Methods for Agnostic Statistical Inference

Kendrick Qijun Li

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

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Kenneth M. Rice, Chair

Lurdes Y. T. Inoue

Noah Simon

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Methods for Agnostic Statistical Inference

Kendrick Qijun Li

Chair of the Supervisory Committee:
Professor Kenneth M. Rice
University of Washington, Department of Biostatistics

A traditional goal of parametric statistics is to estimate some or all of a data-generating model's finite set of parameters, thereby turning data into scientific insights. Point estimates of parameters, and corresponding standard error estimates are used to quantify the information provided by the data. However, in reality the true data-generation rarely follows the assumed model. When the model assumptions are incorrect, though the point estimates' target parameters are often still meaningful quantities, the model-based standard error estimates may be difficult to interpret in any helpful way; they may also be notably biased.

In light of doubts about model assumptions and the known difficulties of checking models, *agnostic* statistics aims to develop statistical inference with only minimal assumptions. Though agnostic statistics has become popular over the past few decades, methods of agnostic inference are yet to be developed in some fundamental application areas.

One such area is meta-analysis. Meta-analysis of 2×2 tables is common and useful in research topics including analysis of adverse events and survey research data. Fixed-effects inference typically centers on measures of association such as the Cochran-Mantel-Haenszel statistic or Woolf's estimator, but to obtain well-calibrated inference when studies are small most methods rely on assuming exact homogeneity across studies, which is often unrealistic. By showing that estimators of several widely-used methods have meaningful estimands even in the presence of heterogeneity, we derive improved confidence intervals for them un-

der heterogeneity. These improvements over current methods are illustrated by simulation. We find that our confidence intervals provide coverage closer to the nominal level when heterogeneity is present, in both small and large-sample settings. The conventional confidence intervals derived under homogeneity are often conservative, though anti-conservative inferences occur in some scenarios. We also apply the proposed methods to a meta-analysis of 19 randomized clinical trials on the effect of sclerotherapy in preventing first bleeding for patients with cirrhosis and esophagogastric varices. Our methods provide a more interpretable approach to meta-analyzing binary data and more accuracy in characterizing the uncertainty of the estimates.

Another area lacking agnostic methods is adaptive shrinkage estimation. Shrinkage estimation attempts to increase the precision of an estimator in exchange for introducing a modest bias. Standard shrinkage estimators in linear models include the James-Stein estimator, Ridge estimator, and LASSO. However, theories regarding the optimal amount of shrinkage and statistical properties of these estimators are often based on stringent distributional assumptions. In Chapter 3, we provide a unified framework of shrinkage estimation – penalized precision-weighted least square estimation. We demonstrate that the James-Stein estimator, Ridge and LASSO are all penalized precision-weighted least-square estimators using model-based precision weights. Using a model-agnostic precision weighting matrix, we propose three shrinkage estimators in the novel framework: Rotated James-Stein estimator, Rotated Ridge, and Rotated LASSO. As we show, the three proposed estimators have theoretical properties and empirical performance that are comparable to the standard shrinkage estimators, while rotated LASSO has improved precision in some situations. We apply these estimators in a prostate cancer example.

The third area is variance estimation in Bayesian inference. Many frequentist parametric statistical methods have large sample Bayesian analogs. However, there is no general Bayesian analog of "robust" covariance estimates, that are widely-used in frequentist work. In Chapter 4, we propose such an analog, produced as the Bayes rule under a form of balanced loss function. This loss combines standard parametric inference's goal of accurate

estimation of the truth with less-standard fidelity of the data to the model. Besides being the large-sample equivalent of its frequentist counterpart, we show by simulation that the Bayesian robust standard error can also be used to construct Wald confidence intervals that improve small-sample coverage. We demonstrate the novel standard error estimates in a Bayesian linear regression model to study the association between systolic blood pressure and age in 2017-2018 NHANES data.

TABLE OF CONTENTS

Page

List of Figures iv

List of Tables vi

Chapter 1: Introduction 1

Chapter 2: Improved inference for fixed-effects meta-analysis of 2×2 tables 3

2.1 Introduction 3

2.2 Notation 4

2.3 Results 8

2.4 Numerical studies 12

2.5 Real Data Example: Sclerotherapy 13

2.6 Discussion 16

Chapter 3: Shrinkage Estimation in Linear Models Using Penalized Precision-Weighted Least Square Estimators 20

3.1 Introduction 20

3.2 Penalized Precision-Weighted Least Square Estimators 21

3.3 Rotated James-Stein Estimator, Ridge Estimator, and LASSO 24

3.4 Computation of rLASSO 27

3.5 Consistency and Asymptotic Distribution 27

3.6 Sign Consistency of rLASSO 29

3.7 Large Sample Behavior of rJS, rRidge, and rLASSO under Nonlinearity . . . 30

3.8 Selecting the Regularizing Parameters 32

3.9 Simulation Study: Point Estimates 33

3.10 Simulation Study: Sign Consistency 35

3.11 Data Application: Prostate Cancer Data 38

3.12 Discussion 43

Chapter 4: Bayesian Variance Estimation and Hypothesis Testing Using Inference Loss Functions . . . 45
4.1 Introduction . . . 45
4.2 Inference Loss functions . . . 46
4.3 Balanced Inference Loss functions for model-robust variance estimation . . . 47
4.4 Examples and simulation studies . . . 49
4.5 Balanced Loss Function for quasi-likelihood regression . . . 56
4.6 Application: Association of Systolic Blood Pressure and Age in 2017-2018 NHANES Data . . . 58
4.7 Discussion . . . 59

Chapter 5: Conclusion . . . 62

Bibliography . . . 63

Appendix A: Large-sample Properties of the Estimators . . . 71
A.1 log-CMH and Woolf results . . . 71
A.2 MLE results . . . 71

Appendix B: Meta-analysis of sclerotherapy studies . . . 76

Appendix C: Zero-cell correction . . . 77

Appendix D: Presentation of James-Stein estimator, LASSO and Ridge estimator and as precision-weighted least-square estimators . . . 78

Appendix E: Proofs of theorems about the large-sample properties of the precision-weighted least square estimators . . . 80
E.1 Proof of Theorem 3.5.1 . . . 80
E.2 Proof of Theorem 3.5.2 . . . 81
E.3 Proof of Theorem 3.6.1 . . . 82
E.4 Proof of Theorem 3.7.1 . . . 84
E.5 Proof of Theorem 3.7.2 . . . 84

Appendix F: Bayes rules for the inference loss function . . . 86

Appendix G: Bayes rule for the Balanced Inference Loss . . . 87

Appendix H: Asymptotic equivalence of Bayesian robust covariance matrix and the large-sample covariance matrix of the Bayes point estimate 88

Vita 91

LIST OF FIGURES

Figure Number Page

2.1 Coverage of the confidence intervals, derived under homogeneity or potential heterogeneity, with varying baseline event probabilities and effect sizes in each study. The vertical dashed line shows the point where $\psi_1 = \psi_2$, i.e. where the effect sizes are homogeneous. 15

2.2 Forest plots of meta-analysis of sclerotherapy studies, [46] with confidence intervals of log-CMH statistics, Woolf's estimator, MLE, derived with or without assuming homogeneity. We include the confidence interval using DerSimonian-Laird method (DSL) as comparison. We also show the chi-square statistics and the p-values for testing the value of chi-square statistics against a chi-square distribution with 1 degree of freedom. 17

3.1 Illustration of Rotated James-Stein estimator (left), Rotated LASSO (middle) and Rotated Ridge estimator (right) in the two-dimension case. 26

3.2 Predictive risks of the ordinary least square estimator (LS), James-Stein estimator (JS), the Positive-Part James-Stein estimator (JS+), LASSO, Ridge, Rotated James-Stein estimator (rJS), Rotated LASSO (rLASSO), and Rotated Ridge (rRidge). See the text for detailed simulation settings. 36

3.3 L_2-loss of the ordinary least square estimator (LS), James-Stein estimator (JS), the Positive-Part James-Stein estimator (JS+), LASSO, Ridge, Rotated James-Stein estimator (rJS), Rotated LASSO (rLASSO), and Rotated Ridge (rRidge). See the text for detailed simulation settings. 37

3.4 Sign consistency of LASSO and rLASSO. (a), (c), (e): Frequencies of LASSO or rLASSO solution path containing the true support versus the rLASSO sign consistency criterion $\eta_1 = 1 - \|D_{21}(D_{11})^{-1}\text{sign}(\beta_S)\|_\infty$, with the variance function $\sigma^2(\cdot)$ given by (i), (ii) or (iii); (b), (d), (f): Relationship of LASSO sign consistency criterion $\eta_0 = 1 - \|C_{21}(C_{11})^{-1}\text{sign}(\beta_S)\|_\infty$ versus rLASSO sign consistency criterion η_1. 39

3.5 Scatter plots of the outcome versus each covariate in the prostate dataset, with univariate LOESS curves. The outcome is standardized to have mean zero; all the covariates are normalized to have mean zero and variance one. BPH: benign prostatic hyperplasia amount; SVI: seminal vesicle invasion; CP: Capsular penetration. 41

3.6 Solution trajectories relative for scaled OLS, rJS, LASSO, Ridge, rLASSO and rRidge relative to the regression coefficient of log(PSA) using the training dataset. 42

4.1 Coverage probabilities of 95% credible interval (red), frequentist (blue) and Bayesian (green) robust confidence intervals. From top to bottom: linear regression, Poisson regression and exponential proportional hazards model. The middle column shows the results with correctly-specified models. θ_0^* denotes the limit of the estimator of interest in each scenario. 57

4.2 Frequentist and Bayesian regression models for analyzing the association between systolic blood pressure and age, adjusting for subjects' gender. SBP: systolic blood pressure; MALE: the indicator for male; AGE: subjects age in years. 59

LIST OF TABLES

Table Number Page

2.1 Notation for a subtable j contributing to a meta-analysis 5

2.2 Estimands and coverage probability of confidence intervals for logarithm of Cochran-Mantel-Haenszel statistics (log-CMH, $\hat{\psi}_{CMH}$), Woolf's estimator ($\hat{\psi}_{Woolf}$) and MLE of a common odds ratio ($\hat{\psi}_{MLE}$), with varying log-odds ratios of the studies. The confidence intervals are the established methods that are derived under homogeneity, and our proposed approaches deriverd under heterogeneity. The estimands are computed using the true values of the parameters. Coverage probabilities are computed over 10,000 replications. 14

3.1 Point estimates, training error and testing error of the ordinary least square estimator (LS), James-Stein estimator (JS), the Positive-Part James-Stein estimator (JS+), LASSO, Ridge, Rotated James-Stein estimator (rJS), Rotated LASSO (rLASSO), and Rotated Ridge (rRidge). We also showed the optimal regularization λ_n of LASSO, Ridge, rJS, rLASSO, and rRidge by three-fold cross validation. 43

4.1 Comparison between posterior standard deviation and the Bayesian robust standard error to the true standard error of the Bayes point estimate $\hat{d}_n$ in linear regression. Gray rows indicate where the model is correctly specified. . 52

4.2 For a Poisson model, comparison of the posterior standard deviation and the Bayesian robust standard error to the true standard error of the posterior mean $\hat{d}_n$. Gray rows indicate where the model is correctly specified. 53

4.3 For an Exponential Proportional Hazards Model, comparison of the posterior standard deviation and the Bayesian robust standard error to the true standard error of the posterior mean $\hat{d}_n$. Gray rows indicate where the model is correctly specified. #events: Average number of events. 55

4.4 Point estimates and standard error estimates of regression coefficients in Frequentist and Bayesian linear regression models using the NHANES data. Post. SD: posterior standard deviation; BRSE: Baysian robust standard error. 60

B.1 Meta-analysis of 19 studies on the effectiveness of sclerotherapy in preventing first bleeding in cirrhosis[46]. Data are shown as number of first bleeding cases/sample sizes. 76

ACKNOWLEDGMENTS

I want to express my gratitude to my dissertation adviser and mentor Professor Kenneth Rice. Thank you for all your time, guidance, suggestions, and your unchanging support even when I am in doubt. I would like to thank the members of my reading committee Lurdes Inoue and Noah Simon, for their feedback, encouragement and precious mentoring throughout my research and study. I would like to thank my GSR Professor Nicholas Smith for your generous time and support. I would like to thank my RA advisors and colleagues Peter Gilbert, Yunda Huang, Youyi Fong, David Benkerser, among many others. The experience of working with you greatly enriched my understanding in applied statistics. I would like to express my sincere appreciation to the University of Washington, for providing amazing support to many international students. To the Department of Biostatistics, for giving me this great opportunity. To all the amazing faculty in the Department of Biostatistics and Department of Statistics, specially to Gary Chan, Mauricio Sadinle, Jon Wellner, Marco Crone, Michael Perlman, Thomas Richardson, and Yen-Chi Chen. Special thanks to Gitana Garofalo, who has always been providing amazing support to all the students. Personally, I would like to thank my family, who I regret to not spend more time with for years, but have always been supportive and happy for my achievements. To my amazing housemates Phuong Vu and Natalie Gasca, who have always inspired and supported me. To Adam Elder and Subodh Selukar, with whom I shared so many happy and encouraging memories over the years. To Jiacheng Wu, who has supported me during my worst moments. To Parker Xie, for your wise advise. To Hongxiang Qiu, who has awed me for your calmness and excellence. To Aaron Hudson, whose diligence and thoughtfulness have always encouraged me to be a more grounded student and a more detailed thinker. To Amarise Little, for your gentleness and elegance and greater strength inside. To Emily Voldal, whose legendary cookies bring all of us joy. To Arjun Sondhi, Brian Williamson, Kelsey Grinde, Tracy Dong,

Angela Zhang, Serge Aleshin-Guendel, Maria Alejandra Valdez Cabrera, and many many others, who have given me precious friendship, company, advice, and other amazing gifts.

.

DEDICATION

to my dear grandmother, Huijuan Wang, to whom I wish the best health and happiest life

Chapter 1

INTRODUCTION

Statistics aims to get insights from data to a scientific problem. A common approach is to assume the data follow a probabilistic model that is characterized by a few parameters, to estimate the parameters using the observed data, and to make interpretations based on the parameter estimate and their uncertainty measures. This approach is often referred to as parametric statistics [10, §7].

However, more often than not the assumed statistical model does not accurately characterize the data-generating distribution. Although methods exist to detect the discrepancies between the data and an assumed model, such as the Pearson's Chi-squared test [49], the Kolmogorov-Smirnov test [42] and others, the practice of using the test for model assumptions to guide estimation has long been controversial [5, 17, 60], not to mention that these tests often require a large sample size to achieve reasonable power. In the cases when an analyst resorts to choosing a model among several candidates, the action of selecting a model itself has non-negligible impact to the following inference and interpretation, and the impact is difficult to account for [40].

Due to these challenges of model-based statistical analysis and many more, model-agnostic statistical methods have grown increasingly popular over the past decades. These methods can roughly be categorized into two groups. One is nonparametric statistics or semiparametric statistics, of which the methods impose very few restrictions on the statistical model. Nonparametric statistics methods don't impose any parametric restrictions on the data-generating distribution except for some mild restrictions such as smoothness or monotonicity, whereas semiparametric statistics imposes parametric restriction on some aspects of the distribution and leaves the rest unspecified. For a formal discussion of these approaches, see [71, 57, 67]. Although these methods are widely popular in the statistics and machine learning community, some challenges remain unaddressed, such as long computa-

tion time, shortage of valid statistical inferential methods, difficulty with the interpretation, less accessibility of user-friendly software, and most importantly, the well-known curse of high-dimensionality (See, for example, Wasserman [71, §4.5]). In contrast, the other approach does not alter the point estimate, which is derived from a statistical model, but instead re-evaluates the results from the parametric analysis, through the lens of model agnoticism, and adjusts for potential model violation. Specifically, this approach acknowledges that the model can be misspecified, clearly defines the estimand under model misspecification, and use an inferential procedure that doesn't require correct model specifications. For example, when the probabilistic model is incorrectly specified, Huber studied the limit of maximum likelihood estimates [29], the interpretation of which was further studied by Akaike [4]. Huber also showed the asymptotic normality of the maximum likelihood estimates under certain conditions [29], which can lead to valid inference.

In this dissertation, we aim to use the methodology of model-agnostic statistical inference to re-evaluate several areas that have broad application but have been missed out in the wave of agnostic statistics research – meta-analysis, shrinkage estimation, and Bayesian parametric regression analysis. In Chapter 2, we revisit the common approaches in meta-analysis and develop model-agnostic inference procedures in meta-analysis for two-by-two tables; in Chapter 3, we propose a novel framework of model-agnostic shrinkage estimation, with special focuses on shrinkage estimators in linear models; in Chapter 4, we develop a Bayesian analog of the Huber-White robust variance matrix [29, 73] as the Bayes rule of a novel loss function.

Chapter 2

IMPROVED INFERENCE FOR FIXED-EFFECTS META-ANALYSIS OF 2×2 TABLES

2.1 Introduction

Meta-analysis methods combine information from multiple studies to make inference and draw conclusions [21], and compared with single studies often have greater statistical power to detect association between two variables. In the common situation where both variables are binary, the data can be represented as an easily-shared 2×2 contingency table, meaning that (unlike most other meta-analyses) we can typically reconstruct individual-level data within the contributing studies. The association between binary variables can be quantified in several ways, including risk differences, risk ratios, and—most commonly-used—odds ratios.

Perhaps the most conventional framework of meta-analysis is ***common-effect*** meta-analysis (sometimes referred to as ***fixed-effect*** meta-analysis, singular), where the odds ratio (or equivalently its logarithm) is homogeneous across studies. In common-effect meta-analysis, the target of inference is the common log odds ratio. The inference is based on consistent estimators of the log odds ratio including the Maximum Likelihood Estimator (MLE) [3], Woolf's inverse-variance estimator [74] and the logarithm of Cochran-Mantel-Haenszel statistic (log-CMH) [41]. But homogeneity is unlikely to hold in practice; differences in study population characteristics and research protocols can induce unequal effect sizes across studies, particularly as odds ratios are non-collapsible across studies [18]. Small-study effects [56] can also induce heterogeneity. When baseline risks differ across studies, risk ratios or risk differences may be homogenous where the corresponding odds ratios are not. It therefore seems prudent to always at least consider the impact of heterogeneity across studies, when meta-analyzing 2×2 tables.

Faced with heterogeneity in practice, it is common to adopt a ***random-effects*** frame-

work. This has two primary justifications [26]. First, one can assume that the meta-analyzed study effects are a random sample from a "super-population". But the implicit independence of studies is dubious, as in many cases later studies' designs and target populations will have been influenced by earlier results. Moreover, the appropriateness of distributional assumptions on the study effects remains debatable [30]. A second justification is to view the random-effects model as an approximate Bayesian statement of exchangeability of the study effects. While this justifies the use of the model, it does not justify drawing inference based on the mean of the random effects distribution. Indeed, it leaves the question of exactly what to estimate unanswered. A distinct concern is the poor performance of inference based on random-effects analysis of a small number of studies. In this common situation, inference can be sensitive to distributional assumptions, and confidence intervals (CIs) may have far from nominal coverage levels.

In light of these difficulties, in this chapter we propose meta-analysis of 2×2 tables under a ***fixed-effects*** (plural) framework, where the effect sizes of single studies are fixed but may be different, thus allowing heterogeneity. Though using the same estimators as in common-effect meta-analysis, the target of inference in fixed-effects meta-analysis is an average effect size across studies, where the exact form of the average depends on the estimator chosen. The fundamental ideas and justification of fixed-effects meta-analysis are discussed elsewhere [54, 50, 38]. With careful statement of what the method estimates under heterogeneity our approach provides inferences, relevant to scientific questions of interest, with good calibration of their frequency properties.

In Section 2.2 we provide the relevant notation, and define the major fixed-effects estimators used in meta-analysis of 2×2 tables. In Section 2.3 we explore the large-sample properties of these estimators, and show how they can be used to construct novel confidence intervals. These are explored through simulation in Section 2.4 and via an applied examples in Section 2.5. We conclude in Section 3.12 with a short discussion.

2.2 Notation

We consider a meta-analysis of k studies where the data in each study has a binary outcome Y and a single binary covariate X. Following convention, we call the population with $X = 1$

		Covariate X		
		1	0	Total
Outcome Y	1	a_j	b_j	n_{1j}
	0	c_j	d_j	n_{0j}
	Total	m_{1j}	m_{0j}	N_j

Table 2.1: Notation for a subtable j contributing to a meta-analysis

the *treatment* group and $X = 0$ the *control* group, though the methods are useful well beyond the setting of controlled trials. The counts of subjects for each combination of levels of outcome and covariate in study j can be laid out in a 2×2 contingency table, as in Table 2.1.

We denote $N = \sum_{j=1}^{K} N_j$ as the total sample size. We condition on the column totals m_{0j}, m_{1j} and assume that a_j and b_j are independent random variables, drawn from $Binomial(m_{1j}, p_{1j})$ and $Binomial(m_{0j}, p_{0j})$ respectively, where p_{1j} and p_{0j} are the risks in the treatment group or control group of study j. For convenience we denote $\delta_j = m_{1j}/N_j$ as the proportion of subjects in treatment group, $N_j/N = \gamma_j$ the ratio of sample size of study j to the total sample size, and also write $\bar{p}_{0j} = 1 - p_{0j}$, $\bar{p}_{1j} = 1 - p_{1j}$ and $\bar{\delta}_j = 1 - \delta_j$.

Within each study j we denote the odds ratio by $\theta_j = [p_{1j}\bar{p}_{0j}] / [p_{0j}\bar{p}_{1j}]$ and its logarithm by $\psi_j = \log(\theta_j)$. (Adjustments made when zero entries occur are deferred until Section 2.3.3.) Again for convenience we denote the log odds of the outcome in the control group of study j by $\eta_j = \log(p_{0j}/\bar{p}_{0j})$. Throughout, bold letters denote vectors; for example, $\boldsymbol{p}_0 = (p_{01}, p_{02}, ..., p_{0K})$.

2.2.1 Estimators: CMH

Using these definitions, we can further define the logarithm of the CMH estimator as

$$\hat{\psi}_{CMH} = \log\left(\sum_{j=1}^{K} \frac{\tau_j}{\sum_{j'=1}^{K} \tau_{j'}} \theta_j\right), \text{ where } \hat{\theta}_j = \frac{a_j d_j}{b_j c_j} \text{ and } \tau_j = \frac{b_j c_j}{N_j},$$

which is frequently re-written as

$$\hat{\psi}_{CMH} = \log\left(\frac{\sum_{j=1}^{K} \frac{a_j d_j}{N_j}}{\sum_{j=1}^{K} \frac{b_j c_j}{N_j}}\right).$$

For inference, Hauck produced the large-sample variance of the CMH statistic [23], from which Breslow developed an "empirical Hauck estimator" of the variance of log-CMH estimator [9], in which the use of the logarithmic scale prevents confidence intervals potentially taking negative values. Breslow's variance estimator is

$$\widehat{\mathrm{Var}}_{Hom}[\,\hat{\psi}_{CMH}\,] = \frac{1}{\left(\sum b_k c_k / N_k\right)^2} \sum_{j=1}^{K} \left(\frac{b_j c_j}{N_j}\right)^2 \left\{\frac{1}{a_j} + \frac{1}{b_j} + \frac{1}{c_j} + \frac{1}{d_j}\right\}, \tag{2.1}$$

where the subscript "Hom" emphasizes that the variance was derived assuming effect homogeneity. To construct approximate confidence intervals, we use the standard formulation

$$\hat{\psi}_{CMH} \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}_{Hom}[\,\hat{\psi}_{CMH}\,]},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2) \times 100\%$ quantile of a standard Normal distribution. The Cochran-Mantel-Haenszel statistics and its confidence interval are available in standard software packages, such as the `mantelhaen.test()` function in base R.

2.2.2 Estimators: Woolf

Woolf's estimator is defined as

$$\hat{\psi}_{Woolf} = \left(\sum_{j=1}^{K} \frac{\hat{w}_j}{\sum_{j'=1}^{K} \hat{w}_{j'}} \hat{\psi}_j\right), \text{ where } \hat{\psi}_j = \log\left(\frac{a_j d_j}{b_j c_j}\right) \text{ and } \hat{w}_j = \left(\frac{1}{a_j} + \frac{1}{b_j} + \frac{1}{c_j} + \frac{1}{d_j}\right)^{-1},$$

and is equivalent to the estimate from a fixed-effects, inverse-variance weighted meta-analysis, that uses the standard point estimate of each subtable's odds ratio and its corresponding variance estimate.

We note how Woolf's differs from the CMH estimator. The CMH estimator is the log of the weighted average of subtable-specific odds ratios, whereas the Woolf estimator is a

weighted average of subtable-specific log odds ratios estimates. The weights do differ but both, broadly, upweight tables that provide more information about their corresponding population-specific odds ratios.

The variance of Woolf's estimator can be approximated using standard results from fixed-effect meta-analysis [21], where the precision (i.e. the inverse-variance) of the overall estimate is the sum of the precisions from each study. Combining odds ratios across 2×2 tables this estimator is

$$\widehat{\text{Var}}_{Hom}[\,\hat{\psi}_{Woolf}\,] = \frac{1}{\sum_{j=1}^{K}\left(\frac{1}{a_j}+\frac{1}{b_j}+\frac{1}{c_j}+\frac{1}{d_j}\right)^{-1}}. \tag{2.2}$$

In Section 2.3.2 we will show how, even in large samples, the above estimator of variance only validly estimate the asymptotic variance of Woolf's estimator when homogeneity holds across all studies. Under heterogeneity the standard confidence interval based on (2.2) is not well calibrated.

2.2.3 Estimators: MLE

The MLE of an assumed-common log odds ratio is not available in closed form, but can instead be found by maximizing the likelihood of the model (described above) with respect to the K logit odds in the control each of the subtable — the α_j — and the assumed-common log odds ratio ψ that relates each α_j to the risk for its corresponding treatment group. To maximize the likelihood we solve the score equations

$$\begin{aligned} a_j + b_j - m_{1j}\text{expit}(\alpha_j+\psi) - m_{0j}\text{expit}(\alpha_j) &= 0, \quad \text{for } j = 1, ..., K, \\ \sum_{j=1}^{K} a_j - m_{1j}\text{expit}(\alpha_k+\psi) &= 0. \end{aligned}$$

where expit() denotes the inverse-logit function, i.e. $\text{expit}(x) = e^x/(1+e^x)$. Like the log-CMH and Woolf estimates, the MLE is unconstrained by issues of non-negativity. Equivalently, the MLE can be obtained by fitting a logistic regression model to the counts data with study-specific fixed-effect intercepts α_j and a fixed-effect slope ψ for the indicator of

treatment. Several authors have criticized the use of the MLE since conventional inference for it assumes a constant odds ratio (i.e. homogeneity) which is untenable in most cases [41].

For inference under homogeneity, standard likelihood inference applied to generalized linear models can be applied to obtain asymptotically-justified confidence intervals. We defer formula to Appendix A, but the basic approach uses the asymptotic Normality of the MLE of the vector of parameters $(\alpha_1, \alpha_2, ...\alpha_K, \psi)$, where the variance of the MLEs can be approximated by the inverse of the observed or expected Fisher information. From this multivariate normality, the approximate variance of the MLE for ψ is used to construct confidence intervals as

$$\hat{\psi}_{MLE} \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}[\,\hat{\psi}_{MLE}\,]},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)\times 100\%$ quantile of a standard normal distribution. (Other inference methods for MLE include conditional or exact inference.) These confidence intervals are widely available, for example via the `glm()` and `confint.default()` functions in base R, but it should be noted their derivation does rely on homogeneity.

2.3 Results

As described in Sections 2.2.1, 2.2.2 and 2.2.3, inference for the CMH estimator, Woolf's estimator and the MLE are well-studied under homogeneity. However, as noted in Section 2.1, while assumptions of exact homogeneity are seldom realistic, corresponding results giving inference under heterogeneity are not available. In this section we set out the behaviour of the three estimators under heterogeneity, where each study estimates its own population's odds ratio and these parameters are not constrained in any way.

