Title
Nonparametric Methods for High-Dimensional Data Analysis

Author
Boileau, Philippe

Publication Date
2023

Peer reviewed|Thesis/dissertation

Nonparametric Methods for High-Dimensional Data Analysis

by

Philippe A Boileau

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sandrine Dudoit, Chair
Professor Mark van der Laan
Professor Alan Hubbard
Professor Andres Cardenas

Summer 2023

Nonparametric Methods for High-Dimensional Data Analysis

Abstract

Nonparametric Methods for High-Dimensional Data Analysis

by

Philippe A Boileau

Doctor of Philosophy in Biostatistics

and the Designated Emphasis in

Computational and Genomic Biology

University of California, Berkeley

Professor Sandrine Dudoit, Chair

Modern biomedical studies generate high-dimensional data, meaning that the number of variables collected is equal to or larger than the number of observations. Examples are numerous, including single-cell transcriptome analyses and clinical trials. The dimensions of these data often prevent the use of traditional statistical methods: the theory motivating their application is no longer applicable. New procedures must be developed for express use in these contexts to ensure trustworthy inference. This dissertation delves into two topics in high-dimensional statistics, the first being covariance matrix estimator selection. Motivated by the need to nonparametrically identify an optimal estimator of the covariance matrix for a given dataset, we propose a cross-validated estimator selection procedure and investigate its finite-sample and high-dimensional asymptotic performance. Our theoretical results, supported by empirical evidence, demonstrate that this procedure selects the optimal estimator asymptotically. Here, optimality is defined in terms of a Frobenius-norm-based risk. Applications are myriad, though we focus on improving exploratory analyses in single-cell transcriptome analyses. The second topic, born of the need to reliably uncover biomarkers that predict clinical trial patients' response to novel therapies, is treatment effect modifier discovery. Treatment effect modifiers are pre-treatment covariates that influence the effect of a treatment on an outcome. While many approaches exist for identifying these effect modifiers in traditional asymptotic settings, few developments have been made for high-dimensional data. We propose a nonparametric framework for defining parameters measuring treatment effect modification, deriving accompanying estimators, and establishing these estimators' asymptotic properties. We derive several such parameters and estimators using our methodology, and assess these estimators' empirical performance through comprehensive simulation studies and real clinical trial data analyses.

To Melissa,
for whom this dissertation is unintelligible
yet has contributed to its content for years.

Contents

List of Figures

List of Tables

Acknowledgments

Completing this journey would have been impossible without the guidance and support of my family, friends, collaborators and mentors. I am particularly indebted to my coauthors of the works presented in this dissertation: Sandrine Dudoit, Mark van der Laan, Nima Hejazi, Ning Leng and Nina Qi.

Chapter 1

Introduction

The biomedical sciences depend on data and statistics to pursue all manner of quantitative investigations. From establishing the efficacy of novel therapies in small clinical trials to uncovering genetic mutations associated with disease susceptibility in genome-wide associate studies of entire populations, statistical reasoning, theory, and methods are required to translate substantive problems into mathematical questions, collect and organize data, perform analyses, and interpret findings.

Whether the resulting findings help answer the original scientific question depends on the statistical models' and methods' ability to faithfully capture the complexity of the problem. In part due to the limited availability of computing power, statistical analyses of biomedical data have traditionally relied on relatively simple parametric models. These models make simplifying assumptions about the data-generating process (DGP) when translating the scientific question into statistical terms, permitting practitioners to perform approximate, computationally efficient inference. The reliability of this inference, however, depends on the quality of the approximation, which is generally unknowable. This uncertainty is unacceptable in high-stake situations, like when assessing the safety and efficacy of potentially life-altering medications in clinical trials.

Semi- and nonparametric modeling approaches have been developed in response to the limitations of parametric procedures. In combination with the increased availability of computing resources, these more flexible frameworks permit the development and use of statistical models that more accurately encode the phenomena under study. Statistical inference performed in these models can therefore provide more realistic approximations of reality. Introductions, reviews, and applications of these frameworks are provided by Vaart [1998], Tsiatis [2006], Bickel and Doksum [2015], van der Laan and Rose [2011b, 2018b], among others.

There are myriad applications of semi- and nonparametric procedures in the biomedical sciences. Much work relating to HIV testing, treatment efficacy, treatment adherence and patient outcomes have relied on these frameworks [Petersen et al., 2007, 2008, Tsai et al., 2010, Geng et al., 2010, Chamie et al., 2016, Petersen et al., 2017, Havlir et al., 2019, , for example]. International and federal health regulatory bodies have recently encouraged the

use of covariate adjustment methods for efficiently assessing therapies' treatment benefits in clinical trials [for Medicinal Products for Human Use, 2023, for Drug Evaluation and Research, 2023], like those proposed and discussed by Rosenblum and van der Laan [2010], Moore and van der Laan [2011], Vermeulen et al. [2015], Díaz et al. [2016], Díaz et al. [2019], Benkeser et al. [2021a], Li et al. [2023], to name but a few. Elucidation COVID-19 vaccines' mechanisms of action have been supported by mediation analyses [Gilbert et al., 2022, Fong et al., 2022, 2023] that are only possible using these flexible approaches. [Benkeser et al., 2021b, Hejazi et al., 2021]. Again, the complexity of these quantitative questions bars the use of simpler parametric-model-based procedures.

Little work has been done, unfortunately, to extend these methodologies for their application to high-dimensional data, data for which the number of features is—at minimum—of the same magnitude as the number of observations. Such data are increasingly generated or collected as part of modern biomedical experiments and studies. Though the aforementioned model mispecification issues remain, practitioners have generally turned to simplifying assumptions and accompanying parametric methods to circumvent or lessen the difficulties associated with high-dimensional data analyses.

A (perceived) challenge driving the adoption of these untenable approaches is the belief that only simple patterns can be reliably estimated from these data. See the "bet on sparsity" by Hastie et al. [2009, Chap. 16]. While the simplicity afforded by penalized parametric methods might be justified in some situations, this precept dissuades researchers from exploring alternative, more appropriate approaches. Assumptions incorporated into statistical models should reflect substantive knowledge about the DGP, not the need to simplify downstream analytical tasks. Again, inference performed with methods depending on unreasonable assumptions are at best unreliable and at worst misleading.

This thesis contains a selection of works demonstrating that nonparametric methods can be developed for high-dimensional data analyses in the biomedical sciences, that they have favorable theoretical properties, and that their application permit rigorous, trustworthy statistical inference.

Motivated by the widespread dependence of statistical methods regularly used in computational biology on the covariance matrix, we propose a general estimator selection procedure of this parameter in Chapter 2. Broadly, the covariance matrix plays a fundamental role in many modern exploratory and inferential procedures, including dimensionality reduction, hypothesis testing, and regression. In low-dimensional regimes, where the number of observations far exceeds the number of variables, the optimality of the sample covariance matrix as an estimator of this parameter is well-established. High-dimensional regimes do not admit such a convenience. As such, a variety of estimators have been derived to overcome the shortcomings of the sample covariance matrix in these settings. The question of selecting an optimal estimator from among the plethora available remains largely unaddressed, however. Using the framework of cross-validated loss-based estimation, we develop the theoretical underpinnings of just such an estimator selection procedure. In particular, we propose a general class of loss functions for covariance matrix estimation and establish finite-sample risk bounds and conditions for the asymptotic optimality of the cross-validated estimator

selector with respect to these loss functions. We evaluate our proposed approach via a comprehensive set of simulation experiments and demonstrate its practical benefits by its use in the exploratory analysis of two single-cell transcriptome sequencing datasets. The contents of this chapter have been published in Boileau et al. [2021b].

