Sparse Partial Least Squares Methods and Extensions for Modeling Heart-Healthy Diets

Natalie C. Gasca

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

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Robyn McClelland, Chair

Noah Simon

Kathleen Kerr

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Sparse Partial Least Squares Methods and Extensions for Modeling Heart-Healthy Diets

Natalie C. Gasca

Chair of the Supervisory Committee:
Dr. Robyn McClelland
Biostatistics

When investigating the link between diet and cardiovascular disease (CVD), nutritional epidemiologists often use unsupervised methods to construct dietary patterns using hundreds of foods. We posit that diet summaries can be better tailored to CVD by incorporating outcome data and sparsity. Partial least squares (PLS) is an appealing supervised method because its patterns are correlated with a continuous response while also capturing covariate variability. However, its statistical and modeling assumptions are not well characterized for non-continuous outcomes. In this dissertation, we clarify the implications of incorporating PLS into linear and Cox models, with the aim of constructing parsimonious patterns to facilitate hypothesis generation. First, we identify an advantageous sparse PLS procedure (SPLS) that targets variable selection for continuous data. We propose using SPLS after fitting the Cox or approximate Cox model to analyze a right-censored survival outcome. To enable proper adjustment for covariates that do not require dimension reduction, we demonstrate that various scientific premises and goals require different types of adjustment when using PLS. These contributions are verified by simulation studies, analytic results, and applications to CVD-related endpoints. Our findings allow for more informed method selection and deeper insight connecting least squares and PLS regression coefficients.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I've learned that people will forget what you said,
people will forget what you did,
but people will never forget how you made them feel.

— Maya Angelou

There are so many people who have shared their wisdom and encouragement with me, all of whom have enabled me to complete this academic achievement. I will name as many as I can, in deep gratitude, acknowledging that there is only so much space allotted.

Living in a multi-generational and multi-ethnic household in Fallbrook, CA (a large town in Southern California known for avocado and orange groves) provided a nurturing childhood for me. I am grateful for my mom's family's values and enveloping warmth at Thanksgiving gatherings, which have encouraged me to pursue my professional goals with thoughtfulness and poise. I also cherish my dad's family's unwavering love, storytelling prowess, and Christmas Eve traditions of dancing cumbias and playing dominos until the early morning—which have encouraged me to make the most of life's opportunities and enjoy the thrill of adventures. I have been blessed with many years of loyal and laughter-filled friendship with my Fallbrook soccer teammates and classmates to this day, including April Doss, Bethany Weber, Deanna Doss, Jazmin Velásquez, and Mayra Rivera. I have also learned tremendously from my friends' parents, especially Lucy Nehlen, Fran Little, Ana Lopez, Cruz Rosales, and Alejandro Villegas. Thank you for providing such an enriching and pragmatic foundation as I began my journey.

My undergraduate years at Cal Poly Pomona were profound in expanding my worldview and passion for helping others. I made life-long friends in the dorms, especially Adriana

Lewis, Anaís Plácido, Cameron Chiu, Chantall Allgood, John Huynh, Mariah Goldbach, and Nolan Whitener, to whom I am extremely grateful for keeping me grounded and optimistic, even in difficult times. Friends in the National Society of Leadership and Success (like Connie Hua and Mayuri Patel), KME Math Club, and Kellogg Honors College also shaped my time there. A warm shout out to my undergrad STEM friends including Itzia Jerónimo, Tam Nguyen, Vanessa Salvary, and Alberto Soto.

One of the greatest gifts of my undergradate experience was meeting incredible mentors. These include Professors Randy Swift, Amber Rosin, Patricia Hale, Chuck Hale, Jeanne Almaraz, Renford Reese, and Stan Becker. Outside of school, I had the fortune of meeting Saba Hafeez and Dr. Anita Jackson, who taught me valuable lessons about the power of community engagement and leading through intuition and sisterhood, respectively. Volunteering was immensely valuable since it introduced me more viscerally to how much potential different underserved groups have, if afforded the right opportunities. To help determine my career path as a math major, I attended conferences such as SACNAS with workshops for underrepresented minorities, which provided guidance for graduate school, enabled me to intern at Johns Hopkins (DSIP) and Cal Poly Pomona (PUMP), and connected me to phenomenal underrepresented peers and professionals such as Vanessa Rivera Quiñones. Panelists inspired me to combine my enthusiasm for interpreting patterns and improving people's well-being by pursuing a biostatistics PhD directly after my bachelor's degree.

Seattle has been a remarkable place to study and live during graduate school. A special thank you to my local family members who helped me feel at home and welcome in Washington. I deeply enjoyed the Seattle International Film Festival; beautifully-long summer days; gorgeous Olympic, Cascades, and Rainier mountain views; and vibrant coffee shops including Starbucks at Wedgwood, Stage Door Café in Greenwood, Chocolati Café near Green Lake, and Big Time Brewery on the Ave. Thank you to the people involved in maintaining these spaces (including the many tribes of Washington state, such as the Coast Salish people). I

am grateful to have pursued my degree in an environment rich with nature.

One of the things that I treasure most from graduate school is the people I have met. Foremost, Gitana Garofalo has played an invaluable role as a deeply-cherished mentor, encouraging me to be intentional about my lifestyle and to find resources to overcome challenges (even beyond academics). She enabled me to pursue deeply-enriching yet less-traditional endeavors while still fulfilling the academic requirements of my program. Of course, I owe a lot to my wonderful cohortmates for their advice, encouragement, trust, and support: Phuong Vu, Rui Zhuang, Kelsey Grinde, Arjun Sondhi, Brian Williamson, Fan Xia, Xiaowen Tian, Yuxiang Xie, and Chaoyu Yu. I treasure many of the conversations, lessons, and stories shared with department professors including Katie Kerr, Mary Lou Thompson, Nancy Tempkin, and Lyn Brumback. In particular, the advice, motivation, and mentorship that Robyn McClelland offered me were instrumental in my completing this program. She often reminded me that "you definitely belong in biostatistics." Everyone deserves an academic mentor like her. I also thank the (bio)statistics students who I befriended, including but not limited to Ernesto Ulloa, Amarise Little, Aaron Hudson, María Valdez, Taylor Okonek, Kendrick Li, Parker Xie, Nanxun Ma, Katie Wilson, Anu Mishra, Anna Plantinga, Anya Mikhalyova, Brayan Ortiz, Davíd Ezekiel-Herrera, Clara Oromendia, Andrew Spieker, Allison Meisner, Max Schneider, Hannah Director, Daphne Liu, and Anupreet Porwal.

I am also indebted to amazing support systems outside of my department, especially those who are a part of CHSCC, GO-MAP (including Vanessa Álvarez and Carolyn Jackson), the Latino Center for Health, Seattle ARCS, Bonderman Fellows, and WIBS. I am grateful for my incredibly inspiring friends including Cassandra Baker, Jessica B. Hernández, Katrina Claw, Jessica Saniġaq Ullrich, Angela R. Fernández, Dalya Perez and her Zumba dance crew, Alec Calac, and Harkirat Sohi. I have also admired professional role models such as Mónica Feliú-Mójer, Yates Coley, Ruth Etzioni, Regina Nuzzo, and Nayak Polissar. Thank you all for paving the way so that I can pursue my goals in a deeply authentic way.

For instance, one of my most impactful experiences has been the Bonderman fellowship. Traveling solo for eight months proved to me that I am more optimistic and curious about people, food, and culture than overpowered by fear of the unknown. The locals and expats I met were so generous with their time, explaining parts of their culture, inviting me to events, and recommending places to visit. Apart from my friends and wonderful connections of friends, I am honored to have spent time with my Samoan aunties, Milagros Muñiz Luaces, Alba Ortiz, Diana Mendoza Curas, Gisela Pérez Fonseca, and José Davíd Gutiérrez Alipio. The people I interacted with reinvigorated me and reminded me that there are so many ways to live life. When I reflect on this adventure, what I miss most is the excitement of connecting with like-minded people, so many styles of transportation, and the appreciation of community traditions. I miss the orange juice ladies (caseritas) in Bolivia, chai from ceramic cups, coconut lemonade, public dance gatherings in plazas, street food carts, and rides on motorbikes. Acknowledging how adaptable I am and how well this trip complemented my personality gave me more confidence in who I am and was meant to be in this world.

On my return, writing groups through GO-MAP (PhDivas and Get PhinisheD!) and Latinas Completing Doctoral Degrees (Collective Resilient Virtual Library for Mujeres) were instrumental in motivating me to complete my research, particularly throughout the pandemic. I give immense thanks for Arianna Gómez, Juandalyn Burke, Antonio Olivas-Martínez, Juliana Sánchez, Sarah Chávez, and Vanessa Martínez. Other sources of encouragement and inspiration came from groups involved in science communication and outreach, in particular Mexicanos en Estadística y Salud and the Diversity Supplement Workshops with Natasha Ludwig-Barrón. I also benefited greatly from my statistical consulting partnerships with Janeth Sánchez and Max Schneider. I am particularly grateful to everyone who has played a role in my health and medical teams throughout these years. Physical therapy, surgery, pandemics, and life events are not always easy to traverse while in a graduate program.

I have also had the fortune of working with an incredible team of academics on my

dissertation. I greatly appreciate the time and insight provided by my committee members: Robyn McClelland, Katie Kerr, Noah Simon, Mandy Fretts, and Mary Lou Biggs. You have each contributed many lessons in my development as a researcher and collaborative biostatistician. Many thanks are also due to the biostatistics staff (such as Deb Nelson, Lisa Sharamitaro, and Carl Riches) who were instrumental in helping me navigate a variety of administrative tasks. I am honored to have received funding from various sources to complete my graduate studies and research, including GO-MAP, ARCS Foundation, National Science Foundation Graduate Research Fellowship Program (Grant # DGE-1256082), and the National Institutes of Health (National Heart, Lung, and Blood Institute) Diversity Supplement (Contract # 75N92020D00001). A warm thank you to the participants, staff, and investigators of the Multi-Ethnic Study of Atherosclerosis (MESA) for their valuable contributions to this dissertation.

The biggest influences throughout my life have been my family—both immediate and extended—who have served as role models and my first support systems. You have inspired and empowered me to develop into an ambitious, caring, warm, and persevering individual. To my parents and brother (Luis, Caroline, and Sito), muchichísimas gracias for nurturing me in your own ways and instilling in me a fire to do what's right; being a bridge for others to access resources, communicating ideas across disciplines, and respecting what everyone brings to the table. Thank you for your sacrifices, lessons, and values that have allowed me to deep dive into academics and activism with conviction, knowing that you will be with me every step of the way. There have been a few losses (and additions) in my family during my time in graduate school; I believe that we can honor their spirits by pursuing our passions and doing our part to help make the world a more accessible and equitable place to live and thrive. To my husband, Zach, thank you for your flexibility, patience, super-human driving abilities, and unwavering support throughout these many long-distance years. It has all paid off, and I am so excited for our bright and joyful future together!

Hermosura de la dialéctica

Estoy viva
como fruta madura
dueña ya de inviernos y veranos,
abuela de los pájaros,
tejedora del viento navegante.

No se ha educado aún mi corazón
y, niña, tiemblo en los atardeceres,
me deslumbran el verde, las marimbas
y el ruido de la lluvia
hermanándose con mi húmedo vientre,
cuando todo es más suave y luminoso.

Crezco y no aprendo a crecer,
no me desilusiono,
ni me vuelvo mujer envuelta en velos,
descreída de todo, lamentando su suerte.
No. Con cada día, se me nacen los ojos del asombro,
de la tierra parida
el canto de los pueblos,
los brazos del obrero construyendo,
la mujer vendedora con su ramo de hijos,
los chavalos alegres marchando hacia el colegio.

Sí.
Es verdad que a ratos estoy triste
y salgo a los caminos,
suelta como mi pelo,
y lloro por las cosas más dulces y más tiernas
y atesoro recuerdos
brotando entre mis huesos
y soy una infinita espieral que se retuerce
entre lunas y soles,
avanzando en los días,
desenrollando el tiempo
con miedo o desparpajo,
desenvainando estrellas
para subir más alto, más arriba,
dándole caza al aire,
gozándome en el ser que me sustenta,
en la eterna marea de flujos y reflujos
que mueve el universo
y que impulsa los giros redondos de la tierra.

Soy la mujer que piensa.
Algún día
mis ojos
encenderán luciérnagas.

— Gioconda Belli

DEDICATION

In memory of my bold and beloved cousin, Dr. Sarah E. Hawley.

Chapter 1

INTRODUCTION

When seeking to characterize the concurrent impact of foods on cardiovascular disease (CVD) risk, nutritional scientists have encountered a few methodological challenges. Especially when utilizing self-reported diet surveys with more than 100 food items, there tend to be strong correlations between some food measurements, which leads to highly variable and sometimes unintuitive associations. Most nutritional studies have used dimension reduction methods to alleviate this problem, but they have predominantly used unsupervised methods, which do not incorporate disease data when constructing dietary patterns. In this dissertation, we evaluate and extend partial least squares (PLS) methods to summarize nutritional information in a way that is tailored to heart disease while still being in a familiar format for nutritional epidemiologists.

As we posit that that not all the hundreds of food items are particularly relevant to heart health, we focus on incorporating sparsity to identify which foods are most related to a univariate outcome. One way to model this is to construct "interpretable" dietary patterns; we propose that "interpretability" entails selecting all of the relevant foods, discarding irrelevant foods, capturing a few dimensions of diet reflecting CVD etiology, and maintaining predictive ability for the CVD-related response, in descending order of importance. We investigate several statistical challenges and considerations that arise when undertaking this scientific goal. While we motivate and test these methods for the purposes of nutritional data, they can be applied to numerous other disciplines whose aim is to reduce the dimension of their predictors, such as bioinformatics, environmental studies, and -omic fields.

In Chapter 2, we pursue the analysis of a univariate continuous outcome using dimension

reduction methods. There are many such methods that can create dietary components reflecting certain features of the data, including those that incorporate different sparsity aims. We draw connections between and discuss which of these methods would be best suited for the nutritional epidemiology context. Via simulations, we evaluate the interpretability and predictive ability of their constructed patterns, varying levels of covariance within blocks of predictors to resemble food groups. As PLS-based methods offer numerous advantages towards accomplishing our aims, we analytically verify the connection between PLS and the linear Conjugate Gradient technique as well as demonstrate the connection to the Krylov subspace. These insights allow us to pursue extensions to PLS in an informed manner.

To more directly study the relationship between diet and a right-censored survival time, we investigate methods that have used PLS to approximate the Cox proportional hazards model in Chapter 3. After clarifying how PLS can serve as both an optimization and summarizing tool for non-continuous outcomes, we focus on our sparsity goal. Existing methods seek to create sparse components, but we presume that overall sparsity is a more pressing aim when considering hundreds of foods. For these settings, we propose incorporating a Sparse PLS procedure that directly promotes variable selection after fitting the Cox model or approximated Cox model.

When constructing dietary patterns to evaluate heart-healthiness, it is important to adjust for potential confounders and precision variables. Though adjustment is straightforward for standard regression analyses, it is particularly tricky when using PLS on one set of covariates, as we will show in Chapter 4. While there have been a few proposed adjustment approaches for PLS methods, their properties and modeling implications have not been thoroughly studied. We demonstrate that there are different ways to adjust for covariates depending on whether PLS is being used to simply summarize the regression model or if its components are considered to be the true scale of interest. We analytically justify that only a subset of these approaches result in the typical covariate adjustment obtained from a multiple regression analysis. Last, we situate all of these adjustment approaches in a data-integration framework, depending on when the sets of covariates are modeled, and we provide corrections

and extensions to certain approaches. These findings can allow for more informed modeling choices and a better understanding of the connections between least squares regression coefficients and those derived from PLS methods.

Our illustrative examples come from the Multi-Ethnic Study of Atherosclerosis (MESA), an observational cohort study which investigates the prevalence and progression of subclinical CVD. In Chapter 2, we apply dimension reduction methods to create dietary patterns related to body mass index, as excess body weight is an important pathway between nutrition and CVD. In Chapter 4, we construct dietary patterns that are related to the time until first incident CVD event, using more than 17 years of follow-up data. We conclude with the implications of this research and directions for future investigations.

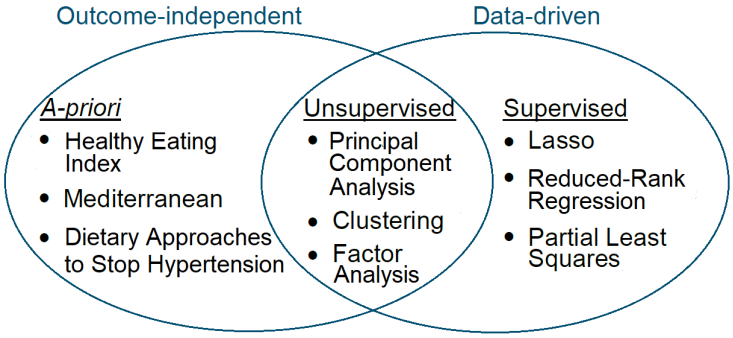
Chapter 2

COMPARISON OF DIMENSION REDUCTION METHODS FOR CREATING HEART-HEALTHY DIETARY PATTERNS

2.1 Introduction

Early investigations of the link between diet and cardiovascular disease (CVD), which has been identified as the leading cause of mortality in American men and women [35], focused on individual nutrients or foods. As research transitioned to studying the concurrent effect of foods, results were often discouraging with frequent unintuitive results due to strong correlations among some food measurements [37]. Attention shifted to dimension reduction methods in order to obtain understandable findings from the most relevant aspects of diet.

Figure 2.1: Classification of dimension reduction methods used to create dietary patterns.



In nutritional epidemiology, a common approach to constructing composite dietary scores is to use previous (*a-priori*) scientific knowledge to assign "heart-healthiness" to a limited number of foods and nutrients (as illustrated in Figure 2.1, adapted from [63]). A few examples include the Dietary Approaches to Stop Hypertension (DASH) diet [66], the Healthy Eating Index [43], and the Mediterranean Diet [70]. However, many of these scores rely on study-specific quantiles to distinguish levels of intake, making it difficult to generalize results

outside of the study population.

Another popular way to construct dietary patterns is to use the information from all food measurements collected in a study, called unsupervised methods. One of the most widespread methods is principal component analysis (PCA), which forms dietary patterns that summarize how people in a study eat overall [58]. Similarly, cluster [24] and factor analysis [12] are examples of unsupervised methods. While these methods describe different aspects of food consumption, their patterns are not inherently predictive of CVD as they are not informed by disease data.

To better characterize links between diet and heart disease, we propose using supervised methods, which incorporate disease data directly into the construction of diet summaries. These techniques have the potential to concisely reduce nutritional information in a way that is tailored to heart disease. Among these techniques, reduced-rank regression (RRR) and canonical-correlation analysis have been the most frequently used in nutritional epidemiology [63]. These methods are applicable when considering a multivariate response (e.g., multiple biomarkers at once), but we focus on a univariate response in this study. One of the most widespread supervised methods across other fields is the Lasso (least absolute shrinkage and selection operator) [68], which is an appealing candidate due to its ability to select variables that are useful for prediction. It accomplishes this property by shrinking its regression coefficients towards zero. It was not until recently that this supervised method was used to explore how well certain foods predicted CVD risk factors [84].

Another supervised method is partial least squares (PLS), which can create dietary patterns that are correlated with a continuous response while also capturing variation in diet composition [81]. Similar to PCA, these patterns are linear combinations of the foods. Although PLS is a standard technique in chemometrics, econometrics and genomics [51], it has only recently begun to gain attention in nutritional epidemiology, particularly in the context of comparing its performance to RRR and PCA when analyzing multivariate responses (see [36, 23, 83]). An appealing extension of PLS is sparse PLS (SPLS) [14], which incorporates variable selection when constructing diet patterns, as we do not presume that every food is

inherently related to heart disease. There has been a similar extension to PCA called sparse PCA (SPCA) [88], which we include in this study to provide a fairer comparison between methods. However there have been other types of sparsity integrated into PLS and PCA (e.g., [40, 44]), so we review their differences to select the one most appropriate for our scientific setting of interest.

Our illustrative example comes from the Multi-Ethnic Study of Atherosclerosis (MESA), an observational cohort study which investigates the prevalence and progression of subclinical CVD [9]. At the baseline exam, 6531 participants reported their consumption of 120 types of foods and beverages during a typical week of the previous year in a food frequency questionnaire (FFQ). We want to extract the most biologically-relevant information to characterize a heart-healthy diet that could be protective against CVD. More specifically, we aim to construct "interpretable" dietary patterns that—in descending order of importance—select all of the relevant foods, discard irrelevant foods, capture a few dimensions of diet, and maintain predictive ability of the CVD-related outcome.

This chapter is organized as follows. In Section 2.2, we review and draw connections between the dimension reduction methods which we will be assessing: PCA, SPCA, PLS, SPLS, and Lasso. Simulation studies are conducted in Section 2.3 to evaluate how the methods perform when blocks of variables (imitating food groups) exhibit varying amounts of covariance within each block. We compare both the predictive ability and scientific interpretability of the constructed patterns. These methods are applied to MESA data in Section 2.4 to create dietary patterns related to BMI, as excess body weight is an important pathway between nutrition and cardiovascular disease [20]. We discuss the study findings in Section 2.5, and additional analytic and simulation results can be found in Appendix A and B, respectively.

2.2 Methodology

The notation convention we use throughout is upper-case bold letters for matrices ($\mathbf{X}$), lower-case bold letters for vectors ($\mathbf{y}$), italic letters for scalars (c), and $\|\mathbf{w}\|_p = \left(\sum_i |\mathbf{w}_i|^p\right)^{1/p}$ for the L_p norm of $\mathbf{w}$. In our setting, we would like to predict a continuous univariate response $\mathbf{y}$ (like

BMI) using a function of the original, potentially highly-correlated predictors in matrix $\mathbf{X}_{n\times p}$ (such as diet), where the number of observations n is greater than the number of predictors p. We assume that each predictor in $\mathbf{X}$ is centered, without loss of generality. In order to create predictive reduced-dimension summaries of $\mathbf{X}$, we consider a class of methods that construct K mutually orthogonal linear combinations of the covariates, denoted by matrix $\mathbf{T}_{n\times K} = [\mathbf{t}_1 \cdots \mathbf{t}_K]$, where each column is a linear combination. These combinations are called *components* or *scores*, though we also use *latent variables* and *patterns* interchangeably. We call this class "component-based methods", and all of the techniques we compare apart from Lasso fall under this category.

2.2.1 Review of select dimension reduction methods

Component-based methods construct weighted sums of the predictors, which are created in descending order of "importance". For example, the first PLS component explains the largest amount of covariance between $\mathbf{X}$ and $\mathbf{y}$ that can be constructed by a weighted sum of the predictors, with higher weights assigned to predictors highly correlated with $\mathbf{y}$. After subtracting out the first component's information from $\mathbf{X}$ and/or $\mathbf{y}$ (to preserve the orthogonality between components), the second PLS component describes the most covariance between the residual predictors and response. This process is repeated for K components, chosen by the researcher, with $K \leq \text{rank}(\mathbf{X}^T\mathbf{X}) \leq p$. If the maximal number of components are constructed, no dimension reduction has occurred, so the PLS solution (and similarly the PCA solution) simplifies to the ordinary least squares (OLS) estimate from linear regression [26]. Indeed, it has been shown that the full set of components form an orthogonal basis for $\mathbf{X}$ [33], so the same information in $\mathbf{X}^T\mathbf{X}$ is contained in $\mathbf{T}^T\mathbf{T}$.

Sparse methods (i.e., methods resulting in either a few non-zero weights or regression coefficients) are useful when not all of the predictors are associated with the response. Indeed, if many irrelevant variables are kept in a component-based model, they could attenuate the estimated regression coefficients of the relevant variables [14]. To improve interpretability, sparse techniques have incorporated variable selection by penalizing a quantity related to

the regression coefficients $\hat{\boldsymbol{\beta}}$—typically the weights in component-based methods.

Principal Component Analysis and Regression

Principal Component Analysis asserts that we can decompose $\mathbf{X}$ into orthogonal components that explain the largest amount of variability in $\mathbf{X}$. Formally, the PCA model is $\mathbf{X} = \mathbf{T}\mathbf{P}^T = \mathbf{T}_K\mathbf{P}_K^T + \mathcal{E}_K$, where the latter notation emphasizes that there is some residual $\mathbf{X}$ information left out when we use K components. As the components are linear combinations of $\mathbf{X}$, we can write this relationship as $\mathbf{T} = \mathbf{X}\mathbf{P}$. Therefore, the key is to find the appropriate weights $\mathbf{P}$ (normally called *loadings* in PCA literature) to obtain the latent variables that maximize their described variability. To ensure that the components are uncorrelated with each other and constructed in descending order of explained variance, the optimal weights correspond to the eigenvectors of $\mathbf{X}^T\mathbf{X}$, ordered by their eigenvalues.

One way to obtain an estimate of $\mathbf{P}$ is by doing a singular value decomposition (SVD) of $\mathbf{X}$ (i.e., $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$) and setting $\hat{\mathbf{P}} = \mathbf{V}$ and $\mathbf{T} = \mathbf{U}\mathbf{D}$. Another more popular way to estimate $\mathbf{P}_{p\times K} = [\mathbf{p}_1 \cdots \mathbf{p}_K]$ is by using the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [81], which can sequentially calculate the most important weights and components until explaining all of the $\mathbf{X}$ variability. Here, the optimization problem for the kth weight is

$$\hat{\mathbf{p}}_k = \underset{\mathbf{p}}{\operatorname{argmax}} \left\{\mathbf{p}^T\mathbf{X}^T\mathbf{X}\mathbf{p}\right\} \qquad \text{s.t. } \mathbf{p}^T\mathbf{p} = 1 \text{ and } \mathbf{p}^T\mathbf{X}^T\mathbf{X}\mathbf{p}_j = 0 \text{ for } j < k.$$

The first constraint is used for the identifiability of the weights and the second ensures that components are orthogonal to all previously constructed components. This optimization problem is equivalent to maximizing $\text{Var}(\mathbf{t}) = \mathbf{t}^T\mathbf{t}/n$ with the same constraints, as we want our latent variables to express as much of the variability of $\mathbf{X}$ as they can.

If one is interested in assessing the relationship between the predictors and an outcome $\mathbf{y}$, principal component regression (PCR) simply uses a chosen number of PCA components as the new covariates. The estimated response becomes $\hat{\mathbf{y}} = \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$, resulting in PCR

regression coefficients $\hat{\boldsymbol{\beta}} = \mathbf{P}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$. The advantage of PCR in comparison to OLS is that one can reduce the dimension of the problem by choosing fewer components than predictors, while still preserving most of the variability of $\mathbf{X}$. In this manner, high-dimensional problems ($n \leq p$) can become tractable. Even in low-dimensional ($n > p$) settings, the PCR estimator is more stable than the OLS estimator when there is multicollinearity (high correlation among some $\mathbf{X}$), which is the setting that we consider.

Sparse Principal Component Analysis

Initially, researchers sought to improve the interpretability of PCA components, so Simplified Component Technique-LASSO (SCoTLASS) was proposed [40], which adds an L_1 penalty to the PCA weight constraints. This constraint shrinks the weights to $\mathbf{0}$, potentially eliminating some variables that describe a small amount of variability in each component step. However, SCoTLASS is often not sparse enough (e.g., many non-0 weights) and it is not a convex problem, meaning that there is no guarantee that the final solution is truly the global optimal solution (rather than a local optimum) and it is more computationally intensive.

To address these issues, a Sparse PCA method was proposed that generalizes a ridge regression formulation of PCA and then adds an L_1 penalty to the PCA weights [88]. First, they demonstrated that PCA can be recast as ridge regressions, by considering the definition $\mathbf{T} = \mathbf{XP}$ as a multivariate regression problem. For the kth component, the "response" is $\mathbf{t}_k = \mathbf{U}_k\mathbf{D}_{kk}$ (from the SVD approach of fitting PCA) and the "regression coefficient" is $\mathbf{p}_k$. Then, the ridge estimator $\hat{\mathbf{p}}^R = \underset{\mathbf{p}}{\operatorname{argmin}} \left\{ \|\mathbf{t}_k - \mathbf{Xp}\|_2^2 + \lambda\|\mathbf{p}\|_2^2 \right\}$ can be used to estimate the kth weight: $\hat{\mathbf{p}}_k = \hat{\mathbf{p}}^R / \|\hat{\mathbf{p}}^R\|_2$.

However, this approach not only requires an estimate of the latent components ($\mathbf{t}_k$) by doing PCA ahead of time, but it still remains difficult to interpret the results, since all of the variables are used to construct the components. The authors suggested generalizing the regression problem to avoid relying on a previous PCA, and then adding an L_1 penalty on the weights in order to impose sparsity for better scientific interpretability. Their Sparse

PCA optimizes

$$\underset{\mathbf{C},\mathbf{P}}{\operatorname{argmin}} \sum_{j=1}^{k} \left\{ \|\mathbf{X}\mathbf{c}_j - \mathbf{X}\mathbf{p}_j\|_2^2 + \lambda\|\mathbf{p}_j\|_2^2 + \lambda_j\|\mathbf{p}_j\|_1 \right\} \qquad \text{s.t. } \mathbf{C}^T\mathbf{C} = \mathbf{I}_k.$$

The estimated sparse weights $\widehat{\mathbf{P}}$ are obtained by iteratively solving $\mathbf{C}$ and $\mathbf{P}$ until convergence. At that point, the sparse weights are used to construct SPCA components, which can be used as new predictors like in PCR (i.e., SPCR). While the authors allow for separate tuning parameters (λ_j) for each component's weight, we simplify this to a single parameter for a fairer comparison with the other sparse methods and reduced computation time.

Though both sparse PCA methods improve interpretation because each component is formed by fewer variables (i.e., component-wise sparsity), they do not guarantee variable selection (i.e., overall sparsity) unless hardly any components are selected with each being composed of a few variables. If we have many potentially irrelevant food variables, such as from a food frequency questionnaire, we posit that overall sparsity is a more beneficial aim.

Partial Least Squares

Partial least squares (specifically, PLS1) is similar to PCA in that it decomposes $\mathbf{X}$ into latent variables, but now we simultaneously model $\mathbf{y}$ via these latent variables, resulting in components that explain the most (squared) covariance between $\mathbf{X}$ and $\mathbf{y}$. Formally, the univariate PLS model is

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathcal{E}_\mathbf{x} \qquad \mathbf{y} = \mathbf{T}\mathbf{q}^T + \epsilon_\mathbf{y}, \qquad \text{where } \mathcal{E}_\mathbf{x}, \epsilon_\mathbf{y} \perp\!\!\!\perp \mathbf{T}.$$

There is debate as to how to interpret these models from a statistical standpoint (see, e.g., [82]), as some view $\mathcal{E}_\mathbf{x}$ as the residual $\mathbf{X}$ information after reducing its dimension to K components and $\epsilon_\mathbf{y}$ as a combination of residual $\mathbf{y}$ information and random noise. However, the "latent variable" perspective posits that $\mathbf{T}$ represents the true underlying phenomena of interest, in which case one would consider both $\mathcal{E}_\mathbf{x}$ and $\epsilon_\mathbf{y}$ as purely random noise.

Regardless of interpretation, the components are a set of mutually orthogonal linear combinations of $\mathbf{X}$, which can be expressed as $\mathbf{T} = \mathbf{XR} = \mathbf{X}\,\mathbf{W}(\mathbf{P^TW})^{-1}$, and consequently $\hat{\boldsymbol{\beta}} = \mathbf{R}\,(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y} = \mathbf{R}\mathbf{q}^T$. Thus, one must find the appropriate weights to obtain $\mathbf{T}$—either Gram–Schmidt-orthogonalized $\mathbf{W}$ from the NIPALS algorithm (which also computes loadings $\mathbf{P}$) or oblique projection $\mathbf{R}$ from the Statistically Inspired Modification of PLS (SIMPLS) algorithm [21]. Both algorithms result in the same estimator for univariate $\mathbf{y}$, but not for multivariate responses. Either PLS algorithm finds the weights in a sequential manner, solving successive optimization problems; the kth (NIPALS) weight solves[1]

$$\widehat{\mathbf{w}}_k = \underset{\mathbf{w}}{\operatorname{argmax}} \left\{\mathbf{w}^T\mathbf{X}^T\mathbf{y}\,\mathbf{y}^T\mathbf{X}\mathbf{w}\right\} \qquad \text{s.t. } \mathbf{w}^T\mathbf{w} = 1 \text{ and } \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}_j = 0 \text{ for } j < k.$$

This is equivalent to maximizing the $\mathrm{Cov}^2(\mathbf{y}, \mathbf{t})$, or more specifically $\mathrm{Cor}^2(\mathbf{y}, \mathbf{t})\ \mathrm{Var}(\mathbf{t})$, with the same constraints [62]. The optimal PLS weights are the eigenvectors of $\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}$ ordered by their eigenvalue, which ensures that the components are uncorrelated with each other and constructed in descending order of covariance between $\mathbf{y}$ and $\mathbf{t}$.

The PLS estimator with K components can actually be interpreted in a few ways, as

$$\hat{\boldsymbol{\beta}}^{(K)} = \mathbf{R}_K\,(\mathbf{T}_K^T\mathbf{T}_K)^{-1}\,\mathbf{T}_K^T\,\mathbf{y} \;=\; \sum_{m=1}^{K} s_m\mathbf{d}_m \;=\; \arg\min_{\beta\in\mathcal{K}_K} \left\|\mathbf{y} - \mathbf{X}\beta\right\|_2^2.$$

The first form demonstrates that the PLS estimator is the projection of the component-scale coefficient $\mathbf{q}^T$. The second form indicates that the PLS estimator is equivalent to the conjugate gradient estimate after K steps (proven in [60] and Appendix A). This provides a useful connection to optimization techniques, which will be utilized in Chapter 3. Specifically, s_m can be considered an optimized step size and $\mathbf{d}_m$ the conjugated residual covariance. Finally, these estimates are restricted to a particular subspace, which provides a convenient way to connect the objective of PLS with that of typical regression models; the K-th order

[1]Each previous component's information is actually subtracted out from $\mathbf{X}$, but we opted to present it in this way to transmit the main point. See Appendix A.1 for the complete NIPALS algorithm.

Krylov subspace $\mathcal{K}_K(\mathbf{S}, \mathbf{s})$ is defined as $\text{span}\{\mathbf{s}, \mathbf{S}\mathbf{s}, \mathbf{S}^2\mathbf{s}, \cdots, \mathbf{S}^{K-1}\mathbf{s}\}$, where we let $\mathbf{s} = \mathbf{X}^T\mathbf{y}$ and $\mathbf{S} = \mathbf{X}^T\mathbf{X}$. For a discussion about the Krylov subspace, see [60], Appendix A.1, and Appendix A.3. Each of these forms are a useful way to consider how PLS creates a reduced-dimension summary of the OLS estimator (with equality when $K = \text{ rank}(\mathbf{X}^T\mathbf{X})$).

Sparse Partial Least Squares

To obtain more interpretable PLS components, a Sparse PLS method was proposed that adds an L_1 penalty to the PLS weight constraints [44]. These soft-thresholded weights are shrunken towards 0 and used to construct the corresponding Sparse PLS components. However, similar to SCoTLASS and SPCA, inducing component-wise sparsity does not guarantee variable selection. Additionally, by shrinking the PLS weights, they are no longer orthogonal, which implies that they do not form an orthonormal basis for the restricted subspace of PLS. Thus, this Sparse PLS estimator is not guaranteed to equal the OLS estimator even when the maximal number of components are constructed.

At first, Chun and Keles tried improving this method by following the idea in SPCA of considering a surrogate weight $\mathbf{c}$ kept close to $\mathbf{r}$, but with both an L_1 and L_2 penalty on $\mathbf{c}$. Letting $\mathbf{M} = \mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}$ for simplicity, they initially considered this optimization problem for the (SIMPLS) weights:

$$\underset{\mathbf{r},\mathbf{c}}{\text{argmin}} \left\{ -\kappa\mathbf{r}^T\mathbf{M}\mathbf{r} + (1-\kappa)(\mathbf{c}-\mathbf{r})^T\mathbf{M}(\mathbf{c}-\mathbf{r}) \; + \; \lambda_1\|\mathbf{c}\|_1 + \lambda_2\|\mathbf{c}\|_2^2 \right\} \qquad \text{s.t. } \mathbf{r}^T\mathbf{r} = 1.$$

However, they estimated the sparse weights using $\hat{\mathbf{c}}$, obtained by iterating between solving for $\mathbf{r}$ and $\mathbf{c}$. They justified considering only the cases where $0 < \kappa \leq 0.5$ to avoid concavity issues. Indeed, when $\kappa = 0.5$ and $\mathbf{y}$ is univariate, $\hat{\mathbf{r}}_1 = \mathbf{X}^T\mathbf{y}/\|\mathbf{X}^T\mathbf{y}\|_2$, the first PLS weight. For $\kappa \in (0, 0.5]$, $\hat{\mathbf{c}}$ can be estimated using the LARS algorithm, but they noted that $\mathbf{M}$ was often singular, particularly for low-dimensional $\mathbf{y}$. They proposed setting $\lambda_2 = \infty$, especially as $\hat{\mathbf{c}}$ does not depend on λ_2 for univariate $\mathbf{y}$. With these simplifications, $\hat{\mathbf{c}} = \text{sign}(\hat{\mathbf{r}}) * (|\hat{\mathbf{r}}| - \lambda_1/2)_+$, the soft-thresholded version of $\hat{\mathbf{r}}$, where $(\,\cdot\,)_+ = \max(\,\cdot\,,\, 0)$. But this is equivalent to the

initial Sparse PLS method for univariate $\mathbf{y}$, so they changed strategies.

Chun and Keles proposed a two-step Sparse PLS procedure that directly induces variable selection and preserves orthogonal weights. For each component, the variables most related to the (residual) response, as measured by covariance, are selected and added to the index set $\mathcal{A}$. The subset of variables $\mathbf{X}_{\mathcal{A}}$ is chosen based on a given sparsity percentage $\eta \in [0, 1)$ and K components. Starting with $\widehat{\boldsymbol{\beta}}^{(0)} = 0$ and indexing set $\mathcal{A}_0 = \{\}$, for $m \in [1, K]$,

$$\mathcal{A}_m = \mathcal{A}_{m-1} \cup \left\{ j \,:\, \left| \mathbf{X}_j^T \left(\mathbf{y} - \mathbf{X}_{\mathcal{A}_{m-1}} \widehat{\boldsymbol{\beta}}_{\mathcal{A}_{m-1}}^{(m-1)} \right) \right| > \eta \max_{1 \le h \le p} \left| \mathbf{X}_h^T \left(\mathbf{y} - \mathbf{X}_{\mathcal{A}_{m-1}} \widehat{\boldsymbol{\beta}}_{\mathcal{A}_{m-1}}^{(m-1)} \right) \right| \right\}$$

where $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_{m-1}}^{(m-1)}$ is the coefficient from an $(m-1)$-component PLS model using the covariates $\mathbf{X}_{\mathcal{A}_{m-1}}$. The final selection of variables $\mathcal{A} = \mathcal{A}_K$ includes those covariates that have the strongest relationship with the response or that best capture the variation among covariates when transformed into K independent linear combinations. Last, this procedure fits a K-component PLS model with the final selected variables. As such, this sparse PLS estimate can be denoted by $\widehat{\beta}_{\mathcal{A}}^{(K)} = \arg\min_{\beta \in \mathcal{K}(\mathcal{A})_K} \left\| \mathbf{y} - \mathbf{X}_{\mathcal{A}}^T \, \beta \right\|_2^2$. Here $\mathcal{K}(\mathcal{A})_K$ indicates the K-th order Krylov subspace with $\mathbf{s} = \mathbf{X}_{\mathcal{A}}^T \, \mathbf{y}$ and $\mathbf{S} = \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}$. We refer to this procedure as SPLS throughout the dissertation, which simplifies to PLS if $\eta = 0$.

SPLS directly selects the variables that are most related to $\mathbf{y}$—even when fitting more than one component—as long as the sparsity level (η) and signal-to-noise ratio are high. In comparison, the initial Sparse PLS method may add in variables with weak relationships to $\mathbf{y}$ as early as the second component if the first component has included most of the information from the strongly related variables (see simulation study in Appendix B.2). Additionally, by fitting PLS on a reduced set of variables, Chun and Keles's weights are still orthogonal, so the SPLS estimator with the full number of components will be equivalent to the OLS estimator with the same set of reduced variables. However, this method no longer promotes component-wise sparsity, since all of the selected variables are used to construct each component.

Lasso

The Lasso is a penalized linear regression method that directly induces variable selection by shrinking all $\hat{\boldsymbol{\beta}}$ towards $\mathbf{0}$ using an L_1 penalty [68]. It can be classified as a dimension reduction method, and its estimate takes the form $\hat{\boldsymbol{\beta}}_{Lasso} = \min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$. If the selected sparsity level $\lambda \geq 0$ is large enough, the smaller coefficients will be set to 0, resulting in variable selection. While the Lasso is optimized for predictive accuracy, it is known to struggle with highly-correlated covariates [86], as it will select only one correlated variable (regardless of etiology). Although the Lasso does not create a latent set of variables, we say that it constructs one pattern ($\mathbf{X}\hat{\boldsymbol{\beta}}_{Lasso}$) to compare it with the component-based methods.

2.2.2 Selection of tuning parameters

After implementing any of these dimension reduction methods, the researcher must decide how many components and/or the sparsity level to select for the final model. Since we want to select a predictive model that is not overfit to our training data, we use 10-fold cross-validation (CV) to be more realistic when assessing a potential model's predictive accuracy, which is measured by the CV root mean squared error of prediction (RMSE).

Once candidate models have been constructed from several (pairs of) tuning parameter values, there is a trade-off between accuracy and complexity. One could select the model with the lowest CV-RMSE, but it still might be modeling some noise. There are typically less complicated models (e.g., with fewer components or more sparsity) that have similar predictive accuracy. To avoid plotting the CV-RMSE curve and selecting an accurate yet (subjectively) parsimonious model, different tests have been developed to select appropriate tuning parameters.

We employ the one standard error (1-se) test to select a reasonable number of components, which aligns with the 1-se test for the Lasso when choosing a reasonable level of sparsity [31]. One first identifies the model with the smallest CV-RMSE, called the reference model. Then, the most parsimonious model whose CV-RMSE is within one standard error of the

reference model's RMSE is selected as the final model. Since SPCR and SPLS additionally incorporate sparsity, their tuning parameter decision becomes two dimensional. We extend the 1-se test to optimize the parameter search across a grid of candidate pairs, favoring a decrease in number of components over an increase in sparsity. More details about this two dimensional 1-se test can be found in Appendix B.1.

2.2.3 Computation

Analyses were performed in R version 3.6.1 [67]. PCR and PLS were run using the `pls` package [52], SPCR was run with the `elasticnet` package [87], SPLS was implemented using a wrapper on functions in the `spls` package [15], and Lasso was run with the `glmnet` package [28]. Both the `pls` and `glmnet` packages have the 1-se test implemented, and a wrapper was written to conduct the 1-se test with two tuning parameters for SPCA and SPLS. For the component-based methods, we follow the data pre-processing convention in `pls` by scaling each predictor by its standard deviation and then centering it, as well as centering the response $\mathbf{y}$. This allows for a more similar comparison between predictors that have different levels of variability. For the Lasso, we only scale each predictor by its standard deviation. We also used the same folds across all methods for 10-fold cross-validation to make the comparisons as fair as possible.

2.3 Simulation studies

2.3.1 Simulation design

As our study is motivated by nutritional epidemiology, we seek to generate predictors that mimic the features of FFQ data. In particular, it is natural to classify foods into food groups, or essentially to create blocks of variables that are more similar to each other. Such "multiblock data" occur in many other scientific domains, since blocks can also represent different types of experiments or different sample collection sites from the same subject, as in some metabolomics studies (see [38, 42]). We have designed the simulation study with

blocks of differing size and structure, as would be found in practice.

We generate a multivariate-normal predictor $\mathbf{X}$ with 1000 observations and 24 variables, only nine of which contribute to the response $\mathbf{y}$. The predictors are grouped into four blocks; Block 1 has two variables (one is relevant), Block 2 has six variables (two are relevant), Block 3 has nine variables (six are relevant), and Block 4 has seven variables (none are relevant). Each predictor has mean zero and unit variance, with the variability representing different relative levels of food consumption in the context of nutritional epidemiology.

The first setting we consider (Case 1) is when all the variables are independent of each other, which is the simplest scenario since relevant and irrelevant variables should be easily distinguishable. But the intake of foods within a food group could be related in practice, so in further settings we also vary the covariance between variables of the same block. We first test scenarios of "matching-covariance block structure", where all blocks have the same level of intra-block covariance ranging from 0.3 to 1 (see Appendix B.3).

Next, we consider "mixed-covariance block structure", where blocks can have different intra-block covariance, using the MESA diet data at baseline as a guide for reasonable correlations in food frequency questionnaires. Thus, the variance of $\mathbf{X}$ has a block diagonal structure, with independence between blocks for simplicity. Small amounts of covariance between blocks were included in a sensitivity analysis, with little change in the results. In Case 2 the intra-block covariances are 0.05, 0.50, 0.30, and 0.10, respectively. In this setting, we would expect that it would be most difficult to separate the relevant variables from the irrelevant ones in Block 2. We also test further scenarios of mixed-covariance block structure by permuting the covariances in Case 2 (see Appendix B.3).

We assign the nine relevant variables to have unit positive or negative associations with the response as follows: the Block 1 and three Block 3 relevant variables are positive while the remaining relevant variables are negative. As the relevant variables have the same magnitude, we are interested in whether methods can distinguish between positive, negative, and null relationships, not differing levels of positive or negative values. The generated response is univariate normal with unit variance and with the sum of the relevant variables (multiplied

by the true association) being its mean. In the context of nutritional epidemiology, the variability here represents the contribution of factors other than diet to any response.

For each of 100 generated data sets, we split our observations into a training (n=700) and test (n=300) set, and we use 10-fold cross validation on the training data to select the tuning parameters for each method (e.g., number of components and/or sparsity level). Due to computational burden, we only perform 5-fold CV for SPCR. Predictions are then made on the test data with the selected tuning parameters. To evaluate predictive performance, we report the average RMSE from the test sets. To compare the interpretability across the methods, we report the average number of variables selected, number of patterns, and percentage of $\mathbf{y}$ explained (R^2).

2.3.2 Results

Table 2.1(a) shows the performance of the five methods across 100 data sets from Case 1, when variables are independent. Among the methods that can induce sparsity, SPLS and Lasso select the relevant variables and up to one irrelevant variable, on average. While SPCR imposes sparsity in the patterns (selecting one or two variables per pattern), it keeps almost all of the predictors because of the large number of patterns selected in order to have competitive predictive ability. When it does discard variables (in 47 data sets), SPCR discards at least one relevant variable half of the time. In terms of prediction, all methods except for SPCA perform well, with SPLS having the lowest RMSE at 1.013, on average. By construction, Lasso produces one pattern, but both PLS and SPLS only create two patterns, the fewest of the component-based methods. PLS's first pattern highlights most of the relevant variables, while PCR describes mostly irrelevant variables at first. Case 1 is the setting in which the methods capture the most information about the response, with most R^2s between 89-90%.

As more covariance is introduced equally in the matching-covariance block structure, the relevant and irrelevant variables become harder to distinguish in the first three blocks (see Appendix B.3). With 0.3 intra-block covariance, variables in each block behave more

Table 2.1: Simulation results across 100 data sets, with mean and (standard deviation).

(a) Case 1: Independent variables (no block structure)

Method	**Number of Variables**[†]	**Number of Patterns**	**RMSE**	R^2 (%)	**First pattern's** R^2 (%)
PCR	24.0 (0.0)	23.8 (0.5)	1.026 (0.040)	90.1 (0.7)	5 (7)
SPCR	23.1 (1.4)	23.0 (1.4)	1.140 (0.211)	87.5 (5.3)	5 (5)
PLS	24.0 (0.0)	2.0 (0.0)	1.027 (0.040)	90.1 (0.8)	88 (1)
SPLS	9.1 (0.6)	2.0 (0.4)	1.013 (0.038)	89.9 (0.8)	89 (1)
Lasso	9.8 (1.0)	1.0 (0.0)	1.043 (0.039)	89.3 (0.8)	89 (1)

(b) Case 2: Mixed-covariance block structure (intra-block covariances of 0.05, 0.50, 0.30, and 0.10)

Method	**Number of Variables**	**Number of Patterns**	**RMSE**	R^2 (%)	**First pattern's** R^2 (%)
PCR	24.0 (0.0)	23.4 (0.8)	1.034 (0.041)	89.3 (0.8)	17 (8)
SPCR	24.0 (0.0)	23.8 (0.5)	1.032 (0.044)	89.3 (0.8)	13 (9)
PLS	24.0 (0.0)	3.5 (0.5)	1.036 (0.042)	89.3 (0.8)	53 (2)
SPLS	13.2 (0.7)	3.7 (0.7)	1.022 (0.041)	89.0 (0.8)	53 (2)
Lasso	10.7 (1.3)	1.0 (0.0)	1.038 (0.040)	88.5 (0.9)	88 (1)

RMSE=root mean squared error. [†] All methods selected the nine relevant variables, except for SPCR in Case 1, which discarded one to two relevant variables in 23 data sets.

similarly, inducing an association between the response and the five irrelevant variables in Block 1 and 2, in particular. SPLS incorrectly selects four of these irrelevant variables, while Lasso only selects 1.5 of them incorrectly on average. SPLS still maintains its predictive ability at an RMSE of 1.019, while Lasso and PLS result in an RMSE of 1.043. It is not until 0.8 intra-block covariance that SPLS provides little improvement in sparsity, as it selects 21 variables on average compared to Lasso which selects 12 variables.

Table 2.1(b) shows the performance of the five methods in Case 2, a mixed-covariance block structure, which is a more realistic setting in nutritional epidemiology. Since we still impose independence across blocks, each block behaves similarly to its respective matching-covariance block structure setting. As Block 2 has the highest covariance (0.5), its four

irrelevant variables behave more similarly to its relevant variables, inducing an association with $\mathbf{y}$. SPLS selects those four irrelevant variables, Lasso only selects 1.7 irrelevant variables, and SPCR includes all variables (selecting only one or two variables per pattern). In terms of prediction accuracy, all methods have RMSEs between 1.032-1.038, except for SPLS, which results in the lowest RMSE at 1.022. Again, Lasso only constructs one pattern, while both PLS and SPLS create 3.6 patterns on average. SPLS's first component emphasizes Block 2 as well as the other relevant variables (to a lesser extent), while both PLS and PCR highlight Block 2 alone in their first component. PCR's second component picks Block 3 variables (the next most correlated block), whereas PLS's second component emphasizes Block 2 variables alongside other relevant variables. In Case 2, all the methods capture roughly 89% of the response information. In the other tested mixed-covariance block structures, we note that the supervised methods perform better when blocks with more relevant variables have lower intra-block covariances.

Across all the tested simulation scenarios, SPLS consistently predicts the best, on average. Regarding variable selection, SPLS outperforms Lasso when predictors are independent (Case 1) by selecting fewer irrelevant variables. Once there is moderate matching covariance within blocks, SPLS selects about 3-5 more irrelevant variables than Lasso. When the covariance structure changes by block, SPLS selects up to three more irrelevant variables than Lasso. Except for in cases of high (0.8-1) matching covariance within blocks, SPLS always selects the relevant variables. On the other hand, SPCR imposes sparsity on the number of variables used in each component, rather than on the actual number of variables selected overall. Thus, SPCR at most drops one variable.

With respect to the number of patterns, Lasso is the most parsimonious, as it creates one pattern by default. In the simplest cases, SPLS selects two components, and in the trickiest (high matching covariance), it selects 6.7 components, on average. In all settings except perfect correlation within blocks, SPLS selects slightly more components than PLS, as it is using far fewer variables to describe the response. Meanwhile, PCA selects between 22 to 24 patterns for all cases except perfect correlation (4 patterns) and the simplest mixed

covariance by block structure (20 patterns). SPCR has similar performance with regard to number of patterns, typically selecting 23-24 patterns. Finally, comparing the R^2 values, most methods explained a similar amount of the response variability, with PCR describing the most and the Lasso usually describing the least (but only about 1% less variability than PCR). Case 1 (independence) resulted in the highest R^2 values (89-90%), and perfect correlation across all blocks resulted in the lowest R^2 values (82-83%).

2.4 Data application

2.4.1 Scientific background and study design

Our motivating illustration is to identify understandable biologically-driven diet patterns related to BMI. In the Multi-Ethnic Study of Atherosclerosis, 6499 participants (95% of the cohort) had complete information on demographics, diet, BMI, and intentional exercise (see [9] for more details). These participants were aged 45-84 (53% women), free of clinical CVD at baseline (2000-2002), and from six United States communities (New York City, Baltimore, Chicago, Los Angeles, the Twin Cities, and Winston Salem). They self-identified as 39% white, 27% Black, 22% Hispanic, and 12% Chinese.

The MESA Block-inspired FFQ was modified from the Insulin Resistance Atherosclerosis Study's FFQ [49], which was validated in their multi-ethnic cohort, and it includes both ethnic and regional foods. In MESA's FFQ, participants reported the frequency and serving size of 120 foods and beverages that they consumed during a typical week of the previous year. There were nine frequency options, ranging from "rare or never" and "once per month" to "multiple times per day", and three serving size options (small, medium, or large compared to others of the same gender and age). The calculated servings per day of each food item were used as our predictors, after standardizing them. By adapting the American MyPlate guidelines [17], we categorized the diet variables into eight food groups: fruits, vegetables, grains, proteins, mixed entrees, dairy, sweets/oils, and beverages.

There are also many factors that could potentially confound the association between BMI

and diet, such as age, gender, and race/ethnicity. Exercise is also important to consider as it is related to BMI independently of diet. In the MESA Typical Week Physical Activity Survey, adapted from the Cross-Cultural Activity Participation Study [1], participants reported the time spent and frequency of various types of physical activity they engaged in during the previous month. Conditioning activities such as walking for exercise, dancing, and sports were categorized as intentional physical activity, and the log of the metabolic equivalent minutes per week was used in our analysis.

In part because relative diet composition is more feasible to change than absolute levels of food consumption—which is influenced by other unalterable factors—studying the effect of diet on a relative scale has been established by nutritional epidemiologists as a useful type of analysis [79]. This can be accomplished by adjusting for total energy (caloric) intake as a potential confounder (see Appendix C for further details on model interpretation), which we chose to perform. One might instead consider modeling food measurements that are divided by total caloric intake. However, as this covariate is a ratio, this type of analysis would not be able to separate the difference between decreasing one's consumption of a particular food item from increasing one's overall food intake.

To construct BMI-related dietary patterns that adjusted for these potential confounders and precision variables, we took the following modeling approach. PCA-based components were created as usual and then included in a multiple linear regression model with the adjustment variables. PLS-based components were constructed on the BMI-residuals from a multiple linear regression with only the adjustment variables, and then these components were included in a multiple linear regression model for BMI with the adjustment variables. More details on this modeling approach will be covered in Chapter 4.

2.4.2 MESA results

Average servings per day of each food item tended to be quite low, since many people did not consume some of the foods. The most popular food was coffee, at 1.13 average servings per day. The next most frequent food was diet soda, with 0.42 average servings. There

were 21 foods that more than 25% of the MESA participants consumed, such as apples and bananas, and similarly 20 foods that less than a quarter of the cohort consumed. BMI ranged from 15 to 62 (median of 27.5) for those with complete data. It has been suggested that a recommended normal BMI range for seniors is $[23, 31)$, and 60% of the MESA cohort were in this category [80]. When modeling without diet data, the adjustment variables explained 15% of the variation in BMI. Foods and beverages were only mildly correlated with BMI residuals, with correlations ranging from -.07 to .11. Foods with the strongest negative correlations (between -.07 and -.05) were natural sweeteners added to coffee and tea, wine, and sweet potatoes, in decreasing order. The most positively correlated foods to BMI residuals (between .11 and .06) were hamburgers, diet soda, light green lettuce, fries, ham, and steak, in decreasing order.

Table 2.3: BMI-related dietary patterns in the Multi-Ethnic Study of Atherosclerosis.

Method	**Number of Variables**	**Number of Patterns**	**CV-RMSE**	$\boldsymbol{R^2}$ **(%)**
PCR	120	17	4.92	17.7
SPCR	116	67	4.86	19.7
PLS	120	1	4.88	19.0
SPLS	5	2	4.91	17.9
Lasso	11	1	4.94	17.5

CV-RMSE=cross-validated root mean squared error (not on a test set).