To obtain large-sample behavior of the estimators in this setting, we consider the limiting regime where the proportion of subjects in treatment group, δ_j's, and the ratio of sample size of single studies to the total sample size, γ_j's, remain fixed as the total sample size N gets large. Less formally, we base inference on the situation where each study is large, not where there are a large number of studies each of fixed small size.

2.3.1 Estimands

Under the large-sample regime described above, we can state the large-sample limits of the three estimators.

Formally, as $N \to \infty$ while the δ_j and γ_j remain fixed, the log-CMH statistic and Woolf's estimator converge to the following values:

$$\psi_{CMH} = \log\left(\frac{\sum_{j=1}^{K} \pi_j \theta_j}{\sum_{j=1}^{K} \pi_j}\right), \qquad \text{where } \pi_j = p_{0j}\bar{p}_{1j}r_j, \quad r_j = \gamma_j\delta_j\bar{\delta}_j; \tag{2.3}$$

$$\psi_{Woolf} = \frac{\sum_{j=1}^{K} w_j\psi_j}{\sum_{j=1}^{K} w_j}, \qquad \text{where } w_j = \gamma_j \Big/ \left(\frac{1}{\delta_j p_{1j}\bar{p}_{1j}} + \frac{1}{\bar{\delta}_j p_{0j}\bar{p}_{0j}}\right). \tag{2.4}$$

The MLE $\hat{\psi}_{MLE}$ converges in probability to the maximizer ψ_{MLE} of the expected log likelihood,

$$L(\alpha, \psi) = \sum_j \gamma_j\{\alpha_j\left(\delta_j p_{1j} + \bar{\delta}_j p_{0j}\right) + \psi\delta_j p_{1j} - \delta_j \log[1 + \exp(\alpha_j + \psi)] - \bar{\delta}_j \log[1 + \exp(\alpha_j)]\} \tag{2.5}$$

which is not available in closed form. In (2.5), the scalar ψ is the assumed-common log odds ratio and the K-vector $\boldsymbol{\alpha}$ consists of the studies' baseline odds in an assumed-homogeneous model. (Here we abuse the notation a little since we use ψ to denote both a common log odds ratio under homogeneity and an assumed-common log odds ratio under heterogeneity.)

One may view the three estimands, broadly, as different versions of average effect size – the estimand ψ_{CMH} is the logarithm of a weighted average of odds ratios with weights proportional to π_j's, and the estimand ψ_{Woolf} is a weighted average of log odds ratios with weights proportional to w_j's. The estimand of MLE ψ_{MLE} is not apparent but its value is between the maximum and the minimum of the study-specific log odds ratios.

Unless homogeneity actually holds, these three large-sample values are all different, i.e. the estimators estimate different things. It is therefore of interest to know which might be easiest to estimate, and if inference on any of them is particularly well- or poorly-calibrated under heterogeneity. From the numerical studies we will see the MLE is more robust to the presence of heterogeneity compared than log-CMH and Woolf's estimator, supporting its

use to infer an overall association. We will detain the discussion until Section 3.12.

2.3.2 Asymptotic distribution

Having established the limiting values of the estimators, we now address the limiting distribution of the estimators – and in particular whether the variance results established under homogeneity are affected by heterogeneity.

As we establish in Appendix A, all three estimators are asymptotically normal under heterogeneity, as they are known to be under homogeneity. Formally, for each estimator $\sqrt{N}(\hat{\theta}-\theta)$ tends in distribution to $N(0,V)$ for some asymptotic variance V that we describe below. Proofs are deferred to Appendix A.

The asymptotic variance of the log-CMH estimator is

$$V_{CMH} = \frac{1}{\left[\sum_{k=1}^{K}\pi_k\right]^2}\sum_{j=1}^{K}\pi_j^2\theta_j^2\rho_j/w_j,$$
$$\text{where } \rho_j = \frac{[\bar{p}_{0j}+p_{0j}e^{\psi_{CMH}}]^2\frac{\bar{p}_{1j}}{\bar{p}_{0j}}\bar{\delta}_j + [p_{1j}+\bar{p}_{1j}e^{\psi_{CMH}}]^2\frac{p_{0j}}{p_{1j}}\delta_j}{\frac{\bar{p}_{0j}}{\bar{p}_{1j}}\bar{\delta}_j + \frac{p_{1j}}{p_{0j}}\delta_j}. \tag{2.6}$$

Unless homogeneity holds—in which case all the ρ_j terms are 1—this value is not equivalent to the version derived by Hauck [23] under homogeneity:

$$V_{CMH,\,Hom} = \frac{\theta^2}{\left[\sum_{j=1}^{K}\pi_j\right]^2}\sum_{j=1}^{K}\pi_j^2/w_j.$$

Consequently, inference using $V_{CMH,Hom}$ (or inference based on it, such as Breslow's weighted variance method [9]) should not be expected to be well-calibrated. Depending on whether the ρ_j terms are generally greater or smaller than 1, methods assuming homogeneity can be expected to give under- or over-coverage respectively, even in large sample sizes.

Similar behavior occurs for the Woolf estimator. Here the asymptotic variance is

$$V_{Woolf} = \frac{1}{\sum_{j=1}^{k}w_j}\left(1 - 2\langle\boldsymbol{\psi},\boldsymbol{\Delta}_1\rangle_{\boldsymbol{w}} + \langle\boldsymbol{\psi},\boldsymbol{\psi}\circ\boldsymbol{\Delta}_2\rangle_{\boldsymbol{w}} - \xi\cdot\langle\boldsymbol{\psi},\boldsymbol{\Delta}_2\rangle_{\boldsymbol{w}}\right), \tag{2.7}$$

where

$$\Delta_{1j} = \frac{(p_{1j}^2 - \bar{p}_{1j}^2)t_j^2 - (p_{0j}^2 - \bar{p}_{0j}^2)s_j^2}{(s_j + t_j)^2};$$
$$\Delta_{2j} = \frac{(p_{1j}^2 - \bar{p}_{1j}^2)^2 t_j^3 + (p_{0j}^2 - \bar{p}_{0j}^2)^2 s_j^3}{(s_j + t_j)^3};$$
$$s_j = \gamma_j \delta_j p_{1j} \bar{p}_{1j};$$
$$t_j = \gamma_j \bar{\delta}_j p_{0j} \bar{p}_{0j}.$$

for $k = 1, ..., K$, with π_j and w_j as defined in Equations (2.3) and (2.4). The notation also uses element-wise multiplication: for two K-vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, the symbol $\circ$ denotes element-wise multiplication of vectors, i.e. $\boldsymbol{a} \circ \boldsymbol{b} = (a_1 b_1, a_2 b_2, ..., a_K b_K)$. We denote by $\langle \boldsymbol{a}, \boldsymbol{b} \rangle_{\boldsymbol{w}}$ the weighted covariance of $\boldsymbol{a}$ and $\boldsymbol{b}$, i.e.,

$$\begin{aligned}\langle \boldsymbol{a}, \boldsymbol{b} \rangle_{\boldsymbol{w}} &= \frac{\sum_{j=1}^K w_j a_j b_j}{\sum_{j=1}^K w_j} - \left(\frac{\sum_{j=1}^K w_j a_j}{\sum_{j=1}^K w_j} \right) \left(\frac{\sum_{j=1}^K w_j b_j}{\sum_{j=1}^K w_j} \right) \\ &= \sum_{i=1}^K \frac{w_i}{\sum_{l=1}^K w_l} \left(a_i - \frac{\sum_{j=1}^K w_j a_j}{\sum_{j=1}^K w_j} \right) \left(b_i - \frac{\sum_{j=1}^K w_j b_j}{\sum_{j=1}^K w_j} \right)\end{aligned}$$

Here again the result departs from that obtained under homogeneity. In this case it is the weighted covariance terms $\langle \boldsymbol{\psi}, \boldsymbol{\Delta}_1 \rangle_{\boldsymbol{w}}$, $\langle \boldsymbol{\psi}, \boldsymbol{\psi} \circ \boldsymbol{\Delta}_2 \rangle_{\boldsymbol{w}}$ and $\langle \boldsymbol{\psi}, \boldsymbol{\Delta}_2 \rangle_{\boldsymbol{w}}$ that determine the impact of heterogeneity on the corresponding inference. Under homogeneity, log odds ratio elements of vector $\boldsymbol{\psi}$ are all identical and so have zero covariance with all of the other terms.

Exact statement of the MLE's asymptotic variance is deferred to Appendix A, but the same pattern holds. In this case the Fisher information matrix under homogeneity is multiplied by another term, which disappears under homogeneity but is otherwise present.

Based on these results, we propose new confidence intervals that use plug-in estimators of the variances under heterogeneity. Specifically, we replace p_{0j} and p_{1j} with their estimates $\hat{p}_{0j} = b_j/m_{0j}$ and $\hat{p}_{1j} = a_j/m_{1j}$, and replace α_j and ψ with their MLEs $\hat{\alpha}_j$ and $\hat{\psi}$, which are simply the estimated study-specific intercepts and the slope of an indicator of the treatment in a logistic regression model. The MLEs can be implemented in standard statistical packages, for example by using *glm()* function in base R.

2.3.3 Zero-entry Correction

As is commonly-encountered in this area, zero entries in 2×2 tables lead to infinite or uncomputable estimates. A common and usually-reasonable 'fix' is to add 1/2 (or other small number) uniformly to all entries in tables with at least one zero, but this procedure is known to have shortcomings [63]. We instead add the reciprocal of the sample size of the opposite arm to the cells in tables with zeroes prior to computing the estimators and their variances, as proposed by Sweeting *et al*[63] along with several alternative corrections. With this zero-entry correction, double-zero subtables produce odds ratios close to one, weakly favoring a conclusion of no association between the covariate X and outcome Y. Full details are given in Appendix C.

2.4 Numerical studies

In this section, we perform simulation studies to compare the coverage of our proposed confidence intervals, that allow for heterogeneity, versus standard approaches that are derived assuming homogeneity.

We simulate data from k=4 or 8 studies, each of which has a sample size of N_j=200. The control group and treatment group in every study both contain 100 subjects. We set the risks of the control group in all studies to be identically p_{0j}=0.05 or 0.10, representing a rare-outcome and a moderate event scenario. We vary the log odds ratios ψ_j in each study; half the studies have one value of ψ_j and the other half a potentially different value. We report the estimand and bias of each estimator for each scenario and the coverage rates of the confidence intervals for that estimand.

Table 2.2 shows the estimands, biases and coverage probabilities of different confidence intervals for the three estimators. In this large-sample setting, under homogeneity all confidence intervals have close to nominal coverage. In the presence of heterogeneity, the standard confidence interval of log-CMH is conservative, while our proposed method maintains the nominal coverage. Our proposed method for Woolf's estimator has coverage probability closer to the nominal level than that of the standard method derived under homogeneity, except when the baseline risks are moderately low and the effects are strongly negative,

where both approaches suffer from under-coverage. The behavior of the two confidence intervals around the MLE is similar.

To further assess the small-sample performance of our proposed confidence intervals, using complete enumeration we compute the exact coverage probability of the confidence intervals for k=2 studies, where the treatment group and control group of both studies both have 20 subjects. The risks of the two control groups are either 0.2 or 0.4. Figure 2.1 shows the results. For such a small sample size and low event probabilities, intervals derived under homogeneity and heterogeneity are almost always conservative. However, with small number of studies and small sample sizes, the coverage probabilities for intervals derived under heterogeneity are closer to nominal than standard methods, that are derived under homogeneity, regardless of the control group risks.

Intervals for the log-CMH derived under homogeneity are almost always conservative, with coverage exceeding that of our proposed method by 2 to 3 percentage points in some cases. For Woolf's estimate there is no such systematic over-coverage of the standard intervals, but our proposed method does in general produce closer-to-nominal levels. The coverage probabilities of the two confidence intervals for the MLE are close in all scenarios considered.

2.5 Real Data Example: Sclerotherapy

To demonstrate the methods we apply the estimators and their confidence intervals to a meta-analysis of 19 randomized trials assessing the effectiveness of endoscopic sclerotherapy in prevention of first bleeding among patients with cirrhosis and esophagogastric varices [46]. The outcome of interest is first bleeding. The full data are attached in the Supplementary Material, as Supplementary Table B.1. A summary of data and results is shown in Figure 2.2.

Cochran's Q statistic [21] from the Woolf-based analysis is 75.62, which compared to χ^2_{18} gives $p = 4.94 \times 10^{-9}$ for the test of homogeneity. The p-values of the Mantel-Haenszel test [41] and Exact Conditional test [7] against a "strong null hypothesis" (log odds ratios of all studies are identically zero) are 6.3×10^{-7} and 5.5×10^{-7} respectively, indicating that sclerotherapy is significantly associated with the bleeding outcome.

Table 2.2: Estimands and coverage probability of confidence intervals for logarithm of Cochran-Mantel-Haenszel statistics (log-CMH, $\hat{\psi}_{CMH}$), Woolf's estimator ($\hat{\psi}_{Woolf}$) and MLE of a common odds ratio ($\hat{\psi}_{MLE}$), with varying log-odds ratios of the studies. The confidence intervals are the established methods that are derived under homogeneity, and our proposed approaches deriverd under heterogeneity. The estimands are computed using the true values of the parameters. Coverage probabilities are computed over 10,000 replications.

K	p_{01}-p_{0K}	ψ_1-$\psi_{K/2}$	$\psi_{(K/2+1)}$-ψ_K	Estimands			Bias			Coverage of nominal 95% CIs					
										log-CMH		Woolf		MLE	
				ψ_{CMH}	ψ_{Woolf}	ψ_{MLE}	ψ_{CMH}	ψ_{Woolf}	ψ_{MLE}	Hom	Het	Hom	Het	Hom	Het
4	0.05	0.0	0.0	0.00	0.00	0.00	0.00	-0.00	0.00	0.972	0.953	0.976	0.964	0.953	0.953
4	0.05	0.0	0.5	0.28	0.27	0.28	0.01	-0.01	0.01	0.970	0.950	0.974	0.962	0.950	0.949
4	0.05	0.5	0.0	0.28	0.27	0.28	0.01	-0.01	0.01	0.971	0.951	0.974	0.963	0.952	0.951
4	0.05	0.5	0.5	0.50	0.50	0.50	0.01	-0.03	0.01	0.967	0.955	0.970	0.960	0.955	0.954
4	0.10	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.963	0.952	0.965	0.955	0.953	0.952
4	0.10	0.0	0.5	0.27	0.27	0.27	0.00	0.00	0.00	0.963	0.951	0.965	0.955	0.951	0.951
4	0.10	0.5	0.0	0.27	0.27	0.27	0.00	0.00	0.00	0.962	0.952	0.964	0.954	0.952	0.952
4	0.10	0.5	0.5	0.50	0.50	0.50	0.00	-0.01	0.01	0.958	0.951	0.961	0.954	0.952	0.952
8	0.05	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.973	0.950	0.976	0.963	0.950	0.950
8	0.05	0.0	0.5	0.28	0.27	0.28	0.00	-0.02	0.00	0.971	0.951	0.975	0.962	0.951	0.951
8	0.05	0.5	0.0	0.28	0.27	0.28	0.00	-0.02	0.00	0.970	0.951	0.973	0.958	0.951	0.951
8	0.05	0.5	0.5	0.50	0.50	0.50	0.00	-0.04	0.01	0.964	0.951	0.968	0.955	0.951	0.951
8	0.10	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.960	0.949	0.962	0.952	0.949	0.949
8	0.10	0.0	0.5	0.27	0.27	0.27	0.00	-0.01	0.00	0.961	0.948	0.963	0.952	0.948	0.948
8	0.10	0.5	0.0	0.27	0.27	0.27	0.00	0.01	0.00	0.961	0.948	0.962	0.952	0.949	0.948
8	0.10	0.5	0.5	0.50	0.50	0.50	0.00	-0.01	0.00	0.957	0.947	0.958	0.950	0.948	0.948

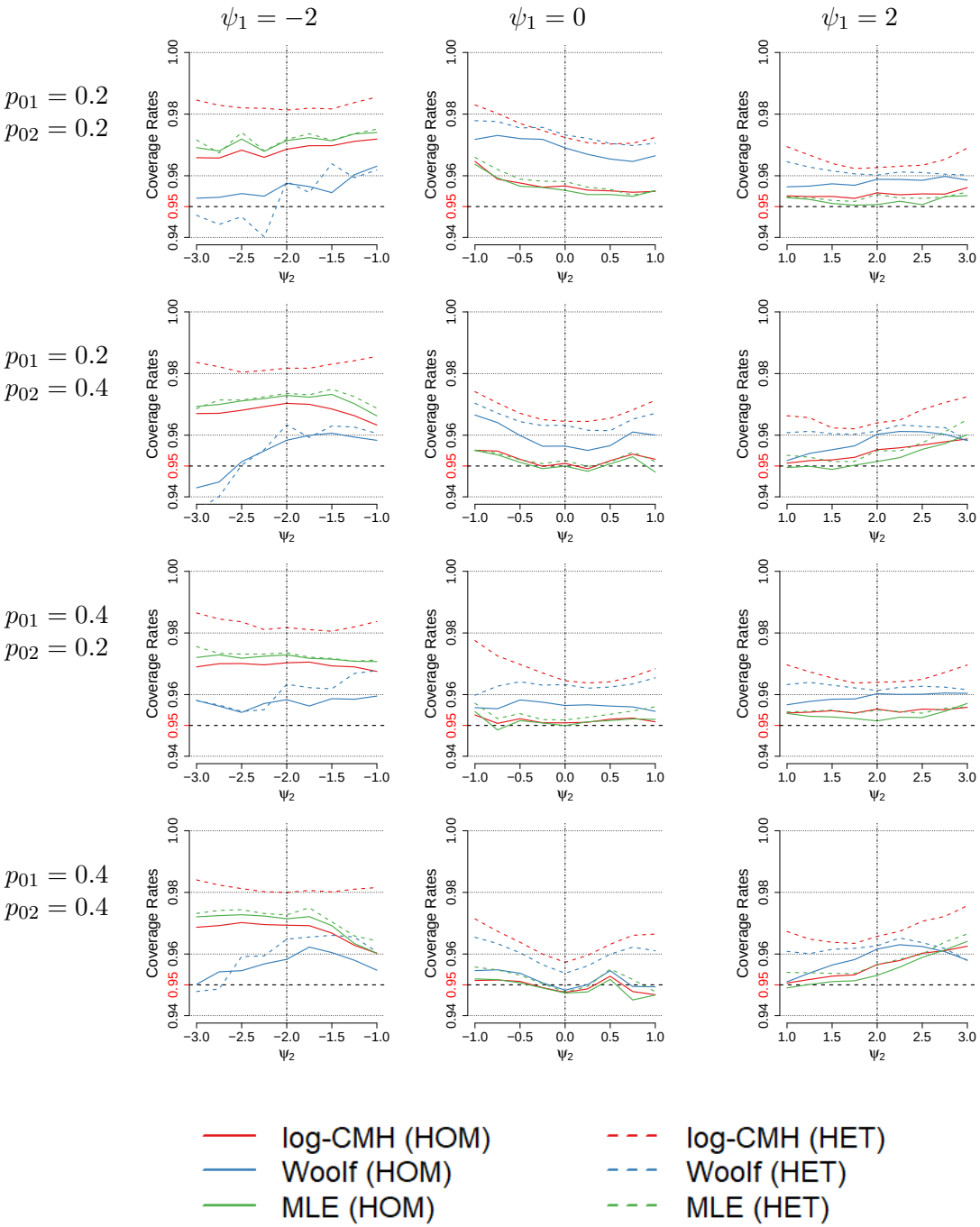

Figure 2.1: Coverage of the confidence intervals, derived under homogeneity or potential heterogeneity, with varying baseline event probabilities and effect sizes in each study. The vertical dashed line shows the point where $\psi_1 = \psi_2$, i.e. where the effect sizes are homogeneous.

Meta-analysis using log-CMH statistics, Woolf's estimator or MLE, with confidence intervals that rely or do not rely on homogeneity assumption, lead to the same conclusion that patients who receive sclerotherapy are less likely to develop first bleeding during follow-up. The confidence interval for log-CMH derived assuming homogeneity is notably wider than that without and the confidence interval for Woolf's estimator assuming homogeneity is a little wider than the one without, while the two confidence intervals for MLE are indistinguishable. These observations coincide with the results illustrated in Section 2.4's simulations.

As a further comparison, we implement a Bayesian version of fixed-effects analysis: putting independent Jeffrey's prior distribution – Beta distribution with parameters $\frac{1}{2}$ and $\frac{1}{2}$ – on the $(p_{1i},\ p_{0i})$'s, and directly sampling the posterior for the parameter of interest (ψ_{CMH}, ψ_{Woolf} and ψ_{MLE}). We obtained 95% credible intervals for three estimators from the 0.025 and 0.975 quantiles of each posterior samples. With 20,000 draws, the 95% credible intervals are (-0.744, -0.344) for log-CMH, (-0.704, -0.268) for Woolf's estimator and (-0.803, -0.356) for MLE. The prior distribution matters less to the credible interval for MLE than for log-CMH and Woolf's estimator. As with the non-Bayesian results this suggests that estimation of MLE's target parameter is more robust to heterogeneity than estimation of the other two average effect sizes, ψ_{CMH} and ψ_{Woolf}.

2.6 Discussion

We have investigated the large-sample behavior of three widely-used estimators for measuring the association between two random variables from data of several potentially heterogeneous studies. We derived their estimands under heterogeneity, and also developed confidence intervals that do not rely on assumptions of homogeneity.

Conventional inference for the inverse-variance method has been criticized for overlooking the uncertainty in the variance estimates of study effects [54, 30]. In a previous study of the large-sample behavior of inverse-variance meta-analysis, Domìnguez and Rice found that for normal outcomes the use of estimated standard errors, rather than standard errors that are known, leads to anti-conservative inference when heterogeneity is present even in large samples [13]. Our work on Woolf's estimator shows that Binomial outcomes are different;

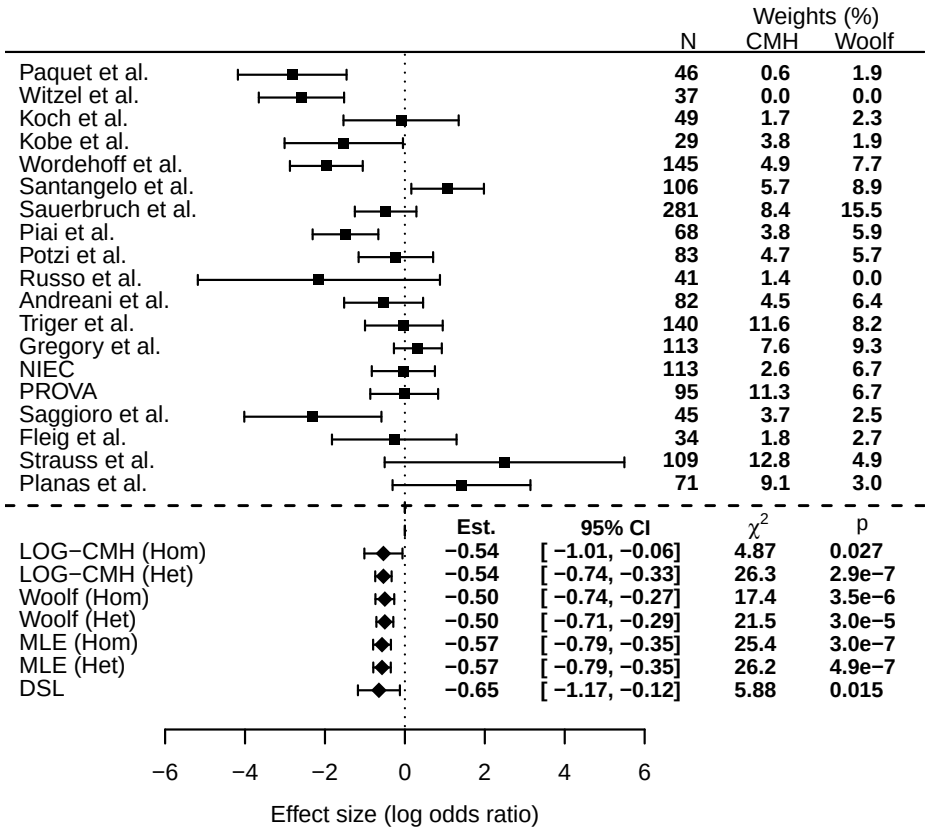

Figure 2.2: Forest plots of meta-analysis of sclerotherapy studies, [46] with confidence intervals of log-CMH statistics, Woolf's estimator, MLE, derived with or without assuming homogeneity. We include the confidence interval using DerSimonian-Laird method (DSL) as comparison. We also show the chi-square statistics and the p-values for testing the value of chi-square statistics against a chi-square distribution with 1 degree of freedom.

when homogeneity holds, the large-sample variance (and corresponding confidence interval width) derived assuming known variances of study effect estimates is indeed consistent. Therefore the conventional confidence interval has correct large-sample coverage over the common log odds ratio under homogeneity. However, in the presence of heterogeneity the large-sample variance of Woolf's estimator derived assuming homogeneity can, when homogeneity does not hold, be larger or smaller than that using no such assumption. While the formulae of Section 2.3 give a full statement of this behavior, it can be explained intuitively, at least in part. Unlike the Normal, the Binomial distribution has a non-zero correlation between the estimated mean and variance. As a result, the fixed-effects confidence interval for Woolf's estimator is not always wider or narrower than the confidence interval assuming known variances.

As we have also shown, the three estimators have different estimands in the presence of heterogeneity. Which to use therefore becomes an issue for practitioners. Ideally, this choice would be made based on context: that there would be some scientific reason to prefer e.g. one form of weighting of study-specific estimates. However having such nuanced information is unlikely in practice. Our results show that, rather than choosing a procedure based on the data (a process known to invalidate many inferential methods) use of the MLE is a pragmatic default, as its coverage seems little affected by whether or not homogeneity is used to derive the CIs — unlike the CMH and Woolf methods. While the MLE can be conservative when small cell counts are expected, it otherwise gives close to nominal coverage, always outperforming at least one of the competing methods.

The question of the meaning of the average effect across studies is an important part of almost all meta-analyses, but we do not advocate for it being the only line of investigation. In particular, the heterogeneity of effects is also of interest, and may be addressed numerically by measures such as Cochran's Q, I^2 and H statistics [25], or graphically via forest plots, funnel plots, scatter plots and line graphs [36]. We do hope, however, that our careful definition of what is being estimated by the across-study average may clarify what results are heterogeneous *around*. Being clear on this should help focus subsequent analyses of, say, non-constant variability in effect sizes as that effect size differs between studies.

A formula of the CMH estimator's variance under heterogeneity was given by Hauck in

in [23], which was derived by writing the CMH estimator as weighted strata-specific odds ratio, although their derivation failed to account for the randomness of the weights. Despite the theoretical shortcoming of the variance formula, in large samples Hauck's variance estimate is indistinguishable to the true variance of the CMH estimator under heterogeneity in simulation studies [45]. Using a different approach, Noma et. al. derived the asymptotic variance of CMH estimator under heterogeneity, both in the scenario of a fixed number of large strata (as is assumed in this article) and in the scenario of many strata with fixed size.

Our work has several limitation. First, though in our methods we use a type of zero-entry correction motivated by rare-outcome data [63], we see it more as a way to stabilize computation. Interested readers may find a more thorough discussion on issues and alternative methods of meta-analyses with rare-outcome binary data elsewhere [62, 37, 39]. Second, our use of asymptotic arguments addresses the common situation of meta-analyses with clinical studies, where each study provides reasonably accurate inference on its own underlying parameter. With a large number of studies of small size, the proposed confidence intervals may not have nominal coverage. This issue, known as "Neyman-Scott Problem", arises because the number of nuisance parameters increases at the same rate as the sample size [44]. In such situations one could instead use random-effects models for better inference [59, 16], though considerable care is needed when choosing the mixing distribution, and a full resolution is beyond the scope of this paper. Finally, though our work focuses on (log) odds ratio, the idea of fixed-effects meta-analysis – fixed and potentially different effect sizes for the studies – can be used for meta-analyses with other pooled effect measures including risk difference, risk ratio and arcsine difference [56].

Code for our methods, including a vignette illustrating their use in practice, are available from Github (`https://github.com/KenLi93/FEMetaBin`).