In Chapter 3, we present a method for reliably uncovering predictive biomarkers from high-dimensional clinical trial data. Predictive biomarker discovery is an endeavor central to precision medicine; they define patient sub-populations which stand to benefit most, or least, from a given therapy. The identification of these biomarkers is often the byproduct of the related but fundamentally different task of treatment rule estimation. Using treatment rule estimation methods to identify predictive biomarkers in clinical trials where the number of covariates exceeds the number of participants often results in high false discovery rates. The higher than expected number of false positives translates to wasted resources when conducting follow-up experiments for drug target identification and diagnostic assay development. Patient outcomes are in turn negatively affected. We propose a variable importance parameter for directly assessing the importance of potentially predictive biomarkers, and develop a flexible nonparametric inference procedure for this estimand. We prove that our estimator is double-robust and asymptotically linear under loose conditions on the DGP, permitting valid inference about the importance metric. The statistical guarantees of the method are verified in a thorough simulation study representative of randomized control trials with moderate and high-dimensional covariate vectors. Our procedure is then used to discover predictive biomarkers from among the tumor gene expression data of metastatic renal cell carcinoma patients enrolled in recently completed clinical trials. We find that our approach more readily discerns predictive from non-predictive biomarkers than procedures whose primary purpose is treatment rule estimation. This chapter has been published in Boileau et al. [2022].

Chapter 4 expands on the method proposed in Chapter 3, providing a general nonparametric inference framework for treatment effect modifier discovery in high dimensions. Heterogeneous treatment effects are driven by treatment effect modifiers, pre-treatment covariates that modify the effect of a treatment on an outcome—like predictive biomarkers. Similarly to procedures used for the more pointed goal of predictive biomarker discovery, current approaches for uncovering treatment effect modifiers are limited to low-dimensional data, data with weakly correlated covariates, or data generated according to specific parametric processes. We resolve these issues by developing a framework for defining model-agnostic treatment effect modifier variable importance parameters applicable to high-dimensional data with arbitrary correlation structure, deriving one-step, estimating equation and targeted maximum likelihood estimators of these parameters, and establishing these estimators' asymptotic properties. This framework is showcased by defining variable importance parameters for DGP with continuous, binary, and time-to-event outcomes with binary treatments, and deriving accompanying multiply-robust and asymptotically linear estimators. Simulation experiments demonstrate that these estimators' asymptotic guarantees are approximately achieved in realistic sample sizes for observational and randomized studies alike. This methodology is applied to gene expression data collected for a clinical trial assessing the effect of a monoclonal antibody therapy on disease-free survival in breast cancer pa-

tients. Genes predicted to have the greatest potential for treatment effect modification have previously been linked to treatment-resistant breast cancer.

Chapter 2

Cross-Validated Covariance Matrix Estimator Selection

2.1 Introduction

The covariance matrix underlies numerous exploratory and inferential statistical procedures central to analyses regularly performed in diverse fields. For instance, in computational biology, this statistical parameter serves as a key ingredient in many popular dimensionality reduction, clustering, and classification methods which are regularly relied upon in quality control assessments, exploratory data analysis, and, recently, the discovery and characterization of novel types of cells. Other important areas in which the covariance matrix is critical include financial economics, climate modeling and weather forecasting, imaging data processing and analysis, probabilistic graphical modeling, and text corpora compression and retrieval. Even more fundamentally, the covariance matrix plays a key role in assessing the strengths of linear relationships within multivariate data, in generating simultaneous confidence bands and regions, and in the construction and evaluation of hypothesis tests. Accurate estimation of this parameter is therefore essential.

When the number of observations in a data set far exceeds the number of features, the estimator of choice for the covariance matrix is the sample covariance matrix: it is an efficient estimator under minimal regularity assumptions on the data-generating distribution [Anderson, 2003, Smith, 2005]. In high-dimensional regimes, however, this simple estimator has undesirable properties. When the number of features outstrips the number of observations, the sample covariance matrix is singular. Even when the number of observations slightly exceeds the number of features, the sample covariance matrix is numerically unstable on account of an overly large condition number [Golub and Van Loan, 1996]. Its eigenvalues are also generally over-dispersed when compared to those of the population covariance matrix [Johnstone, 2001, Ledoit and Wolf, 2004]: the leading eigenvalues are positively biased, while the trailing eigenvalues are negatively biased [Marčenko and Pastur, 1967].

High-dimensional data have become increasingly widespread in myriad scientific domains,

with many examples arising from challenges posed by cutting-edge biological sequencing technologies. Accordingly, researchers have turned to developing novel covariance matrix estimators to remediate the deficiencies of the sample covariance matrix. Under certain sparsity assumptions, Bickel and Levina [2008b,c], Rothman et al. [2009], Lam and Fan [2009], Cai et al. [2010b], and Cai and Liu [2011], among others, demonstrated that the true covariance matrix can be estimated consistently under specific losses by applying element-wise thresholding or tapering functions to the sample covariance matrix. Another thread of the literature, which includes notable contributions by Stock and Watson [2002], Bai [2003], Fan et al. [2008, 2013, 2016b, 2019], and Onatski [2012], has championed methods employing factor models in covariance matrix estimation. Other popular proposals include the families of estimators inspired by the empirical Bayes framework [Robbins, 1964, Efron, 2012], formulated by Schäfer and Strimmer [2005] and Ledoit and Wolf [2004, 2012, 2015, 2018].

Despite the flexibility afforded by the apparent wealth of candidate estimators, this variety poses many practical issues. Namely, identifying the most appropriate estimator from among a collection of candidates is itself a significant challenge. A partial answer to this problem has come in the form of data-adaptive approaches designed to select the optimal estimator within a particular class [for example, Bickel and Levina, 2008b,c, Cai and Liu, 2011, Fan et al., 2013, Fang et al., 2016, Bartz, 2016]. Such approaches, however, are inherently limited by their focus on relatively narrow families of covariance matrix estimators. The successful application of such estimator selection frameworks requires, as a preliminary step, that the practitioner make a successful choice among estimator families, injecting a degree of subjectivity in their deployment. The broader question of selecting an optimal estimator from among a diverse library of candidates has remained unaddressed. We offer a general loss-based framework building upon the seminal work of van der Laan and Dudoit [2003b], Dudoit and van der Laan [2005], van der Vaart et al. [2006] for asymptotically optimal covariance matrix estimator selection based upon cross-validation.

In the cross-validated loss-based estimation framework, the parameter of interest is defined as the risk minimizer, with respect to the data-generating distribution, based on a loss function chosen to reflect the problem at hand. Candidate estimators may be generated in a variety of manners, including as empirical risk minimizers with respect to an empirical distribution over parameter subspaces corresponding to models for the data-generating distribution. One would ideally select as optimal estimator that which minimizes the "true" risk with respect to the data-generating distribution. However, as this distribution is typically unknown, one turns to cross-validation for risk estimation. van der Laan and Dudoit [2003b], Dudoit and van der Laan [2005], van der Vaart et al. [2006] have shown that, under general conditions on the data-generating distribution and loss function, the cross-validated estimator selector is asymptotically optimal in the sense that it performs asymptotically as well in terms of risk as an optimal oracle selector based on the true, unknown data-generating distribution. These results extend prior work summarized by Györfi et al. [2002, Ch. 7–8].

Focusing specifically on the covariance matrix as the parameter of interest, we address the choice of loss function and candidate estimators, and derive new, high-dimensional

asymptotic optimality results for multivariate cross-validated estimator selection procedures. Requiring generally nonrestrictive assumptions about the structure of the true covariance matrix, the proposed framework accommodates arbitrary families of covariance matrix estimators. The method therefore enables the objective selection of an optimal estimator while fully taking advantage of the plethora of available estimators. As such, it generalizes existing, but more limited, data-adaptive estimator selection frameworks where the library of candidate estimators is narrowed based on available subject matter knowledge, or, as is more commonly the case, for convenience's sake.

The remainder of the chapter is organized as follows. In Section 2.2, we formulate the problem within the general framework of cross-validated loss-based estimation. In Section 2.3, we propose a general class of loss functions for covariance matrix estimation and establish finite-sample risk bounds and conditions for the asymptotic optimality of the cross-validated estimator selector with respect to these loss functions. We briefly review and detail several of covariance matrix estimator families in Section 2.4. Section 2.5 provides a thorough interrogation of the proposed cross-validated estimator selector in a series of simulation experiments. The practical benefits of using the cross-validated estimator selection strategy are then demonstrated by its application in a dimensionality reduction pipeline for novel cell-type identification with two distinct single-cell RNA-seq datasets in Section 2.6. We conclude the chapter with a brief discussion in Section 2.7. All proofs are relegated to Section 2.8 for expository clarity.