Table 2.3 displays the results of the adjusted BMI-related dietary patterns in MESA using five dimension reduction methods. From the 120 foods and beverages, PCR constructed 17 patterns, which had a cross-validated RMSE of 4.92 and described 17.7% of the BMI variability. The foods with the largest magnitude associations were (in order) hamburgers, steak, fries, and diet soda, all positively related to BMI. Sparse PCR achieved the best CV-RMSE and R^2 values, but it required 67 components to do so. As expected, SPCR created sparse patterns (each component being constructed from one or a handful of foods) which allows for easier scientific interpretation, but it only discarded four foods completely. The first three components described the consumption of Western/fast foods, stir fried Asian

meals, and Hispanic meals, respectively. The most important foods overall according to SPCR were hamburgers, diet soda, light green lettuce, and fries, all positively associated with BMI and listed by magnitude.

PLS selected only one diet pattern using all of the foods (RMSE=4.88, R^2=19.0%). The most relevant foods here were hamburgers (+), diet soda (+), natural sweeteners added to coffee or tea (-), and light green lettuce (+), where (+) indicates a positive association and (-) a negative association. Sparse PLS constructed two diet patterns with only five foods. With this reduction in covariates, it still had comparable predictive accuracy and R^2 to the other methods, even outperforming PCR. The first diet pattern emphasized hamburger and diet soda intake, typical fast food staples. The second diet pattern again highlighted hamburger consumption, but it also focused on natural sweeteners added to coffee and tea. By order of magnitude, the most relevant foods overall were hamburgers (+), diet soda (+), light green lettuce (+), natural sweeteners (-), and wine (-).

Figure 2.2: SPLS's first BMI-related diet pattern in MESA, ordered by strength of relationship with BMI.

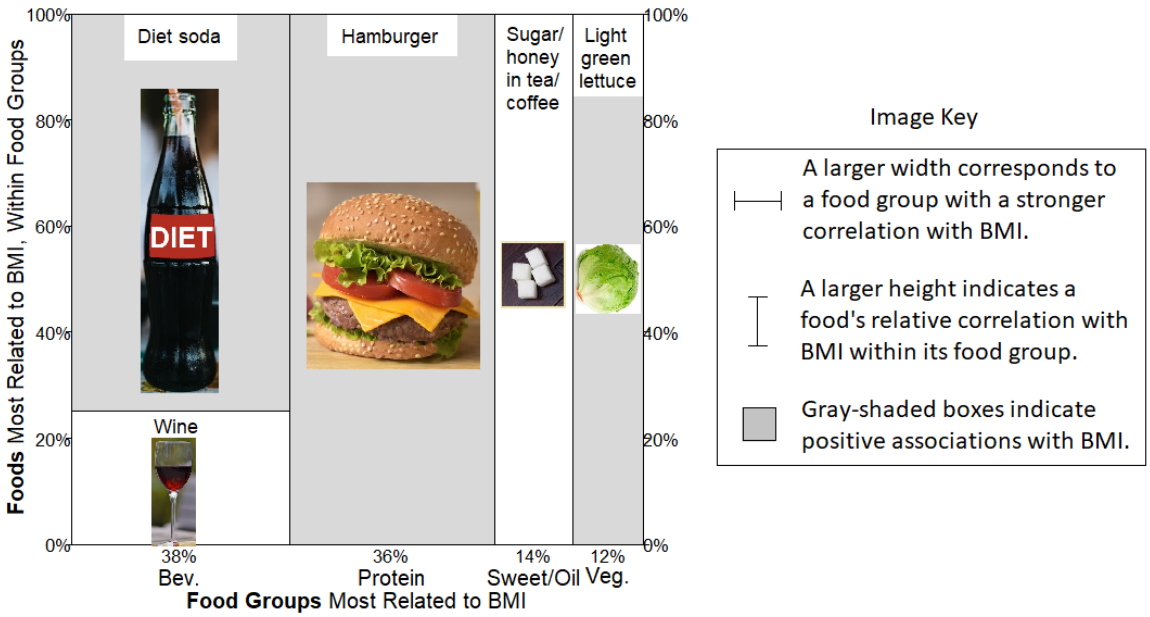


Figure 2.2 displays the relative contribution of the selected foods to the construction of

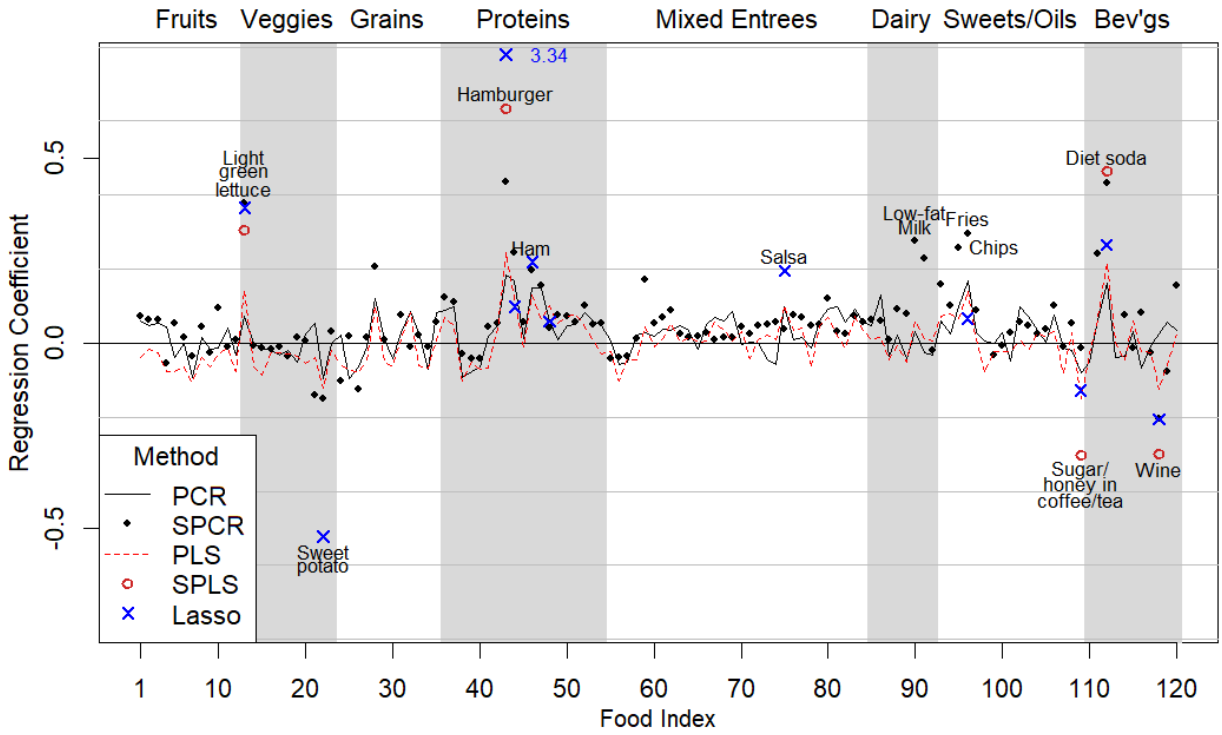
the first SPLS component. Along the horizontal axis, food groups are sorted by their squared correlation with BMI (known as global weights within Multiblock PLS literature [77]). The vertical axis represents the relative importance of the selected foods within each food group (i.e., block weights), ordered by their squared correlation with BMI. It has been shown that these multiblock weights are simply scaled and grouped versions of their PLS counterparts (when there is no missing data) [78, 61], and Chun and Keles's Sparse PLS method can be viewed as a PLS on a reduced set of foods. However, we opt to focus on the $\mathbf{R}$ weights instead of $\mathbf{W}$ so as to be analogous to plotting PCA loadings. Hence, this type of graphic can provide an added perspective about how each food group contributes to the creation of (S)PLS components.

The Lasso's diet pattern was composed of 11 foods, the most important of which were hamburgers (+), light green lettuce (+), diet soda (+), and sweet potatoes (-). Although the Lasso selected the same five foods as SPLS did along with six others, its CV-RMSE and R^2 performance was the worst among the five methods. In particular, most of the foods selected by Lasso but discarded by SPLS tended to be highly correlated with total energy intake. Figure 2.3 displays the resulting regression coefficients from the considered dimension reduction methods. PCR and PLS (solid and dashed lines) are forced to model all 120 foods, but the inclusion of potentially many irrelevant foods could contribute to the attenuated coefficients. SPCR (filled dots) selects most of the foods, but it does give more magnitude to some of the same foods highlighted by SPLS (open dots) and Lasso (diagonal crosses). Lasso's regression coefficients vary the most, with the coefficient for hamburger being four times larger than any other method's coefficients.

2.5 Discussion

When the scientific goal is to construct patterns that reflect the underlying relationship between predictors and a continuous outcome, PLS-based methods have useful properties to consider (as further investigated in Appendix A). If variable selection is not needed, PLS is a good alternative to PCR since it requires fewer components to describe a similar amount

Figure 2.3: Adjusted associations between foods and BMI in MESA, ordered by food groups.



of variability in the outcome, demonstrated across all of our simulation settings and the data application. If not all predictors are believed to be relevant to the relationship, then SPLS can create more interpretable patterns than PCR or PLS because it uses fewer predictors and components without much loss of predictive ability, as measured by RMSE. This was illustrated in our data application, as SPLS provided an impressive reduction in selected foods with relatively few components while maintaining predictive performance. The Lasso also excelled at discarding most foods, but it can arbitrarily drop some variables if they are highly correlated regardless of biological mechanism.

We were surprised by the performance of certain methods across the simulation settings. SPCR in particular had trouble with variable selection, convergence in several settings, and computational burden. Though SPCR created the most parsimonious patterns (selecting 1-2 variables per pattern), it required almost all of the variables and patterns to achieve

reasonable predictions. Next, PCR often selected 24 components, which is equivalent to a linear regression. These are both due to the fact that we selected the number of components based on its predictive capability, as prediction is part of the study goal. If the goal were only to explain the most variability across predictors, the PCA-based methods would have selected far fewer components, but with poorer predictive performance. Last, we expected Lasso to outperform the component-based methods in out-of-sample predictive ability (RMSE column in Table 2.1), but most methods had similar predictive performance taking into account the standard deviations of RMSE. Perhaps Lasso's advantage in prediction would become apparent if the signal-to-noise ratio were smaller, when PLS methods can behave poorly [65], or if we constructed covariates with differing magnitudes of true associations, which Lasso can more easily model [29].

Fundamental to our study is what we consider to be "interpretability". We propose that an interpretable model should, in descending order of priority, select all of the relevant variables, discard irrelevant variables, capture a few dimensions of the scientific relationship, and maintain predictive ability of the outcome. We view this as a useful way to form patterns that can help us better understand the scientific mechanisms that relate a large set of covariates to a disease or disease risk factor. The Lasso accomplishes most of these goals, except that its selection of variables is solely focused on how well a subset of variables can predict an outcome, making it less appealing for settings of potentially moderately correlated food measurements. Additionally, the field of nutritional epidemiology has favored reducing the dimension of diet variables into scores or dietary patterns when the goal is to interpret complex diet-disease associations. These prompted us to investigate PLS-based methods, and we conclude that SPLS components are a strong contender when considering these goals, particularly when the signal-to-noise ratio is sufficiently large. Another advantage of SPLS is that it can easily switch the priority between decreasing the number of components and increasing the sparsity level by adjusting the weights in the 1-se test, if the scientific setting favors this approach. However, interpreting SPLS patterns still requires reasonable scientific knowledge, as the components are combinations of (a subset of) the predictors.

Our data application illustrates how SPLS can express scientifically-interpretable relationships. Dietary pattern research has found that consuming more vegetables, fruits, whole grains, low-fat dairy, seafood, legumes, nuts, and alcohol (in moderation) is associated with lower risks of obesity and chronic diseases, as is consuming less red and processed meats, refined grains, and sugar-sweetened food and beverages [18]. The SPLS diet patterns complement this health recommendation, for the most part. The main discrepancies are our negative association between BMI and natural sweeteners added to coffee or tea and the positive association between BMI and light green lettuce. A possible explanation for this finding is that people who add honey or sugar to their beverages are conscious of how much sweetener they are consuming, as opposed to someone drinking a pre-sweetened beverage, so they could be consuming less sweetener in that way. Additionally, the "light green lettuce" category only includes lettuces similar to iceberg, not romaine. Apart from iceberg lettuce being relatively low in nutrients, this trend could reflect more of a fast food diet, as iceberg lettuce salads tend to be served at pizza parlors or fast food restaurants as a "healthy" side. The SPLS trends are fairly conventional, given what we know about obesity, so it is reassuring that this method seems to select reasonable aspects of nutrition—in this case highlighting the association between "fast food" or a more "Western" diet with BMI in MESA. However, we acknowledge that any dietary survey is prone to measurement error, which should be taken in consideration when interpreting these results (regardless of method used).

One limitation of using supervised methods (e.g., PLS, SPLS, Lasso) instead of unsupervised methods like PCA is that one should no longer use the derived components to test associations with $\mathbf{y}$. The resulting inference will be overly-optimistic since the same outcome data will be re-used. Instead, one can stop at the component stage and interpret the scientific relevance of the pattern to generate hypotheses for future studies, as we have illustrated with MESA data. Using components in a prediction setting is also a reasonable next step. If inference is the ultimate goal, researchers in practice have compared a disease outcome with PLS components that were derived from a different loosely-related response (see for example [83, 23]). Another alternative is to derive components using a hold-out sample and

then test associations in the remainder of the data, or even to derive components using one study population and then perform inference in an independent population, if the selected predictors are available.

Regardless of method or aim, we emphasize that when using data from observational studies such as MESA, results can only be interpreted as associations between predictors and a disease, not as causal relationships. However, these studies still serve a valuable role in increasing our understanding of complex diseases. For instance, if similar results are found in future studies, this could lend support to the idea that only a few foods contribute most to obesity. On a practical note, this could motivate investigating whether the current 120-item FFQ could be shortened to asking about a subset of foods, reducing the burden on participants and data centers when collecting detailed food consumption information.

Overall, we recommend that future nutritional epidemiology studies consider using sparse PLS methods if their scientific goal is to succinctly summarize aspects of predictor data that are most related to an outcome. These results can complement the previous nutritional literature, which is predominantly based on unsupervised methods, and provide further insight into how diet and diseases are related, as previously suggested [63, 39]. While we have focused on illustrating the usefulness of SPLS in the context of nutritional epidemiology, these dimension reduction methods also apply to many other disciplines whose aim is to reduce the dimension of their predictors, such as bioinformatics, environmental studies, omics fields, and other settings with high-throughput data.

Chapter 3

PLS WITH VARIABLE SELECTION FOR SURVIVAL OUTCOMES

3.1 Introduction

In nutritional epidemiology, food frequency questionnaires (FFQ) can be used to obtain information on people's typical levels of food intake. However, if there are a substantial number of food items, it is likely that certain measurements will be correlated (e.g., within food groups) and that not all foods will be strongly related to disease risk. To overcome multi-collinearity issues due to correlated food measurements, investigators interested in characterizing the relationship between diet and disease risk have predominantly utilized dimension reduction methods. Common approaches to constructing dietary patterns [63] include using previous scientific knowledge to assign "disease risk" scores to a limited number of foods and nutrients or summarizing the information from all food measurements in orthogonal linear combinations (e.g., principal component analysis [58]). However, neither of these types of patterns are inherently predictive of a disease, as they are not informed by disease data.

To concisely summarize nutritional information in a way that is tailored to a disease, we consider using partial least squares (PLS) [81], which incorporates outcome data directly into the construction of its orthogonal linear combinations (e.g., components). Specifically, PLS can create dietary patterns that are correlated with a continuous response while also capturing variation in diet composition. By using a subset of its patterns, PLS can be considered a shrinkage method, with the aim of reducing the variance of its coefficient estimates while highlighting the most prominent outcome-related aspects of the covariates.

To better understand the relationship between diet and cardiovascular disease (CVD) risk, investigators have focused on relating dietary patterns to the occurrence of heart attacks or

cardiac death, for instance, or to risk factors such as high-density lipoprotein cholesterol, fasting insulin, or C-reactive protein levels [63]. Although these outcomes can provide certain insight regarding heart-health, a more direct endpoint would be the time until someone experiences a heart disease event (if at all). However, the statistical complexities of modeling survival outcomes have largely prevented scientists from evaluating this relationship. Apart from the difficulty of accounting for right-censored data (e.g., not everyone experiences an event during the study period and some people drop-out), it is not obvious how best to integrate component-based methods into standard survival methods, such as the Cox proportional hazards model [19]. Moreover, it is typical in many biomedical settings to observe a relatively low event rate (roughly 10%), so it can be challenging and computationally restrictive to consider so many covariates.

Initial methods tried to approximate PLS using successive fits of Cox models [6] or even ignoring censorship altogether and modeling survival time as a continuous outcome [54]. Once PLS was extended to generalized linear models (GLMs) [48], the Cox (partial) likelihood could be approximated by fitting a Poisson PLS [57]. In 2008, the GLM approach was modified for the survival setting by iteratively maximizing the Cox likelihood using weighted PLS approximations in place of weighted least squares (WLS) updates [55]. Other researchers went a step further to avoid the computational burden of an iterative procedure by approximating the WLS problem with a linear one (using deviance residuals from a Cox model without covariates) [4]. However, its estimates could be poor for individuals whose survival times are not well predicted by the proportional hazards model, the estimate of the baseline hazard in the null Cox model ignores all the covariate information, and computational capability is no longer as prohibitive as before. Therefore, we focus on methods that iteratively fit the Cox model, and we improve upon computational speed in other manners.

Unfortunately, there are relatively few methods that utilize sparsity in conjunction with PLS adaptations of the Cox model or GLMs. A simple option would be to perform a univariate variable selection procedure before fitting a Cox model (as in [10]), but this is not ideal since it does not account for any correlation between covariates. More refined methods

incorporate different forms of sparsity within the iterative maximization procedure [45, 16]. However, if most of the covariates are related to the outcome (even if weakly in some cases), then it seems reasonable to incorporate sparsity after the maximization method to make use of the entire covariate information when estimating the non-linear model.

Since it is plausible that many foods are (somewhat) related to heart-health, we propose a one-step sparse PLS (SPLS) summary of the (log) hazard ratios from either the Cox or PLS-approximated Cox model. We use Chun and Keles's SPLS procedure [14] to select the most relevant foods, as it provided an improvement in model parsimony in the linear context, as demonstrated in Chapter 2. Our proposal delineates a way to analyze the relationship between potentially correlated covariates and the time-to-an-event while using PLS. Specifically, our approach provides a lower-dimensional (PLS) summary of the log hazard ratios from a Cox model while incorporating variable selection. If implementing the PLS-approximated Cox model, one additionally make use of PLS as an optimization tool, which is particularly useful if there are many subjects or covariates.

To understand the applicable scenarios of our proposed approach and its connections with standard methods, we first discuss the implications of using WLS updates or PLS-approximations for the Cox model in Section 3.2.1. Next, we clarify the interpretation of PLS summaries of Cox models and describe different ways to integrate SPLS variable selection in Section 3.2.2. Last, we compare the performance of our proposed sparse methods with other sparse Cox methods via a simulation study in Section 3.3. Further simulation studies can also be found in Appendix D.4 and E.

3.2 Methodology

The notation convention we use is upper-case bold letters for matrices ($\mathbf{X}$), lower-case bold letters for vectors ($\mathbf{z}$), italic letters for scalars (c), and $\|\mathbf{w}\|_2 = \left(\sum_j |\mathbf{w}_j|^2\right)^{1/2}$ for the L_2 norm of $\mathbf{w}$. Consider the context of a univariate right-censored survival outcome; let $\mathbf{x}_i$ denote the i-th person's p covariates, t_i their event time, and c_i their censorship time, where $c_i \perp\!\!\!\perp t_i|\mathbf{x}_i$. Let $y_i = \min(t_i, c_i)$ be their observed time and $\delta_i = \mathbb{1}\{t_i \leq c_i\}$ indicate whether their event

was observed. Denote the set of observed events by $\tau_1, ..., \tau_{\tilde{n}}$. The observed data for n participants (with $n > \tilde{n} > p$) is contained in $\{\mathbf{X}, \mathbf{y}, \boldsymbol{\delta}\}$, with the aim of relating the time-to-an-event (with no tied times) to a function of the relevant and potentially highly-correlated centered predictors using adaptations of the Cox proportional hazards model [19].

3.2.1 Maximization techniques for the Cox partial likelihood

First, we review the standard approach to estimate the regression coefficients $\boldsymbol{\beta}$ using the (semi-parametric) Cox proportional hazards model. The hazard function takes the form $\lambda(t|\mathbf{x}_i) = \lambda_0(t)\ \exp(\mathbf{x}_i^T\boldsymbol{\beta})$, where $\lambda_0(t)$ is an underlying baseline hazard and the effect of the covariates on risk does not vary with time. The baseline hazard can be approximated non-parametrically using Breslow's estimator [11] $\widehat{\lambda}_0(\tau_k) = \left(\sum_{j \in R_k} e^{\mathbf{X}_j^T\boldsymbol{\beta}}\right)^{-1}$, where R_k is the set of individuals still at observable risk at time τ_k (i.e., $j\ :\ y_j \geq \tau_k$). Cox proposed to estimate the effect of the covariates without modeling the (nuisance) change in risk over time by using only a portion of the full log likelihood. The use of this log partial likelihood $\ell(\boldsymbol{\beta}) = \sum_{i=1}^n\ \delta_i\ \log\left[\exp(\mathbf{x}_i^T\boldsymbol{\beta})\ \widehat{\lambda}_0(y_i)\right]$ has been justified theoretically since then (see [3, 71]).

To obtain the maximum (log) partial likelihood estimator (MPLE) $\widehat{\boldsymbol{\beta}}$, the standard procedure is to use iterative Newton updates—specifically Fisher scoring. First, the Cox model's estimate of whether a person experienced an event at their observed time, given their covariates, is $\widehat{\mu}_i(\boldsymbol{\beta}) = \exp(\mathbf{x}_i^T\boldsymbol{\beta}) \sum_{j\ :\ \tau_j \leq y_i} \widehat{\lambda}_0(\tau_j)$. The sum estimates the cumulative baseline hazard. Next, given an estimate $\widehat{\boldsymbol{\beta}}_t$, the score and Hessian matrices are respectively $\mathbf{U}_t = \mathbf{X}^T\left(\boldsymbol{\delta} - \widehat{\boldsymbol{\mu}}_t\right)$ and $\mathbf{H}_t = -\mathbf{X}^T\,\mathbf{V}_t\,\mathbf{X}$, where $\mathbf{V}_t = \text{diag}\left(\widehat{\mu}_{t,i}\right)_{i=1}^n$ is the first order approximation of the working response's variance for computational speed [32]. Given an initial estimate $\widehat{\boldsymbol{\beta}}_0$, Iteratively Reweighted Least Squares (IRLS) can be used to maximize $\ell(\boldsymbol{\beta})$ using the Newton update

$$\widehat{\boldsymbol{\beta}}_{t+1} = \widehat{\boldsymbol{\beta}}_t\ +\ (-\mathbf{H}_t)^{-1}\,\mathbf{U}_t\ =\ (-\mathbf{H}_t)^{-1}\ \mathbf{X}^T\mathbf{V}_t\left[\mathbf{X}\widehat{\boldsymbol{\beta}}_t + \mathbf{V}_t^{-1}(\boldsymbol{\delta} - \widehat{\boldsymbol{\mu}}_t)\right]\ :=\ (-\mathbf{H}_t)^{-1}\ \mathbf{X}^T\mathbf{V}_t\,\mathbf{z}_t. \tag{3.1}$$

The standard Cox model will typically converge to a local minima in quadratic time if

the initial guess $\widehat{\boldsymbol{\beta}}_0$ is close to the true solution. However, if n or p is moderately large, it can be computationally costly to compute and invert the Hessian matrix at each iteration. In high-dimensional settings ($p > \tilde{n}$), it is infeasible to find the MPLE with the standard Cox model because the Hessian will not be invertible.

To facilitate obtaining the MPLE, we could solve a simpler approximation to the Newton update using a Truncated Newton method [53]. One such optimization technique is Conjugate Gradient (CG) Accelerated IRLS [25]. At each IRLS loop, CG—an algorithm that solves systems of linear equations—is used to obtain a close approximation to the Newton solution without taking the extra time to completely solve it. This can speed up the search particularly when the Hessian is hard or even impossible to invert. A version of this algorithm was shown to converge to the MPLE as long as the error tolerance was small enough [25]. Since Conjugate Gradient and PLS are equivalent (as proven in [60] and verified in Appendix A), this approach is one way to incorporate PLS into the Cox model, but it does not provide a PLS summary of $\boldsymbol{\beta}$.

If we allow for some shrinkage when estimating $\boldsymbol{\beta}$ in order to obtain a summary of reduced dimension, we could instead fix the number of PLS components used to approximate the Newton solution within each IRLS loop. Each iteration with this "Hessian-free" Newton update takes less time than the standard Newton update because it avoids inverting the Hessian matrix, enabling it to solve high-dimensional problems. Nygard proposed this method for the survival setting [55], which I will refer to as approximate MPLE (aMPLE) throughout the rest of this dissertation, though I specifically mean a fixed-rank PLS approximation. Alternatively, if one presumed that PLS components were the true covariate scale of interest (see [82] and Chapter 4 for more discussion), this would also be the appropriate method to use. Using k PLS components, the update within each IRLS loop can be expressed as

$$\widehat{\boldsymbol{\beta}}_{t+1}^{(k)} \;=\; \left(\mathbf{X}^T\mathbf{V}_t^{(k)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}_t^{(k)}\mathbf{z}_t^{(k)} \;-\; \boldsymbol{\xi}_t^{(k)} \;=\; \mathbf{R}_t^{(k)}\left(\mathbf{T}_t^{(k)T}\mathbf{V}_t^{(k)}\mathbf{T}_t^{(k)}\right)^{-1}\mathbf{T}_t^{(k)T}\mathbf{V}_t^{(k)}\mathbf{z}_t^{(k)}. \quad (3.2)$$

The label (k) indicates either that k components were used to construct that term or that

$\widehat{\boldsymbol{\beta}}_t^{(k)}$ was plugged in to attain that term. As in Chapter 2, the set of components is denoted by $\mathbf{T}_{n\times k}$, and $\mathbf{R}_{p\times k}$ is the oblique projection that transforms the regression coefficients from the component-scale to original $\mathbf{X}$-scale. The difference that arises from using fewer than $\text{rank}(\mathbf{X}^T\mathbf{V}_t^{(k)}\mathbf{X})$ components is denoted by $\boldsymbol{\xi}_t^{(k)}$, which will be discussed later. Under regularity assumptions, conditions on the PLS search directions, and appropriate use of a line search strategy, this type of Truncated Newton method will also converge to(wards) the MPLE [53]. Approximate MPLE makes use of PLS both as a shrinkage method and as a means to quicken the estimation using rank-constrained approximation.

However, if the covariates are not well approximated by the chosen number of components, aMPLE can converge to a poor estimate of $\widehat{\boldsymbol{\beta}}$. This can occur as with linear PLS if the predictors are over-simplified (e.g., too much dimension-reduction), resulting in overly attenuated coefficients [26, 60]. Additionally, due to the non-linearity of the survival model, certain aMPLE terms $\left(\left[\widehat{\mu}_i\left(\widehat{\boldsymbol{\beta}}^{(k)}\right)\right]^{-1} \text{ or } \mathbf{z}_i^{(k)}\right)$ can be noticeably different from their MPLE counterparts $\left(\left[\widehat{\mu}_i\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1} \text{ or } \mathbf{z}_i\right)$. This situation is more likely to occur if only a few components are being used, but this is also typically the objective of practitioners to enable easier interpretation of results. Appendix D.1 further characterizes when aMPLE is a good approximation to MPLE for the Cox model.

As a formal overview of how we consider maximizing the Cox partial likelihood, denote the MPLE and aMPLE terms upon convergence by a hat (e.g., $\widehat{\mathbf{V}}$ or $\hat{\mathbf{z}}^{(k)}$). Then,

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left[-\ell(\boldsymbol{\beta})\right] = \arg\min_{\boldsymbol{\beta}}\left\|\widehat{\mathbf{V}}^{1/2}\left(\hat{\mathbf{z}} - \mathbf{X}^T\boldsymbol{\beta}\right)\right\|_2^2 \;\approx\; \arg\min_{\boldsymbol{\beta}\in\mathcal{K}_k}\left\|\widehat{\mathbf{V}}^{(k)1/2}\left(\hat{\mathbf{z}}^{(k)} - \mathbf{X}^T\boldsymbol{\beta}\right)\right\|_2^2.$$

Here $\mathcal{K}_k$ indicates the k-th order Krylov subspace that spans $\{\mathbf{s}, \mathbf{S}\mathbf{s}, \mathbf{S}^2\mathbf{s}, ..., \mathbf{S}^{k-1}\mathbf{s}\}$, with $\mathbf{s} = \mathbf{X}^T\widehat{\mathbf{V}}^{(k)}\hat{\mathbf{z}}^{(k)}$ and $\mathbf{S} = \mathbf{X}^T\widehat{\mathbf{V}}^{(k)}\mathbf{X}$. As mentioned in Chapter 2, this expanding affine subspace denotes the restricted space in which the PLS approximate solution lies, which is exemplified further in Appendix A.3.

3.2.2 Approaches to incorporate (sparse) PLS into the Cox model

We first describe how to incorporate PLS after maximizing the Cox partial likelihood. Then, we present Chun and Keles's SPLS procedure for non-linear models. Last, we investigate when to include SPLS in the modeling procedure for nutritional epidemiology settings.

PLS summaries of the (approximate) Cox model

If one were to use the standard Cox model to maximize the partial likelihood, one could subsequently perform a single PLS fit $\left(\text{using } \widehat{\mathbf{V}} \text{ and } \hat{\mathbf{z}}\right)$ to obtain a PLS summary of $\widehat{\boldsymbol{\beta}}$. If aMPLE were used instead, other researchers have opted to use the last iteration as the PLS summary [55, 16, 45]. However, because aMPLE needs to incorporate a line search strategy to satisfy convergence requirements, this allows for the possibility that the converged $\widehat{\boldsymbol{\beta}}^{(k)}$ is actually a weighted sum of multiple PLS fits (see Appendix D.2 for more details). Therefore, we perform a single PLS fit after aMPLE converges $\left(\text{using } \widehat{\mathbf{V}}^{(k)} \text{ and } \hat{\mathbf{z}}^{(k)}\right)$ to obtain a proper PLS summary.

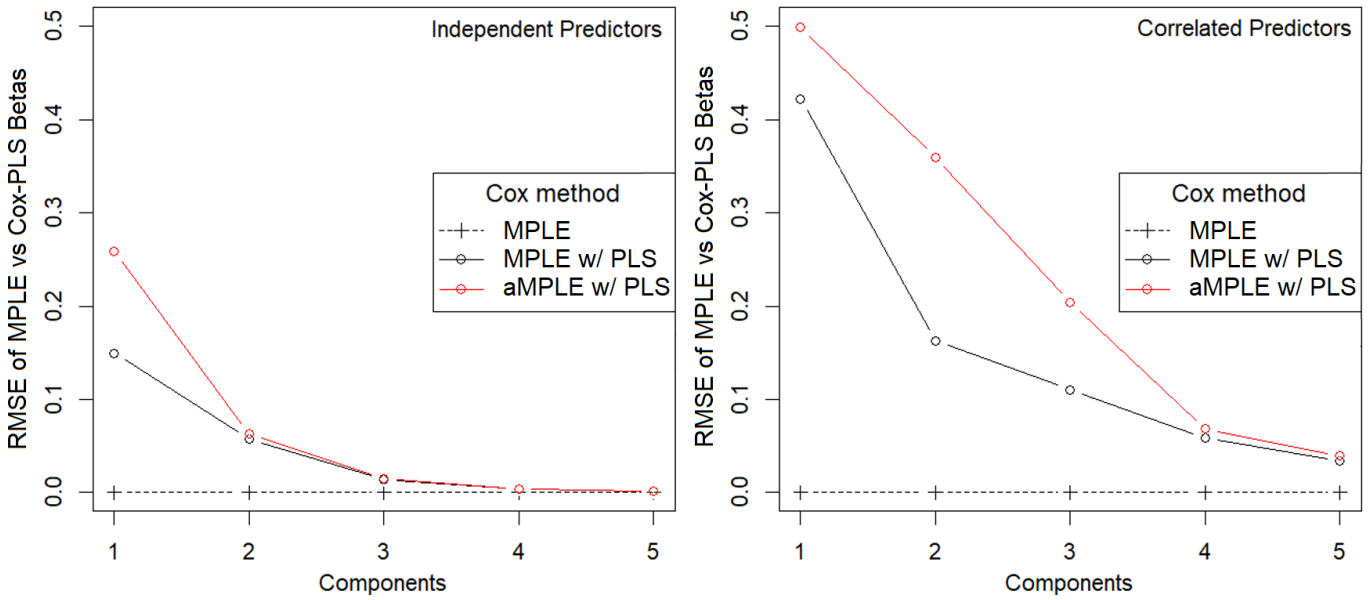
More formally, after maximizing the Cox partial likelihood $\left(\text{with converged } \breve{\mathbf{V}} \text{ and } \breve{\mathbf{z}}\right)$, a PLS summary refers to a k-component PLS model that minimizes the squared Pearson residuals (equivalent to the squared scaled working residuals). The PLS summary of (a)MPLE can be denoted by $\breve{\beta}^{(k)} = \arg\min_{\beta \in \mathcal{K}_k} \left\| \breve{\mathbf{V}}^{1/2} \left(\breve{\mathbf{z}} - \mathbf{X}^T\beta\right) \right\|_2^2$. Here $\mathcal{K}_k$ indicates the k-th order Krylov subspace with $\mathbf{s} = \mathbf{X}^T\breve{\mathbf{V}}\breve{\mathbf{z}}$ and $\mathbf{S} = \mathbf{X}^T\breve{\mathbf{V}}\mathbf{X}$. Table 3.1 highlights the different ways we consider using PLS in a Cox proportional hazards model. To differentiate between general maximization techniques and candidate Cox methods that potentially include dimension reduction, we denote the latter by this font.

Table 3.1: Cox methods with different types of PLS incorporation

Title	**Purpose of PLS in Cox method**
MPLE (standard Cox)	Not applicable
MPLE with PLS summary	Dimension reduction in summary
aMPLE with PLS summary	Fixed-rank approximation to MPLE and dimension reduction in summary

Overall, MPLE with PLS summary can be viewed as the attenuated[1] projection of MPLE, and aMPLE with PLS summary is an attenuated[1] version of MPLE with PLS summary (see Appendix D.3 for a more detailed visual comparison). Regardless of the number of components or covariate correlation structure, MPLE with PLS summary tends to result in coefficient magnitudes that are closest to MPLE. Figure 3.1 presents a simulated example with 24 predictors (with the same design as in Chapter 2) that demonstrates this trend. As a reminder, the coefficients across methods are equivalent when the full set of components are used (here, when k=24). Given these observations, MPLE with PLS summary is closer to the standard Cox fit that most researchers are familiar with. However, if MPLE is computationally costly or infeasible or if the priority is to model latent variables, aMPLE with PLS summary provides a good alternative (provided enough components are used) for a "Cox-PLS" method.

Figure 3.1: Differences between MPLE coefficients and PLS summaries of 24 covariates



[1]While the norm of the PLS solution will be smaller than the OLS solution, there could be a few predictors whose coefficients are actually magnified. However, especially when only a few components are used relative to p, most coefficients will be attenuated.

SPLS for non-linear models

We now present Chun and Keles's SPLS procedure in the context of a non-linear model (that can be solved by Weighted Least Squares) in Algorithm 1. The only difference from the linear setting is that the covariance $\mathbf{X}^T\mathbf{y}$ is replaced by an estimated weighted covariance (i.e., gradient) $\mathbf{X}^T\mathbf{V}\mathbf{z}$. We note that at the mth step, all m components are (re)constructed, not simply the mth component. For a visual representation of this procedure, Figure 3.2 displays an example of MPLE with a 3 component PLS summary and 80% sparsity. The truly relevant variables are indicated by gray-filled circles, so this model selected 7 true variables and 2 false variables.

Algorithm 1: Chun and Keles's Sparse PLS selection for a non-linear model

Input : covariates $\mathbf{X}$, weighting matrix $\mathbf{V}$, transformed response $\mathbf{z}$, k number of components, and sparsity percentage $\gamma \in [0, 1)$

Output: Index of SPLS selected covariates $\mathcal{A}$

Initialize $\widehat{\boldsymbol{\beta}}^{(0)} = 0$ and $\mathcal{A} = \{\}$

for $m = 1$ *to* k **do**

$$\sigma^{(m)} = \mathbf{X}^T\mathbf{V}\left(\mathbf{z} - \mathbf{X}_{\mathcal{A}}^T\,\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(m-1)}\right)$$

$$\mathcal{A} = \mathcal{A} \cup \left\{j \text{ s.t. } \left|\sigma_j^{(m)}\right| > \gamma \max_{1\le h\le p}\left|\sigma_h^{(m)}\right|\right\}$$

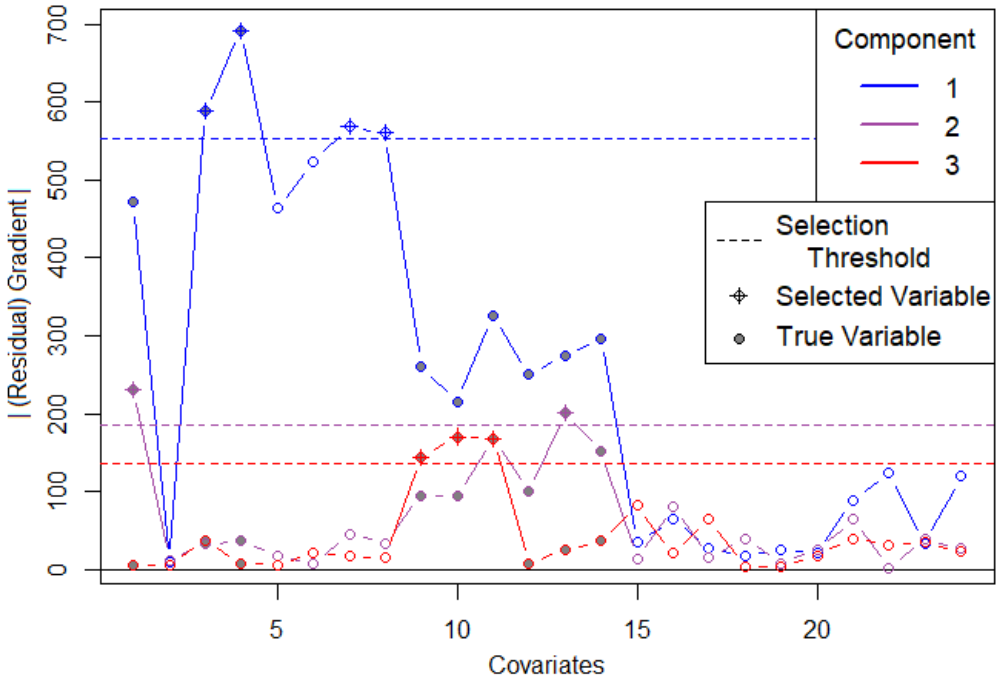
$$\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(m)} = \arg\min_{\boldsymbol{\beta}\in\mathcal{K}_m} \left\|\mathbf{V}^{1/2}\left(\mathbf{z} - \mathbf{X}_{\mathcal{A}}^T\,\boldsymbol{\beta}\right)\right\|_2^2 \equiv \mathrm{WPLS}\left(\mathbf{X}_{\mathcal{A}}, \mathbf{V}, \mathbf{z}, m\right)$$

end

return $\mathcal{A}$

Although this sparsity procedure is a bit more involved, it is better equipped to discard false variables than simply thresholding the original gradient magnitude. If that were the criterion, the procedure would be forced to falsely select variables 5–8 in Figure 3.2 before adding the other true variables. By adding a few variables at a time and subtracting the current predicted response out at each step, SPLS is able to select the variables that best describe the gradient via a set of independent weighted PLS components.

Figure 3.2: Variable selection for MPLE with a 3 component PLS summary and 80% sparsity



In Step 1 (blue), variables 3, 4, 7, and 8 are selected since their gradient magnitudes are at least 80% of the largest magnitude (dashed line). The information from a 1-component PLS model with the selected variables is subtracted from the gradient. In Step 2 (purple), variables 1 and 13 are chosen, since their residual gradient magnitudes are at least 80% of the largest residual magnitude. The information from an updated 2-component PLS model with all 6 selected variables is subtracted from the original gradient. In Step 3 (red), variables 9–11 are chosen; the final set of selected variables include 7 true and 2 false ones.

Options to incorporate SPLS variable selection into Cox-PLS

Next we present four ways to incorporate Chun and Keles's SPLS procedure into the Cox proportional hazards model in Algorithm 2, but these approaches can be applied to any generalized linear model by substituting the appropriate working response and weighting matrix. After sorting the survival data $\{\mathbf{X}, \mathbf{y}, \boldsymbol{\delta}\}$ by time $\mathbf{y}$ and centering (and potentially scaling) $\mathbf{X}$ by columns, we find the (a)MPLE using IRLS iterations until the log partial likelihood converges. In addition, we have the option to use SPLS sparsity during or after the maximization procedure.

Algorithm 2: Four approaches to incorporate Sparse PLS for a survival outcome

Input : sparsity percentage $\gamma \in [0, 1)$, k components, covariates $\mathbf{X}$, observed times $\mathbf{y}$, and indicators for observed events $\boldsymbol{\delta}$, SPLS.after, MPLE, and aMPLE

Output: SPLS regression coefficient $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(k)}$

Initialize $\boldsymbol{\beta}_0 = 0$, tol=1, and t=0.

while $tol > \epsilon$ **do**

$$\widehat{\mu}_{i,t}(\boldsymbol{\beta}_t) = \exp(\mathbf{X}_i^T \boldsymbol{\beta}_t) \sum_{j \,:\, \tau_j \leq y_i} \widehat{\lambda}_0(\tau_j) \qquad \text{and} \qquad \mathbf{V}_t = \text{diag}\left(\widehat{\mu}_{i,t}\right)_{i=1}^n$$

$$\mathbf{z}_t = \mathbf{X}\boldsymbol{\beta}_t + {\mathbf{V}_t}^{-1}\left(\boldsymbol{\delta} - \widehat{\boldsymbol{\mu}}_t\right)$$

$$\mathcal{A} = \mathbb{1}\{\text{SPLS.after}\} * \{1, ..., p\} \;\; + \;\; (1 - \mathbb{1}\{\text{SPLS.after}\}) * \text{SPLS}\left(\mathbf{X}, \mathbf{V}_t, \mathbf{z}_t, k, \gamma\right)$$

$$\breve{\boldsymbol{\beta}} = \mathbb{1}\{\text{MPLE}\} * \left(\mathbf{X}_{\mathcal{A}}^T \mathbf{V}_t \mathbf{X}_{\mathcal{A}}\right)^{-1} \mathbf{X}_{\mathcal{A}}^T \mathbf{V}_t \; \mathbf{z}_t \;\; + \;\; \mathbb{1}\{\text{aMPLE}\} * \text{WPLS}\left(\mathbf{X}_{\mathcal{A}}, \mathbf{V}_t, \mathbf{z}_t, k\right)$$

$$\alpha_t = \arg\max_{\alpha \in [-\mathbb{1}\{\text{aMPLE}\},\; 1]} \left\{ \ell\left((1 - \alpha)\; \boldsymbol{\beta}_t \;+\; \alpha\; \breve{\boldsymbol{\beta}} \right) \right\}$$

$$\boldsymbol{\beta}_{t+1} = (1 - \alpha_t)\; \boldsymbol{\beta}_t \;+\; \alpha_t\; \breve{\boldsymbol{\beta}}$$

$$tol = |\ell(\boldsymbol{\beta}_{t+1}) - \ell(\boldsymbol{\beta}_t)| \qquad \text{and} \qquad t = t + 1$$

end

Let $\widehat{\mathbf{V}}$ and $\hat{\mathbf{z}}$ be the corresponding values upon convergence.

$$\mathcal{A} = \mathbb{1}\{\text{SPLS.after}\} * \text{SPLS}\left(\mathbf{X}, \widehat{\mathbf{V}}, \hat{\mathbf{z}}, k, \gamma\right) \;\; + \;\; (1 - \mathbb{1}\{\text{SPLS.after}\}) * \mathcal{A}$$

$$\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(k)} = \text{WPLS}\left(\mathbf{X}_{\mathcal{A}}, \widehat{\mathbf{V}}, \hat{\mathbf{z}}, k\right)$$

return $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(k)}$

This leads to four potential alternatives to incorporate SPLS sparsity within Cox-PLS, depending on whether MPLE or aMPLE is fit and when sparsity is applied. As we have found that these methods have similar predictive ability by the third PLS component, we focus on their variable selection and estimation error to assess their properties and relative advantages. Table 3.2 presents these observations and acronyms for each method. For simplicity, we often refer to the methods by their maximization label, which is composed of the following pieces: [if applicable, author label of sparsity procedure + "s-"] + [Newton-type update]. Unless otherwise specified, SPLS refers to Chun and Keles's version.

If the majority of included predictors are believed to be unrelated to the outcome (as

Table 3.2: Summary of sparse Cox-PLS methods using the SPLS procedure

Maximization label	**Full method name**	**Most advantageous setting and aim**
MPLE	MPLE (standard Cox) with SPLS summary	*Setting:* many weakly relevant variables. *Aim:* select almost all true variables, discard more false variables, and apply less shrinkage. *Summary:* MPLE more favorable in selection and estimation error across settings.
aMPLE	Approximate MPLE with SPLS summary	
CHs-MPLE	Chun's sparsity within MPLE with PLS summary	*Setting:* many noise variables with some correlation. *Aim:* select all true variables. *Summary:* CHs-MPLE induces less shrinkage. CHs-aMPLE discards more false variables.
CHs-aMPLE	Chun's sparsity within approximate MPLE with PLS summary	

in certain -omics settings like [42]), there could be a preference to exclude them from the maximization procedure in order to avoid introducing noise into $\widehat{\mathbf{V}}$ and $\hat{\mathbf{z}}$. However, if most of the predictors are related to the outcome but some are only weakly related, it would seem reasonable to use all of the covariates to precisely estimate $\mathbf{V}$ and $\mathbf{z}$, and then identify the strongly-related predictors in the SPLS summary. This hypothesis was confirmed after testing simulation settings with 80 noise or weakly-relevant predictors in addition to the 24 multiblock predictors generated as in Chapter 2.

Across all tested settings, MPLE was a strong contender, regardless of covariate correlation structure. Not only did it induce the least amount of shrinkage, leading to less estimation error, but it selected the fewest or close to the fewest false variables while selecting almost all true variables. Therefore, we suggest using MPLE with SPLS summary as the preferred Cox-PLS method with SPLS sparsity, followed by aMPLE with SPLS summary when computation prohibits the former fit. Full results are presented in Appendix D.4, with the summary included in the last column of Table 3.2.

For a formal review of our proposed approach, MPLE with SPLS summary applies Chun and Keles's SPLS procedure to select the variables that describe the most estimated gradient information ($\mathbf{X}^T\widehat{\mathbf{V}}\hat{\mathbf{z}}$) via a set of independent weighted PLS components. The subset of variables $\mathbf{X}_{\mathcal{A}}$ is chosen based on a given sparsity percentage $\gamma \in [0, 1)$ and k number of

components. Starting with $\widehat{\boldsymbol{\beta}}^{(0)} = 0$ and indexing set $\mathcal{A}_0 = \{\}$, for $m \in [1, k]$,

$$\mathcal{A}_m = \mathcal{A}_{m-1} \cup \left\{ j \ : \ \left| \mathbf{X}_j^T \ \widehat{\mathbf{V}} \ \left[\hat{\mathbf{z}} - \mathbf{X}_{\mathcal{A}_{m-1}}^T \widehat{\boldsymbol{\beta}}_{\mathcal{A}_{m-1}}^{(m-1)} \right] \right| \ > \ \gamma \ \max_{1 \le h \le p} \left| \mathbf{X}_h^T \ \widehat{\mathbf{V}} \ \left[\hat{\mathbf{z}} - \mathbf{X}_{\mathcal{A}_{m-1}}^T \widehat{\boldsymbol{\beta}}_{\mathcal{A}_{m-1}}^{(m-1)} \right] \right| \right\}$$

where $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_{m-1}}^{(m-1)}$ is the coefficient from an $(m-1)$-component PLS model using the covariates $\mathbf{X}_{\mathcal{A}_{m-1}}$. The final selection of variables $\mathcal{A} = \mathcal{A}_k$ includes those covariates that either have the strongest relationship with the hazard of an event or that best capture the variation among covariates when transformed into k independent linear combinations. This procedure then fits a K-component PLS model with the final selected variables. As such, the sparse PLS summary of MPLE can be denoted by $\widehat{\beta}_{\mathcal{A}}^{(k)} = \arg\min_{\beta \in \mathcal{K}(\mathcal{A})_k} \left\| \widehat{\mathbf{V}}^{1/2} \left(\hat{\mathbf{z}} - \mathbf{X}_{\mathcal{A}}^T \ \beta \right) \right\|_2^2$. Here $\mathcal{K}(\mathcal{A})_k$ indicates the k-th order Krylov subspace with $\mathbf{s} = \mathbf{X}_{\mathcal{A}}^T \widehat{\mathbf{V}} \hat{\mathbf{z}}$ and $\mathbf{S} = \mathbf{X}_{\mathcal{A}}^T \widehat{\mathbf{V}} \mathbf{X}_{\mathcal{A}}$.

We briefly mention the only other existing approach that includes sparsity in a Cox-PLS model: LEs-aMPLE (Lee's sparsity within approximate MPLE) [45]. Within each IRLS loop, rank-constrained approximation is accomplished by using penalized PLS components. Specifically, a random effect penalty shrinks the PLS direction vectors (e.g., PLS weights $\mathbf{W}$ from Chapter 2) via a mean-0 Gaussian distribution, with its variability composed of a sparsity parameter and a mean-1 Gamma random variable. When a covariate's direction vector is 0, it is not included in that particular component, but it can be considered for other components. As seen in Chapter 2, though, component-wise sparsity does not guarantee sufficient overall sparsity. We note that we expect LEs-aMPLE to perform similarly to our proposed aMPLE with SPLS summary because of the related maximization procedure.

3.3 Simulation studies

3.3.1 Study design

To evaluate the sparse and non-sparse Cox-PLS methods in the context of nutritional epidemiology, we generated predictors that exhibit similar patterns of correlation as foods do in MESA. The predictor matrix $\mathbf{X}_{n \times 24}$ ($n \in \{1000, 4000, 7000\}$) had four blocks, where variables in the same block were more similar to each other to imitate food groups, and we

varied the strength of these covariances. We typically present results from the independent covariate setting and a correlated setting when intra-block covariances are {.05, .5, .3, .1}, respectively, though other settings are included in Appendix E. Only 9 of the 24 predictors and slight random Gaussian noise contributed to survival, with the true $\boldsymbol{\beta}$'s either unit positive or negative. Exponential survival times were generated according to [8] such that roughly a 10% event rate and 15% censorship rate were observed at the end of the study period, which was informed by MESA CVD event data. Across 100 generated data sets, we split our observations into a training (70%) and test (30%) set, and we used 5-fold CV on the training data to select the tuning parameters for each method. We used the k-fold approximation [74] of the CV log-partial likelihood (CVLL) [75] to compute our deviance criteria (-2*CVLL). All splits of the data were balanced by the survival times (based on its quartiles) using the `caret` package in R in order for results to be less sensitive to fold splits.

On the test data, we evaluated variable selection via the number of true variables (TP, out of 9) and false variables (FP, out of 15) selected, parsimony by the number of selected components (K, out of 5), model discrimination ability via Harrell's C-statistic [30], and explained response variability by a pseudo R^2 [56] $\left(\text{a modified version of Nagelkerke's } R^2\text{: } 1-\exp\left(\frac{-2(\ell_{\hat{\boldsymbol{\beta}}}-\ell_{\mathbf{0}})}{\#\text{events}}\right)\right)$. We also included the Lasso-Cox method [69] to compare its prediction-based variable selection with that of the sparse PLS methods.

3.3.2 Results

In Table 3.3, we compare the performance of five Cox methods on 100 simulated data sets with 4000 observations, a 10% event rate, and correlated predictors. The proposed sparse Cox-PLS methods selected the fewest false variables while fitting slightly fewer components than LEs-aMPLE, and they maintained both predictive ability and explained response

Table 3.3: Average simulation results across 100 data sets ($n = 4000$, $p = 24$), using minimum deviance to select tuning parameters.

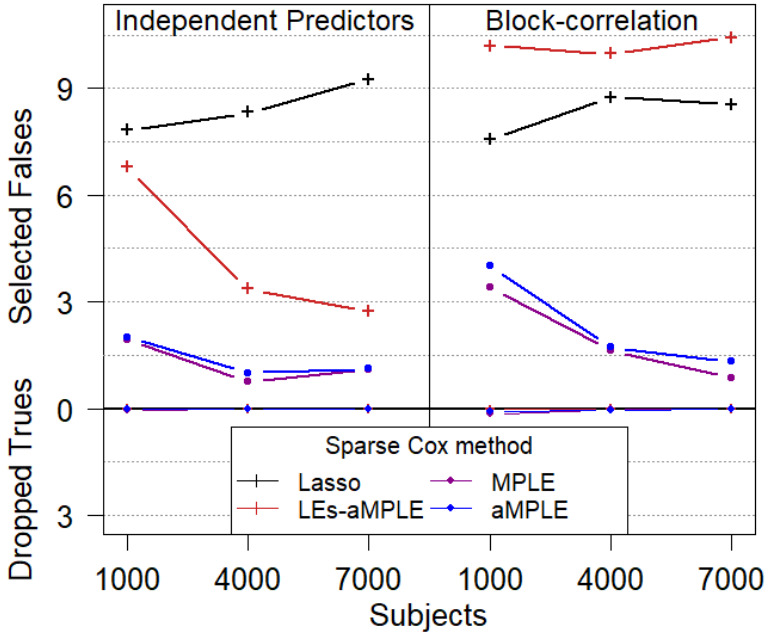
Cox Method	TP	FP	K	C	R^2
Lasso	9.0	8.8	1.0	0.92	0.96
LEs-aMPLE	9.0	10.0	4.8	0.91	0.96
MPLE w/ SPLS	9.0	1.6	4.1	0.92	0.96
aMPLE w/ SPLS	8.9	1.7	4.2	0.91	0.96
aMPLE w/ PLS	9.0	15.0	4.7	0.91	0.96

variability. The proposed sparse methods did drop true variables in a few data sets (2 with MPLE, 5 with aMPLE), but the average performance was quite similar to the other methods. The proposed methods without sparsity selected a similar number of components as LEs-aMPLE, and they also had similar C and R^2 levels. The most notable improvement that the proposed sparse methods provide over LEs-aMPLE is their ability to discard false variables, regardless of predictor correlation. Even in the simplest correlation setting to distinguish between relevant and irrelevant variables (independence; see Figure 3.3), the proposed methods outperformed the existing methods; Lasso picked 8.3 false variables and LEs-aMPLE selected 3.4, while the proposed sparse methods picked 0.8 (MPLE) and 1.0 (aMPLE), on average.

Across all the tested simulation scenarios (see Appendix E), LEs-aMPLE had consistently worse prediction performance, based on either deviance (on the test sets) or the C-statistic. Our proposed sparse methods had a slightly higher deviance than Lasso did when predictors were independent, while the opposite held when predictors had varying block covariances. Though the C-statistic did not change much between methods, the proposed PLS methods and LEs-aMPLE sometimes had a .01 lower C-statistic than the rest of the methods, particularly when the predictors had higher covariances or the sample size was small. Likewise, most methods explained a similarly high amount of the response variability within each simulation setting ($\approx 95\%$), with the largest difference occurring with low events; here, Lasso and MPLE with SPLS had a 1% higher pseudo R^2 than the three other Cox methods.

