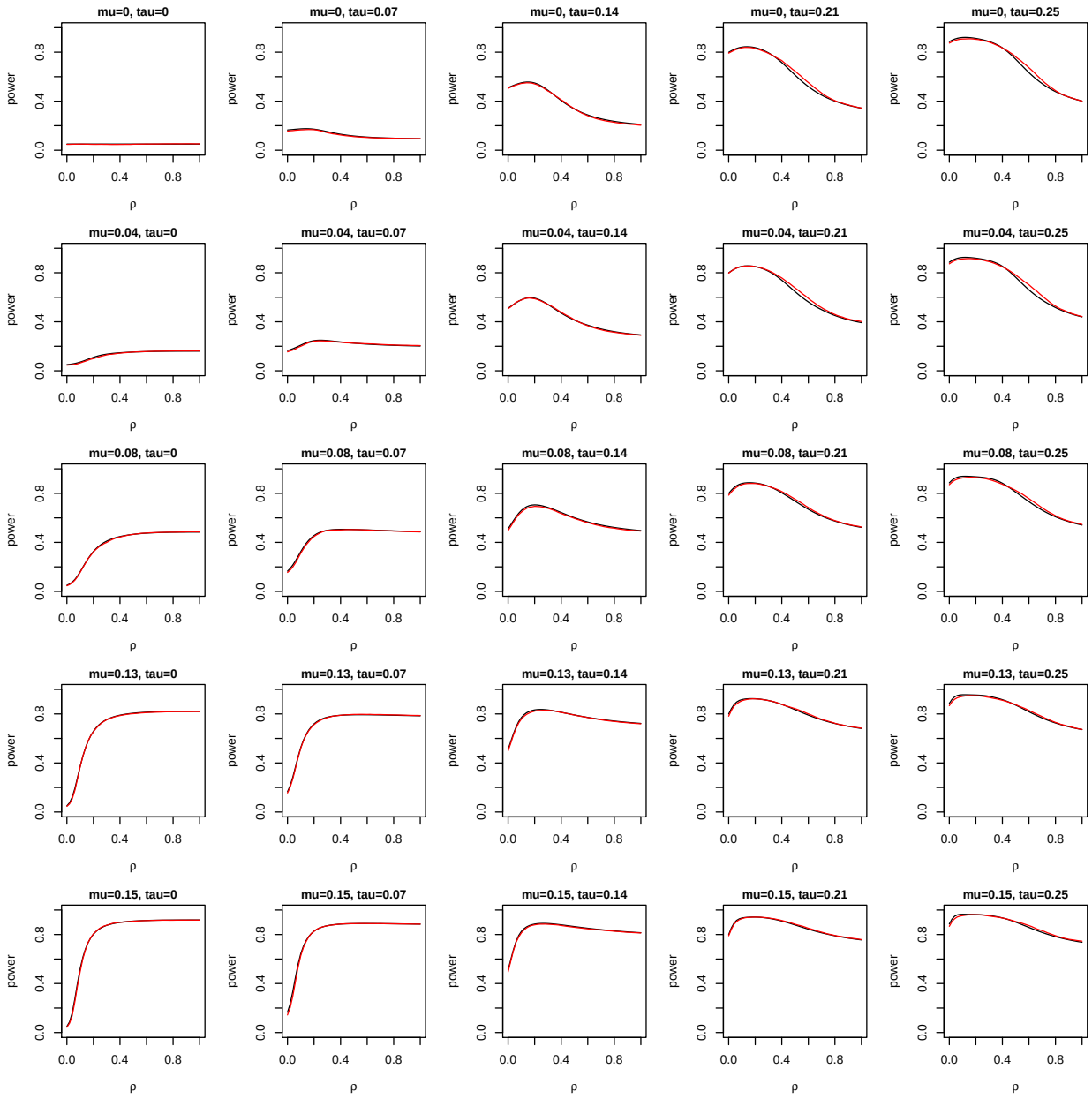


Figure A.1: Continuous trait. Black line is the analytical power and red line is the empirical power.