2.2 Problem Formulation and Background

Consider a data set $X_{n\times J} = \{X_1, \ldots, X_n : X_i \in \mathbb{R}^J\}$, comprising n independent and identically distributed (i.i.d.) random vectors, where $n \approx J$ or $n < J$. Let $X_i \sim P_0 \in \mathcal{M}$, where P_0 denotes the true data-generating distribution and $\mathcal{M}$ the statistical model, that is, a collection of possible data-generating distributions P for X_i. We assume a nonparametric statistical model $\mathcal{M}$ for P_0, minimizing assumptions on the form of P_0. We denote by P_n the empirical distribution of the n random vectors forming $X_{n\times J}$. Letting $\mathbb{E}[X_i] = 0$ without loss of generality and defining $\psi_0 \equiv \mathrm{Var}[X_i]$, we take as our goal the estimation of the covariance matrix ψ_0. This is accomplished by identifying the "optimal" estimator of ψ_0 from among a collection of candidates, where, as detailed below, optimality is defined in terms of risk.

For any distribution $P \in \mathcal{M}$, define its covariance matrix as $\psi = \Psi(P)$, where Ψ is a mapping from the model $\mathcal{M}$ to the set of $J \times J$ symmetric, positive semi-definite matrices. Furthermore, candidate estimators of the covariance matrix are defined as $\hat{\psi}_k \equiv \hat{\Psi}_k(P_n)$ for $k = 1, \ldots, K$ in terms of mappings $\hat{\Psi}_k$ from the empirical distribution P_n to $\Psi \equiv \{\psi \in \mathbb{R}^{J\times J} | \psi = \psi^\top\}$. While this notation suggests that the number of candidate estimators K is fixed, and we treat it as such throughout, this framework may be extended such that K grows as a function of n and J. It also follows that $\{\psi = \Psi(P) : P \in \mathcal{M}\} \subset \Psi$; that is, the set of all covariance matrices corresponding to the data-generating distributions P belonging to the model $\mathcal{M}$ is a subset of Ψ.

In order to assess the optimality of estimators in the set Ψ, we introduce a generic loss function $L(X;\psi,\eta)$ characterizing a cost applicable to any $\psi \in \Psi$ and $X \sim P \in \mathcal{M}$, and where η is a (possibly empty) nuisance parameter. Specific examples of loss functions for the covariance estimation setting are proposed in Section 2.3. Define H as the mapping from the model $\mathcal{M}$ to the nuisance parameter space $H \equiv \{\eta = H(P) : P \in \mathcal{M}\}$ and let $\hat{\eta} \equiv \hat{H}(P_n)$ denote a generic nuisance parameter estimator, where $\hat{H}$ is a mapping from P_n to H. Given any $\eta \in H$, the risk under $P \in \mathcal{M}$ for any $\psi \in \Psi$ is defined as the expected value of $L(X;\psi,\eta)$ with respect to P:

$$\begin{aligned}\Theta(\psi,\eta,P) &\equiv \int L(x;\psi,\eta)dP(x) \\ &= \mathbb{E}_P\left[L(X;\psi,\eta)\right].\end{aligned}$$

Under the additional constraint on the loss function that a risk minimizer exists under the true data-generating distribution P_0, the minimizer is given by the parameter of interest

$$\psi_0 \equiv \underset{\psi\in\Psi}{\operatorname{argmin}}\, \Theta(\psi,\eta_0,P_0), \tag{2.1}$$

where $\eta_0 \equiv H(P_0)$. The risk minimizer need not be unique. The optimal risk under P_0 is

$$\theta_0 \equiv \min_{\psi\in\Psi}\ \Theta(\psi,\eta_0,P_0),$$

which is to say that a risk minimizer ψ_0 attains risk θ_0.

For any given estimator $\hat{\psi}_k$ of ψ_0, its conditional risk given P_n with respect to the true data-generating distribution P_0 is

$$\begin{aligned}\tilde{\theta}_n(k,\eta_0) &\equiv \mathbb{E}_{P_0}[L(X;\hat{\Psi}_k(P_n),\eta_0) \mid P_n] \\ &= \Theta(\hat{\psi}_k,\eta_0,P_0).\end{aligned}$$

Defining the risk difference of the k^{th} estimator as $\tilde{\theta}_n(k,\eta_0) - \theta_0$, the index of the estimator that achieves the minimal risk difference is

$$\tilde{k}_n \equiv \underset{k\in\{1,\ldots,K\}}{\operatorname{argmin}}\ \tilde{\theta}_n(k,\eta_0) - \theta_0.$$

The subscript n emphasizes that the risk and optimal estimator index are conditional on the empirical distribution P_n. They are therefore random variables.

Given the high-dimensional nature of the data, it is generally most convenient to study the performance of estimators of ψ_0 using *Kolmogorov asymptotics*, that is, in the setting in which both $n \to \infty$ and $J \to \infty$ such that $J/n \to m < \infty$. Historically, estimators have been derived within this high-dimensional asymptotic regime to improve upon the finite sample results of estimators brought about by traditional asymptotic arguments. After all,

the sample covariance matrix retains its asymptotic optimality properties when J is fixed, even though it is known to perform poorly in high-dimensional settings.

Naturally, it would be desirable for an estimator selection procedure to select the estimator indexed by $\tilde{k}_n$; however, this quantity depends on the true, unknown data-generating distribution P_0. As a substitute for the candidates' true conditional risks, we employ instead the cross-validated estimators of these same conditional risks.

Cross-validation (CV) consists of randomly, and possibly repeatedly, partitioning a data set into a training set and a validation set. The observations in the training set are fed to the candidate estimators and the observations in the validation set are used to evaluate the performance of these estimators [Breiman and Spector, 1992, Friedman et al., 2001]. A range of CV schemes have been proposed and investigated, both theoretically and computationally; Dudoit and van der Laan [2005] provide a thorough review of popular CV schemes and their properties. Among the variety, V-fold stands out as an approach that has gained traction on account of its relative computational feasibility and good performance. Any CV scheme can be expressed in terms of a binary random vector B_n, which assigns observations into either the training or validation set. Observation i is said to lie in the training set when $B_n(i) = 0$ and in the validation set otherwise. The training set therefore contains $\sum_i (1 - B_n(i)) = n(1 - p_n)$ observations and the validation set $\sum_i B_n(i) = np_n$ observations, where p_n is the fixed validation set proportion corresponding to the chosen CV procedure. The empirical distributions of the training and validation sets are denoted by P^0_{n,B_n} and P^1_{n,B_n}, respectively, for any given realization of B_n. B_n, we emphasize, is independent of P_n.

Using this general definition, the cross-validated estimator of a candidate $\hat{\Psi}_k$'s risk is

$$\begin{aligned}\hat{\theta}_{p_n,n}(k, \hat{H}(P^0_{n,B_n})) &\equiv \mathbb{E}_{B_n}\left[\Theta(\hat{\Psi}_k(P^0_{n,B_n}), \hat{H}(P^0_{n,B_n}), P^1_{n,B_n})\right] \\ &= \mathbb{E}_{B_n}\left[\frac{1}{np_n}\sum_{i=1}^{n}\mathbb{I}(B_n(i) = 1)L(X_i; \hat{\Psi}_k(P^0_{n,B_n}), \hat{H}(P^0_{n,B_n}))\right],\end{aligned}$$

for a nuisance parameter estimator mapping $\hat{H}$. Here, $\mathbb{E}_{B_n}[\cdot]$ denotes the expectation with respect to B_n. The corresponding cross-validated selector is

$$\hat{k}_{p_n,n} \equiv \underset{k\in\{1,\ldots,K\}}{\operatorname{argmin}}\ \hat{\theta}_{p_n,n}(k, \hat{H}(P^0_{n,B_n})).$$

As a benchmark, the unknown cross-validated conditional risk of the k^{th} estimator is

$$\tilde{\theta}_{p_n,n}(k, \eta_0) \equiv \mathbb{E}_{B_n}\left[\Theta(\hat{\Psi}_k(P^0_{n,B_n}), \eta_0, P_0)\right].$$

The cross-validated oracle selector is then

$$\tilde{k}_{p_n,n} \equiv \underset{k\in\{1,\ldots,K\}}{\operatorname{argmin}}\ \tilde{\theta}_{p_n,n}(k, \eta_0).$$

As before, the p_n and n subscripts highlight the dependence of these objects on the CV procedure and the empirical distribution P_n, respectively, thus making them random variables.