Chapter 3

SHRINKAGE ESTIMATION IN LINEAR MODELS USING PENALIZED PRECISION-WEIGHTED LEAST SQUARE ESTIMATORS

3.1 Introduction

When estimating the mean vector for a multivariate Normal distribution in three or more dimensions, James and Stein [31] discovered that the maximum likelihood estimator (MLE) is inadmissable, being dominated by a biased estimator, later called the *James-Stein estimator*, that can be expressed as the MLE multiplied by a shrinkage factor. This seminal work showed that pursuit of unbiased estimators alone was misguided, as accepting a little bias in exchange for better variance could give better performance overall. Since James and Stein [31], shrinkage estimators – that make the tradeoff of bias for variance in some way – and their statistical properties have become a major research topic. Baranchik [6] showed that the James-Stein estimator was itself inadmissible and proposed an improvement by lower-truncating the James-Stein estimator at zero. Cohen [11] studied the admissibility of shrinkage estimators and gave the sufficient and necessary conditions for a linear estimator to be admissible. They further proposed a family of minimax shrinkage estimators. Gruber [19] provides a comprehensive book-length overview of James-Stein type estimators and their extensions.

The James-Stein estimator can be extended to provide shrinkage estimates for linear regression, as discussed in the Chapter IV of [19]. In the same chapter, the author also discusses Bayesian and frequentist justifications for the James-Stein shrinkage estimator. In regression problems, however, penalized least-square estimators are more popular. These minimize penalized least-square objective functions, and include the Ridge estimator [27] and LASSO [65]. For discussion of many extensions see Hastie et al. [22].

The aforementioned shrinkage estimators are often motivated and derived under certain distributional assumptions. For example, the James-Stein estimator was proposed for

the multivariate Normal mean vector, and can also be motivated from an empirical Bayes perspective with a multivariate Normal model [14, 15]. The James-Stein estimator is also an approximate minimizer of MSE among all the scaled ordinary least square estimators if the data are multivariate Normal [19]. Both the Ridge and LASSO can be seen as posterior modes, under classical linear models with specific priors [19, 47]. When the true data-generating distribution cannot be assumed to be the classical Normal form, LASSO and the Ridge are simply viewed as the minimizer of a penalized mean squared error.

In this article, we introduced the penalized precision-weighted shrinkage estimators, that are the minimizers of penalized precision-weighted squared distance from a preliminary estimator. Heuristically, the proposed estimators can be viewed as the shrunken version of the preliminary estimator, where (1) the entries in the preliminary estimator with higher uncertainty are shrunken more aggressively, and (2) the shrinkage is induced and regularized by an added penalty. In Section 3.2, we present the penalized precision-based least square estimators and show that the James-Stein estimator, Ridge and LASSO are all examples of precision-based shrinkage estimators, adding penalties of different forms. In Section 3.3, we proposed the *Rotated James-Stein* estimator (rJS), *Rotated Ridge* (rRidge) and *Rotated LASSO* (rLASSO) as three precision-based shrinkage estimators, that using a heteroskedasticity-consistent precision weighting matrix in place of the standard weight derived under homoskedasticity. We study the theoretical properties of rJS, rRidge and rLASSO in Section 3.5-3.7, and study their empirical performances in simulation studies in comparison with the standard shrinkage estimators in Section 3.9. In Section 3.6 and 3.10, we investigate the sign consistency of rLASSO compared with LASSO. We demonstrate the application of the proposed estimators using the prostate cancer dataset in Section 3.11.

3.2 Penalized Precision-Weighted Least Square Estimators

Suppose the data are independent and identically distributed random variables Z_1, Z_2, ..., Z_n, with probability measure denoted P_0. We are interested in learning some summary of the data generating distribution in the form of a population parameter $\theta = \theta(P_0) \in \mathbb{R}^d$. Assume there exists an initial estimator for θ, denoted by $\hat{\theta}_I$. We assume the variance of $\hat{\theta}_I$, denoted by $\hat{\Sigma}_n$, is either known or can be estimated from the data. We propose the

penalized precision-weighted least square estimators:

$$\hat{\theta}_{\mathcal{P}} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} (\theta - \hat{\theta}_I)^T \hat{\Sigma}_n^{-1} (\theta - \hat{\theta}_I) + \mathcal{P}(\theta), \tag{3.1}$$

for some penalty $\mathcal{P}$. As we discuss later in Section 3.3, the penalized precision-weighted least square estimators can be viewed as methods that search for the value of θ that has the smallest Mahalanobis distance from $\hat{\theta}_I$ within a constrained parameter space defined by $\mathcal{P}(\theta)$.

The definition in (3.1) involves three components:

1. The initial estimate $\hat{\theta}_I$ that determines the origin of shrinkage;

2. The precision matrix $\hat{\Sigma}_n^{-1}$;

3. The penalizing function $\mathcal{P}(\theta)$.

The latter two components determine the direction and strength of the shrinkage. While the form of $\mathcal{P}$ is for now left as quite general, we note that – informally – the use of the precision matrix in (3.1) means we generally shrink components in the initial estimates in $\hat{\theta}_I$ more aggressively when they have higher uncertainty. In this way we intend to accept bias where the payoff in terms of variance reduction can be greatest.

In the linear regression problem, the data are the random sample $Z_i = (Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, where $Y_i \in \mathbb{R}$ is a scalar outcome and $\boldsymbol{X}_i \in \mathbb{R}^p$ is a vector of p covariates. The parameter of interest is often the "linear trend"

$$\beta_0 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} E\left[(Y_1 - \boldsymbol{X}_1^T \beta)^2\right] = \left[E(\boldsymbol{X}_1 \boldsymbol{X}_1^T)\right]^{-1} E(\boldsymbol{X}_1 Y_1).$$

Without loss of generality, we omit the intercept term. The default estimator for β_0 is the ordinary least squares (OLS) estimator

$$\hat{\beta}_{LS} = \left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i Y_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\mathbf{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)^T$.

The OLS estimator $\hat{\beta}_{LS}$ is popular for several reasons. It is the maximum likelihood estimator of β_0 if $Y_i \overset{i.i.d}{\sim} N(\boldsymbol{X}_i^T\beta_0, \sigma^2)$ and enjoys the corresponding optimalities [69]. With the common assumptions in linear models that

$$Y_i = \boldsymbol{X}_i^T\beta_0 + \epsilon_i, \quad E(\epsilon_i|\boldsymbol{X}_i) = 0 \qquad \text{(Linear Mean Model)} \tag{3.2}$$

$$\text{Var}(\epsilon_i|\boldsymbol{X}_i) = \sigma^2 \qquad \text{(homoskedasticity)} \tag{3.3}$$

for $i = 1, \ldots, n$, the OLS estimator is the best unbiased linear estimator by the well-known Gaussian-Markov theorem [33, §7]. Therefore, in terms of shrinkage estimation, we will consider $\hat{\beta}_{LS}$ to be the initial estimator $\hat{\theta}_I$ as in the display 3.1.

Assuming a linear mean model and homoskedasticity, the variance of $\hat{\beta}_{LS}$ is consistently estimated by $\hat{\Sigma}_0 = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$, where $\hat{\sigma}^2 = (n-p)^{-1}\sum_{i=1}^n(Y_i - \boldsymbol{X}_i^T\hat{\beta}_{LS})^2$ is an unbiased estimator of σ^2. Therefore, the penalized precision-weighted least-square estimators in linear regression, using the OLS estimator $\hat{\beta}_{LS}$ and model-based variance $\hat{\Sigma}_0$, has the common form

$$\underset{\beta\in\mathbb{R}^p}{\text{argmin}}(\beta - \hat{\beta}_{LS})^T\mathbf{X}^T\mathbf{X}(\beta - \hat{\beta}_{LS})/\hat{\sigma}^2 + \mathcal{P}(\beta) \tag{3.4}$$

for penalty $\mathcal{P}(\cdot)$.

To obtain JS, Ridge and LASSO estimates under this framework, we use penalties

$$\begin{array}{cccc} \text{Estimator} & & \text{Penalty, } \mathcal{P}(\boldsymbol{\beta}) & \\ \hline \text{JS} & \mathcal{P}_{\text{JS}}(\boldsymbol{\beta}) & = & \frac{\lambda_{JS}}{\hat{\sigma}^2}||\mathbf{X}\boldsymbol{\beta}||_2^2, \\ \text{Ridge} & P_{\text{Ridge}}(\boldsymbol{\beta}, \lambda_r) & = & \frac{\lambda_r}{\hat{\sigma}^2}\|\boldsymbol{\beta}\|_2^2, \\ \text{LASSO} & \mathcal{P}_{\text{LASSO}}(\boldsymbol{\beta}, \lambda_l) & = & \frac{\lambda_l}{\hat{\sigma}^2}\|\boldsymbol{\beta}\|_1. \end{array} \tag{3.5}$$

where the classic James-Stein estimate that dominates the sample mean is given by setting

$$\lambda_{\text{JS}} = \frac{\hat{\sigma}^2}{\frac{n-p+2}{(n-p)(p-2)}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \hat{\sigma}^2}.$$

For proofs see Appendix D.

3.3 Rotated James-Stein Estimator, Ridge Estimator, and LASSO

Here and throughout, we consider the case where the design matrix $\mathbf{X}$ is random. The presentation above uses the model-based variance $\hat{\Sigma}_0$ for the OLS estimator $\hat{\beta}_{LS}$, under the assumptions of the linear mean model (3.2) and homoskedasticity (3.3). However, the model-based variance fails to quantify the variability of $\hat{\beta}_{LS}$ under heteroskedasticity (i.e. non-constant variance) and/or non-linearity. This affects the motivation of all three shrinkage estimators as penalized precision-weighted least square estimators; the inverse-variance weighting term in the least squares, and also in the penalty for JS, does not reflect the actual variance under the true data-generation model. Any efficiency or other benefits that might hold under homoskedasticity should not therefore be expected to apply under heteroskedasticity.

To consistently estimate the variance of OLS estimates under heteroskedasticity, Huber et al. [29] and White [73] proposed the *model-robust variance estimate*

$$\hat{\Sigma}_n = \mathbf{A}_n^{-1}\mathbf{B}_n\mathbf{A}_n^{-1}, \text{ where } \mathbf{A}_n = \mathbf{X}^T\mathbf{X}, \mathbf{B}_n = \mathbf{X}^T\text{diag}(\boldsymbol{e}^2)\mathbf{X},$$

and $\text{diag}(\boldsymbol{e}^2)$ denotes the $n \times n$ diagonal matrix with diagonal entries given by the squared residuals $e_i^2 = (Y_i - \boldsymbol{X}_i^T\hat{\beta}_{LS})^2$. The model-robust variance estimate, also known as the Huber-White estimator or the sandwich variance estimate, is consistent with the asymptotic variance of the OLS estimator $\hat{\beta}_{LS}$ even in the presence of model violation. In other words, the convergence

$$\hat{\Sigma}_n^{-\frac{1}{2}}(\hat{\beta}_{LS} - \beta_0) \to_d N(0, I_p) \tag{3.6}$$

holds even under heteroskedasticity or non-linearity. (If the association is truly nonlinear, the parameter β_0 would indicate the linear trend "closest" to the true data-generating distribution in mean-square error, assuming uncorrelated and homoskedastic residual errors [24]). Replacing the model-based variance $\hat{\Sigma}_0 = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$ with the model-robust variance $\hat{\Sigma}_n$, we therefore obtain model-agnostic precision-weighted variants of the James-Stein estima-

tor, the Ridge estimator, and the LASSO. We denote them as

$$\hat{\beta}^0_{rJS} = \underset{\beta}{\operatorname{argmin}}(\beta - \hat{\beta}_{LS})^T \hat{\Sigma}_n^{-1}(\beta - \hat{\beta}_{LS}) + \frac{\lambda}{n}\|\mathbf{X}\beta\|_2^2, \tag{3.7}$$

$$\hat{\beta}^0_{rLASSO} = \underset{\beta}{\operatorname{argmin}}(\beta - \hat{\beta}_{LS})^T \hat{\Sigma}_n^{-1}(\beta - \hat{\beta}_{LS}) + \lambda\|\beta\|_1, \tag{3.8}$$

$$\hat{\beta}^0_{rRidge} = \underset{\beta}{\operatorname{argmin}}(\beta - \hat{\beta}_{LS})^T \hat{\Sigma}_n^{-1}(\beta - \hat{\beta}_{LS}) + \lambda\|\beta\|_2^2, \tag{3.9}$$

or equivalently

$$\hat{\beta}_{rJS,\lambda_n} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{Y} - \mathbf{X}\beta)^T \mathbf{X}\left[\mathbf{X}^T \operatorname{diag}(\boldsymbol{e}^2)\mathbf{X}\right]^{-1}\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\beta) + \frac{\lambda_n}{n}\|\mathbf{X}\beta\|_2^2, \tag{3.10}$$

$$\hat{\beta}_{rLASSO,\lambda_n} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{Y} - \mathbf{X}\beta)^T \mathbf{X}\left[\mathbf{X}^T \operatorname{diag}(\boldsymbol{e}^2)\mathbf{X}\right]^{-1}\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\beta) + \lambda_n\|\beta\|_1, \tag{3.11}$$

$$\hat{\beta}_{rRidge,\lambda_n} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{Y} - \mathbf{X}\beta)^T \mathbf{X}\left[\mathbf{X}^T \operatorname{diag}(\boldsymbol{e}^2)\mathbf{X}\right]^{-1}\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\beta) + \lambda_n\|\beta\|_2^2. \tag{3.12}$$

We call the estimators defined in (3.10), (3.11) and (3.12) the *Rotated James-Stein* estimator (rJS), the *Rotated Ridge* estimator (rRidge), and the *Rotated LASSO* estimator (rLASSO) respectively. We use the subscript λ_n to emphasize that the resulting estimators depend on the regularizing parameter λ_n. Note that the regularizing parameter in the classic James-Stein estimator has a well-known form given in (3.5), but in rJS we consider selecting λ_n adaptively.

We denote the estimators as "rotated" due to the Karush-Kuhn-Tucker conditions [8, Chap 5]). Both the LASSO and rLASSO estimators are the intersection of the some ellipsoid and the diamond $\|X\beta\|_1 = t$ for suitable tuning parameters $t > 0$. The ellipsoid for LASSO is $\{\beta : (\beta - \hat{\beta}_{LS})\hat{\Sigma}_0^{-1}(\beta - \hat{\beta}_{LS}) < t\}$ and for rLASSO is $\{\beta : (\beta - \hat{\beta}_{LS})\hat{\Sigma}_n^{-1}(\beta - \hat{\beta}_{LS}) < t\}$. Both ellipsoids are centered at $\hat{\beta}_{LS}$ and yet their axes point to different directions. Similar distinctions hold for rRidge versus classic Ridge and rJS versus classic JS. Figure 3.1 illustrates all three relationships, for simplicity in two dimensions.

The precision-weighted least square estimators can also be motivated through Wald tests. To test the null hypothesis $H_{0,\boldsymbol{\beta}_0} : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ versus the alternative hypothesis $H_{1,\boldsymbol{\beta}_0} : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, a Wald test based on the OLS estimator $\hat{\boldsymbol{\beta}}_{LS}$ rejects $H_{0,\boldsymbol{\beta}_0}$ if the value of the Wald statistics $(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}_0)^T \hat{\Sigma}_n^{-1}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}_0)$ is large. On the other hand, the penalty term

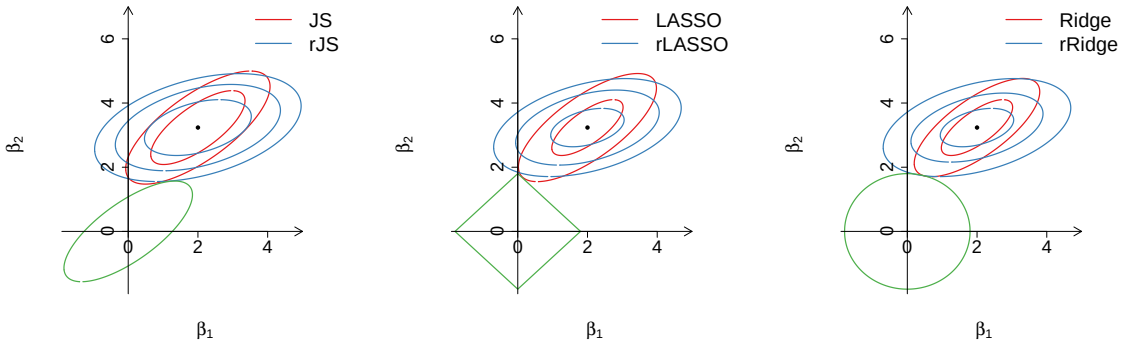

Figure 3.1: Illustration of Rotated James-Stein estimator (left), Rotated LASSO (middle) and Rotated Ridge estimator (right) in the two-dimension case.

$\mathcal{P}(\boldsymbol{\beta})$ in the penalized precision-weighted least square estimators can often be viewed as a constraint on the parameter space. For example, an alternative representation of the rLASSO optimization problem is as a constrained optimization, where we seek

$$\min_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta})^T\hat{\Sigma}_n^{-1}(\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta}) \quad \text{such that} \quad \|\boldsymbol{\beta}\|_1 \leq t,$$

for some specified $t > 0$. We see that rLASSO searches for the value of $\boldsymbol{\beta}_0$ in the space $\{\boldsymbol{\beta}_0 : \|\boldsymbol{\beta}_0\|_1 \leq t\}$ such that the Wald statistic for testing the null hypothesis H_{0,β_0} has the smallest value, where this search is also constrained to null values for which $\{H_{0,\beta_0}|\|\beta_0\|_1 \leq t\}$. In other words, from the perspective of a Wald test, the rLASSO estimator gives the null value of the parameter $\boldsymbol{\beta}$ within the constrained parameter space that is most concordant with the OLS estimate. The penalties in rJS and rRidge can also be viewed as constraints on the parameter space, where the constraints are $\|\mathbf{X}\beta\|_2^2 \leq t$ and $\|\beta\|_2^2 \leq t$, respectively, for some $t > 0$.

For penalized precision-weighted least square estimators, the rotated shrinkage estimate are model-agnostic in the following sense. The form penalization applied to the linear trend estimate $\hat{\boldsymbol{\beta}}_{LS}$ depends on empirically-derived variance estimates for that linear trend. Neither the original estimate nor the shrinkage therefore require a prespecified mean model, or assumptions about the variance such as homoskedasticity. The sole assumption we are

relying on to motivate the variance estimates used in the shrinkage is independence of the observations.

In the following sections, we present some findings regarding the theoretical properties of rJS, rLASSO, and rRidge.

3.4 Computation of rLASSO

Among the proposed estimators, rJS and rRidge have closed-form expressions and are easy to compute once the regularizing parameter λ_n is determined. Here we briefly describe the computation of rLASSO (3.11) given the regularizing parameter λ_n.

We first note that (3.11) is equivalent to

$$\hat{\beta}_{rLASSO,\lambda_n} = \underset{\beta}{\operatorname{argmin}} (\beta - \beta_{LS})^T \hat{\Sigma}_n^{-1} (\beta - \beta_{LS}) + \lambda \|\beta\|_1.$$

We assume the estimated precision matrix $\hat{\Sigma}_n^{-1}$ is positive definite and thus allows the Cholesky decomposition $\hat{\Sigma}_n^{-1} = L_n L_n^T$, where L_n is a $p \times p$ lower-triangular matrix with positive diagonal entries. Writing $\tilde{\gamma} = L_n^T \hat{\beta}_{LS}$ and $\gamma = L_n^T \beta$, we can rewrite the rLASSO objective function as

$$\|\tilde{\gamma} - \gamma\|_2^2 + \lambda_n \|(L_n^T)^{-1}\gamma\|.$$

This is the so-called generalized LASSO problem and can be efficiently solved by the path algorithm [66].

3.5 Consistency and Asymptotic Distribution

We first assume the observed data form an i.i.d. random sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ where $Y_i = \boldsymbol{X}_i^T \beta_0 + \epsilon_i$, and the residual ϵ_i satisfies $E(\epsilon_i | \boldsymbol{X}_i) = 0$ and $E(\epsilon_i^2 | \boldsymbol{X}_i) = \sigma^2(\boldsymbol{X}_i)$, and $\sigma : \mathbb{R}^p \to (0, \infty)$ is an unknown, positive, and measurable function of $\boldsymbol{X}_i$.

We further write

$$D_n = \frac{1}{n} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \operatorname{diag}(e^2) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}$$

and

$$D = E(\boldsymbol{X}_1 \boldsymbol{X}_1^T) \left[E(\sigma(\boldsymbol{X}_1)^2 \boldsymbol{X}_1 \boldsymbol{X}_1^T) \right]^{-1} E(\boldsymbol{X}_1 \boldsymbol{X}_1^T).$$

We also write $C_n = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ and $C = E[\boldsymbol{X}_1\boldsymbol{X}_1^T]$. It is well-known that $D_n \xrightarrow{p} D$ and $C_n \xrightarrow{p} C$.

The theorems in this section are inspired by the investigation of bridge estimators in Knight and Fu [35]. Proofs are all deferred to Appendix E.

Theorem 3.5.1 states the convergence of $\hat{\beta}_{rLASSO,\lambda_n}$, which implies that rLASSO is consistent if tuning parameter does not grow too quickly with sample size.

Theorem 3.5.1. *If D is positive definite and $\lambda_n/n \to \lambda_0 \geq 0$, then*

$$\begin{aligned}
\hat{\beta}_{rJS} &\xrightarrow{p} \operatorname*{argmin}_{\beta}(\beta-\beta_0)^T D(\beta-\beta_0) + \lambda_0\beta^T C\beta = (D+\lambda_0 C)^{-1}D\beta_0, \\
\hat{\beta}_{rLASSO,\lambda_n} &\xrightarrow{p} \operatorname*{argmin}_{\beta}(\beta-\beta_0)^T D(\beta-\beta_0) + \lambda_0\|\beta\|_1, \\
\hat{\beta}_{rRidge,\lambda_n} &\xrightarrow{p} \operatorname*{argmin}_{\beta}(\beta-\beta_0)^T D(\beta-\beta_0) + \lambda_0\|\beta\|_2^2 = (D+\lambda_0 I_p)^{-1}D\beta_0.
\end{aligned}$$

In Theorem 3.5.1, we see that if $\lambda_n = o(n)$, then all three estimators converge in probability to β_0 and are consistent. When $\lambda_0 > 0$, the three estimators are inconsistent, and their bias depend on the value of λ_0 and the particular form of the penalty.

Studying the asymptotic distribution of the proposed estimators further, we obtain

Theorem 3.5.2. *If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and C and D are nonsingular, then*

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{rJS,\lambda_n} - \beta_0) &\xrightarrow{d} \operatorname{argmin}(V_0) = D^{-1}(W - \lambda_0 C\beta_0) \sim N(-\lambda_0 D^{-1}C\beta_0, D^{-1}). \\
\sqrt{n}(\hat{\beta}_{rLASSO,\lambda_n} - \beta_0) &\to_d \operatorname{argmin}(V_1) \\
\sqrt{n}(\hat{\beta}_{rRidge,\lambda_n} - \beta_0) &\to_d \operatorname{argmin}(V_2) = D^{-1}(W - \lambda_0\beta_0) \sim N(-\lambda_0 D^{-1}\beta_0, D^{-1}),
\end{aligned}$$

where

$$\begin{aligned}
V_0(u) &= -2u^T W + u^T D u + 2\lambda_0\beta_0^T C u, \\
V_1(u) &= -2u^T W + u^T D u + \lambda_0 \sum_{j=1}^{p}[u_j \operatorname{sign}(\beta_{0j}) I(\beta_j \neq 0) + |u_j| I(\beta_{0j} = 0)], \\
V_2(u) &= -2u^T W + u^T D u + 2\lambda_0\beta_0^T u, \\
W &\sim N(0, D).
\end{aligned}$$

From Theorem 3.5.2, we see that with $\lambda_n = O(\sqrt{n})$, the rJS and rRidge are asymptotically biased, normally distributed estimators, with identical asymptotic covariance matrix as the OLS estimator. When $\lambda_n = o(\sqrt{n})$ or $\lambda_0 = 0$, however, the rJS and rRidge are asymptotically equivalent to the OLS estimator. Similarly, the rLASSO is asymptotically biased unless $\lambda_n = o(\sqrt{n})$, under which the rLASSO has the same asymptotic distribution has the OLS estimator.

3.6 Sign Consistency of rLASSO

When the linear mean model assumption (3.2) holds and there are many candidate covariates, of which only a subset are associated with the outcome, it is desirable that variable-selection methods have high probability of selecting the truly-associated covariates. For regression methods, the it is desirable that the signs of the identified nonzero regression coefficients should, with high probability, match those of the true regression coefficients, and thus give correction direction of the effects. The LASSO, viewed as a variable-selection method, is known to have these properties [77]. In this section, we show that rLASSO has similar properties. That is, under certain conditions, the rLASSO will recover the nonzero regression coefficients and their signs with high probability, in large samples.

We say an estimate $\hat{\beta}$ is *equal in sign* with β_0, written as $\hat{\beta} \stackrel{s}{=} \beta_0$, if $\text{sign}(\hat{\beta}_j) = \text{sign}(\beta_{0j})$ for $\beta_{0j} \neq 0$ and $\hat{\beta}_j = 0$ for $\beta_{0j} = 0$. We say an estimator $\hat{\beta}$ is *sign consistent* if $P(\hat{\beta} \stackrel{s}{=} \beta_0) \to 1$ as $n \to \infty$.

We denote the support of β_0 as $S = \{i \in \{1, \ldots, p\} : \beta_{0i} \neq 0\}$. Without loss of generality, we assume $S = \{1, \ldots, q\}$ for some $q \geq 1$ and less than p. For a p-vector v, we denote the subvector with entries $\{v_i : i \in S\}$ as v_S. We write

$$D_n = \frac{1}{n}\mathbf{X}^T\mathbf{X}[\mathbf{X}^T\text{diag}(\boldsymbol{e}^2)\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{X} = \begin{pmatrix} D_{11}^n & D_{12}^n \\ D_{21}^n & D_{22}^n \end{pmatrix},$$

where $D_{11}^n \in \mathbb{R}^{q\times q}$, $D_{12}^n \in \mathbb{R}^{q\times(p-q)}$, $D_{21}^n \in \mathbb{R}^{(p-q)\times q}$ and $D_{22}^n \in \mathbb{R}^{(p-q)\times(p-q)}$. Similarly we

decompose D as

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}.$$

Theorem 3.6.1 below states the **Generalized Irrepresentable Condition** that is sufficient for rLASSO's sign consistency. This condition is similar to the Irrepresentable Condition (due to Zhao and Yu [77]) which is sufficient for LASSO's sign consistency.

Theorem 3.6.1. *For fixed p, q and β_0, define the Generalized Irrepresentable Condition as*

$$|D_{21}D_{11}^{-1}\text{sign}(\beta_{0S})| < 1 \tag{3.13}$$

where the absolute value and the inequality applies elementwise. If this condition holds than for any λ_n that satisfies $\lambda_n/n \to 0$ and $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ with $0 \leq c < 1$, we have

$$P(\hat{\beta}_{rLASSO,\lambda_n} \overset{s}{=} \beta_0) \to 1.$$

Theorem 3.6.1 shows implies that under the Generalized Irrepresentable Condition, in large samples rLASSO will identify the covariates that are associated with the outcome with high probability and the direction of their effects. In Section 3.10, we show that the Generalized Irrepresentable Condition is numerically indistinguishable with the Irrepresentable Condition in [77] under a wide range of scenarios, indicating that rLASSO and LASSO perform similarly in terms of sign consistency.

3.7 Large Sample Behavior of rJS, rRidge, and rLASSO under Nonlinearity

In Sections 3.5 and 3.6 we have assumed the mean model is linear, i.e. that $E[\,\boldsymbol{Y}|\mathbf{X} = \mathbf{X}\,] = \mathbf{X}^T\boldsymbol{\beta}_0$, for some fixed $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. In reality, however, this linearity is unlikely to hold except when all covariates are binary. In this section, we consider the general setting where $Y_i = \mu(\boldsymbol{X}_i) + \epsilon_i$ for $i = 1, \ldots, n$ for some measurable function $\mu : \mathbb{R}^p \mapsto \mathbb{R}$, with $E(\epsilon_i|\boldsymbol{X}_i) = 0$ and $Var(\epsilon_i|\boldsymbol{X}_i) = \sigma^2(\boldsymbol{X}_i)$ for some positive measurable function $\sigma : \mathbb{R}^p \to \mathbb{R}$. We write $Y = \boldsymbol{\mu}(\mathbf{X}) + \epsilon$, where $\boldsymbol{\mu}(\mathbf{X}) = (\mu(\boldsymbol{X}_1), \ldots, \mu(\boldsymbol{X}_n))^T$.