A.10 Independence of scores under the general setting

Consider a likelihood function (can also be partial likelihood or pseudo likelihood) $L(\alpha, \beta; y)$, where y is the data, α is a m dimensional nuisance parameter and β is a p dimensional parameter of interest. Sometimes we use a mixed-effects model for β to combine information, $\beta = Z\mu + \delta$, where Z is a $p \times q$ design matrix, μ is a q dimensional fixed-effects parameter and δ is a p dimensional random effect. Typically we assume δ follows a multivariate Gaussian distribution with mean 0 and covariance matrix $\tau^2\Omega$, in which Ω is known and τ^2 is a random-effects variance component. Let $\ell(\alpha, \mu, \delta)$ be the log likelihood based on α and $\beta = Z\mu + \delta$. We use the log marginal likelihood for inference, namely,

$$\ell_m(\alpha, \mu, \tau^2) = \log \left[\int_\delta \exp\left\{\ell(\alpha, \mu, \delta)\right\} \exp\left(-\frac{1}{2\tau^2} \delta^T \Omega^{-1} \delta \right) d\delta \right].$$

Denote by $\tilde{\alpha}$ the MLE under $H_0 : \mu = 0, \tau^2 = 0$. Denote by $\hat{\alpha}, \hat{\mu}$ be the MLE under $\tau^2 = 0$. The score for μ under $H_0 : \mu = 0, \tau^2 = 0$ is

$$\left.\frac{\partial \ell_m}{\partial \mu}\right|_{\alpha=\tilde{\alpha},\mu=0,\tau^2=0} = Z^T \left.\frac{\partial \ell}{\partial \beta}\right|_{\alpha=\tilde{\alpha},\mu=0,\delta=0} = Z^T \frac{\partial \ell}{\partial \beta}(\tilde{\alpha}, 0, 0).$$

The score for τ^2 under $\tau^2 = 0$ is

$$\left.\frac{\partial \ell_m}{\partial \tau^2}\right|_{\alpha=\hat{\alpha},\mu=\hat{\mu},\tau^2=0} = \left.\frac{\partial \ell}{\partial \beta^T} \Omega \frac{\partial \ell}{\partial \beta}\right|_{\alpha=\hat{\alpha},\mu=\hat{\mu},\delta=0} = \frac{\partial \ell}{\partial \beta^T}(\hat{\alpha}, \hat{\mu}, 0) \Omega \frac{\partial \ell}{\partial \beta}(\hat{\alpha}, \hat{\mu}, 0).$$

The next proposition indicates the independence of two score statistics under the general setting.

Proposition 3.

$$cov\left\{ \frac{\partial \ell}{\partial \beta}(\hat{\alpha}, \hat{\mu}, 0), Z^T \frac{\partial \ell}{\partial \beta}(\tilde{\alpha}, 0, 0) \right\} = o_p(1).$$

Proof. From Taylor expansion,

$$\frac{\partial \ell}{\partial \beta}(\hat{\alpha}, \hat{\mu}, 0) = \frac{\partial \ell}{\partial \beta}(\tilde{\alpha}, 0, 0) + \begin{pmatrix} \frac{\partial^2 \ell}{\partial \beta^2}(\tilde{\alpha}, 0, 0) Z & \frac{\partial^2 \ell}{\partial \beta \partial \alpha} \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\alpha} - \tilde{\alpha} \end{pmatrix} + O_p(1). \tag{A.4}$$

Since $(\hat{\alpha}, \hat{\mu})$ are the MLE under $\tau^2 = 0$, using Taylor expansion we have

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\partial \ell}{\partial \mu}(\hat{\alpha}, \hat{\mu}, 0) \\ \frac{\partial \ell}{\partial \alpha}(\hat{\alpha}, \hat{\mu}, 0) \end{pmatrix} = \begin{pmatrix} \frac{\partial \ell}{\partial \mu}(\tilde{\alpha}, 0, 0) \\ \frac{\partial \ell}{\partial \alpha}(\tilde{\alpha}, 0, 0) \end{pmatrix} + \begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu^2}(\tilde{\alpha}, 0, 0) & \frac{\partial^2 \ell}{\partial \mu \partial \alpha}(\tilde{\alpha}, 0, 0) \\ \frac{\partial^2 \ell}{\partial \alpha \partial \mu}(\tilde{\alpha}, 0, 0) & \frac{\partial^2 \ell}{\partial \alpha^2}(\tilde{\alpha}, 0, 0) \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\alpha} - \tilde{\alpha} \end{pmatrix} + O_p(1).$$