Figure 3.3 demonstrates the variable selection performance of the sparse methods across sample sizes and correlation settings with a 10% event rate. With the goal of discarding as few true variables and selecting as few false variables as possible, an ideal method would have both values close to the "0" line in the plot. Most methods selected the 9 true variables across all simulations, with the only exceptions being the proposed sparse methods discarding a fraction of a true variable, on average, when there were fewer events and/or more correlated predictors (as seen in the figure). In terms of the 15 false variables, Lasso actually increased the number of false selections as sample size increased, but its performance was not affected much by the correlation structure of the predictors. LEs-aMPLE did perform competitively

Figure 3.3: Incorrect variable selection performance of sparse Cox methods for 100 data sets

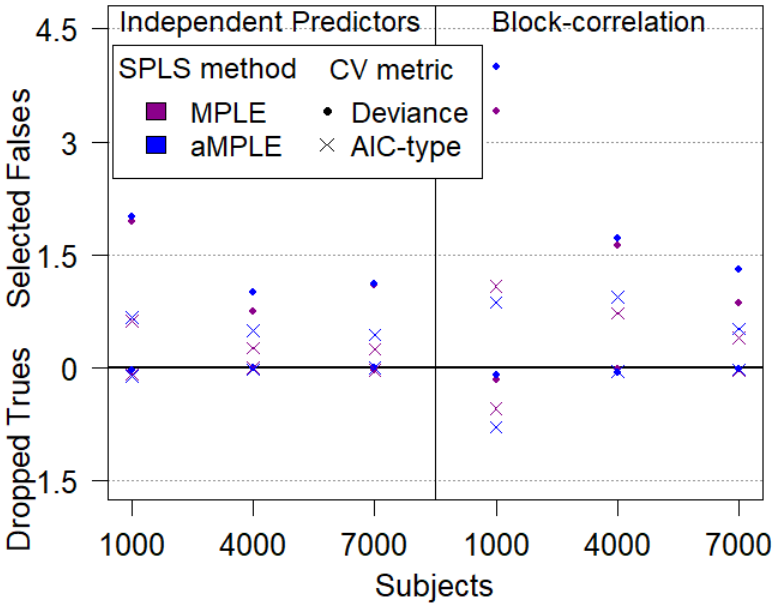


for independent predictors and more events ($\approx$ 3 false variables), but it quickly added false variables for predictors with even a small amount of correlation. The proposed sparse PLS methods, on the other hand, maintained the false variable selection quite low ($\approx$ 1.5 on average), with a slight increase for correlated predictors or low sample sizes. If there exist sufficient events per predictor variable in a data set, the proposed sparse methods provide a noticeable advantage in discarding false variables while still selecting all truly relevant ones.

To choose an even sparser model that is still reasonably-predictive (with the aim of systematically discarding more false variables), we considered changing the CV criterion to the 1-se deviance test for all methods. We additionally considered a criterion similar to AIC [2] for our proposed methods, which penalizes for the number of predictors in the model: -2*CVLL + 2*(# predictors). Lasso benefited the most from its 1-se test without impacting its overall performance, enabling it to select only 2 false variables on average. LEs-aMPLE, on the other hand, only tended to discard up to 1 or 2 more false variables when using the

1-se test[2], with a slight loss of predictability and R^2 in lower event settings. Although our proposed methods also discarded up to 1-2 more false variables with the 1-se or AIC-type test, their already low false selection rates makes this feat notable. Figure 3.4 demonstrates the differences in variable selection when using the AIC-type criterion[3] and (minimum) deviance criterion across sample sizes. With the AIC-type criterion, false variable selection

Figure 3.4: Incorrect variable selection when using different CV criterion on 100 data sets



was never greater than 1.5, on average, but sufficient events-per-variables were needed for either parsimony test to select all the true variables (e.g., $\tilde{n}/p \geq 5$ or 10 [76]) if correlation was present. If this condition was met, AIC selected a more (accurate) parsimonious model than the minimum deviance criterion in all settings, whereas the 1-se test did not outperform the minimum deviance criterion when predictors had mixed intra-block covariance. Therefore, we suggest using the AIC-type criterion with our proposed sparse methods if there are a

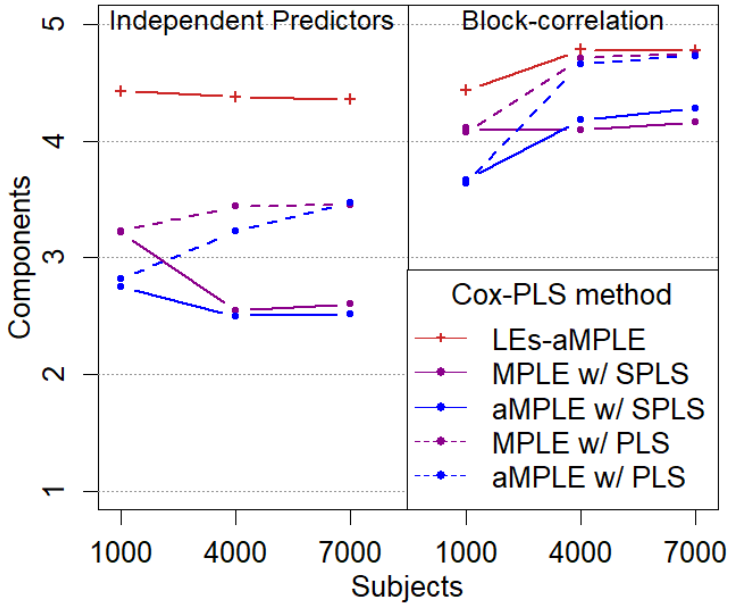
[2]Indeed, the prespecified LEs-aMPLE threshold to determine negligible coefficients had a larger impact on discarding false variables than the 1-se test did. Although the authors used a threshold of 5e-5 in [45], we used 5e-4 to give it a more reasonable advantage and be comparable to our proposed methods.

[3]We call this an AIC-type criterion because with PLS models, one is actually using less information than the number of selected predictors but more information than the number of selected components. More details on approximating the degrees of freedom in a PLS model can be found in [59].

sizable number of events and accurate variable selection is a motivating consideration.

With respect to the number of components selected (out of 5), our proposed sparse PLS methods (≈ 3.5) were generally more parsimonious than our proposed PLS methods (≈ 4), both of which were more parsimonious than LEs-aMPLE (≈ 4.5) in all tested settings. Figure 3.5 demonstrates the component selection performance of the Cox methods across sample sizes with a 10% event rate and varying predictor correlation. In general, all PLS methods selected more components when predictors were more correlated. Additionally, the proposed methods tended to select slightly more components as events increased, whereas LEs-aMPLE was less impacted by event sizes. While the proposed sparse methods used all selected variables in each of their components, LEs-aMPLE tended to use most of the selected variables in the first two components and then fewer than half of the selected variables in the remaining components, allowing for a different type of interpretability (e.g., some component-wise sparsity). However, this also explains why more LEs-aMPLE components are needed than the proposed SPLS components for a reasonably predictive model.

Figure 3.5: Component selection performance of Cox-PLS methods for 100 data sets



Overall, the proposed sparse Cox-PLS methods exhibited an impressive reduction in false variables selected, as SPLS did in the continuous setting of Chapter 2. Though LEs-aMPLE performed competitively when predictors were independent, it also tended to select more components than the proposed methods in all tested settings. With a sufficient number of events per predictor variables, MPLE w/ SPLS can result in a parsimonious summary of the relationship between the most relevant potentially-correlated predictors and survival time.

3.4 Discussion

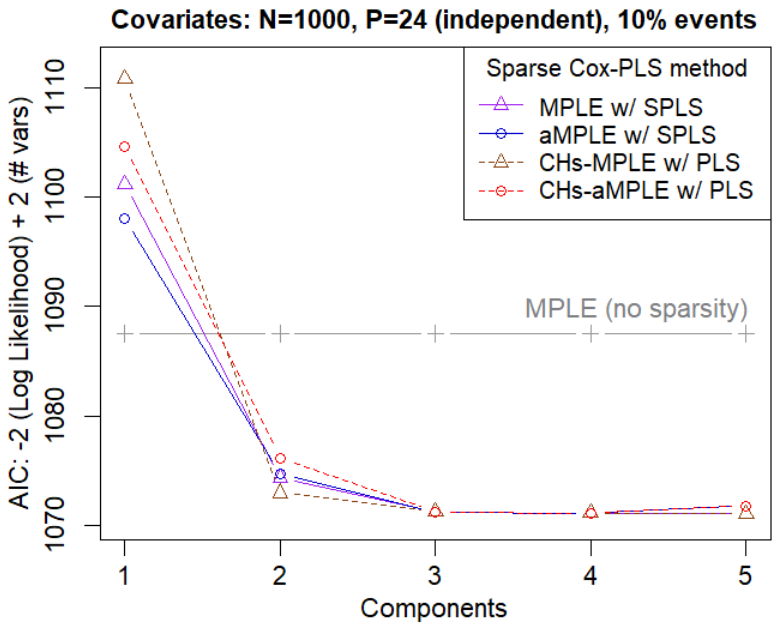
Although the Cox proportional hazards model is useful in many settings, there are times when practitioners prefer to summarize predictors via components, whether because there are a vast number of predictors that they want to simplify while still maintaining most of the relevant covariance information or because they believe that there are truly a few underlying factors of interest. As such, there have been many previous attempts to apply PLS to survival data, whether by approximating survival time [54, 57] or engaging with the Cox MPLE in some manner [6, 55, 4]. Expanding upon methods that directly estimate the MPLE using IRLS, we found that MPLE with PLS summary provided a reasonable dimension-reduced transformation of the Cox MPLE. The aMPLE with PLS summary allowed for faster computation and tractable solutions for high-dimensional data, but it did rely on the inclusion of enough components for reasonable PLS approximations.

When many covariates are presumed to be irrelevant, researchers have recommended performing a variable selection step before fitting Cox-PLS models (such as choosing the highest-ranked variables based on univariate score test p-values [55]), but very few methodological developments have focused explicitly on this goal. We examined different ways to incorporate Chun and Keles's SPLS procedure, anticipating an improvement in model parsimony as was seen in the linear context in Chapter 2. We found that MPLE with SPLS summary discarded more false variables than the existing LEs-aMPLE (and Lasso-Cox) while maintaining a similar predictive and explanatory ability, selecting almost all true variables, and choosing typically fewer components. Likewise, aMPLE with SPLS summary performed comparably, only dropping a fraction more a true variable under certain settings.

Though the SPLS authors extended their method to generalized linear models by incorporating SPLS into the maximization procedure (e.g., CHs-aMPLE) [16], our investigations in Section 3.2.2 indicate that using sparsity after the maximization procedure is more appropriate when many of the covariates are weakly related to the outcome. Apart from selecting fewer false variables and inducing less estimation shrinkage across a variety of settings, MPLE

with SPLS summary also provides a computational advantage. The maximization procedure is the most computationally expensive part—especially for relatively large n and more than 3 PLS components—even after using `C++` to perform matrix multiplications. However, to select the most appropriate SPLS model, we would ideally like to test a fine grid of sparsity percentages and components. As demonstrated in Figure 3.6, selecting 1–2 sparse PLS components *a-priori* for non-linear models can lead to unnecessarily poor predictive performance. Methods that apply sparsity after maximization can use the same (a)MPLE fit instead of having to fit one model for each tuning candidate pair within cross-validation, allowing for less burdensome testing.

Figure 3.6: Example of predictive ability of sparse Cox-PLS methods, across components



Comparing our proposed sparse methods with LEs-aMPLE, we observe that (as in Chapter 2) sparsity within components does not adequately promote variable selection, even if it does provide a different type of interpretability. Although LEs-aMPLE did recognize that false variables should have a much smaller magnitude coefficient than true variables in simple settings, it did not completely drop them as our proposed methods did. While one could

simply drop covariates with small magnitudes (above the already pre-specified threshold), it comes at the risk of starting to drop some true variables, particularly when covariates are correlated. On a different note, the distributional assumptions of LEs-aMPLE needed to impose sparsity may not be reasonable, and they require one more tuning parameter than our proposed sparse methods (for more details on the random effect penalty, see [46]).

Nevertheless, there are still questions remaining regarding the performance of (a)MPLE with SPLS summary. For aMPLE with SPLS summary (as with PLS in general), it is still unclear as to how many components are required to obtain a reasonable maximization estimate without relying on cross-validation. In fact, we restricted our search to 1-5 components in this chapter because most practitioners in our settings of interest prioritize interpretability, but more components could be considered if the goal is instead to obtain a close lower-rank approximation. Additionally, aMPLE with SPLS summary has the potential to be quite useful in high-dimensional settings, so an in-depth study comparing its performance to LEs-aMPLE and CHs-aMPLE with PLS summary would be needed. Further simulation studies would also be helpful to verify whether incorporating sparsity after the maximization step is advantageous for generalized linear models as well as to focus on data correlation structures that are common in other fields.

In conclusion, when the dominating goal of a survival data analysis is to identify strongly relevant covariates while incorporating the dimension reduction benefits of PLS, we propose using (a)MPLE with SPLS summary, depending on the covariate properties. If there are many events relative to the total number of covariates, one could use an AIC-type cross-validation criterion instead of deviance to promote even more model parsimony.

Chapter 4

COVARIATE ADJUSTMENT WHEN USING PLS TO SUMMARIZE OTHER COVARIATES

4.1 Introduction

In epidemiological studies where the primary goal is hypothesis generation or inference, an important consideration when estimating the association between an outcome and covariates of interest is the appropriate adjustment for potential confounders and precision variables, among others. When reducing "higher-dimensional" covariates ($\mathbf{X}_{n\times p}$; $n > p$ but p large) such as diet, it can be important to appropriately adjust for "low-dimensional" covariates ($\mathbf{L}_{n\times q}$; $n \gg q$), like demographics. In regression analyses without component-based dimension reduction, adjustment is straightforward, as we can include all covariate types in the model. However, it is not as apparent when and how to adjust for covariates when the motivating goal is to reduce the dimension of only one set of variables, particularly in the context of non-linear models. This can be particularly tricky for partial least squares (PLS), as its coefficients are highly dependent on which (residual) outcome is used in the model. A thorough study of covariate adjustment in the context of PLS would clarify best practices.

With recent technological and computational advances, there has been a similar interest in considering how to develop proper clinico-genomic models, where the goal is to understand how the combination of low-dimensional clinical covariates and "high-dimensional" genomic variables (or other -omic data; $\mathbf{H}_{n\times r}$; $n < r$) contribute to complex diseases [22, 47]. Maturana de Lopez *et al.* presented a helpful classification for data integration strategies: independent, conditional, and joint [47]. Briefly, "independent" means that each covariate type is regressed separately on the outcome, and the resulting selected or reduced variables from each are combined into a final model. While this is the simplest approach to implement,

it does not consider any correlation between the data types. On the other end, "joint" refers to applying selection and/or dimension reduction to both sets of covariates simultaneously. While this strategy does account for correlation, it is less applicable to our epidemiological setting of interest where we often have a predetermined set of low-dimensional covariates that we want to adjust for, even if they are somewhat correlated.

Thus, we focus on "conditional" methods, in which we have a pre-specified set of low-dimensional covariates, and we want to incorporate a function of the high-dimensional variables. We propose a further sub-classification of this category: 1-stage and 2-stage. "Two-stage-conditional" refers to fitting the low-dimensional (independent) model first and then modeling the high-dimensional variables on the residuals. "One-stage-conditional" means that we combine the process of modeling (e.g., shrinking and/or selecting variables from) the high-dimensional variables in conjunction with using the pre-specified low-dimensional variables to obtain a final model. Note that this differs from "joint" adjustment because there is no model decision-making occurring for the low-dimensional variables.

We focus on the setting when we have "low-dimensional" covariates that we want to adjust for, such as confounders or precision variables, while reducing the dimension of "higher-dimensional" covariates in a "conditional" model. As we will demonstrate, it turns out that the purpose for using PLS informs the type of adjustment to undertake. Our aim is to identify a reasonable approach to summarize the relationship ($\boldsymbol{\beta}$) between $\mathbf{X}$ and an outcome using sparse PLS (SPLS) while accounting for $\mathbf{L}$. We highlight that the priority is on preserving the parsimonious properties of Chun and Keles's SPLS [14] procedure that were observed in Chapters 2 and 3.

Other studies that have considered data integration for clinico-genomic models have focused on the prognostic (e.g., predictive) benefit that results from using multiple types of data [22]. For instance, a comparison of supervised and unsupervised methods applied to three survival data sets found that clinico-genomic models outperformed purely genomic models, while they typically resulted in insubstantial to moderate predictive improvement over the clinical model [10]. Others have highlighted the classification and predictive im-

provements (for binary outcomes) when using "conditional" PLS and principal component regression models as opposed to either clinical or genomic models in both simulations and real data [7]. However, our main interest is not to enhance predictive performance. Moreover, in many epidemiological settings, the adjustment covariates $\mathbf{L}$ are already well-established (and usually much more relevant than $\mathbf{X}$). We focus instead on providing a reasonable and parsimonious summary of $\boldsymbol{\beta}$ that can lead to hypothesis generation and can be interpreted as being adjusted for $\mathbf{L}$. A comprehensive study and evaluation of current adjustment approaches would allow for more informed modeling by practitioners as well as allow for a better understanding of the connection between common regression coefficients and those derived from PLS dimension-reduction methods.

We first provide a framework for covariate adjustment when using either the original covariates or PLS components in a multiple (weighted) least squares (LS) regression model in Section 4.2. Next, we present the model implications and interpretations of various conditional adjustment schemes for a continuous outcome in Section 4.3. We extend these lessons to the survival setting in Section 4.4. Apart from describing the model implications of different 1-stage-conditional approaches, we connect and extend 2-stage-conditional adjustment techniques for survival data. To evaluate how the performance of the top conditional adjustment approaches compare to current sparse Cox-PLS methods, we conduct a simulation study in Section 4.5 that varies the correlation between and relative importance of $\mathbf{L}$ and $\mathbf{X}$. We focus on which approach results in better variable selection, closer $\boldsymbol{\beta}$ estimates, and reasonable predictive ability. Last, we apply the principles learned to our analysis of the relationship between the time to a first cardiovascular disease (CVD) event and Food Frequency Questionnaire data in the Multi-Ethnic Study of Atherosclerosis (MESA) in Section 4.6.

4.2 Framework for statistical adjustment with PLS

To analyze the relationship between an outcome and a set of covariates of interest $\mathbf{X}$ while adjusting for a set of covariates $\mathbf{L}$, one can simply fit a multiple (weighted) linear regression, which I will refer to as "statistical adjustment". We begin by presenting three different

ways to obtain the resulting statistically-adjusted coefficients. If using iterative Newton updates with diagonal weighting matrix $\mathbf{V}_t$ and estimated mean $\boldsymbol{\mu}_t$, the t-th update for the working response takes the form $\mathbf{z}_t = \mathbf{X}\boldsymbol{\beta}_{t-1} + \mathbf{L}\boldsymbol{\gamma}_{t-1} + \mathbf{V}_t^{-1}(\mathbf{y} - \boldsymbol{\mu}_t)$. At each iteration, let $\mathcal{L} = \mathbf{L}(\mathbf{L}^T\mathbf{V}_t\mathbf{L})^{-1}\mathbf{L}^T\mathbf{V}_t$ be the generalized (e.g., weighted) projection of $\mathbf{L}$, which is idempotent but not symmetric, and similarly define $\mathcal{X}$ and $\mathcal{T}$. In linear regression (setting $\mathbf{V} = \mathbf{I}$), $\mathcal{L}$ is an orthogonal projection that obtains the linear predictor of the outcome associated with $\mathbf{L}$: $\mathcal{L}\mathbf{z} = \mathbf{L}\ (\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{z} = \mathbf{L}\widehat{\boldsymbol{\gamma}}_0$. By using properties of matrix blockwise inversion and assuming that both $\mathbf{X}^T\mathbf{V}\mathbf{X}$ and $\mathbf{L}^T\mathbf{V}\mathbf{L}$ are invertible, we can express the coefficient estimates from a multiple weighted least squares (WLS) regression as follows, given a weighting matrix $\mathbf{V}$ and outcome $\mathbf{z}$:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{V}\mathbf{X} & \mathbf{X}^T\mathbf{V}\mathbf{L} \\ \mathbf{L}^T\mathbf{V}\mathbf{X} & \mathbf{L}^T\mathbf{V}\mathbf{L} \end{bmatrix}^{-1} = \begin{bmatrix} [\mathbf{X}^T\mathbf{V}\mathbf{X} - \mathbf{X}^T\mathbf{V}\mathcal{L}\mathbf{X}]^{-1} & \mathbf{0} \\ \mathbf{0} & [\mathbf{L}^T\mathbf{V}\mathbf{L} - \mathbf{L}^T\mathbf{V}\mathcal{X}\mathbf{L}]^{-1} \end{bmatrix} * \begin{bmatrix} \mathbf{I} & -\mathbf{X}^T\mathbf{V}\mathbf{L}(\mathbf{L}^T\mathbf{V}\mathbf{L})^{-1} \\ -\mathbf{L}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1} & \mathbf{I} \end{bmatrix},$$

$$\text{so} \quad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{V}\mathbf{X} & \mathbf{X}^T\mathbf{V}\mathbf{L} \\ \mathbf{L}^T\mathbf{V}\mathbf{X} & \mathbf{L}^T\mathbf{V}\mathbf{L} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^T\mathbf{V}\mathbf{z} \\ \mathbf{L}^T\mathbf{V}\mathbf{z} \end{bmatrix} = \begin{bmatrix} [\mathbf{X}^T\mathbf{V}(\mathbf{I} - \mathcal{L})\mathbf{X}]^{-1}\ \mathbf{X}^T\mathbf{V}(\mathbf{I} - \mathcal{L})\mathbf{z} \\ [\mathbf{L}^T\mathbf{V}(\mathbf{I} - \mathcal{X})\mathbf{L}]^{-1}\ \mathbf{L}^T\mathbf{V}(\mathbf{I} - \mathcal{X})\mathbf{z} \end{bmatrix}.$$

These statistically-adjusted coefficients can be obtained two other ways using separate regressions. First we introduce the concept of "orthogonalizing", which refers to projecting a block of covariates off of the column space of the other set of covariates (e.g., $(\mathbf{I} - \mathcal{L})\mathbf{X}$). Remark 1 outlines three equivalent ways to obtain the coefficient estimates from a multiple weighted least squares model, with the proof given in Appendix F. I use the convention `model.fit[` outcome $\sim$ covariates `]`, where weighting by matrix $\mathbf{V}$ is implied.

Remark 1. *Given a weighting matrix $\boldsymbol{V}$ and outcome $\boldsymbol{z}$, statistical adjustment via a multiple weighted least squares regression can be accomplished by any of the following ways:*

1. ***Concurrent fit:*** $\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{pmatrix} = \texttt{WLS[}\boldsymbol{z}\sim \boldsymbol{X}+\boldsymbol{L}\texttt{]} = \begin{bmatrix}[\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{L})\boldsymbol{X}]^{-1}\ \boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{L})\boldsymbol{z}\\ [\boldsymbol{L}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{X})\boldsymbol{L}]^{-1}\ \boldsymbol{L}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{X})\boldsymbol{z}\end{bmatrix}$

2. ***Orthogonalized fits:*** $\widehat{\boldsymbol{\beta}} = \texttt{WLS[}\boldsymbol{z}\sim(\boldsymbol{I}-\mathcal{L})\boldsymbol{X}\texttt{]} = [\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{L})\boldsymbol{X}]^{-1}\ \boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{L})\boldsymbol{z}$
   $\widehat{\boldsymbol{\gamma}} = \texttt{WLS[}\boldsymbol{z}\sim(\boldsymbol{I}-\mathcal{X})\boldsymbol{L}\texttt{]} = [\boldsymbol{L}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{X})\boldsymbol{L}]^{-1}\ \boldsymbol{L}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{X})\boldsymbol{z}$

3. ***Residual fits:*** $\widehat{\boldsymbol{\beta}} = \texttt{WLS[}\boldsymbol{z}-\boldsymbol{L}\widehat{\boldsymbol{\gamma}}\sim\boldsymbol{X}\texttt{]} = (\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\ \boldsymbol{X}^T\boldsymbol{V}\left(\boldsymbol{z}-\boldsymbol{L}\widehat{\boldsymbol{\gamma}}\right)$
   $\widehat{\boldsymbol{\gamma}} = \texttt{WLS[}\boldsymbol{z}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\sim\boldsymbol{L}\texttt{]} = (\boldsymbol{L}^T\boldsymbol{V}\boldsymbol{L})^{-1}\ \boldsymbol{L}^T\boldsymbol{V}\left(\boldsymbol{z}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$.

Although it is clear that the concurrent regression coefficients are equivalent to those obtained by orthogonalizing the blocks of coefficients and regressing them separately on the outcome[1], the latter form allows for a partition between the sets of covariates if we want to solve the coefficients differently (e.g., only using PLS for $\mathbf{X}$). The statistically-adjusted coefficients can also be obtained by separate regressions without orthogonalization if the opposite adjusted coefficient is known. This can be shown by using matrix identities and properties of the inverse of a sum of matrices [34]. Note that $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}_0 - (\mathbf{L}^T\mathbf{V}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{V}\mathbf{X}\widehat{\boldsymbol{\beta}}$, so regressing $\mathbf{X}$ on the residual $\mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}}_0$ would result in a biased estimate unless $\mathbf{L}$ and $\mathbf{X}$ were orthogonal with respect to $\mathbf{V}$ (e.g., $\mathbf{L}^T\mathbf{V}\mathbf{X}=\mathbf{0}$).

Next, we present these same connections for statistical adjustment when the covariates of interest are PLS components constructed from $\mathbf{X}$. First, we review the mechanics of PLS in a weighted metric. To apply dimension-reduction to centered $\mathbf{X}$ when evaluating any outcome $\mathbf{z}$, PLS constructs weights $\mathbf{R}_{p\times k}$ that maximize $\mathbf{R}^T\mathbf{X}^T\mathbf{V}\mathbf{z}$. These weights are used to (obliquely) project $\mathbf{X}$ to form components $\mathbf{T}_{n\times k} = \mathbf{X}\mathbf{R}$ and coefficients[2] $\widehat{\mathbf{q}} = (\mathbf{T}^T\mathbf{V}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{V}\mathbf{z}$. To transform between predictor scales, it holds that $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{R}\widehat{\mathbf{q}}) = \mathbf{T}\widehat{\mathbf{q}}$.

Now, if one assumes that the PLS components are the true scale of interest, the ultimate goal would be to evaluate a regression model with both $\mathbf{T}$ and $\mathbf{L}$. The PLS com-

[1]We could obtain the same adjusted coefficients by regressing, for example, the orthogonalized $\mathbf{X}$ on a residual of $\mathbf{z}$ using any projection f of $\mathbf{L}$, since $(\mathbf{I}-\mathcal{L})[\mathbf{z}-f(\mathbf{L})] = (\mathbf{I}-\mathcal{L})\mathbf{z}$.

[2]For ease of presentation in this chapter, we change notation for the PLS coefficient from $\widehat{\mathbf{q}}^T$ to $\widehat{\mathbf{q}}$.

ponent should then be constructed using the outcome projected off of the column space of $\mathbf{L}$ (e.g., $(\mathbf{I}-\mathcal{L})\mathbf{z}$). Given weights $\mathbf{R}$ from this PLS fit, we can similarly express the adjusted coefficients by substituting $\mathbf{X}$ with $\mathbf{T}$: $\widehat{\mathbf{q}} = [\mathbf{T}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{T}]^{-1}\,\mathbf{T}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z}$ and $\widehat{\boldsymbol{\gamma}} = [\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{T})\mathbf{L}]^{-1}\,\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{T})\mathbf{z}$. These coefficients are also statistically-adjusted and can be obtained two other ways using separate regressions. Remark 2 outlines these approaches, with the proof given in Appendix F. The symbol $\longrightarrow \mathbf{T}$ signifies that we construct the component $\mathbf{T} = \mathbf{XR}$ using the corresponding weights from the indicated PLS fit.

Remark 2. *Given a weighting matrix $\boldsymbol{V}$ and outcome $\boldsymbol{z}$, statistical adjustment via a multiple weighted least squares regression with weighted PLS (WPLS) components $\boldsymbol{T}$ can be accomplished by any of the following ways:*

1. ***Concurrent fit:*** $\texttt{WPLS}[(\boldsymbol{I}-\mathcal{L})\boldsymbol{z} \sim \boldsymbol{X}] \longrightarrow \boldsymbol{T}$
$$\begin{pmatrix}\widehat{\boldsymbol{q}}\\ \widehat{\boldsymbol{\gamma}}\end{pmatrix} = \texttt{WLS}[\boldsymbol{z} \sim \boldsymbol{T}+\boldsymbol{L}] = \begin{bmatrix}[\boldsymbol{T}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{L})\boldsymbol{T}]^{-1}\;\boldsymbol{T}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{L})\boldsymbol{z}\\ [\boldsymbol{L}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{T})\boldsymbol{L}]^{-1}\;\boldsymbol{L}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{T})\boldsymbol{z}\end{bmatrix}$$

2. ***Orthogonalized fits:*** $\widehat{\boldsymbol{q}} = \begin{cases}\texttt{WPLS}[\boldsymbol{z} \sim (\boldsymbol{I}-\mathcal{L})\boldsymbol{X}] \longrightarrow \boldsymbol{T}\\ [\boldsymbol{T}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{L})\boldsymbol{T}]^{-1}\;\boldsymbol{T}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{L})\boldsymbol{z}\end{cases}$
$\widehat{\boldsymbol{\gamma}} = \texttt{WLS}[\boldsymbol{z} \sim (\boldsymbol{I}-\mathcal{T})\boldsymbol{L}] = [\boldsymbol{L}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{T})\boldsymbol{L}]^{-1}\;\boldsymbol{L}^T\boldsymbol{V}(\boldsymbol{I}-\mathcal{T})\boldsymbol{z}$

3. ***Residual fits:*** $\texttt{WPLS}[(\boldsymbol{I}-\mathcal{L})\boldsymbol{z} \sim \boldsymbol{X}] \longrightarrow \boldsymbol{T}$
$\widehat{\boldsymbol{q}} = \texttt{WLS}[\boldsymbol{z}-\boldsymbol{L}\widehat{\boldsymbol{\gamma}} \sim \boldsymbol{T}] = (\boldsymbol{T}^T\boldsymbol{V}\boldsymbol{T})^{-1}\;\boldsymbol{T}^T\boldsymbol{V}\left(\boldsymbol{z}-\boldsymbol{L}\widehat{\boldsymbol{\gamma}}\right)$
$\widehat{\boldsymbol{\gamma}} = \texttt{WLS}[\boldsymbol{z}-\boldsymbol{T}\widehat{\boldsymbol{q}} \sim \boldsymbol{L}] = (\boldsymbol{L}^T\boldsymbol{V}\boldsymbol{L})^{-1}\;\boldsymbol{L}^T\boldsymbol{V}\left(\boldsymbol{z}-\boldsymbol{T}\widehat{\boldsymbol{q}}\right)$.

Instead of fitting a WPLS and then subsequently performing a multiple WLS, the same adjusted coefficients can be efficiently obtained by orthogonalizing[3] each of the covariates before regressing them separately on the outcome $\mathbf{z}$. This holds because the PLS weights

[3]We briefly note that the equivalence between the concurrent fit and the orthogonalized fits for PLS does not hold for PCR because PCA is not a supervised method. PCA weights (e.g., loadings) that maximize $\mathbf{R}^T\mathbf{X}^T\mathbf{XR}$ are not equivalent to those maximizing $\mathbf{R}^T\mathbf{X}^T(\mathbf{I}-\mathcal{L})\mathbf{XR}$.

are maximizing the same quantity, since $\mathbf{X}^T\mathbf{V}[(\mathbf{I}-\mathcal{L})\mathbf{z}] = [\mathbf{X}^T(\mathbf{I}-\mathcal{L})^T]\mathbf{V}\mathbf{z}$. The main distinction[4] from the WLS case is that $\mathbf{L}$ should be orthogonalized off of $\mathcal{T}$ instead of $\mathcal{X}$.

The adjusted coefficients can also be obtained by separate regressions on residual outcomes after fitting a WPLS. Although it is less ideal to take three steps to achieve the desired coefficients, it is required to obtain the same values. If one instead tried to combine the first two steps by fitting a WPLS model using the residual $\mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}}$, the resulting PLS weight would be maximizing a different objective, since $\mathbf{X}^T\mathbf{V}[\mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}}] \neq \mathbf{X}^T\mathbf{V}[\mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}}_0]$.

In this section, we have presented three ways to obtain the same statistically-adjusted coefficients from a multiple WLS regression, regardless of whether one opts to use $\mathbf{X}$ or $\mathbf{T}$ as the preferred covariate scale. The concurrent fit highlights the typical coefficient interpretation from a multiple regression model with two sets of covariates. The orthogonalized fits outline a computationally expedient way to obtain the same coefficients. Finally, the residual fits demonstrate a different way to interpret the regression coefficients without relying on orthogonalization. Additionally, for linear settings with PLS, this approach demonstrates a way to interpret the connection between 2-stage and 1-stage conditional approaches.

4.3 Conditional adjustment for a continuous outcome

To gain insight about the performance of conditional adjustment techniques, first consider a centered continuous outcome constructed from two sets of covariates and random independent mean-0 noise: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{L}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$. Let $\mathcal{L} = \mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T$ be the orthogonal projection onto the column space of $\mathbf{L}$, and similarly define orthogonal projections $\mathcal{X}$ and $\mathcal{T}$. We show in Appendix F that the multiple linear regression coefficient estimates are

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{bmatrix} = \begin{bmatrix}[\mathbf{X}^T(\mathbf{I}-\mathcal{L})\mathbf{X}]^{-1}\,\mathbf{X}^T(\mathbf{I}-\mathcal{L})\mathbf{y}\\ [\mathbf{L}^T(\mathbf{I}-\mathcal{X})\mathbf{L}]^{-1}\,\mathbf{L}^T(\mathbf{I}-\mathcal{X})\mathbf{y}\end{bmatrix} = \begin{bmatrix}[\mathbf{X}^T\mathbf{X}]^{-1}\,\mathbf{X}^T(\mathbf{y}-\mathbf{L}\widehat{\boldsymbol{\gamma}})\\ [\mathbf{L}^T\mathbf{L}]^{-1}\,\mathbf{L}^T(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})\end{bmatrix}.$$

[4] Additionally, the "pseudo-component" from the PLS fit that orthogonalizes $\mathbf{X}$ corresponds to $(\mathbf{I}-\mathcal{L})\mathbf{X}\mathbf{R}$, so it is important to recalculate it as $\mathbf{T} = \mathbf{X}\mathbf{R}$.

Table 4.1 outlines four ways to create a conditional PLS summary of these regression coefficients, depending on which type of adjustment is preferred and the purpose for using PLS to model $\mathbf{X}$. As before, I use the convention `model.fit[` outcome $\sim$ covariates `]`. The center column lists the steps taken to obtain the PLS summary, while the last column provides the subsequent $\boldsymbol{\gamma}$ estimate for 1-stage-conditional approaches. Note that only the last adjustment is an iterative procedure, as indicated by the subscript t. By Remark 2, these techniques will all achieve the statistically-adjusted PLS coefficients ($\widehat{\mathbf{q}}$) if $\mathbf{X}$ is orthogonalized, but only the 1-stage-conditional approaches will obtain the statistically-adjusted coefficients for the adjustment variables ($\widehat{\boldsymbol{\gamma}}$).

Table 4.1: Conditional techniques to perform PLS summary of linear regression

Adjustment	**(P)LS covariate estimates**
Pre (PRE)	$\widehat{\boldsymbol{\gamma}}_0 = \texttt{LS[} \mathbf{y} \sim \mathbf{L} \texttt{]}$ $\widehat{\mathbf{q}}_1 = \texttt{PLS[} \mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}_0 \sim \mathbf{X} \texttt{]} \longrightarrow \mathbf{T}$
Sequential (SQN)	$\widehat{\mathbf{q}}_1 = \texttt{PLS[} \mathbf{y} \sim (\mathbf{I} - \mathcal{L})\mathbf{X} \texttt{]} \longrightarrow \mathbf{T}_1 \quad \widehat{\boldsymbol{\gamma}}_1 = \texttt{LS[} \mathbf{y} - \mathbf{T}_1\widehat{\mathbf{q}}_1 \sim \mathbf{L} \texttt{]}$
Combination (CMB)	$\boldsymbol{\gamma}^* = \texttt{LS[} \mathbf{y} \sim (\mathbf{I} - \mathcal{X})\mathbf{L} \texttt{]}$ $\widehat{\mathbf{q}}_1 = \texttt{PLS[} \mathbf{y} - \mathbf{L}\boldsymbol{\gamma}^* \sim \mathbf{X} \texttt{]} \longrightarrow \mathbf{T}_1 \quad \widehat{\boldsymbol{\gamma}}_1 = \texttt{LS[} \mathbf{y} - \mathbf{T}_1\widehat{\mathbf{q}}_1 \sim \mathbf{L} \texttt{]}$
Incremental (INC)	$\widehat{\mathbf{q}}_t = \texttt{PLS[} \mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}_{t-1} \sim \mathbf{X} \texttt{]} \longrightarrow \mathbf{T}_t \quad \widehat{\boldsymbol{\gamma}}_t = \texttt{LS[} \mathbf{y} - \mathbf{T}_t\widehat{\mathbf{q}}_t \sim \mathbf{L} \texttt{]}$

4.3.1 Two-stage-conditional adjustment

The Pre-Adjustment (PRE) technique begins by fitting a linear model using only the adjustment variables, which results in the estimate $\widehat{\boldsymbol{\gamma}}_0$. In the second stage, a PLS model is fit on the response residuals $\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}_0$, so its PLS weights maximize $\mathbf{R}^T\mathbf{X}^T(\mathbf{I} - \mathcal{L})\mathbf{y}$—as do the weights from the statistically-adjusted PLS fit. However, its PLS coefficients are biased (regardless of how many are used), as they take the form $[\mathbf{T}^T\mathbf{T}]^{-1}\ \mathbf{T}^T(\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}_0)$ instead of $[\mathbf{T}^T\mathbf{T}]^{-1}\ \mathbf{T}^T(\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}})$. Another way to interpret the difference is that the PRE PLS estimate is not accounting for the correlation between $\mathbf{L}$ and $\mathbf{T}$ in its variance calculation (e.g., assuming independence). Finally, by Remark 2, we could consider the PRE components to be correctly constructed, but they are not subsequently being used in a multiple linear regression with $\mathbf{L}$ to obtain the statistically-adjusted PLS coefficient. Hence, the PRE approach

obtains a PLS fit on a residual outcome, but not statistical adjustment. Note that the PRE $\boldsymbol{\gamma}$ estimate will always be biased unless the covariate sets are independent.

4.3.2 One-stage-conditional adjustment

To improve estimation, one could construct a PLS model while concurrently modeling the adjustment variables. However, we must now consider the assumptions and purpose for using PLS. Formally, the univariate PLS model[5] with adjustment variables is

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathcal{E}_{\mathbf{x}} \qquad \mathbf{y} = \mathbf{T}\mathbf{q} + \mathbf{L}\boldsymbol{\gamma} + \epsilon_{\mathbf{y}}, \qquad \text{where } \mathcal{E}_{\mathbf{x}}, \epsilon_{\mathbf{y}} \perp\!\!\!\perp \mathbf{T}.$$

As alluded to in Chapter 2, there is debate as to how to interpret these models from a statistical standpoint (see e.g., [82]). While some view PLS as a tool to summarize $\mathbf{X}$ using a smaller dimension (e.g., $\mathcal{E}_{\mathbf{x}}$ is residual $\mathbf{X}$ information), others posit that $\mathbf{T}$ represents the true underlying phenomena of interest (e.g., $\mathcal{E}_{\mathbf{x}}$ is random noise). This distinction will influence how much of the residual outcome is used when fitting PLS, and as PLS is a supervised method, it will affect the resulting coefficient estimates.

Component-summary with PLS

With the "latent variable" approach, the goal is to to construct PLS components that are statistically-adjusted for $\mathbf{L}$. As such, a PLS model is fit on the outcome projected off of the column space of $\mathbf{L}$, with weights maximizing $\mathbf{R}^T\mathbf{X}^T(\mathbf{I} - \mathcal{L})\mathbf{y}$. Then, the constructed PLS components are included in a linear model with $\mathbf{L}$. By the concurrent fit framework in Remark 2, this adjustment technique could be considered as an additional step beyond PRE, so we refer to this as Sequential-Adjustment (SQN). In Table 4.1, we have presented a computationally expedient way to achieve statistically-adjusted estimates via the orthogonalized and residual fits frameworks, regardless of the number of selected components.

An iterative PLS adjustment [41] has been proposed based on the concurrent fit frame-

[5]Typically the PLS coefficient is denoted by $\mathbf{q}^T$, but we opt for this form to reduce notation burden.

work. If it orthogonalizes $\mathbf{X}$ (pointed out by [7]), it is equivalent to SQN by Remark 2.

Covariate-summary with PLS

If we instead viewed PLS as a tool to create a lower-dimensional summary of $\mathbf{X}$, we would begin with a multiple linear regression using $\mathbf{X}$ and $\mathbf{L}$. Then PLS could be used to model the residual outcome after subtracting out $\mathbf{L}\widehat{\boldsymbol{\gamma}}$. Note that this residual is inherently larger than the one used in SQN (unless the covariate sets are independent). Also, any $\mathbf{X}$ variable that is correlated with $\mathbf{L}$ will have a larger magnitude coefficient estimate as compared to the SQN estimate, which "orthogonalizes" $\mathbf{X}$. Due to the difference in procedures, this technique does not obtain statistically-adjusted PLS coefficients, but it can still can be a meaningful measure, as it results in a shrinkage summary of the adjusted relationship between $\mathbf{y}$ and $\mathbf{X}$.

If we subsequently refit the $\boldsymbol{\gamma}$ estimate using the constructed PLS component, this could be interpreted as seeking to use PLS for both objectives. We call this Combination-Adjustment (CMB), as it is begins by using PLS as a summary of $\mathbf{X}$ but finishes by treating the components as the covariate of interest. Although they can sometimes be similar, the PLS estimates using CMB will only equal the SQN estimates when all components are fit, as they both reduce to the multiple least squares coefficients estimates.

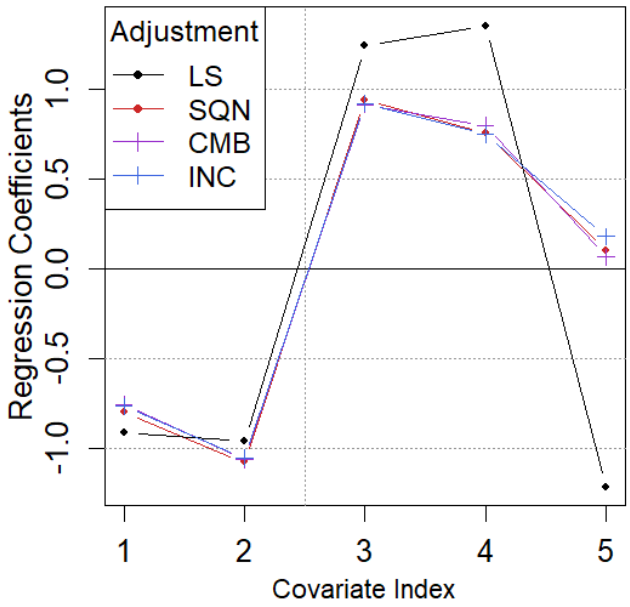
Last, we consider an iterative adjustment approach that consecutively updates its coefficient estimates (starting with initial estimate $\widehat{\boldsymbol{\gamma}}_0$). Although iterates are less appealing for a continuous outcome, these insights will be useful for the survival setting. This Incremental-Adjustment (INC) is inspired by the least squares residual fits in Remark 1, but as explained in Section 4.2, combining PLS residual fits steps does not result in statistically-adjusted coefficients. Instead, they are an approximation of CMB (shown in Appendix F). If we let $\mathbf{T}_\infty$ denote the component $\mathbf{XR}_\infty$ upon convergence, INC becomes a better approximation as the number of components used increases, as $\mathcal{T}_1$ through $\mathcal{T}_\infty$ better approximate $\mathcal{X}$. When all components are fit, this approach will almost[6] converge to the multiple LS estimates.

[6]As the INC estimate is constructed from all component iterates, the initial poorly fit estimates will slightly bias the final result, though these initial components are quite down-weighted.

4.3.3 Implications for the conditional adjustment of a continuous response

We have presented two types of conditional adjustment approaches when using PLS with continuous data, further classified by the objective of modeling with PLS. Although true statistical adjustment is only achieved when the goal is to use components as the scale of interest, PLS can also be used to summarize **X** while adjusting for other variables. Figure 4.1 illustrates the differences in adjusted PLS coefficients in a simple example with unit associations. The first two covariates were moderately correlated, as were the last three. The outcome was constructed from these five base variables, a precision variable and confounder that were five times as important, and random Gaussian noise. When using two components, the PLS summaries were quite similar as compared to the LS fit. However, there are some differences in estimation, particularly for the first and fifth coefficient. Last, though INC is close to CMB, there are still discrepancies in the last two coefficients.

Figure 4.1: Example of conditional adjustment for a continuous outcome with 2 components.



Simulation results are presented in Appendix G to demonstrate how Chun and Keles's Sparse PLS (SPLS) [14] procedure performs when replacing the PLS step with SPLS in these adjustment approaches. Overall, SPLS resulted in better variable selection when applying Combination-Adjustment, though the other approaches were generally similar. They were most sensitive to the amount of noise added to the mean model, followed by the presence or not of a confounder—which impacted PLS coefficient estimation performance.

4.4 Conditional adjustment for a survival outcome

We begin by briefly presenting the Cox proportional hazards model [19] with adjustment variables, though many insights in this section can be extended to generalized linear models (GLMs). Let $\mathbf{x}_i$ denote the i-th person's p "higher-dimensional" covariates, $\mathbf{l}_i$ denote the i-th person's q "low-dimensional" covariates, t_i their event time, and c_i their censorship time, where $c_i \perp\!\!\!\perp t_i | \mathbf{x}_i, \mathbf{l}_i$. Let $y_i = \min(t_i, c_i)$ be their observed time and $\delta_i = \mathbb{1}\{t_i \leq c_i\}$ indicate whether their event was observed. Denote the set of observed events by $\tau_1, ..., \tau_{\tilde{n}}$. The observed data for n participants (with $n > \tilde{n} > p + q$) is contained in $\{\mathbf{X}, \mathbf{L}, \mathbf{y}, \boldsymbol{\delta}\}$. The hazard function takes the form $\lambda(t|\mathbf{x}_i, \mathbf{l}_i) = \lambda_0(t)\ \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{l}_i^T\boldsymbol{\gamma})$, where $\lambda_0(t)$ is an underlying baseline hazard and the effect of the covariates on risk does not vary with time. Breslow's estimator [11] non-parametrically estimates the baseline hazard via $\widehat{\lambda}_0(\tau_k) = \left(\sum_{j \in R_k} e^{\mathbf{x}_j^T\boldsymbol{\beta} + \mathbf{l}_j^T\boldsymbol{\gamma}} \right)^{-1}$, where R_k is the set of individuals still at observable risk at time τ_k (i.e., $j\ :\ y_j \geq \tau_k$).

Cox proposed to estimate the effect of the covariates without modeling the change in risk over time by using only a portion of the log likelihood, called the log partial likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n}\ \delta_i \log\Big[\exp(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{l}_i^T\boldsymbol{\gamma})\ \widehat{\lambda}_0(y_i)\Big]$. The Cox model's estimate of whether a person experienced an event at their observed time, given their covariates, is $\widehat{\mu}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{l}_i^T\boldsymbol{\gamma}) \sum_{j\ :\ \tau_j \leq y_i} \widehat{\lambda}_0(\tau_j)$. The sum estimates the cumulative baseline hazard. Now, to obtain the maximum (log) partial likelihood estimator (MPLE) $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$, the standard procedure is to use iterative Newton updates. Given a current estimate $(\widehat{\boldsymbol{\beta}}_{t-1}, \widehat{\boldsymbol{\gamma}}_{t-1})$, this can be achieved by first updating $\widehat{\boldsymbol{\mu}}_t$ and $\mathbf{V}_t = \text{diag}\,(\widehat{\mu}_{t,i})_{i=1}^n$, the latter of which is the first order approximation of the working response's variance. These enable us to update the working response via $\mathbf{z}_t = \mathbf{X}\widehat{\boldsymbol{\beta}}_{t-1} + \mathbf{L}\widehat{\boldsymbol{\gamma}}_{t-1} + \mathbf{V}_t^{-1}(\boldsymbol{\delta} - \widehat{\boldsymbol{\mu}}_t)$ to ultimately improve our MPLE estimate.

We now transition to applying the conditional adjustment insights learned for a continuous outcome to survival data. We first discuss 1-stage-conditional approaches, in which the survival outcome is modeled by estimating both sets of coefficients at once. Then we present 2-stage-conditional approaches, in which a restricted Cox model is fit before using PLS.

4.4.1 One-stage-conditional adjustment

Consider iteratively modeling a survival outcome as described above by updating both sets of coefficient estimates. At each iteration, let $\mathcal{L} = \mathbf{L}(\mathbf{L}^T\mathbf{V}_t\mathbf{L})^{-1}\mathbf{L}^T\mathbf{V}_t$ be the projection corresponding to the (weighted) column space of $\mathbf{L}$, which is idempotent but not symmetric, and similarly define $\mathcal{X}$ and $\mathcal{T}$. Upon convergence with weighting matrix $\mathbf{V}$ and outcome $\mathbf{z}$, we can express the Cox MPLE as follows, as proven in Appendix F:

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{bmatrix} = \begin{bmatrix}[\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{X}]^{-1}\ \mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z}\\ [\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{X})\mathbf{L}]^{-1}\ \mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{X})\mathbf{z}\end{bmatrix} = \begin{bmatrix}[\mathbf{X}^T\mathbf{V}\mathbf{X}]^{-1}\ \mathbf{X}^T\mathbf{V}(\mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}})\\ [\mathbf{L}^T\mathbf{V}\mathbf{L}]^{-1}\ \mathbf{L}^T\mathbf{V}(\mathbf{z}-\mathbf{X}\widehat{\boldsymbol{\beta}})\end{bmatrix}.$$

As in Chapter 3, we divide the modeling procedure of a survival outcome into two parts—maximizing the Cox MPLE and then applying a one-step (S)PLS summary. As WLS can be considered a WPLS with the full set of components, we only present the 1-stage-conditional adjustment approaches that use WPLS to estimate the Cox MPLE in Table 4.2. I use the convention `model.fit[` outcome $\sim$ covariates `]`, where weighting by matrix $\mathbf{V}_t$ is implied and $\longrightarrow \mathbf{T}_t$ indicates that we construct the component $\mathbf{T}_t = \mathbf{X}\mathbf{R}_t$ using the weights from the corresponding PLS fit. As in the continuous case, these techniques will all reduce to the statistically-adjusted SQN coefficients if $\mathbf{X}$ is orthogonalized, by Remark 2. We first discuss the interpretation and performance of these adjustments when using all components to obtain the MPLE and then how they differ when selecting fewer components (e.g., different versions of approximate MPLE; aMPLE).

Table 4.2: One-stage-conditional techniques to estimate MPLE using PLS

Adjustment	**W(P)LS covariate estimate updates**
Sequential (SQN)	$\widehat{\mathbf{q}}_t = \texttt{WPLS[}\ \mathbf{z} \sim (\mathbf{I}-\mathcal{L})\mathbf{X}\ \texttt{]} \longrightarrow \mathbf{T}_t \quad \widehat{\boldsymbol{\gamma}}_t = \texttt{WLS[}\ \mathbf{z} - \mathbf{T}_t\widehat{\mathbf{q}}_t \sim \mathbf{L}\ \texttt{]}$
Combination (CMB)	$\boldsymbol{\gamma}_t^* = \texttt{WLS[}\ \mathbf{z} \sim (\mathbf{I}-\mathcal{X})\mathbf{L}\ \texttt{]}$ $\widehat{\mathbf{q}}_t = \texttt{WPLS[}\ \mathbf{z} - \mathbf{L}\boldsymbol{\gamma}_t^* \sim \mathbf{X}\ \texttt{]} \longrightarrow \mathbf{T}_t \quad \widehat{\boldsymbol{\gamma}}_t = \texttt{WLS[}\ \mathbf{z} - \mathbf{T}_t\widehat{\mathbf{q}}_t \sim \mathbf{L}\ \texttt{]}$
Incremental (INC)	$\widehat{\mathbf{q}}_t = \texttt{WPLS[}\ \mathbf{z} - \mathbf{L}\widehat{\boldsymbol{\gamma}}_{t-1} \sim \mathbf{X}\ \texttt{]} \longrightarrow \mathbf{T}_t \quad \widehat{\boldsymbol{\gamma}}_t = \texttt{WLS[}\ \mathbf{z} - \mathbf{T}_t\widehat{\mathbf{q}}_t \sim \mathbf{L}\ \texttt{]}$

PLS summaries of MPLE

If these adjustment techniques use all the components to estimate the MPLE, SQN and CMB converge to the standard Cox regression coefficients using $\mathbf{X}$ and $\mathbf{L}$. However, the INC technique again leads to biased estimates. Specifically, $\mathbf{X}$ variables that are correlated with $\mathbf{L}$ result in coefficient estimates that are magnified beyond the Cox estimates. This behavior is particularly concerning, as this technique can result in PLS summaries that have a larger magnitude than the Cox estimates, which is contradictory to the dimension reduction property of PLS. Unfortunately, a version of INC with a slower $\boldsymbol{\gamma}$ update (i.e., regressing $\mathbf{L}$ on $\mathbf{z} - \mathbf{T}_{t-1}\widehat{\mathbf{q}}_{t-1}$) has become the default approach for covariate adjustment within Cox aMPLE methods [55, 16, 45].

By performing a one-step PLS summary after maximizing the MPLE, the reasonable adjustment techniques can be interpreted as in the continuous case; SQN creates components and estimates that are statistically-adjusted for $\mathbf{L}$, while CMB provides a dimension-reduced summary of $\mathbf{X}$. As the MPLE model is used when $\mathbf{X}$ is considered to be the predictor scale of interest, we propose using Combination-Adjustment when PLS is only being used to summarize $\mathbf{X}$.

PLS summaries of distinct aMPLEs

When using fewer components to approximate the MPLE, SQN again creates components and estimates that are statistically-adjusted for $\mathbf{L}$. If operating under the assumption that latent components are the true scale of interest, we propose using Sequential-Adjustment when performing aMPLE. There has been a proposed adjustment scheme [7] for PLS in the logistic setting that uses the concurrent fit framework, and it is equivalent to SQN with the aMPLE model by Remark 2. We briefly note that SQN with the MPLE model simply results in less attenuated coefficients than the aMPLE ones.

If applying CMB, however, its objective is two-fold. While it begins by performing a WLS update and using PLS to summarize $\mathbf{X}$, its $\boldsymbol{\gamma}$ update implies that the components are the scale of interest. In simulations, this adjustment technique results in PLS estimates that

are a combination of the SQN version of aMPLE estimates and those of the standard Cox model (MPLE). In particular, $\mathbf{X}$ variables that are correlated with $\mathbf{L}$ result in estimates that are closer to (but never larger than) the Cox estimates, while those that are uncorrelated resemble the SQN aMPLE estimates. In addition, the aMPLE estimates are quite similar if only a few components are being fit relative to rank($\mathbf{X}^T\mathbf{V}\mathbf{X}$).

4.4.2 Two-stage-conditional adjustment

Another way to apply conditional adjustment for a non-linear model is to begin with the restricted model using only the $\mathbf{L}$ covariates and subsequently fit a PLS model on the residuals. To differentiate the converged restricted Cox model terms from those derived from a Cox model that includes $\mathbf{X}$, we denote the restricted versions by $\breve{\mathbf{V}}$ and $\breve{\mathbf{z}}$. Although less ideal than a 1-stage-conditional adjustment from a model interpretation and estimation perspective, 2-stage-conditional adjustments have been considered for their computational advantages, particularly in high-dimensional settings. The simplest approach involves fitting weighted PLS on the Pearson residuals of a restricted Cox model, which we refer to as Restricted-Adjustment (RSTR) in Table 4.3. This approach resembles a score test, which fits the restricted model and tests whether including $\mathbf{X}$ results in significantly closer estimates to the M(P)LE.

Table 4.3: Two-stage-conditional adjustment techniques for a survival outcome

Adjustment	**W(P)LS covariate estimates**	
Restricted (RSTR)	1. Fit Cox model on $\mathbf{L}$ alone: $\breve{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \left\|\breve{\mathbf{V}}^{1/2}\ (\breve{\mathbf{z}} - \mathbf{L}\boldsymbol{\gamma})\right\|_2^2$.	2. $\breve{\mathbf{q}} = \texttt{WPLS[}\ \breve{\mathbf{z}} - \mathbf{L}\breve{\boldsymbol{\gamma}} \sim \mathbf{X}\ \texttt{]} \longrightarrow \mathbf{T}$.
Deviance Residual (DR)	1. Fit Cox model on $\mathbf{L}$ alone: $\breve{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \left\|\breve{\mathbf{V}}^{1/2}\ (\breve{\mathbf{z}} - \mathbf{L}\boldsymbol{\gamma})\right\|_2^2$.	2. Fit Deviance Residuals $\mathbf{D}_i = \text{sign}(\delta_i - \breve{\mu}_i)\ \sqrt{2\left[\breve{\mu}_i - \delta_i - \delta_i \log(\breve{\mu}_i)\right]}$. 3. $\breve{\mathbf{q}} = \texttt{PLS[}\ \mathbf{D} \sim \breve{\mathbf{V}}^{1/2}\ \mathbf{X}\ \texttt{]} \longrightarrow \mathbf{T}$.