With large samples, the ordinary least square estimator $\hat{\beta}_{LS}$ converges to

$\beta^* = E(\boldsymbol{X}_1\boldsymbol{X}_1^T)^{-1}E(\boldsymbol{X}_1\mu(\boldsymbol{X}_1))$, a p-vector describing the linear trend that quantifies association between outcome and covariates. Under homoskedascitcy the variance function is $\sigma^2(\cdot) = \sigma_0^2 > 0$ and the OLS estimator $\hat{\beta}_{LS}$ gives the best linear unbiased prediction [24].

In large samples for random $\mathbf{X}$ the Huber-White estimator $\hat{\Sigma}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{diag}(\boldsymbol{e}^2)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$ will generally provide a good approximation of the covariance matrix of $\hat{\beta}_{LS}$, but its large-sample limit:

$$\hat{\Sigma}_n/n \to_p E(\boldsymbol{X}_1\boldsymbol{X}_1^T)^{-1}E\left\{\boldsymbol{X}_1\boldsymbol{X}_1^T[(\mu(\boldsymbol{X}_1) - \boldsymbol{X}_1^T\beta^*)^2 + \sigma^2(\boldsymbol{X}_1)]\right\}E(\boldsymbol{X}_1\boldsymbol{X}_1^T)^{-1},$$

involves a term describing model-misspecification, and not variance directly. For fixed $\mathbf{X}$ this leads to conservative statements of the asymptotic variance, albeit usually only very mildly conservative.

We denote the inverse of the large-sample limit as D^*; that is,

$$D^* = E(\boldsymbol{X}_1\boldsymbol{X}_1^T)\left(E\left\{\boldsymbol{X}_1\boldsymbol{X}_1[(\mu(\boldsymbol{X}_1) - \boldsymbol{X}_1^T\beta^*)^2 + \sigma^2(\boldsymbol{X}_1)]\right\}\right)^{-1}E(\boldsymbol{X}_1\boldsymbol{X}_1^T).$$

Following Theorem 3.5.1, it is natural to speculate that $\hat{\beta}_{rJS,\lambda_n}$, $\hat{\beta}_{rRidge,\lambda_n}$, and $\hat{\beta}_{rLASSO,\lambda_n}$ converge to β^* in large samples with certain regularizing parameters (for proof, see Appendix E.4).

Theorem 3.7.1. *For any $\lambda_n \geq 0$ such that $\lambda_n/n \to \lambda_0 \geq 0$, we have*

$$\begin{aligned}
\hat{\beta}_{rJS,\lambda_n} &\xrightarrow{p} \operatorname*{argmin}_{\beta}(\beta-\beta^*)^T D^*(\beta-\beta^*) + \lambda_0\beta^T C\beta = (D^*+\lambda_0 C)^{-1}D^*\beta^*,\\
\hat{\beta}_{rLASSO,\lambda_n} &\xrightarrow{p} \operatorname*{argmin}_{\beta}(\beta-\beta^*)^T D^*(\beta-\beta^*) + \lambda_0\|\beta\|_1,\\
\hat{\beta}_{rRidge,\lambda_n} &\xrightarrow{p} \operatorname*{argmin}_{\beta}(\beta-\beta^*)^T D^*(\beta-\beta^*) + \lambda_0\|\beta\|_2^2 = (D^*+\lambda_0 I_p)^{-1}D^*\beta^*.
\end{aligned}$$

The following theorem further gives the asymptotic distributions of the three estimators if $\lambda_n = O(\sqrt{n})$.

Theorem 3.7.2. *If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and D^* is nonsingular, then*

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{rJS,\lambda_n} - \beta^*) &\xrightarrow{d} \operatorname{argmin}(V_0^*) = (D^*)^{-1}(W^* - \lambda_0 C^*\beta_0) \sim N(-\lambda_0(D^*)^{-1}C\beta^*, (D^*)^{-1}). \\
\sqrt{n}(\hat{\beta}_{rLASSO,\lambda_n} - \beta^*) &\to_d \operatorname{argmin}(V_1^*) \\
\sqrt{n}(\hat{\beta}_{rRidge,\lambda_n} - \beta^*) &\to_d \operatorname{argmin}(V_2^*) = (D^*)^{-1}(W^* - \lambda_0\beta^*) \sim N(-\lambda_0(D^*)^{-1}\beta^*, (D^*)^{-1}),
\end{aligned}$$

where

$$\begin{aligned}
V_0^*(u) &= -2u^T W^* + u^T D^* u + 2\lambda_0(\beta^*)^T C u, \\
V_1^*(u) &= -2u^T W^* + u^T D^* u + \lambda_0 \sum_{j=1}^{p} [u_j \operatorname{sign}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)], \\
V_2^*(u) &= -2u^T W^* + u^T D^* u + 2\lambda_0(\beta^*)^T u, \\
W^* &\sim N(0, D^*).
\end{aligned}$$

Theorems 3.7.1 and 3.7.2 imply that with large samples and small enough regularizing parameters, the proposed three shrinkage estimators converge to the "best linear estimator" β^*. However, with large regularizing parameters, the three shrinkage estimators converge to different quantities and have different asymptotic distributions.

3.8 Selecting the Regularizing Parameters

The proposed estimators in (3.10), (3.11) and (3.12) all involve a regularizing parameter λ_n. In general, the optimal regularizing parameter λ_n differs with the optimality criteria considered, as is the case in LASSO [78].

Shrinkage estimators are often used to improve prediction accuracy. With this goal in mind, we propose selecting the regularizing parameters to minimize the predictive risk $E[(\mu(\tilde{X}) - \tilde{X}\hat{\beta}_\lambda)^2]$, where $\hat{\beta}_\lambda$ is an estimator of regression coefficients computed using Y, $\mathbf{X}$ that involves the regularizing parameter λ, and $\tilde{X}$ is a p-dimensional random vector drawn from the same distribution of X_i's and independent with the data (Y_i, X_i), $i = 1, \ldots, n$. Writing $\tilde{Y}$ as the outcome variable and $\tilde{\epsilon}$ as the residual error associated with $\tilde{X}$, then we

can decompose prediction error of the outcomes as

$$E[(\tilde{Y} - \tilde{\boldsymbol{X}}^T\hat{\beta}_\lambda)^2] = E[(\mu(\tilde{\boldsymbol{X}}) - \tilde{\boldsymbol{X}}^T\hat{\beta}_\lambda)^2] + E[\sigma^2(\tilde{\boldsymbol{X}})].$$

To minimize the predictive risk we will therefore select the optimal regularizing parameter λ from a grid of candidate values, choosing the one that minimizes the mean out-of-sample error $E[(\tilde{Y} - \tilde{\boldsymbol{X}}^T\hat{\beta}_\lambda)^2]$. This quantity can be approximated using, for example, cross-validation [51] or nonparametric bootstrap [20].

3.9 Simulation Study: Point Estimates

In this section we investigate the risk properties of the novel shrinkage estimators – rJS, rRidge, and rLASSO – through simulation. With sample size n ranging from 100 to 1000, we simulate 10 observation-specific covariates $X_{i,1}$-$X_{i,10}$ that are independent standard normal random variables. Each observation's outcome is then generated from $Y_i|X_i \sim N(X_i^T\beta_0, \sigma(X_i)^2)$, where the true regression coefficients are one of the following;

Name	Regression coefficients	Description
(A)	$\beta_{0,1} = \cdots = \beta_{0,4} = 2, \beta_{0,5} = \cdots = \beta_{0,10} = 0$	A few covariates have nonzero effects
(B)	$\beta_{0,1} = \cdots = \beta_{0,10} = 0.4$	All covariates have small effects
(C)	$\beta_{0,1} = \cdots = \beta_{0,10} = 2$	All covariates have medium effects

We consider three forms of the residual variance given the covariates:

Name	$\sigma(\mathbf{x})^2$	Description
(i)	0.5	homoskedastic
(ii)	$\frac{1}{4}\left(\lvert x_1\rvert + \lvert x_2\rvert\right)^2$	heteroskedastic; residual variance increases with $\lvert x_1\rvert + \lvert x_2\rvert$
(iii)	$\exp\left(-2\left(\lvert x_1\rvert + \lvert x_2\rvert\right)\right)$	heteroskedastic; residual variance decreases with $\lvert x_1\rvert + \lvert x_2\rvert$

We simulate and analyze 2000 datasets from each scenario, and report the L_2-error of the estimates $\|\hat{\beta} - \beta_0\|_2^2 = \sum_{j=1}^p (\hat{\beta}_j - \beta_{0j})^2$ and the predictive risks for the following estimators:

(1) the OLS estimator: $\hat{\beta}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$;

(2) James-Stein estimator: $\hat{\beta}_{JS} = \left[1 - \dfrac{(n-p)(p-2)\hat{\sigma}^2}{(n-p+2)\hat{\beta}_{LS}^T\mathbf{X}^T\mathbf{X}\hat{\beta}_{LS}}\right]\hat{\beta}_{LS}$;

(3) Positive-part James-Stein estimator: $\hat{\beta}_{JS+} = \left[1 - \dfrac{(n-p)(p-2)\hat{\sigma}^2}{(n-p+2)\hat{\beta}_{LS}^T X^T X\hat{\beta}_{LS}}\right]_+ \hat{\beta}_{LS}$, where $(a)_+ = \max(a, 0)$;

(4) LASSO: $\hat{\beta}_{LASSO,\lambda_n} = \underset{\beta}{\operatorname{argmin}}\|\boldsymbol{Y} - \mathbf{X}\beta\|_2^2 + \lambda_n\|\beta\|_1$;

(5) Ridge estimator: $\hat{\beta}_{Ridge,\lambda_n} = \underset{\beta}{\operatorname{argmin}}\|\boldsymbol{Y} - \mathbf{X}\beta\|_2^2 + \lambda_n\|\beta\|_2^2$.

(6) rJS: $\hat{\beta}_{rJS,\lambda_n} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{Y} - \mathbf{X}\beta)^T\mathbf{X}\left[\mathbf{X}^T\mathrm{diag}(\boldsymbol{e}^2)\mathbf{X}\right]^{-1}\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\beta) + \dfrac{\lambda_n}{n}\|\mathbf{X}\beta\|_2^2$.

(7) rLASSO: $\hat{\beta}_{rLASSO,\lambda_n} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{Y} - \mathbf{X}\beta)^T\mathbf{X}\left[\mathbf{X}^T\mathrm{diag}(\boldsymbol{e}^2)\mathbf{X}\right]^{-1}\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\beta) + \lambda_n\|\beta\|_1$;

(8) rRidge: $\hat{\beta}_{rRidge,\lambda_n} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{Y} - \mathbf{X}\beta)^T\mathbf{X}\left[\mathbf{X}^T\mathrm{diag}(\boldsymbol{e}^2)\mathbf{X}\right]^{-1}\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\beta) + \lambda_n\|\beta\|_2^2$.

Among the above estimators, the last three are penalized precision-weighted least square estimators. Estimators (4)-(8) involve regularizing parameters, the values of which are chosen to minimize the five-fold cross validation errors.

The prediction risks are computed by drawing an additional $B = 10,000$ covariate vectors $\tilde{X}_b$ $(b = 1, \ldots, B)$ from the same distribution as the X_b's and are calculated as

$$\sum_{b=1}^{B}(\hat{\beta} - \beta_0)^T\mathbf{X}^T\mathbf{X}(\hat{\beta} - \beta_0).$$

Figures 3.2 and 3.3 respectively show the estimators' predictive risks and L_2-losses. The investigated methods perform similarly in most scenarios, and are essentially indistinguishable in the scenarios where all the regression coefficients are moderate and nonzero. When only some of the regression coefficients are nonzero (scenario (A)), LASSO and rLASSO outperform the other estimators, and rLASSO further has smallest predictive risks and L_2-loss if the residual variance $\sigma(X_i)^2$ decreases with $(|X_{i1}| + |X_{i2}|)$. When all the covariates

have small and nonzero effects (scenario (B)), Ridge and rRidge appear to have the smallest predictive risks L_2-losses on average. The differences between the estimators disappear with moderate to large samples.

3.10 Simulation Study: Sign Consistency

In Section 3.6 we showed that the Generalized Irrepresentable Condition is a sufficient condition for $rLASSO$ to be sign consistent. In this section, we demonstrate by simulation the relationship between the Generalized Irrepresentable Condition and the sign consistency of rLASSO.

The simulation setting is based on that of Zhao and Yu [77]. We have $p = 10$ covariates, out of which the support of true regression coefficients is $S = \{1, 2, 3\}$ and the values of nonzero coefficients are $\beta_{0S} = (5, 4, 1)$. We consider three residual variance functions :

Name	$\sigma(x)^2$	Description
(i)	0.1^2	homoskedastic
(ii)	$\sigma(x)^2 = (0.01 + 0.05(\lvert x_1\rvert + \lvert x_2\rvert))^2$	$\sigma(x)^2$ increases with $\lvert x_1\rvert + \lvert x_2\rvert$
(iii)	$\sigma(x)^2 = \left(\min(0.01 + \frac{0.2}{\lvert x_1\rvert + \lvert x_2\rvert}, 10)\right)^2$	$\sigma(x)^2$ decreases with $\lvert x_1\rvert + \lvert x_2\rvert$

For $b = 1, \ldots, 100$, we generate X of different designs as follows:

1. Generate a wide range of the covariance matrices of X, $\mathbf{S}_b = \mathrm{Cov}(X)$, from $\mathrm{Wishart}_p(I_p, p)$

2. Compute $E(\sigma(X_i)^2 X_i X_i^T)$ by Monte Carlo approximation. That is, we draw an additional sample of $x_j^* \sim N_p(0, S_B)$, $j = 1, \ldots, M = 10,000$, and calculate the quantity by $E(\sigma(X_i)^2 X_i X_i) = \frac{1}{M} \sum_{j=1}^{M} \sigma(X_i^*)^2 X_j^* (X_j^*)^T$

3. Compute the matrix $D = S_b^{-1} E(\sigma(X_i)^2 \mathbf{x}_i X_i^T) S_b^{-1}$

4. For the b-th design, compute the criterion for rLASSO's sign consistency $\eta_1 = 1 - \|D_{21}(D_{11})^{-1}\mathrm{sign}(\beta_S)\|_\infty$. We expect sign consistency in large samples if $\eta_1 > 0$.

5. Let $C = S_b^{-1}$. For each design also compute the criterion for LASSO's sign consistency $\eta_0 = 1 - \|C_{21}(C_{11})^{-1}\mathrm{sign}(\beta_S)\|_\infty$ as given in [77]

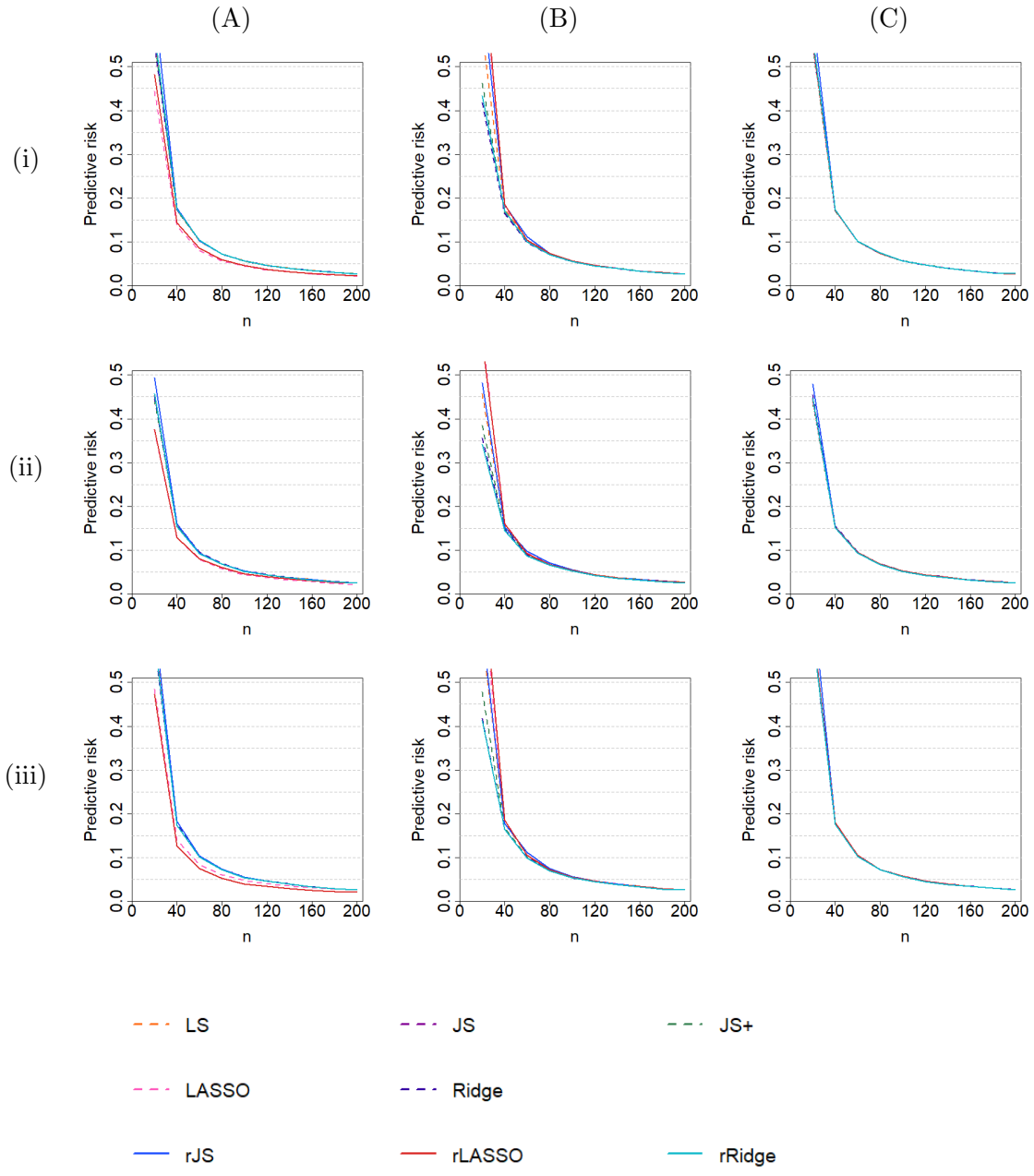

Figure 3.2: Predictive risks of the ordinary least square estimator (LS), James-Stein estimator (JS), the Positive-Part James-Stein estimator (JS+), LASSO, Ridge, Rotated James-Stein estimator (rJS), Rotated LASSO (rLASSO), and Rotated Ridge (rRidge). See the text for detailed simulation settings.

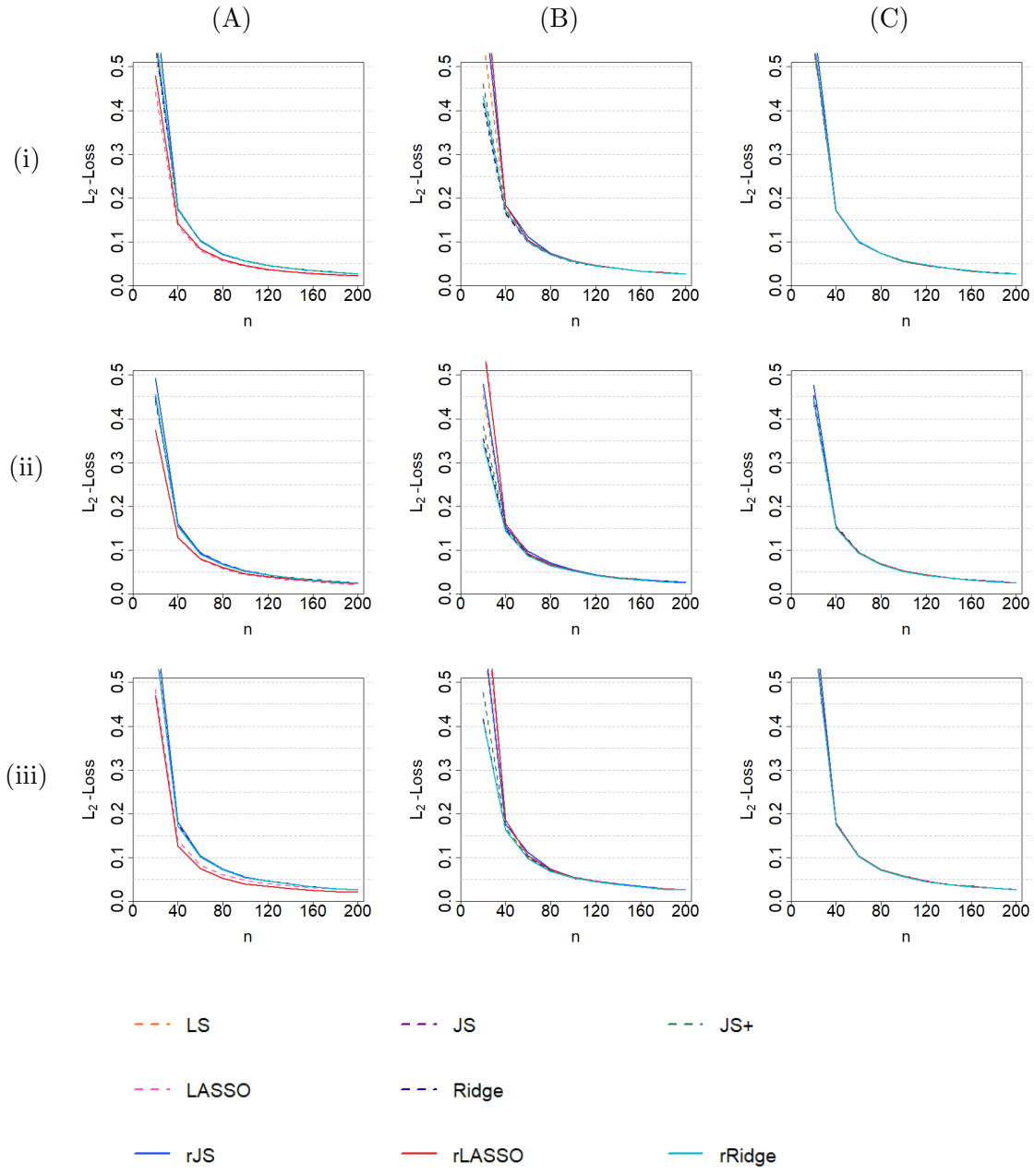

Figure 3.3: L_2-loss of the ordinary least square estimator (LS), James-Stein estimator (JS), the Positive-Part James-Stein estimator (JS+), LASSO, Ridge, Rotated James-Stein estimator (rJS), Rotated LASSO (rLASSO), and Rotated Ridge (rRidge). See the text for detailed simulation settings.

6. For the b-th design, we draw $n = 100$ samples of X_i from $N_p(0, S_b)$ and draws the outcome y_i from $N(X_i^T\beta_0, \sigma(X_i)^2)$ for $i = 1, \ldots, 100$

For each design, we draw 100 replicates, and plot against η_1 the frequency of those cases where the solution path of rLASSO or LASSO contains the true support and the signs of the estimates match those of the true coefficients.

Figure 3.4 shows the results. We first notice that the sign consistency criteria for rLASSO and LASSO have almost identical values for all scenarios considered. For each variance function, the frequency with which both the support and the signs are contained in the solution path is close to zero when $\eta_1 < 0$ and close to one when $\eta_1 > 0$, for both LASSO and rLASSO. In the neighborhood of $\eta_1 = zero$ we see a rapid growth of both LASSO and rLASSO's sign recovery rate, as η_1 increases.

3.11 Data Application: Prostate Cancer Data

To demonstrate the practical use of penalized precision-weighted least square estimators, we apply the proposed estimators to a prostate cancer dataset.

The prostate cancer dataset, from a small observational study, contain 97 observations. The recorded variables are cancer volume, prostate weight, age, benign prostatic hyperplasia amount, seminal vesicle invasion, capsular penetration, Gleason score, and prostate specific antigen [61]. Our goal is to identify the demographic or clinical variables associated with cancer volume. We dichotomized benign prostatic hyperplasia amount and capsular penetration by their presence, where the value is one if benigh prostatic hyperplasia or capsular penetration is present for the patient. As Figure 3.5 shows, the data exhibits heteroskedasticity and the outcome may have a nonlinear association with the covariates such as age.

We randomly split the data $7:3$ into training ($n = 67$) and testing ($n = 30$) data. We estimated the regression coefficients with the training data using the standard shrinkage estimators as well as the proposed penalized precision-weighted least square estimators, where we determined the optimal regularizing parameters λ_n of LASSO, Ridge, rJS, rLASSO and rRidge via three-fold cross validation. We computed the prediction error of the estimators

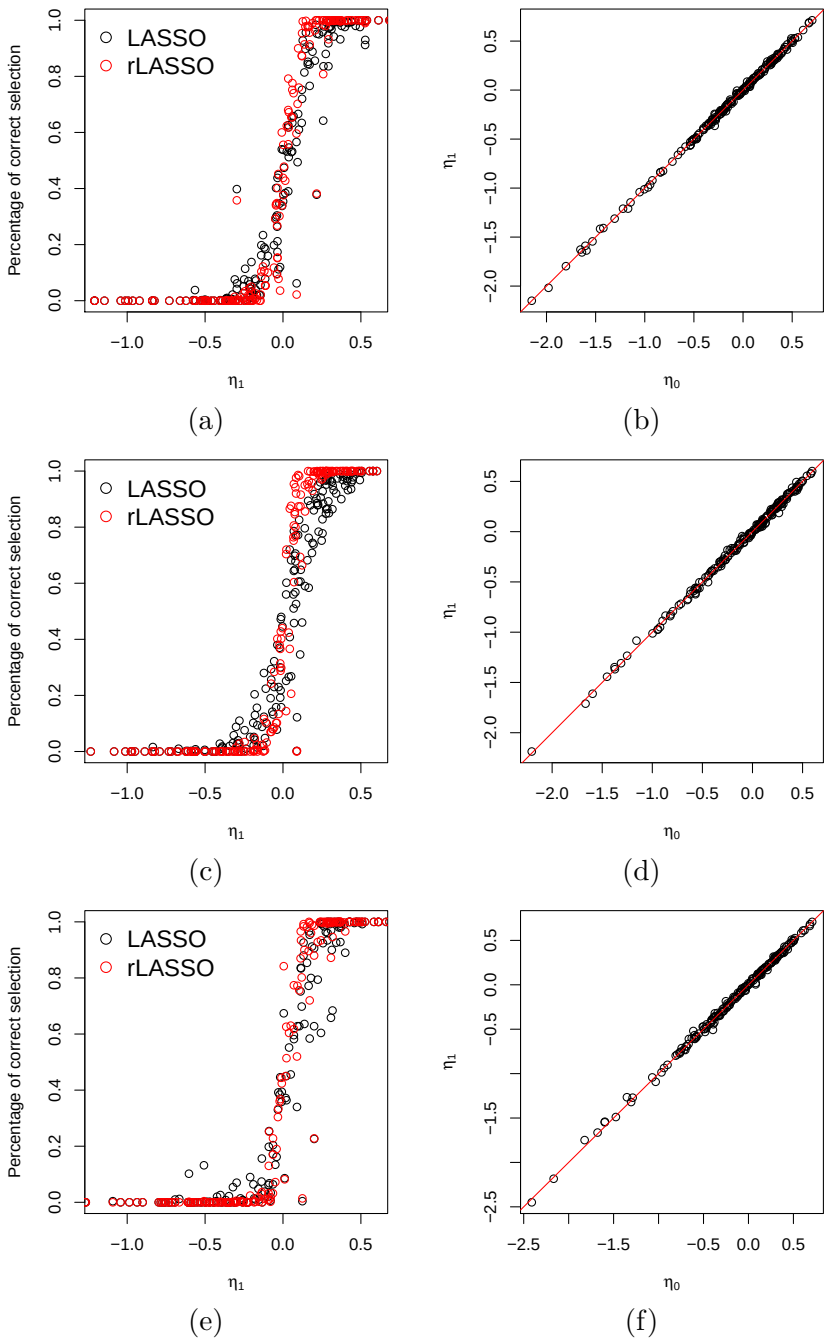

Figure 3.4: Sign consistency of LASSO and rLASSO. (a), (c), (e): Frequencies of LASSO or rLASSO solution path containing the true support versus the rLASSO sign consistency criterion $\eta_1 = 1 - \|D_{21}(D_{11})^{-1}\text{sign}(\beta_S)\|_\infty$, with the variance function $\sigma^2(\cdot)$ given by (i), (ii) or (iii); (b), (d), (f): Relationship of LASSO sign consistency criterion $\eta_0 = 1 - \|C_{21}(C_{11})^{-1}\text{sign}(\beta_S)\|_\infty$ versus rLASSO sign consistency criterion η_1.

on the testing data.