We simplify the notation and ignore the evaluation at $(\tilde{\alpha}, 0, 0)$. Thus

$$\begin{pmatrix} \hat{\mu} \\ \hat{\alpha} - \tilde{\alpha} \end{pmatrix} = - \begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \alpha} \\ \frac{\partial^2 \ell}{\partial \alpha \partial \mu} & \frac{\partial^2 \ell}{\partial \alpha^2} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial \ell}{\partial \mu} \\ \frac{\partial \ell}{\partial \alpha} \end{pmatrix} + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{A.5}$$

Substituting equation (A.5) into equation (A.4) gives

$$\frac{\partial \ell}{\partial \beta}(\hat{\alpha}, \hat{\mu}, 0) = \frac{\partial \ell}{\partial \beta} - \begin{pmatrix} \frac{\partial^2 \ell}{\partial \beta^2} Z & \frac{\partial^2 \ell}{\partial \beta \partial \alpha} \end{pmatrix} \begin{pmatrix} Z^T \frac{\partial^2 \ell}{\partial \beta^2} Z & Z^T \frac{\partial^2 \ell}{\partial \beta \partial \alpha} \\ \frac{\partial^2 \ell}{\partial \alpha \partial \beta} Z & \frac{\partial^2 \ell}{\partial \alpha^2} \end{pmatrix}^{-1} \begin{pmatrix} Z^T \frac{\partial \ell}{\partial \beta} \\ \frac{\partial \ell}{\partial \alpha} \end{pmatrix} + O_p(1),$$

where the right side is evaluated at $(\tilde{\alpha}, 0, 0)$ under H_0. Let

$$\begin{pmatrix} I_{\beta\beta} & I_{\beta\alpha} \\ I_{\alpha\beta} & I_{\alpha\alpha} \end{pmatrix} = \begin{pmatrix} E\left\{-\frac{\partial^2 \ell}{\partial \beta^2}\right\} & E\left\{-\frac{\partial^2 \ell}{\partial \beta \partial \alpha}\right\} \\ E\left\{-\frac{\partial^2 \ell}{\partial \alpha \partial \beta}\right\} & E\left\{-\frac{\partial^2 \ell}{\partial \alpha^2}\right\} \end{pmatrix}$$

be the Fisher information matrix evaluated at $(\tilde{\alpha}, 0, 0)$ under H_0. Then

$$\frac{\partial \ell}{\partial \beta}(\hat{\alpha}, \hat{\mu}, 0) = \frac{\partial \ell}{\partial \beta} - \begin{pmatrix} I_{\beta\beta} Z & I_{\beta\alpha} \end{pmatrix} \begin{pmatrix} Z^T I_{\beta\beta} Z & Z^T I_{\beta\alpha} \\ I_{\alpha\beta} Z & I_{\alpha\alpha} \end{pmatrix}^{-1} \begin{pmatrix} Z^T \frac{\partial \ell}{\partial \beta} \\ \frac{\partial \ell}{\partial \alpha} \end{pmatrix} + O_p(1). \tag{A.6}$$

Thus

$$\begin{aligned} \operatorname{cov}\left\{ \frac{\partial \ell}{\partial \beta}(\hat{\alpha}, \hat{\mu}, 0), Z^T \frac{\partial \ell}{\partial \beta}(\tilde{\alpha}, 0, 0) \right\} &= I_{\beta\beta} Z - \begin{pmatrix} I_{\beta\beta} Z & I_{\beta\alpha} \end{pmatrix} \begin{pmatrix} Z^T I_{\beta\beta} Z & Z^T I_{\beta\alpha} \\ I_{\alpha\beta} Z & I_{\alpha\alpha} \end{pmatrix}^{-1} \begin{pmatrix} Z^T I_{\beta\beta} Z \\ I_{\alpha\beta} Z \end{pmatrix} + o_p(1) \\ &= I_{\beta\beta} Z - \begin{pmatrix} I_{\beta\beta} Z & I_{\beta\alpha} \end{pmatrix} \begin{pmatrix} I \\ 0 \end{pmatrix} + o_p(1) = o_p(1) \end{aligned}$$

□