We now extend and correct the deviance residual method of [4] (DR-PLS) to allow for adjustment variables $\mathbf{L}$. Specifically, instead of fitting a null Cox model, we propose fitting a Cox model only with $\mathbf{L}$ to estimate $\boldsymbol{\gamma}$. For martingale ($\mathbf{M}$) and deviance ($\mathbf{D}$) residuals,

the same argument as demonstrated in [64] can be used to state that

$$\left\|\breve{\mathbf{V}}^{1/2}\ (\breve{\mathbf{z}} - \mathbf{L}\breve{\boldsymbol{\gamma}})\right\|_2^2 \ = \ \left\|\breve{\mathbf{V}}^{-1/2}\ (\boldsymbol{\delta} - \breve{\boldsymbol{\mu}})\right\|_2^2 = \sum_i \frac{\mathbf{M}_i^2}{\breve{\mu}_i} \ \approx \ \sum_i \frac{\mathbf{M}_i^2}{\delta_i} \approx \sum_i \mathbf{D}_i^2.$$

As $|D_i| \approx \left|\breve{\mathbf{V}}_{ii}^{1/2}(\breve{\mathbf{z}}_i - \mathbf{L}_i\breve{\boldsymbol{\gamma}})\right|$, one way to approximate RSTR is to use the deviance residuals as the new "linear" response, which we refer to as Deviance-Residual-Adjustment (DR). However, covariates $\mathbf{X}$ need to be pre-multiplied by $\breve{\mathbf{V}}^{1/2}$ to be scaled correctly for the outcome, which has not been indicated by the authors of this method [4, 5].

To connect these methods to the 1-stage-conditional adjustment techniques, we consider the interpretation of the resulting PLS estimates. As we can express the $\boldsymbol{\gamma}$ estimate from the restricted model as $\breve{\boldsymbol{\gamma}} = \breve{\mathcal{L}}\breve{\mathbf{z}}$, RSTR is the generalized version of Pre-Adjustment that we saw in Section 4.3 for a continuous outcome. Although the PLS weights using RSTR are maximizing the correct quantity (albeit using different converged survival terms), its PLS and $\boldsymbol{\gamma}$ estimates are biased. But, if we were to orthogonalize $\mathbf{X}$ before fitting WPLS, we would obtain a component summary just like Sequential-Adjustment. The same patterns hold for DR, except that the approximation $\breve{\mathbf{V}}^{1/2}\mathbf{D} \approx \breve{\mathbf{V}}(\mathbf{I} - \breve{\mathcal{L}})\breve{\mathbf{z}}$ is being used to construct both the PLS components and coefficient estimates.

In simulation studies (presented in Appendix G), SPLS resulted in far better variable selection, PLS coefficient estimation, and model discrimination when applying RSTR with orthogonalized $\mathbf{X}$. Although DR with orthogonalized $\mathbf{X}$ was the next best contender, its computational advantage was outweighted by its worse estimation and prediction properties in the context of low-dimensional and pre-specified $\mathbf{L}$. Last, we note that any 2-stage-conditional adjustment will suffer in its PLS estimation performance the more that $\mathbf{L}$ dominates the relationship with survival time, as the contribution of $\mathbf{X}$ will largely be masked by the strongly relevant adjustment variables in the restricted model.

4.5 Simulation studies for a survival outcome

4.5.1 Study design

In this section, we evaluate Cox methods that integrate sparse PLS dimension reduction and adjust for other covariates. Specifically, we compare our proposed MPLE w/ SPLS (using Combination-Adjustment) and aMPLE w/ SPLS (using Sequential-Adjustment) with the two existing Cox-PLS methods that incorporate sparsity within aMPLE—LEs-aMPLE [45] and CHs-aMPLE w/ PLS [16], both of which use a version of Incremental-Adjustment. We also present results for the top 2-stage-conditional option (RSTR with Sequential-Adjustment).

To tailor the simulations to a nutritional epidemiology context, we generated multiblock predictors that exhibit similar patterns of correlation as foods do in MESA. The "higher-dimensional" covariates $\mathbf{X}_{1000\times 24}$ had four blocks, where variables in the same block were more similar to each other to imitate food groups (intra-block covariances of .05, .5, .3, and .1, respectively). Adjustment variables $\mathbf{L}_{1000\times 4}$ were generated such that the first three were always precision variables and the role of the last one varied. Specifically, we present settings in which the last adjustment variable is related (magnitude of .30) to no $\mathbf{X}$ ("pure" precision), two false $\mathbf{X}$ ("spurious" precision), and two true $\mathbf{X}$ (confounder). Survival times were constructed from all 4 adjustment variables, only 9 of the 24 $\mathbf{X}$ predictors, and slight random Gaussian noise (sd=1). The true $\boldsymbol{\beta}$'s were either unit positive or negative, and we varied the relative importance of the variable sets by testing $\boldsymbol{\gamma}$ magnitudes between 1 to 6. Exponential survival time (without ties) was generated according to [8] such that roughly a 10% event rate and 15% censorship rate were observed at the end of the study period.

Across 50 generated data sets, we split our observations into a training (n=700) and test (n=300) set, and we used 5-fold cross-validation (CV) on the training data to select the tuning parameters for each method. The k-fold approximation [74] of the CV log-partial likelihood (CVLL) [75] was used to compute our deviance criterion (-2*CVLL). All splits of the data were balanced by survival time (based on its quartiles) using the `caret` package in R in order for results to be less sensitive to fold splits. On the test data, we evaluated variable

selection via the number of true $\mathbf{X}$ variables (TP, out of 9) and false variables (FP, out of 15) selected, PLS estimation error via the RMSE of $\widehat{\boldsymbol{\beta}}$, model discrimination via Harrell's C-statistic (C), and parsimony by the number of selected components (K, out of 5). Note that the RMSE of $\widehat{\boldsymbol{\beta}}$ considers the estimation of both true and false coefficients. Often, the over-shrinkage of true estimates will outweigh the benefit of discarding false variables (if they were assigned a low magnitude), which will be reflected in a larger RMSE.

4.5.2 Results

In Table 4.4, we compare five sparse Cox methods that apply PLS to 24 covariates while conditionally adjusting for four strongly relevant covariates, and we evaluate their performance on 50 simulated data sets with 1000 observations and a 10% event rate. Cells that are particularly advantageous are listed in blue, while the disadvantageous ones are in red. First we discuss the setting when one of the adjustment variables mimics a confounder, which can cause Sequential methods to incorrectly drop the related $\mathbf{X}$ variables because they restrict the PLS estimates to be outside of the column space of $\mathbf{L}$.

Table 4.4: Average simulation results across 50 multiblock data sets ($n = 1000,\ p = 24$) in the **presence of confounding**, using minimum deviance to select tuning parameters.

	Setting 1: $\lvert\widehat{\boldsymbol{\gamma}}\rvert = \mathbf{5}$					**Setting 2:** $\lvert\widehat{\boldsymbol{\gamma}}\rvert = \mathbf{2}$				
Cox Maximization	TP	FP	RMSE($\widehat{\boldsymbol{\beta}}$)	C	K	TP	FP	RMSE($\widehat{\boldsymbol{\beta}}$)	C	K
MPLE (CMB)	8.8	**4.6**	0.32	0.988	4.0	8.8	4.4	**0.23**	0.974	3.9
aMPLE (SQN)	8.6	**4.6**	0.34	0.988	4.4	8.5	4.5	0.27	0.973	3.7
LEs-aMPLE (INC)	**9.0**	**13.5**	**0.29**	0.988	4.7	**9.0**	**10.7**	0.28	0.971	4.4
CHs-aMPLE (INC)	8.9	5.9	0.32	0.988	4.5	8.7	4.5	0.25	0.972	4.1
RSTR (SQN)	8.5	5.7	**0.39**	0.987	3.5	8.5	4.3	**0.33**	0.968	3.7

Note: Adjustment approaches are listed in parenthesis. CMB summarizes $\mathbf{X}$, SQN summarizes $\mathbf{T}$, and INC tries to achieve both but in a biased way.

As the conditional adjustment methods reach essentially the same high level of model discrimination, we focus on variable selection and PLS estimation performance. As expected, the methods that use Sequential-Adjustment to construct statistically-adjusted components

select fewer true variables, but only about half of a variable is dropped on average. Although LEs-aMPLE is the only method that consistently selects all true variables, it has a difficult time discarding any false variables (FP $\approx$ 11–14). In comparison, both of our proposed methods select the fewest (or close to the fewest) false variables ($\approx$ 4–5) while including almost all true variables. When $\mathbf{L}$ was five times as informative to survival time as $\mathbf{X}$ (left panel), CHs-aMPLE selected one more false variable than our proposed method, while its other performance metrics remained similar. When $\mathbf{L}$ was only twice as informative as $\mathbf{X}$ (right panel), our proposed MPLE resulted in the closest PLS estimates. However, as the relative importance of $\mathbf{L}$ increased, LEs-aMPLE enforced the least amount of shrinkage (although not in a statistically principled manner), leading to the closest PLS estimates. The 2-stage-conditional method (RSTR) overshrunk its PLS estimates across all settings, making it the least ideal method.

Next, in Table 4.5 we present results when all adjustment variables are independent of $\mathbf{X}$ (e.g., "pure" precision variables). This setting should be easier for methods to distinguish between true and false $\mathbf{X}$ variables, especially when $\mathbf{X}$ is almost as important as $\mathbf{L}$ (right panel). Even in this simpler case, LEs-aMPLE struggles to discard false variables (FP $\approx$ 10), while our proposed methods again select 4–5 false variables. Another distinction with the confounder setting is that our MPLE method results in equally as close PLS coefficient estimates as LEs-aMPLE when $\mathbf{L}$ is five times as informative as $\mathbf{X}$. Last, RSTR drops almost one more true variable than the 1-stage adjustments, so we exclude its performance from the subsequent plots. More adjustment variable settings are studied in Appendix G.

Figure 4.2 illustrates the variable selection and PLS estimation performance of the 1-stage-conditional adjustments across three adjustment variable settings, fixing $|\frac{\gamma}{\beta}| = 5$. The first setting corresponds to adjusting for only "pure" precision variables, the second to including a variable that induces spurious associations between two false $\mathbf{X}$ and survival time, and the last setting to including a confounder. Sequential-Adjustment is particularly equipped to do well in the second setting (as illustrated by aMPLE's estimation error), since it can discard the two false $\mathbf{X}$ that lie partly in the column space of $\mathbf{L}$. In this figure, the most

Table 4.5: Average simulation results across 50 multiblock data sets ($n = 1000$, $p = 24$) **without confounding**, using minimum deviance to select tuning parameters.

	Setting 1: $\lvert\frac{\gamma}{\beta}\rvert = 5$					**Setting 2:** $\lvert\frac{\gamma}{\beta}\rvert = 2$				
Cox Maximization	TP	FP	RMSE($\widehat{\boldsymbol{\beta}}$)	C	K	TP	FP	RMSE($\widehat{\boldsymbol{\beta}}$)	C	K
MPLE (CMB)	8.7	5.1	**0.31**	0.988	4.3	8.9	**4.1**	**0.21**	0.966	4.0
aMPLE (SQN)	8.5	**4.7**	0.33	0.988	4.3	8.9	4.9	0.22	0.966	3.9
LEs-aMPLE (INC)	**9.0**	**12.7**	**0.31**	0.987	4.7	9.0	**10.0**	0.28	0.963	4.5
CHs-aMPLE (INC)	8.7	6.0	0.35	0.988	4.3	9.0	4.2	0.22	0.966	4.2
RSTR (SQN)	**8.1**	**4.4**	**0.39**	0.986	3.6	8.7	**3.3**	**0.30**	0.963	3.8

Note: Adjustment approaches are listed in parenthesis. CMB summarizes $\mathbf{X}$, SQN summarizes $\mathbf{T}$, and INC tries to achieve both but in a biased way.

notable features is that LEs-aMPLE has a difficult time discarding any variables, but by enforcing the least amount of $\widehat{\boldsymbol{\beta}}$ shrinkage when $\lvert\frac{\gamma}{\beta}\rvert = 5$, its PLS estimates tend to be the closest to the truth. We also point out that using the MPLE instead of approximated MPLE will inherently result in a larger magnitude PLS summary, which is why our proposed MPLE method tends to achieve closer PLS estimates than our aMPLE method, regardless of the type of adjustment conducted. The right panel shows that CHs-aMPLE also tends to shrink its PLS estimates more than our proposed methods which utilize sparsity after the maximization step. Overall, our methods provide a reduction in false variable selection while selecting most true variables and conducting a more principled adjustment strategy.

In Figure 4.3 we compare the number of components selected out of 5. LEs-aMPLE generally selected the most, as expected. Because it strives for component-wise sparsity, it tended to use half of the selected variables in the first component and another component, while it used roughly a third of the selected variables in the remaining components. Thus, it requires more components than the methods targeting variable selection to construct a reasonably predic-

Figure 4.3: Average component selection across 50 multiblock data sets ($\lvert\frac{\gamma}{\beta}\rvert = 5$).

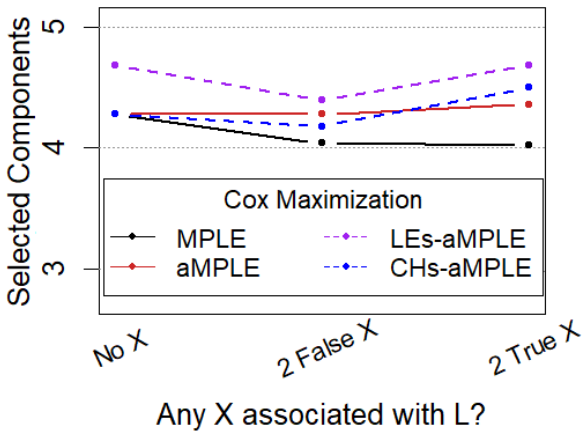
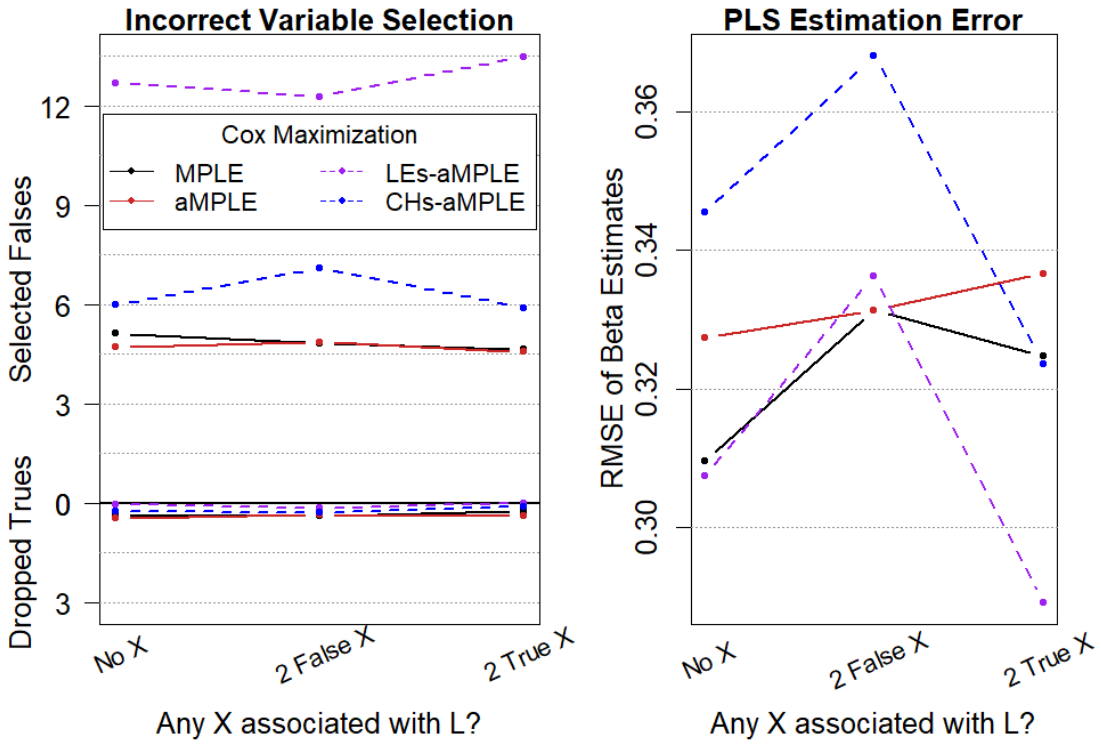


Figure 4.2: Average simulation results for 50 multiblock data sets ($n = 1000$, $p = 24$, $|\frac{\gamma}{\beta}| = 5$), across adjustment variable settings.



tive model. The remaining methods tended to select a similar number of components across settings, with all methods selecting a fraction less of a component when $|\frac{\gamma}{\beta}| = 2$.

Overall, the proposed sparse methods demonstrate a noticeable reduction in false variable selection while maintaining a reasonably high level of true variable selection. While LEs-aMPLE is beneficial if the aim is to construct PLS components that utilize only a few variables each, its covariate adjustment should be modified to apply Combination-Adjustment so as not to over amplify PLS coefficients. The same suggestion applies to CHs-aMPLE, which has only been investigated fully for logistic models. These simulation results suggest that our proposed methods offer a useful alternative to current methods by incorporating sparsity after the maximization procedure and adjusting for other covariates in a principled manner.

4.6 Data application

We return to our motivating scientific goal to identify understandable diet patterns that relate to CVD risk. In the Multi-Ethnic Study of Atherosclerosis, 5,881 participants (86% of the cohort) had complete information on demographics, diet, time of CVD events, and our adjustment variables, resulting in 877 CVD events (including angina) during more than 17 years of follow-up. There was a roughly 15% censorship rate (which we assume is non-informative), and these participants had an 84% probability of surviving beyond 15 years. At baseline (2000-2002), they were aged 44-84 (53% women), free of clinical CVD, and from six United States communities (New York City, Baltimore, Chicago, Los Angeles, the Twin Cities, and Winston Salem). They self-identified as 40% white, 26% Black, 22% Hispanic, and 12% Chinese. Additionally, 42% of the sample had a family history of CVD, 13% had previously or currently smoked, the median educational attainment was a high school diploma, and the median annual income category was \$25,000-\$49,000. The MESA Food Frequency Questionnaire (FFQ) data included the frequency and serving size of 120 foods and beverages that participants consumed during a typical week of the previous year. The calculated servings per day of each food item were used as our "higher-dimensional" predictors, after standardizing them. We also categorized the diet variables into eight food groups: fruits, vegetables, grains, proteins, mixed entrees, dairy, sweets/oils, and beverages.

In our analysis, we adjusted for age, sex, race/ethnicity, education, income, smoking status (ever/never), and family history of heart disease (yes/no). Apart from these factors, we adjusted for their conditioning exercise (log of metabolic equivalent minutes per week) and their total energy (caloric) intake. The latter was included to compare the effect of relative changes in food consumption (see Appendix C for further model interpretation details), which has been established by nutritional epidemiologists a useful scale of analysis [79]. Because total energy intake was calculated using the FFQ data, it was moderately (.10-.40) correlated with almost all the foods and even more strongly correlated with the intake of fries, chips, fried chicken, and steak. Thus, we have strong evidence to suggest that caloric intake

was related to some true food items and may play the role of a confounder. Moreover, we suspect that heart disease risk is far more related to our adjustment variables than the diet measurements, so summarizing the food-scale associations is likely to provide more insight than constructing a statistically-adjusted component.

Table 4.6 displays the results of the CVD-related dietary patterns in MESA when using PLS in three manners. The proposed aMPLE with SPLS summary that constructs statistically-adjusted components selected the sparsest model possible, indicating that there is likely not much benefit to including a food pattern after accounting for the adjusted variables. This model selected the food item including ham hocks, chicharrones (e.g., pork rinds), and pig's feet (positively associated with CVD risk; +), and it resulted in a 0.1% increase in C-statistic over the restricted model. Our proposed MPLE with SPLS summary that provides a dimension-reduced summary of the 120 food and beverages constructed four diet pattern using only 16 items. By order of magnitude, it highlighted Hispanic rice and chicken (-), seeds (-), meat tacos/nachos/enchiladas/tamales (-), refried beans (+), and nuts (except for peanuts; +). The selected model had a 1.7% higher C-statistic than the restricted model, leading to the highest training C-statistic. Last, the diet pattern created by LEs-aMPLE, which seeks to combine both PLS aims in a biased manner, selected 34 foods, most notably eggs (+), refried beans (+), nuts (except for peanuts; +), ham hocks, chicharrones, and pig's feet (+), and Asian dumplings and buns (+). This selection resulted in a 0.1% increase in C-statistic over its restricted model.

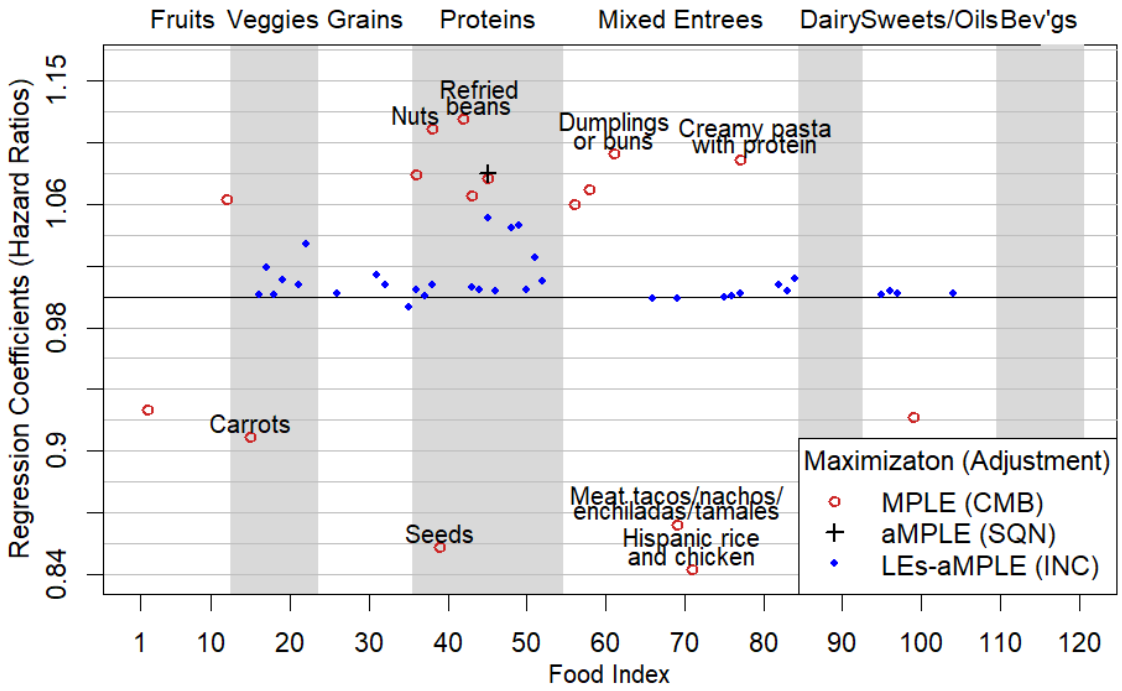
Table 4.6: Summary of CVD-related diet patterns in the Multi-Ethnic Study of Atherosclerosis.

Cox Maximization	Foods	K	CV-Dev.	C
MPLE (CMB)	16	4	18.1	0.731
aMPLE (SQN)	1	1	18.1	0.716
LEs-aMPLE (INC)	34	1	18.3	0.674

K=Number of components. CV-Dev.=Cross-validated deviance. C=Training C-statistic.

Figure 4.4 displays the adjusted hazard ratios from the three Cox methods presented. While the latter two models selected 6 of the same items, it is reasonable to question whether some of the selected foods are driven by the consumption patterns of just a few people, given that some of these items are regional or ethnic specific. Of MPLE's 16 selected food items,

Figure 4.4: Adjusted associations between selected foods and time to CVD events in MESA, ordered by food groups.



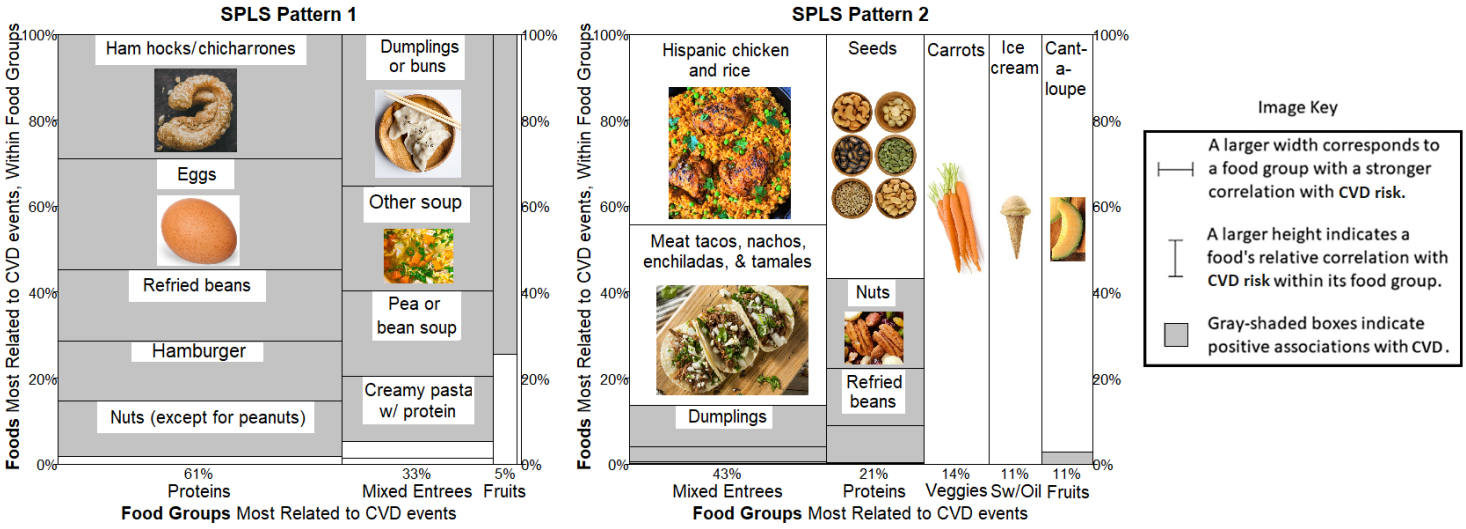
only three were consumed by fewer than 20% of the MESA cohort. These were seeds (19%), Hispanic chicken and rice (17%), and ham hocks/chicharrones/pig's feet (15%). Seeds were consumed fairly evenly across racial/ethnic categories ($\approx$20% of each group except Blacks), whereas Hispanic chicken and rice was predominantly consumed by Hispanics (75% of the consumers). Interestingly, ham hocks/chicharrones/pig's feet were consumed evenly across the non-white groups ($\approx$25% of each group), and this food item was still selected when restricting to the non-white MESA participants in a sensitivity analysis (see Appendix C).

We reiterate that the presented associations are adjusted for total caloric intake, so they represent relative changes in food consumption. Additionally, this model only highlights correlations as opposed to indicating causality. Even so, we note that LEs-aMPLE (blue dots) mainly identified foods that were detrimental to heart-health, and it selected a considerable number of foods to which it attributed very small magnitudes. On the other hand, it is

plausible that MPLE with SPLS summary (open red circles) is over-simplifying the relationship between CVD and diet, but it would seem useful to investigate these foods (or at least those selected by most models) further for potential health implications.

Another way to visualize MPLE's dietary patterns is to look at how the selected foods contribute to the constructed component—analogous to plotting PCA loadings. This can be measured by the PLS weight $\mathbf{R}$, which is the projection that transforms the predictors between the original and component scale (since $\mathbf{T} = \mathbf{XR}$). Foods with larger magnitude contributions to a PLS component will have larger weights, and by grouping and scaling them by food group, practitioners can gain an added perspective about how each food group contributes to the creation of PLS components. Figure 4.5 displays the relative relationships between the selected foods and CVD risk. Along the horizontal axis, food groups are sorted

Figure 4.5: CVD-related diet patterns in MESA, ordered by strength of relationship.



by their squared correlation with (pseudo) CVD events. The vertical axis represents the relative importance of the selected foods within each food group, ordered from top to bottom. The first pattern particularly highlights food items that include fattier preparations of proteins and are associated with higher CVD risk. The second pattern is more reflective of foods associated with lower CVD risk. This type of graphic can be useful to practitioners as

it allows for easier integration of previous scientific knowledge when generating hypotheses.

While certain foods selected by MPLE are more intuitively related to heart-health (e.g., fried pork skins/feet, eggs, hamburgers, creamy pasta with meat/seafood, and refried beans), others were surprising. Seeds are generally viewed as heart-healthy, but so are nuts (if consumed in moderation), though nuts may often be consumed with more salt on them. Both carrots and cantaloupe are considered to be protective of heart health due to their elevated levels of beta carotene [73]. On a similar note, dumplings and buns are one of the few Asian food items without any vegetables. Due to the consumption pattern of Hispanic chicken and rice, this selection may be reflective of a more traditional way of living, which is in line with the hypothesized "Hispanic paradox" [50]. Meat tacos/nachos/enchiladas/tamales might also have a similar explanation, or it could be that these meals tend to prepare meat in a less fatty manner, as they are usually cooked in a sauce as opposed to fried. The negative association between ice cream and CVD risk is the least clear. However, there is another food item in the FFQ that includes low-fat ice cream and frozen yogurt. It may be that people who are healthier are consuming more full-fat ice cream, similar to our BMI analysis that highlighted the negative impact of diet soda instead of regular soda. Although these results would need to be corroborated in other studies, this type of analysis may reveal foods that should be considered for their impact on heart health.

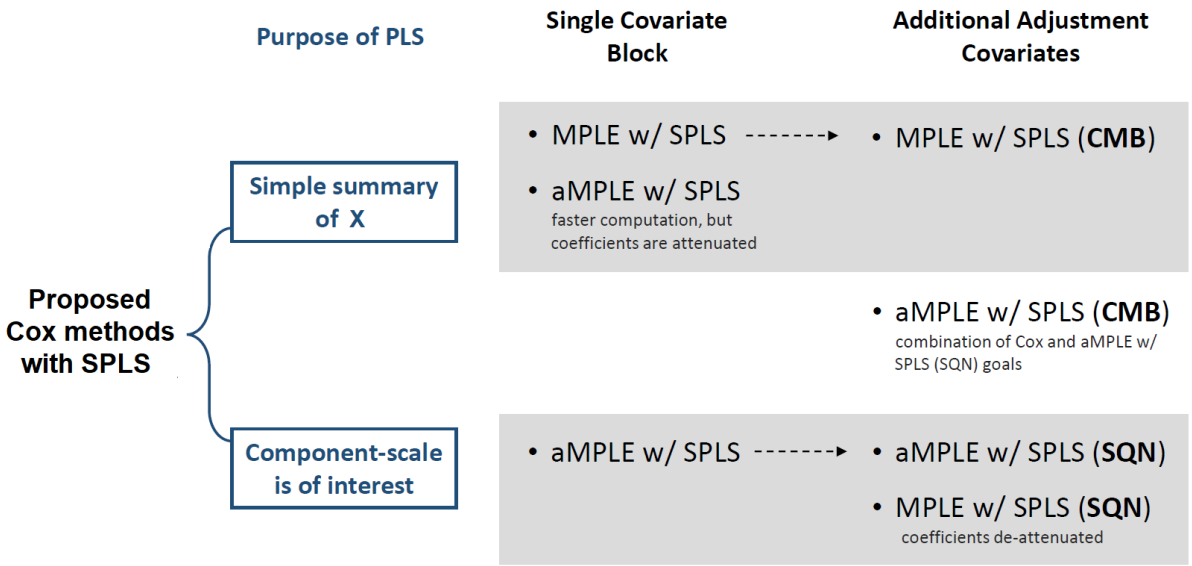
4.7 Discussion

In this chapter, we have investigated how to appropriately adjust for "low-dimensional" covariates when constructing a PLS summary of "higher-dimensional" covariates. In Section 4.2, we presented three ways to obtain the same statistically-adjusted coefficients from a multiple (weighted) least squares regression model, even when using PLS components as one of the sets of covariates. In essence, the concurrent fit allows for the standard interpretation of regression coefficients when including two sets of covariates in a model. The orthogonalized fits outlined a computationally expedient way to obtain those same coefficients, additionally providing insight as to when statistical adjustment will improve variable selection perfor-

mance. Finally, the residual fits demonstrated a way to partition the coefficient estimation by covariate set without relying on orthogonalization. These connections have not been presented as clearly or solved (in the PLS case) previously.

Through our investigations, we found that the most widespread method of covariate adjustment within Cox-PLS methods can unfortunately lead to amplified PLS estimates [55, 16, 45]. After considering the model implications and estimation properties of various conditional adjustments, we suggest that their use of aMPLE implies that Sequential-Adjustment would be best suited for their aims. Alternatively, Combination-Adjustment is most closely aligned to their current approach, while guaranteeing a shrunk estimate. When using our proposed sparse methods, Figure 4.6 outlines which adjustment technique is most suitable depending on one's reason for using PLS. We further connect our proposed adjustment strategy to the corresponding Cox-PLS model from Chapter 3. Recall that the model nomenclature we use is: [("a"pproximated) "MPLE"] + ["SPLS" summary] + [(Adjustment)].

Figure 4.6: Aims of proposed Cox models that incorporate Chun and Keles's Sparse PLS.



In the first row, PLS is being used to summarize $\mathbf{X}$ using fewer dimensions; since $\mathbf{X}$ is the predictor scale of interest, it follows that a standard Cox model (e.g., MPLE) would provide

the maximization aligned with this aim. In the second gray row, the goal is to construct latent components and use those when modeling the predictors. In this case, approximated MPLE achieves this when maximizing the Cox partial likelihood. We place aMPLE with Combination-Adjustment in between these aims, as it is trying to accomplish both, but more research is needed to fully characterize its properties. We include a second model under certain sections if they can also be used to achieve that goal, with any notable differences listed below in a smaller font.

In this chapter, we also extended and corrected 2-stage-conditional adjustment techniques for the survival setting. This work clarifies the underpinnings of the deviance residual model [4] and enables researchers (particularly in the "high-dimensional" setting) to take advantage of its computational speed while more appropriately modeling their covariates. When modeling "lower-dimensional" data, though, we recommend applying 1-stage-conditional adjustment to better model the relative contributions of each covariate type.

Though component-wise sparsity can provide a certain type of interpretability, we posit that our motivating epidemiological setting calls for an emphasis on variable selection. In agreement with our findings in Chapter 3, we saw that utilizing sparsity during the maximization of the log partial likelihood did not generally result in parsimony improvements when compared to applying a sparse PLS summary after the maximization procedure. Although this may not hold in settings where a majority of covariates are completely unrelated to the outcome (as opposed to weakly related), the computational advantage and general utility of (a)MPLE with SPLS summary lead us to prefer these methods. Across a variety of simulation settings, both of our proposed methods provided a notable reduction in false variables selected while including almost all true variables, as SPLS did in the continuous setting in Chapter 2.

Despite these observed advantages, there are still certain limitations to consider. Across all tested settings, it was difficult for any method to select fewer than four false variables, especially as the relative importance of $\mathbf{L}$ increased. This is largely due to the multiblock correlation structure of $\mathbf{X}$ in which four false variables in Block 2 were strongly related ($\approx .50$

magnitude) to two true variables. However, this restriction was not as apparent in Chapter 3 when there were no adjustment variables. This implies that any sparse Cox-PLS method with substantial adjustment variables is likely to select a certain number of irrelevant covariates if they are strongly correlated with relevant covariates. Next, we allowed all models to construct up to 5 (S)PLS components because most practitioners in our settings of interest prioritize interpretability, but additional components could be considered if the goal were instead to obtain a close lower-rank approximation. Finally, we restricted our focus to dimension reduction of "higher-dimensional" covariates as opposed to "high-dimensional" ones. In the latter setting, MPLE with SPLS summary is no longer tractable, but a Conjugate Gradient Accelerated IRLS approach [25] could be used if $\mathbf{X}$ is considered to be the scale of interest. In a future study, it would be interesting to assess which of the conditional adjustment techniques provides better parsimonious properties for "high-dimensional" data.

When modeling different-sized data sets, researchers have actively studied the combination of "low-dimensional" clinical covariates such as age, sex, and test results with "high-dimensional" -omic data. However, it still poses a challenge to properly integrate "higher-dimensional" clinical covariates due to differing correlation structures, potentially very sparse data, and considerations about how to weight variables from each data type [47]. Our research findings can provide insight into how to combine such data sets into current clinico-genomic models to better understand nuances with complex diseases such as heart disease. To that end, we analyzed the relationship between CVD risk and diet, adjusting for relevant risk factors and demographics. Though preliminary in nature, this "multi-dimensional" clinical model can help inform future CVD hypotheses, inspire more disease-tailored diet patterns, and encourage others to pursue more detailed clinico-genomic models. We close by emphasizing that PLS models are particularly utilized in many -omic settings, and many of these research findings can easily be adapted to generalized linear models.

Chapter 5

CONCLUSIONS AND IMPLICATIONS FOR FUTURE WORK

In this dissertation, we have investigated the statistical and modeling implications of using PLS to construct dietary patterns that are tailored to specific diseases. We have further proposed sparse PLS methods and adjustment strategies to facilitate hypothesis generation by interpreting more parsimonious patterns. In Chapter 2, we drew connections between and discussed which dimension reduction method would be best suited for nutritional epidemiology when analyzing a continuous response. We then analytically characterized the relationship between PLS and other relevant numerical algorithms. Using these insights, we evaluated the various rationale for incorporating PLS into the Cox proportional hazards model in Chapter 3. We proposed utilizing a Sparse PLS procedure that directly selects variables after fitting the Cox (or approximate Cox) model to study the relationship between diet and a right-censored survival time. Next, to enable proper adjustment for covariates that do not require dimension reduction, we demonstrated in Chapter 4 that the reason for using PLS dictates which type of adjustment to perform. We also provided a data-integration framework to understand the differences between adjustment strategies.

Combined with our illustrations of proposed methods and adjustment strategies using data from the Multi-Ethnic Study of Atherosclerosis (MESA), these findings can enable more informed selection of methods. For instance, when using PLS and adjusting for low-dimensional covariates, we recommend avoiding the Incremental-Adjustment scheme that has been popularized among Cox-PLS methods since it was proposed [55]. Not only is this iterative scheme less efficient when updating the adjustment coefficients, it has the unfortunate property of de-attenuating certain PLS coefficients even beyond their corresponding least squares estimates. Instead, we found that this procedure was targeting the Combined-

Adjustment approach, which does not suffer from over-amplification. However, if wanting to approximate the Cox model using PLS, more investigation is needed to fully characterize the performance of aMPLE with SPLS when using Combined-Adjustment. If a 2-stage-conditional adjustment is instead preferred, we recommend simply fitting a Cox model on the adjustment variables, and then using PLS to minimize its (sum of squared) Pearson residuals, as opposed to the deviance residual approach [4].

Our work has focused on extending appropriate methods to model food frequency questionnaire data, which paves the way for analyzing the relationship between diet and time to incidence of angina, heart attack, stroke, or CVD mortality. There are also other types of heart disease related endpoints to consider. In particular, the relationship between the presence of coronary artery calcium (CAC) and diet has not been well established, so this could be a valuable scientific contribution. Additional simulation studies would be helpful to confirm whether applying sparsity after the maximization technique is as useful in GLMs as in the survival context. Moreover, as PLS methods are used in many other disciplines, such as chemometrics, environmental studies, and -omics settings, it would be fruitful to test other covariate correlation structures in simulations to better mimic those data sets.

Furthermore, as a dimension reduction method, PLS has the potential to contribute greatly to certain high-dimensional analyses. If the goal is to conduct a 1-stage-conditional adjustment, it is no longer tractable to solve for the M(P)LE of non-linear models. Instead, one could utilize the approximate M(P)LE method or even Conjugate Gradient Accelerated IRLS (e.g., not fixing the number of components at each iteration) from Chapter 3. If there is interest in conducting a 2-stage-conditional adjustment, the amount of methodological adaptations (if any) would depend on how many adjustment variables are being considered (e.g., are there enough observations to fit a restricted model or not). Therefore, it would be interesting to undertake a full investigation of the utility of PLS in high-dimensional settings.

Another aspect of our investigation involved identifying a suitable sparsity procedure. Although Chun and Keles's Sparse PLS technique has been quite advantageous in both simulations and real world examples, it would be helpful to more precisely characterize its model

selection consistency to ensure that the true variables will be recovered under certain conditions. Specifically, future work could entail verifying whether a version of the Irrepresentable Condition holds as it does for the Lasso [85]. A related aim of ours was to automate the selection of a more parsimonious model during cross-validation. We adapted the one standard error test [31] for Sparse PLS methods with two and three tuning parameters. With survival data, this test selected overly-simplified models, so we opted for an AIC-type test in this setting. More studies would be useful to investigate these automatic parsimony tests for GLMs and determine when they are most applicable.

Overall, it has been rewarding to consider how best to serve the goals of practitioners throughout this dissertation. To enable scientists to more easily compare results with their domain knowledge, we developed graphics that facilitate examining PLS components at the food group level (inspired by Multiblock PLS literature [77, 78, 61, 42]). To facilitate their use in practice, we will develop our proposed methods into an `R` package, particularly taking advantage of the optimized `coxph` code for our MPLE with SPLS summary.

On a last note, when constructing statistically-adjusted dietary patterns within MESA, it was reassuring that SPLS highlighted the positive association between BMI and "fast foods" or a more "Western" diet. This gave us confidence to extend this procedure to more complex outcomes. Beyond using survival analysis to construct a heart-healthy diet in MESA, our second data analysis highlighted the importance of conducting studies in multi-ethnic cohorts and with food surveys that reflect regional and ethnic options. Although our proposed sparse Cox-PLS method generally selected items with fattier preparations of proteins, the food item with the strongest signal for CVD risk included ham hocks, chicharrones, and pig's feet. Only 3% of white participants consumed these foods, but more than 25% of Asians, Blacks and Hispanics did. While this food category could simply be a surrogate of non-white status, it also has the potential of truly being related to heart-health—and it would not have been identified in a predominantly white cohort. Thus, observational studies such as MESA can help build evidence for which foods should be investigated further to improve heart-health.

Appendix A

VERIFICATION THAT (W)PLS AND CONJUGATE GRADIENT ARE EQUIVALENT

A.1 Equivalence between PLS and Conjugate Gradient

A.1.1 Introduction

Linear Conjugate Gradient (CG) is an iterative approach to solving the system of linear equations $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is symmetric and positive definite. In the context of solving the normal estimating equations, this is useful for solving the system $(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$. This approach constructs search directions that are guaranteed to be descent directions by using conjugacy with respect to $\mathbf{A}$; a set of vectors $\{\mathbf{v}_i\}$ are $\mathbf{A}$-conjugate if $\mathbf{v}_i^T\mathbf{A}\mathbf{v}_j = 0 \; \forall \; i \neq j$. These linearly independent directions form a basis that diagonalizes $\mathbf{A}$, and so CG finds the minimizer $\hat{\mathbf{x}}$ over the entire space in rank($\mathbf{A}$) steps. At the k-th CG step, its solution $\hat{\mathbf{x}}^{(k)}$ minimizes $\left\{\left(\hat{\mathbf{x}} - \mathbf{x}^{(k)}\right)^T \mathbf{A} \left(\hat{\mathbf{x}} - \mathbf{x}^{(k)}\right)\right\}$ over the current rank-k (Krylov) subspace, with the norm of each error monotonically decreasing with k.

Though PLS was developed with the purpose of constructing orthogonal linear combinations of a predictor (e.g. components) that maximize the covariance with an outcome, it turns out that these algorithms are equivalent if CG is initialized with $\mathbf{x}_0 = \mathbf{0}$ [60]. In the following section we provide some intuition for the CG terms and we verify the relationship between these and the PLS terms. These relationships will be useful for derivations throughout this dissertation.

To clarify notation, we begin by specifying each algorithm in the context of solving the normal equations, and WLOG assume $\mathbf{X}$ is centered. The product form for the subsequent residual $\mathbf{X}_k$ and $\mathbf{y}_k$ were initially proven by Helland [33].

Conjugate gradient with $\hat{\boldsymbol{\beta}}_0 = 0$

Initialize: $\mathbf{r}_0 = -\mathbf{X}^T\mathbf{y}$ and $\mathbf{d}_0 = 0$.

For k in 1, ..., rank($\mathbf{X}^T\mathbf{X}$) {

$$\mathbf{r}_k = \mathbf{r}_{k-1} + s_{k-1}\ \mathbf{X}^T\mathbf{X}\ \mathbf{d}_{k-1}$$
$$\mathbf{d}_k = -\mathbf{r}_k + \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\ \mathbf{d}_{k-1}$$
$$s_k = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{d}_k^T\ \mathbf{X}^T\mathbf{X}\ \mathbf{d}_k}$$
$$\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\beta}}_{k-1} + s_k\ \mathbf{d}_k$$

}

Partial Least Squares (NIPALS)

Initialize: $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{y}_1 = \mathbf{y}$.

For k in 1, ..., rank($\mathbf{X}^T\mathbf{X}$)=K {

$$\tilde{\mathbf{w}}_k = \mathbf{X}_k^T\ \mathbf{y}_k$$
$$\mathbf{w}_k = \tilde{\mathbf{w}}_k / \sqrt{\tilde{\mathbf{w}}_k^T\ \tilde{\mathbf{w}}_k}$$
$$\mathbf{t}_k = \mathbf{X}_k\ \mathbf{w}_k$$
$$\mathbf{p}_k = \mathbf{X}_k^T\mathbf{t}_k/(\mathbf{t}_k^T\ \mathbf{t}_k)$$
$$q_k = \mathbf{y}_k^T\mathbf{t}_k/(\mathbf{t}_k^T\ \mathbf{t}_k)$$
$$\begin{aligned}\mathbf{X}_{k+1} &= \mathbf{X}_k\ -\ \mathbf{t}_k\mathbf{p}_k^T \\ &= \mathbf{X}_k\ -\ \mathbf{t}_k\mathbf{t}_k^T\ \mathbf{X}_k/(\mathbf{t}_k^T\ \mathbf{t}_k) \\ &= \left[\prod_{m=1}^{k}\left(\mathbf{I}_n - \frac{\mathbf{t}_m\mathbf{t}_m^T}{\mathbf{t}_m^T\mathbf{t}_m}\right)\right]\mathbf{X}\end{aligned}$$
$$\begin{aligned}\mathbf{y}_{k+1} &= \mathbf{y}_k\ -\ \mathbf{t}_k q_k^T \\ &= \left[\prod_{m=1}^{k}\left(\mathbf{I}_n - \frac{\mathbf{t}_m\mathbf{t}_m^T}{\mathbf{t}_m^T\mathbf{t}_m}\right)\right]\mathbf{y}\end{aligned}$$

}

$$\mathbf{W}_K = \begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_K\end{bmatrix}$$
$$\mathbf{P}_K = \begin{bmatrix}\mathbf{p}_1 & \cdots & \mathbf{p}_K\end{bmatrix}$$
$$\mathbf{Q}_K = \begin{bmatrix}q_1 & \cdots & q_K\end{bmatrix}$$
$$\hat{\boldsymbol{\beta}}_K = \mathbf{W}_K\left(\mathbf{P}_K^T\mathbf{W}_K\right)^{-1}\mathbf{Q}_K^T\ :=\ \mathbf{R}_K\ \mathbf{Q}_K^T$$

Note that PLS doesn't require the calculation of $\hat{\boldsymbol{\beta}}_k$ until all the components are constructed. Additionally, we restrict our study to univariate $\mathbf{y}$, so $\mathbf{Q}$ is actually a vector.

Below, we provide some intuition for the CG terms:

- s_k is the inverse of the variance of the selected components, which is the approximation of $(\mathbf{X}^T\mathbf{X})^{-1}$ in OLS. It is also the step size that maximizes the log likelihood when using k CG steps:

$$\begin{aligned} s_k &= \min_{s\in(0,\infty)} \; -\ell\left(\hat{\boldsymbol{\beta}}_{k-1} + s\ \mathbf{d}_k\right) \;=\; \min_{s\in(0,\infty)} \; \left\|\mathbf{y} - \mathbf{X}\left(\hat{\boldsymbol{\beta}}_{k-1} + s\ \mathbf{d}_k\right)\right\|_2^2 \\ &= \min_{s\in(0,\infty)} \; \left\{ -2s\ (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{k-1})^T\mathbf{X}\mathbf{d}_k \;+\; s^2\ \mathbf{d}_k^T\mathbf{X}^T\mathbf{X}\mathbf{d}_k\right\} \\ &= \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{k-1})^T\mathbf{X}\mathbf{d}_k}{\mathbf{d}_k^T\mathbf{X}^T\mathbf{X}\mathbf{d}_k} \;=\; \frac{\mathbf{d}_1^T\mathbf{d}_k \;- s_1\mathbf{d}_1^T\mathbf{X}^T\mathbf{X}\mathbf{d}_k - \ldots - s_{k-1}\mathbf{d}_{k-1}^T\mathbf{X}^T\mathbf{X}\mathbf{d}_k}{\mathbf{d}_k^T\mathbf{X}^T\mathbf{X}\mathbf{d}_k} \;=\; \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{d}_k^T\mathbf{X}^T\mathbf{X}\mathbf{d}_k} \end{aligned}$$

- $\mathbf{r}_k$ are the residual directions of steepest descent (e.g. negative residual gradients), provided that they are orthogonal to all previous directions. They can also be constructed as $\mathbf{r}_k = -\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\ \widehat{\boldsymbol{\beta}}_{k-1} = -\mathbf{X}^T\mathbf{X}\ (\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \widehat{\boldsymbol{\beta}}_{k-1})$.

- $\mathbf{d}_k$ are the conjugated residual gradients, guaranteeing that they are $\mathbf{X}^T\mathbf{X}$-orthogonal from previous directions. They are the oblique projection of $-\mathbf{r}_k$ onto the space orthogonal to span($\mathbf{X}^T\mathbf{X}\ \mathbf{d}_{k-1}$) along $\mathbf{d}_{k-1}$. One can obtain PLS components with $\mathbf{X}\ \mathbf{d}_k$, up to a scaling factor.

For ease of labeling, we sometimes refer to $\mathbf{r}_k$ as "gradients" and $\mathbf{d}_k$ as "conjugated gradients" in the following proofs.

A.1.2 Proving PLS and CG are equivalent

Now, we demonstrate that the following relationships hold at the kth CG and PLS step, for any $k \in [1, \text{rank}(\mathbf{X}^T\mathbf{X})]$. Specifically, we will verify the terms that are underlined, as the others hold by definition. For clarity, the terms on the left are from PLS and those on the right are from CG.

- residual gradients (weights): $\tilde{\mathbf{w}}_k = -\mathbf{r}_k$ and $\mathbf{w}_k = -\mathbf{r}_k/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- orthogonal projection: $\left(\mathbf{I}_n - \frac{\mathbf{t}_k\mathbf{t}_k^T}{\mathbf{t}_k^T\mathbf{t}_k}\right) = \left(\mathbf{I}_n - s_k\ \mathbf{X}_k\ \frac{\mathbf{r}_k\mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k}\ \mathbf{X}_k^T\right)$
- components: $\mathbf{t}_k = \mathbf{X}_k\ \mathbf{w}_k\ =\ -\left[\prod_{m=1}^{k}\left(\mathbf{I}_n - s_m\ \mathbf{X}_m\ \frac{\mathbf{r}_m\mathbf{r}_m^T}{\mathbf{r}_m^T\mathbf{r}_m}\ \mathbf{X}_m^T\right)\right]\mathbf{X}\mathbf{r}_k\ /\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$ $= \mathbf{X}\ \mathbf{R}_k^{(k)} = \mathbf{X}\ \mathbf{d}_k/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- component variance: $\mathbf{t}_k^T\mathbf{t}_k = 1/s_k$
- loadings: $\mathbf{p}_k = \frac{s_k}{\sqrt{\mathbf{r}_k^T\mathbf{r}_k}}\ \mathbf{X}^T\mathbf{X}\ \mathbf{d}_k\ =\ (\mathbf{r}_{k+1} - \mathbf{r}_k)/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- component-scale regression coefficient: $q_k = s_k\ \sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- conjugated gradients: $\tilde{\mathbf{w}}_k + \frac{\tilde{\mathbf{w}}_k^T\tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_{k-1}^T\tilde{\mathbf{w}}_{k-1}}\tilde{\mathbf{w}}_{k-1} + \frac{\tilde{\mathbf{w}}_k^T\tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_{k-2}^T\tilde{\mathbf{w}}_{k-2}}\tilde{\mathbf{w}}_{k-2} + \cdots + \frac{\tilde{\mathbf{w}}_k^T\tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_1^T\tilde{\mathbf{w}}_1}\tilde{\mathbf{w}}_1\ =\ \mathbf{d}_k$
- X-scale residual gradients (weights): $\mathbf{R}_k^{(k)} = \mathbf{d}_k/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- X-scale regression coefficient: $\hat{\boldsymbol{\beta}}_k = \mathbf{R}_k\mathbf{Q}_k^T = \left[\mathbf{R}_k^{(1)}\ \ \mathbf{R}_k^{(2)}\ \ ...\ \ \mathbf{R}_k^{(k)}\right]\left[q_1\ \ ...\ \ q_k\right]^T = \sum_{m=1}^{k} s_m\mathbf{d}_m$

To verify these statements, we will make use of certain **orthogonality properties and equalities**, letting $\mathcal{S} = \mathbf{X}^T\mathbf{X}$ and $j, k \in [1, K]$:

1. gradients are orthogonal: $\mathbf{r}_j^T\ \mathbf{r}_k = 0 \qquad \forall\ j \neq k$

2. conjugated gradients are $\mathcal{S}$-orthogonal: $\mathbf{d}_j^T\ \mathcal{S}\ \mathbf{d}_k = 0 \qquad \forall\ j \neq k$

3. gradients are $\mathcal{S}$-orthogonal if far-enough apart: $\mathbf{r}_j^T\ \mathcal{S}\ \mathbf{r}_k = 0 \qquad \forall\ |j-k| \geq 2$

4. conjugated gradients are $\mathcal{S}^2$-orthogonal if far-enough apart: $\mathbf{d}_j^T\ \mathcal{S}^2\ \mathbf{d}_k = 0\ \forall\ |j-k| \geq 2$

5. $\mathbf{r}_k^T\mathbf{r}_k\ =\ -s_{k-1}\left(\mathbf{r}_{k-1}^T\ \mathbf{X}_{k-1}^T\ \mathbf{X}_{k-1}\ \mathbf{r}_k\right)\ =\ \mathbf{d}_1^T\mathbf{d}_k$

6. $\mathbf{X}^T\mathbf{X}\ \mathbf{d}_k = -\mathbf{X}_k^T\mathbf{X}_k\ \mathbf{r}_k \qquad$ (e.g. Conjugate Gradient and PLS are equivalent)

The first two properties are based on the construction of the gradients and conjugated gradients, so we simply provide examples of these. We verify the remaining properties, then return to justifying the PLS and Conjugate Gradient relationships.