Ideally, the cross-validated estimator selection procedure should identify a $\hat{k}_{p_n,n}$ that is asymptotically (in n, J, and possibly K) equivalent in terms of risk to the oracle $\tilde{k}_{p_n,n}$, under a set of nonrestrictive assumptions based on the choice of loss function, target parameter space, estimator ranges, and, if applicable, nuisance parameter space, in the sense that

$$\frac{\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}, \eta_0) - \theta_0}{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}, \eta_0) - \theta_0} \xrightarrow{P} 1 \text{ as } n, J \to \infty. \tag{2.2}$$

That is, the estimator selected via CV is equivalent in terms of risk to the CV scheme's oracle estimator chosen from among all candidates.

When Equation (2.2) holds, a further step may be taken by relating the performance of the cross-validated selector to that of the full-dataset oracle selector, $\tilde{k}_n$:

$$\frac{\tilde{\theta}_n(\hat{k}_{p_n,n}, \eta_0) - \theta_0}{\tilde{\theta}_n(\tilde{k}_n, \eta_0) - \theta_0} \xrightarrow{P} 1 \text{ as } n, J \to \infty. \tag{2.3}$$

When the cross-validated selection procedure's full-dataset conditional risk difference converges in probability to that of the full-dataset oracle's, the chosen estimator is *asymptotically optimal*. In other words, the data-adaptively selected estimator performs asymptotically as well, with respect to the chosen loss, as the candidate that would be picked from the collection of estimators if the true data-generating distribution were known.

2.3 Loss Functions and Estimator Selection

Proposed Loss Function

The choice of loss function should reflect the goals of the estimation task. While loss functions based on the sample covariance matrix and either the Frobenius or the spectral norms are often employed in the covariance matrix estimation literature, Dudoit and van der Laan [2005]'s estimator selection framework is more amenable to loss functions that operate over random vectors. Accordingly, we propose the observation-level Frobenius loss:

$$\begin{aligned} L(X; \psi, \eta_0) &\equiv \|XX^\top - \psi\|^2_{F,\eta_0} \\ &= \sum_{j=1}^{J} \sum_{l=1}^{J} \eta_0^{(jl)} (X^{(j)} X^{(l)} - \psi^{(jl)})^2, \end{aligned} \tag{2.4}$$

where $X^{(j)}$ is the j^{th} element of a random vector $X \sim P \in \mathcal{M}$, $\psi^{(jl)}$ is the entry in the j^{th} row and l^{th} column of an arbitrary covariance matrix $\psi \in \Psi$, and η_0 is a $J \times J$ matrix acting as a scaling factor, that is, a potential nuisance parameter. For an estimator $\hat{\eta}$ of η_0, the cross-validated risk estimator of the k^{th} candidate estimator $\hat{\Psi}_k$ under the observation-level Frobenius loss is

$$\hat{\theta}_{p_n,n}(k, \hat{H}(P^0_{n,B_n})) = \mathbb{E}_{B_n} \left[\frac{1}{np_n} \sum_{i=1}^{n} \mathbb{I}(B_n(i) = 1) \|X_i X_i^\top - \hat{\Psi}_k(P^0_{n,B_n})\|^2_{F,\hat{H}(P^0_{n,B_n})} \right].$$

Ledoit and Wolf [2004], Bickel and Levina [2008c], and Rothman et al. [2009], among others, have employed analogous (scaled) Frobenius losses to prove various optimality results, defining $\eta_0^{(jl)} = 1/J$, $\forall j, l$. This particular choice of scaling factor is such that whatever the value of J, $\|I_{J\times J}\|_{F,\eta_0} = 1$. With such a scaling factor, the loss function may be viewed as a relative loss whose yardstick is the $J \times J$ identity matrix. A similarly reasonable option for when the true covariance matrix is assumed to be dense is $\eta_0^{(jl)} = 1/J^2$. This weighting scheme effectively computes the average squared error across every entry of the covariance matrix; however, when the scaling factor is constant, it only impacts the interpretation of the loss. Constant scaling factors have no impact on our asymptotic analysis. Since it need not be estimated, it is not a nuisance parameter in the conventional sense.

When the scaling factor of Equation (2.4) is constant, the risk minimizers are identical for the cross-validated observation-level Frobenius risk and the common cross-validated Frobenius risk [used by, for example, Bickel and Levina, 2008c, Rothman et al., 2009, Fan et al., 2013, Fang et al., 2016].

Proposition 2.1. *Define the cross-validated Frobenius risk for an estimator* $\hat{\Psi}_k$ *as*

$$\hat{R}_n(\hat{\Psi}_k, \eta_0) \equiv \mathbb{E}_{B_n}\left[\|S_n(P^1_{n,B_n}) - \hat{\Psi}_k(P^0_{n,B_n})\|^2_{F,\eta_0}\right], \tag{2.5}$$

where $S_n(P^1_{n,B_n})$ *is the sample covariance matrix computed over the validation set* P^1_{n,B_n}*, and* η_0 *is some constant scaling matrix. Then,* $\hat{R}_n(\hat{\Psi}_k, \eta_0) - \hat{\theta}_{p_n,n}(k, \eta_0)$ *is constant with respect to* $\hat{\Psi}_k(P^0_{n,B_n})$ *such that*

$$\begin{aligned}\hat{k}_{p_n,n} &= \underset{k\in\{1,\ldots,K\}}{\operatorname{argmin}}\ \hat{\theta}_{p_n,n}(k, \eta_0) \\ &= \underset{k\in\{1,\ldots,K\}}{\operatorname{argmin}}\ \hat{R}_n(\hat{\Psi}_k, \eta_0).\end{aligned}$$

Note that the traditional Frobenius loss corresponds to the sum of the squared eigenvalues of the difference between the sample covariance matrix and the estimate. Proposition 2.1 therefore implies the existence of a similar relationship for our observation-level Frobenius loss. It may therefore serve as a surrogate for a loss based on the spectral norm.

We are not restricted to a constant scaling factor matrix. One might consider weighting the covariance matrix's off-diagonal elements' errors by their corresponding diagonal entries, especially useful when the random variables are of different scales. Such a scaling factor might offer a more equitable evaluation across all entries of the parameter:

$$L_{\text{weighted}}(X; \psi, \eta_0) \equiv \sum_{j=1}^{J}\sum_{l=1}^{J} \frac{1}{\sqrt{\psi_0^{(jj)}\psi_0^{(ll)}}}(X^{(j)}X^{(l)} - \psi^{(jl)})^2.$$

Here, $\eta_0 = \text{diag}(\psi_0)$ is a genuine nuisance parameter which can be estimated via the diagonal entries of the sample covariance matrix.

Finally, the covariance matrix ψ_0 is the risk minimizer of the observation-level Frobenius loss if the integral with respect to X and the partial differential operators with respect to ψ are interchangeable.

Proposition 2.2. *Let the integral with respect to X and the partial differential operators with respect to ψ be interchangeable, and let η be some fixed $J \times J$ matrix. Then*

$$\psi_0 = \underset{\psi \in \Psi}{\operatorname{argmin}} \ \Theta(\psi, \eta, P_0)$$

for $\Theta(\cdot)$ defined under the observation-level Frobenius loss.

Our proposed loss therefore satisfies the condition of Equation (2.1). The main results of the paper, however, relate only to the constant scaling factor case. In a minor abuse of notation, we set $\eta_0 = 1$, and suppress dependence of the loss function on the scaling factor wherever possible throughout the remainder of the chapter.