Table 3.11 shows the coefficients of the different estimators using the training data, as well as the optimal regularization parameters, training error and testing error. With the prostate cancer data, all the estimators give similar regression coefficients, among which rLASSO has the smallest testing error. Compared with LASSO, the optimal rLASSO by cross validation shrunk the regression coefficients for prostate weight and benigh prostatic hyperplasia to zero. Among all the covariates considered, PSA appears to have the strongest association with the prostate cancer volume, whereas the prostate weight and benign prostatic hyperplasia appear to have the weakest, if any, association.

Figure 3.6 shows the solution trajectory of the estimates that involve an regularizing parameter: the scaled OLS estimator, LASSO, Ridge, rJS, rLASSO, and rRidge. We include the solution trajectory of a simple scaled OLS estimator

$$\hat{\beta}_{sLS,\lambda_n} = \text{argmin}_{\beta\in\mathbb{R}^p}\|Y - \mathbf{X}\beta\|_2^2 + \lambda_n\|\mathbf{X}\beta\|_2^2 = \frac{1}{\lambda_n + 1}\hat{\beta}_{LS}$$

as a comparison to rJS, since the regular James-Stein estimator is the scaled OLS estimator with

$$\lambda_n = \frac{(n-p)(p-2)\hat{\sigma}^2}{(n-p+2)\hat{\beta}_{LS}^T X^T X\hat{\beta}_{LS} - (n-p)(p-2)\hat{\sigma}^2}.$$

When $\lambda_n = 0$, there is no shrinkage and all five estimates are equal to those of ordinary least squares. Similarly to LASSO, rLASSO has piecewise linear solution trajectories and all the coordinates become zero with a sufficiently large λ_n. In contrast, rRidge has smooth linear solution trajectories that converge to but never reach zero. The relative size of rLASSO coordinates are a little different from those of LASSO, and similar differences can be seen between rRidge and Ridge. While both the scaled OLS estimate and rJS have smooth solution trajectories and coordinates that approach but never reach zero, the relative magnitude of the scaled OLS estimates remain the same for every value of λ_n. The trajectories of rJS have changing relative magnitude and crossovers that are not present in the scaled OLS trajectories.

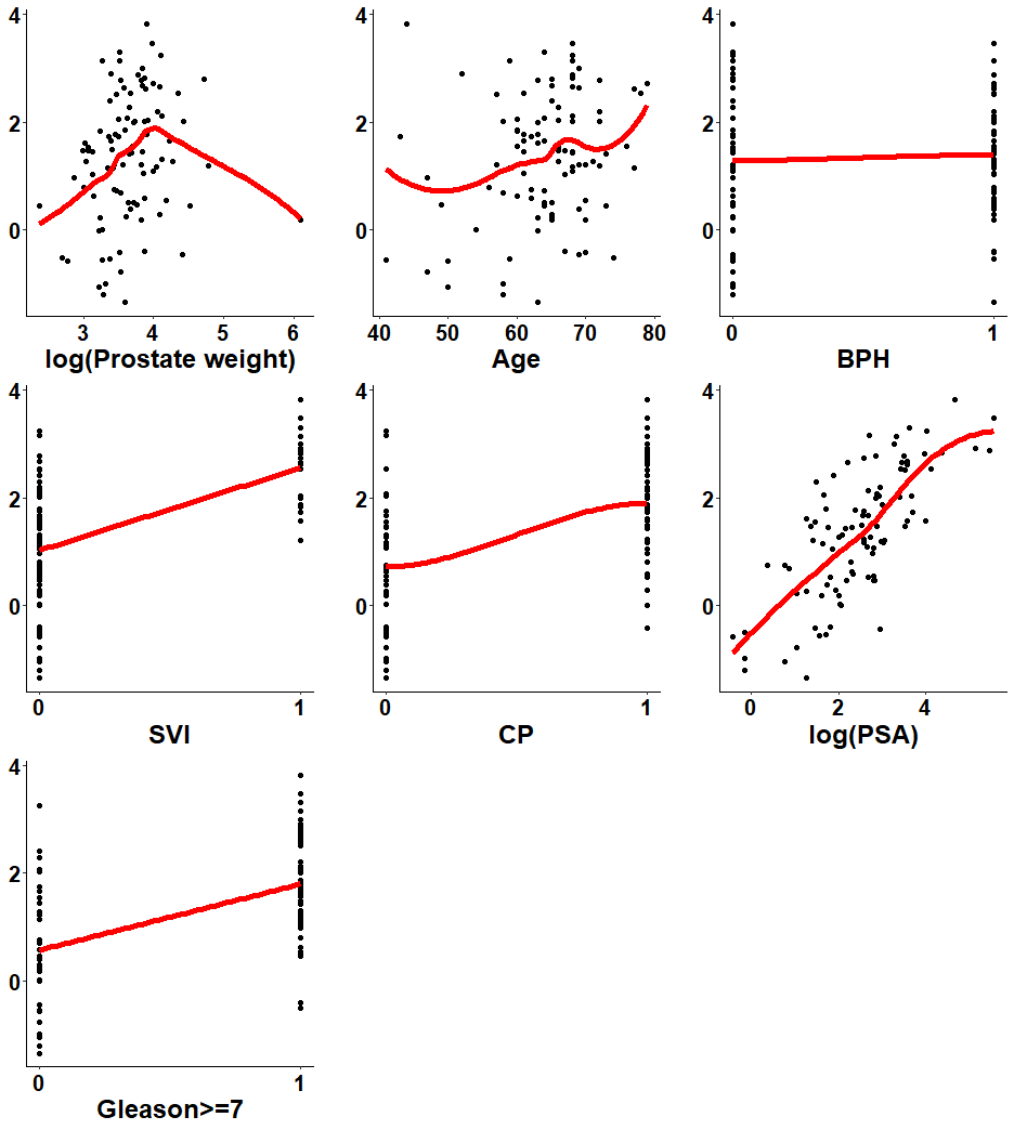

Figure 3.5: Scatter plots of the outcome versus each covariate in the prostate dataset, with univariate LOESS curves. The outcome is standardized to have mean zero; all the covariates are normalized to have mean zero and variance one. BPH: benign prostatic hyperplasia amount; SVI: seminal vesicle invasion; CP: Capsular penetration.

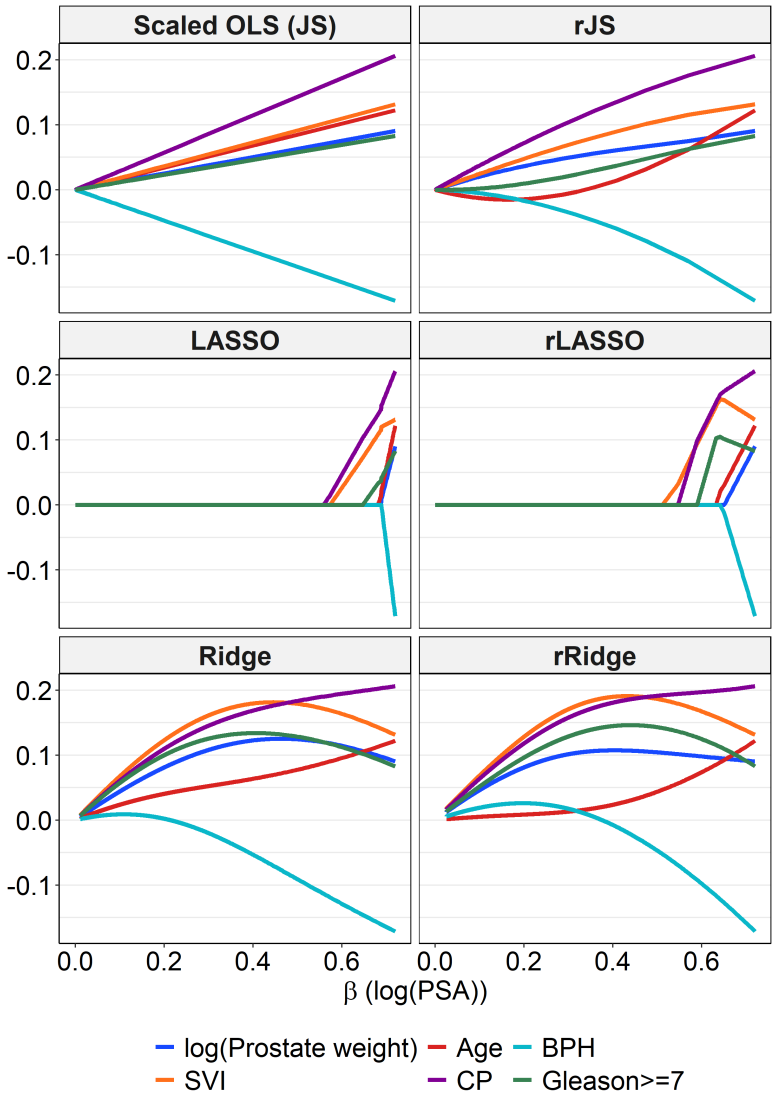

Figure 3.6: Solution trajectories relative for scaled OLS, rJS, LASSO, Ridge, rLASSO and rRidge relative to the regression coefficient of log(PSA) using the training dataset.

Table 3.1: Point estimates, training error and testing error of the ordinary least square estimator (LS), James-Stein estimator (JS), the Positive-Part James-Stein estimator (JS+), LASSO, Ridge, Rotated James-Stein estimator (rJS), Rotated LASSO (rLASSO), and Rotated Ridge (rRidge). We also showed the optimal regularization λ_n of LASSO, Ridge, rJS, rLASSO, and rRidge by three-fold cross validation.

	LS	JS	JS+	LASSO	Ridge	rJS	rLASSO	rRidge
log(Prostate weight)	0.090	0.084	0.084	0.019	0.119	0.090	0	0.099
Age	0.122	0.113	0.113	0.035	0.089	0.122	0.004	0.067
BPH	-0.171	-0.158	-0.158	-0.022	-0.117	-0.171	0	-0.087
SVI	0.131	0.121	0.121	0.122	0.170	0.131	0.154	0.172
CP	0.206	0.191	0.191	0.160	0.191	0.206	0.160	0.196
log(PSA)	0.720	0.666	0.666	0.693	0.568	0.720	0.634	0.581
Gleason$\geq$7	0.083	0.077	0.077	0.047	0.118	0.083	0.103	0.130
λ_n	/	/	/	16.528	11.505	0	19.283	18.273
training error	0.591	0.597	0.597	0.622	0.606	0.591	0.634	0.607
testing error	0.473	0.445	0.445	0.412	0.481	0.473	0.393	0.463

3.12 Discussion

In this chapter we have proposed a unified framework of shrinkage estimation in linear models – penalized precision-weighted least square estimation – in which a preliminary estimator is shrunk based on the precision with which we can estimate the coefficients. Unlike previous work on shrinkage estimators, the proposed framework of shrinkage estimation does not rely on distributional assumptions. Using a heteroskedasticity-consistent variance estimate, we propose 'rotated' variants of three popular shrinkage estimators: James-Stein estimator, LASSO, and Ridge, and in particular investigated a sufficient condition for the sign consistency of rLASSO. Our simulation studies show that the proposed estimators have comparable performances with the traditional shrinkage estimators under a wide range of scenarios. In particular, when the data is heteroskedastic and only a few covariates are associated with the outcome, rLASSO can have lower predictive risks than other estimators.

To our knowledge, by far no one else has expressed the James-Stein estimator as a penalized least square problem with the penalty proportional to the L_2-norm of the predicted value $\|\mathbf{X}\beta\|_2^2$. This is tantamount to introducing a bias to the OLS estimator so that the predictive value is shrunk towards zero, and in turn shrinking the OLS estimator towards

zero if the matrix $\mathbf{X}^T\mathbf{X}$ is positive definite.

To improve the efficiency of high-dimensional linear regression under heteroskedasticity, Wagener and Dette [70] proposed Penalized Weighted Least Squared Estimator (WLSE). Their method estimates the variance function $\sigma^2(\cdot)$ using nonparametric regression and solves a constrained weighted least square problem. The WLSE outperforms LASSO in their simulation studies in terms of mean squared errors. The estimator of σ^2, however, may not perform well when p is large and is badly biased if the linear mean model is misspecified. In contrast, our method does not attempt to estimate $\sigma^2(\cdot)$ and simply views the resulted estimator as a shrunken $\hat{\beta}_{LS}$, so should provide better robustness, albeit at the potential cost of using a noisier estimate of the variance of the OLS estimator.

An important limitation of our approach is that the three proposed estimators (rJS, rLASSO and rRidge) rely on the availability of the OLS estimator $\hat{\beta}_{LS}$ and the Huber-White sandwich estimator $\hat{\Sigma}_n$, which can only be applied to problems with $p < n$ and ideally with n much larger than p. In situations where $n \geq p$, the analog of the penalized precision-weighted least square estimator framework is unclear. One possible shrinkage estimator in linear model is given by choosing debiased LASSO [68, 76] as the initial estimator $\hat{\theta}_I$. However, to the authors' knowledge there doesn't exist a variance estimate for the debiased LASSO estimators that doesn't assume the homoscedasticity of the data.

Chapter 4

BAYESIAN VARIANCE ESTIMATION AND HYPOTHESIS TESTING USING INFERENCE LOSS FUNCTIONS

4.1 Introduction

Standard parametric statistics assumes a probabilistic model determined by a finite number of parameters. Those parameters characterize the data generating distribution, and the parameters or some transformation of them represent quantities of interest. Frequentist parametric statistical methods view the parameters as fixed quantities and use statements about replicate datasets to make inference, while Bayesian parametric methods uses the "language" of probability to directly describe uncertainty about parameters. This philosophical distinction can lead to practical differences in analytic methods, computation, and interpretation.

Yet despite their differences, in analysis of large samples many frequentist approaches have close Bayesian analogs. For example, maximum likelihood estimates and posterior means, or standard error estimates and posterior standard deviations — fundamentally due to Bernstein-von Mises' theorem [69, §10.2].

Under model violations, it is known that in large samples the posterior mean approaches the point in the model family closest (in Kullback-Leibler divergence) to the true data-generating distribution [34]. While the point estimates can therefore be interpreted as some form of 'best guess', the corresponding posterior variance is extremely hard to interpret, even in large samples. This contrasts sharply with frequentist use of "robust" covariance estimates [55], which are widely-used and provide valid large-sample inference for the parameter estimated consistently by the MLE, even under model misspecification. A general parametric Bayesian analog of this popular frequentist method is yet to be developed.

In this article, we proposed a Bayesian analog of robust covariance estimates, as the Bayes rule with respect to a form of *balanced* loss function [75]. The balanced loss function comprises two terms, penalizing estimation error and lack of model fit respectively. We show

by analytic examples and simulation studies that the proposed Bayesian robust covariance matrix converges asymptotically to the true covariance matrix of the posterior mean of the parameter of interest. We also show by simulation studies how, in small samples, the Wald-type confidence interval obtained using Bayes robust standard deviations can improve the frequentist coverage, compared to some standard methods.

4.2 Inference Loss functions

For general observations we denote $\mathbf{Z_n} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ as an independent and identically distributed random sample from distribution P_{θ_0} with density p_{θ_0} with respect to a measure μ, indexed by a real parameter $\theta_0 \in \Theta \subset \mathbb{R}^p$. In our major focus on regression models, the $\boldsymbol{Z}_i$'s are the combined outcome and explanatory variables, i.e. $\boldsymbol{Z}_i = (Y_i, \boldsymbol{X}_i)$ where $Y_i \in \mathbb{R}$ is the outcome variable and $\boldsymbol{X}_i$ is a real vector of explanatory variables. For Bayesian analysis, we use ϑ to denote the parameter, and write $\pi(\cdot)$ as the density function of the assumed prior, supported on Θ. We write $\pi_n(\cdot|\mathbf{Z_n})$ as the posterior density of ϑ, given the observed data. We denote E_{Π_n} and Cov_{Π_n} as the posterior expectation and posterior covariance matrix, respectively, with respect to the posterior measure Π_n.

Suppose $\theta \in \mathbb{R}^p$ is the parameter of interest and d a corresponding estimate. Using decision theory, optimal Bayesian estimates, known as *Bayes rules*, are obtained by minimizing the posterior risk $E_{\Pi_n}[L(\vartheta, d)|\mathbf{Z_n}]$, where $L(\cdot, d)$ is a loss function describing the discrepancy between the decision rule and the parameter. Common loss functions include the L_1-loss $\|\theta - d\|_1$ and L_2-loss $\|\theta - d\|_2$, for which the Bayes rules are the posterior median and posterior mean, respectively. For a more comprehensive review of decision theory and optimality of Bayes rules, see Parmigiani and Inoue [48].

To give rules that estimate θ but also give an indication of the precision of that estimate, we need more general loss functions. To achieve both those goals we consider what we shall call the *inference loss function*

$$L_I(\theta, d, \Sigma) = \log|\Sigma| + (\theta - d)^T \Sigma^{-1} (\theta - d), \tag{4.1}$$

where Σ is a $p \times p$ positive definite matrix. The right-hand term is a sum of adaptively-

weighted losses, each of the form $(\Sigma^{-1})_{ij}(\theta_i - d_i)(\theta_j - d_j)$, where the weights are determined by the elements of the matrix Σ^{-1}. To discourage the weight for any combination of the $(\theta_i - d_i)(\theta_j - d_j)$ from reaching zero, we penalize by the left term, the log of the determinant of Σ.

By Theorem F.0.1 in the Appendix, the Bayes rules for d and Σ with respect to L_I are $\hat{d}_n = E_{\Pi_n}[\vartheta|\mathbf{Z_n}]$ and $\hat{\Sigma}_n = Cov_{\Pi_n}[\vartheta|\mathbf{Z_n}]$, i.e. the posterior mean and variance. Bernstein von Mises Theorem implies that, in the frequentist sense, the posterior covariance matrix and the covariance matrix of the posterior mean are asymptotically equivalent, or $\hat{\Sigma}_n = Cov_{\Pi_n}[\vartheta|\mathbf{Z_n}] \approx Cov_{P_{\theta_0}}(\hat{d}_n)$. As a result, the inference loss function returns estimates and approximations of those estimates' error in both the Bayesian and frequentist senses.

We note that the inference loss function is not new: it was previously discussed in Dawid and Sebastiani [12], who noted that it corresponds to using the log-determinant of the covariance matrix as a criterion to compare study designs. Holland and Ikeda [28] also considered the inference loss as an example of a proper loss function. However, to our knowledge, the simple and general form of its Bayes rule is not known, nor are the extensions we provide below.

4.3 Balanced Inference Loss functions for model-robust variance estimation

We extend the inference loss function and present a Bayesian analog of robust covariance matrices, as Bayes rules for a loss which penalizes both the lack of model fit and estimation error. Our loss function is motivated by the balanced loss functions, originally proposed by Zellner [75]. Balanced loss functions have since been developed considerably, but all share a common form in which the loss is a weighted average of a term indicating lack-of-fit in some way, and one indicating estimation error. The general form of the balanced loss function we will consider is

$$L_{BI}(\theta, d, \Sigma, \Omega) = \log|\Sigma| + \underbrace{(\theta - d)^T \Omega \Sigma^{-1} (\theta - d)}_{\text{Estimation error}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \dot{l}_i(\theta)^T \left(\Omega I_n(\theta)\right)^{-1} \dot{l}_i(\theta)}_{\text{Lack of fit}}, \qquad (4.2)$$

where $\dot{l}_i(\theta) = \frac{\partial}{\partial\theta}\log\left[p_\theta(\boldsymbol{Z}_i)\right]$ is the score function based on a single observation, $I_n(\theta) = -\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log\left[p_\theta(\boldsymbol{Z}_i)\right]$ is the Fisher information, and we make decisions $d \in \mathbb{R}^p$, and Σ and Ω which are $p \times p$ positive-definite weighting matrices. We call (4.2) the *balanced inference loss* function.

The first term in L_{BI} is essentially the inference loss from (4.1), described earlier. However the balanced inference loss describes an additional decision, matrix Ω, that sets rates of tradeoff between elements of the estimation component of the inference loss versus the penalty $\frac{1}{n}\sum_{i=1}^n \dot{l}_i(\theta)^T I_n(\theta)^{-1}\dot{l}_i(\theta)$, a form of 'signal to noise' ratio for deviations from the model; the score evaluated at any $\theta \in \mathbb{R}^p$ describes discrepancies between the data and corresponding model, and the Fisher information matrix describes the uncertainty in those deviations. Furthermore, the penalty term is invariant to one-to-one transformations of θ. In Theorem **??** of the Appendix, we show that the balanced inference loss function's Bayes rules are

$$\hat{d}_n = E_{\Pi_n}(\vartheta|\mathbf{Z_n}), \quad \hat{\Omega} = E\left(\frac{1}{n}\sum_{i=1}^n \dot{l}_i(\vartheta)\dot{l}_i(\vartheta)^T I_n(\vartheta)^{-1}|\mathbf{Z_n}\right), \quad \hat{\Sigma}_n = Cov_{\Pi_n}(\vartheta|\mathbf{Z_n})\hat{\Omega}.$$

In Appendix H, we further show that $\hat{\Sigma}_n$ is asymptotically equivalent to the asymptotic covariance matrix of $\hat{d}_n$, under only mild regularity conditions that do not require fully-correct model specification, and is thus a Bayesian analog of the frequentist "robust" covariance matrix estimate.

As well as providing a Bayesian approach to these empirically-useful methods, our work also allows us to construct Bayesian analogs of Wald-type confidence intervals. We define the two-sided Bayesian robust confidence interval at level $(1-\alpha)$ for the jth entry of $\hat{d}_n$ as

$$\hat{d}_{n,j} \pm z_{1-\frac{\alpha}{2}}\hat{\Sigma}^{\frac{1}{2}}_{n,jj},$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. As we describe below, intervals constructed in this way can have appealing operating characteristics.

4.4 Examples and simulation studies

In this section we show that the Bayesian robust covariance matrix quantifies the variability of the Bayes point estimates in large samples, even in the presence of model misspecification, through analytic examples or simulation studies.

4.4.1 Estimating a normal mean, but misspecifying the variance

As a deliberately straightforward first example, suppose a random sample $X_1, \ldots, X_n$ is drawn from a normal distribution with unknown mean θ and variance σ_0^2. We are interested in estimating the mean θ but wrongly assume the observations have variance 1. In other words, we use the model $X_i|\vartheta = \theta_0 \sim N(\theta_0, 1)$, in which the variance is mis-specified. The analysis also uses prior $\vartheta \sim N(\mu, \eta^2)$, for known μ, η.

The balanced loss function in this setting is

$$L_{BI}(\theta, d, \sigma^2, \omega) = \log|\sigma^2| + \omega(\theta - d)^2/\sigma^2 + \frac{1}{n\omega}\sum_{i=1}^{n}(X_i - \theta)^2$$

and the Bayes rule sets

$$\hat{d}_n = \frac{\mu + \eta^2 \sum_{i=1}^n X_i}{n\eta^2 + 1}, \quad \hat{\sigma}_n^2 = \frac{\eta^2}{n(n\eta^2+1)}\sum_{i=1}^{n}(X_i - \bar{X})^2 + \left(\frac{\eta^2}{n\eta^2+1}\right)^2 + \frac{\eta^2(\mu - \bar{X})^2}{(n\eta^2+1)^3}.$$

Letting n go to infinity, the point estimate is equivalent to $\frac{1}{n}\sum_{i=1}^n X_i$ and the variance estimate is equivalent to σ^2/n, i.e. the asymptotic inference that would be obtained under a correct model-specification. The underlying approach remains fully parametric, and retains the (valuable) coherence of Bayesian analysis, together with the appealing normative element of decision theory, where a clear statement about the goal of the analysis (here, balancing estimation of θ with fidelity to the data) makes the optimal choice of what to report automatic.

4.4.2 Estimating the variance of Generalized Linear Model regression coefficients

The balanced inference loss function takes an appealingly straightforward form when the data are independent observations $Y_1, \ldots, Y_n$, from a generalized linear model (GLM), i.e.

an exponential family where the distribution has the density

$$p(Y_i|\theta,\alpha) = \exp\left(\frac{Y_i\theta_i - b(\theta_i)}{\alpha} + c(Y_i,\alpha)\right)$$

for functions $b(\cdot)$, $c(\cdot,\cdot)$ and scalars θ_i and α [43]. We assume the presence of p-dimensional explanatory variables $\boldsymbol{x}_i$ augmenting each observation y_i, with a link function $g(\cdot)$ connecting the mean function $\mu_i = E[Y_i|\theta_i,\alpha]$ and the linear predictor $\boldsymbol{x}_i^T\beta$ via $g(\mu_i) = \boldsymbol{x}_i^T\beta$, where β is a p-vector of regression coefficients. We write $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^T$ and $\mathbf{X} = (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)^T$.

With a GLM, the score function with respect to β for a single observation is $\dot{l}(\beta;\boldsymbol{X}_i) = \dfrac{\partial\mu_i}{\partial\beta}\dfrac{y_i-\mu_i}{\alpha V_i}$ and the empirical Fisher information is $I_n(\beta) = E[S(\beta)S(\beta)^T] = \mathbf{D}^T\mathbf{V}^{-1}\mathbf{D}/\alpha$, where $\mathbf{D}$ is the $n\times p$ matrix with elements $\partial\mu_i/\partial\beta_j$, $i=1,\ldots,n$, $j=1,\ldots,p$ and V is the $n\times n$ diagonal matrix with the i-th diagonal element $V_i = \dfrac{\partial\mu_i}{\partial\theta_i}$. The balanced inference loss is therefore

$$L_{BI}(\beta,d,\Sigma,\Omega) = \log|\Sigma| + (\beta-d)^T\Omega\Sigma^{-1}(\beta-d) + \sum_{i=1}^{n}\frac{(Y_i-\mu_i)^2}{\alpha V_i}\cdot\left(\frac{\partial\mu_i}{\partial\beta}\right)^T\left[\Omega\sum_{j=1}^{n}\left(\frac{\partial\mu_j}{\partial\beta}\right)\left(\frac{\partial\mu_j}{\partial\beta}\right)^T\right]^{-1}\left(\frac{\partial\mu_i}{\partial\beta}\right), \quad (4.3)$$

in which we observe that the data only enters the balancing term via the terms $(Y_i-\mu_i)^2/\alpha V_i$, which can be thought of as a Bayesian analog of (squared) Pearson residuals. The other components in the balancing term determine how much weight these squared '' terms receive. The balanced inference loss's Bayes rule sets d to be the usual posterior mean, and

$$\hat{\Sigma}_n = Cov_{\Pi_n}(\beta|\mathbf{Z_n})\cdot E_{\Pi_n}\left\{\mathbf{D}^T\mathbf{V}^{-1}\mathrm{diag}\{[\boldsymbol{Y}-\boldsymbol{\mu}]^2\}\mathbf{V}^{-1}\mathbf{D}[\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D}]^{-1}/\alpha\Big|\,\mathbf{Z_n}\right\},$$

which can be thought of as the model-based posterior variance 'corrected' by a term that assesses fidelity of the data to the model. As with the general case, asymptotic equivalence with the classical "robust" approach holds.