1. We begin by giving examples of how gradients are orthogonal.

$$\begin{aligned}
\mathbf{r}_1^T\mathbf{r}_2 &= \mathbf{r}_1^T(\mathbf{r}_1 + s_1\mathcal{S}\mathbf{d}_1) \;=\; \mathbf{r}_1^T\mathbf{r}_1 - s_1\,\mathbf{d}_1^T\mathcal{S}\mathbf{d}_1 \;=\; \mathbf{r}_1^T\mathbf{r}_1 - \mathbf{r}_1^T\mathbf{r}_1 \;=\; 0\\
\mathbf{r}_1^T\mathbf{r}_3 &= \mathbf{r}_1^T(\mathbf{r}_2 + s_2\mathcal{S}\mathbf{d}_2) \;=\; \mathbf{r}_1^T\mathbf{r}_2 - s_2\,\mathbf{d}_1^T\mathcal{S}\mathbf{d}_2 \;=\; 0\\
\mathbf{r}_2^T\mathbf{r}_3 &= (\mathbf{r}_1 + s_1\mathcal{S}\mathbf{d}_1)^T\;\mathbf{r}_3 \;=\; \mathbf{r}_1^T\mathbf{r}_3 - s_1\,\mathbf{r}_1^T\mathcal{S}\mathbf{r}_3 \;=\; 0\\
\mathbf{r}_2^T\mathbf{r}_4 &= (\mathbf{r}_1 + s_1\mathcal{S}\mathbf{d}_1)^T\;\mathbf{r}_4 \;=\; \mathbf{r}_1^T\mathbf{r}_4 - s_1\,\mathbf{r}_1^T\mathcal{S}\mathbf{r}_4 \;=\; 0\\
\mathbf{r}_3^T\mathbf{r}_4 &= (\mathbf{r}_2 + s_2\mathcal{S}\mathbf{d}_2)^T\;\mathbf{r}_4 \;=\; \mathbf{r}_2^T\mathbf{r}_4 + s_2\,\mathbf{d}_2^T\mathcal{S}\mathbf{r}_4 \;=\; s_2\left(-\mathbf{r}_2 + \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\mathbf{d}_1\right)^T\mathcal{S}\mathbf{r}_4\\
&= s_2\left(-\mathbf{r}_2^T\mathcal{S}\mathbf{r}_4 - \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\,\mathbf{r}_1^T\mathcal{S}\mathbf{r}_4\right) \;=\; 0\\
r_4^T\mathbf{r}_6 &= (\mathbf{r}_3 + s_3\mathcal{S}\mathbf{d}_3)^T\;\mathbf{r}_6 \;=\; \mathbf{r}_3^T\mathbf{r}_6 + s_3\,\mathbf{d}_3^T\mathcal{S}\mathbf{r}_6 \;=\; s_3\left(-\mathbf{r}_3^T\mathcal{S}\mathbf{r}_6 + \frac{\mathbf{r}_3^T\mathbf{r}_3}{\mathbf{r}_2^T\mathbf{r}_2}\,\mathbf{d}_2^T\mathcal{S}\mathbf{r}_6\right)\\
&= s_3\left(-\mathbf{r}_3^T\mathcal{S}\mathbf{r}_6 + \frac{\mathbf{r}_3^T\mathbf{r}_3}{\mathbf{r}_2^T\mathbf{r}_2}\left[-\mathbf{r}_2^T - \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\mathbf{r}_1^T\right]\mathcal{S}\mathbf{r}_6\right)\\
&= s_3\left(-\mathbf{r}_3^T\mathcal{S}\mathbf{r}_6 - \frac{\mathbf{r}_3^T\mathbf{r}_3}{\mathbf{r}_2^T\mathbf{r}_2}\,\mathbf{r}_2^T\mathcal{S}\mathbf{r}_6 - \frac{\mathbf{r}_3^T\mathbf{r}_3}{\mathbf{r}_1^T\mathbf{r}_1}\,\mathbf{r}_1^T\mathcal{S}\mathbf{r}_6\right) = 0
\end{aligned}$$

2. Now we prove that gradients are $\mathcal{S}$-orthogonal if they are far-enough (2 or more) apart. For any $k < j$ such that $|k-j| \geq 2$,

$$\begin{aligned}
\mathbf{r}_{k+1}^T\mathbf{r}_j &= (\mathbf{r}_k + s_k\mathcal{S}\mathbf{d}_k)^T\,\mathbf{r}_j \;=\; \mathbf{r}_k^T\mathbf{r}_j + s_k\,\mathbf{d}_k^T\mathcal{S}\mathbf{r}_j \;=\; s_k\left(-\mathbf{r}_k + \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\mathbf{d}_{k-1}\right)^T\mathcal{S}\mathbf{r}_j\\
&= s_k\left(-\mathbf{r}_k^T\mathcal{S}\mathbf{r}_j + \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\,\mathbf{d}_{k-1}^T\mathcal{S}\mathbf{r}_j\right)\\
&= s_k\left(-\mathbf{r}_k^T\mathcal{S}\mathbf{r}_j - \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\,\mathbf{r}_{k-1}^T\mathcal{S}\mathbf{r}_j - \cdots - \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_1^T\mathbf{r}_1}\,\mathbf{r}_1^T\mathcal{S}\mathbf{r}_j\right) \;=\; 0 \qquad \text{by gradient defn.}\\
\implies \mathbf{r}_k^T\mathcal{S}\mathbf{r}_j &= -\mathbf{r}_{k+1}^T\mathbf{r}_j/s_k \;-\frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\,\mathbf{r}_{k-1}^T\mathcal{S}\mathbf{r}_j - \cdots - \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_1^T\mathbf{r}_1}\,\mathbf{r}_1^T\mathcal{S}\mathbf{r}_j \;=\; 0 \qquad \text{by 3.}
\end{aligned}$$

Note that this orthogonal property does not hold for gradients that are consecutive; the orthogonality proof relies on $\mathbf{r}_{k+1}^T \mathbf{r}_j$ being 0, which does not hold if $k = j - 1$.

3. Next, we give examples of how conjugate gradients are $\mathcal{S}$-orthogonal.

$$\begin{aligned}
\mathbf{d}_1^T \mathcal{S} \mathbf{d}_2 &= \mathbf{r}_1^T \mathcal{S} \left(\left[1 + \frac{\mathbf{r}_2^T \mathbf{r}_2}{\mathbf{r}_1^T \mathbf{r}_1} \right] \mathbf{I}_p - s_1 \mathcal{S} \right) \mathbf{r}_1 \;=\; s_1^2 \, \frac{\mathbf{r}_1^T \mathcal{S}^2 \mathbf{r}_1}{\mathbf{r}_1^T \mathbf{r}_1} \, \mathbf{r}_1^T \mathcal{S} \mathbf{r}_1 - s_1 \, \mathbf{r}_1^T \mathcal{S}^2 \mathbf{r}_1 \\
&= s_1 \, \mathbf{r}_1^T \mathcal{S}^2 \mathbf{r}_1 \left(s_1 \, \frac{\mathbf{r}_1^T \mathcal{S} \mathbf{r}_1}{\mathbf{r}_1^T \mathbf{r}_1} - 1 \right) \;=\; 0 \\
\mathbf{d}_1^T \mathcal{S} \mathbf{d}_3 &= \left(\mathbf{r}_1^T \mathbf{r}_3 - \mathbf{r}_1^T \mathbf{r}_4 \right) / s_3 \;=\; 0 \\
\mathbf{d}_2^T \mathcal{S} \mathbf{d}_3 &= \left(-\mathbf{r}_2 + \frac{\mathbf{r}_2^T \mathbf{r}_2}{\mathbf{r}_1^T \mathbf{r}_1} \mathbf{d}_1 \right)^T \mathcal{S} \mathbf{d}_3 \;=\; -\mathbf{r}_2^T \mathcal{S} \mathbf{d}_3 \;=\; \mathbf{d}_1^T \mathcal{S} \mathbf{d}_3 - s_1 \, \mathbf{d}_1^T \mathcal{S}^2 \mathbf{d}_3 \;=\; 0 \\
\mathbf{d}_2^T \mathcal{S} \mathbf{d}_4 &= \left(-\mathbf{r}_2 + \frac{\mathbf{r}_2^T \mathbf{r}_2}{\mathbf{r}_1^T \mathbf{r}_1} \mathbf{d}_1 \right)^T \mathcal{S} \mathbf{d}_4 \;=\; -\mathbf{r}_2^T \mathcal{S} \mathbf{d}_4 \;=\; \mathbf{d}_1^T \mathcal{S} \mathbf{d}_4 - s_1 \, \mathbf{d}_1^T \mathcal{S}^2 \mathbf{d}_4 \;=\; 0 \\
\mathbf{d}_3^T \mathcal{S} \mathbf{d}_4 &= \left(-\mathbf{r}_3 + \frac{\mathbf{r}_3^T \mathbf{r}_3}{\mathbf{r}_2^T \mathbf{r}_2} \mathbf{d}_2 \right)^T \mathcal{S} \mathbf{d}_4 \;=\; -\mathbf{r}_3^T \mathcal{S} \mathbf{d}_4 \;=\; -\mathbf{r}_2^T \mathcal{S} \mathbf{d}_4 - s_2 \, \mathbf{d}_2^T \mathcal{S}^2 \mathbf{d}_4 \;=\; 0 \\
\mathbf{d}_4^T \mathcal{S} \mathbf{d}_6 &= \left(-\mathbf{r}_4 + \frac{\mathbf{r}_4^T \mathbf{r}_4}{\mathbf{r}_3^T \mathbf{r}_3} \mathbf{d}_3 \right)^T \mathcal{S} \mathbf{d}_6 \;=\; -\mathbf{r}_4^T \mathcal{S} \mathbf{d}_6 \;=\; -\mathbf{r}_3^T \mathcal{S} \mathbf{d}_6 - s_3 \, \mathbf{d}_3^T \mathcal{S}^2 \mathbf{d}_6 \;=\; 0
\end{aligned}$$

4. Now we prove that conjugated gradients are $\mathcal{S}^2$-orthogonal if they are far-enough (2 or more) apart. For any $k < j$ such that $|k - j| \geq 2$,

$$\begin{aligned}
\mathbf{d}_{k+1}^T \mathcal{S} \mathbf{d}_j &= \left(-\mathbf{r}_{k+1} + \frac{\mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{r}_k} \mathbf{d}_k \right)^T \mathcal{S} \mathbf{d}_j \;=\; -\mathbf{r}_{k+1}^T \mathcal{S} \mathbf{d}_j \;=\; -\mathbf{r}_k^T \mathcal{S} \mathbf{d}_j - s_k \, \mathbf{d}_k^T \mathcal{S}^2 \mathbf{d}_j \\
&= \mathbf{d}_k^T \mathcal{S} \mathbf{d}_j - s_k \, \mathbf{d}_k^T \mathcal{S}^2 \mathbf{d}_j \;=\; 0 \qquad \text{by definition of conjugated gradients.} \\
\implies \mathbf{d}_k^T \mathcal{S}^2 \mathbf{d}_j &= \left(\mathbf{d}_k^T \mathcal{S} \mathbf{d}_j - \mathbf{d}_{k+1}^T \mathcal{S} \mathbf{d}_j \right) / s_k \;=\; 0
\end{aligned}$$

Again, this orthogonality property does not hold for conjugated gradients that are

consecutive; the proof relies on $\mathbf{d}_{k+1}^T \mathcal{S} \mathbf{d}_j$ being 0, which does not hold if $k = j - 1$.

5. Next, we show that $\mathbf{r}_k^T \mathbf{r}_k = s_{k-1} \left(\mathbf{d}_{k-1}^T \mathcal{S} \mathbf{r}_k \right) = -s_{k-1} \left(\mathbf{r}_{k-1}^T \ \mathbf{X}_{k-1}^T \ \mathbf{X}_{k-1} \ \mathbf{r}_k \right)$ for any k.

$$\mathbf{r}_k^T \mathbf{r}_k \ = \ (\mathbf{r}_{k-1} + s_{k-1} \ \mathcal{S} \mathbf{d}_{k-1})^T \mathbf{r}_k \ = \ s_{k-1} \left(\mathbf{d}_{k-1}^T \mathcal{S} \mathbf{r}_k \right) \ \stackrel{6.}{=} \ -s_{k-1} \left(\mathbf{r}_{k-1}^T \ \mathbf{X}_{k-1}^T \ \mathbf{X}_{k-1} \ \mathbf{r}_k \right).$$

For example, $\mathbf{r}_1^T \mathcal{S} \mathbf{r}_2 = \mathbf{r}_1^T \mathcal{S} \mathbf{r}_1 - \frac{\mathbf{r}_1^T \mathbf{r}_1}{\mathbf{r}_1^T \mathcal{S} \mathbf{r}_1} (\mathbf{r}_1^T \mathcal{S}^2 \mathbf{r}_1)$, so $s_1 (\mathbf{r}_1^T \mathcal{S} \mathbf{r}_2) = \mathbf{r}_1^T \mathbf{r}_1 - s_1^2 \ (\mathbf{r}_1^T \mathcal{S}^2 \mathbf{r}_1) = -\mathbf{r}_2^T \mathbf{r}_2$.

Additionally, we show that $\mathbf{r}_k^T \mathbf{r}_k = \mathbf{d}_1^T \mathbf{d}_k$ for any k. By definition and induction,

$$\mathbf{d}_1^T \mathbf{d}_k \ = \ -\mathbf{d}_1^T \mathbf{r}_k + \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \ \mathbf{d}_1^T \mathbf{d}_{k-1} \ = \ \mathbf{r}_1^T \mathbf{r}_k + \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \ (\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}) \ = \ \mathbf{r}_k^T \mathbf{r}_k.$$

6. Finally, we show that $\mathbf{X}^T \mathbf{X} \ \mathbf{d}_k = -\mathbf{X}_k^T \mathbf{X}_k \ \mathbf{r}_k$, which is the main statement needed to equate PLS with CG. To make this proof more concise, we highlight some useful relationships.

$$\mathbf{d}_k^T \mathcal{S} \mathbf{d}_k \ = \ -\mathbf{r}_k^T \mathcal{S} \mathbf{d}_k \ = \ -\mathbf{r}_{k-1}^T \mathcal{S} \mathbf{d}_k - s_{k-1} \ \mathbf{d}_{k-1}^T \mathcal{S}^2 \mathbf{d}_k \ = \mathbf{d}_{k-1}^T \mathcal{S} \mathbf{d}_k - s_{k-1} \ \mathbf{d}_{k-1}^T \mathcal{S}^2 \mathbf{d}_k \ = \ -s_{k-1} \ \mathbf{d}_{k-1}^T \mathcal{S}^2 \mathbf{d}_k$$

$$\implies s_k \ = \ \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{d}_k^T \mathcal{S} \mathbf{d}_k} \ = \ \frac{s_{k-1} \ \mathbf{d}_{k-1}^T \mathcal{S} \mathbf{r}_k}{-s_{k-1} \ \mathbf{d}_{k-1}^T \mathcal{S}^2 \mathbf{d}_k} \ = \ \frac{-\mathbf{d}_{k-1}^T \mathcal{S} \mathbf{r}_k}{\mathbf{d}_{k-1}^T \mathcal{S}^2 \mathbf{d}_k}$$

Previously we showed that $-s_k \left(\mathbf{r}_k^T \ \mathbf{X}_k^T \ \mathbf{X}_k \ \mathbf{r}_j\right) = r_j^T \mathbf{r}_j$ when $j = k+1$.

$$\begin{aligned}
\text{If } j = k+2, \quad -s_k \left(\mathbf{r}_k^T \ \mathbf{X}_k^T \ \mathbf{X}_k \ \mathbf{r}_j\right) &= -s_k \left(\mathbf{r}_k^T \ \mathbf{X}_k^T \ \mathbf{X}_k \ \mathbf{r}_{k+1}\right) - s_k s_{k+1} \left(\mathbf{r}_k^T \ \mathbf{X}_k^T \ \mathbf{X}_k \mathcal{S} \mathbf{d}_{k+1}\right) \\
&= \mathbf{r}_{k+1}^T \mathbf{r}_{k+1} + s_{k+1} s_k \left(\mathbf{d}_k^T \mathcal{S}^2 \mathbf{d}_{k+1}\right) \\
&= \mathbf{r}_{k+1}^T \mathbf{r}_{k+1} + \frac{-\mathbf{r}_{k+1}^T \mathbf{r}_{k+1} \ \left(\mathbf{d}_{k+1}^T \mathcal{S} \mathbf{d}_{k+1}\right)}{\mathbf{d}_{k+1}^T \mathcal{S} \mathbf{d}_{k+1}} \ = \ 0
\end{aligned}$$

$$\begin{aligned}
\text{If } j > k+2, \quad -s_k \left(\mathbf{r}_k^T \ \mathbf{X}_k^T \ \mathbf{X}_k \ \mathbf{r}_j\right) &= -s_k \left(\mathbf{r}_k^T \ \mathbf{X}_k^T \ \mathbf{X}_k \ \mathbf{r}_{j-1}\right) - s_k s_{k+1} \left(\mathbf{r}_k^T \ \mathbf{X}_k^T \ \mathbf{X}_k \mathcal{S} \mathbf{d}_{j-1}\right) \\
&= 0 \ + \ s_k s_{k+1} \ \left(\mathbf{d}_k^T \mathcal{S}^2 \mathbf{d}_{j-1}\right) \ = \ 0
\end{aligned}$$

For ease, we also use the conjugated gradient relationship (in different notation):

$$-d_k \ = \ \mathbf{r}_k - \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \ \mathbf{d}_{k-1} \ = \ \mathbf{r}_k \ + \ \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \ \mathbf{r}_{k-1} \ + \ \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-2}^T \mathbf{r}_{k-2}} \ \mathbf{r}_{k-2} \ + \cdots + \ \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_1^T \mathbf{r}_1} \ \mathbf{r}_1.$$

Now we can more clearly prove $\mathbf{X}^T\mathbf{X} \ \mathbf{d}_k \ = \ -\mathbf{X}_k^T \mathbf{X}_k \ \mathbf{r}_k$ by induction, for any $k \leq \text{rank}(\mathbf{X}^T\mathbf{X})$.

When $k = 1$, by definition we have that $-\mathbf{X}_1^T \mathbf{X}_1 \ \mathbf{r}_1 = -\mathbf{X}^T\mathbf{X} \ \mathbf{r}_1 = \mathbf{X}^T\mathbf{X} \ \mathbf{d}_1$.

$$\begin{aligned}
\text{For } k=2,\ \mathbf{X}_2^T\mathbf{X}_2\ \mathbf{r}_2 = \ & \mathbf{X}^T\left(\mathbf{I}_n - \frac{\mathbf{t}_1\mathbf{t}_1^T}{\mathbf{t}_1^T\mathbf{t}_1}\right)\mathbf{X}\ \mathbf{r}_2 \ = \ \mathbf{X}^T\left(\mathbf{I}_n - s_1\ \mathbf{X}_1\ \frac{\mathbf{r}_1\mathbf{r}_1^T}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{X}_1^T\right)\mathbf{X}\ \mathbf{r}_2 \\
= \ & \mathbf{X}^T\mathbf{X}\left[\mathbf{I}_n - s_1\ \frac{\mathbf{r}_1\mathbf{r}_1^T}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{X}^T\mathbf{X}\right]\ \mathbf{r}_2 \\
= \ & \mathbf{X}^T\mathbf{X}\left[\mathbf{r}_2 - \frac{s_1}{\mathbf{r}_1^T\mathbf{r}_1}\ (\mathbf{r}_1^T\mathcal{S}\mathbf{r}_2)\ \mathbf{r}_1\right] \\
= \ & \mathbf{X}^T\mathbf{X}\left[\mathbf{r}_2 + \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\right] \ = \ -\mathbf{X}^T\mathbf{X}\ \mathbf{d}_2.
\end{aligned}$$

For k=3,

$$\begin{aligned}
\mathbf{X}_3^T\mathbf{X}_3\ \mathbf{r}_3 = \ & \mathbf{X}_2^T\left(\mathbf{I}_n - s_2\ \mathbf{X}_2\ \frac{\mathbf{r}_2\mathbf{r}_2^T}{\mathbf{r}_2^T\mathbf{r}_2}\ \mathbf{X}_2^T\right)\mathbf{X}_2\ \mathbf{r}_3 \ = \ \left[\mathbf{X}_2^T\mathbf{X}_2 - s_2\ \mathbf{X}_2^T\mathbf{X}_2\ \frac{\mathbf{r}_2\mathbf{r}_2^T}{\mathbf{r}_2^T\mathbf{r}_2}\ \mathbf{X}_2^T\mathbf{X}_2\right]\ \mathbf{r}_3 \\
= \ & \mathcal{S}\left[\mathbf{I}_n - s_1\ \frac{\mathbf{r}_1\mathbf{r}_1^T}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{X}^T\mathbf{X}\ -\ \frac{s_2}{\mathbf{r}_2^T\mathbf{r}_2}\left\{\mathbf{I}_n - s_1\ \frac{\mathbf{r}_1\mathbf{r}_1^T}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{X}^T\mathbf{X}\right\}\mathbf{r}_2\ \mathbf{r}_2^T\ \mathbf{X}_2^T\mathbf{X}_2\right]\ \mathbf{r}_3 \\
= \ & \mathcal{S}\left[\mathbf{I}_n - \frac{s_1}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\ \mathbf{r}_1^T\mathcal{S}\ -\ \frac{s_2}{\mathbf{r}_2^T\mathbf{r}_2}\left\{\mathbf{r}_2\ \mathbf{r}_2^T\ \mathbf{X}_2^T\mathbf{X}_2 + \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\ \mathbf{r}_2^T\mathbf{X}_2^T\mathbf{X}_2\right\}\right]\ \mathbf{r}_3 \\
= \ & \mathcal{S}\left[\mathbf{r}_3 - \frac{s_1}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\ (\mathbf{r}_1^T\mathcal{S}\mathbf{r}_3)\ -\ \frac{s_2}{\mathbf{r}_2^T\mathbf{r}_2}\left\{\mathbf{r}_2\ (\mathbf{r}_2^T\mathbf{X}_2^T\mathbf{X}_2\mathbf{r}_3) + \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\ (\mathbf{r}_2^T\mathbf{X}_2^T\mathbf{X}_2\mathbf{r}_3)\right\}\right] \\
= \ & \mathcal{S}\left[\mathbf{r}_3\ +\ \frac{\mathbf{r}_3^T\mathbf{r}_3}{\mathbf{r}_2^T\mathbf{r}_2}\ \mathbf{r}_2\ +\ \frac{\mathbf{r}_3^T\mathbf{r}_3}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\right] \ = \ -\mathbf{X}^T\mathbf{X}\ \mathbf{d}_3.
\end{aligned}$$

Now assume that $\mathbf{X}^T\mathbf{X}\ \mathbf{d}_k = -\mathbf{X}_k^T\mathbf{X}_k\ \mathbf{r}_k$, by a similar pattern. Then,

$$
\begin{aligned}
\mathbf{X}_{k+1}^T\mathbf{X}_{k+1}\ \mathbf{r}_{k+1} &= \mathbf{X}_k^T\left(\mathbf{I}_n - s_k\ \mathbf{X}_k\ \frac{\mathbf{r}_k\mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k}\ \mathbf{X}_k^T\right)\mathbf{X}_k\ \mathbf{r}_{k+1} \;=\; \left[\mathbf{X}_k^T\mathbf{X}_k - s_k\ \mathbf{X}_k^T\mathbf{X}_k\ \frac{\mathbf{r}_k\mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k}\ \mathbf{X}_k^T\mathbf{X}_k\right]\ \mathbf{r}_{k+1} \\
&= \mathcal{S}\Bigg[\mathbf{I}_n - \frac{s_1}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\ \mathbf{r}_1^T\mathcal{S} \;-\; \frac{s_2}{\mathbf{r}_2^T\mathbf{r}_2}\left(\mathbf{r}_2 + \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\right)\ \mathbf{r}_2^T\mathbf{X}_2^T\mathbf{X}_2 \\
&\qquad - \frac{s_3}{\mathbf{r}_3^T\mathbf{r}_3}\left[\mathbf{r}_3 \;+\; \frac{\mathbf{r}_3^T\mathbf{r}_3}{\mathbf{r}_2^T\mathbf{r}_2}\left(\mathbf{r}_2 + \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\right)\right]\ \mathbf{r}_3^T\mathbf{X}_3^T\mathbf{X}_3 - \cdots \\
&\quad - \frac{s_k}{\mathbf{r}_k^T\mathbf{r}_k}\Bigg\{\mathbf{I}_n - \frac{s_1}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\ \mathbf{r}_1^T\mathcal{S} \;-\; \frac{s_2}{\mathbf{r}_2^T\mathbf{r}_2}\left(\mathbf{r}_2 + \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\right)\ \mathbf{r}_2^T\mathbf{X}_2^T\mathbf{X}_2 \\
&\qquad - \frac{s_3}{\mathbf{r}_3^T\mathbf{r}_3}\left[\mathbf{r}_3 \;+\; \frac{\mathbf{r}_3^T\mathbf{r}_3}{\mathbf{r}_2^T\mathbf{r}_2}\left(\mathbf{r}_2 + \frac{\mathbf{r}_2^T\mathbf{r}_2}{\mathbf{r}_1^T\mathbf{r}_1}\ \mathbf{r}_1\right)\right]\ \mathbf{r}_3^T\mathbf{X}_3^T\mathbf{X}_3 - \cdots\Bigg\}\ \mathbf{r}_k\ \mathbf{r}_k^T\mathbf{X}_k^T\mathbf{X}_k\Bigg]\ \mathbf{r}_{k+1} \\
&= \mathcal{S}\left[\mathbf{r}_{k+1} - \frac{s_k(\mathbf{r}_k^T\mathbf{X}_k^T\mathbf{X}_k\mathbf{r}_{k+1})}{\mathbf{r}_k^T\mathbf{r}_k}\left\{\mathbf{r}_k + \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\left(\mathbf{r}_{k-1} + \frac{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}{\mathbf{r}_{k-2}^T\mathbf{r}_{k-2}}\ [\mathbf{r}_{k-2} + \cdots]\right)\right\}\right] \\
&= \mathcal{S}\left[\mathbf{r}_{k+1} + \frac{\mathbf{r}_{k+1}^T\mathbf{r}_{k+1}}{\mathbf{r}_k^T\mathbf{r}_k}\left\{\mathbf{r}_k + \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\left(\mathbf{r}_{k-1} + \frac{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}{\mathbf{r}_{k-2}^T\mathbf{r}_{k-2}}\ [\mathbf{r}_{k-2} + \cdots]\right)\right\}\right] \\
&= -\mathbf{X}^T\mathbf{X}\ \mathbf{d}_{k+1}.
\end{aligned}
$$

As $\mathbf{X}^T\mathbf{X}\ \mathbf{d}_{k+1} = -\mathbf{X}_{k+1}^T\mathbf{X}_{k+1}\ \mathbf{r}_{k+1}$, the proof of the induction step is complete. Therefore, relationship 6. holds for any $k \leq \text{rank}(\mathbf{X}^T\mathbf{X})$. □

Now we return to the original task of justifying the relationships between PLS and Conjugate Gradient. We list the 5 substantial equalities again for ease of reading, and then we plug these results into the remaining equalities.

A. (residual) gradients: $\tilde{\mathbf{w}}_k = \mathbf{X}_k^T \mathbf{y}_k = -\mathbf{r}_k$

B. conjugated gradients: $\tilde{\mathbf{w}}_k + \frac{\tilde{\mathbf{w}}_k^T \tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_{k-1}^T \tilde{\mathbf{w}}_{k-1}} \tilde{\mathbf{w}}_{k-1} + \frac{\tilde{\mathbf{w}}_k^T \tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_{k-2}^T \tilde{\mathbf{w}}_{k-2}} \tilde{\mathbf{w}}_{k-2} + \cdots + \frac{\tilde{\mathbf{w}}_k^T \tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_1^T \tilde{\mathbf{w}}_1} \tilde{\mathbf{w}}_1 = \mathbf{d}_k$

C. component variance: $\mathbf{t}_k^T \mathbf{t}_k = 1/s_k$

D. orthogonal projection: $\left(\mathbf{I}_n - \frac{\mathbf{t}_k \mathbf{t}_k^T}{\mathbf{t}_k^T \mathbf{t}_k}\right) = \left(\mathbf{I}_n - s_k \, \mathbf{X}_k \, \frac{\mathbf{r}_k \mathbf{r}_k^T}{\mathbf{r}_k^T \mathbf{r}_k} \, \mathbf{X}_k^T\right)$

E. X-scale gradients: $\mathbf{R}_k^{(k)} = \mathbf{d}_k / \sqrt{\mathbf{r}_k^T \mathbf{r}_k}$

For the first component (k=1) these easily hold, since by the initialization steps:

$$\tilde{\mathbf{w}}_1 = \mathbf{X}^T \mathbf{y} = \mathbf{d}_1 = -\mathbf{r}_1.$$

$$\text{Therefore} \quad \mathbf{w}_1 = \mathbf{R}_1^{(1)} = -\mathbf{r}_1 / \sqrt{\mathbf{r}_1^T \mathbf{r}_1} \quad \text{and} \quad \mathbf{t}_1 = -\mathbf{X}\mathbf{r}_1 / \sqrt{\mathbf{r}_1^T \mathbf{r}_1}.$$

$$\text{Then,} \quad \mathbf{t}_1^T \mathbf{t}_1 \;=\; \mathbf{r}_1^T \, \mathbf{X}^T \, \mathbf{X} \, \mathbf{r}_1 / (\mathbf{r}_1^T \mathbf{r}_1) \;=\; \mathbf{d}_1^T \, \mathbf{X}^T \mathbf{X} \, \mathbf{d}_1 / (\mathbf{r}_1^T \mathbf{r}_1) \;=\; 1/s_1.$$

$$\text{Last,} \quad \mathbf{t}_1 \mathbf{t}_1^T \;= \mathbf{X} \, \mathbf{r}_1 \, \mathbf{r}_1^T \, \mathbf{X}^T / (\mathbf{r}_1^T \mathbf{r}_1) \;=\; \mathbf{X} \frac{\mathbf{r}_1 \, \mathbf{r}_1^T}{\mathbf{r}_1^T \mathbf{r}_1} \mathbf{X}^T, \quad \text{so it holds that}$$

$$\left(\mathbf{I}_n - \frac{\mathbf{t}_1 \mathbf{t}_1^T}{\mathbf{t}_1^T \mathbf{t}_1}\right) \;=\; \left(\mathbf{I}_n - s_1 \, \mathbf{X} \frac{\mathbf{r}_1 \, \mathbf{r}_1^T}{\mathbf{r}_1^T \mathbf{r}_1} \mathbf{X}^T\right).$$

For the future components, more algebra is required.

A. We start by proving that $\tilde{\mathbf{w}}_k = -\mathbf{r}_k$. For two components,

$$\begin{aligned}\tilde{\mathbf{w}}_2 = \mathbf{X}_2^T\mathbf{y}_2 \;&=\; \mathbf{X}^T\left(\mathbf{I}_n - \frac{\mathbf{t}_1\mathbf{t}_1^T}{\mathbf{t}_1^T\mathbf{t}_1}\right)\mathbf{y} \;=\; \mathbf{X}^T\mathbf{y} \;-\; \frac{1}{\mathbf{t}_1^T\mathbf{t}_1}\,\mathbf{X}^T\mathbf{t}_1\,\mathbf{t}_1^T\mathbf{y} \\ &=\; -\mathbf{r}_1 \;-\; s_1\,\mathbf{X}^T\left[X\frac{\mathbf{r}_1\,\mathbf{r}_1^T}{\mathbf{r}_1^T\mathbf{r}_1}\mathbf{X}^T\right]\,\mathbf{y} \;=\; -\mathbf{r}_1 \;+\; s_1\,\mathbf{X}^T\left[X\frac{\mathbf{r}_1\,\mathbf{r}_1^T}{\mathbf{r}_1^T\mathbf{r}_1}\mathbf{r}_1\right] \\ &=\; -\mathbf{r}_1 \;+\; s_1\,\mathbf{X}^T\,[X\mathbf{r}_1] \;=\; -\mathbf{r}_1 \;-\; s_1\,\mathbf{X}^T\mathbf{X}\mathbf{d}_1 \;=\; -\mathbf{r}_2.\end{aligned}$$

For three components, using that $\mathbf{t}_2\mathbf{t}_2^T \;= \mathbf{X}_2\,\mathbf{r}_2\,\mathbf{r}_2^T\,\mathbf{X}_2^T/(\mathbf{r}_2^T\mathbf{r}_2) \;=\; \mathbf{X}_2\frac{\mathbf{r}_2\,\mathbf{r}_2^T}{\mathbf{r}_2^T\mathbf{r}_2}\mathbf{X}_2^T$,

$$\begin{aligned}\tilde{\mathbf{w}}_3 = \mathbf{X}_3^T\mathbf{y}_3 \;&=\; \mathbf{X}_2^T\left(\mathbf{I}_n - \frac{\mathbf{t}_2\mathbf{t}_2^T}{\mathbf{t}_2^T\mathbf{t}_2}\right)\mathbf{y}_2 \;=\; \mathbf{X}_2^T\mathbf{y}_2 \;-\; s_2\,\mathbf{X}_2^T\left[\mathbf{X}_2\frac{\mathbf{r}_2\,\mathbf{r}_2^T}{\mathbf{r}_2^T\mathbf{r}_2}\mathbf{X}_2^T\right]\mathbf{y}_2 \\ &=\; -\mathbf{r}_2 \;+\; s_2\,\mathbf{X}_2^T\mathbf{X}_2\,\mathbf{r}_2 \;=\; -\mathbf{r}_2 \;-\; s_2\,\mathbf{X}^T\mathbf{X}\,\mathbf{d}_2 \quad=\quad -\mathbf{r}_3.\end{aligned}$$

Assuming that $\tilde{\mathbf{w}}_k = -\mathbf{r}_k$ by a similar pattern, we have that

$$\begin{aligned}\tilde{\mathbf{w}}_{k+1} = \mathbf{X}_{k+1}^T\mathbf{y}_{k+1} \;&=\; \mathbf{X}_k^T\left(\mathbf{I}_n - \frac{\mathbf{t}_k\mathbf{t}_k^T}{\mathbf{t}_k^T\mathbf{t}_k}\right)\mathbf{y}_k \;=\; \mathbf{X}_k^T\mathbf{y}_k \;-\; s_k\,\mathbf{X}_k^T\left[\mathbf{X}_k\frac{\mathbf{r}_k\,\mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k}\mathbf{X}_k^T\right]\mathbf{y}_k \\ &=\; -\mathbf{r}_k \;+\; s_k\,\mathbf{X}_k^T\mathbf{X}_k\,\mathbf{r}_k \;=\; -\mathbf{r}_k \;-\; s_k\,\mathbf{X}^T\mathbf{X}\,\mathbf{d}_k \quad=\quad -\mathbf{r}_{k+1}.\end{aligned}$$

Therefore, since the induction step holds, $\tilde{\mathbf{w}}_k = -\mathbf{r}_k$ for any $k \leq \text{rank}(\mathbf{X}^T\mathbf{X})$. □

B. Next, we show the relationship between $\mathbf{d}_k$ and $\tilde{\mathbf{w}}_k$, using equality A. and the definition $\mathbf{d}_k = -\mathbf{r}_k + \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \mathbf{d}_{k-1}$.

$$\begin{aligned}
\mathbf{d}_2 = -\mathbf{r}_2 + \frac{\mathbf{r}_2^T \mathbf{r}_2}{\mathbf{r}_1^T \mathbf{r}_1} \mathbf{d}_1 &= \tilde{\mathbf{w}}_2 + \frac{\tilde{\mathbf{w}}_2^T \tilde{\mathbf{w}}_2}{\tilde{\mathbf{w}}_1^T \tilde{\mathbf{w}}_1} \tilde{\mathbf{w}}_1. \\
\mathbf{d}_3 = -\mathbf{r}_3 + \frac{\mathbf{r}_3^T \mathbf{r}_3}{\mathbf{r}_2^T \mathbf{r}_2} \mathbf{d}_2 &= \tilde{\mathbf{w}}_3 + \frac{\tilde{\mathbf{w}}_3^T \tilde{\mathbf{w}}_3}{\tilde{\mathbf{w}}_2^T \tilde{\mathbf{w}}_2} \left(\tilde{\mathbf{w}}_2 + \frac{\tilde{\mathbf{w}}_2^T \tilde{\mathbf{w}}_2}{\tilde{\mathbf{w}}_1^T \tilde{\mathbf{w}}_1} \tilde{\mathbf{w}}_1 \right).
\end{aligned}$$

Assume this relationship holds for step $k-1$. Then,

$$\mathbf{d}_k = -\mathbf{r}_k + \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \mathbf{d}_{k-1} = \tilde{\mathbf{w}}_k + \frac{\tilde{\mathbf{w}}_k^T \tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_{k-1}^T \tilde{\mathbf{w}}_{k-1}} \left\{ \tilde{\mathbf{w}}_{k-1} + \frac{\tilde{\mathbf{w}}_{k-1}^T \tilde{\mathbf{w}}_{k-1}}{\tilde{\mathbf{w}}_{k-2}^T \tilde{\mathbf{w}}_{k-2}} \left[\tilde{\mathbf{w}}_{k-2} + \cdots \left(\tilde{\mathbf{w}}_2 + \frac{\tilde{\mathbf{w}}_2^T \tilde{\mathbf{w}}_2}{\tilde{\mathbf{w}}_1^T \tilde{\mathbf{w}}_1} \tilde{\mathbf{w}}_1 \right) \right] \right\}$$

$$= \tilde{\mathbf{w}}_k + \frac{\tilde{\mathbf{w}}_k^T \tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_{k-1}^T \tilde{\mathbf{w}}_{k-1}} \tilde{\mathbf{w}}_{k-1} + \frac{\tilde{\mathbf{w}}_k^T \tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_{k-2}^T \tilde{\mathbf{w}}_{k-2}} \tilde{\mathbf{w}}_{k-2} + \cdots + \frac{\tilde{\mathbf{w}}_k^T \tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_1^T \tilde{\mathbf{w}}_1} \tilde{\mathbf{w}}_1$$

Therefore, since the induction step holds, equality B. holds for any $k \leq \text{rank}(\mathbf{X}^T \mathbf{X})$. □

C. Next, we show that $s_k = 1/(\mathbf{t}_k^T \mathbf{t}_k)$. We rely on equality A. and relationship 6. For any $k \leq \text{rank}(\mathbf{X}^T \mathbf{X})$,

$$\begin{aligned}
\mathbf{t}_k^T \mathbf{t}_k = \mathbf{w}_k^T \mathbf{X}_k^T \mathbf{X}_k \mathbf{w}_k &= \frac{\mathbf{r}_k^T}{\sqrt{\mathbf{r}_k^T \mathbf{r}_k}} \mathbf{X}_k^T \mathbf{X}_k \frac{\mathbf{r}_k}{\sqrt{\mathbf{r}_k^T \mathbf{r}_k}} = \frac{\mathbf{r}_k^T}{\mathbf{r}_k^T \mathbf{r}_k} \left[\mathbf{X}_k^T \mathbf{X}_k \mathbf{r}_k \right] \\
&= \frac{-\mathbf{r}_k^T \mathcal{S} \mathbf{d}_k}{\mathbf{r}_k^T \mathbf{r}_k} = \frac{\mathbf{d}_k^T \mathcal{S} \mathbf{d}_k}{\mathbf{r}_k^T \mathbf{r}_k} = \frac{1}{s_k}.
\end{aligned}$$

□

D. We now show that $\left(\mathbf{I}_n - \frac{\mathbf{t}_k \mathbf{t}_k^T}{\mathbf{t}_k^T \mathbf{t}_k} \right) = \left(\mathbf{I}_n - s_k \mathbf{X}_k \frac{\mathbf{r}_k \mathbf{r}_k^T}{\mathbf{r}_k^T \mathbf{r}_k} \mathbf{X}_k^T \right)$. Using equality C., we only need to show that $\mathbf{t}_k \mathbf{t}_k^T = \mathbf{X}_k \frac{\mathbf{r}_k \mathbf{r}_k^T}{\mathbf{r}_k^T \mathbf{r}_k} \mathbf{X}_k^T$. For any $k \leq \text{rank}(\mathbf{X}^T \mathbf{X})$,

$$\mathbf{t}_k \mathbf{t}_k^T = \mathbf{X}_k \frac{\mathbf{r}_k}{\sqrt{\mathbf{r}_k^T \mathbf{r}_k}} \frac{\mathbf{r}_k^T}{\sqrt{\mathbf{r}_k^T \mathbf{r}_k}} \mathbf{X}_k^T = \mathbf{X}_k \frac{\mathbf{r}_k \mathbf{r}_k^T}{\mathbf{r}_k^T \mathbf{r}_k} \mathbf{X}_k^T$$

□

Therefore, we can use the relationship shown by Helland to state that

$$\mathbf{X}_{k+1} \;=\; \left[\prod_{m=1}^{k}\left(\mathbf{I}_n - \frac{\mathbf{t}_m\mathbf{t}_m^T}{\mathbf{t}_m^T\mathbf{t}_m}\right)\right]\mathbf{X} \;=\; \left[\prod_{m=1}^{k}\left(\mathbf{I}_n - s_m\;\mathbf{X}_m\;\frac{\mathbf{r}_m\mathbf{r}_m^T}{\mathbf{r}_m^T\mathbf{r}_m}\;\mathbf{X}_m^T\right)\right]\mathbf{X},$$

and similarly, $\mathbf{y}_{k+1} \;=\; \left[\prod_{m=1}^{k}\left(\mathbf{I}_n - s_m\;\mathbf{X}_m\;\frac{\mathbf{r}_m\mathbf{r}_m^T}{\mathbf{r}_m^T\mathbf{r}_m}\;\mathbf{X}_m^T\right)\right]\mathbf{y}.$

E. Finally, we show that $\mathbf{R}_k^{(k)} = \mathbf{d}_k/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$. First, we clarify the notation below. Let $\mathbf{R}_k^{(k)}$ be the kth column of $\mathbf{R}_k = \mathbf{W}_k\left(\mathbf{P}_k^T\mathbf{W}_k\right)^{-1}$, where $\mathbf{W}_k = \Big[\mathbf{w}_1 \;\; \mathbf{w}_2 \;\; \ldots \;\; \mathbf{w}_k\Big]$ and $\mathbf{P}_k = \Big[\mathbf{p}_1 \;\; \ldots \;\; \mathbf{p}_k\Big]$.

First, we use the facts that $\mathbf{w}_k = -\mathbf{r}_k/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$ and

$$\mathbf{p}_k = \mathbf{X}_k^T\mathbf{t}_k/(\mathbf{t}_k^T\mathbf{t}_k) = \frac{-s_k}{\sqrt{\mathbf{r}_k^T\mathbf{r}_k}}\;\mathbf{X}_k^T\mathbf{X}_k\;\mathbf{r}_k = \frac{s_k}{\sqrt{\mathbf{r}_k^T\mathbf{r}_k}}\;\mathbf{X}^T\mathbf{X}\;\mathbf{d}_k \;=\; \frac{\mathbf{r}_{k+1}-\mathbf{r}_k}{\sqrt{\mathbf{r}_k^T\mathbf{r}_k}} := \frac{\mathbf{r}_{k+1}-\mathbf{r}_k}{\|\mathbf{r}_k\|}.$$

$$\left(\mathbf{P}_k^T\mathbf{W}_k\right)^{-1} = \left(\begin{bmatrix} (\mathbf{r}_2-\mathbf{r}_1)^T/\|\mathbf{r}_1\| \\ \ldots \\ (\mathbf{r}_{k+1}-\mathbf{r}_k)^T/\|\mathbf{r}_k\| \end{bmatrix} \quad \begin{bmatrix} \frac{-\mathbf{r}_1}{\|\mathbf{r}_1\|} & \ldots & \frac{-\mathbf{r}_k}{\|\mathbf{r}_k\|} \end{bmatrix}\right)^{-1}$$

$$= \begin{pmatrix} 1 & \frac{-\|\mathbf{r}_2\|}{\|\mathbf{r}_1\|} & 0 & \ldots & 0 \\ 0 & 1 & \frac{-\|\mathbf{r}_3\|}{\|\mathbf{r}_2\|} & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & \cdots & & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & \frac{\|\mathbf{r}_2\|}{\|\mathbf{r}_1\|} & \frac{\|\mathbf{r}_3\|}{\|\mathbf{r}_1\|} & \ldots & \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_1\|} \\ 0 & 1 & \frac{\|\mathbf{r}_3\|}{\|\mathbf{r}_2\|} & \ldots & \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_2\|} \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & 1 \end{pmatrix}$$

$$\mathbf{R}_k = \mathbf{W}_k \left(\mathbf{P}_k^T \mathbf{W}_k\right)^{-1} = \left[\mathbf{w}_1 \quad \mathbf{w}_2 + \frac{\|\mathbf{r}_2\|}{\|\mathbf{r}_1\|}\mathbf{w}_1 \quad \cdots \quad \mathbf{w}_k + \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_{k-1}\|}\mathbf{w}_{k-1} + ... + \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_1\|}\mathbf{w}_1\right]$$

$$= \left[\frac{-\mathbf{r}_1}{\|\mathbf{r}_1\|} \quad \frac{-\mathbf{r}_2}{\|\mathbf{r}_2\|} - \frac{\|\mathbf{r}_2\|\mathbf{r}_1}{\|\mathbf{r}_1\|^2} \quad \cdots \quad \frac{-\mathbf{r}_k}{\|\mathbf{r}_k\|} - \frac{\|\mathbf{r}_k\|\mathbf{r}_{k-1}}{\|\mathbf{r}_{k-1}\|^2} - ... - \frac{\|\mathbf{r}_k\|\mathbf{r}_1}{\|\mathbf{r}_1\|^2}\right]$$

$$\mathbf{R}_k^{(j)} = \left(-\mathbf{r}_j - \frac{\|\mathbf{r}_j\|^2}{\|\mathbf{r}_{j-1}\|^2}\ \mathbf{r}_{j-1} - ... - \frac{\|\mathbf{r}_j\|^2}{\|\mathbf{r}_1\|^2}\ \mathbf{r}_1\right) / \|\mathbf{r}_j\|$$

$$= \mathbf{d}_j / \sqrt{\mathbf{r}_j^T \mathbf{r}_j}$$

$$\implies \mathbf{R}_k^{(k)} = \mathbf{d}_k / \sqrt{\mathbf{r}_k^T \mathbf{r}_k}. \qquad \square$$

The remaining equalities can be seen by plugging in the previous results. For example, see that $q_k = \mathbf{y}_k^T \mathbf{t}_k / (\mathbf{t}_k^T \mathbf{t}_k) = s_k(-\mathbf{r}_k\ \mathbf{w}_k) = s_k \sqrt{\mathbf{r}_k^T \mathbf{r}_k}$, so we obtain that

$$\hat{\boldsymbol{\beta}}_k = \mathbf{R}_k \mathbf{Q}_k^T = \left[\mathbf{R}_k^{(1)} \quad \mathbf{R}_k^{(2)} \quad ... \quad \mathbf{R}_k^{(k)}\right] \left[q_1 \quad ... \quad q_k\right]^T = \sum_{m=1}^{k} s_m \mathbf{d}_m.$$

This concludes the verification that PLS and Conjugate Gradient are equivalent procedures at the kth step.

A.2 Equivalence between WPLS and Conjugate Gradient

As mentioned at the beginning of the last section, Linear Conjugate Gradient solves the general system of linear equations $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is symmetric and positive definite. When solving generalized estimating equations, CG can similarly be applied to the system $(\mathbf{X}^T\mathbf{VX})\ \hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Vz}$. The counterpart for PLS is now weighted PLS (WPLS), and since very similar mechanics are in play to prove the equivalence of these methods, we only briefly show the relationship of the PLS residual variance and covariance matrices. Then we list the algorithms and full relationship between CG and WPLS terms for reference. The product form for the subsequent residuals $\mathbf{X}_k$ and $\mathbf{z}_k$ were proven by Marx [48] as an adaptation of Helland's proof. Again, WLOG we assume $\mathbf{X}$ is centered.

The only change to PLS is that components are now created in a weighted metric. We provide justification that the Weighted PLS variance and covariance matrices take a similar form as the PLS matrices. For any component $k+1$, with $\mathcal{S}_k = \mathbf{X}_k^T\mathbf{V}\mathbf{X}_k$,

$$\begin{aligned}
\mathbf{X}_{k+1}^T\mathbf{V}\mathbf{X}_{k+1} &= \mathbf{X}_k^T\left(\mathbf{I}_n - \frac{\mathbf{t}_k\mathbf{t}_k^T\mathbf{V}}{\mathbf{t}_k^T\mathbf{V}\mathbf{t}_k}\right)^T\ \mathbf{V}\ \left(\mathbf{I}_n - \frac{\mathbf{t}_k\mathbf{t}_k^T\mathbf{V}}{\mathbf{t}_k^T\mathbf{V}\mathbf{t}_k}\right)\ \mathbf{X}_k \\
&= \mathbf{X}_k^T\left(\mathbf{I}_n - \frac{\mathbf{V}\mathbf{t}_k\mathbf{t}_k^T}{\mathbf{t}_k^T\mathbf{V}\mathbf{t}_k}\right)\ \left(\mathbf{V} - \frac{\mathbf{V}\mathbf{t}_k\mathbf{t}_k^T\mathbf{V}}{\mathbf{t}_k^T\mathbf{V}\mathbf{t}_k}\right)\ \mathbf{X}_k\ =\ \mathbf{X}_k^T\left(\mathbf{I}_n - \frac{\mathbf{V}\mathbf{t}_k\mathbf{t}_k^T}{\mathbf{t}_k^T\mathbf{V}\mathbf{t}_k}\right)\ \mathbf{V}\mathbf{X}_k \\
&= \mathcal{S}_k - s_k\ \mathcal{S}_k\ \frac{\mathbf{r}_k\ \mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k}\ \mathcal{S}_k \\
\\
X_{k+1}^T\mathbf{V}\mathbf{z}_{k+1} &= \mathbf{X}_k^T\left(\mathbf{I}_n - \frac{\mathbf{V}\mathbf{t}_k\mathbf{t}_k^T}{\mathbf{t}_k^T\mathbf{V}\mathbf{t}_k}\right)\ \mathbf{V}\mathbf{z}_k \\
&= \mathbf{X}_k^T\mathbf{V}\mathbf{z}_k - s_k\ \mathcal{S}_k\ \frac{\mathbf{r}_k\ \mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k}\ \mathbf{X}_k^T\mathbf{V}\mathbf{z}_k\ =\ -\mathbf{r}_k + s_k\ \mathcal{S}_k\ \mathbf{r}_k\ =\ -\mathbf{r}_{k+1}.
\end{aligned}$$

The last equality—and all the remaining proofs—hold as in the PLS case because we see the same form emerge. For the weighted PLS proof, simply replace $\mathbf{X}^T\mathbf{X}$ with $\mathbf{X}^T\mathbf{VX}$ and $\mathbf{X}^T\mathbf{y}$ with $\mathbf{X}^T\mathbf{Vz}$.

Conjugate gradient with $\hat{\boldsymbol{\beta}}_0 = 0$

Initialize: $\mathbf{r}_0 = -\mathbf{X}^T\mathbf{V}\mathbf{z}$ and $\mathbf{d}_0 = 0$.

For k in 1, ..., rank($\mathbf{X}^T\mathbf{V}\mathbf{X}$) {

$$\mathbf{r}_k = \mathbf{r}_{k-1} + s_{k-1}\ (\mathbf{X}^T\mathbf{V}\mathbf{X})\ \mathbf{d}_{k-1}$$
$$\mathbf{d}_k = -\mathbf{r}_k + \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\ \mathbf{d}_{k-1}$$
$$s_k = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{d}_k^T\ (\mathbf{X}^T\mathbf{V}\mathbf{X})\ \mathbf{d}_k}$$
$$\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\beta}}_{k-1} + s_k\ \mathbf{d}_k$$

}

Weighted Partial Least Squares

Initialize: $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{z}_1 = \mathbf{z}$.

For k in 1, ..., rank($\mathbf{X}^T\mathbf{V}\mathbf{X}$)=K {

$$\tilde{\mathbf{w}}_k = \mathbf{X}_k^T\ \mathbf{V}\ \mathbf{z}_k$$
$$\mathbf{w}_k = \tilde{\mathbf{w}}_k / \sqrt{\tilde{\mathbf{w}}_k^T\ \tilde{\mathbf{w}}_k}$$
$$\mathbf{t}_k = \mathbf{X}_k\ \mathbf{w}_k$$
$$\mathbf{p}_k = \mathbf{X}_k^T\mathbf{V}\mathbf{t}_k / (\mathbf{t}_k^T\ \mathbf{V}\ \mathbf{t}_k)$$
$$q_k = \mathbf{z}_k^T\mathbf{V}\mathbf{t}_k / (\mathbf{t}_k^T\ \mathbf{V}\ \mathbf{t}_k)$$

$$\begin{aligned}\mathbf{X}_{k+1} &= \mathbf{X}_k\ -\ \mathbf{t}_k\mathbf{p}_k^T \\ &=\ \mathbf{X}_k\ -\ \mathbf{t}_k\mathbf{t}_k^T\mathbf{V}\ \mathbf{X}_k/(\mathbf{t}_k^T\mathbf{V}\mathbf{t}_k) \\ &= \left[\prod_{m=1}^{k}\left(\mathbf{I}_n - \frac{\mathbf{t}_m\mathbf{t}_m^T\mathbf{V}}{\mathbf{t}_m^T\mathbf{V}\mathbf{t}_m}\right)\right]\mathbf{X}\end{aligned}$$

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \mathbf{t}_k q_k^T = \left[\prod_{m=1}^{k}\left(\mathbf{I}_n - \frac{\mathbf{t}_m\mathbf{t}_m^T\mathbf{V}}{\mathbf{t}_m^T\mathbf{V}\mathbf{t}_m}\right)\right]\mathbf{z}$$

}

$$\mathbf{W}_K = \begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_K\end{bmatrix}$$
$$\mathbf{P}_K = \begin{bmatrix}\mathbf{p}_1 & \cdots & \mathbf{p}_K\end{bmatrix}$$
$$\mathbf{Q}_K = \begin{bmatrix}q_1 & \cdots & q_K\end{bmatrix}$$
$$\hat{\boldsymbol{\beta}}_K = \mathbf{W}_K\left(\mathbf{P}_K^T\mathbf{W}_K\right)^{-1}\mathbf{Q}_K^T\ :=\ \mathbf{R}_K\ \mathbf{Q}_K^T$$

The following relationships hold at the kth CG and WPLS step, for any $k \in [1, \text{rank}(\mathbf{X}^T\mathbf{V}\mathbf{X})]$, by the proofs in the previous section. For clarity, the terms on the left are from WPLS and those on the right are from CG.