Optimality of the Cross-validated Estimator Selector

Having defined a suitable loss function, we turn to a discussion of the theoretical properties of the cross-validated estimator selection procedure. Specifically, we present, in Theorem 2.1, sufficient conditions under which the method is asymptotically equivalent in terms of risk to the commensurate CV oracle selector (as per Equation (2.2)). This theorem extends the general framework of Dudoit and van der Laan [2005] for use in high-dimensional multivariate estimator selection. Adapting their existing theory to this setting requires a judicious choice of loss function, new assumptions, and updated proofs reflecting the use of high-dimensional asymptotics. Corollary 2.1 then builds on Theorem 2.1 and details conditions under which the procedure produces asymptotically optimal selections in the sense of Equation (2.3). Again, all proofs are provided in Section 2.8.

Theorem 2.1. *Let $X_1, \ldots, X_n$ be a random sample of n i.i.d. random vectors of dimension J, such that $X_i \sim P_0 \in \mathcal{M}, i = 1, \ldots, n$. Assume, without loss of generality, that $\mathbb{E}[X_i] = 0$, and define $\psi_0 \equiv Var[X_i]$. Denote the set of K candidate estimators by $\{\hat{\Psi}_k(\cdot) : k = 1, \ldots, K\}$. Next, define the observation-level Frobenius loss function as $L(X; \psi) \equiv \|X^\top X - \psi\|^2_{F,1}$. Finally, designate p_n as the proportion of observations in any given cross-validated validation set. Consider the following assumptions:*

Assumption 2.1. *For any $P \in \mathcal{M}$ and $X \sim P$, $\max_{j=1,\ldots,J} (|X^{(j)}|) < \sqrt{M_1} < \infty$ almost surely (a.s.).*

Assumption 2.2. *Define $\Psi \equiv \{\psi \in \mathbb{R}^{J \times J} |\ \psi = \psi^\top, |\psi^{(jl)}| < M_2 < \infty, \ \forall\ j, l = 1, \ldots, J\}$, and assume that $\hat{\Psi}_k(P_n), \psi_0 \in \Psi$.*

Finite-Sample Result. *Let* $\overline{M}(J) \equiv 4(M_1+M_2)^2J^2$ *and* $c(\delta, \overline{M}(J)) \equiv 2(1+\delta)^2\overline{M}(J)(1/3+1/\delta)$. *Then, for any* $\delta > 0$,

$$0 \leq \mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta_0] \leq (1+2\delta)\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0] + 2c(\delta, \overline{M}(J))\frac{1+\log(K)}{np_n}. \quad (2.6)$$

High-Dimensional Asymptotic Result. *The finite-sample result in Equation* (2.6) *has the following asymptotic implications: If* $c(\delta, \overline{M}(J))(1+\log(K))/(np_n\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})-\theta_0]) \to 0$ *and* $J/n \to m < \infty$ *as* $n, J \to \infty$, *then*

$$\frac{\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta_0]}{\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0]} \to 1. \quad (2.7)$$

Further, if $c(\delta, \overline{M}(J))(1+\log(K))/(np_n(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0)) \xrightarrow{P} 0$ *as* $n, J \to \infty$, *then*

$$\frac{\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta_0}{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0} \xrightarrow{P} 1. \quad (2.8)$$

The proof relies on special properties of the random variable $Z_k \equiv L(X; \hat{\Psi}_k(P_n)) - L(X; \psi_0)$ and on an application of Bernstein's inequality [Bennett, 1962]. Together, they are used to show that $2c(\delta, \overline{M}(J))(1+\log(K))/(np_n)$ is a finite-sample bound for comparing the performance of the cross-validated selector, $\hat{k}_{p_n,n}$, against that of the oracle selector over the training sets, $\tilde{k}_{p_n,n}$. Once this bound is established, the high-dimensional asymptotic results follow immediately.

Only a few sufficient conditions are required to provide finite-sample bounds on the expected risk difference of the estimator selected via our CV procedure. First, that each element of the random vector X be bounded, and, second, that the entries of all covariance matrices in the parameter space and the set of possible estimates be bounded. Together, these assumptions allow for the definition of $\overline{M}(J)$, the object permitting the extension of the loss-based estimation framework to the high-dimensional covariance matrix estimation problem.

The first assumption is technical in nature — it makes the proofs tractable. While it may appear stringent, and, for instance, is not satisfied by Gaussian distributions, we believe it to be generally applicable. We stress that parametric data-generating distributions, like those exhibiting Gaussianity, rarely reflect reality, that is, they are merely mathematical conveniences[1] Most random variables, or transformations thereof, arising in scientific practice are bounded by limitations of the physical, electronic, or biological measurement process; thus, our method remains widely applicable. For example, in microarray and next-generation sequencing experiments, the raw data are images on a 16-bit scale, constraining them to

[1]Anecdotally, one cannot help but find themself reminded that "Everyone is sure of this [that errors are normally distributed] ... since the experimentalists believe that it is a mathematical theorem, and the mathematicians that it is an experimentally determined fact." [Poincaré, 1912, p. 171].

$[0, 2^{16})$. Similarly, the measurement of immunologic markers, of substantial interest in vaccine efficacy trials of HIV-1, COVID-19, and other infectious diseases, are bounded by the limits of detection and/or quantitation imposed by assay biotechnology.

Verifying that the additional assumptions required by Theorem 2.1's asymptotic results hold proves to be more challenging. We write $f(y) = O(g(y))$ if $|f|$ is bounded above by g, $f(y) = o(g(y))$ if f is dominated by g, $f(y) = \Omega(g(y))$ if f is bounded below by g, and $f(y) = \omega(g(y))$ if f dominates g, all in asymptotics with respect to n and J. Further, a subscript "P" might be added to these bounds, denoting a convergence in probability. Now, note that $c(\delta, \overline{M}(J))(1 + \log(K))/(np_n) = O(J)$ for fixed p_n and as $J/n \to m > 0$. Then the conditions associated with Equation (2.7) and Equation (2.8) hold so long as $\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0] = \omega(J)$ and $\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0 = \omega_P(J)$, respectively.

These requirements do not seem particularly restrictive given that the complexity of the problem generally increases as a function of the number of features. There are many more entries in the covariance matrix requiring estimation than there are observations. This intuition is corroborated by our extensive simulation study in the following section. Consistent estimation in terms of the Frobenius risk is therefore not possible in high-dimensions without additional assumptions about the data-generating process.

Some additional insight might be gained by identifying conditions under which these assumptions are *unmet* for popular structural beliefs about the true covariance matrix. In particular, we consider the sparse covariance matrices defined in Bickel and Levina [2008c] and accompanying hard-thresholding estimators (see Section 2.4):

Proposition 2.3. *In addition to Assumptions 2.1 and 2.2 of Theorem 2.1, assume that ψ_0 is a member of the following set of matrices:*

$$\left\{ \psi : \psi^{(jj)} < M_2, \sum_{l=1}^{J} I(\psi^{(jl)} \neq 0) < s(J) \text{ for all }, j = 1, \ldots, J \right\}$$

where $s(J)$ is the row sparsity, that the hard-thresholding estimator is in the library of candidates, and that it uses a "sufficiently large" thresholding hyperparameter value in the sense of Bickel and Levina [2008c]. Then, by Theorem 2 of Bickel and Levina [2008c], we have $\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0] = o(J)$ if $s(J)/J = o(1/\log J)$ asymptotically in n and J.

Proposition 2.3 states that the conditions for achieving the asymptotic results of Theorem 2.1 are not met if the proportion of non-zero elements in the covariance matrix's row with the most non-zero elements converges to zero faster than $1/\log J$ and the library of candidates possesses a hard-thresholding estimator whose thresholding hyperparameter is reasonable in the sense of Bickel and Levina [2008c]'s Theorem 2 and its subsequent discussion. Plainly, the true covariance matrix cannot be too sparse if the collection of considered estimators contains the hard-thresholding estimator with a correctly specified thresholding hyperparameter value.

This implies that banded covariance matrices whose number of bands are fixed for J do not meet the criteria for our theory to apply, assuming that one of the candidate estimators

correctly specifies the number of bands. Nevertheless, we observe empirically in Section 2.5 that our cross-validated procedure selects an optimal estimator when the true covariance matrix is banded or tapered more quickly in terms of n and J than any other type of true covariance matrix.