Using the canonical link function is used, i.e. where $\theta_i = \boldsymbol{x}_i^T\beta$, the Bayesian rule further simplifies to

$$\hat{\Sigma}_n = Cov_{\Pi}(\beta|\mathbf{Z_n})\cdot E_{\Pi_n}\left\{\mathbf{X}^T\mathrm{diag}\left\{[\boldsymbol{Y}-\boldsymbol{\mu}]^2\right\}\mathbf{X}[\mathbf{X}^T\mathbf{V}\mathbf{X}]^{-1}/\alpha|\mathbf{Z_n}\right\},$$

where the interpretation above can also be seen to hold.

Example: Linear regression

For observations $Y_1, \ldots, Y_n$ and the explanatory variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, a Bayesian linear regression model assumes $Y_i|\boldsymbol{X}_i, \beta, \sigma^2 \sim N(\boldsymbol{X}_i^T\beta, \sigma^2)$. The corresponding balanced inference loss is

$$L_{BI}(\beta, d, \Sigma, \Omega) = \log|\Sigma| + (\beta - d)^T\Omega\Sigma^{-1}(\beta - d) + \sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{X}_i^T\beta)^2}{\sigma^2} \cdot \boldsymbol{X}_i^T \left[\Omega \sum_{j=1}^{n} \boldsymbol{X}_j\boldsymbol{X}_j^T\right]^{-1} \boldsymbol{X}_i,$$

in which the novel penalty term is a sum of squared Pearson errors, weighted by a modified form of leverage ([43] Section 12.7.1, page 405) which includes a term Ω inside the familiar 'hat' matrix $X(X^TX)^{-1}X^T$. The Bayes rule for Σ is the Bayesian robust covariance matrix

$$\hat{\Sigma}_n = Cov_{\Pi_n}(\beta|\mathbf{Z_n}) \cdot E_{\Pi_n}\left(\mathbf{X}^T \mathrm{diag}\left((\boldsymbol{Y} - \mathbf{X}\beta)^2\right)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}/\sigma^2 \,|\, \mathbf{Z_n}\right).$$

We perform a simulation study to investigate the performance of Bayesian robust covariance matrices for linear regression. We generate univariate explanatory variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ as $\boldsymbol{X}_i = (1, U_i)^T$ where $U_i \sim U(0, 3)$ independently, and generate outcome variables as $Y_i|\boldsymbol{X}_i \sim N(U_i + aU_i^2, 1)$ for $i = 1, \ldots, n$, for various constants a. When $a \neq 0$, the mean model, $E(Y|\boldsymbol{X} = \boldsymbol{x})$ is quadratic in $\boldsymbol{x}$, and hence the model is misspecified.

We suppose interest lies in linear trend parameter β_2 and $\hat{d}_n$ is the posterior mean, i.e. the Bayes rule for its estimation. We use weakly informative priors $\beta_j \sim N(0, 10^3)$ for $j = 1, 2$ and $\sigma^{-2} \sim \mathrm{Gamma}(0.1, 0.1)$. For 1000 simulation under each setting, we report the average of posterior mean ($Ave(\hat{d}_n)$), the standard error of $\hat{d}_n$ ($SE(\hat{d}_n)$), the average posterior standard deviation of ϑ ($Ave(SD_{\Pi_n}(\vartheta|\mathbf{Z_n}))$), and the average Bayesian robust standard error ($Ave(\widehat{BRSE}(\hat{d}_n))$).

Table 4.1 compares the posterior standard deviation and the Bayesian robust standard error in estimating the standard error of the Bayes rule. In all the scenarios except when $a = 0$ (noted in grey, indicating that the linear regression model is correctly specified), while there is a recognizable difference between the true standard error and the posterior standard deviation, the Bayesian robust standard error is notably closer to the true standard error.

n	a	Ave$(\hat{d}_n)$	$SE(\hat{d}_n)$	Ave$(SD_{\Pi_n}(\vartheta\|\boldsymbol{Z_n}))$	$Ave(\widehat{BRSE}(\hat{d}_n))$
50	-2	-4.996	0.332	0.282	0.331
50	-1	-2.000	0.216	0.203	0.220
50	0	1.009	0.171	0.167	0.167
50	1	3.999	0.227	0.203	0.220
50	2	6.997	0.349	0.282	0.334
100	-2	-4.980	0.233	0.198	0.233
100	-1	-2.001	0.150	0.141	0.153
100	0	1.001	0.115	0.117	0.117
100	1	4.008	0.152	0.141	0.153
100	2	6.994	0.223	0.197	0.231

Table 4.1: Comparison between posterior standard deviation and the Bayesian robust standard error to the true standard error of the Bayes point estimate $\hat{d}_n$ in linear regression. Gray rows indicate where the model is correctly specified.

Example: Poisson regression

Poisson regression is a default method for studying the association between count data and the explanatory variables ([43] Chapter 6). With count observations $Y_1, \ldots, Y_n$ and corresponding explanatory variables (which may contain an intercept) $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, a Poisson regression model assumes $Y_i|\boldsymbol{X}_i \sim \text{Poisson}\left(\exp(\boldsymbol{X}_i^T\beta)\right)$. The balanced inference loss is

$$L_{BI}(\beta, d, \Sigma, \Omega) = \log|\Sigma| + (\beta - d)^T\Omega\Sigma^{-1}(\beta - d) + \sum_{i=1}^{n}(Y_i - \exp(\boldsymbol{X}_i^T\beta))^2\boldsymbol{X}_i^T\left[\sum_{j=1}^{n}\boldsymbol{X}_j\boldsymbol{X}_j^T\exp(\boldsymbol{X}_j^T\beta)\right]^{-1}\boldsymbol{X}_i,$$

which is again interpretable as a weighted sum of squared Pearson errors, and the weights are again modified versions of leverage. The Bayes rule for Σ is another Bayesian robust covariance matrix,

$$\hat{\Sigma}_n = Cov_{\Pi_n}(\beta|\mathbf{Z_n}) \cdot E_{\Pi_n}\left\{\mathbf{X}^T\text{diag}([\boldsymbol{Y} - \exp(\mathbf{X}\beta)]^2)\mathbf{X}\left[\mathbf{X}^T\text{diag}(\exp(\mathbf{X}\beta))X\right]^{-1}|\mathbf{Z_n}\right\}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $X = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$. As before, we generate explanatory variables $\boldsymbol{X}_i = (1, U_i)^T$, where $U_i \sim U(-3, 3)$, and generate the outcomes as $Y_i|\boldsymbol{X}_i \sim \text{Poisson}(U_i + aU_i^2)$, for various constants a and $i = 1, \ldots, n$. The model is misspecified when

$a \neq 0$.

We assume the parameter of interest is the log risk ratio $\theta = \beta_1$ and $\hat{d}_n$ is the posterior mean, the corresponding Bayes rule. We use weakly informative priors $\beta_j \sim N(0, 10^3)$ for $j = 1, 2$. For each setting we perform 1000 replicates, and report the same measures as for linear regression.

Table 4.2 shows the results. Again, the Bayesian robust standard error is a notably better estimate of the true standard error of $\hat{d}_n$ than the posterior standard deviation of ϑ; it performs better when the model is misspecified and no worse when correctly specified.

n	a	Ave$(\hat{d}_n)$	$SE(\hat{d}_n)$	Ave$(SD_{\Pi_n}(\vartheta \mid \boldsymbol{Z_n}))$	$Ave(\widehat{BRSE}(\hat{d}_n))$
50	-0.50	0.348	0.101	0.115	0.100
50	-0.25	0.576	0.091	0.098	0.090
50	0.00	1.014	0.086	0.085	0.084
50	0.25	1.788	0.134	0.070	0.118
50	0.50	2.870	0.296	0.049	0.181
100	-0.50	0.344	0.066	0.080	0.068
100	-0.25	0.567	0.060	0.068	0.062
100	0.00	1.005	0.059	0.058	0.058
100	0.25	1.808	0.089	0.049	0.084
100	0.50	2.954	0.163	0.034	0.136

Table 4.2: For a Poisson model, comparison of the posterior standard deviation and the Bayesian robust standard error to the true standard error of the posterior mean $\hat{d}_n$. Gray rows indicate where the model is correctly specified.

4.4.3 Estimating the variance of log hazards ratio with an exponential proportional hazards model

We demonstrate how our framework can be extended beyond GLMs with an association analysis for time-to-event outcomes and explanatory variables.

Suppose the observed data are $(T_i, \boldsymbol{X}_i, \Delta_i)$ for $i = 1, \ldots, n)$, where $T_i = \min(S_i, C_i)$ is the observed survival time, $\Delta_i = I(S_i \leq C_i)$ is the event indicator, S_i is the true survival time, C_i is the censoring time and X_i is a p-vector of covariates.

Suppose we intend to model the association between the survival time and the covariates

using the exponential proportional hazards model

$$S_i|\boldsymbol{X}_i, \beta \sim \text{Exponential}(\exp(\boldsymbol{X}_i^T\beta)).$$

The balanced inference loss is

$$L_{BI}(\beta, d, \Sigma, \Omega) = \log|\Sigma| + (\beta - d)^T\Omega\Sigma^{-1}(\beta - d) + \frac{1}{n}\sum_{i=1}^{n}(\Delta_i - T_i\exp(\boldsymbol{X}_i^T\beta))^2\boldsymbol{X}_i^T\left\{\Omega\left[\sum_{j=1}^{n}T_j\exp(\boldsymbol{X}_j^T\beta)\boldsymbol{X}_j\boldsymbol{X}_j^T\right]\right\}^{-1}\boldsymbol{X}_i$$

In this case, the interpretation of the balancing term is less clear than before. We note, however, for those subjects who experienced events (i.e. $\Delta_i = 1$), their contribution to the balancing term can be written as

$$\frac{(T_i - \Delta_i\exp(-\boldsymbol{X}_i^T\beta))^2}{\exp(-2\boldsymbol{X}_i^T\beta)}\boldsymbol{X}_i^T\left\{\Omega\left[\sum_{j=1}^{n}T_j\exp(\boldsymbol{X}_j^T\beta)\boldsymbol{X}_j\boldsymbol{X}_j^T\right]\right\}^{-1}\boldsymbol{X}_i,$$

which, again, is their Pearson residual weighted by some version of their leverage.

The Bayes rule for Σ is a Bayesian robust covariance matrix

$$\hat{\Sigma}_n = Cov_{\Pi_n}(\beta|\mathbf{Z_n})E_{\Pi_n}\left\{\boldsymbol{X}^T\text{diag}\left((\Delta_i - T_i\exp(\boldsymbol{X}_i^T\beta))^2\right)\mathbf{X}\left[\mathbf{X}^T\text{diag}\left(T_i\exp(\boldsymbol{X}_i^T\beta)\right)\mathbf{X}\right]^{-1} \mid \mathbf{Z_n}\right\}$$

We conduct a simulation study to investigate the robustness of different variance estimates in exponential proportional hazards models. We generated the covariates as $\boldsymbol{X}_i = (1, U_i)^T$ where $U_i \sim U(0, 3)$, for $i = 1, \ldots, n$, and the survival times by a Weibull proportional hazards model

$$S_i|\boldsymbol{X}_i, \beta \sim \text{Weibull}(\kappa, \exp(\boldsymbol{X}_i^T\beta)),$$

where the pdf of a Weibull(κ, λ) random variable is given by

$$f(t) = \lambda\kappa t^{\kappa-1}\exp(-\lambda t^\kappa), \qquad t > 0.$$

Our use of Exponential proportional model deviates from the true data generating distribution if $\kappa \neq 1$.

We further performed administrative censoring at $C_i = 10$, $i = 1, \ldots, n$.

We suppose the parameter of interest is the log hazards ratio $\theta = \beta_2$ in the model. We use the prior distribution $\beta_1 \sim N(0, 10^3)$ and $\beta_2 \sim N(0, 10^3)$.

Again, we conduct 1000 simulation under each setting considered. We report the average of posterior mean ($Ave(\hat{d}_n)$), the standard error of $\hat{d}_n$ ($SE(\hat{d}_n)$), the average posterior standard deviation of ϑ ($Ave(SD_{\Pi_n}(\vartheta|\boldsymbol{Z_n}))$), and the average Bayesian robust standard error ($Ave(\widehat{BRSE}(\hat{d}_n))$). Results are given in Table 4.3. Again, although slightly under estimating the standard error of the Bayes estimator $\hat{d}_n$ when the model is correctly specified, in general the Bayesian robust standard error is a better estimate of the true standard error of $\hat{d}_n$ than the posterior standard deviation of ϑ.

Table 4.3: For an Exponential Proportional Hazards Model, comparison of the posterior standard deviation and the Bayesian robust standard error to the true standard error of the posterior mean $\hat{d}_n$. Gray rows indicate where the model is correctly specified. #events: Average number of events.

n	κ	β	#events	Ave($\hat{d}_n$)	$SE(\hat{d}_n)$	Ave($SD_{\Pi_n}(\vartheta\|\boldsymbol{Z_n})$)	$Ave(\widehat{BRSE}(\hat{d}_n))$
50	0.8	0.00	49.9	-0.004	0.187	0.151	0.164
50	0.8	-0.25	49.8	-0.317	0.179	0.150	0.165
50	0.8	-0.50	49.4	-0.608	0.185	0.152	0.165
50	1.0	0.00	50.0	0.000	0.146	0.148	0.132
50	1.0	-0.25	49.8	-0.250	0.151	0.149	0.134
50	1.0	-0.50	49.9	-0.503	0.146	0.149	0.134
50	1.5	0.00	50.0	0.006	0.101	0.148	0.094
50	1.5	-0.25	50.0	-0.165	0.101	0.148	0.093
50	1.5	-0.50	50.0	-0.337	0.102	0.146	0.093
100	0.8	0.00	99.8	0.003	0.130	0.103	0.119
100	0.8	-0.25	99.6	-0.311	0.131	0.103	0.119
100	0.8	-0.50	98.8	-0.614	0.128	0.105	0.119
100	1.0	0.00	100.0	-0.006	0.102	0.102	0.095
100	1.0	-0.25	100.0	-0.248	0.102	0.103	0.096
100	1.0	-0.50	99.8	-0.497	0.104	0.102	0.097
100	1.5	0.00	100.0	-0.001	0.071	0.102	0.066
100	1.5	-0.25	100.0	-0.170	0.070	0.102	0.066
100	1.5	-0.50	100.0	-0.332	0.072	0.102	0.067

4.4.4 Bayesian robust confidence intervals

While appealingly robust in large samples, standard frequentist model-agnostic variance estimates often have unstable behavior in small samples. We investigate whether their Bayesian analog might produce better small sample behavior, by the shrinkage/stability provided through the prior distribution. Using the posterior mean as estimate $\hat{d}_n$ and Bayesian robust covariance matrix estimate $\hat{\Sigma}_n$, we construct the Bayesian analog of Wald-type "robust" confidence intervals. The proposed Bayesian robust confidence interval at confidence level $1-\alpha$ for θ_j, $j = 1, \ldots, p$, is

$$\hat{d}_{n,j} \pm z_{1-\frac{\alpha}{2}} \hat{\Sigma}_{n,jj}^{\frac{1}{2}}.$$

Figure 4.1 shows the coverage probabilities of these intervals, together with standard model-based 95% credible intervals and standard robust confidence intervals. We assume the same models and data distributions as in the previous sections. In all three regression examples, when the model is correctly specified, the posterior credible intervals has nominal coverage, whereas the Frequentist and Bayesian robust confidence intervals suffer from slight under coverage in small samples. When the model is misspecified, the posterior credible intervals will, as expected, suffer from under- or over-coverage even in large samples, depending on the specific form of model violation. With moderate to large sample sizes, the coverage probabilities of both the Frequentist and Bayesian robust confidence intervals tend to the nominal levels. With small sample sizes, compared with the standard robust confidence intervals, the Bayesian robust confidence intervals have coverage probabilities closer to the nominal level. This suggests that the stability provided by even the relatively weak priors we have used is sufficient to improve the performance of the standard purely-empirical frequentist methods.

4.5 Balanced Loss Function for quasi-likelihood regression

The Balanced Inference Loss function can easily be modified to provide a Bayesian analog for the variance estimate in a quasilikelihood regression model. Consider the same setup as in Section 4.4.2, with the additional assumption that $E[Y_i|\boldsymbol{X}_i] = \alpha V_i$ for some $\alpha > 0$. We

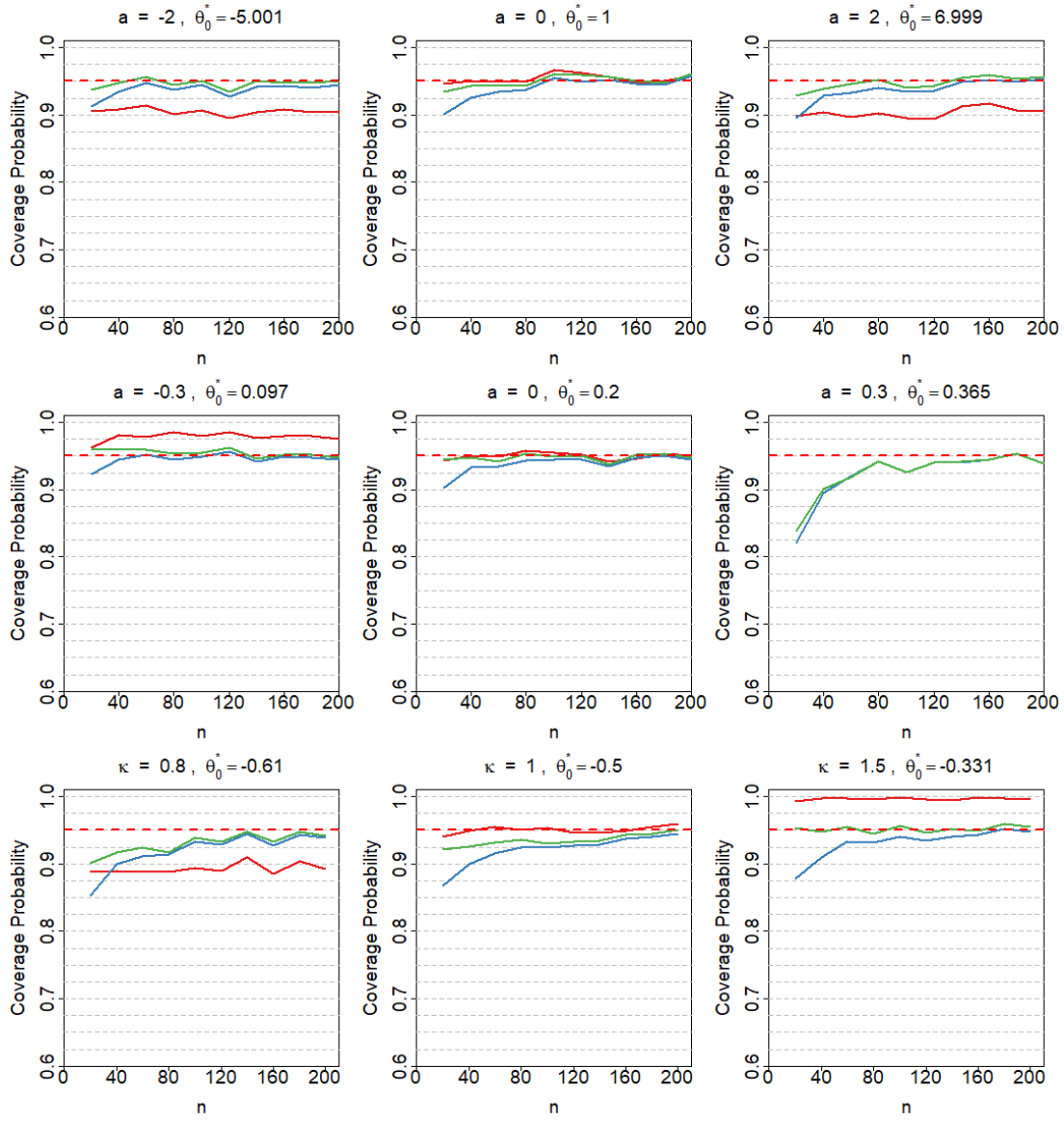

Figure 4.1: Coverage probabilities of 95% credible interval (red), frequentist (blue) and Bayesian (green) robust confidence intervals. From top to bottom: linear regression, Poisson regression and exponential proportional hazards model. The middle column shows the results with correctly-specified models. θ_0^* denotes the limit of the estimator of interest in each scenario.

propose the following balanced inference loss function

$$L_{BI}(\beta, d, \Sigma, \omega) = \log|\Sigma| + (\beta - d)^T \omega \Sigma^{-1} (\beta - d) + \frac{p}{n\omega} \sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{V_i}, \tag{4.4}$$

where $\omega > 0$ is a scaling factor. In contrast to the general loss function for GLMs (4.3), in (4.4) the correction term accounting for lack of model fit is proportional to the average Pearson residual $\frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{V_i}$. The Bayes rules for the loss function (4.4) set d to be the usual posterior mean of β, ω to be the posterior mean of average Pearson residual

$$\hat{\omega} = E_{\Pi_n}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{V_i}\middle| \boldsymbol{Y}, \mathbf{X}\right], \tag{4.5}$$

and Σ to be the posterior variance scaled by $\hat{\omega}_n$, i.e. $\hat{\Sigma}_n = \hat{\omega}_n \cdot \mathrm{Var}_{\Pi_n}[\beta|\boldsymbol{Y}, \mathbf{X}]$. It is easy to see that $\hat{\Sigma}_n$ gives a Bayes analog of the variance estimate of the frequentist quasi-likelihood method [72].

4.6 Application: Association of Systolic Blood Pressure and Age in 2017-2018 NHANES Data

We demonstrate the Bayesian robust standard error estimator using data from 2017-2018 National Health and Nutrition Examination Survey (NHANES) data. We are interested in the association between subjects' systolic blood pressure systolic and their age among the subpopulation with age between 18 and 60 years old. To make the difference between the variance estimates more distinguishable, we randomly select a subset of 200 subjects for our analysis.

We used a simple linear regression model as a working model, with the average of two systolic blood pressure readings as the outcome and age and gender as covariates (Figure 4.2 left panel). We computed both the model-based standard error estimates (assuming homoscedasticity) and robust standard error estimates (resistent to heteroscedasticity and/or model misspecification) of the regression coefficients. For the corresponding Bayesian model, we postulate a normal linear regression model with the same outcome and covariates. We use weakly informative prior distributions for the unknown regression coefficients and the residual variance (Figure 4.2 right panel). We report the posterior mean as point estimates,

and compute the posterior standard deviations and Bayesian standard errors for the regression coefficients. For the Bayesian analysis, we used Gibbs sampling with 3 chains and 30,000 iterations for each chain, where the first 18,000 iterations in each chain are burn-in and discarded from the analysis. We carried out the Bayesian analysis using JAGS [52].

Frequentist	Bayesian
$E[\text{SBP}] = \beta_0 + \beta_1 \text{MALE} + \beta_2 \text{AGE}$	$\text{SBP} \sim N(\beta_0 + \beta_1 \text{MALE} + \beta_2 \text{AGE}, \sigma^2)$ $\beta_j \sim N(0, 1, 000), \quad j = 0, 1, 2$ $\sigma^2 \sim InvGamma(0.01, 0.01)$

Figure 4.2: Frequentist and Bayesian regression models for analyzing the association between systolic blood pressure and age, adjusting for subjects' gender. SBP: systolic blood pressure; MALE: the indicator for male; AGE: subjects age in years.

In Table 4.4, we report the point estimates and the variance estimates of the regression coefficients for the frequentist and the Bayesian linear regression models. We focus on the comparison of the variance estimates. The model-based standard errors/Bayesian posterior standard deviations given higher estimates of the standard error for all three regression coefficients than the frequentist/Bayesian robust standard errors. For the intercept and the regression coefficient for age, the Bayesian posterior standard deviations are nearly equal to the corresponding frequentist model-based standard errors, whereas the Bayesian robust standard error are approximately equal to the frequentist robust standard error. The four versions of standard error estimates for the regression coefficient of gender is not as distinguishable.

Code for the analysis and the dataset are available on GitHub (`https://github.com/KenLi93/BRSE`).

4.7 Discussion

In this chapter, we have proposed a balanced inference loss function, whose Bayes rules provide an Bayesian analog of robust covariance matrix. We have proven how the Bayesian robust covariance matrix converges to the true covariance matrix of the estimator of interest

Table 4.4: Point estimates and standard error estimates of regression coefficients in Frequentist and Bayesian linear regression models using the NHANES data. Post. SD: posterior standard deviation; BRSE: Baysian robust standard error.

	Frequentist			Bayesian		
	Est.	SE	Robust SE	Est.	Post. SD	BRSE
(Intercept)	94.067	2.350	1.781	93.481	2.307	1.767
Male	4.817	2.063	2.032	5.005	2.062	2.050
Age (yrs)	0.570	0.046	0.042	0.580	0.046	0.042

in large samples, under only mild regularity conditions. In several examples, including GLMs and other forms of models, the corresponding loss function is seen to be a straightforward balance between losses for standard model-based inference, and a term that assesses how well the corresponding model actually fits the data that have been observed; the penalty employs terms and components that are familiar from diagnostic use of Pearson residuals, and from study of leverage. Using these methods we also proposed the Wald confidence intervals using Bayesian robust standard error, and found more stable behavior than their frequentist counterparts.

Our development of the robust covariance estimates is fully parametric, pragmatically adjusting the standard model-based inference with a measure of how the corresponding model fits the data. This is not the only motivation available for robust covariance estimates however; focusing only on linear regression, Szpiro et al. [64] proposed a Bayesian model-robust variance estimate by using highly-flexible models for the data distribution and clearly defining the parameter of interest in a model-agnostic manner. The extension of this essentially non-parametric approach beyond linear regression is unclear, and moreover its motivation can be challenging – much like the challenge of a frequentist development built purely on estimating equations, without parametric assumptions. In comparison, our proposed framework is general and can be applied to obtain the Bayesian robust covariance matrix in any regular parametric models. In comparison, our proposed method does not require redefining the parameter of interest during the modeling, and the interpretation of the point estimates stays the same as in regular parametric statistics. The proposed Bayesian variance estimates do not require nonparametric Bayes components in the model specifica-

tion as in Szpiro et al. [64], thus greatly reducing the computation time and increasing the algorithmic stability. Our framework is general and can be applied to obtain the Bayesian robust covariance matrix in any regular parametric models.

Our work demonstrated that robust standard error estimates are ubiquitous both in Frequentist and in Bayesian statistics. Our approach enables the proper quantification of the variability of parameter estimates in Bayesian parametric models. The proposed balance inference loss function, through which the Bayesian robust standard error was derived, also provides insights on the source of discrepancy between the model-based and model-agnostic variance estimates.

Chapter 5

CONCLUSION

In this dissertation, we added to the rich literature of model-agnostic statistics novel methods in several overlooked areas and provided improved methods. These methods include heterogeneity-resistant variance estimates and confidence intervals in meta-analysis, shrinkage estimators in linear models with model-agnostic motivations, and a model-robust Bayesian variance estimate. We hoped these methods have model-agnostic motivations, model-violation robust properties, and are easily computable for practitioners.

The raising complexity in the real world applications calls for novel statistical methods. The study of model-agnostic methods may continue to facilitate more principled inference and data-driven decision making.

BIBLIOGRAPHY

[1] *Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2017-2018.*

[2] Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[3] Alan Agresti et al. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, 1992.

[4] Hirotsugu Akaike. Information theory and an extension of the likelihood principle. In *Proceedings of the Second International Symposium of Information Theory*, 1973.

[5] TA Bancroft and Chien-Pai Han. Inference based on conditional specification: a note and a bibliography. *International Statistical Review/Revue Internationale de Statistique*, pages 117–127, 1977.

[6] Alvin J Baranchik. Multiple regression and estimation of the mean of a multivariate normal distribution. Technical report, STANFORD UNIV CALIF, 1964.

[7] MW Birch. The detection of partial association, i: the 2× 2 case. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):313–324, 1964.

[8] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[9] NE Breslow and KY Liang. The variance of the mantel-haenel estimator. *Biometrics*, pages 943–952, 1982.

[10] G Casella and RL Berger. Statistical inference. 2002.

[11] Arthur Cohen. All admissible linear estimates of the mean vector. *The Annals of Mathematical Statistics*, pages 458–463, 1966.