- gradients (weights): $\tilde{\mathbf{w}}_k = -\mathbf{r}_k$ and $\mathbf{w}_k = -\mathbf{r}_k/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- orthogonal projection: $\left(\mathbf{I}_n - \frac{\mathbf{t}_k\mathbf{t}_k^T\mathbf{V}}{\mathbf{t}_k^T\mathbf{V}\mathbf{t}_k}\right) = \left(\mathbf{I}_n - s_k \left[\mathbf{X}_k \frac{\mathbf{r}_k\mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k} \mathbf{X}_k^T\right]\mathbf{V}\right)$
- components: $\mathbf{t}_k = \mathbf{X}_k\,\mathbf{w}_k = -\left[\prod_{m=1}^{k}\left(\mathbf{I}_n - s_m \left[\mathbf{X}_m \frac{\mathbf{r}_m\mathbf{r}_m^T}{\mathbf{r}_m^T\mathbf{r}_m} \mathbf{X}_m^T\right]\mathbf{V}\right)\right]\mathbf{X}\mathbf{r}_k\,/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$ $= \mathbf{X}\,\mathbf{R}_k^{(k)} = \mathbf{X}\,\mathbf{d}_k/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- component variance: $\mathbf{t}_k^T\mathbf{V}\mathbf{t}_k = 1/s_k$
- loadings: $\mathbf{p}_k = \frac{s_k}{\sqrt{\mathbf{r}_k^T\mathbf{r}_k}}\,\mathbf{X}^T\mathbf{V}\mathbf{X}\,\mathbf{d}_k = (\mathbf{r}_{k+1} - \mathbf{r}_k)/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- component-scale regression coefficient: $q_k = s_k\,\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- conjugated gradients: $\tilde{\mathbf{w}}_k + \frac{\tilde{\mathbf{w}}_k^T\tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_{k-1}^T\tilde{\mathbf{w}}_{k-1}}\tilde{\mathbf{w}}_{k-1} + \frac{\tilde{\mathbf{w}}_k^T\tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_{k-2}^T\tilde{\mathbf{w}}_{k-2}}\tilde{\mathbf{w}}_{k-2} + \cdots + \frac{\tilde{\mathbf{w}}_k^T\tilde{\mathbf{w}}_k}{\tilde{\mathbf{w}}_1^T\tilde{\mathbf{w}}_1}\tilde{\mathbf{w}}_1 = \mathbf{d}_k$
- X-scale gradients (weights): $\mathbf{R}_k^{(k)} = \mathbf{d}_k/\sqrt{\mathbf{r}_k^T\mathbf{r}_k}$
- X-scale regression coefficient: $\hat{\boldsymbol{\beta}}_k = \mathbf{R}_k\mathbf{Q}_k^T = \left[\mathbf{R}_k^{(1)} \quad \mathbf{R}_k^{(2)} \quad \ldots \quad \mathbf{R}_k^{(k)}\right]\left[q_1 \quad \ldots \quad q_k\right]^T = \sum_{m=1}^{k} s_m\mathbf{d}_m$

A.3 Illustration of Conjugate Gradient approximating the full-rank solution

As Conjugate Gradient (and consequently (W)PLS) finds the minimizer $\hat{\mathbf{x}}$ over the entire space in rank($\mathbf{A}$) steps, we demonstrate what occurs when using fewer (component) steps to solve $\mathbf{Ax} = \mathbf{b}$. At the k-th CG step, its solution $\hat{\mathbf{x}}^{(k)}$ minimizes $\{\left(\hat{\mathbf{x}} - \mathbf{x}^{(k)}\right)^T \mathbf{A} \left(\hat{\mathbf{x}} - \mathbf{x}^{(k)}\right)\}$ over the current rank-k (Krylov) subspace, with the norm of each error monotonically decreasing with k. As the full-rank solution can be expressed as $\hat{\boldsymbol{x}} = \sum_{m=1}^{p} s_m \, \mathbf{d}_m = \mathbf{A}^{-1} \, \mathbf{b}$, we first express $\mathbf{d}_m$ in terms of $\mathbf{b}$:

$$\text{Let } \boldsymbol{\omega}_k = \left(1 + \frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2}\right) \mathbf{I} - s_{k-1}\mathbf{A} \qquad \text{for } k \geq 2, \quad \text{and} \quad \boldsymbol{\omega}_1 = \mathbf{I}. \quad \text{Let } \boldsymbol{\Omega}_k = \prod_{m=1}^{k} \boldsymbol{\omega}_k.$$

$$\text{Then} \quad \mathbf{d}_k = \boldsymbol{\omega}_k \, \mathbf{d}_{k-1} - \frac{\|\mathbf{r}_{k-1}\|^2}{\|\mathbf{r}_{k-2}\|^2} \, \mathbf{d}_{k-2} \qquad \text{for } k \geq 3. \tag{A.1}$$

$$\text{Let } \boldsymbol{\Xi}_k = \boldsymbol{\omega}_k \, (\boldsymbol{\Xi}_{k-1}) \; - \; \frac{\|\mathbf{r}_{k-1}\|^2}{\|\mathbf{r}_{k-2}\|^2} (\boldsymbol{\Omega}_{k-2} + \boldsymbol{\Xi}_{k-2}) \qquad \text{for } k \geq 3, \quad \text{and} \quad \boldsymbol{\Xi}_1 = \boldsymbol{\Xi}_2 = \mathbf{0}.$$

$$\text{Then} \quad \mathbf{d}_k = (\boldsymbol{\Omega}_k + \boldsymbol{\Xi}_k) \, \mathbf{d}_1 \; = \; (\boldsymbol{\Omega}_k + \boldsymbol{\Xi}_k) \, \mathbf{b} \qquad \text{for } k \geq 1. \tag{A.2}$$

Now, we can apply the Cayley-Hamilton Theorem [13] to state that the inverse of symmetric matrix $\mathbf{A}$ can be expressed as a sum of polynomials with respect to $\mathbf{A}$:
$\mathbf{A}^{-1} = \sum_{m=0}^{p-1} c_m \, \mathbf{A}^m$. By equation A.2, we can claim that Conjugate Gradient approximates this inverted matrix and provide a few low-rank examples: $\mathbf{A}^{-1} \approx \sum_{m=0}^{k-1} (c_m - e_m) \, \mathbf{A}^m$. Consider a simple example with 2 predictors. Cayley-Hamilton states that $\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}[\text{tr}(\mathbf{A}) \, \mathbf{I} - \mathbf{A}]$. Since $\det(\mathbf{A}) = \frac{1}{s_1 \; s_2}$ and $\text{trace}(\mathbf{A}) = \frac{1}{s_2} + \frac{1}{s_1}\left(1 + \frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2}\right)$,

$$\mathbf{A}^{-1} = s_1 \; s_2 \left[\left\{\frac{1}{s_2} + \frac{1}{s_1}\left(1 + \frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2}\right)\right\} \; \mathbf{I} - \mathbf{A}\right] = s_1 \; s_2 \left[\frac{1}{s_2}\mathbf{I} + \left\{\frac{1}{s_1}\left(1 + \frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2}\right)\mathbf{I} - \mathbf{A}\right\}\right],$$

where the second form breaks apart the contribution of the first and second CG iterate. With three predictors, we first present the Cayley-Hamilton form and then distinguish each

CG iterate's contribution:

$$\begin{aligned}
\mathbf{A}^{-1} = \; & s_1\ s_2\ s_3 \Bigg[\left\{ \frac{1}{s_2\ s_3} + \frac{1}{s_1\ s_3}\left(1+\frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2}\right) + \frac{1}{s_1\ s_2}\left(1+\frac{\|\mathbf{r}_3\|^2}{\|\mathbf{r}_2\|^2}+\frac{\|\mathbf{r}_3\|^2}{\|\mathbf{r}_1\|^2}\right) \right\} \mathbf{I} \\
& - \left\{ \frac{1}{s_3} + \frac{1}{s_2}\left(1+\frac{\|\mathbf{r}_3\|^2}{\|\mathbf{r}_2\|^2}\right) + \frac{1}{s_1}\left(1+\frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2}\right) \right\} \mathbf{A} + \mathbf{A}^2 \Bigg] \\
= \; & s_1\ s_2\ s_3 \Bigg[\frac{1}{s_2\ s_3}\mathbf{I} + \left\{ \frac{1}{s_1\ s_3}\left(1+\frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2}\right)\mathbf{I} - \frac{1}{s_3}\mathbf{A} \right\} + \left\{ \frac{1}{s_1\ s_2}\left(1+\frac{\|\mathbf{r}_3\|^2}{\|\mathbf{r}_2\|^2}+\frac{\|\mathbf{r}_3\|^2}{\|\mathbf{r}_1\|^2}\right)\mathbf{I} \right. \\
& \left. - \left[\frac{1}{s_2}\left(1+\frac{\|\mathbf{r}_3\|^2}{\|\mathbf{r}_2\|^2}\right) + \frac{1}{s_1}\left(1+\frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2}\right) \right] \mathbf{A} + \mathbf{A}^2 \right\} \Bigg]
\end{aligned}$$

Next we express the difference between the full-rank solution and the partial Conjugate Gradient solution after k steps, with respect to different $\mathbf{d}_j$ of interest for interpretation.

$$\text{Let } \boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k+1}\ \boldsymbol{\omega}_{k+1} - \frac{\|\mathbf{r}_{k+1}\|^2}{\|\mathbf{r}_k\|^2}\boldsymbol{\alpha}_{k+2} + s_k\mathbf{I} \quad \text{for } k \le p, \quad \text{and} \quad \boldsymbol{\alpha}_j = \mathbf{0} \quad \text{for } j > p.$$

$$\text{Then } \hat{\mathbf{x}} - \mathbf{x}^{(k)} = \boldsymbol{\alpha}_{k+1}\ \mathbf{d}_{k+1} - \frac{\|\mathbf{r}_{k+1}\|^2}{\|\mathbf{r}_k\|^2}\ \boldsymbol{\alpha}_{k+2}\ \mathbf{d}_k = (\boldsymbol{\alpha}_k - s_k\mathbf{I})\ \mathbf{d}_k - \frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2}\ \boldsymbol{\alpha}_{k+1}\ \mathbf{d}_{k-1}. \tag{A.3}$$

We can re-express $\boldsymbol{\alpha}_k$ as a sum of step sizes $\left(\sum_{m=k}^{p} s_m\ \boldsymbol{\gamma}_m^k\right)$, letting

$$\boldsymbol{\gamma}_m^k = \boldsymbol{\omega}_m \boldsymbol{\gamma}_{m-1}^k - \frac{\|\mathbf{r}_{m-1}\|^2}{\|\mathbf{r}_{m-2}\|^2}\ \boldsymbol{\gamma}_{m-2}^k \quad \text{for } m > k, \quad \boldsymbol{\gamma}_k^k = \mathbf{I}, \quad \text{and} \quad \boldsymbol{\gamma}_m^k = \mathbf{0} \quad \text{for } m < k.$$

$$\text{Then } \hat{\mathbf{x}} - \mathbf{x}^{(k)} = \sum_{m=k+1}^{p} s_m \left[\left(\boldsymbol{\gamma}_m^{k+1} \boldsymbol{\omega}_{k+1} - \frac{\|\mathbf{r}_{k+1}\|^2}{\|\mathbf{r}_k\|^2}\ \boldsymbol{\gamma}_m^{k+2} \right) \mathbf{d}_k - \left(\frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2}\ \boldsymbol{\gamma}_m^{k+1} \right) \mathbf{d}_{k-1} \right] \tag{A.4}$$

$$= \sum_{m=k+1}^{p} s_m \left[\boldsymbol{\Omega}_m + \boldsymbol{\Xi}_m \right] \mathbf{d}_1 \ = \ \sum_{m=k+1}^{p} s_m\ \mathbf{d}_m.$$

We now verify that the first two equations hold.

Using the definitions of the residual and conjugated gradients to show equation A.1,

$$\begin{aligned}
\mathbf{d}_k &= -\mathbf{r}_k + \frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2}\ \mathbf{d}_{k-1} && \mathbf{r}_k = \mathbf{r}_{k-1} + s_{k-1}\mathbf{A}d_{k-1} \\
&= -\mathbf{r}_{k-1} + \left(\frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2}\ \mathbf{I} - s_{k-1}\mathbf{A} \right) \mathbf{d}_{k-1} \\
&= \left[\left(1 + \frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2} \right) \mathbf{I} - s_{k-1}\mathbf{A} \right] \mathbf{d}_{k-1} - \frac{\|\mathbf{r}_{k-1}\|^2}{\|\mathbf{r}_{k-2}\|^2}\ \mathbf{d}_{k-2} && \text{for } k \geq 3 \\
&= \boldsymbol{\omega}_k\ \mathbf{d}_{k-1} - \frac{\|\mathbf{r}_{k-1}\|^2}{\|\mathbf{r}_{k-2}\|^2}\ \mathbf{d}_{k-2} && \text{for } k \geq 3
\end{aligned}$$

□

We show equation A.2 holds by induction.

$$\mathbf{b} = \mathbf{d}_1 = \mathbf{I}\ \mathbf{d}_1 = (\boldsymbol{\Omega}_1 + \boldsymbol{\Xi}_1)\,\mathbf{d}_1 \qquad\qquad \mathbf{d}_2 = \boldsymbol{\omega}_2\ \mathbf{d}_1 = (\boldsymbol{\Omega}_2 + \boldsymbol{\Xi}_2)\,\mathbf{d}_1$$

$$\mathbf{d}_3 = \boldsymbol{\omega}_3\ \mathbf{d}_2 - \frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2}\mathbf{d}_1 \;=\; \boldsymbol{\Omega}_3\mathbf{d}_1 - \frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2}\boldsymbol{\Omega}_1\mathbf{d}_1 \;=\; (\boldsymbol{\Omega}_3 + \boldsymbol{\Xi}_3)\,\mathbf{d}_1$$

$$\mathbf{d}_4 = \boldsymbol{\omega}_4\ \mathbf{d}_3 - \frac{\|\mathbf{r}_3\|^2}{\|\mathbf{r}_2\|^2}\mathbf{d}_2 \;=\; \boldsymbol{\Omega}_4\mathbf{d}_1 + \left[\boldsymbol{\omega}_4(\boldsymbol{\Xi}_3) - \frac{\|\mathbf{r}_3\|^2}{\|\mathbf{r}_2\|^2}\,(\boldsymbol{\Omega}_2 + \boldsymbol{\Xi}_2)\right]\mathbf{d}_1 \;=\; (\boldsymbol{\Omega}_4 + \boldsymbol{\Xi}_4)\,\mathbf{d}_1$$

Assume this pattern holds for component k. Then, for step $k+1$ we have that

$$\mathbf{d}_{k+1} = \boldsymbol{\omega}_{k+1}\ \mathbf{d}_k - \frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2}\mathbf{d}_{k-1} \;=\; \boldsymbol{\Omega}_{k+1}\mathbf{d}_1 + \left[\boldsymbol{\omega}_{k+1}(\boldsymbol{\Xi}_k) - \frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2}\,(\boldsymbol{\Omega}_{k-1} + \boldsymbol{\Xi}_{k-1})\right]\mathbf{d}_1$$

$$= \;(\boldsymbol{\Omega}_{k+1} + \boldsymbol{\Xi}_{k+1})\,\mathbf{d}_1 \;=\; (\boldsymbol{\Omega}_{k+1} + \boldsymbol{\Xi}_{k+1})\,\mathbf{b} \qquad \square$$

Appendix B

EXTENDED SIMULATION RESULTS FOR CHAPTER 2

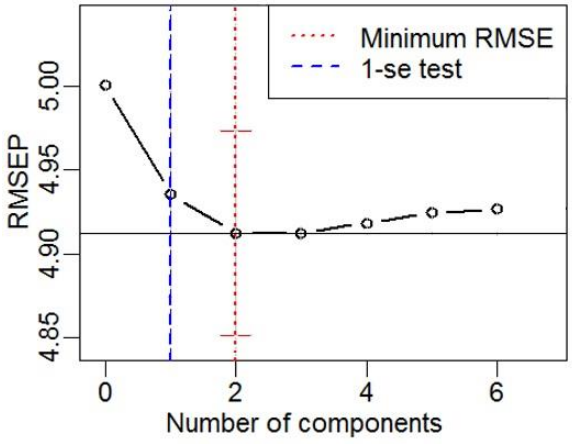
B.1 Extended one standard error test for cross-validation

Section 1: One standard error test for multiple tuning parameters

Background: To assess the predictive ability of candidate models, we use cross-validation (CV) and compare each model's root mean squared error of prediction (RMSE), where the mean is taken across the CV folds. There is a trade-off between model accuracy and complexity, as there are typically more parsimonious models (with fewer components or more sparsity) that have similar predictive accuracy as the model with the minimum RMSE.

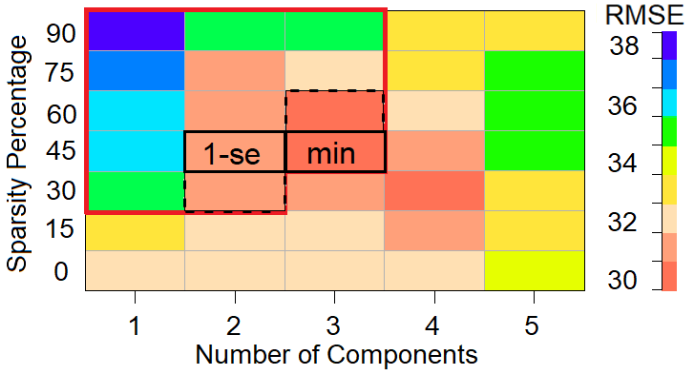
In the context of component-based methods, the one standard error (1-se) test chooses the model with the fewest components whose RMSE is within one standard error of the minimum RMSE. In the example of Figure A1, two components results in the lowest RMSE (red dotted line). After calculating the standard error of the squared error of prediction for the 2-component model across the CV folds, we search for the most parsimonious model whose predictive ability is within 1-se of the minimum RSME (i.e., whose RMSE is no larger than the upper red bar). Here, the model with one component (blue dashed line) satisfies this criterion, so we select this simpler model.

FIGURE A1 Illustration of the one standard error test for one tuning parameter (for PLS).



Extension: Sparse component-based methods have two tuning parameters, so we extended the 1-se test to optimize the parameter search across a grid of candidate pairs. Figure A2 illustrates our approach in the context of Sparse PLS, where its sparsity parameter is a fraction {$\eta \in [0,1)$}. In this example, the model with three components and 45% sparsity results in the lowest RMSE (black solid square with "min"). We then consider all models in the region of potential parsimony improvement (red rectangle), which emphasizes a decrease in the number of components for our scientific goal. There are three pairs of candidates (black dashed squares and "1-se" square) whose RMSEs are within 1-se of the minimum RMSE, so we use a weighted metric to determine which pair provides the best improvement in parsimony. Since we favor a decrease in components over an increase in sparsity, we select the model with two components and 45% sparsity (black solid square with "1-se").

FIGURE A2 Illustration of the one standard error test for two tuning parameters (for Sparse PLS).



Formally, the weighted metric is given by $\text{improve} = \alpha_1 * (\Delta K) + \alpha_2 * (\Delta\lambda)$.

$\Delta K = K_{min} - K_{candidate}$ refers to the change in number of components,

$\Delta\lambda = \lambda_{min} - \lambda_{candidate}$ refers to the change in sparsity level,

αs are the weights assigned to each change in tuning parameter, and

$improve > 0$ means that the candidate model provides an increase in parsimony over the model with minimum RMSE.

The candidate model with the largest *improve* value is selected by the extended 1-se test.

Example with Sparse PLS:
In our setting, we would like fewer components, so we set $\alpha_1 = +1$, meaning that the greater the decrease in components, the larger *improve* will be.

We would also like more sparsity, so we set $\alpha_2 = -5$; the greater the increase in sparsity, the larger *improve* will be. For Sparse PLS, we use its sparsity fraction for the sparsity unit (e.g., $\lambda = \eta$).

We choose these specific magnitudes because we prefer a decrease in components over an increase in sparsity. Specifically, $\left|\frac{\alpha_1}{\alpha_2}\right| = .20$ implies that we will select a model with a 1-unit decrease in components even if its sparsity level is smaller by up to 19% (absolute difference).

So, if we were considering between a model with 4 components and 70% sparsity and a model with 3 components and 55% sparsity, we would select the latter model. However, we would not select a model with 3 components and 50% sparsity.

Example with Sparse PCR:
In Sparse PCR, its sparsity parameter takes a much larger (positive) range, based on the size of the weight **p**. We strategically select 9 sparsity values to test for each data set based on the size of the largest element in $\mathbf{p_1}$. To make this test comparable across all data sets, we instead use the index of sparsity value as our λ (e.g., $\lambda = \{1, 2, \ldots, 9\}$).

Again, we prefer a decrease in components over an increase in sparsity, but now we change the magnitudes. We set $\alpha_1 = +2$ and $\alpha_2 = -1$ (with $\left|\frac{\alpha_1}{\alpha_2}\right| = 2$), meaning that we will select a model with a 1-unit decrease in components even if its sparsity index is 1 unit smaller.

B.2 Simulations comparing sparse PLS methods

As the sparse PLS methods mentioned in Chapter 2 have not been formally compared, we conducted a simulation study to determine which method selected more interpretable models in contexts similar to nutritional epidemiology. Specifically, we consider a method that includes component-wise sparsity (using an L_1 penalty; denoted by subscript 1) [44] and one that targets variable selection (using Chun and Keles's procedure; denoted by subscript 2) [14]. Furthermore, we applied two parsimony tests to identify the most advantageous setting for these methods: one-standard-error test (denoted by "se") [31] and permutation test (denoted by "perm") [72]. We generated a continuous response from 9 of 24 predictors in $\mathbf{X}$, with varying amounts noise. We favored a decrease in the number of components over an increase in sparsity if there were multiple candidate pairs that satisfied the parsimony criterion (see Appendix B.1 for more details).

Both sparse PLS methods selected all relevant variables until the response had a relatively high variance ($\text{sd}(\mathbf{y}) > 5$), but only SPLS_2 with the 1-se test was able to discard most irrelevant variables. If SPLS_1's first component included most of the information from the relevant variables, then its subsequent components were much more likely to select irrelevant variables than with SPLS_2. Additionally, SPLS_2 with the 1-se test had the best predictive performance (as indicated by the RMSE plot) for a wide range of response variabilities. Thus, Chun and Keles's sparse PLS with the 1-se test was most conducive to our setting.

We next illustrate how increasing the variability of added noise affects the sample covariance between our multiblock predictors and the generated response in Figure B.2. Recall that the within-block covariance levels are 0.05, 0.50, 0.30, and 0.10, respectively. True variables (with non-null associations) are denoted by blue triangles and false variables by black circles. Though true and false variables are fairly easily distinguished for lower levels of noise variability, it becomes quite difficult to determine variable relevance once $\text{sd}(\mathbf{y}) > 5$. This will cause difficulty for any sparse method.

Figure B.1: Average prediction and variable selection performance for four SPLS methods by response variability.

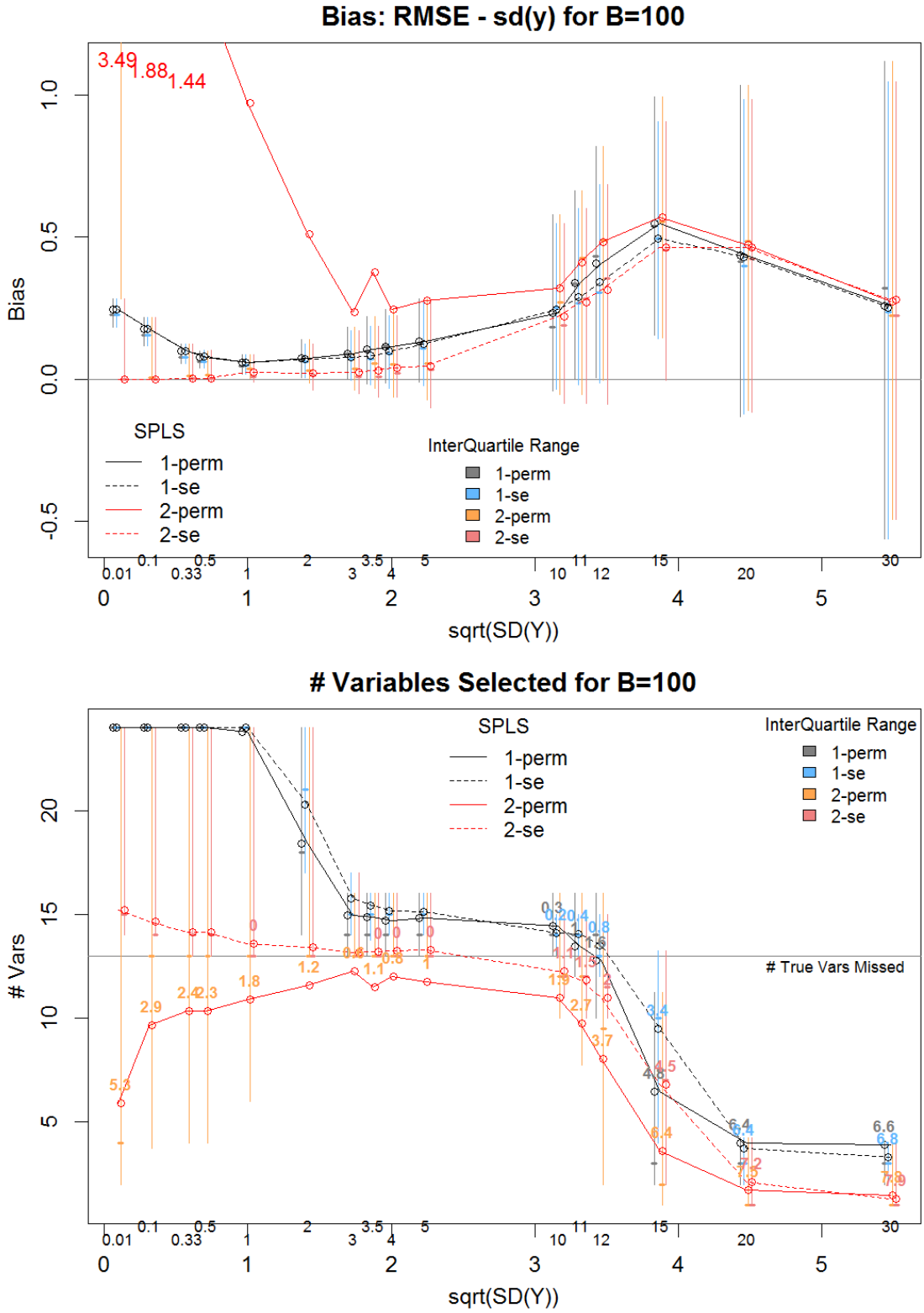
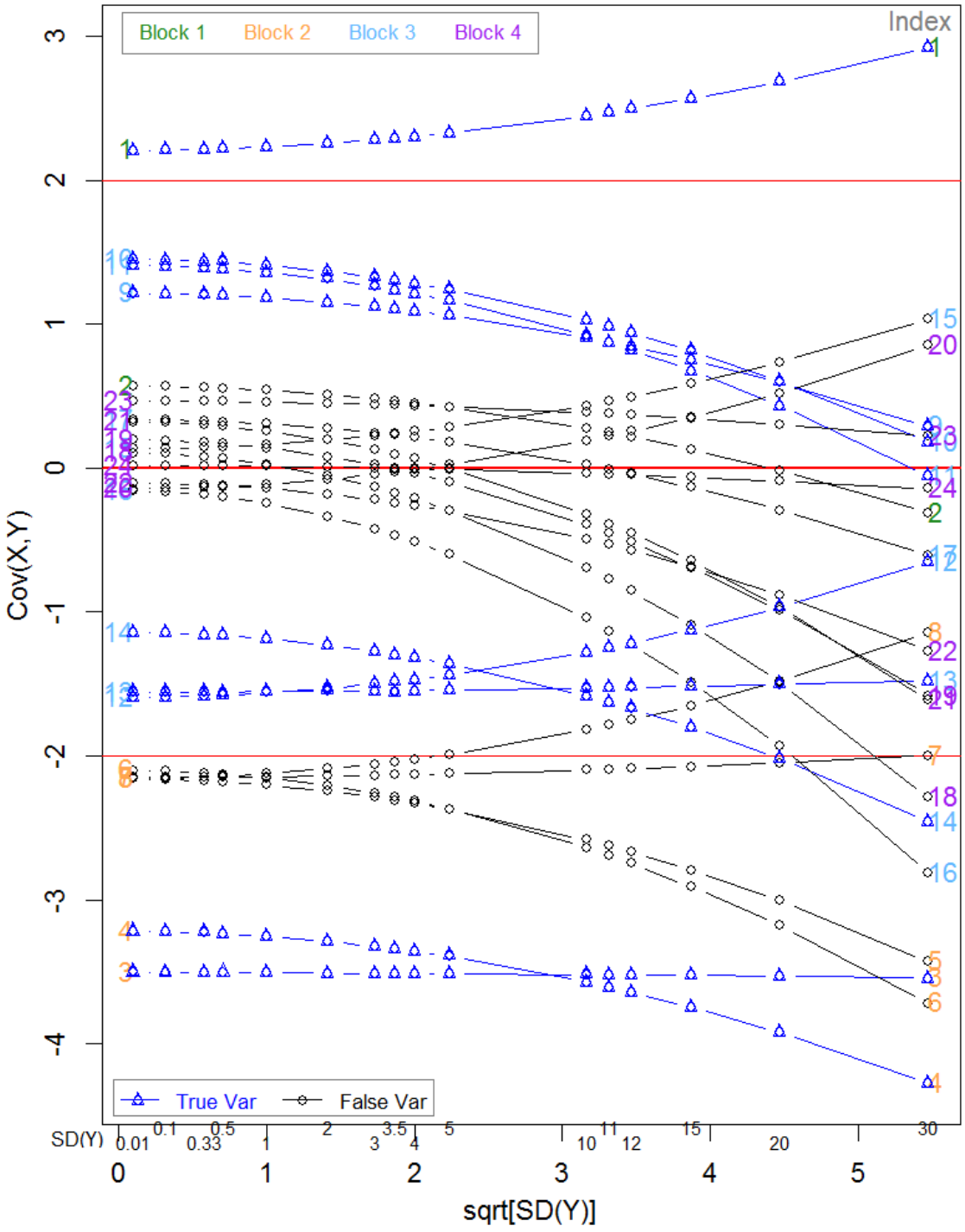


Figure B.2: Illustration of sample covariance between response and multiblock predictors as response variability changes.



B.3 Simulations comparing sparse dimension reduction methods

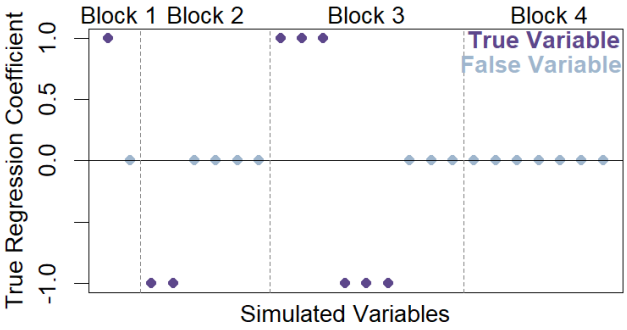
Section 2: Further Simulation Results

Background: We present results from seven "matching-covariance block structure" settings. Here, all blocks are generated with the same level of intra-block covariance (ρ), specifically with ρ={0, .1, .3, .5, .6, .8, 1}. Formally, the predictors **X** are generated as

$$\mathbf{X}_{n\times p} = \begin{bmatrix}\text{Block 1}\\ \text{Block 2}\\ \text{Block 3}\\ \text{Block 4}\end{bmatrix}^{T} \sim N_n\left(\mathbf{0}, \begin{bmatrix} 1 & \rho & & & & & \dots & 0 \\ \rho & 1 & & & & & & \vdots \\ & & 1 & \rho & & & & \\ & & \rho & 1 & & & & \\ & & & & 1 & \rho & & \\ & & & & \rho & 1 & & \\ & & & & & & 1 & \rho \\ \vdots & & & & & & \rho & 1 \\ 0 & \dots & & & & & & \end{bmatrix}\right).$$

The univariate response **y** is then generated as a normally distributed variable with unit variance and mean of **Xβ**, where the true regression coefficient is given by Figure B1.

FIGURE B1 True relationship between 24 simulated predictors and the response.



Matching-covariance results: On the next page, Figure B2 graphically presents average results across 100 generated data sets, as well as 95% confidence intervals.

Panel A displays the RMSE of the test sets, and we see that SPLS consistently has the lowest prediction error. This difference is not statistically significant, apart from the independent case (ρ=0), when SPCR has a statistically worse prediction error compared to all the other methods.

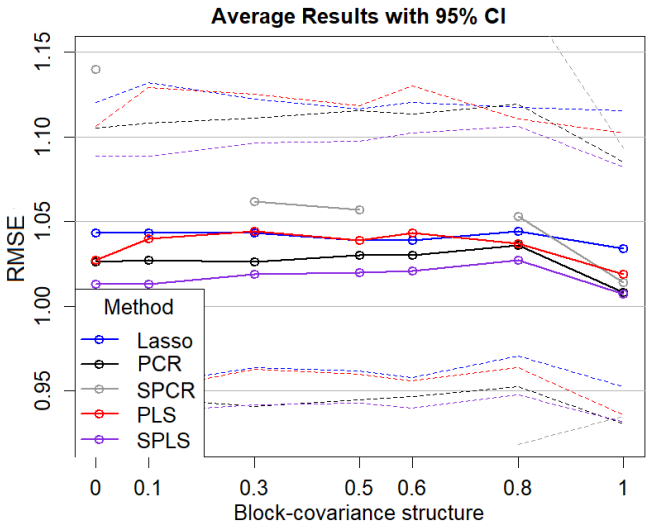
Panel B displays the number of true variables selected. While SPCR has some difficulty when ρ=0, SPLS does not discard any true variables until ρ=0.8, though it is most noticeable for SPLS and Lasso when ρ=1.

Panel C displays the number of false variables selected. SPLS and Lasso select less than one false variable until ρ>0.1. Lasso is less sensitive to the block-covariance structure, selecting at most 2.6 false variables when ρ=0.8, whereas SPLS selects 12 false variables. However, when the block-covariance is more realistic to food frequency questionnaire data (ρ≤0.5), SPLS tends to select less than a third of the false variables (e.g., 5).

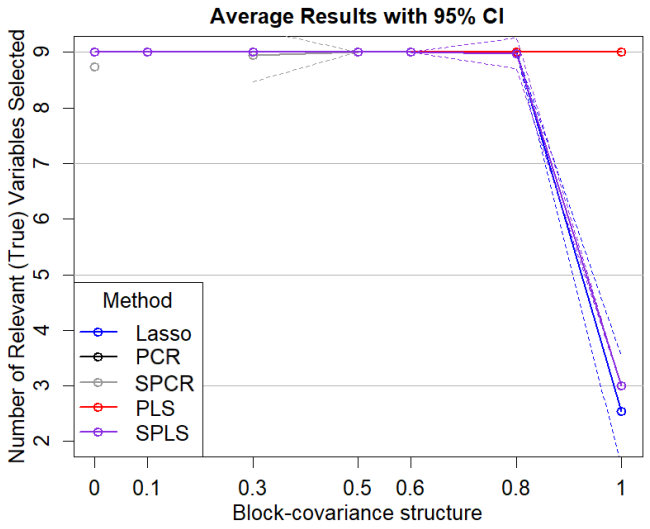
Panel D displays the number of components selected. Except for with perfect correlation (ρ=1), PCA-based methods select 23-24 components. PLS-based methods tend to select 2-7 components, with SPLS generally selecting half a component more than PLS.

FIGURE B2 Simulation results across 100 data sets for seven matching-covariance block structures.

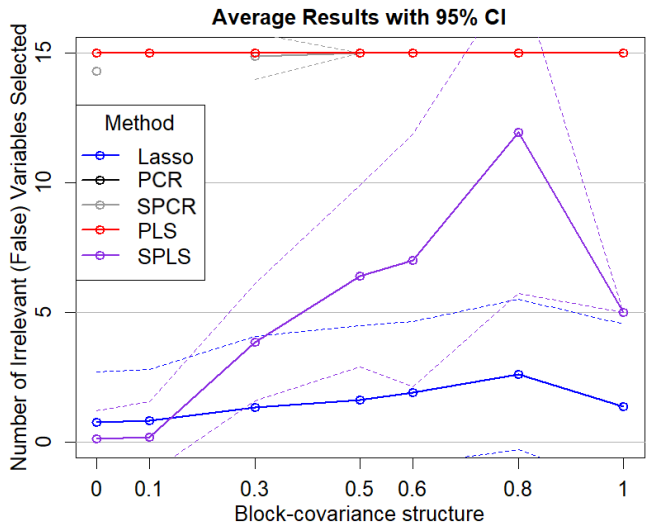
(a)



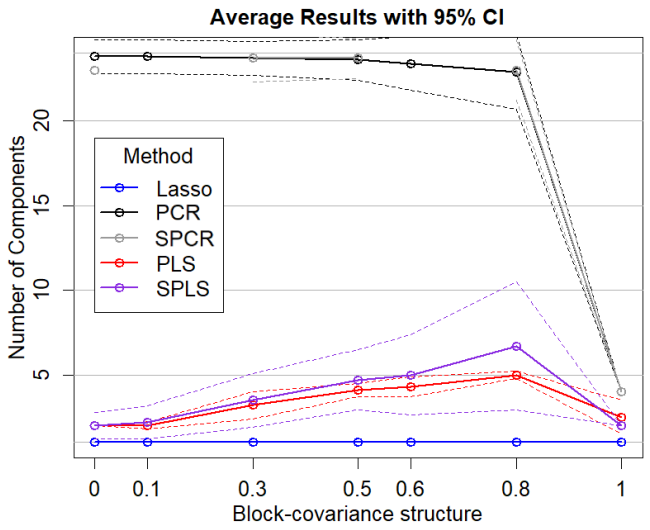
(b)



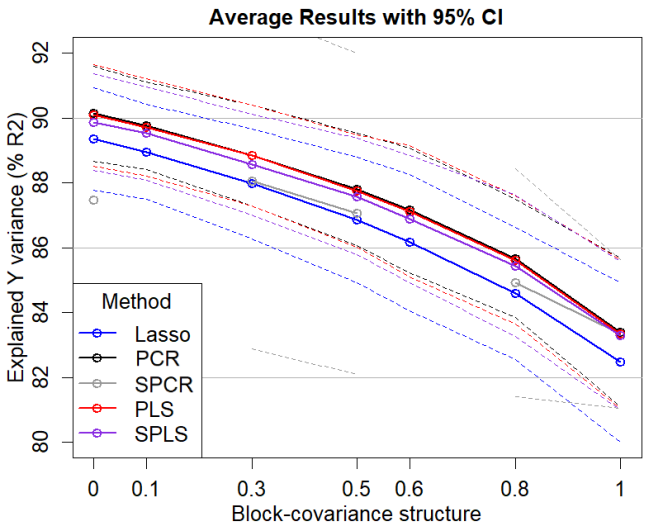
(c)



(d)



(e)



Panel E displays the percent of **y** variance explained (R^2). Most methods exhibit a gradual decrease in R^2 as the block-covariances increase. Typically, PCR and PLS explain the most **y** variance, followed by SPLS, SPCR, and Lasso.

Mixed-covariance results: We also tested "mixed-covariance block structures", in which we permuted each block's covariance from among $\rho=\{0.05, 0.10, 0.30, 0.50\}$. Since we still impose independence across blocks, each block behaves similarly to the corresponding covariance structure from the "matching" scenarios.

In general, the supervised methods have a harder time distinguishing between true and false variables when the blocks with more relevant variables (Blocks 2 and 3) have higher intra-block covariances. In the easiest mixed settings, SPLS and Lasso select less than one false variable; in the hardest mixed setting, SPLS selects 5 false variables while Lasso selects 2.

SPLS again has the lowest RMSE across all mixed settings, though PCR comes close when the blocks with more relevant variables have high intra-block covariances. Interestingly, in that setting when the blocks with more relevant variables have low intra-block covariances, PCR and SPCR select only 20 components, as compared to their usual 23-24 components.

Appendix C

EXTENDED DATA ANALYSES FOR CHAPTERS 2 AND 4

Interpretation when adjusting for total energy intake

To clarify the interpretation of a model relating diet to a CVD-related endpoint that adjusts for total energy (caloric) intake, we present two models seeking to link (centered) BMI with two food items, burgers and wine.

First, consider the linear regression model that adjusts for the variable Adjustment but not for total calories (Cals); we subsequently provide the interpretation of the first coefficient.

$$\text{BMI} = \beta_1 \,\text{Burger} + \beta_2 \,\text{Wine} + \gamma_1 \,\text{Adjustment}$$

$\widehat{\beta}_1$ = `expected change in BMI for every 1 unit change in Burger (and thus total caloric) intake, keeping Adjustment and Wine levels constant`.

Hence, this model analyzes the contribution of its diet variables on an absolute scale.

Now, consider the interpretation of a linear model that does adjust for total calories.

$$\text{BMI} = \beta_1 \,\text{Burger} + \beta_2 \,\text{Wine} + \gamma_1 \,\text{Adjustment} + \gamma_2 \,\text{Cals}$$

$\widehat{\beta}_1$ = `expected change in BMI for every 1 unit change in Burger intake, keeping Adjustment, Wine, and Cals levels constant`.

For the previous coefficient interpretation to make scientific sense, we need to assume that there are other food items that contribute to the total calories consumed. For example, if Cals=Burger+Wine+Cheese, the adjustment can be interpreted as increasing Burger intake while decreasing Cheese intake to keep Cals constant. Therefore, we can rewrite the coefficient interpretation as the expected change in BMI when **substituting 1 unit of**

Cheese intake with Burger intake, keeping Adjustment, Wine, and Cals levels constant. Thus, this model analyzes the contribution of its diet variables on a relative scale, comparing its contribution to that of consumed foods not included in the model. This interpretation concurs with that presented in the discussion of Willett, Howe, and Kushi's 1997 paper [27].

When fitting a Cox model for CVD that adjusts for total caloric intake, we can similarly interpret e^{β_1} as the expected hazard ratio for a CVD event when comparing two groups of people where one group has **substituted 1 unit of Cheese intake with Burger intake**, keeping Adjustment, Wine, and Cals levels constant.

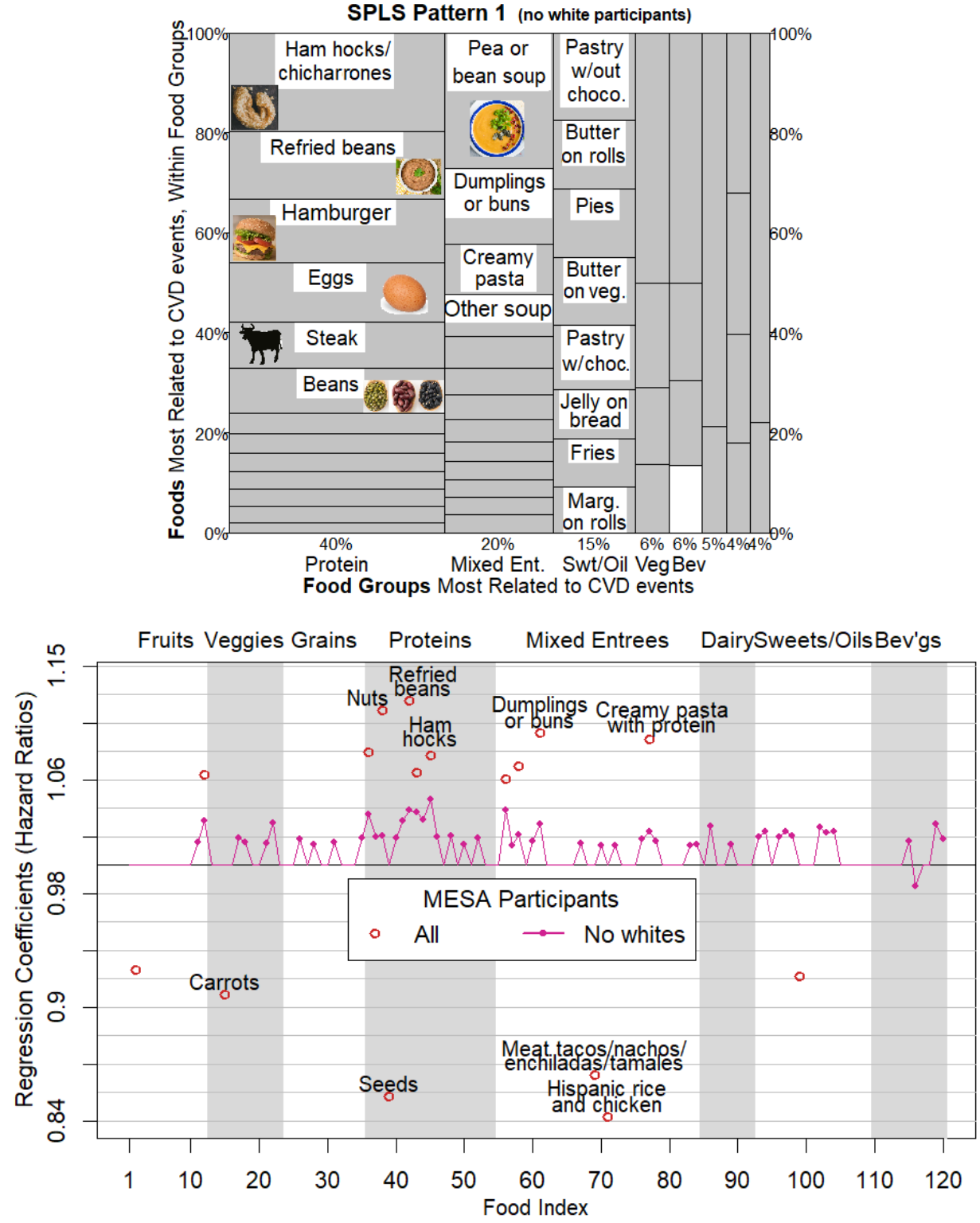
CVD sensitivity analysis

We conducted a sensitivity analysis for our CVD data analysis to give evidence as to whether or not ham hocks/chicharrones/pig's feet was still considered a relevant food item when restricting the sample to non-white MESA participants. Table C.1 provides some details about the differences between the samples and corresponding Cox models with SPLS summary using Combination-Adjustment. Figure C.1 illustrates the contribution of the selected foods to the construction of the first SPLS component in the restricted sample, which is fairly similar to the first pattern in the complete case analysis (see Figure 4.5). We again see that ham hocks/chicharrones/pig's feet contributes the most to the construction of the first SPLS component; this leads to the largest magnitude (log) hazard ratio in the analysis without whites (pink line), as it only selected the first pattern as the most relevant to CVD risk.

Table C.1: Summary of samples and CVD-related diet patterns in MESA.

Metric	All groups	Non-white groups
Sample Size	5881	3545
Event Size	877	506
Events per Variable	6.7	3.9
Foods	16	50
Patterns	4	1
C-statistic	73	72

Figure C.1: Sensitivity analysis of CVD risk excluding white MESA participants.



Appendix D

CONSIDERATIONS WHEN APPROXIMATING THE COX MPLE

D.1 Investigation of when aMPLE approximates MPLE well

Below we provide a brief investigation for when the Cox approximate MPLE regression coefficients are reasonably close to the MPLE. We begin by specifying these terms more formally, and assume that the initial estimate for either algorithm is $\boldsymbol{\beta}_0 = \mathbf{0}$. Both algorithms can be expressed as follows, given their respective estimate $\boldsymbol{\beta}_t$ after the t-th iteration:

$$\hat{\mu}_{i,t} = \exp(\mathbf{x}_i^T \boldsymbol{\beta}_t)\, \widehat{\Lambda}_0(\mathbf{y}_i|\boldsymbol{\beta}_t), \qquad \widehat{\mathbf{V}}_t = \operatorname{diag}\left(\hat{\mu}_{i,t}\right)_{i=1}^n, \qquad \hat{z}_{i,t} = \mathbf{x}_i\boldsymbol{\beta}_t + \left(\delta_i - \hat{\mu}_{i,t}\right)/\hat{\mu}_{i,t}$$

$$\boldsymbol{\beta}_{t+1} = \begin{cases} \left(\mathbf{X}^T\, \widehat{\mathbf{V}}_t\, \mathbf{X}\right)^{-1}\, \mathbf{X}^T \widehat{\mathbf{V}}_t\, \hat{\mathbf{z}}_t, & \text{if MPLE} \\ \left(\mathbf{X}^T\, \widehat{\mathbf{V}}_t\, \mathbf{X}\right)^{-1}\, \mathbf{X}^T \widehat{\mathbf{V}}_t\, \hat{\mathbf{z}}_t \;-\; \boldsymbol{\xi}_t^{(k)}, & \text{if aMPLE} \end{cases}$$

At the first iteration, both algorithms start at $\boldsymbol{\beta}_0 = \mathbf{0}$, so they have the same $\widehat{\mathbf{V}}_1$ and $\widehat{\mathbf{z}}_1$. Therefore, adopting the notation of (k) for the aMPLE estimate, we can specifically state that $\mathcal{N}_1 := \mathbf{X}^T\left[\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_1^{(k)}\right] = \mathbf{X}^T\boldsymbol{\xi}_0^{(k)}$. Consider the estimate of the next aMPLE baseline hazard; for a person r at a generic time τ_g when they are at risk (e.g. $r \in R_g$),

$$\widehat{\lambda}_0(\tau_g)_2^{(k)} = \left(\sum_{i\in R_g} \exp\left[\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_1 - \mathbf{x}_i^T\boldsymbol{\xi}_0^{(k)}\right]\right)^{-1} = \left(\sum_{i\in R_g} \exp\left[\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_1 - \mathcal{N}_{i,1}\right]\right)^{-1}.$$

If $\sum_{i\in R_g}\left(e^{-\mathcal{N}_{i,1}} - e^{-\mathcal{N}_{r,1}}\right)\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_1) < 0$, then $\widehat{\lambda}_0(\tau_g)_2^{(k)} > \widehat{\lambda}_0(\tau_g)_2$ (e.g., the second iteration of the MPLE baseline hazard at time τ_g). After summing over all times that person r is at risk to obtain their cumulative baseline hazard, we can more formally express the relationship

between their MPLE and aMPLE estimated risk of an event:

$$\begin{aligned}\hat{\mu}_{r,2}^{(k)} &= \exp(\mathbf{x}_r^T \hat{\boldsymbol{\beta}}_1^{(k)}) \;\; \widehat{\Lambda}_0(\mathbf{y}_r|\hat{\boldsymbol{\beta}}_1^{(k)}) \;\; = \;\; \exp(\mathbf{x}_r^T \hat{\boldsymbol{\beta}}_1) \;\; \exp(-\mathcal{N}_{r,1}) \;\; \widehat{\Lambda}_0(\mathbf{y}_r|\hat{\boldsymbol{\beta}}_1^{(k)}) \\ &= \hat{\mu}_{r,2} \left[\frac{\widehat{\Lambda}_0(\mathbf{y}_r|\hat{\boldsymbol{\beta}}_1^{(k)})}{\exp(\mathcal{N}_{r,1}) \;\; \widehat{\Lambda}_0(\mathbf{y}_r|\hat{\boldsymbol{\beta}}_1)} \right] \;\; := \;\; \hat{\mu}_{r,2} \;\; \zeta_{r,2}\end{aligned}$$

The constant $\zeta_{r,2}$ determines which of the cumulative hazard estimates is larger for person r. In essence, this constant is comparing the effect of replacing the i-th person's difference in their linear predictor due to PLS ($\mathcal{N}_{i,1}$) with that of the r-th person ($\mathcal{N}_{r,1}$) in all baseline hazards. Using this notation, we can similarly compare the second $\hat{\mathbf{z}}$ values as follows:

$$\begin{aligned}\hat{z}_{r,2}^{(k)} &= \mathbf{x}_r \hat{\boldsymbol{\beta}}_1^{(k)} + \frac{\delta_r}{\hat{\mu}_{r,2}^{(k)}} - 1 \;\; = \;\; \mathbf{x}_r \hat{\boldsymbol{\beta}}_1 + \frac{\delta_r}{\hat{\mu}_{r,2} \;\; \zeta_{r,2}} - 1 - \mathcal{N}_{r,1} \\ \hat{z}_{r,2} &= \mathbf{x}_r \hat{\boldsymbol{\beta}}_1 + \frac{\delta_r}{\hat{\mu}_{r,2}} - 1\end{aligned}$$

The difference between the second working responses is due to both the scaling factor $\zeta_{r,2}$ and the bias term $\mathcal{N}_{r,1}$.

When considering future iterations, most of this notation stays the same, except that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_{t+1}^{(k)} = \;& (-\mathbf{H}_t)^{-1} \mathbf{X}^T \mathbf{V}_t \; \mathbf{z}_t \;\; - \;\; \left(-\mathbf{H}_t^{(k)}\right)^{-1} \mathbf{X}^T \mathbf{V}_t^{(k)} \mathbf{z}_t^{(k)} \;\; + \;\; \boldsymbol{\xi}_t^{(k)} \\ &= \hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_t^{(k)} \;\; + \;\; (-\mathbf{H}_{t+1})^{-1} \mathbf{U}_{t+1} - \left(-\mathbf{H}_{t+1}^{(k)}\right)^{-1} \mathbf{U}_{t+1}^{(k)} + \;\; \boldsymbol{\xi}_t^{(k)} \\ &= (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_1^{(k)}) \;\; + \;\; \sum_{m=2}^{t+1} \left[(\hat{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}}_{m-1}) - (\hat{\boldsymbol{\beta}}_m^{(k)} - \hat{\boldsymbol{\beta}}_{m-1}^{(k)}) \right] \qquad \text{(D.1)} \\ \mathcal{N}_{t+1} = \;& \mathbf{X}^T \left[\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_{t+1}^{(k)} \right]\end{aligned}$$

For MPLE and aMPLE to converge to similar regression coefficient estimates, it must hold that $\zeta_{r,t} \to 1$ and $\mathcal{N}_{r,t} \to 0$ for all people r as iteration t increases. However, notice from equation D.1 that $\|\mathcal{N}_{t+1}\|$ is at least the size of $\|\mathbf{X}^T \boldsymbol{\xi}_0^{(k)}\|$ (Triangle Inequality), which

approaches 0 as k increases, but not necessarily as t increases. Hence, we cannot describe convergence bounds so strictly, but it does provide some intuition as to when aMPLE results in a close approximation of the Cox MPLE.

D.2 Computational improvements via Line Search

Although MPLE and aMPLE generally converge, they can sometimes exhibit oscillating behavior or extra slow convergence of the regression coefficients if the "amount" of an update at each iteration is not optimized, particularly for aMPLE (also seen in [57, 55, 45]). In the MPLE update $\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + \alpha_t \, (-\mathbf{H}_t)^{-1}\mathbf{U}_t$, the positive constant α_t is called the step size, and it determines how far in the direction of the negative gradient $(-\mathbf{H}_t)^{-1}\mathbf{U}_t$ to move for the next update. The default for (a)MPLE is to set $\alpha_t = 1$ for all iterations t, but improvements in stability and speed can be made if this step size is chosen by a line search to optimize the log partial likelihood at each iteration:

$$\alpha_t = \arg\max_{\alpha \in (0,1]} \ell\left(\hat{\boldsymbol{\beta}}_t + \alpha \, (-\mathbf{H}_t)^{-1}\mathbf{U}_t\right) \qquad \text{for MPLE and}$$

$$\alpha_t = \arg\max_{\alpha \in [-1,1]} \ell\left(\hat{\boldsymbol{\beta}}_t^{(k)} + \alpha \, \left[\left(-\mathbf{H}_t^{(k)}\right)^{-1}\mathbf{U}_t^{(k)} - \boldsymbol{\xi}_t^{(k)}\right]\right) \qquad \text{for aMPLE.}$$

Below we provide justification for why aMPLE benefits from considering a negative step size, after reformulating the coefficient estimate updates (3.1) and (3.2), with the latter using the conjugate gradient notation.

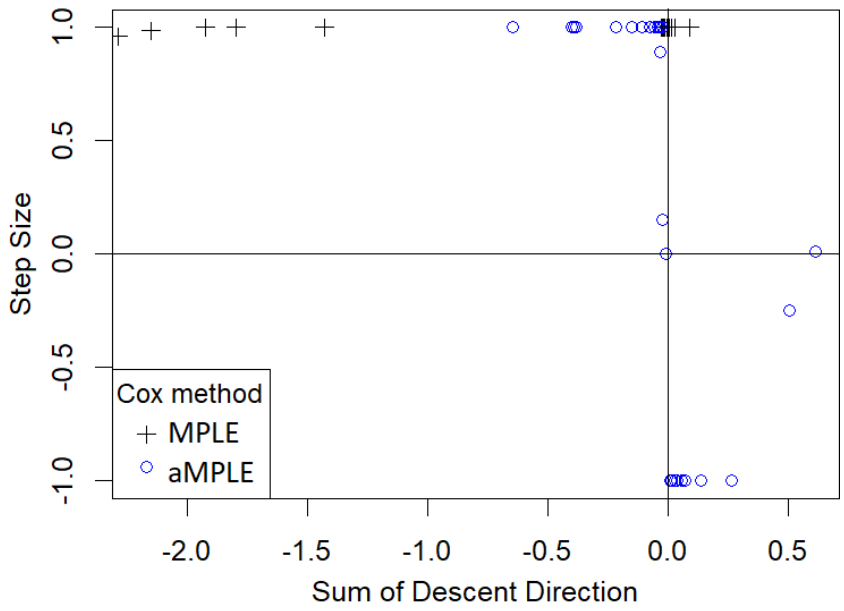
$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{t+1} &= \hat{\boldsymbol{\beta}}_t + \alpha_t \ (-\mathbf{H}_t)^{-1}\mathbf{U}_t \ = \ \hat{\boldsymbol{\beta}}_t - \alpha_t\hat{\boldsymbol{\beta}}_t \ + \alpha_t\hat{\boldsymbol{\beta}}_t \ + \alpha_t(-\mathbf{H}_t)^{-1}\mathbf{U}_t \\
&= (1-\alpha_t)\ \hat{\boldsymbol{\beta}}_t + \alpha_t \left[(-\mathbf{H}_t)^{-1}\ \mathbf{X}^T\mathbf{V}_t \left(\mathbf{X}\hat{\boldsymbol{\beta}}_t + \mathbf{V}_t^{-1}\ \frac{\partial \ell(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}}\right)\right] \\
&= (1-\alpha_t)\ \hat{\boldsymbol{\beta}}_t + \alpha_t\ (-\mathbf{H}_t)^{-1}\ \mathbf{X}^T\mathbf{V}_t\ \mathbf{z}_t
\end{aligned}
$$

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{t+1}^{(k)} &= (1-\alpha_t)\ \hat{\boldsymbol{\beta}}_t^{(k)} + \alpha_t \ \sum_{m=1}^{k} s_{m,t}\ \mathbf{d}_{m,t} \\
&= (1-\alpha_t)\ \hat{\boldsymbol{\beta}}_t^{(k)} + \alpha_t \ \left[(-\mathbf{H}_t^{(k)})^{-1}\ \mathbf{X}^T\mathbf{V}_t^{(k)}\ \mathbf{z}_t^{(k)} - \boldsymbol{\xi}_t^{(k)}\right] \\
&= (1-\alpha_t)\ \hat{\boldsymbol{\beta}}_t^{(k)} + \alpha_t \ \left[\hat{\boldsymbol{\beta}}_t^{(k)} + (-\mathbf{H}_t^{(k)})^{-1}\ \mathbf{U}_t^{(k)} - \boldsymbol{\xi}_t^{(k)}\right] \\
&= \hat{\boldsymbol{\beta}}_t^{(k)} \ + \ \alpha_t \left[(-\mathbf{H}_t^{(k)})^{-1}\ \mathbf{U}_t^{(k)} - \boldsymbol{\xi}_t^{(k)}\right]
\end{aligned}
$$

Now, the sum of the update direction $\mathbf{1}^T(-\mathbf{H}_t)^{-1}\mathbf{U}_t$ is generally negative because MPLE searches for a consecutively smaller local minimum of $-\ell(\boldsymbol{\beta})$ by choosing a descent direction (negative gradient). We say "generally" because at the 2nd iteration, the sum of the descent direction can change from very negative to slightly positive, but by the 4th iteration and beyond, it returns to a small negative value. As the term $\left(-\mathbf{H}_t^{(k)}\right)^{-1}\mathbf{U}_t^{(k)}$ with the aMPLE coefficient behaves similarly, the sum of this descent direction is also generally negative. Similarly, $\mathbf{1}^T\boldsymbol{\xi}_t^{(k)} < 0$ because it is also constructed to point towards a descent direction.