These results are likely explained by the relatively low complexity of the estimation problem in this setting. High-dimensional asymptotic arguments are perhaps unnecessary when the proportion of entries needing to be estimated in the true covariance matrix quickly converges to zero. These limitations of our theory reflect stringent, and typically unverifiable, structural assumptions about the estimand. We reiterate that the conditions of Theorem 2.1 are generally satisfied. In situations where the true covariance matrix is known to possess this level sparsity, practitioners might instead appeal to Equation (39) of Bickel and Levina [2008c] to support their use of a cross-validated estimator selection procedure. This result, coupled with that of Proposition 2.1, likely explains the aforementioned simulation findings of the banded and tapered covariance matrices.

Now, Theorem 2.1's high-dimensional asymptotic results relate the performance of the cross-validated selector to that of the oracle selector for the CV scheme. As indicated by the expression in Equation (2.3), however, we would like our cross-validated procedure to be asymptotically equivalent to the oracle over the *entire* data set. The conditions to obtain this desired result are provided in Corollary 2.1, a minor adaptation of previous work by Dudoit and van der Laan [2005]. This extension accounts for increasing J, thereby permitting its use in high-dimensional asymptotics.

Corollary 2.1. *Building upon Assumptions 2.1 and 2.2 of Theorem 2.1, we introduce the additional assumptions that, as* $n, J \to \infty$ *and* $J/n \to m < \infty$, $p_n \to 0$, $c(\delta, \overline{M}(J))(1 + \log(K))/(np_n(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0)) \xrightarrow{P} 0$, *and*

$$\frac{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0}{\tilde{\theta}_n(\tilde{k}_n) - \theta_0} \xrightarrow{P} 1. \tag{2.9}$$

Under these assumptions, it follows that

$$\frac{\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta_0}{\tilde{\theta}_n(\tilde{k}_n) - \theta_0} \xrightarrow{P} 1. \tag{2.10}$$

The proof is a direct application of the asymptotic results of Theorem 2.1.

As before, the assumption that $c(\delta, \overline{M}(J))(1 + \log(K))/(np_n(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0)) \xrightarrow{P} 0$ remains difficult to verify, but essentially requires the estimation error of the oracle to increase quickly as the number of features grows. That is, $np_n(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0) = \omega_P(J)$. We posit that this condition is generally satisfied, similarly to the asymptotic results of Theorem 2.1.

Now, a sufficient condition for Equation (2.9) is that there exists a $\gamma > 0$ such that

$$\Big(n^\gamma(\tilde{\theta}_n(\tilde{k}_n) - \theta_0),\ (n(1-p_n))^\gamma(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0)\Big) \xrightarrow{d} (Z, Z), \tag{2.11}$$

for a single random variable Z with $\mathbb{P}(Z > a) = 1$ for some $a > 0$. For single-split validation, where $\mathbb{P}(B_n = b) = 1$ for some $b \in \{0, 1\}^n$, it suffices to assume that there exists a $\gamma > 0$ such that $n^{\gamma}(\tilde{\theta}_n(\tilde{k}_n) - \theta_0) \xrightarrow{d} Z$ for a random variable Z with $\mathbb{P}(Z > a) = 1$ for some $a > 0$.

Equation (2.9) essentially requires that the (appropriately scaled) distributions of the cross-validated and full-dataset conditional risk differences of their respective oracle selections converge in distribution as $p_n \to 0$. Again, this condition is unrestrictive. As $p_n \to 0$, the composition of each training set becomes increasingly similar to that of the full-dataset. The resulting estimates produced by each candidate estimator over the training sets and the full-dataset will correspondingly converge. Naturally, so too will the cross-validated and full-dataset conditional risk difference distributions of their respective selections.

While the number of candidates in the estimator library K has been assumed to be fixed in this discussion of the proposed method's asymptotic results, it may be allowed to grow as a function of n and J. Of course, this will negatively impact the convergence rates of $c(\delta, \overline{M}(J))(1+\log(K))/(np_n\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})-\theta_0])$ and $c(\delta, \overline{M}(J))(1+\log(K))/(np_n(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})-\theta_0))$. The sufficient conditions outlined in the asymptotic results of Theorem 2.1 are achieved so long as the library of candidates does not grow too aggressively. That is, we can make the additional assumptions that $K = o(\exp\{\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0]/J\})$ and $K = o_P(\exp\{(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})-\theta_0)/J\})$ such that the results of Equations (2.7) and (2.8) are achieved, respectively.

Finally, we have assumed thus far that $\mathbb{E}_{P_0}[X]$ is known. This is generally not the case in practice. In place of a random vector centered at zero, we might instead consider the set of n demeaned random vectors $\tilde{X}_{n\times J}$ where $\tilde{X}_i = X_i - \bar{X}$ and $\bar{X}^{(j)} = 1/n \sum X_i^{(j)}$. It follows from the details given in Remark 2.1 that the asymptotic results of Theorem 2.1 and Corollary 2.1 apply to $\tilde{X}_{n\times J}$.

Remark 2.1. *We assume throughout this work that $\mathbb{E}_{P_0}[X] = 0$ without loss of generality. In practice, however, the mean vector is generally unknown. Consider the uniformly bounded random vector Y such that $X = Y - \mathbb{E}_{P_0}[Y]$. We might therefore consider using the demeaned random vector $\tilde{Y} = Y - \bar{Y}$ instead, where $\bar{Y}^{(j)} = 1/n \sum Y_i^{(j)}$. Employing $\tilde{Y}$ in place of X in Lemma 2.1, and denoting $\tilde{Z}_k \equiv L(\tilde{Y}; \hat{\Psi}_k(P^0_{n,B_n})) - L(\tilde{Y}; \psi_0)$, we find that $\mathbb{E}_{P_0}[\tilde{Z}_k | P^0_{n,B_n}, B_n] = \sum\sum(\psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))((n-2)\psi_0^{(jl)}/n - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))$. It then follows that, as $n \to \infty$, $\mathbb{E}_{P_0}[\tilde{Z}_k | P^0_{n,B_n}, B_n] = \mathbb{E}_{P_0}[Z_k | P^0_{n,B_n}, B_n]$. The asymptotic results of Theorem 2.1 and Corollary 2.1 are therefore achievable when $\mathbb{E}_{P_0}[Y]$ is unknown. The same cannot be said for the finite sample result of Theorem 2.1: $\mathbb{E}_{P_0}[\tilde{Z}_k | P^0_{n,B_n}, B_n]$ is not strictly nonegative. For large enough values of n, however, we do expect these finite bounds to be approximately correct.*

Software Implementation

The cross-validated covariance matrix estimator selection procedure is implemented in `cvCovEst` [Boileau et al., 2021a], an open-source `R` package [R Core Team, 2021]. This package is avail-

able via the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=cvCovEst`.

2.4 Candidate Covariance Matrix Estimators

Using the proposed cross-validated selection procedure effectively requires a large and diverse set of candidate covariance matrix estimators. In this spirit, we provide in the sequel a brief overview of select covariance matrix estimators that have proven to be effective in a variety of settings. Note, however, that the proposed selection framework need not be limited to those described here. Thorough reviews of estimators have been provided by Pourahmadi [2013], Fan et al. [2016a], and Ledoit and Wolf [2020], to name only a few.

Thresholding Estimators

An often natural, simplifying assumption about the true covariance matrix's structure is that it is sparse, that is, a non-negligible portion of its off-diagonal elements have a value near zero or equal to zero. This assumption is not altogether unreasonable: Given a system of numerous variables, it seems unlikely in many settings that a majority of these variables would depend on one another.