[12] A Philip Dawid and Paola Sebastiani. Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, pages 65–81, 1999.

[13] Clara Domínguez Islas and Kenneth M Rice. Addressing the estimation of standard errors in fixed effects meta-analysis. *Statistics in Medicine*, 37(11):1788–1809, 2018.

[14] Bradley Efron and Carl Morris. Limiting the risk of bayes and empirical bayes estimators—part i: the bayes case. *Journal of the American Statistical Association*, 66(336): 807–815, 1971.

[15] Bradley Efron and Carl Morris. Limiting the risk of bayes and empirical bayes estimators—part ii: The empirical bayes case. *Journal of the American Statistical Association*, 67(337):130–139, 1972.

[16] Malay Ghosh. On some bayesian solutions of the neyman-scott problem. In *Statistical Decision Theory and Related Topics V*, pages 267–276. Springer, 1994.

[17] Judith A Giles and David EA Giles. Pre-test estimation and testing in econometrics: recent developments. *Journal of Economic Surveys*, 7(2):145–197, 1993.

[18] Sander Greenland, James M Robins, Judea Pearl, et al. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, 1999.

[19] Marvin Gruber. *Improving efficiency by shrinkage: The James–stein and ridge regression estimators*. Routledge, 2017.

[20] Peter Hall. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of multivariate analysis*, 32(2):177–203, 1990.

[21] Joachim Hartung, Guido Knapp, and Bimal K Sinha. *Statistical meta-analysis with applications*, volume 738. John Wiley & Sons, 2011.

[22] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[23] Walter W Hauck. The large sample variance of the mantel-haenel estimator of a common odds ratio. *Biometrics*, pages 817–819, 1979.

[24] Charles R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975.

[25] Julian PT Higgins and Simon G Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558, 2002.

[26] Julian PT Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, 2009.

[27] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[28] Matthew J Holland and Kazushi Ikeda. Minimum proper loss estimators for parametric models. *IEEE Transactions on Signal Processing*, 64(3):704–713, 2015.

[29] Peter J Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press, 1967.

[30] Dan Jackson and Ian R White. When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, 60(6):1040–1058, 2018.

[31] W James and C Stein. Estimation with quadratic loss: Proc. 4th berkeley symp. math. stat. and prob. 1961.

[32] William James and Charles Stein. Estimation with quadratic loss. In *Breakthroughs in statistics*, pages 443–460. Springer, 1992.

[33] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ, 2002.

[34] Bastiaan Jan Korneel Kleijn, Adrianus Willem Van der Vaart, et al. The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

[35] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

[36] Jochem König, Ulrike Krahn, and Harald Binder. Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons. *Statistics in Medicine*, 32(30):5414–5429, 2013.

[37] O Kuss. Statistical methods for meta-analyses including information from studies without any events—add nothing to nothing and succeed nevertheless. *Statistics in Medicine*, 34(7):1097–1116, 2015.

[38] Nan M Laird and Frederick Mosteller. Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care*, 6(1): 5–30, 1990.

[39] Peter W Lane. Meta-analysis of incidence of rare events. *Statistical methods in medical research*, 22(2):117–132, 2013.

[40] Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, pages 21–59, 2005.

[41] Nathan Mantel and William Haenel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

[42] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[43] P McCullagh and John A Nelder. *Generalized Linear Models*, volume 37. CRC Press, 1989.

[44] Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

[45] Hisashi Noma and Kengo Nagashima. A note on the mantel-haenszel estimators when the common effect assumptions are violated. *Epidemiologic Methods*, 5(1):19–35, 2016.

[46] Luigi Pagliaro, Gennaro D'Amico, Thorkild IA Sörensen, Didier Lebrec, Andrew K Burroughs, Alberto Morabito, Fabio Tiné, Flavia Politi, and Mario Traina. Prevention of first bleeding in cirrhosis: a meta-analysis of randomized trials of nonsurgical treatment. *Annals of Internal Medicine*, 117(1):59–70, 1992.

[47] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[48] Giovanni Parmigiani and Lurdes Inoue. *Decision theory: Principles and approaches*, volume 812. John Wiley & Sons, 2009.

[49] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[50] Richard Peto. Why do we need systematic overviews of randomized trials? (transcript of an oral presentation, modified by the editors). *Statistics in Medicine*, 6(3):233–240, 1987.

[51] Richard R Picard and R Dennis Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984.

[52] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria., 2003.

[53] David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

[54] Kenneth Rice, Julian PT Higgins, and Thomas Lumley. A re-evaluation of fixed effect (s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(1):205–227, 2018.

[55] Richard M Royall. Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review/Revue Internationale de Statistique*, pages 221–226, 1986.

[56] Gerta Rücker, Guido Schwarzer, James R Carpenter, Harald Binder, and Martin Schumacher. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics*, 12(1):122–142, 2010.

[57] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003.

[58] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[59] Thomas A Severini. On the relationship between bayesian and non-bayesian elimination of nuisance parameters. *Statistica Sinica*, pages 713–724, 1999.

[60] M Iqbal Shamsudheen and Christian Hennig. Should we test the model assumptions before running a model-based test? *arXiv preprint arXiv:1908.02218*, 2019.

[61] Thomas A Stamey, John N Kabalin, John E McNeal, Iain M Johnstone, Fuad Freiha, Elise A Redwine, and Norman Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of urology*, 141(5):1076–1083, 1989.

[62] Alexander J Sutton, Nicola J Cooper, Paul C Lambert, David R Jones, Keith R Abrams, and Michael J Sweeting. Meta-analysis of rare and adverse event data. *Expert review of pharmacoeconomics & outcomes research*, 2(4):367–379, 2002.

[63] Michael J Sweeting, Alexander J Sutton, and Paul C Lambert. What to add to nothing?

use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, 23(9):1351–1375, 2004.

[64] Adam A Szpiro, Kenneth M Rice, Thomas Lumley, et al. Model-robust regression and a bayesian "sandwich" estimator. *The Annals of Applied Statistics*, 4(4):2099–2113, 2010.

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

[66] Ryan J Tibshirani, Jonathan Taylor, et al. The solution path of the generalized lasso. *The annals of statistics*, 39(3):1335–1371, 2011.

[67] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[68] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

[69] Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[70] Jens Wagener and Holger Dette. Bridge estimators and the adaptive lasso under heteroscedasticity. *Mathematical Methods of Statistics*, 21(2):109–126, 2012.

[71] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[72] Robert WM Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447, 1974.

[73] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.

[74] Barnet Woolf. On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19(4):251–253, 1955.

[75] Arnold Zellner. Bayesian and non-bayesian estimation using balanced loss functions. In *Statistical Decision Theory and Related Topics V*, pages 377–390. Springer, 1994.

[76] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242, 2014.

[77] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

[78] Hui Zou, Trevor Hastie, Robert Tibshirani, et al. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

Appendix A

LARGE-SAMPLE PROPERTIES OF THE ESTIMATORS

The probabilistic limits of log-CMH, Woolf's estimator and MLE can be found using Slutsky's Theorem. Below we show the sketch proof of results stated in Section 2.3.2.

A.1 log-CMH and Woolf results

The asymptotic distribution of log-CMH and Woolf's estimator are obtained by multivariate Delta-method[69] and the fact that

$$\sqrt{N}\left(\begin{pmatrix} \frac{a_1}{m_{11}} \\ \frac{b_1}{m_{01}} \\ \vdots \\ \frac{a_K}{m_{1K}} \\ \frac{b_K}{m_{0K}} \end{pmatrix} - \begin{pmatrix} p_{11} \\ p_{01} \\ \vdots \\ p_{1K} \\ p_{0K} \end{pmatrix}\right) \rightarrow_d$$

$$N_{2K}\left(0, \begin{pmatrix} \frac{p_{11}(1-p_{11})}{\delta_1\gamma_1}) & & & & \\ & \frac{p_{01}(1-p_{01})}{(1-\delta_1)\gamma_1} & & & \\ & & \ddots & & \\ & & & \frac{p_{1K}(1-p_{1K})}{\delta_K\gamma_K} & \\ & & & & \frac{p_{0K}(1-p_{0K})}{(1-\delta_K)\gamma_K} \end{pmatrix}\right).$$

A.2 MLE results

For MLE, under effect homogeneity the standard likelihood inference theory [69] gives

$$\sqrt{N}\left(\begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_k \\ \hat{\psi} \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \\ \psi \end{pmatrix}\right) \rightarrow_d N_K(0, J^{-1}), \tag{A.1}$$

where J is the Fisher information matrix

$$\begin{aligned} J &\equiv J(\boldsymbol{\alpha}, \psi) \\ &= \begin{bmatrix} u_1 + v_1 & 0 & ... & 0 & u_1 \\ 0 & u_2 + v_2 & ... & 0 & u_2 \\ \vdots & \vdots & & \vdots & \vdots \\ u_1 & u_2 & ... & u_K & \sum_k u_k \end{bmatrix} \end{aligned} \tag{A.2}$$

and

$$u_j = -\gamma_j \delta_j \frac{exp(\alpha_j + \psi)}{[1 + \exp(\alpha_j + \psi)]^2}$$
$$v_j = -\gamma_j \bar{\delta}_k \frac{exp(\alpha_j)}{[1 + \exp(\alpha_j)]^2}.$$

We denote the plug-in Fisher information by $\hat{J} \equiv J(\hat{\boldsymbol{\alpha}}, \hat{\psi})$. Assuming homogeneity, the estimator of variance of $\hat{\psi}$, $\hat{var}(\hat{\psi})$, is given by the entry in $\hat{J}^{-1}$ corresponding to ψ divided by $\sqrt{N}$.

Under heterogeneity, we proceed as follows:

If we wrongly assume effect homogeneity, i.e. $\psi_1 = ... = \psi_K = \psi$, we have $\text{logit}(p_{0k}) = \alpha_k$ and $\text{logit}(p_{1k}) = \alpha_k + \psi$ for $k = 1, ..., K$, where $\text{logit}(x) = \log(\frac{x}{1-x})$ and $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$ The mis-specified likelihood is given by

$$R(X; \boldsymbol{\alpha}, \beta) = \prod_{k=1}^{K} \begin{pmatrix} m_{1k} \\ a_k \end{pmatrix} \begin{pmatrix} m_{0k} \\ b_k \end{pmatrix} [\text{expit}(\alpha_k + \psi)]^{a_k} [1 - \text{expit}(\alpha_k + \psi)]^{c_k}$$
$$[\text{expit}(\alpha_k)]^{b_k} [1 - \text{expit}(\alpha_k)]^{d_k}$$

The log-likelihood is

$$\rho(X; \boldsymbol{\alpha}, \psi) = \sum_{k=1}^{K} (\alpha_k(a_k + b_k) + \psi a_k - m_{1k} \log(1 + \exp(\alpha_k + \psi)) - m_{0k} \log(1 + \exp(\alpha_k)))$$

The score equations are given by

$$\frac{\partial}{\partial \alpha_k}\rho(X;\boldsymbol{\alpha},\psi) = a_k + b_k - m_{1k}\text{expit}(\alpha_k+\psi) - m_{0k}\text{expit}(\alpha_k) = 0$$
$$\frac{\partial}{\partial \psi}\rho(X;\boldsymbol{\alpha},\psi) = \sum_{k=1}^{K}\left[a_k - m_{1k}\text{expit}(\alpha_k+\psi)\right] = 0$$

The second-order derivatives of $\rho(X;\boldsymbol{\alpha},\psi)$ are

$$\frac{\partial^2}{\partial \alpha_i \partial \alpha_j}\rho(X;\boldsymbol{\alpha},\psi) = 0$$
$$\frac{\partial^2}{\partial \alpha_k \partial \psi}\rho(X;\boldsymbol{\alpha},\psi) = -m_{1k}\frac{\exp(\alpha_k+\psi)}{\left[1+\exp(\alpha_k+\psi)\right]^2}$$
$$\frac{\partial^2}{\partial \alpha_k^2}\rho(X;\boldsymbol{\alpha},\psi) = -m_{1k}\frac{\exp(\alpha_k+\psi)}{\left[1+\exp(\alpha_k+\psi)\right]^2} - m_{0k}\frac{\exp(\alpha_k)}{\left[1+\exp(\alpha_k)\right]^2}$$
$$\frac{\partial^2}{\partial \psi^2}\rho(X;\boldsymbol{\alpha},\psi) = -\sum_k m_{1k}\frac{\exp(\alpha_k+\psi)}{\left[1+\exp(\alpha_k+\psi)\right]^2}$$

We expect the true value of $(\boldsymbol{\alpha},\psi)$ maximizes

$$E\left[\frac{1}{N}\rho(X;\boldsymbol{\alpha},\psi)\right] = \sum_k \gamma_k\{\alpha_k\left(\delta_k\text{expit}(\eta_k+\psi_k) + (1-\delta_k)\text{expit}(\eta_k)\right) +$$
$$\psi\delta_k\text{expit}(\eta_k+\psi_k) - \delta_k\log[1+\exp(\alpha_k+\psi)] -$$
$$(1-\delta_k)\log[1+\exp(\alpha_k)]\}$$

and that $(\hat{\boldsymbol{\alpha}},\hat{\psi})$ maximizes $\frac{1}{N}\rho(X;\boldsymbol{\alpha},\psi)$.

Let

$$\Psi(X;\boldsymbol{\alpha},\psi) \equiv \nabla_{\boldsymbol{\alpha},\psi}\left(\frac{1}{N}\rho(X;\boldsymbol{\alpha},\psi)\right)$$
$$= \frac{1}{N}\begin{bmatrix} a_1 + b_1 - m_{11}\text{expit}(\alpha_1+\psi) - m_{01}\text{expit}(\alpha_1) \\ \vdots \\ a_K + b_K - m_{1K}\text{expit}(\alpha_K+\psi) - m_{0K}\text{expit}(\alpha_K) \\ \sum_k a_k - \sum_k m_{1k}\text{expit}(\alpha_k+\psi) \end{bmatrix}$$

We expect that

$$
\begin{aligned}
0 &= \Psi(X; \hat{\boldsymbol{\alpha}}, \hat{\psi}) \\
&= \frac{1}{N} \begin{bmatrix} a_1 + b_1 - m_{11}\text{expit}(\hat{\alpha}_1 + \hat{\psi}) - m_{01}\text{expit}(\hat{\alpha}_1) \\ \vdots \\ a_K + b_K - m_{1K}\text{expit}(\hat{\alpha}_K + \hat{\psi}) - m_{0K}\text{expit}(\hat{\alpha}_K) \\ \sum_k a_k - \sum_k m_{1k}\text{expit}(\hat{\alpha}_k + \hat{\psi}) \end{bmatrix}
\end{aligned}
$$

By Taylor expansion of $\Psi(X; \hat{\boldsymbol{\alpha}}, \hat{\psi})$ around $(\boldsymbol{\alpha}, \psi)$,

$$
\begin{aligned}
0 &= \Psi(X; \hat{\boldsymbol{\alpha}}, \hat{\psi}) \\
&= \Psi(X; \boldsymbol{\alpha}, \psi) + \dot{\Psi}(X; \boldsymbol{\alpha}^*, \psi^*)(\begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_K \\ \hat{\psi} \end{bmatrix} - \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \\ \psi \end{bmatrix})
\end{aligned}
$$

where $(\boldsymbol{\alpha}^*, \psi^*)$ is between $(\hat{\boldsymbol{\alpha}}, \hat{\psi})$ and $\dot{\Psi}(X; \boldsymbol{\alpha}^*, \psi^*)$ is the Hessian matrix of $\rho(X; \boldsymbol{\alpha}, \psi)$ evaluated at $(\boldsymbol{\alpha}^*, \psi^*)$.

We have

$$
\begin{aligned}
\sqrt{N}\Psi(X; \boldsymbol{\alpha}, \psi) &= \frac{1}{\sqrt{N}} \begin{bmatrix} a_1 + b_1 - m_{11}\text{expit}(\alpha_1 + \psi) - m_{01}\text{expit}(\alpha_1) \\ \vdots \\ a_K + b_K - m_{1K}\text{expit}(\alpha_K + \psi) - m_{0K}\text{expit}(\alpha_K) \\ \sum_k a_k - \sum_k m_{1k}\text{expit}(\alpha_k + \psi) \end{bmatrix} \\
&\to_d N(0, U)
\end{aligned}
$$

where

$$
U = \begin{bmatrix} s_1 + t_1 & 0 & ... & 0 & s_1 \\ 0 & s_2 + t_2 & ... & 0 & s_2 \\ \vdots & \vdots & & \vdots & \vdots \\ s_1 & s_2 & ... & s_K & \sum_k s_k \end{bmatrix}
$$

and

$$s_k = \gamma_k \delta_k p_{1k}(1 - p_{1k})$$
$$t_k = \gamma_k (1 - \delta_k) p_{0k}(1 - p_{0k})$$

for $k = 1, ..., K$.

Furthermore,

$$\dot{\Psi}(X; \boldsymbol{\alpha}, \psi) = \frac{1}{N} \begin{bmatrix} \frac{\partial^2}{\partial \alpha_1^2}\rho(X; \boldsymbol{\alpha}, \psi) & 0 & ... & 0 & \frac{\partial^2}{\partial \alpha_1 \partial \psi}\rho(X; \boldsymbol{\alpha}, \psi) \\ 0 & \frac{\partial^2}{\partial \alpha_2^2}\rho(X; \boldsymbol{\alpha}, \psi) & ... & 0 & \frac{\partial^2}{\partial \alpha_2 \partial \psi}\rho(X; \boldsymbol{\alpha}, \psi) \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{\partial^2}{\partial \alpha_1 \partial \psi}\rho(X; \boldsymbol{\alpha}, \psi) & \frac{\partial^2}{\partial \alpha_2 \partial \psi}\rho(X; \boldsymbol{\alpha}, \psi) & ... & \frac{\partial^2}{\partial \alpha_K \partial \psi}\rho(X; \boldsymbol{\alpha}, \psi) & \frac{\partial^2}{\partial \psi^2}\rho(X; \boldsymbol{\alpha}, \psi) \end{bmatrix}$$

$$\rightarrow_d \begin{bmatrix} u_1 + v_1 & 0 & ... & 0 & u_1 \\ 0 & u_2 + v_2 & ... & 0 & u_2 \\ \vdots & \vdots & & \vdots & \vdots \\ u_1 & u_2 & ... & u_K & \sum_k u_k \end{bmatrix}$$

$$\equiv J \equiv J(\boldsymbol{\alpha}, \psi)$$

where

$$u_k = -\gamma_k \delta_k \frac{\exp(\alpha_k + \psi)}{[1 + \exp(\alpha_k + \psi)]^2}$$
$$v_k = -\gamma_k (1 - \delta_k) \frac{\exp(\alpha_k)}{[1 + \exp(\alpha_k)]^2}$$

for $k = 1, ..., K$. We conclude that

$$\sqrt{N}\left(\begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_k \\ \hat{\psi} \end{bmatrix} - \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_k \\ \psi \end{bmatrix} \right) \rightarrow_d N_K(0, J^{-1} U J^{-1}) \tag{A.3}$$

Appendix B

META-ANALYSIS OF SCLEROTHERAPY STUDIES

Table B.1 shows the data of a meta-analysis of 19 sclerotherapy studies.[46]

Table B.1: Meta-analysis of 19 studies on the effectiveness of sclerotherapy in preventing first bleeding in cirrhosis[46]. Data are shown as number of first bleeding cases/sample sizes.

Study	Sclerotherapy	Control
Paquet et al.	3/35	22/36
Witzel et al.	5/56	30/53
Koch et al.	5/16	6/18
Kobe et al.	3/23	9/22
Wordehoff et al.	11/49	31/46
Santangelo et al.	18/53	9/60
Sauerbruch et al.	17/53	26/60
Piai et al.	10/71	29/69
Potzi et al.	12/41	14/41
Russo et al.	0/21	3/20
Andreani et al.	9/42	13/41
Triger et al.	13/33	14/35
Gregory et al.	31/143	23/138
NIEC	20/55	19/51
PROVA	13/73	13/72
Saggioro et al.	3/13	12/16
Fleig et al.	3/21	5/28
Strauss et al.	4/18	0/19
Planas et al.	6/22	2/24

Appendix C

ZERO-CELL CORRECTION

If study j with four cell counts a_j, b_j, c_j and d_j contains at least one zero cell, we replace the cell counts with a'_j, b'_j, c'_j and d'_j prior to computing the approximate variances, where

$$a'_j = a_j + k_T, \qquad b'_j = b_j + k_C,$$
$$c'_j = c_j + k_T, \qquad d'_j = d_j + k_C.$$

The added numbers k_T and k_C satisfy

$$k_T/k_C = m_{1j}/m_{0j}, \qquad k_T + k_C = .01$$

We choose $k_T + k_C$ to be small so that the added number would only lead to a small bias of the estimator.

Appendix D

PRESENTATION OF JAMES-STEIN ESTIMATOR, LASSO AND RIDGE ESTIMATOR AND AS PRECISION-WEIGHTED LEAST-SQUARE ESTIMATORS

The LASSO is

$$\begin{aligned}\hat{\beta}_{LASSO} &= \underset{\beta}{\operatorname{argmin}} \|\boldsymbol{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \\ &= \underset{\beta}{\operatorname{argmin}} (\hat{\beta}_{LS} - \beta)^T \mathbf{X}^T \mathbf{X} (\hat{\beta}_{LS} - \beta) + \lambda\|\beta\|_1 \\ &= \underset{\beta}{\operatorname{argmin}} \hat{\sigma}^2 (\hat{\beta}_{LS} - \beta)^T \hat{\Sigma}_0^{-1} (\hat{\beta}_{LS} - \beta) + \lambda\|\beta\|_1 \\ &= \underset{\beta}{\operatorname{argmin}} (\hat{\beta}_{LS} - \beta)^T \hat{\Sigma}_0^{-1} (\hat{\beta}_{LS} - \beta) + \lambda'\|\beta\|_1\end{aligned}$$

where $\lambda' = \lambda/\hat{\sigma}^2$ and $\hat{\Sigma}_0 = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$. The same calculation gives

$$\begin{aligned}\hat{\beta}_{Ridge} &= \underset{\beta}{\operatorname{argmin}} \|\boldsymbol{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 \\ &= \underset{\beta}{\operatorname{argmin}} (\hat{\beta}_{LS} - \beta)^T \hat{\Sigma}_0^{-1} (\hat{\beta}_{LS} - \beta) + \lambda'\|\beta\|_2^2.\end{aligned}$$

Finally, the James-Stein estimator is

$$\hat{\beta}_{JS} = \left[1 - \frac{(n-p)(p-2)\hat{\sigma}^2}{(n-p+2)\hat{\beta}_{LS}^T \mathbf{X}^T \mathbf{X} \hat{\beta}_{LS}}\right] \hat{\beta}_{LS}.$$

On the other hand, the solution for the optimization problem

$$\min_{\beta} (\hat{\beta}_{LS} - \beta)^T \hat{\Sigma}_0^{-1} (\hat{\beta}_{LS} - \beta) + \frac{\lambda}{n}\|\mathbf{X}\beta\|_2^2 \tag{D.1}$$

is

$$\left(\frac{\lambda}{n}\hat{\Sigma}_n \mathbf{X}^T\mathbf{X} + I_p\right)^{-1} \hat{\beta}_{LS} = \frac{n}{\lambda\hat{\sigma}^2 + n}\hat{\beta}_{LS}. \tag{D.2}$$

The James-Stein estimator is recovered by letting

$$\frac{n}{\lambda\hat{\sigma}^2 + n} = 1 - \frac{(n-p)(p-2)\hat{\sigma}^2}{(n-p+2)\hat{\beta}_{LS}^T \mathbf{X}^T \mathbf{X} \hat{\beta}_{LS}}$$

in (D.2).

Appendix E

PROOFS OF THEOREMS ABOUT THE LARGE-SAMPLE PROPERTIES OF THE PRECISION-WEIGHTED LEAST SQUARE ESTIMATORS

E.1 Proof of Theorem 3.5.1

Proof. Define

$$Z_0(\beta) = (\beta - \beta_0)^T D(\beta - \beta_0) + \lambda_0 \beta^T C \beta$$

and

$$Z_j(\beta) = (\beta - \beta_0)^T D(\beta - \beta_0) + \lambda_0 \|\beta\|_j^j, \quad j = 1, 2.$$

For brevity, we write $\hat{\beta}_n^j = \text{argmin}_\beta Z_{nj}(\beta)$, $j = 0, 1, 2$, where

$$Z_{n0} = \frac{1}{n}(Y - \mathbf{X}\beta)^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T(Y - \mathbf{X}\beta) + \frac{\lambda_n}{n}\beta^T \left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\beta$$

and

$$Z_{nj} = \frac{1}{n}(Y - \mathbf{X}\beta)^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T(Y - \mathbf{X}\beta) + \frac{\lambda_n}{n}\|\beta\|_j^j, \quad j = 1, 2.$$

Then $\hat{\beta}_n^0 = \hat{\beta}_{rJS,\lambda_n}$ $\hat{\beta}_n^1 = \hat{\beta}_{rLASSO,\lambda_n}$ and $\hat{\beta}_n^2 = \hat{\beta}_{rRidge,\lambda_n}$.