However, while $\left(-\mathbf{H}_t^{(k)}\right)^{-1}\mathbf{U}_t^{(k)}$ tends toward $\mathbf{0}$ since $\mathbb{E}(\mathbf{U}) = \mathbf{0}$, $\boldsymbol{\xi}_t^{(k)}$ is not necessarily $\mathbf{0}$, particularly if some of the discarded components are truly relevant. In these aMPLE cases, the discarded components can account for a larger (negative) magnitude than $\left(-\mathbf{H}_t^{(k)}\right)^{-1}\mathbf{U}_t^{(k)}$, which leads to an overall positive sum of the update direction. Therefore, to preserve a negative contribution for the update direction and reliably optimize the partial log likelihood, it is useful for aMPLE to consider step sizes that are negative. In Figure D.1, we plot a few examples of MPLE and aMPLE converging for the same data sets, demonstrating that aMPLE tends to select negative step sizes when the sum of the update direction

Figure D.1: Comparing the range of step sizes for Cox methods



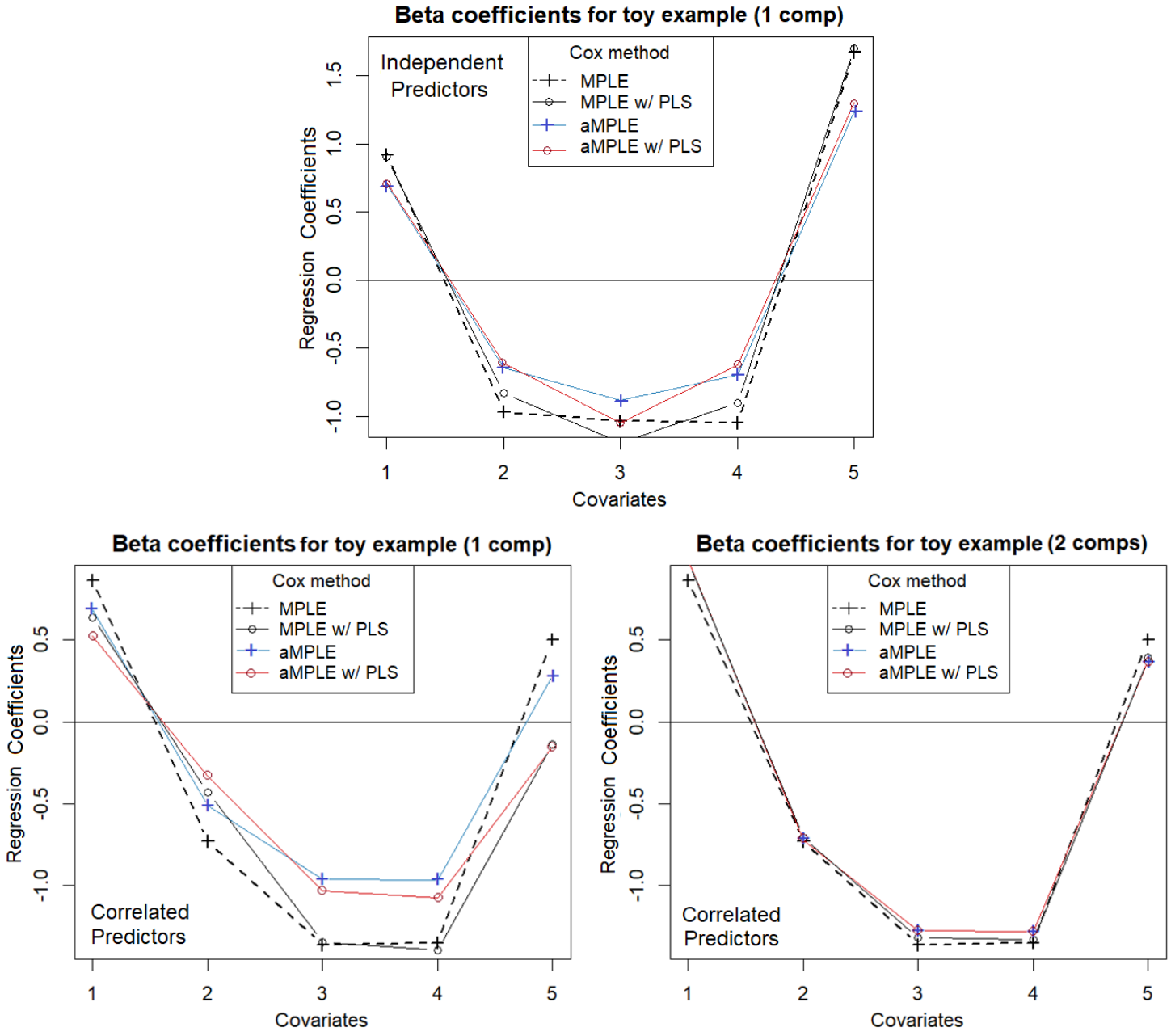
is positive. Though MPLE will occasionally result in small positive directions, the predominant sum of the update direction will be negative, so a positive step size is reasonable for convergence to a local minima.

To obtain PLS components from a Cox model, some researchers have considered only using the last iteration of aMPLE with a line search [55, 16, 45]. However, if the last step size is not 1, the resulting converged coefficient is a sum of multiple PLS fits: $\widehat{\boldsymbol{\beta}}^{(k)} = \sum_{i=1}^{t} \psi_i \left(\sum_{m=1}^{k} s_{m,i} \, \mathbf{d}_{m,i} \right)$, where $\psi_i = \left[\prod_{m=i+1}^{t} (1 - \alpha_m) \right] \alpha_i$. To ensure that the final coefficient is derived from a single PLS fit so that future predictions can easily be made, we refit PLS on the converged (a)MPLE coefficient.

D.3 Visual comparison between (a)MPLE with a PLS summary

To get a better sense of the difference between Cox-PLS methods, we present the regression coefficients across components for a toy example in Figure D.2. The data is generated such

Figure D.2: Cox-PLS regression coefficients for toy example (p=5), across predictor correlation settings and components



that there are five predictors, the first and last of which have a (true) unit positive regression coefficient and the remaining have a unit negative regression coefficient. We present results when predictors are independent of each other and when they are correlated—with the first two being related (0.1 intra-block covariance) and the last three being related (0.3 intra-block covariance). The setting presented has 1000 subjects, a roughly 10% event rate, and a 15% censorship rate.

In the upper panel (independent covariates), we first notice that aMPLE has a similar shape to MPLE (just attenuated) and same with aMPLE with PLS having an attenuated similar shape to MPLE with PLS. However, because of the amount of attenuation, the MPLE with PLS coefficients have a closer magnitude to the MPLE coefficients than do the aMPLE estimates.

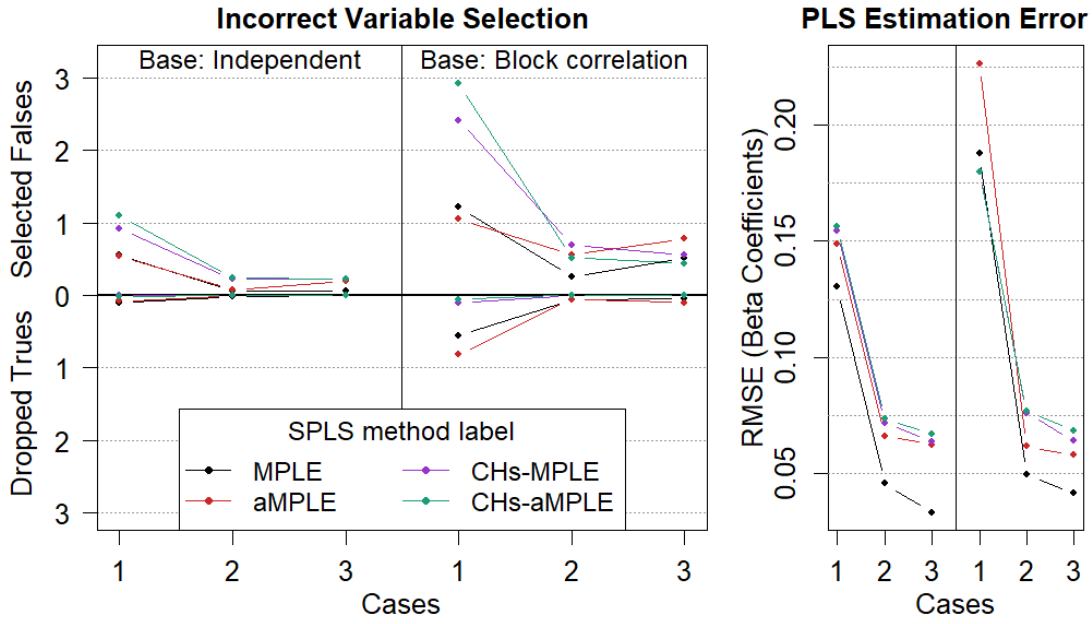
The relationship between the magnitudes of the methods changes when there are correlated predictors (bottom panel). While the mentioned pattern holds for variables 3 and 4 when 1 component is used, aMPLE is actually closer in magnitude to MPLE for the remaining variables. The attenuation pattern still holds (and perhaps even more clearly) in the 1 component model for correlated covariates. We include the 2 component model on the bottom right to demonstrate that the methods are producing very similar coefficients, and in fact they become indistinguishable (to sight) for all subsequent components.

D.4 Simulations to compare four sparse Cox-PLS methods

If the majority of one's predictors are believed to be unrelated to the outcome, there could be a preference to exclude them from the maximization procedure in order to avoid introducing noise into $\widehat{\mathbf{V}}$ and $\hat{\mathbf{z}}$. However, if most of the predictors are related to the outcome but some are only weakly related, it would seem reasonable to use all of the covariates to precisely estimate $\mathbf{V}$ and $\mathbf{z}$, and then identify the strongly-related predictors in the SPLS summary. We test this hypothesis in a simulation study with 50 generated data sets. Base predictors are constructed as in Chapter 2 (e.g. number of events $\tilde{n} \approx 100$, p=24, $|\beta| = \{1, 0\}$, 9 relevant variables), but we also consider adding 80 independent variables with small signals

($|\beta| \in [.005, .05]$) or no signal (e.g. noise variables). We generate a roughly 10% event rate and 15% censorship rate. As Cox methods are sensitive to the number of events-per-variables, we increased the sample size for the latter two cases to obtain $\tilde{n} \approx 400$ for parity. For all methods, we select their tuning parameters using a deviance criterion that penalizes for the number of selected variables (see AIC-type criterion in Section 3.3).

Figure D.3: Simulation results (B=50) for sparse Cox-PLS models across correlation amounts when using **AIC-type criterion** and keeping $\tilde{n}/p \approx 4$.



Case 1 setting: $n = 1000$, $\tilde{n} \approx 100$, $p = 24$. **Case 2 and 3 settings:** $n = 4000$, $\tilde{n} \approx 400$, $p = 104$ (same base predictors as Case 1, with 80 independent predictors weakly related [Case 2] or not related [Case 3] to the outcome).

Figure D.3 illustrates the variable selection and estimation error performance across these predictor cases, targeting the recovery of all strongly relevant variables and omission of the less relevant variables (e.g. as close to the "0" line as possible). In the simplest correlation setting to distinguish between variable types (independence; left column of plots), almost all true variables were selected, while the methods that used sparsity within the maximization procedure selected more false variables. In the block-correlation setting (right column of

plots), the post-maximization sparsity methods dropped more false variables in Cases 1 and 2 as expected, though they also dropped a fraction of a true variable in Case 1. The methods that used sparsity within the maximization procedure showed a marked variable selection advantage in Case 3, as they were able to discard almost all noise variables earlier on.

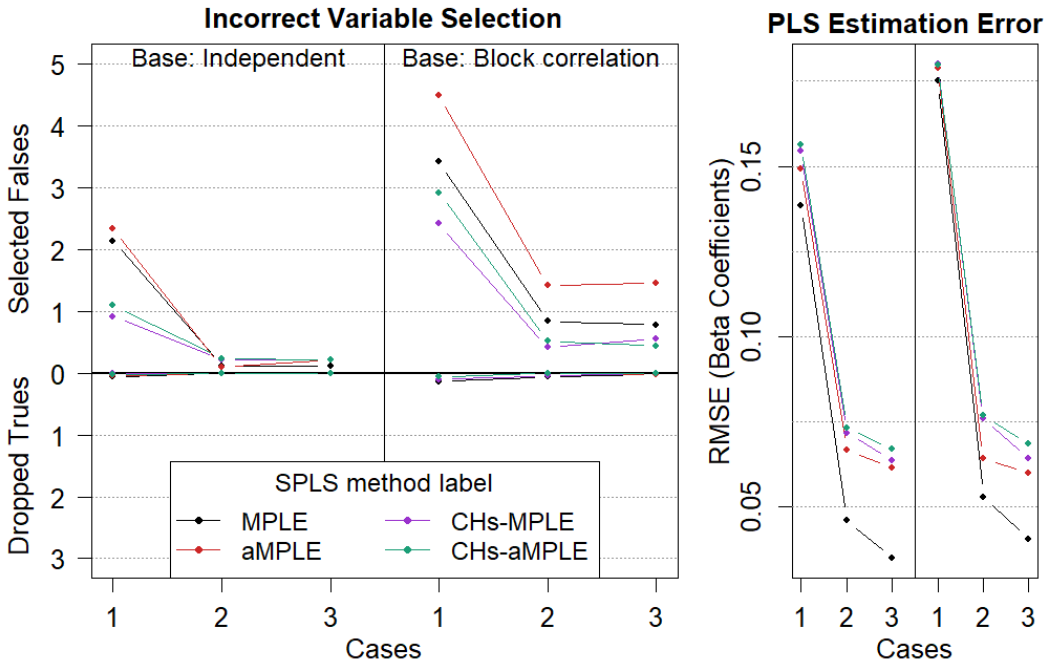
In terms of estimation error, MPLE (followed by aMPLE) resulted in the best performance across almost all settings. This was largely due to the post-maximization sparsity methods inducing less shrinkage on the true variables' coefficient estimates. This advantage was particularly noticeable in the settings with 80 additional covariates, which is closer aligned to Food Frequency Questionnaire data. Even when using deviance as the CV criterion (instead of this AIC-inspired one), the post-maximization sparsity methods resulted in less estimation error, although they did select more false variables in the correlated setting (see Figure D.4).

Overall, MPLE was a strong contender across all cases, regardless of correlation structure. Not only did it induce the least amount of shrinkage, leading to less estimation error, but it selected the fewest or close to the fewest false variables while selecting almost all true variables. The most difficult case for MPLE was when modeling only 24 multi-block correlated predictors, since it dropped roughly half of a true variable, on average. However, in the data sets with more events and covariates, this property became negligible. Therefore, we suggest using MPLE with SPLS summary as the preferred Cox-PLS method with SPLS sparsity, followed by aMPLE with SPLS summary when computation prohibits the former fit.

D.5 Computational considerations for (sparse) Cox-PLS methods

If certain Cox-PLS algorithms do not implement a proper line search, there is a chance that the coefficient estimate will get stuck oscillating between two values. When comparing the performance with other algorithms, such as Nygard's [55], we implemented a check for oscillation, and we selected the value that resulted in the larger log partial likelihood if oscillation occurred. We opted to stop the algorithm when the difference in log partial likelihoods was smaller than .01[1]. We decided on this comparison rather than the difference

Figure D.4: Simulation results (B=50) for four sparse Cox-PLS models across correlation amounts when using **minimum deviance criterion** and keeping $\tilde{n}/p \approx 4$.



Case 1 setting: $n = 1000$, $\tilde{n} \approx 100$, $p = 24$. **Case 2 and 3 settings:** $n = 4000$, $\tilde{n} \approx 400$, $p = 104$ (same base predictors as Case 1, with 80 independent predictors weakly related [Case 2] or not related [Case 3] to the outcome).

in $\hat{\boldsymbol{\beta}}$s because it was also reasonable to use when considering jointly-modeling additional covariates (see Chapter 4). If convergence was not reached within 100 iterations, we stopped the algorithm and simply used the last iterate for the remainder of the computations. For a variety of Cox settings, MPLE was more prone to not reaching convergence than aMPLE.

In order to select the most appropriate sparse Cox-PLS model, we would ideally like to test a fine grid of sparsity percentages and components, but the computational burden can greatly increase with large n and more than 3 components. This is especially true if we have to fit the estimation method for every single candidate pair. One way to improve the computational speed is to use `C++` to perform matrix multiplication and matrix-vector

[1]Although our code converges to the same values as `coxph` when the likelihood error tolerance is set to e^{-11}, we used a larger tolerance due to the time involved (especially for simulations in Section 3.2.2).

multiplication when solving either the WLS or weighted PLS problem at each iteration.

Our proposed sparse methods additionally benefit from the computational advantage of requiring fewer fits of (a)MPLE than the methods that incorporate sparsity within the maximization procedure. Specifically, MPLE w/ SPLS only requires fitting MPLE once (per fold) and aMPLE w/ SPLS only requires fitting aMPLE once per component candidate (per fold). Then SPLS can be performed on the converged gradient using a very fine grid of candidate pairs, without much computational repercussion. On the other hand, because the methods that use sparsity within their maximization step could be selecting different covariates for each candidate pair, they require refitting the maximization procedure for every single pair (per fold). Even when employing a form of "warm starts" for these methods, it was found that these starting values would overly influence the result of subsequent coefficients. Therefore, MPLE w/ SPLS in particular is able to test a fine grid of candidate tuning parameters without as much computational burden as other sparse Cox-PLS alternatives, even surpassing aMPLE w/ SPLS in speed if considering PLS models with 4 or more components. Moreover, one could take full advantage of the implemented optimization of the `coxph` function if we were to write a wrapper of `coxph` for MPLE with SPLS summary.

In general, we note that Cox-PLS models with 1-2 components tend to require lower levels of sparsity (20-50%) to maintain predictive ability, while models with 4-5 PLS components tend to use more sparsity (70-80%). However, as tempting as it might be to *a priori* select a sparse Cox-PLS model with only a few components, we advocate for testing a handful of components (e.g. 1-5) via cross validation to attain a useful model rather than choosing a model with unnecessarily poor predictive performance.

Appendix E

EXTENDED SIMULATION RESULTS FOR CHAPTER 3

Results presented are based on the average of 100 data sets. All simulations have a roughly 10% event rate (unless specified otherwise) and a 15% censorship rate.

When implementing the one standard error (1-se) test for sparse Cox-PLS methods with multiple tuning parameters, we used a weighted metric to determine which pair provides the best improvement in parsimony (see Appendix B.1 for more details). In Chapter 3, we used the following metric for all models using Chun and Keles's sparsity procedure:

$$improve = 1 * (\Delta K) - 5 * (\Delta \gamma).$$

A candidate model provides an increase in parsimony if it results in $improve > 0$. Specifically, the change Δ subtracts the candidate value from the value corresponding to the minimum CV criterion. The units for the tuning parameters are number of components K and sparsity percentage ($\gamma * 100$). Here, we value a unit decrease in components over a sparsity percentage decrease of at most 19%.

The LEs-aMPLE model has three tuning parameters: number of component K and two sparsity terms w and τ, which enforce more sparsity with larger values. We expand the parsimony metric as follows, using the index of sparsity values (e.g., {1,2,3,...}) as opposed to absolute values to make comparisons fairer (due to large differences in scale):

$$improve = 3 * (\Delta K) - 2.75 * (\Delta w) - 1.33 * (\Delta \tau).$$

Again, we prioritize a unit decrease in components over at most a 1-unit decrease in w or a 2-unit decrease in τ. Additionally, we value a 2-unit increase in τ over at most a 1-unit decrease in w.

Figure E.1: Simulation results across sample sizes when using **minimum deviance criteria**

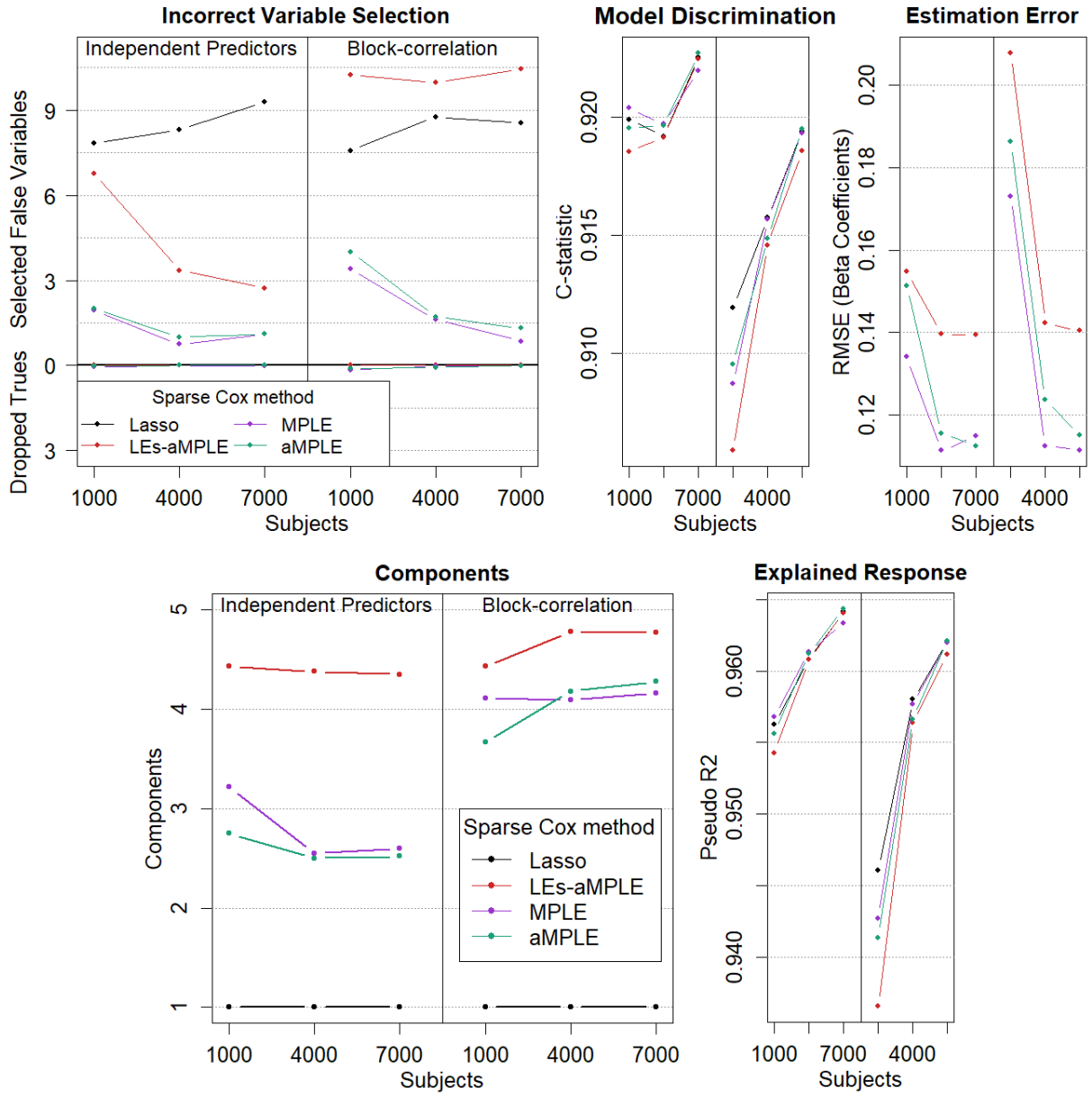


Figure E.2: Simulation results across sample sizes when using **1-se deviance criteria**

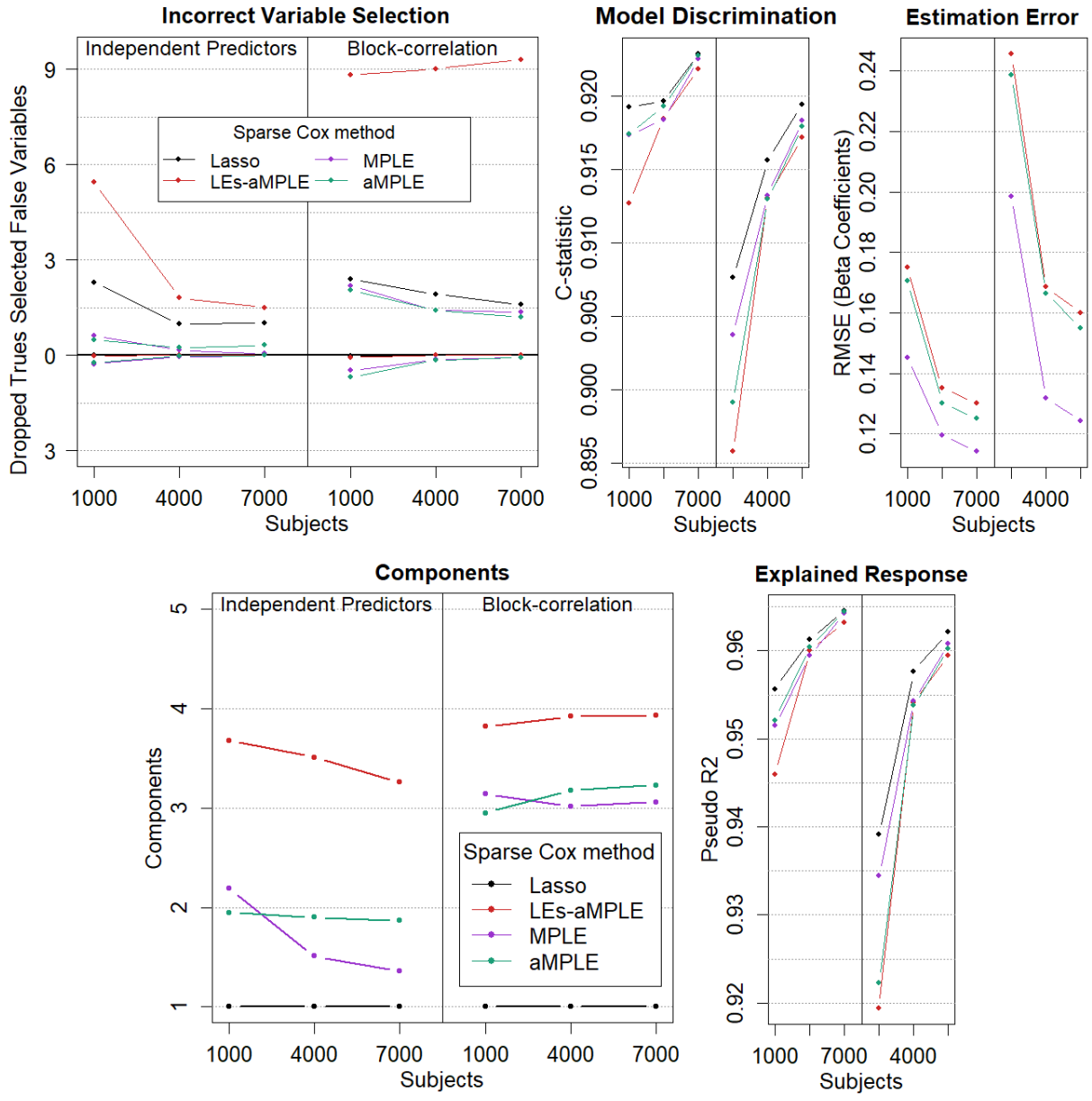


Figure E.3: Simulation results across sample sizes when using **AIC-type criteria**

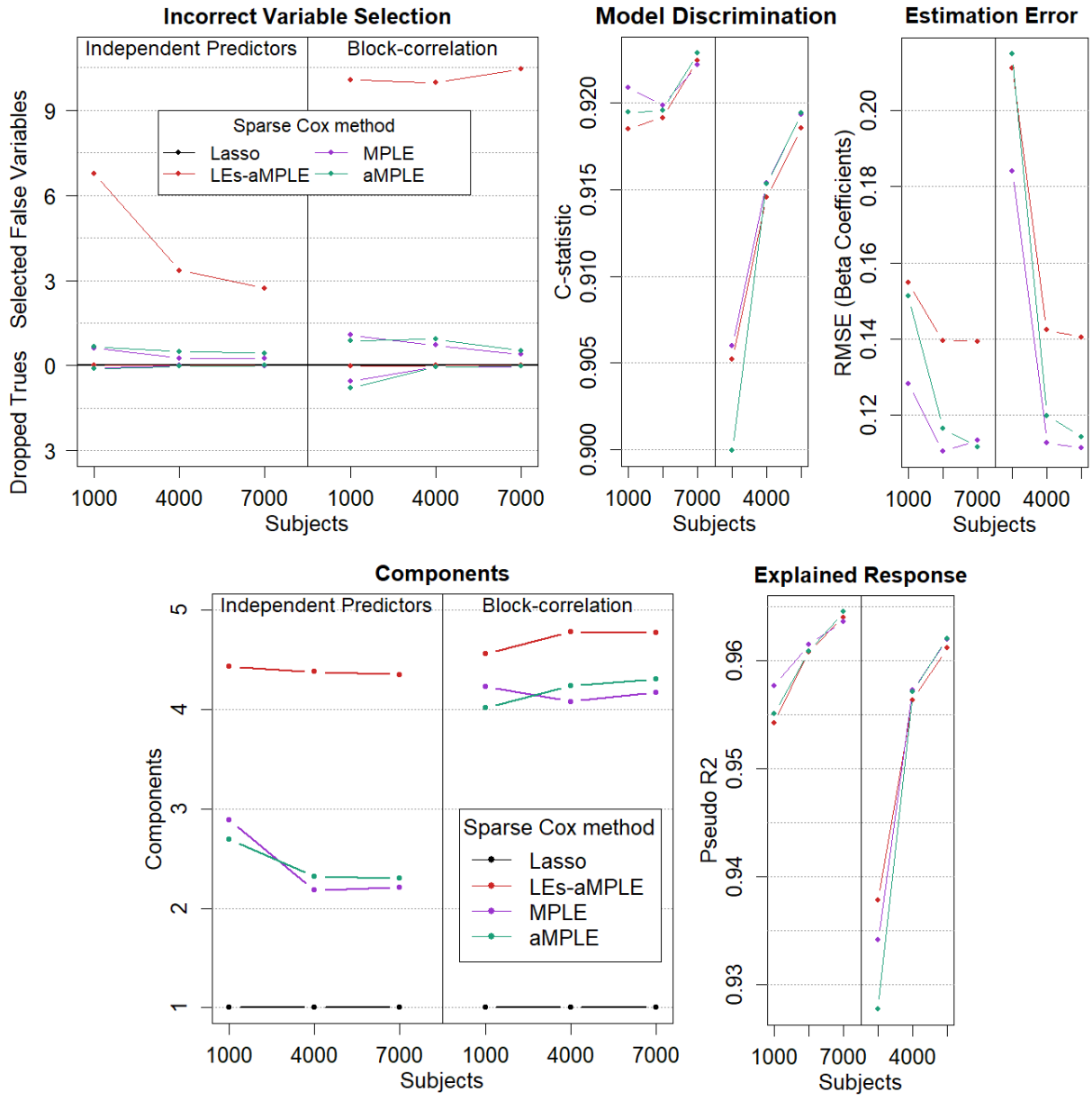


Figure E.4: Simulation results across sample sizes and criteria type. The first row corresponds to the **minimum value** and the second row to the **1-se value** for each CV criteria.

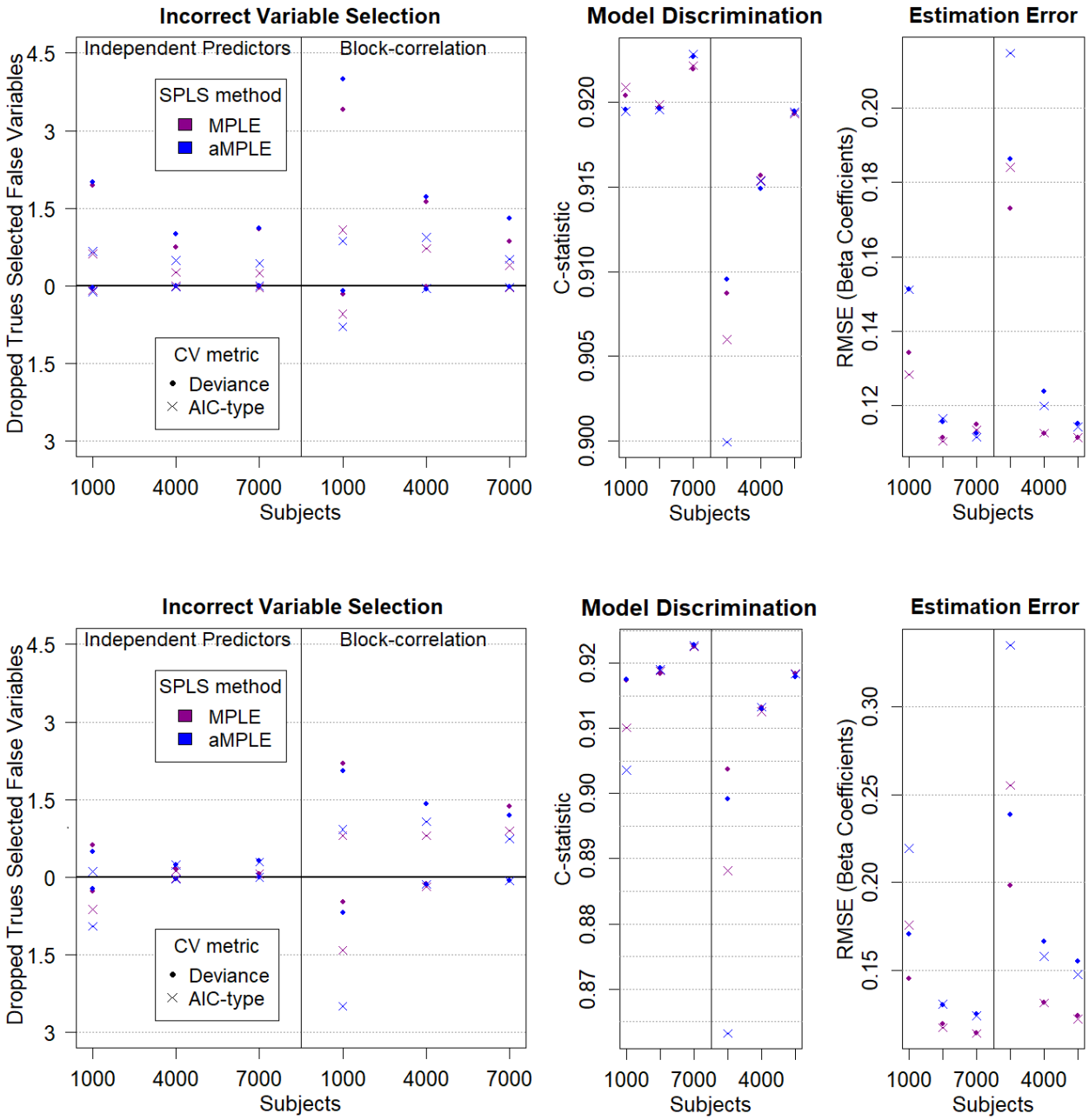


Figure E.5: Simulation results for 1000 subjects across event rates and correlation amounts when using **minimum deviance criteria**

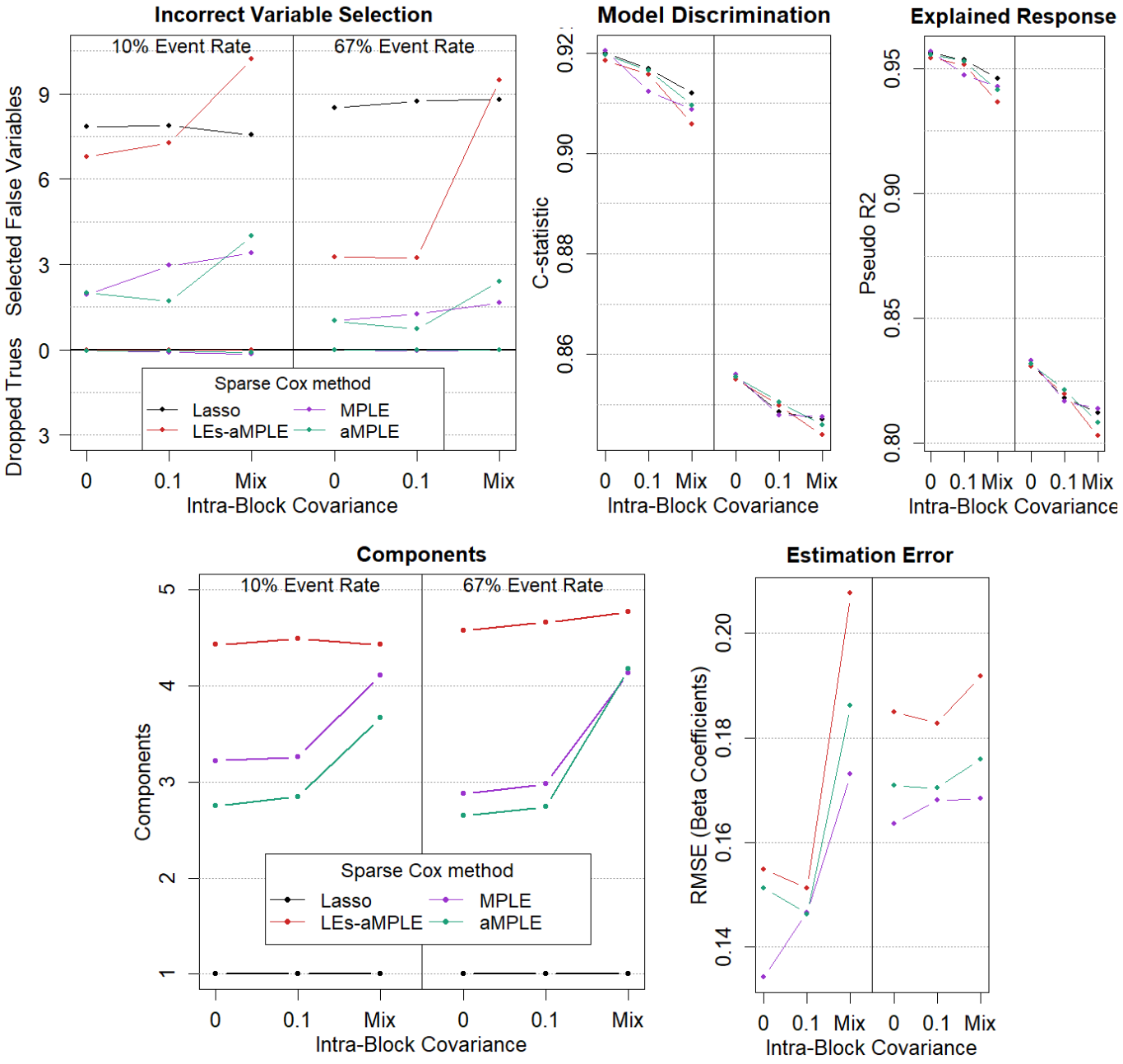


Figure E.6: Simulation results for 1000 subjects across event rates and correlation amounts when using **1-se deviance criteria**

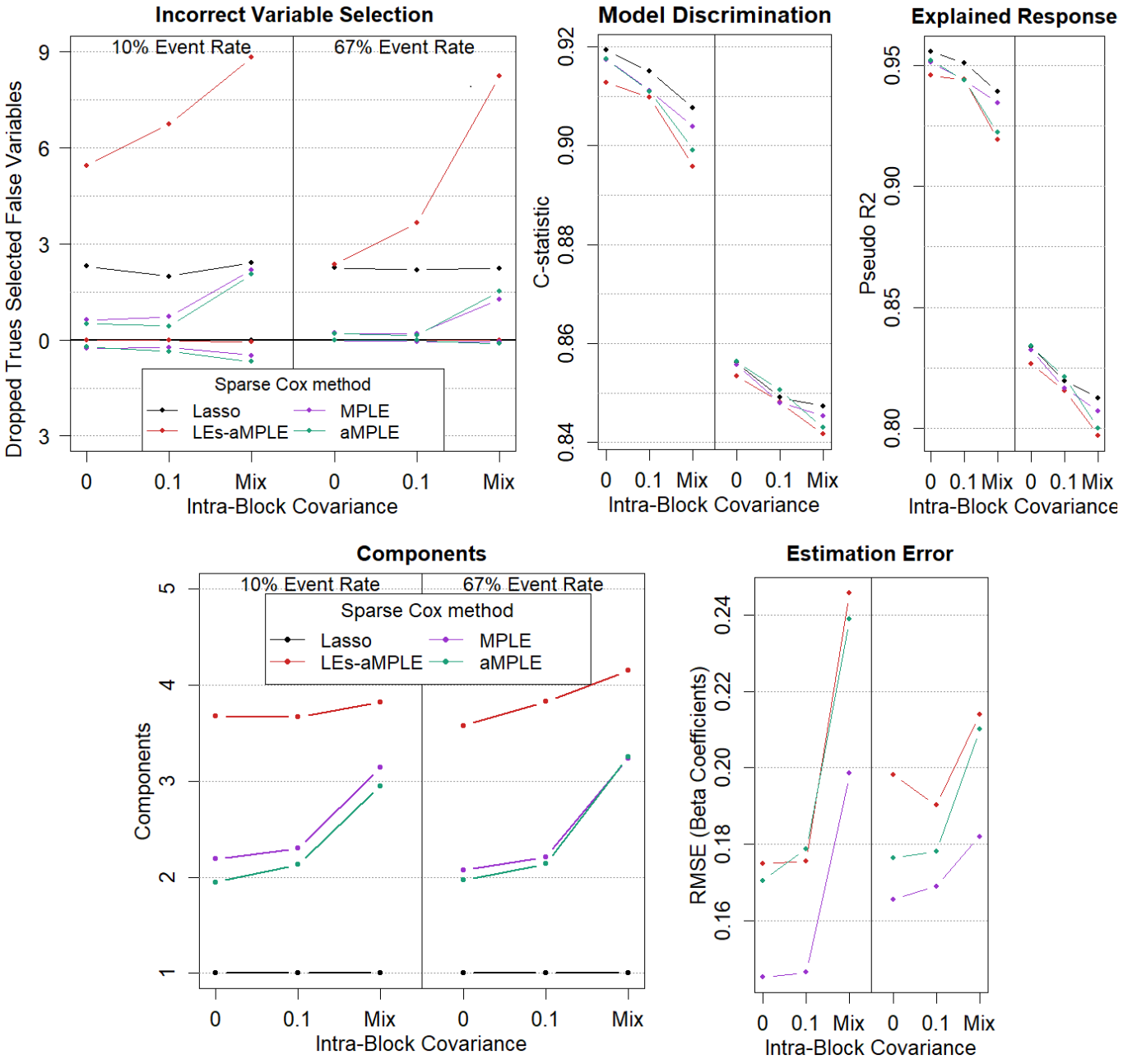
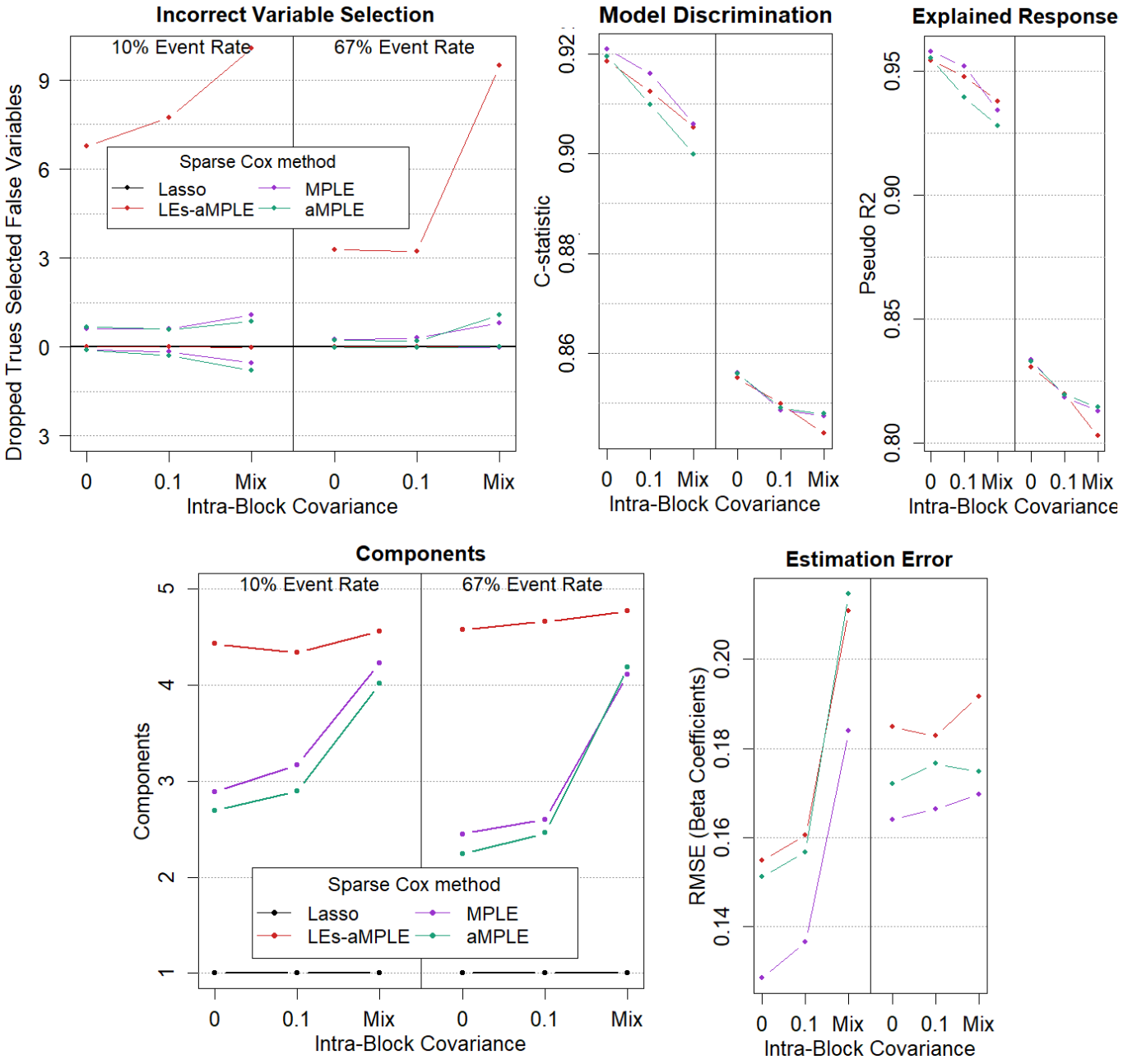


Figure E.7: Simulation results for 1000 subjects across event rates and correlation amounts when using **AIC-type criteria**



Appendix F

ANALYTICAL RESULTS FOR CHAPTER 4

F.1 Useful algebraic properties and assumptions

We begin by presenting several properties and assumptions that will be useful in the proofs of the following sections. When possible, we will present the weighted versions of the properties, with linear results following by taking the weighting matrix to be $\mathbf{I}_n$ and the working response to be simply the linear response. When in a weighted setting, we consider the Newton update $\mathbf{z}_t = \mathbf{X}\boldsymbol{\beta}_{t-1} + \mathbf{L}\boldsymbol{\gamma}_{t-1} + \mathbf{V}_t^{-1}(\boldsymbol{\delta} - \boldsymbol{\mu}_t)$. At each iteration, let $\mathcal{L} = \mathbf{L}(\mathbf{L}^T\mathbf{V}_t\mathbf{L})^{-1}\mathbf{L}^T\mathbf{V}_t$ be the hat matrix corresponding to $\mathbf{L}$, which is idempotent but not symmetric, and similarly define hat matrices $\mathcal{X}$ and $\mathcal{T}$.

First we point out that in the linear case, $(\mathbf{I} - \mathcal{L})$ is symmetric and idempotent; in the weighted case $(\mathbf{I}-\mathcal{L})^T\mathbf{V}_t = \mathbf{V}_t - \mathbf{V}_t\mathbf{L}(\mathbf{L}^T\mathbf{V}_t\mathbf{L})^{-1}\mathbf{L}^T\mathbf{V}_t = \mathbf{V}_t(\mathbf{I}-\mathcal{L})$, and it is idempotent since $(\mathbf{I} - \mathcal{L})^2 = \mathbf{I} - 2\mathcal{L} + \mathcal{L} = (\mathbf{I} - \mathcal{L})$. Also, $\mathbf{L}^T\mathbf{V}_t(\mathbf{I} - \mathcal{L}) = \mathbf{0}$.

Given a weighting matrix $\mathbf{V}$, we can use properties of matrix blockwise inversion and the assumption that both $\mathbf{X}^T\mathbf{V}\mathbf{X}$ and $\mathbf{L}^T\mathbf{V}\mathbf{L}$ are invertible to show:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{V}\mathbf{X} & \mathbf{X}^T\mathbf{V}\mathbf{L} \\ \mathbf{L}^T\mathbf{V}\mathbf{X} & \mathbf{L}^T\mathbf{V}\mathbf{L} \end{bmatrix}^{-1} = \begin{bmatrix} [\mathbf{X}^T\mathbf{V}\mathbf{X} - \mathbf{X}^T\mathbf{V}\mathcal{L}\mathbf{X}]^{-1} & \mathbf{0} \\ \mathbf{0} & [\mathbf{L}^T\mathbf{V}\mathbf{L} - \mathbf{L}^T\mathbf{V}\mathcal{X}\mathbf{L}]^{-1} \end{bmatrix} * \begin{bmatrix} \mathbf{I} & -\mathbf{X}^T\mathbf{V}\mathbf{L}(\mathbf{L}^T\mathbf{V}\mathbf{L})^{-1} \\ -\mathbf{L}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1} & \mathbf{I} \end{bmatrix}$$

$$= \begin{bmatrix} [\mathbf{X}^T\mathbf{V}(\mathbf{I} - \mathcal{L})\mathbf{X}]^{-1} & -[\mathbf{X}^T\mathbf{V}(\mathbf{I} - \mathcal{L})\mathbf{X}]^{-1}\ \mathbf{X}^T\mathbf{V}\mathbf{L}(\mathbf{L}^T\mathbf{V}\mathbf{L})^{-1} \\ -[\mathbf{L}^T\mathbf{V}(\mathbf{I} - \mathcal{X})\mathbf{L}]^{-1}\ \mathbf{L}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1} & [\mathbf{L}^T\mathbf{V}(\mathbf{I} - \mathcal{X})\mathbf{L}]^{-1} \end{bmatrix}$$

Thus, given a working outcome $\mathbf{z}$, we can express the coefficient estimates from a multiple

weighted least squares regression as follows:

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^T\mathbf{V}\mathbf{X} & \mathbf{X}^T\mathbf{V}\mathbf{L}\\ \mathbf{L}^T\mathbf{V}\mathbf{X} & \mathbf{L}^T\mathbf{V}\mathbf{L}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}^T\mathbf{V}\mathbf{z}\\ \mathbf{L}^T\mathbf{V}\mathbf{z}\end{bmatrix} = \begin{bmatrix}[\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{X}]^{-1}\ \mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z}\\ [\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{X})\mathbf{L}]^{-1}\ \mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{X})\mathbf{z}\end{bmatrix}$$
$$= \begin{bmatrix}[\mathbf{X}^T\mathbf{V}\mathbf{X}]^{-1}\ \mathbf{X}^T\mathbf{V}(\mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}})\\ [\mathbf{L}^T\mathbf{V}\mathbf{L}]^{-1}\ \mathbf{L}^T\mathbf{V}(\mathbf{z}-\mathbf{X}\widehat{\boldsymbol{\beta}})\end{bmatrix}. \tag{F.1}$$

We will prove Equation F.1 in the next section. In addition, a similar expression can be derived when using weighted PLS to estimate $\widehat{\boldsymbol{\beta}}$ by replacing $\mathbf{X}$ with components $\mathbf{T} = \mathbf{X}\mathbf{R}$ whose PLS weights are constructed to maximize $\mathbf{R}^T\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z}$.

We will be making use of the identity $(\mathbf{I}-\mathbf{P})^{-1} = \mathbf{I} + (\mathbf{I}-\mathbf{P})^{-1}\mathbf{P}$ and two forms of the inverse of a difference of matrices presented (for sums) in [34]:

$$(\Sigma - \mathbf{ABC})^{-1} = \Sigma^{-1} + \Sigma^{-1}\mathbf{A}\left(\mathbf{I}-\mathbf{BC}\Sigma^{-1}\mathbf{A}\right)^{-1}\mathbf{BC}\Sigma^{-1} \qquad (\dagger)$$
$$= \Sigma^{-1} + \Sigma^{-1}\mathbf{AB}\left(\mathbf{I}-\mathbf{C}\Sigma^{-1}\mathbf{AB}\right)^{-1}\mathbf{C}\Sigma^{-1}. \qquad (\dagger\dagger)$$

F.2 Proof of Remarks 1 and 2

Assume the same given $\mathbf{X}, \mathbf{L}, \mathbf{V}$, and outcome $\mathbf{z}$ as in Appendix F.1. We will first show that the multiple Weighted Least Squares coefficient estimates can be recovered via orthogonalized fits. Using the projection properties mentioned previously, see that modeling orthogonalized covariates $(\mathbf{I}-\mathcal{L})\mathbf{X}$ on the full outcome $\mathbf{z}$ will result in $\widehat{\boldsymbol{\beta}}$:

$$[\mathbf{X}^T(\mathbf{I}-\mathcal{L})^T\ \mathbf{V}\ (\mathbf{I}-\mathcal{L})\mathbf{X}]^{-1}\ \mathbf{X}^T(\mathbf{I}-\mathcal{L})^T\ \mathbf{V}\mathbf{z}\ =\ [\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{X}]^{-1}\ \mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z}.$$

The same holds for $\widehat{\boldsymbol{\gamma}}$ by substituting $\mathbf{X}$s and $\mathbf{L}$s.

Next we show that the statistically-adjusted coefficients can be derived using residual fits (i.e., Equation F.1). Specifically, we prove that

$$\left[\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z} = \left[\mathbf{X}^T\mathbf{V}\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{V}(\mathbf{z}-\mathbf{L}\widehat{\gamma}).$$

For purposes of space, let $\Sigma := \mathbf{X}^T\mathbf{V}\mathbf{X}$. We begin by proving the equality with all terms except $\mathbf{z}$. We indicate the lines that use the matrix inverse forms of [34] by the corresponding daggers.

$$
\begin{aligned}
&\left[\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L}) \;=\; \left[\Sigma - \mathbf{X}^T\mathbf{V}\mathcal{L}\mathbf{X}\right]^{-1}\;\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L}) \\
&= \Sigma^{-1}\;\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L}) + \Sigma^{-1}\mathbf{X}^T\mathbf{V}\mathcal{L}\left(\mathbf{I}-\mathbf{X}\Sigma^{-1}\mathbf{X}^T\mathbf{V}\mathcal{L}\right)^{-1}\mathbf{X}\Sigma^{-1}\;\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L}) \qquad (\dagger\dagger) \\
&= \Sigma^{-1}\mathbf{X}^T\mathbf{V} - \Sigma^{-1}\mathbf{X}^T\mathbf{V}\mathcal{L}\left[\mathbf{I}-(\mathbf{I}-\mathcal{X}\mathcal{L})^{-1}\mathcal{X}(\mathbf{I}-\mathcal{L})\right] \\
&= \Sigma^{-1}\mathbf{X}^T\mathbf{V} - \Sigma^{-1}\mathbf{X}^T\mathbf{V}\mathcal{L}\left[\mathbf{I}-\mathcal{X}\mathcal{L}\right]^{-1}(\mathbf{I}-\mathcal{X}\mathcal{L}-\mathcal{X}+\mathcal{X}\mathcal{L}) \\
&= \Sigma^{-1}\mathbf{X}^T\mathbf{V} - \Sigma^{-1}\mathbf{X}^T\mathbf{V}\mathcal{L}\left[\mathbf{I}-\mathcal{X}\mathcal{L}\right]^{-1}(\mathbf{I}-\mathcal{X}) \\
&= \Sigma^{-1}\mathbf{X}^T\mathbf{V} - \Sigma^{-1}\mathbf{X}^T\mathbf{V}\mathcal{L}\left(\mathbf{I}+\left[\mathbf{I}-\mathcal{X}\mathcal{L}\right]^{-1}\mathcal{X}\mathcal{L}\right)(\mathbf{I}-\mathcal{X}) \\
&= \Sigma^{-1}\mathbf{X}^T\mathbf{V} - \Sigma^{-1}\mathbf{X}^T\mathbf{V}\;\mathbf{L}\;(\mathbf{L}^T\mathbf{V}\mathbf{L})^{-1}\left[\mathbf{I}+\mathbf{L}^T\mathbf{V}\left[\mathbf{I}-\mathcal{X}\mathcal{L}\right]^{-1}\mathcal{X}\;\mathbf{L}(\mathbf{L}^T\mathbf{V}\mathbf{L})^{-1}\right]\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{X}) \\
&= \Sigma^{-1}\mathbf{X}^T\mathbf{V} - \Sigma^{-1}\mathbf{X}^T\mathbf{V}\mathbf{L}\;\left[\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{X})\mathbf{L}\right]^{-1}\;\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{X}) \qquad (\dagger) \\
&= \Sigma^{-1}\;\mathbf{X}^T\mathbf{V}\;\left(\mathbf{I}-\mathbf{L}\left[\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{X})\mathbf{L}\right]^{-1}\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{X})\right)
\end{aligned}
$$

Multiplying by $\mathbf{z}$ on the right side of the equations gives the final $\widehat{\beta}$ equality. Substituting all $\mathbf{X}$s and $\mathbf{L}$s results in the $\widehat{\gamma}$ equality to finish the proof of Equation F.1 and Remark 1. The linear case holds by setting $\mathbf{V} = \mathbf{I}$ and $\mathbf{z} = \mathbf{y}$.