The class of generalized thresholding estimators [Bickel and Levina, 2008c, Rothman et al., 2009, Cai and Liu, 2011] is one collection of covariance matrix estimators based upon this structural assumption. Given the sample covariance matrix S_n, the generalized thresholding estimator is defined as

$$\hat{\Psi}_{\text{gt}}(P_n; t) \equiv \left\{ t\left(S_n^{(jl)}\right) : \ j, l = 1, \ldots, J \right\}, \tag{2.12}$$

where $t(\cdot)$ is a thresholding function that often requires one or more hyperparameters dictating the amount of regularization. The hard, soft, smoothly clipped absolute deviation (SCAD) [Fan and Li, 2001], and adaptive LASSO [Rothman et al., 2009] functions are among the collection of suitable thresholding operators. As an example, the hard-thresholding function is defined as $t_u(S_n^{(jl)}) \equiv S_n^{(jl)} I(S_n^{(jl)} > u)$ for some threshold $u > 0$. Cai and Liu [2011] also demonstrated that element-specific thresholding functions might be useful when the features' variances are highly variable. The hyperparameters for any specific thresholding function are often selected via CV. Regardless of the choice of $t(\cdot)$, these estimators preserve the symmetry of the sample covariance matrix and are invariant under permutations of the features' order.

Bickel and Levina [2008c] and Rothman et al. [2009] have shown that these estimators are consistent under the spectral norm (which is defined as a square matrix's largest absolute eigenvalue), assuming that $\log J/n \to 0$, that the observations' marginal distributions satisfy a tail condition, and that the true covariance matrix is a member of the class of matrices satisfying a particular notion of "approximate sparsity." Cai and Liu [2011] have derived

similar results for their entry-specific thresholding estimator over an even broader parameter space of sparse covariance matrices.

Banding and Tapering Estimators

A family of estimators related to thresholding estimators are banding and tapering estimators [Bickel and Levina, 2008b, Cai et al., 2010b]. Like thresholding estimators, these estimators rely on the assumption that the true covariance matrix is sparse. However, the structural assumption of these estimators on ψ_0 is much more rigid than that of thresholding estimators. Specifically, such estimators assume that the true covariance matrix is a *band matrix*, that is, a sparse matrix whose non-zero entries are concentrated about the diagonal. These estimators therefore require a natural ordering of the observations' features, operating under the assumption that "distant" variables are uncorrelated. Such structure is often encountered in longitudinal and time-series data.

Given the sample covariance matrix S_n, the banding estimator of Bickel and Levina [2008b] is defined as

$$\hat{\Psi}_{\text{band}}(P_n; b) \equiv \left\{ S_n^{(jl)} \mathbb{I}(|j-l| \leq b) : \ j, l = 1, \ldots, J \right\},$$

where b is a hyperparameter that determines the number of bands to retain from the sample covariance matrix and is chosen via a CV procedure. For the class of "bandable" covariance matrices, i.e., the set of well-conditioned matrices whose elements not in the central bands of the matrix decay rapidly, this banding estimator has been shown to be uniformly consistent in the spectral norm so long as $\log(J)/n \to 0$.

The tapering estimator of Cai et al. [2010b] is the smooth generalization of the banding estimator, gradually shrinking the off-diagonal bands of the sample covariance matrix towards zero. It is defined as

$$\hat{\Psi}_{\text{tap}}(P_n; b) \equiv W_b \circ S_n,$$

for some weight matrix W_b. Here, "$\circ$" denotes the Hadamard (element-wise) matrix product. Clearly, letting $W_b^{(jl)} = \mathbb{I}(|j-l| \leq b)$ for some positive integer b results in the banding estimator. A popular weighting scheme derived by Cai et al. [2010b] is

$$W_b^{(jl)} \equiv \begin{cases} 1, & \text{when } |j-l| \leq \frac{b}{2} \\ 2 - \frac{|j-l|}{b}, & \text{when } \frac{b}{2} < |j-l| \leq b \\ 0, & \text{otherwise} \end{cases},$$

which we use in our simulation study presented in Section 2.5. Cai et al. [2010b] also derived the optimal rates of convergence for this estimator under the Frobenius and spectral norms, considering a class of bandable covariance matrices that is more general than that considered by Bickel and Levina [2008b]: The smallest eigenvalue of the covariance matrices in this class can take on a value of zero. However, this estimator does not improve upon the bounds set by the banding estimator.

Shrinkage Estimators

We next consider the linear and non-linear shrinkage estimators inspired by Stein's work on empirical Bayesian methods. These estimators are rotation-equivariant, shrinking the eigenvalues of the sample covariance matrix towards a set of target values, whilst leaving its eigenvectors untouched. In doing so, the resultant estimators are better-conditioned than the sample covariance matrix in a manner guaranteeing that the resultant covariance matrix estimator be non-singular. Further, these estimators do not rely on sparsity assumptions about the true covariance matrix, setting them apart from those previously discussed.

One of the first shrinkage estimators, the linear shrinkage estimator of the sample covariance matrix, was proposed by Ledoit and Wolf [2004]. This estimator is defined as the convex combination of the sample covariance matrix and the identity matrix. Hence, it represents a compromise between S_n, an unbiased but highly variable estimator of ψ_0 in high dimensions, and $I_{J\times J}$, a woefully biased but fixed estimator. Ledoit and Wolf [2004] found that, under mild conditions, the asymptotically optimal shrinkage intensity with respect to the scaled (by J) Frobenius norm can be estimated consistently in high dimensions. This estimator is defined as

$$\hat{\Psi}_{\text{ls}}(P_n) \equiv \frac{b_n^2}{d_n^2} m_n I + \frac{a_n^2}{d_n^2} S_n, \tag{2.13}$$

for $m_n = \text{tr}(S_n)/J$, $d_n^2 = \|S_n - m_n I\|_{F,1/J}^2$, $\bar{b}_n^2 = n^{-2}\sum_i \|X_i X_i^\top - S_n\|_{F,1/J}^2$, $b_n^2 = \min(\bar{b}_n^2, d_n^2)$, and $a_n^2 = d_n^2 - b_n^2$.

Bespoke shrinkage targets may be used in place of the identity. For example, one might consider a dense matrix target whose diagonal elements are the average of the sample covariance matrix's diagonal elements and whose off-diagonal elements are equal to the average of all the sample covariance matrix's off-diagonal elements. For the sake of brevity, discussion of such estimators is omitted, but examples are provided by, among others, Ledoit and Wolf [2003] and Schäfer and Strimmer [2005], particularly for use in financial economics and statistical genomics applications, respectively.

When assumptions about the true covariance matrix's structure are unfounded, it can become impossible to select an appropriate linear shrinkage target. Instead, one might consider generalizing these estimators to shrink the eigenvalues of the sample covariance matrix in a non-linear fashion. That is, an estimator that shrinks the sample covariance matrix's eigenvalues not by a common shrinkage factor (as with linear shrinkage estimators) but with shrinkage factors tailored to each sample eigenvalue. As with the aforementioned linear shrinkage estimators, such non-linear shrinkage estimators produce positive-definite estimates so long as the shrunken sample eigenvalues are positive and rotation-equivariant. These estimators belong to a class initially introduced by Stein [1986] and have since seen a resurgence in the work of Ledoit and Wolf [2012, 2015]. More recently, Ledoit and Wolf [2018] derived an analytical non-linear shrinkage estimator that is asymptotically optimal in high dimensions and more computationally efficient than their previously formulated estimators.

Estimators Based on Factor Models

Covariance matrix estimators based on factor models form another broad family of estimators that do not assume sparsity of the true covariance matrix. Often encountered in econometrics and finance, these estimators utilize the operating assumption that the dataset's observations are functions of a few common, often latent, factors. The factor model can be described as follows:

$$X = \mu + \beta F + \epsilon, \tag{2.14}$$

where $X_{J\times 1}$ represents a random observation, $\mu_{J\times 1}$ a mean vector, $\beta_{J\times L}$ a matrix of factor coefficients, $F_{L\times 1}$ a random vector of L common factors, and $\epsilon_{J\times 1}$ a random error vector. Assuming that F and ϵ are uncorrelated, the covariance matrix of X is given by

$$\psi = \beta \text{Cov}(F)\beta^\top + \psi_\epsilon, \tag{2.15}$$

where ψ_ϵ is the covariance matrix of the random error.