Now we have

$$\begin{aligned} Z_{n0}(\beta) - Z_0(\beta) = (\beta_0 - \beta)^T [&\frac{1}{n}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} - D](\beta_0 - \beta) + \\ &\frac{1}{n}\epsilon^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\epsilon + \frac{2}{n}(\beta_0 - \beta)^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\epsilon + \\ &\frac{\lambda_n}{n}\beta^T(C_n - C)\beta + \left(\frac{\lambda_n}{n} - \lambda_0\right)\beta^T C\beta \\ \xrightarrow{p} 0 \end{aligned}$$

$$\begin{aligned} Z_{nj}(\beta) - Z_j(\beta) = (\beta_0 - \beta)^T [&\frac{1}{n}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} - D](\beta_0 - \beta) + \\ &\frac{1}{n}\epsilon^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\epsilon + \frac{2}{n}(\beta_0 - \beta)^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\epsilon + \\ &(\frac{\lambda_n}{n} - \lambda_0)\|\beta\|_j^j \\ \xrightarrow{p} 0, \quad j = 1, 2 \end{aligned}$$

By Convexity Lemma in [53], we have $\sup_{\beta \in K} |Z_{nj}(\beta) - Z(\beta)| \to_p 0$, $j = 1, 2$, for every open

and convex set K and $\hat{\beta}_n^j = O_p(1)$. Therefore $\hat{\beta}_n^j = \text{argmin}(Z_{nj}) \to_p \text{argmin}(Z_j)$. □

E.2 Proof of Theorem 3.5.2

Proof. Define

$$V_{n0}(u) = (\epsilon - \mathbf{X}u/\sqrt{n})^T\mathbf{X}(\mathbf{X}^T\text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T(\epsilon - \mathbf{X}u/\sqrt{n}) - \epsilon^T\mathbf{X}(\mathbf{X}^T\text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\epsilon + \lambda_n\left[(\beta_0 + u/\sqrt{n})^T C_n(\beta_0 + u/\sqrt{n}) - \beta_0^T C_n\beta_0\right]$$

and

$$V_{nj}(u) = (\epsilon - \mathbf{X}u/\sqrt{n})^T\mathbf{X}(\mathbf{X}^T\text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T(\epsilon - \mathbf{X}u/\sqrt{n}) - \epsilon^T\mathbf{X}(\mathbf{X}^T\text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\epsilon + \lambda_n\sum_{k=1}^{p}[|\beta_{0k} + u_k/\sqrt{n}|^j - |\beta_{0k}|^j], \qquad j = 1, 2.$$

Note that

$$\text{argmin}_{u\in\mathbb{R}^p}V_{n0}(u) = \sqrt{n}(\hat{\beta}_{rJS,\lambda_n} - \beta_0),$$
$$\text{argmin}_{u\in\mathbb{R}^p}V_{n1}(u) = \sqrt{n}(\hat{\beta}_{rLASSO,\lambda_n} - \beta_0),$$

and

$$\text{argmin}_{u\in\mathbb{R}^p}V_{n2}(u) = \sqrt{n}(\hat{\beta}_{rRidge,\lambda_n} - \beta_0).$$

Some calculation leads to

$$V_{n0}(u) = -2u^T\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\left[\frac{\mathbf{X}^T\text{diag}(\mathbf{e}^2)\mathbf{X}}{n}\right]^{-1}\cdot\frac{1}{\sqrt{n}}\mathbf{X}^T\epsilon + u^T\left(\frac{\mathbf{X}^T\mathbf{X}[\mathbf{X}^T\text{diag}(\mathbf{e}^2)\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{X}}{n}\right)u + \lambda_n\left[(\beta_0 + u/\sqrt{n})^T C_n(\beta_0 + u/\sqrt{n}) - \beta_0^T C_n\beta_0\right]$$

and

$$V_{nj}(u) = -2u^T\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\left[\frac{\mathbf{X}^T\text{diag}(\mathbf{e}^2)\mathbf{X}}{n}\right]^{-1}\cdot\frac{1}{\sqrt{n}}\mathbf{X}^T\epsilon + u^T\left(\frac{\mathbf{X}^T\mathbf{X}[\mathbf{X}^T\text{diag}(\mathbf{e}^2)\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{X}}{n}\right)u + \lambda_n\sum_{k=1}^{p}(|\beta_{0k} + u_k/\sqrt{n}|^j - |\beta_{0k}|^j), \qquad j = 1, 2.$$

By CLT, we have $\frac{1}{\sqrt{n}}\mathbf{X}^T\epsilon \to_d N(0, E(\sigma(\mathbf{x}_i)^2\mathbf{x}_i\mathbf{x}_i^T)^{-1})$. Since we also have

$$\lambda_n\left[(\beta_0 + u/\sqrt{n})^T C_n(\beta_0 + u/\sqrt{n}) - \beta_0^T C_n\beta_0\right] = \frac{\lambda_n}{\sqrt{n}}\left(2\beta_0^T C_n u + u^T C_n u/\sqrt{n}\right) \xrightarrow{p} 2\lambda_0\beta_0^T C u,$$

$$\lambda_n \sum_{k=1}^{p}[|\beta_{0k} + u_k/\sqrt{n}| - |\beta_{0k}|] = \frac{\lambda_n}{\sqrt{n}} \sum_{k=1}^{p}[|\sqrt{n}\beta_{0k} + u_k| - |\sqrt{n}\beta_{0k}|]$$
$$\to \lambda_0 \sum_{k=1}^{p}[u_k \text{sign}(\beta_{0k}) I(\beta_{0k} \neq 0) + |u_k| I(\beta_k = 0)],$$

and

$$\lambda_n \sum_{k=1}^{p}[|\beta_{0k} + u_k/\sqrt{n}|^2 - |\beta_{0k}|^2] = \frac{\lambda_n}{\sqrt{n}} \sum_{k=1}^{p}(2\beta_{0k}u_k + \frac{u_k^2}{\sqrt{n}}) \to 2\lambda_0 \beta_0^T u.$$

we have $V_{nj}(u) \to_d V_j(u)$, $j = 0, 1, 2$. Since all V_{nj}'s are convex and V_j's have a unique minimum, we conclude that

$$\text{argmin}_{u \in \mathbb{R}^p} V_{nj}(u) \to_d \text{argmin}_{u \in \mathbb{R}^p} V_j(u), \qquad j = 0, 1, 2.$$

□

E.3 Proof of Theorem 3.6.1

Lemma 1. *Assume the generalized irrepresentable condition holds with a constant vector* $\eta > 0$ *then* $P(\hat{\beta}_{rLASSO,\lambda_n} =_s \beta_0) \geq P(A_n \cap B_n \cap S_n \cap R_n)$ *for*

$$A_n = \{|(D_{11}^n)^{-1} W_S| < \sqrt{n}\left(|\beta_{0S}| - \frac{\lambda_n}{2n}|(D_{11}^n)^{-1}\text{sign}(\beta_{0S})|\right)\},$$

$$B_n = \{D_{21}^n (D_{11}^n)^{-1} W_S - W_{S^C} \leq \frac{\lambda_n}{2\sqrt{n}}\eta\},$$

$$S_n = \{D_n \textit{ is positive definite}\},$$

$$R_n = \{|D_{21}^n (D_{11}^n)^{-1}\text{sign}(\beta_{0S}) - D_{21}(D_{11})^{-1}\text{sign}(\beta_{0S})| \leq 1 - \eta - |D_{21}(D_{11})^{-1}\text{sign}(\beta_{0S})|\}$$

where $W = \frac{1}{\sqrt{n}}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\epsilon$.

Proof. Write $\hat{u} = \hat{\beta}_{rLASSO,\lambda_n} - \beta_0$ and define

$$V_n(u) = (\epsilon - Xu)^T X(X^T\text{diag}(e^2)X)^{-1}X^T(\epsilon - Xu) - \epsilon^T X[X^T\text{diag}(e^2)X]^{-1}X^T\epsilon + \lambda_n\|u + \beta_0\|_1.$$

We have $\hat{u} = \text{argmin}_{u \in \mathbb{R}^p} V_n(u)$.

Now we want to find a set of sufficient conditions for $\hat{\beta}_{rLASSO,\lambda_n} =_S \beta$. Such conditions can be: $|\hat{u}_S| < |\beta_S|$ (here the inequality holds elementwise) and $\hat{u}_{S^C} = \beta_{S^C} = 0$. When D_n is positive definite, $\hat{u} = (\hat{u}_S, 0)$ is the unique minimizer of V_n and satisfies the zero-subgradient condition:

$$D_{11}^n(\sqrt{n}\hat{u}_S) - W_S = -\frac{\lambda_n}{2\sqrt{n}}\text{sign}(\beta_{0S})$$

$$-\frac{\lambda_n}{2\sqrt{n}}\mathbf{1} \leq D_{21}^n\sqrt{n}\hat{u}_S - W_{S^C} \leq \frac{\lambda_n}{2\sqrt{n}}\mathbf{1}.$$

With the extra condition that $|\hat{u}_S| < |\beta_S|$, the above conditions require

$$|\frac{1}{\sqrt{n}}(D_{11}^n)^{-1}W_S - \frac{\lambda_n}{2\sqrt{n}}(D_{11}^n)^{-1}\text{sign}(\beta_S)| < |\beta_S| \tag{E.1}$$

and

$$-\frac{\lambda_n}{2\sqrt{n}}(\mathbf{1}-D_{21}^n(D_1 1^n)^{-1}\text{sign}(\beta_S)) \leq D_{21}^n(D_{11}^n)^{-1}W_S - W_{S^C} \leq \frac{\lambda_n}{2\sqrt{n}}(\mathbf{1}+D_{21}^n(D_{11}^n)^{-1}\text{sign}(\beta_S)) \tag{E.2}$$

Note that when $\{|D_{21}^n(D_{11}^n)^{-1}\text{sign}(\beta_{0S})| < 1-\eta\}$, B_n implies

$$|D_{21}^n(D_{11}^n)^{-1}W_S - W_{S^C}| \leq \frac{\lambda_n}{2\sqrt{n}}(\mathbf{1} - |D_{21}^n(D_{11}^n)^{-1}\text{sign}(\beta_{0S}))|$$

which further implies (E.2). The event $\{|D_{21}^n(D_{11}^n)^{-1}\text{sign}(\beta_{0S})| < 1-\eta\}$ is a result of R_n.

Also A_n implies (E.1). The conclusion follows.

□

Below we prove Theorem 3.6.1:

Proof. By Lemma 1 we have

$$P(\hat{\beta}_{rLASSO,\lambda_n} = \beta_0) \geq P(A_n \cap B_n \cap S_n \cap R_n)$$

whereas

$$\begin{aligned}
&1 - P(A_n \cap B_n \cap S_n \cap R_n) \leq P(A_n^C) + P(B_n^C) + P(S_n^C) + P(R_n^C) \\
\leq &\sum_{i=1}^{q} P(|z_i| \geq \sqrt{n}(|\beta_{0i}| - \frac{\lambda_n}{2n}b_i)) + \sum_{i=1}^{p-q} P(|\xi_i| \geq \frac{\lambda_n}{2\sqrt{n}}\eta_i) + \\
&P(D_n \text{ is not positive definite}) + \sum_{i=1}^{p-q} P(|\kappa_i| \geq d_i)
\end{aligned}$$

where $z = (z_1, z_2, \ldots, z_q)^T = (D_{11}^n)^{-1}W_S$, $\xi = (\xi_1, \ldots, \xi_{p-q})^T = D_{21}^n(D_{11}^n)^{-1}W_S - W_{S^C}$, $\kappa = (\kappa_1, \ldots, \kappa_{p-q})^T = |D_{21}^n(D_{11}^n)^{-1}\text{sign}(\beta_{0S}) - D_{21}(D_{11})^{-1}\text{sign}(\beta_{0S})|$, $b = (b_1, \ldots, b_q) = (D_{11}^n)^{-1}\text{sign}(\beta_S)$, $d = (d_1, \ldots, d_{p-q})^T = \mathbf{1} - \eta - |D_{21}(D_{11})^{-1}\text{sign}(\beta_{0S})|$.

By standard large sample theories we have $(D_{11}^n)^{-1}W_S \to_d N(0, D_{11}^{-1})$ and $D_{21}^n(D_{11}^n)^{-1}W_S - W_{S^C} \to_d N(0, D_{22} - D_{21}D_{11}^{-1}D_{12})$. Hence all the z_i's, ξ_i's and $\sqrt{n}\kappa_i$'s are assymptotically normal random variables.

We thus have

$$P(A_n^C) + P(B_n^C) + P(S_n^C) + P(R_n^C) \to 0.$$

The result follows. □

E.4 Proof of Theorem 3.7.1

Proof. Define

$$Z^*(\beta) = (\beta - \beta^*)^T D^*(\beta - \beta^*) + \lambda_0 \beta^T C \beta$$

and

$$Z_j^*(\beta) = (\beta - \beta^*)^T D^*(\beta - \beta^*) + \lambda_0 \|\beta\|_j^j, \quad j = 1, 2.$$

We also write $\hat{\beta}_n^j = \text{argmin}_\beta Z_{nj}^*(\beta)$, $j = 0, 1, 2$, where

$$Z_{n0}^* = \frac{1}{n}(Y - \mathbf{X}\beta)^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T(Y - \mathbf{X}\beta) + \frac{\lambda_n}{n}\beta^T \left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\beta$$

and

$$Z_{nj}^* = \frac{1}{n}(Y - \mathbf{X}\beta)^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T(Y - \mathbf{X}\beta) + \frac{\lambda_n}{n}\|\beta\|_j^j, \quad j = 1, 2.$$

Then $\hat{\beta}_n^0 = \hat{\beta}_{rJS,\lambda_n}$ $\hat{\beta}_n^1 = \hat{\beta}_{rLASSO,\lambda_n}$ and $\hat{\beta}_n^2 = \hat{\beta}_{rRidge,\lambda_n}$.

Using the same arguments in Appendix E.1, we can show that $\hat{\beta}_n^j = \underset{\beta}{\text{argmin}}\, Z_{nj}^*(\beta) \xrightarrow{p}$ $\underset{\beta}{\text{argmin}}\, Z_j^*(\beta)$. □

E.5 Proof of Theorem 3.7.2

Proof. Define

$$V_{n0}^*(u) = (\epsilon - \mathbf{X}u/\sqrt{n})^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T(\epsilon - \mathbf{X}u/\sqrt{n}) - \epsilon^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\epsilon +$$
$$\lambda_n \left[(\beta_0 + u/\sqrt{n})^T C_n(\beta_0 + u/\sqrt{n}) - \beta_0^T C_n \beta_0\right]$$

and

$$V_{nj}^*(u) = (\epsilon - \mathbf{X}u/\sqrt{n})^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T(\epsilon - \mathbf{X}u/\sqrt{n}) - \epsilon^T \mathbf{X}(\mathbf{X}^T \text{diag}(\mathbf{e}^2)\mathbf{X})^{-1}\mathbf{X}^T\epsilon +$$
$$\lambda_n \sum_{k=1}^{p} [|\beta_{0k} + u_k/\sqrt{n}|^j - |\beta_{0k}|^j], \qquad j = 1, 2.$$

Note that

$$\text{argmin}_{u \in \mathbb{R}^p} V_{n0}^*(u) = \sqrt{n}(\hat{\beta}_{rJS,\lambda_n} - \beta_0),$$
$$\text{argmin}_{u \in \mathbb{R}^p} V_{n1}^*(u) = \sqrt{n}(\hat{\beta}_{rLASSO,\lambda_n} - \beta_0),$$

and

$$\text{argmin}_{u \in \mathbb{R}^p} V_{n2}^*(u) = \sqrt{n}(\hat{\beta}_{rRidge,\lambda_n} - \beta_0).$$

Using the same arguments as in Appendix 3.7.2, we can conclude that

$$\text{argmin}_{u \in \mathbb{R}^p} V_{nj}^*(u) \to_d \text{argmin}_{u \in \mathbb{R}^p} V_j^*(u), \qquad j = 0, 1, 2.$$

□

Appendix F

BAYES RULES FOR THE INFERENCE LOSS FUNCTION

Theorem F.0.1. *The Bayes rule with respect to the inference loss in (4.1) is* $\hat{d} = E_{\Pi_n}(\vartheta|\boldsymbol{Z_n})$ *and* $\hat{\Sigma} = Cov_{\Pi_n}[\vartheta|\boldsymbol{Z_n}]$*, which gives the minimized posterior risk* $E_{\Pi_n}\left[L(\vartheta, \hat{d}, \hat{\Sigma})|\boldsymbol{X_n}\right] = \log|Cov_{\Pi_n}[\vartheta|\boldsymbol{Z_n}]| + p$.

Proof. The posterior risk is

$$E_{\Pi_n}[L|\boldsymbol{Z_n}] = \log|\Sigma| + E_{\Pi_n}\left[(\vartheta - d)^T\Sigma^{-1}(\vartheta - d)|\boldsymbol{Z_n}\right] = \log|\Sigma| + tr\left(\Sigma^{-1}E_{\Pi_n}\left[(\vartheta - d)(\vartheta - d)^T|\boldsymbol{Z_n}\right]\right).$$

Writing $A = \Sigma^{-1}E_{\Pi_n}\left[(\vartheta - d)(\vartheta - d)^T|\boldsymbol{Z_n}\right]$, the posterior risk is

$$\log\left|E_{\Pi_n}\left[(\vartheta - d)(\vartheta - d)^T|\boldsymbol{Z_n}\right]\right| - \log|A| + tr(A)$$

in which the last two terms are respectively the sum of the $-\log$ eigenvalues and the eigenvalues of A. This is minimized by setting A so that all eigenvalues equal to 1, which occurs if and only if A is the identity matrix. Hence $\Sigma = E_{\Pi_n}\left[(\vartheta - d)(\vartheta - d)^T|\boldsymbol{Z_n}\right]$.

To minimize the posterior risk with respect to d, it remains to consider

$$\log|E_{\Pi_n}\left[(\vartheta - d)(\vartheta - d)^T|\boldsymbol{Z_n}\right]| + p = \log|Cov_{\Pi_n}(\vartheta|\boldsymbol{X_n}) + (E_{\Pi_n}[\vartheta|\boldsymbol{Z_n}] - d)(E_{\Pi_n}[\vartheta|\boldsymbol{Z_n}] - d)^T| + p.$$

By matrix determinant lemma, the above expression equals

$$\log|Cov_{\Pi_n}(\vartheta|\boldsymbol{Z_n})| + \log\left(1 + (E_{\Pi_n}[\vartheta|\boldsymbol{Z_n}] - d)^T Cov_{\Pi_n}[\vartheta|\boldsymbol{Z_n}]^{-1}(E_{\Pi_n}[\vartheta|\boldsymbol{Z_n}] - d)\right) + p$$

which is minimized by setting $d = E_{\Pi_n}[\vartheta|\boldsymbol{Z_n}]$, which in turn means setting $\Sigma = Cov_{\Pi_n}[\vartheta|\boldsymbol{Z_n}]$. This rule achieves minimized posterior risk $E_{\Pi_n}[L|\boldsymbol{Z_n}] = \log|Cov_{\Pi_n}[\vartheta|\boldsymbol{Z_n}]| + p$.

□

Appendix G

BAYES RULE FOR THE BALANCED INFERENCE LOSS

Theorem G.0.1. *The Bayes rule with respect to the inference loss in (4.2) is*

$$\hat{d}_n = E_{\Pi_n}(\vartheta|\boldsymbol{Z_n}), \quad \hat{\Omega} = E\left(\frac{1}{n}\sum_{i=1}^{n} \dot{l}_i(\vartheta)\dot{l}_i(\vartheta)^T I_n(\vartheta)^{-1}|\boldsymbol{Z_n}\right), \quad \hat{\Sigma}_n = Cov_{\Pi_n}(\vartheta|\boldsymbol{Z_n})\hat{\Omega}.$$

Proof. The expected posterior loss is

$$E_{\Pi_n}[L_{BI}|\boldsymbol{Z_n}] = \log|\Sigma| + E_{\Pi_n}\left[(\vartheta-d)^T\Omega\Sigma^{-1}(\vartheta-d)\Big|\boldsymbol{Z_n}\right] + E_{\Pi_n}\left[\frac{1}{n}\sum_{i=1}^{n}\dot{l}_i(\vartheta)^T(\Omega I_n(\vartheta))^{-1}\dot{l}_i(\vartheta)\Big|\boldsymbol{Z_n}\right] \tag{G.1}$$

Similar to the proof of Theorem F.0.1, the minimum of (G.1) with respect to Σ with Ω fixed is achieved by setting $\Sigma = E\left\{(\vartheta-d)(\theta-d)^T\Big|\boldsymbol{Z_n}\right\}\Omega$ and $d = E[\vartheta|\boldsymbol{Z_n}]$. Substituting $E\left\{(\vartheta-d)(\theta-d)^T\Big|\boldsymbol{Z_n}\right\}\Omega$ and $\hat{d} = E[\vartheta|\boldsymbol{Z_n}]$ in (G.1), we obtain that the minimal expected posterior loss with Ω fixed is

$$\log|Var_{\Pi_n}[\vartheta|\boldsymbol{Z_n}]| + p + \log|\Omega| + E_{\Pi_n}\left[\frac{1}{n}\sum_{i=1}^{n}\dot{l}_i(\vartheta)^T(\Omega I_n(\vartheta))^{-1}\dot{l}_i(\vartheta)\Big|\boldsymbol{Z_n}\right] \tag{G.2}$$

Again using the same method as in the proof of Theorem F.0.1, the minimum of (G.2) is achieved by setting $\Omega = E\left[\frac{1}{n}\sum_{i=1}^{n}\dot{l}_i(\vartheta)\dot{l}_i(\vartheta)^T I_n^{-1}(\vartheta)\Big|\boldsymbol{Z_n}\right]$, which in turn gives the Bayes rule of Σ:

$$\hat{\Sigma} = Var_{\Pi_n}[\vartheta|\boldsymbol{Z_n}]E\left[\frac{1}{n}\sum_{i=1}^{n}\dot{l}_i(\vartheta)\dot{l}_i(\vartheta)^T I_n^{-1}(\vartheta)\Big|\boldsymbol{Z_n}\right].$$

□

Appendix H

ASYMPTOTIC EQUIVALENCE OF BAYESIAN ROBUST COVARIANCE MATRIX AND THE LARGE-SAMPLE COVARIANCE MATRIX OF THE BAYES POINT ESTIMATE

Let $I(\theta) = -E_0\left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(Z_1)\right]$ be the Fisher information. By the Bernstein von Mises Theorem under model misspecification, $nCov_{\Pi_n}(\vartheta|\boldsymbol{Z_n}) \approx I_{\theta^*}^{-1}$, which $\theta^* = \text{argmin}_\theta E_0\left[\frac{p_0(Z_i)}{p_\theta(Z_i)}\right]$ is the minimal Kullback-Leibler point. We see that

$$\sum_{i=1}^{n} \dot{l}_i(\vartheta)\dot{l}_i(\vartheta)^T I_n(\vartheta)^{-1} \to_{P_0} E_0[\dot{l}_i(\theta)\dot{l}_i(\theta)^T]I(\theta)^{-1}$$

for any $\theta \in \Theta$. By the theorem below, we show that under certain conditions,

$$E_{\Pi_n}\left\{\sum_{i=1}^{n} \dot{l}_i(\vartheta)\dot{l}_i(\vartheta)^T I_n(\vartheta)^{-1}\middle|\boldsymbol{Z_n}\right\} \to_{P_0} E_0[\dot{l}_i(\theta^*)\dot{l}_i(\theta^*)^T]I(\theta^*)^{-1},$$

where the right-hand-side is the asymptotic variance of the Bayes point estimate by [34].

Theorem H.0.1. *Suppose $A_n(\theta)$ is a random function and measurable on $(Z_1, \ldots, Z_n)$. Suppose the conditions in [34] Theorem 2.3 and Corollary 4.2 hold. Further assume the following conditions hold:*

1. *A is continuous in a neighborhood of θ^* and $|A(\theta^*)| < \infty$.*

2. *Uniform convergence of A_n in a neighborhood of θ^*: for any $\epsilon, \gamma > 0$, there exists $\eta_0 > 0$ and positive integer N such that for any $n > N$, we have*

$$P_0(\sup_{\theta_1:\|\theta_1-\theta^*\|<\eta_0} |A_n(\theta_1) - A(\theta_1)| > \epsilon) < \gamma.$$

3. *Asymptotic posterior uniform integrability: for any $\epsilon > 0$, we have*

$$\lim_{M\to\infty} \limsup_{n\to\infty} P_0\left(E_{\Pi_n}\left\{\|A_n(\theta)\| I\left(\|A_n(\theta)\| > M\right)\middle|Z_1, \ldots, Z_n\right\} > \epsilon\right) = 0 \qquad \text{(H.1)}$$

Then $E_{\Pi_n}(A_n(\vartheta)|\boldsymbol{Z_n}) \to_{P_0} A(\theta^*)$.

Proof. **Step 1:**

We show that $\Pi_n(|A_n(\vartheta) - A_n(\theta^*)| > \epsilon | Z_1, \ldots, Z_n) \to_{P_0} 0$ as $n \to \infty$, for any $\epsilon > 0$. In other words, for any ϵ, γ, and $\xi > 0$, there exists a positive integer $N > 0$, such that for any $n > N$, we have

$$P_0\left(\Pi_n\left(\|A_n(\vartheta) - A_n(\theta^*)\| > \epsilon \middle| Z_1, \ldots, Z_n\right) > \xi\right) < 2\gamma. \tag{H.2}$$

Note that the additional conditions 1, 2 imply that for any $\epsilon, \gamma > 0$, there exists $\eta_0 > 0$ and positive integer N_1 such that for any $n > N_1$, we have

$$P_0(\sup_{\theta_1 : \|\theta_1 - \theta^*\| < \eta_0} |A_n(\theta_1) - A_n(\theta^*)| > \epsilon) < \gamma. \tag{H.3}$$

By Corollary 4.2 in [34], there exists a positive integer N_2 such that

$$P_0\left(\Pi_n\left(\|\vartheta - \theta^*\|_1 \geq \eta_0 \middle| Z_1, \ldots, Z_n\right) > \xi\right) < \gamma$$

for any $n > N_2$.

Now that the left-hand-side of (H.2) is

$$\begin{aligned}
&P_0\left(\Pi_n\left(\|A_n(\vartheta) - A_n(\theta^*)\| > \epsilon, \|\vartheta - \theta^*\|_1 < \eta_0 \middle| Z_1, \ldots, Z_n\right) > \xi\right) + \\
&\quad P_0\left(\Pi_n\left(\|A_n(\vartheta) - A_n(\theta^*)\| > \epsilon, \|\vartheta - \theta^*\|_1 \geq \eta_0 \middle| Z_1, \ldots, Z_n\right) > \xi\right) \\
\leq &P_0\left(\sup_{\theta : \|\theta - \theta^*\| < \eta_0} \|A_n(\theta) - A_n(\theta^*)\| > \epsilon, \Pi_n\left(\|\vartheta - \theta^*\|_1 < \eta_0 \middle| Z_1, \ldots, Z_n\right) > \xi\right) + \\
&\quad P_0\left(\Pi_n\left(\|\vartheta - \theta^*\|_1 \geq \eta_0 \middle| Z_1, \ldots, Z_n\right) > \xi\right) \\
\leq &P_0\left(\sup_{\theta : \|\theta - \theta^*\| < \eta_0} \|A_n(\theta) - A_n(\theta^*)\| > \epsilon\right) + P_0\left(\Pi_n\left(\|\vartheta - \theta^*\|_1 \geq \eta_0 \middle| Z_1, \ldots, Z_n\right) > \xi\right) \\
< &2\gamma.
\end{aligned}$$

Step 2: We show that

$$E_{\Pi_n}(A_n(\vartheta) | Z_1, \ldots, Z_n) = A_n(\theta^*) + O_{P_0}(1).$$

Note that the uniform continuity (H.2) also holds for functions $A_n(\cdot) \wedge M$ with $M > 0$. The results of Step 1 also applies to $A_n(\cdot) \wedge M$:

$$\Pi_n(\|A_n(\vartheta) \wedge M - A_n(\theta^*) \wedge M\| > \epsilon | Z_1, \ldots, Z_n) \to_{P_0} 0 \tag{H.4}$$

for any $\epsilon > 0$.

It suffices to show that

$$E_{\Pi_n}(\|A_n(\vartheta) - A_n(\theta^*)\| | Z_1, \ldots, Z_n) \to_{P_0} 0.$$

By triangular inequality we have

$$\begin{aligned} E_{\Pi_n}(\|A_n(\vartheta) - A_n(\theta^*)\||Z_1,\ldots,Z_n) = {} & E_{\Pi_n}(\|A_n(\vartheta) - A_n(\vartheta)\wedge M\||Z_1,\ldots,Z_n)+ \\ & E_{\Pi_n}(\|A_n(\vartheta)\wedge M - A_n(\theta^*)\wedge M\||Z_1,\ldots,Z_n)+ \\ & \|A_n(\theta^*)\wedge M - A_n(\theta^*)\|. \end{aligned}$$

Now on the right-hand side, the third term converges in P_0 to $\|A(\theta^*)\wedge M - A(\theta^*)\|$ as $n\to\infty$ and then converges in P_0 to 0 by letting $M\to\infty$.

The first term satisfies

$$E_{\Pi_n}\left(\|A_n(\vartheta) - A_n(\theta^*)\|\middle|Z_1,\ldots,Z_n\right) \le E_{\Pi_n}\left\{\|A_n(\theta)\|I\left(\|A_n(\theta)\| > M\right)\middle|Z_1,\ldots,Z_n\right\},$$

which converges to zero by letting $n\to\infty$ followed by letting $M\to\infty$.

Finally, the second term is

$$\begin{aligned} & E_{\Pi_n}\left(\|A_n(\vartheta)\wedge M - A_n(\vartheta)\wedge M\|I\left(\|\theta-\theta^*\|\le\eta_0\right)\middle|Z_1,\ldots,Z_n\right)+ \\ & \quad E_{\Pi_n}\left(\|A_n(\vartheta)\wedge M - A_n(\vartheta)\wedge M\|I\left(\|\vartheta-\theta^*\|\le\eta_0\right)\middle|Z_1,\ldots,Z_n\right) \\ \le {} & \sup_{\theta_1:\|\theta_1-\theta^*\|<\eta_0}\|A_n(\theta_1)\wedge M - A_n(\theta^*)\wedge M\|\cdot\Pi_n\left(\|\vartheta-\theta^*\|<\eta_0\middle|Z_1,\ldots,Z_n\right)+ \\ & \quad M\cdot\Pi_n\left(\|\vartheta-\theta^*\|<\eta_0\middle|Z_1,\ldots,Z_n\right)\to_{P_0} 0+0=0 \end{aligned}$$

by letting $n\to\infty$ followed by letting $M\to\infty$. □

VITA

Kendrick Qijun Li was born in Shanghai in 1993 and raised in Shanghai, China. He received a Bachelor degree in Sciences with Specialty in Biological Sciences from Peking University in 2016 with a double major in Applied Mathematics. He entered the Biostatistics Master Thesis program in the University of Washington in September of 2016, and switched to the Doctoral program the next year. Starting in July of 2021, he will be a post-doctoral research fellow at the University of Michigan, Department of Biostatistics.