We will next show that the multiple Weighted Least Squares coefficient estimates with PLS components $\mathbf{T}$ as covariates can be obtained from a residual fit as well. First, recall that we fit a WPLS model on the residual outcome $\mathbf{z} - \mathbf{L}\widehat{\gamma}_0 = (\mathbf{I}-\mathcal{L})\mathbf{z}$. Thus, the corresponding PLS weight $\mathbf{R}$ maximizes $\mathbf{R}^T\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z}$. Next, we consider the coefficient estimates from a multiple WLS with constructed components $\mathbf{T}$ and $\mathbf{L}$. These can be obtained easily by

substituting $\mathbf{T}$ for $\mathbf{X}$:

$$\begin{bmatrix} \widehat{\mathbf{q}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{T}^T\mathbf{V}\mathbf{T} & \mathbf{T}^T\mathbf{V}\mathbf{L} \\ \mathbf{L}^T\mathbf{V}\mathbf{T} & \mathbf{L}^T\mathbf{V}\mathbf{L} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{T}^T\mathbf{V}\mathbf{z} \\ \mathbf{L}^T\mathbf{V}\mathbf{z} \end{bmatrix} = \begin{bmatrix} [\mathbf{T}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{T}]^{-1}\ \mathbf{T}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z} \\ [\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{T})\mathbf{L}]^{-1}\ \mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{T})\mathbf{z} \end{bmatrix} \tag{F.2}$$

$$= \begin{bmatrix} [\mathbf{T}^T\mathbf{V}\mathbf{T}]^{-1}\ \mathbf{T}^T\mathbf{V}(\mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}}) \\ [\mathbf{L}^T\mathbf{V}\mathbf{L}]^{-1}\ \mathbf{L}^T\mathbf{V}(\mathbf{z}-\mathbf{T}\widehat{\mathbf{q}}) \end{bmatrix}. \tag{F.3}$$

While Equation F.3 holds due to the same proof as for Equation F.1, we point out one distinction of note. Can we simplify the coefficient residual fits to two lines as in WLS? Specifically, does it hold that $\widehat{\mathbf{q}} = \texttt{WPLS[} \mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}} \sim \mathbf{X} \texttt{]}$? Here, the PLS weight $\breve{\mathbf{R}}$ is maximizing $\breve{\mathbf{R}}^T\mathbf{X}^T\mathbf{V}(\mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}}) \neq \breve{\mathbf{R}}^T\mathbf{X}^T\mathbf{V}(\mathbf{z}-\mathbf{L}\widehat{\boldsymbol{\gamma}}_0)$. Thus, $\breve{\mathbf{R}} \neq \mathbf{R}$, so the constructed components will differ. These will only be equivalent when $\mathbf{X}^T\mathbf{V}\mathbf{L} = \mathbf{0}$.

However, there is a way that we can recover the adjusted coefficients using two orthogonalized fits. Consider modeling orthogonalized covariates $(\mathbf{I}-\mathcal{L})\mathbf{X}$ on the full outcome $\mathbf{z}$. Its PLS weight $\breve{\mathbf{R}}$ is maximizing $\breve{\mathbf{R}}^T\mathbf{X}^T(\mathbf{I}-\mathcal{L})^T\mathbf{V}\mathbf{z} = \breve{\mathbf{R}}^T\mathbf{X}^T\mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z}$, so $\breve{\mathbf{R}}$ is equivalent to $\mathbf{R}$. The resulting coefficient is $[\breve{\mathbf{R}}^T\mathbf{X}^T(\mathbf{I}-\mathcal{L})^T\ \mathbf{V}\ (\mathbf{I}-\mathcal{L})\mathbf{X}\breve{\mathbf{R}}]^{-1}\ \breve{\mathbf{R}}^T\mathbf{X}^T(\mathbf{I}-\mathcal{L})^T\ \mathbf{V}\mathbf{z}$, which reduces to $[\breve{\mathbf{R}}^T\mathbf{X}^T\ \mathbf{V}\ (\mathbf{I}-\mathcal{L})\mathbf{X}\breve{\mathbf{R}}]^{-1}\ \breve{\mathbf{R}}^T\mathbf{X}^T\ \mathbf{V}(\mathbf{I}-\mathcal{L})\mathbf{z}$ after using properties of projections. As $\breve{\mathbf{R}}$ is equivalent to $\mathbf{R}$, this coefficient is equal to $\widehat{\mathbf{q}}$ in Equation F.2. After defining the PLS component $\mathbf{T} = \mathbf{X}\breve{\mathbf{R}}$, the same projection properties allow us to recover the correct $\widehat{\boldsymbol{\gamma}}$ by modeling orthogonalized covariates $(\mathbf{I}-\mathcal{T})\mathbf{L}$ on the full outcome $\mathbf{z}$:

$$[\mathbf{L}^T(\mathbf{I}-\mathcal{T})^T\ \mathbf{V}\ (\mathbf{I}-\mathcal{T})\mathbf{L}]^{-1}\ \mathbf{L}^T(\mathbf{I}-\mathcal{T})^T\ \mathbf{V}\mathbf{z}\ =\ [\mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{T})\mathbf{L}]^{-1}\ \mathbf{L}^T\mathbf{V}(\mathbf{I}-\mathcal{T})\mathbf{z}.$$

Therefore, we have shown that exact equalities hold between the three regression procedures presented in Remark 2; the linear case can be seen by setting $\mathbf{V} = \mathbf{I}$ and $\mathbf{z} = \mathbf{y}$.

F.3 Connection between linear 1-stage-conditional adjustment techniques

In this section, we will demonstrate the connection between a multiple linear regression fit and 1-stage-conditional adjustment techniques for a linear outcome $\mathbf{y}$. Specifically, Table F.1 summarizes the linear predictors associated with each subset of covariates by adjustment technique. The SQN components are denoted by the symbol $\breve{\mathbf{T}}$ to emphasize that their PLS weights are optimizing a different value than CMB's weights. Although the INC weights will also be distinct, they approximate those of CMB (although they will not be equivalent regardless of the number of components and iterates).

Table F.1: Summary of linear predictors for 1-stage-conditional techniques

Adjustment	**Linear predictors when using (P)LS estimates**
Least Squares (LS)	$\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathcal{X}\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{X} + \mathcal{X}\mathcal{L} - \mathcal{X}\mathcal{L}\mathcal{X} + ...\right]\right)\mathbf{y}$ $\mathbf{L}\widehat{\boldsymbol{\gamma}} = \mathcal{L}\left(\mathbf{I} - \mathcal{X}\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{X} - \mathcal{L}\mathcal{X}\mathcal{L} + ...\right]\right)\mathbf{y}$
Sequential (SQN)	$\breve{\mathbf{T}}\widehat{\mathbf{q}} = \breve{\mathcal{T}}\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \breve{\mathcal{T}} + \breve{\mathcal{T}}\mathcal{L} - \breve{\mathcal{T}}\mathcal{L}\breve{\mathcal{T}} + ...\right]\right)\mathbf{y}$ $\mathbf{L}\widehat{\boldsymbol{\gamma}} = \mathcal{L}\left(\mathbf{I} - \breve{\mathcal{T}}\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\breve{\mathcal{T}} - \mathcal{L}\breve{\mathcal{T}}\mathcal{L} + ...\right]\right)\mathbf{y}$
Combination (CMB)	$\mathbf{T}\widehat{\mathbf{q}} = \mathcal{T}\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{X} + \mathcal{X}\mathcal{L} - \mathcal{X}\mathcal{L}\mathcal{X} + ...\right]\right)\mathbf{y}$ $\mathbf{L}\widehat{\boldsymbol{\gamma}} = \mathcal{L}\left(\mathbf{I} - \mathcal{T}\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{X} - \mathcal{L}\mathcal{X}\mathcal{L} + ...\right]\right)\mathbf{y}$
Incremental (INC)	$\mathbf{T}_t\widehat{\mathbf{q}}_t = \mathcal{T}_t\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{T}_{t-1} + \mathcal{T}_{t-1}\mathcal{L} - \mathcal{T}_{t-1}\mathcal{L}\mathcal{T}_{t-2} + ...\right]\right)\mathbf{y}$ $\mathbf{L}\widehat{\boldsymbol{\gamma}}_t = \mathcal{L}\left(\mathbf{I} - \mathcal{T}_t\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{T}_{t-1} - \mathcal{L}\mathcal{T}_{t-1}\mathcal{L} + ...\right]\right)\mathbf{y}$

We prove these linear predictor forms below. First, we demonstrate the telescoping pattern for the Least Squares linear predictors. We will make use of plugging in the following relationships:

$$\mathbf{L}\widehat{\boldsymbol{\gamma}} = \mathcal{L}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathcal{L}(\mathbf{y} - \mathcal{X}\left[\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}\right]) \qquad \text{and} \qquad \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathcal{X}(\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}).$$

$$\begin{aligned}
\text{Now, } \mathbf{L}\widehat{\boldsymbol{\gamma}} = \mathcal{L}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \;&=\; \mathcal{L}(\mathbf{y} - \mathcal{X}\left[\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}\right]) \;=\; \mathcal{L}\left(\mathbf{y} - \mathcal{X}\mathbf{y} + \mathcal{X}\mathcal{L}\left[\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right]\right) \\
&= \mathcal{L}\left(\mathbf{y} - \mathcal{X}\mathbf{y} + \mathcal{X}\mathcal{L}\mathbf{y} - \mathcal{X}\mathcal{L}\mathcal{X}\mathbf{y} + \mathcal{X}\mathcal{L}\mathcal{X}[\mathbf{L}\widehat{\boldsymbol{\gamma}}]\right) \\
&= \mathcal{L}\left(\mathbf{I} - \mathcal{X}\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{X} - \mathcal{L}\mathcal{X}\mathcal{L} + ...\right]\right)\mathbf{y}.
\end{aligned}$$

By substituting the $\mathbf{X}$ and $\mathbf{L}$ matrices, it follows that

$$\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathcal{X}\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{X} + \mathcal{X}\mathcal{L} - \mathcal{X}\mathcal{L}\mathcal{X} + \ldots\right]\right)\mathbf{y}.$$

The same proof holds for the Sequential-Adjustment linear predictors, simply by substituting $\breve{\mathbf{T}}$ for $\mathbf{X}$.

Now, Combination-Adjustment applies PLS to the residual $\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}$. The linear predictor from this PLS fit is $\mathbf{T}\widehat{\mathbf{q}} = \mathcal{T}\left(\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}\right)$. By applying the same telescoping pattern as with Least Squares, it follows that this linear predictor can be expressed as $\mathcal{T}\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{X} + \mathcal{X}\mathcal{L} - \mathcal{X}\mathcal{L}\mathcal{X} + \ldots\right]\right)\mathbf{y}$. Thus, Combination-Adjustment PLS coefficient is simply a projection of the LS coefficient. Substituting this value into the $\boldsymbol{\gamma}$ refit results in the following linear predictor: $\mathbf{L}\widehat{\boldsymbol{\gamma}} = \mathcal{L}\left(\mathbf{I} - \mathcal{T}\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{X} - \mathcal{L}\mathcal{X}\mathcal{L} + \ldots\right]\right)\mathbf{y}$.

Incremental-Adjustment is an iterative procedure that begins with initial estimates $(\widehat{\mathbf{q}}_0, \widehat{\boldsymbol{\gamma}}_0) = (\mathbf{0},\ [\mathbf{L}^T\mathbf{L}]^{-1}\mathbf{L}^T\mathbf{y})$. Its first iterates result in the following linear predictors.

$$\begin{aligned}
\mathbf{T}_1\widehat{\mathbf{q}}_1 &= \mathcal{T}_1\left(\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}_0\right) = \mathcal{T}_1\left(\mathbf{I} - \mathcal{L}\right)\mathbf{y} & \mathbf{L}\widehat{\boldsymbol{\gamma}}_1 &= \mathcal{L}\left(\mathbf{y} - \mathbf{T}_1\widehat{\mathbf{q}}_1\right) = \mathcal{L}\left(\mathbf{I} - \mathcal{T}_1 + \mathcal{T}_1\mathcal{L}\right)\mathbf{y} \\
\mathbf{T}_2\widehat{\mathbf{q}}_2 &= \mathcal{T}_2\left(\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}_1\right) & \mathbf{L}\widehat{\boldsymbol{\gamma}}_2 &= \mathcal{L}\left(\mathbf{y} - \mathbf{T}_2\widehat{\mathbf{q}}_2\right) \\
&= \mathcal{T}_2\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{T}_1 + \mathcal{T}_1\mathcal{L}\right]\right)\mathbf{y} & &= \mathcal{L}\left(\mathbf{I} - \mathcal{T}_2\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{T}_1 - \mathcal{L}\mathcal{T}_1\mathcal{L}\right]\right)\mathbf{y}
\end{aligned}$$

$$\begin{aligned}
\mathbf{T}_3\widehat{\mathbf{q}}_3 &= \mathcal{T}_3\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{T}_2 + \mathcal{T}_2\mathcal{L} - \mathcal{T}_2\mathcal{L}\mathcal{T}_1 + \mathcal{T}_2\mathcal{L}\mathcal{T}_1\mathcal{L}\right]\right)\mathbf{y} \\
\mathbf{L}\widehat{\boldsymbol{\gamma}}_3 &= \mathcal{L}\left(\mathbf{I} - \mathcal{T}_3\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{T}_2 - \mathcal{L}\mathcal{T}_2\mathcal{L} + \mathcal{L}\mathcal{T}_2\mathcal{L}\mathcal{T}_1 - \mathcal{L}\mathcal{T}_2\mathcal{L}\mathcal{T}_1\mathcal{L}\right]\right)\mathbf{y}.
\end{aligned}$$

$$\begin{aligned}
\text{Thus, } \mathbf{T}_t\widehat{\mathbf{q}}_t &= \mathcal{T}_t\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{T}_{t-1} + \mathcal{T}_{t-1}\mathcal{L} - \mathcal{T}_{t-1}\mathcal{L}\mathcal{T}_{t-2} + \ldots\right]\right)\mathbf{y} \\
\mathbf{L}\widehat{\boldsymbol{\gamma}}_t &= \mathcal{L}\left(\mathbf{I} - \mathcal{T}_t\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{T}_{t-1} - \mathcal{L}\mathcal{T}_{t-1}\mathcal{L} + \ldots\right]\right)\mathbf{y}.
\end{aligned}$$

Although the iterates of Incremental-Adjustment approximate the CMB coefficients, they will never exactly be equivalent because INC uses its iterate $\mathcal{T}$s in place of $\mathcal{X}$. As more components are used, the INC coefficients will better approximate those of CMB. However, as both INC coefficients contain all iterates in their solution, these will be slightly biased.

In fact, the slower Incremental-Adjustment (analogous to the current strategy used by Cox-PLS methods) forms the linear predictors listed below. This slower version will converge to the INC estimates, though it requires more iterates than INC in the linear setting.

$$\begin{aligned}
\mathbf{T}_1\widehat{\mathbf{q}}_1 &= \mathcal{T}_1\mathbf{y} & \mathbf{L}\widehat{\boldsymbol{\gamma}}_1 &= \mathcal{L}\mathbf{y} \\
\mathbf{T}_2\widehat{\mathbf{q}}_2 &= \mathcal{T}_2\left(\mathbf{y} - \mathbf{L}\widehat{\boldsymbol{\gamma}}_1\right) = \mathcal{T}_2\left(\mathbf{I} - \mathcal{L}\right)\mathbf{y} & \mathbf{L}\widehat{\boldsymbol{\gamma}}_2 &= \mathcal{L}\left(\mathbf{I} - \mathcal{T}_1\right)\mathbf{y} \\
\mathbf{T}_3\widehat{\mathbf{q}}_3 &= \mathcal{T}_3\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{T}_1\right]\right)\mathbf{y} & \mathbf{L}\widehat{\boldsymbol{\gamma}}_3 &= \mathcal{L}\left(\mathbf{I} - \mathcal{T}_2\left[\mathbf{I} - \mathcal{L}\right]\right)\mathbf{y} \\
\mathbf{T}_4\widehat{\mathbf{q}}_4 &= \mathcal{T}_4\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{T}_2(\mathbf{I} - \mathcal{L})\right]\right)\mathbf{y} & \mathbf{L}\widehat{\boldsymbol{\gamma}}_4 &= \mathcal{L}\left(\mathbf{I} - \mathcal{T}_3\left[\mathbf{I} - \mathcal{L}(\mathbf{I} - \mathcal{T}_1)\right]\right)\mathbf{y} \\
&= \mathcal{T}_4\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{T}_2 + \mathcal{T}_2\mathcal{L}\right]\right)\mathbf{y} & &= \mathcal{L}\left(\mathbf{I} - \mathcal{T}_3\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{T}_1\right]\right)\mathbf{y}
\end{aligned}$$

$$\text{Thus, } \mathbf{T}_t\widehat{\mathbf{q}}_t = \mathcal{T}_t\left(\mathbf{I} - \mathcal{L}\left[\mathbf{I} - \mathcal{T}_{t-2} + \mathcal{T}_{t-2}\mathcal{L} - \mathcal{T}_{t-2}\mathcal{L}\mathcal{T}_{t-4} + \ldots\right]\right)\mathbf{y}$$

$$\mathbf{L}\widehat{\boldsymbol{\gamma}}_t = \mathcal{L}\left(\mathbf{I} - \mathcal{T}_{t-1}\left[\mathbf{I} - \mathcal{L} + \mathcal{L}\mathcal{T}_{t-3} - \mathcal{L}\mathcal{T}_{t-3}\mathcal{L} + \ldots\right]\right)\mathbf{y}.$$

This concludes our derivation of the connection between a multiple linear regression fit and 1-stage-conditional adjustment techniques for a linear outcome.

Appendix G

EXTENDED SIMULATION RESULTS FOR CHAPTER 4

Simulation study for a continuous outcome

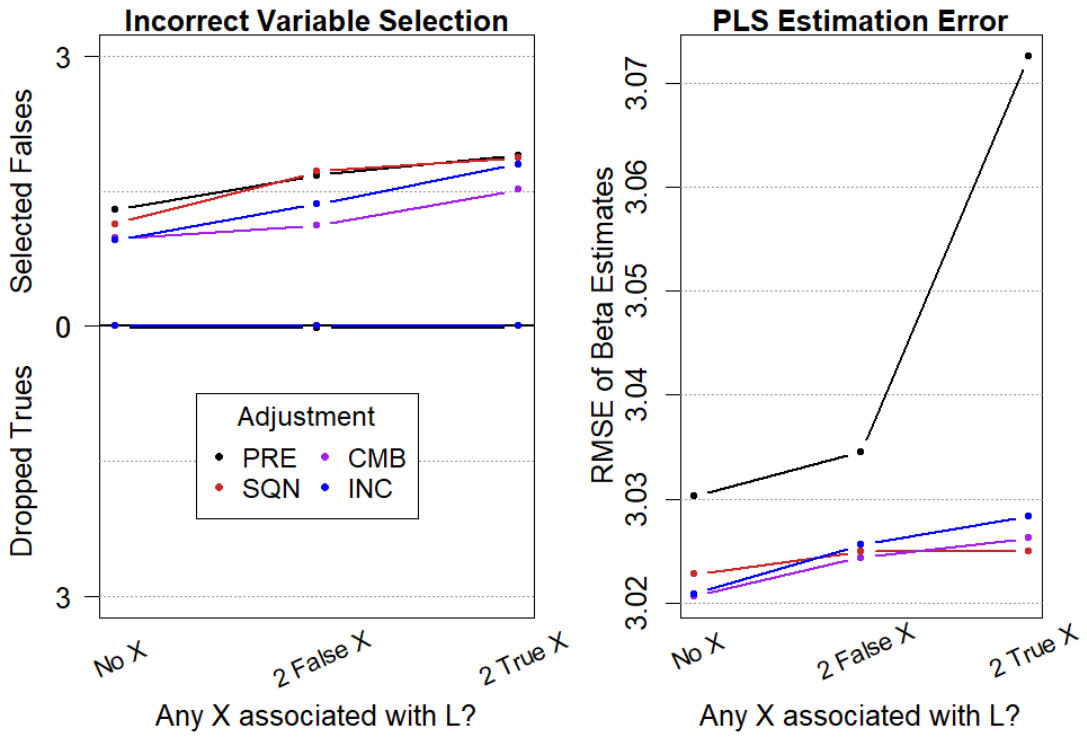
To compare adjustment techniques for linear models while incorporating sparsity, we fit Chun's Sparse PLS procedure instead of PLS to model the **X** covariates, and we use least squares to model the **L** covariates. We generate the same multiblock **X** predictors as in previous chapters, with intra-block covariance of .05, .50, .30, and .10 respectively, and only 9 of the 24 predictors contributing to the linear outcome. $\mathbf{L}_{1000\times 4}$ is generated such that the first three variables emulate precision variables and the last either another precision variable or a confounder. Specifically, the first two variables are binomial and Poisson distributed with roughly .40 correlation but independent of **X**. The second two are normally-distributed with .18 correlation, the last of which is either independent or correlated (magnitude of .30) with two true **X** variables. The mean model of our linear outcome is constructed from the true **X** covariates (with $\beta = \pm 1$) and **L** covariates (with $\gamma = \pm 5$). We vary the amount of Gaussian noise added to the outcome, ranging from sd=1 to sd=11.

Across 50 generated data sets, we split our observations into a training (n=700) and test (n=300) set, and we use 10-fold CV on the training data to select our tuning parameters (sparsity and number of components) using the criteria of minimum CV RMSE. On the test data, we evaluate the variable selection properties (true variables of 9 and false variables of 15), predictive ability (RMSE of outcome; not shown), and PLS estimation error (RMSE of $\widehat{\boldsymbol{\beta}}$). Note that the last metric compares $\widehat{\boldsymbol{\beta}}$ with the known $\boldsymbol{\beta}$; an ideal method will select all true variables and not over-shrink their estimate while discarding most false variables and only allowing small magnitudes for the estimates of those it falsely selects.

We see that with smaller variability Gaussian noise, all methods are able to select the

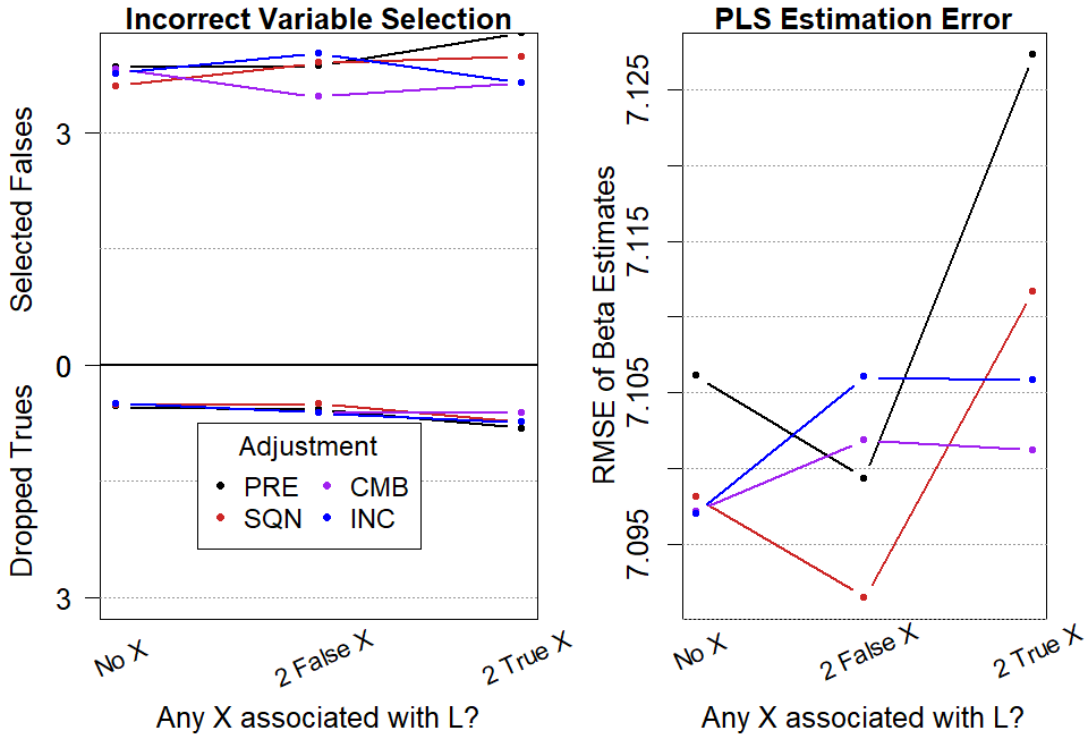
entire true set of covariates, with Combination-Adjustment best equipped to discard false variables. However, the largest discrepancy is that Pre-Adjustment results in the worst PLS estimation error regardless of correlation structure.

Figure G.1: Average simulation results for conditional adjustment on 50 multiblock data sets ($n = 1000,\ p = 28,\ |\frac{\gamma}{\beta}| = 5$) with **sd(y)=3**, across adjustment variable settings.



With more substantial noise (sd=7), the methods drop a fraction of a true variable while selecting roughly four false variables. Again, Combination-Adjustment tends to do the best in variable selection across correlation settings. This technique also performs the best in PLS estimation error when there is the presence of a confounder (as expected), but Sequential-Adjustment outperfoms CMB when **L** is associated with two false **X** variables (also as expected). Moreover, Pre-Adjustment always performs worse than Sequential-Adjustment in terms of estimation error.

Figure G.2: Average simulation results for conditional adjustment on 50 multiblock data sets ($n = 1000,\ p = 28,\ |\frac{\gamma}{\beta}| = 5$) with **sd(y)=7**, across adjustment variable settings.



Simulation study for a survival outcome

To compare adjustment techniques for Cox models while incorporating sparsity, we keep most of the same setup as in the linear model case. The main difference is that we set the variability of the added Gaussian noise to the linear predictor to have sd=1, and we now vary the relative importance between $\mathbf{X}$ and $\mathbf{L}$ by considering $|\frac{\gamma}{\beta}|$ from 1 to 6. Full details on the simulation set up can be found in Section 4.5.

Figure G.3: Average simulation results for **1-stage-conditional adjustment** on 50 multi-block data sets ($n = 1000,\ p = 28,\ |\frac{\gamma}{\beta}| = \mathbf{2}$), across adjustment variable settings.

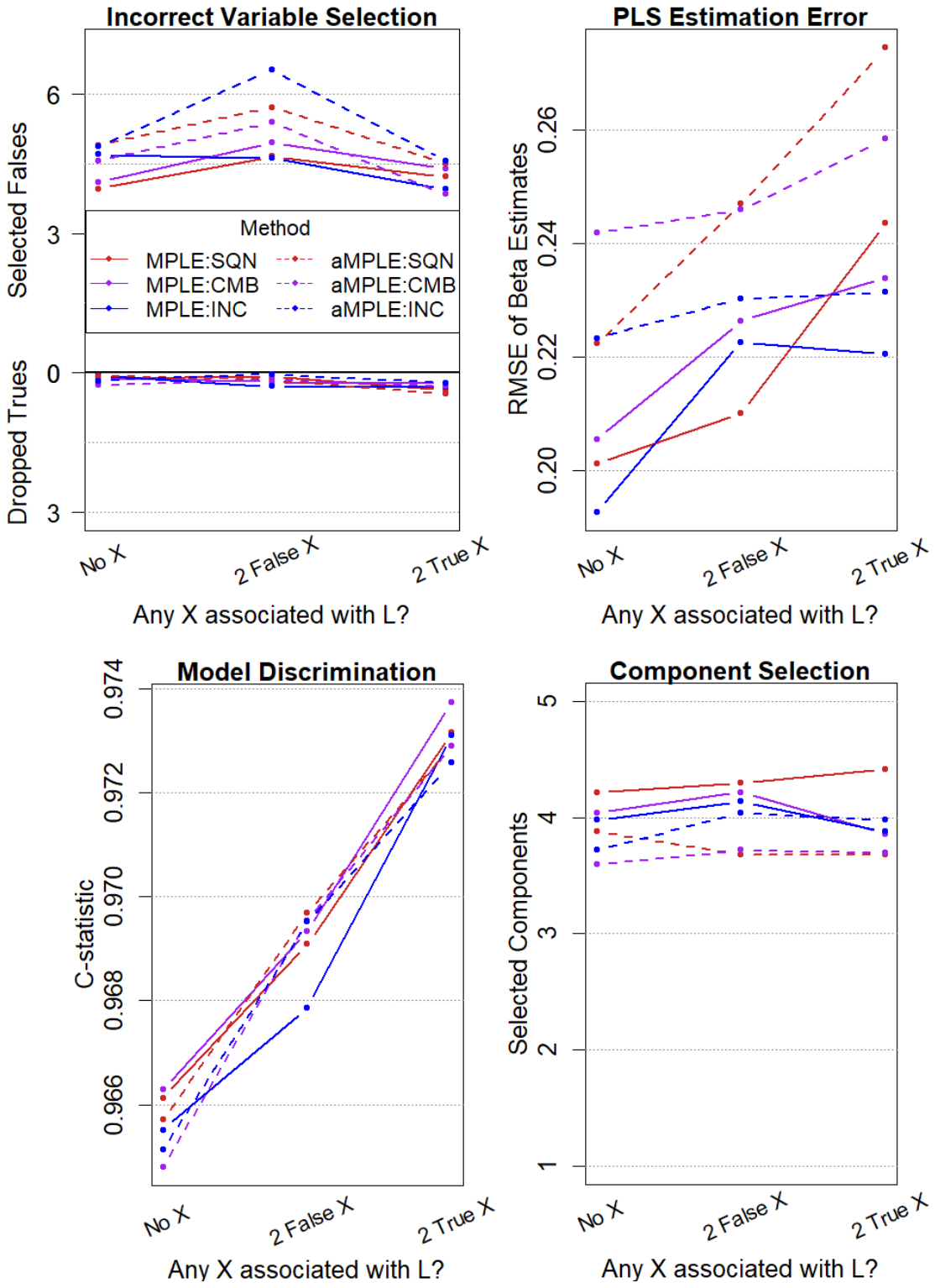


Figure G.4: Average simulation results for **1-stage-conditional adjustment** on 50 multi-block data sets ($n = 1000,\ p = 28,\ |\frac{\gamma}{\beta}| = \mathbf{5}$), across adjustment variable settings.

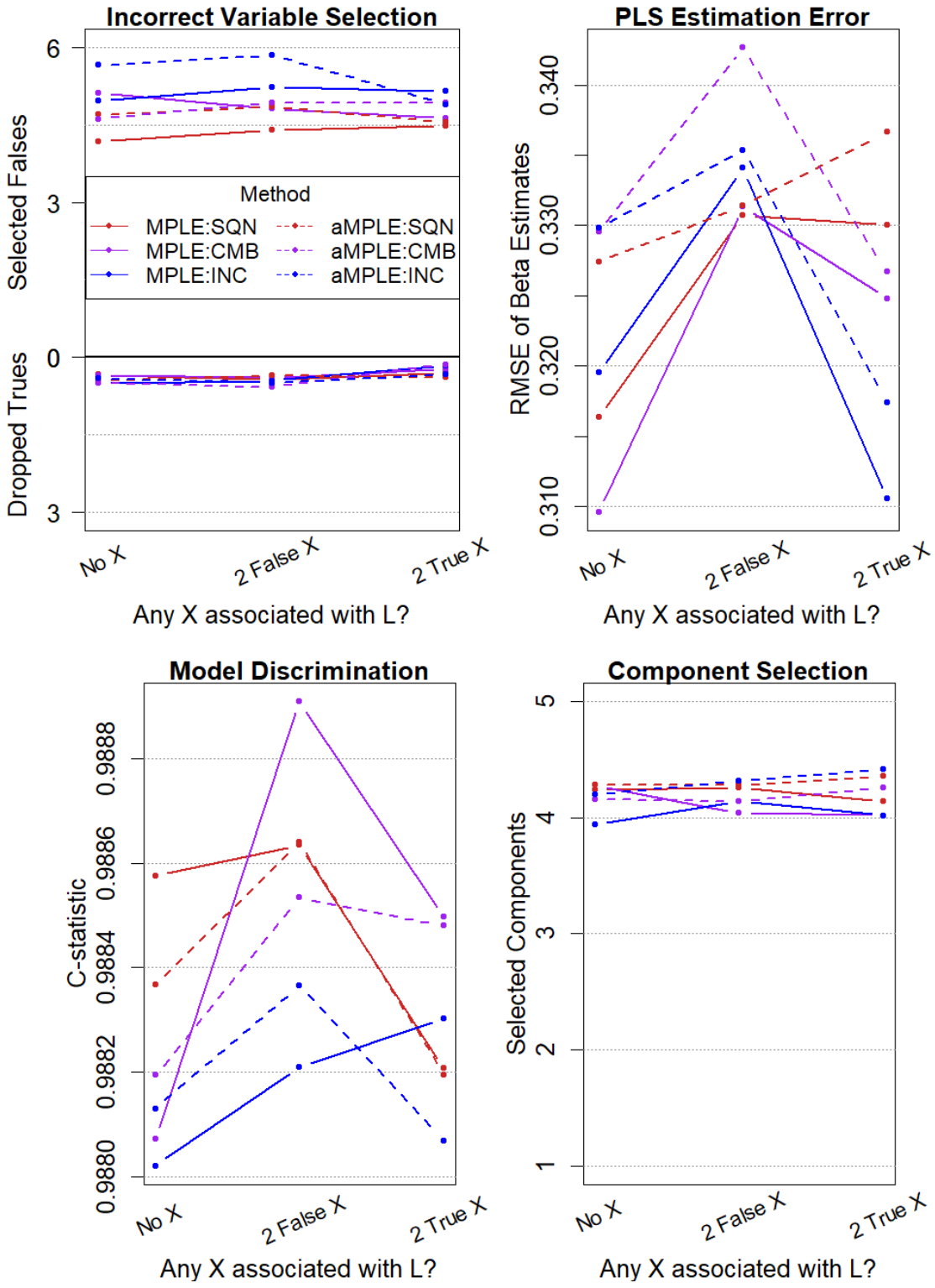


Figure G.5: Average simulation results for **2-stage-conditional adjustment** on 50 multi-block data sets ($n = 1000,\ p = 28,\ |\frac{\gamma}{\beta}| = \mathbf{2}$), across adjustment variable settings.

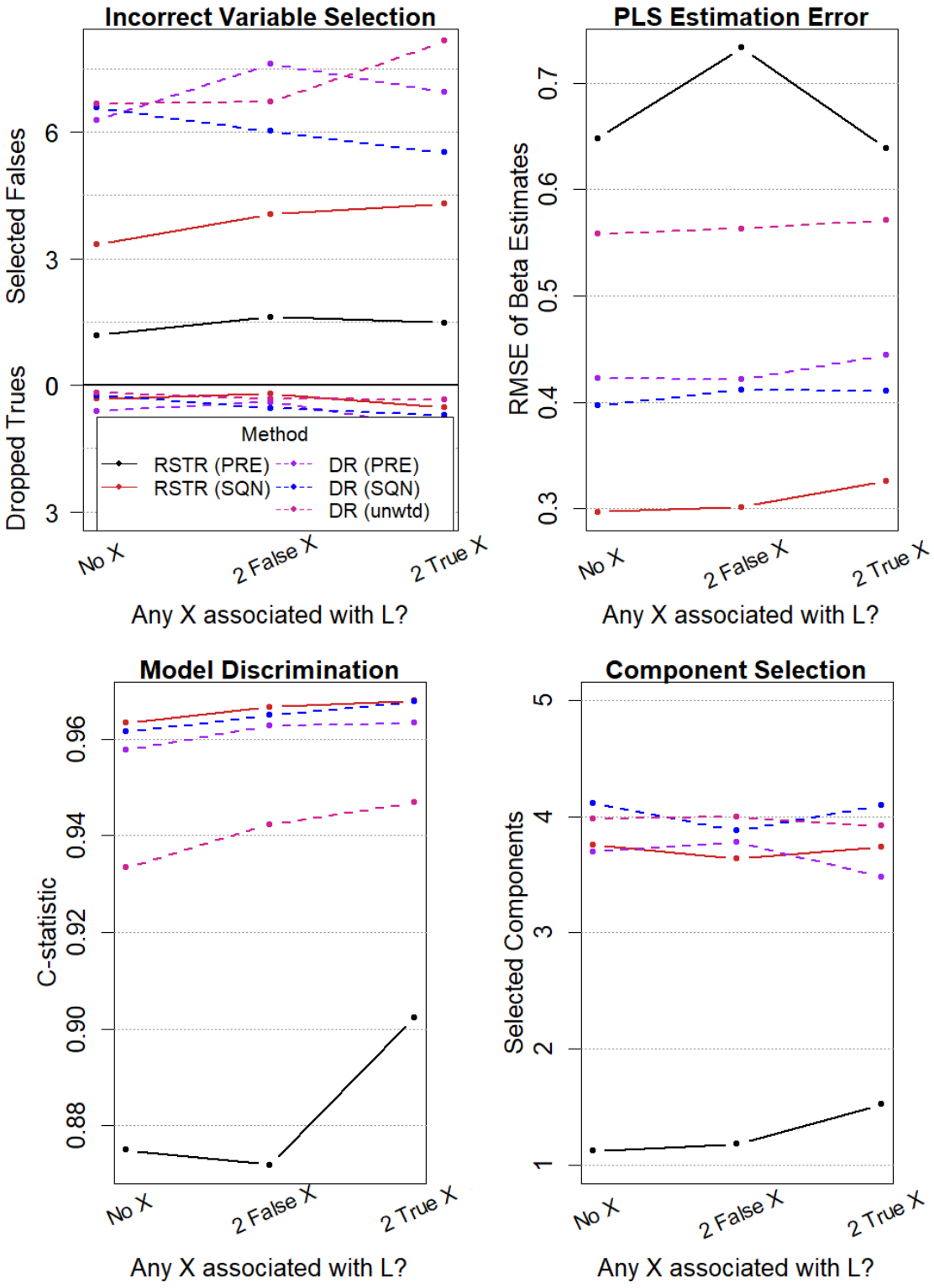
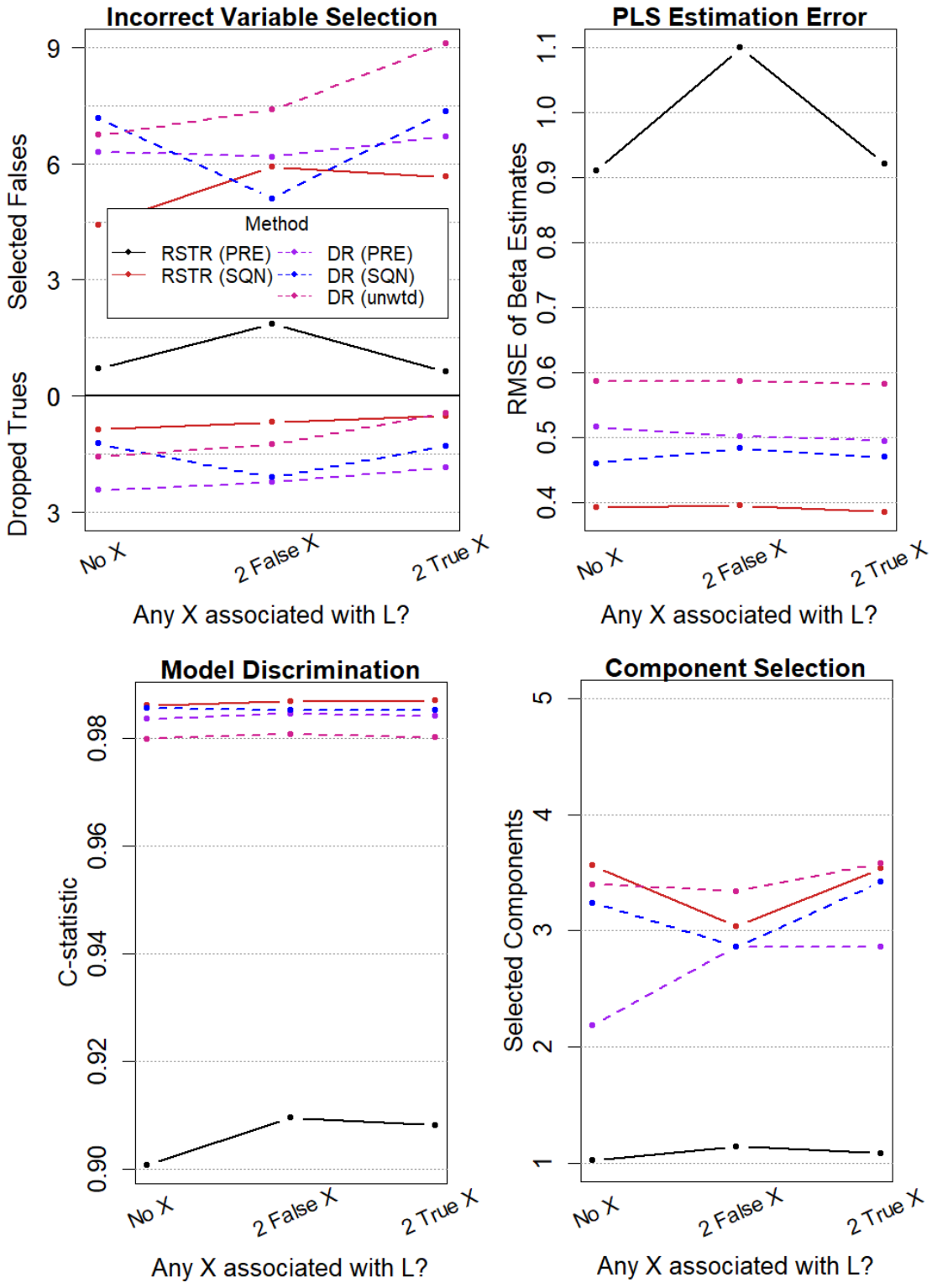


Figure G.6: Average simulation results for **2-stage-conditional adjustment** on 50 multi-block data sets ($n = 1000,\ p = 28,\ |\frac{\gamma}{\beta}| = \mathbf{5}$), across adjustment variable settings.



BIBLIOGRAPHY

[1] Barbara E Ainsworth, Melinda L Irwin, Cheryl L Addy, Melicia C Whitt, and Lisa M Stolarczyk. Moderate physical activity patterns of minority women: The Cross-Cultural Activity Participation study. *Journal of Women's Health & Gender-based Medicine*, 8(6):805–813, 1999.

[2] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[4] Philippe Bastien. Deviance residuals based PLS regression for censored data in high dimensional setting. *Chemometrics and Intelligent Laboratory Systems*, 91(1):78–86, 2008.

[5] Philippe Bastien, Frédéric Bertrand, Nicolas Meyer, and Myriam Maumy-Bertrand. Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics*, 31(3):397–404, 2015.

[6] Philippe Bastien and Michel Tenenhaus. PLS generalised linear regression: Application to the analysis of life time data. In *PLS and Related Methods, Proceedings of the PLS '01 International Symposium, CISIA-CERESTA, Paris*, pages 131–140, 2001.

[7] Caroline Bazzoli and Sophie Lambert-Lacroix. Classification based on extensions of LS-PLS using logistic regression: Application to clinical and multiple genomic data. *BMC Bioinformatics*, 19(1):1–13, 2018.

[8] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

[9] Diane E Bild, David A Bluemke, Gregory L Burke, Robert Detrano, Ana V Diez Roux, Aaron R Folsom, Philip Greenland, David R Jacobs Jr, Richard Kronmal, Kiang Liu, et al. Multi-Ethnic Study of Atherosclerosis: Objectives and design. *American Journal of Epidemiology*, 156(9):871–881, 2002.

[10] Hege M Bøvelstad, Ståle Nygård, and Ørnulf Borgan. Survival prediction from clinico-genomic models—a comparative study. *BMC bioinformatics*, 10(1):1–9, 2009.

[11] Norman E Breslow. Contribution to discussion of paper by Dr. Cox. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34:216–217, 1972.

[12] John B Carroll. *Exploratory factor analysis: A tutorial*. Ablex Publishing Corporation, 1985.

[13] Arthur Cayley. A memoir on the theory of matrices. *Philosophical transactions of the Royal Society of London*, II(148):17–37, 1858.

[14] Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

[15] Dongjun Chung, Hyonho Chun, and Sunduz Keleş. Spls: Sparse Partial Least Squares (SPLS) regression and classification. R package version 2.2–1, 2013.

[16] Dongjun Chung and Sunduz Keles. Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[17] Dietary Guidelines Advisory Committee et al. *Dietary Guidelines for Americans 2015-2020*. Government Printing Office, 2015.

[18] Dietary Guidelines Advisory Committee et al. Scientific report of the 2015 Dietary Guidelines Advisory Committee: Advisory report to the Secretary of Health and Human Services and the Secretary of Agriculture. *Agricultural Research Service*, 2015.

[19] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):187–220, 1972.

[20] Raffaele De Caterina, Antonella Zampolli, Serena Del Turco, Rosalinda Madonna, and Marika Massaro. Nutritional mechanisms that influence cardiovascular disease. *The American Journal of Clinical Nutrition*, 83(2):421S–426S, 2006.

[21] Sijmen De Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

[22] Sanjoy Dey, Rohit Gupta, Michael Steinbach, and Vipin Kumar. Integration of clinical and genomic data: A methodological survey. Technical report, 13-005. University of Minnesota Digital Conservancy, 2013.

[23] Julia R DiBello, Peter Kraft, Stephen T McGarvey, Robert Goldberg, Hannia Campos, and Ana Baylin. Comparison of 3 methods for identifying dietary patterns associated with risk of disease. *American Journal of Epidemiology*, 168(12):1433–1443, 2008.

[24] Harold Edson Driver and Alfred Louis Kroeber. *Quantitative Expression of Cultural Relationships*, volume 31. University of California Press, 1932.

[25] Massimo Fornasier, Steffen Peter, Holger Rauhut, and Stephan Worm. Conjugate gradient acceleration of iteratively re-weighted least squares methods. *Computational Optimization and Applications*, 65(1):205–259, 2016.

[26] Ildiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[27] Laurence S Freedman, Victor Kipnis, Charles C Brown, Arthur Schatzkin, Sholom Wacholder, and Anne M Hartman. Comments on "Adjustment for total energy intake in epidemiologic studies". *The American Journal of Clinical Nutrition*, 65(4):1229S–1231S, 1997.

[28] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.4, 2009.

[29] Jerome Friedman and Bogdan E Popescu. Gradient directed regularization for linear regression and classification. Technical report, Statistics Department, Stanford University, 2003.

[30] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

[31] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[32] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

[33] Inge S Helland. On the structure of partial least squares regression. *Communications in Statistics—Simulation and Computation*, 17(2):581–607, 1988.

[34] Harold V Henderson and Shayle R Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60, 1981.

[35] M Heron and RN Anderson. Changes in the leading cause of death: Recent patterns in heart disease and cancer mortality. Technical report, NCHS Data Brief, No 254. Hyattsville, MD: National Center for Health Statistics, 2016.

[36] Kurt Hoffmann, Matthias B Schulze, Anja Schienkiewitz, Ute Nöthlings, and Heiner Boeing. Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *American Journal of Epidemiology*, 159(10):935–944, 2004.

[37] Frank B Hu. Dietary pattern analysis: A new direction in nutritional epidemiology. *Current Opinion in Lipidology*, 13(1):3–9, 2002.

[38] Anne Krog Ingerslev, Ibrahim Karaman, Murat Bagcıoglu, Achim Kohler, Peter Kappel Theil, Knud Erik Bach Knudsen, and Mette Skou Hedemann. Whole grain consumption increases gastrointestinal content of sulfate-conjugated oxylipins in pigs—a multicompartmental metabolomics study. *Journal of Proteome Research*, 14(8):3095–3110, 2015.

[39] David R Jacobs Jr and Lyn M Steffen. Nutrients, foods, and dietary patterns as exposures in research: A framework for food synergy. *The American Journal of Clinical Nutrition*, 78(3):508S–513S, 2003.

[40] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[41] Kjetil Jørgensen, Vegard Segtnan, Kari Thyholt, and Tormod Næs. A comparison of methods for analysing regression models with both spectral and designed variables. *Journal of Chemometrics*, 18(10):451–464, 2004.

[42] Ibrahim Karaman, Natalja P Nørskov, Christian Clement Yde, Mette Skou Hedemann, Knud Erik Bach Knudsen, and Achim Kohler. Sparse multi-block PLSR for biomarker discovery when integrating data from LC-MS and NMR metabolomics. *Metabolomics*, 11(2):367–379, 2015.

[43] Susan M Krebs-Smith, TusaRebecca E Pannucci, Amy F Subar, Sharon I Kirkpatrick, Jennifer L Lerman, Janet A Tooze, Magdalena M Wilson, and Jill Reedy. Update of the Healthy Eating Index: HEI-2015. *Journal of the Academy of Nutrition and Dietetics*, 118(9):1591–1602, 2018.

[44] Kim-Anh Lê Cao, Debra Rossouw, Christele Robert-Granié, and Philippe Besse. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

[45] Donghwan Lee, Youngjo Lee, Yudi Pawitan, and Woojoo Lee. Sparse partial least-squares regression for high-throughput survival data analysis. *Statistics in Medicine*, 32(30):5340–5352, 2013.

[46] Youngjo Lee and Hee-Seok Oh. A new sparse variable selection via random-effect model. *Journal of Multivariate Analysis*, 125:89–99, 2014.

[47] Evangelina López de Maturana, Lola Alonso, Pablo Alarcón, Isabel Adoración Martín-Antoniano, Silvia Pineda, Lucas Piorno, M Luz Calle, and Núria Malats. Challenges in the integration of omics and non-omics data. *Genes*, 10(3):238, 2019.

[48] Brian D Marx. Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, 38(4):374–381, 1996.

[49] Elizabeth J Mayer-Davis, Mara Z Vitolins, Suzan L Carmichael, Sandra Hemphill, Georgia Tsaroucha, Julia Rushing, and Sarah Levin. Validity and reproducibility of a food frequency interview in a multi-cultural epidemiologic study. *Annals of Epidemiology*, 9(5):314–324, 1999.

[50] Jose Medina-Inojosa, Nathalie Jean, Mery Cortes-Bergoderi, and Francisco Lopez-Jimenez. The Hispanic paradox in cardiovascular disease and total mortality. *Progress in Cardiovascular Diseases*, 57(3):286–292, 2014.

[51] Tahir Mehmood and Bilal Ahmed. The diversity in the applications of partial least squares: An overview. *Journal of Chemometrics*, 30(1):4–17, 2016.

[52] Bjørn-Helge Mevik, Ron Wehrens, and Kristian Hovde Liland. pls: Partial least squares and principal component regression. R package, version 2.3, 2011.

[53] Stephen G. Nash and Ariela Sofer. *Linear and Nonlinear Programming*. McGraw-Hill, 1996.

[54] Danh V Nguyen and David M Rocke. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, 18(12):1625–1632, 2002.

[55] Ståle Nygård, Ørnulf Borgan, Ole Christian Lingjærde, and Hege Leite Størvold. Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis*, 14(2):179–195, 2008.

[56] John O'Quigley, Ronghui Xu, and Janez Stare. Explained randomness in proportional hazards models. *Statistics in Medicine*, 24(3):479–489, 2005.

[57] Peter J Park, Lu Tian, and Isaac S Kohane. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18(Suppl 1):S120–S127, 2002.

[58] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[59] A Phatak, PM Reilly, and A Penlidis. The asymptotic variance of the univariate PLS estimator. *Linear Algebra and its Applications*, 354(1-3):245–253, 2002.

[60] Aloke Phatak and Frank de Hoog. Exploiting the connection between PLS, Lanczos methods and conjugate gradients: Alternative proofs of some properties of PLS. *Journal of Chemometrics*, 16(7):361–367, 2002.

[61] S Joe Qin, Sergio Valle, and Michael J Piovoso. On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics*, 15(9):715–742, 2001.

[62] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In Craig Saunders, Marko Grobelnik, Steve Gunn, and John Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection*, pages 34–51. Springer, 2006.

[63] Matthias B Schulze and Kurt Hoffmann. Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. *British Journal of Nutrition*, 95(5):860–869, 2006.

[64] Mark R Segal. Microarray gene expression data with linked survival phenotypes: Diffuse large-B-cell lymphoma revisited. *Biostatistics*, 7(2):268–285, 2006.

[65] Petre Stoica and Torsten Söderström. Partial least squares: A first-order analysis. *Scandinavian Journal of Statistics*, 25(1):17–24, 1998.

[66] Laura P Svetkey, Frank M Sacks, Eva Obarzanek, William M Vollmer, Lawrence J Appel, Pao-Hwa Lin, Njeri M Karanja, David W Harsha, George A Bray, Mikel Aickin, et al. The DASH diet, sodium intake and blood pressure trial (DASH-sodium): Rationale and design. *Journal of the American Dietetic Association*, 99(8):S96–S104, 1999.

[67] R Core Team et al. R: A language and environment for statistical computing, 2013.

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

[69] Robert Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

[70] Antonia Trichopoulou, Tina Costacou, Christina Bamia, and Dimitrios Trichopoulos. Adherence to a Mediterranean diet and survival in a Greek population. *New England Journal of Medicine*, 348(26):2599–2608, 2003.

[71] Anastasios A Tsiatis et al. A large sample study of Cox's regression model. *The Annals of Statistics*, 9(1):93–108, 1981.

[72] Hilko van der Voet. Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems*, 25(2):313–323, 1994.

[73] Mary Ann S Van Duyn and Elizabeth Pivonka. Overview of the health benefits of fruit and vegetable consumption for the dietetics professional: Selected literature. *Journal of the American Dietetic Association*, 100(12):1511–1521, 2000.

[74] Hans C Van Houwelingen, Tako Bruinsma, Augustinus AM Hart, Laura J Van't Veer, and Lodewyk FA Wessels. Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, 25(18):3201–3216, 2006.

[75] Pierre JM Verweij and Hans C Van Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12(24):2305–2314, 1993.

[76] Eric Vittinghoff and Charles E McCulloch. Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, 165(6):710–718, 2007.

[77] LE Wangen and BR Kowalski. A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, 3(1):3–20, 1989.

[78] Johan A Westerhuis, Theodora Kourti, and John F MacGregor. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12(5):301–321, 1998.

[79] Walter C Willett, Geoffrey R Howe, and Lawrence H Kushi. Adjustment for total energy intake in epidemiologic studies. *The American Journal of Clinical Nutrition*, 65(4):1220S–1228S, 1997.

[80] Jane E Winter, Robert J MacInnis, Naiyana Wattanapenpaiboon, and Caryl A Nowson. BMI and all-cause mortality in older adults: A meta-analysis. *The American Journal of Clinical Nutrition*, 99(4):875–890, 2014.

[81] Herman Wold. Nonlinear estimation by iterative least squares procedures. In FN David and J Neyman, editors, *Research Papers in Statistics*, pages 411–444. Wiley, 1966.

[82] Svante Wold, Martin Høy, Harald Martens, Johan Trygg, Frank Westad, John MacGregor, and Barry M Wise. The PLS model space revisited. *Journal of Chemometrics*, 23(2):67–68, 2009.

[83] Tiffany C Yang, Lorna S Aucott, Garry G Duthie, and Helen M Macdonald. An application of partial least squares for identifying dietary patterns in bone health. *Archives of Osteoporosis*, 12(1):63, 2017.

[84] Fengqing Zhang, Tinashe M Tapera, and Jiangtao Gou. Application of a new dietary pattern analysis method in nutritional epidemiology. *BMC Medical Research Methodology*, 18(1):119, 2018.

[85] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[86] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[87] Hui Zou and Trevor Hastie. Elasticnet: elastic-net for sparse estimation and sparse PCA. R package, version 1.1, 2012.

[88] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.