For a review on estimating the covariance matrix in the presence of observable factors, see Fan et al. [2016a]. We now briefly discuss the estimation of ψ when the factors are unobservable. Notice that when ψ_ϵ is not assumed to be diagonal, the decomposition of ψ in Equation (2.15) is not identifiable for fixed n and J, since the random vector X is the only observed component of the model. By letting $J \to \infty$, and assuming that the eigenvalues of ψ_ϵ are uniformly bounded or grow at a slow rate relative to J and that the eigenvalues of $(1/J)\ \beta^\top\beta$ are uniformly bounded away from zero and infinity, it can be shown that $\beta\text{Cov}(F)\beta^\top$ is asymptotically identifiable [Cai et al., 2010b]. It follows from these assumptions that the signal in the factors increases as the number of features increases, while the noise contributed by the error term remains constant. The eigenvalues associated with $\beta\text{Cov}(F)\beta^\top$ therefore become easy to differentiate from those of ψ_ϵ.

Now, even with $\beta\text{Cov}(F)\beta^\top$ being asymptotically identifiable, β and F cannot be distinguished. As a solution, the following constraint is imposed on F: $\text{Cov}(F) = I_{L\times L}$. It then follows that

$$\psi = \beta\beta^\top + \psi_\epsilon. \tag{2.16}$$

Under the additional assumption that the columns of β be orthogonal, Fan et al. [2013] found that the leading L eigenvalues of ψ are spiked, meaning that they are bounded below by some constant [Johnstone, 2001], and grow at rate $O(J)$ as the dimension of ψ increases. The remaining $J-L$ eigenvalues are then either bounded above or grow slowly. This implies that the latent factors and their loadings can be approximated via the eigenvalues and eigenvectors of ψ.

Fan et al. [2013] therefore proposed the Principal Orthogonal compleEment Thresholding (POET) estimator of ψ, which was motivated by the spectral decomposition of the sample

covariance matrix

$$\begin{aligned} S_n &= \sum_{j=1}^{J} \lambda_j V_{\cdot,j} V_{\cdot,j}^\top \\ &\approx \sum_{j=1}^{L} \beta_{\cdot,j} \beta_{\cdot,j}^\top + \sum_{j=L+1}^{J} \lambda_j V_{\cdot,j} V_{\cdot,j}^\top, \end{aligned}$$

where λ_j and $V_{\cdot,j}$ are the j^{th} eigenvalues and eigenvectors of S_n, respectively, and $\beta_{\cdot,j}$ is the j^{th} column of β. For ease of notation, we denote the second term by S_ϵ and refer to this matrix as the principal orthogonal complement. The estimator for L latent variables is then defined as

$$\hat{\Psi}_{\text{POET}}(P_n; L, s) \equiv \sum_{j=1}^{L} \lambda_j V_{\cdot,j} V_{\cdot,j}^\top + T_{\epsilon,s}, \tag{2.17}$$

where $T_{\epsilon,s}$ is the generalized thresholding matrix of S_ϵ

$$T_{\epsilon,s}^{(jl)} \equiv \begin{cases} S_\epsilon^{(jj)}, & \text{when } j = l \\ s\left(S_\epsilon^{(jl)}\right), & \text{otherwise} \end{cases},$$

for some generalized thresholding function s.

Although this estimator is computationally efficient, the assumptions encoding the factor based model under which it is derived are such that the latent features' eigenvalues grow in J. This results in a poor convergence rate under the spectral norm [Fan et al., 2016a].

2.5 Simulation Study

Simulation Study Design

We conducted a series of simulation experiments using prominent covariance models to verify the theoretical results of our cross-validated estimator selection procedure. These models are described below.

Model 1: A dense covariance matrix, where

$$\psi^{(jl)} = \begin{cases} 1, & j = l \\ 0.5, & \text{otherwise} \end{cases}.$$

Model 2: An AR(1) model, where $\psi^{(jl)} = 0.7^{|j-l|}$. This covariance matrix, corresponding to a common timeseries model, is approximately sparse for large J, since the off-diagonal elements quickly shrink to zero.

Model 3: An MA(1) model, where $\psi^{(jl)} = 0.7^{|j-l|} \cdot \mathbb{I}(|j-l| \leq 1)$. This covariance matrix, corresponding to another common timeseries model, is truly sparse. Only the diagonal, subdiagonal, and superdiagonal contain non-zero elements.

Model 4: An MA(2) model, where

$$\psi^{(jl)} = \begin{cases} 1, & j = l \\ 0.6, & |j-l| = 1 \\ 0.3, & |j-l| = 2 \\ 0, & \text{otherwise} \end{cases}.$$

This timeseries model is similar to Model 3, but slightly less sparse.

Model 5: A random covariance matrix model. First, a $J \times J$ random matrix whose elements are i.i.d. Uniform$(0, 1)$ is generated. Next, entries below $1/4$ are set to 1, entries between $1/4$ and $1/2$ are set to -1, and the remaining entries are set to 0. The square of this matrix is then computed and added to the identity matrix $I_{J \times J}$. Finally, the corresponding correlation matrix is computed and used as the model's covariance matrix.

Model 6: A Toeplitz covariance matrix, where

$$\psi^{(jl)} = \begin{cases} 1, & j = l \\ 0.6|j-l|^{-1.3}, & \text{otherwise} \end{cases}.$$

Like the AR(1) model, this covariance matrix is approximately sparse for large J. However, the off-diagonal entries decay less quickly as their distance from the diagonal increases.

Model 7: A Toeplitz covariance matrix with alternating signs, where

$$\psi^{(jl)} = \begin{cases} 1, & j = l \\ (-1)^{|j-l|}0.6|j-l|^{-1.3}, & \text{otherwise} \end{cases}.$$

This model is almost identical to Model 6, though the signs of the covariance matrix's entries are alternating.

Model 8: A covariance matrix inspired by the latent variable model described in Equation (2.14). Let $\beta_{J \times 3} = (\beta_1, \ldots, \beta_J)^\top$, where β_j are randomly generated using a $N(0, I_{3 \times 3})$ distribution for $j = 1, \ldots, J$. Then $\psi = \beta\beta^\top + I_{J \times J}$ is the covariance matrix of a model with three latent factors.

Estimator	Hyperparameters
Sample covariance matrix	Not applicable
Hard thresholding [Bickel and Levina, 2008c]	Thresholds $= \{0.1, 0.2, \ldots, 1.0\}$
SCAD thresholding [Fan and Li, 2001, Rothman et al., 2009]	Thresholds $= \{0.1, 0.2, \ldots, 1.0\}$
Adaptive LASSO [Rothman et al., 2009]	Thresholds $= \{0.1, 0.2, \ldots, 0.5\}$; exponential weights $= \{0.1, 0.2, \ldots, 0.5\}$
Banding [Bickel and Levina, 2008b]	Bands $= \{1, 2, \ldots, 5\}$
Tapering [Cai et al., 2010b]	Bands $= \{2, 4, \ldots, 10\}$
Linear shrinkage [Ledoit and Wolf, 2004]	Not applicable
Dense linear shrinkage [Schäfer and Strimmer, 2005]	Not applicable
Nonlinear shrinkage [Ledoit and Wolf, 2018]	Not applicable
POET [Fan et al., 2013] using hard thresholding	Latent factors $= \{1, 2, \ldots, 5\}$; thresholds $= \{0.1, 0.2, 0.3\}$

Table 2.1: Families of candidate estimators used by cvCovEst in the simulation study (74 distinct estimators in total)

Each covariance model was used to generate data sets consisting of $n \in \{50, 100, 200, 500\}$ i.i.d. multivariate Gaussian, mean-zero observations. The uniform boundedness condition of Theorem 2.1's Assumption 2.1 is therefore not satisfied; we do this purposefully to further stress that this assumption is not limiting in many practical settings. For each model and sample size, five data dimension ratios were considered: $J/n \in \{0.3, 0.5, 1, 2, 5\}$. Together, the eight covariance models, four sample sizes, and five dimensionality ratios result in 160 distinct simulation settings. For each such setting, the performance of the cross-validated selector with respect to the various oracle selectors and several well-established estimators is evaluated based on aggregation across 200 Monte Carlo repetitions.

We applied our estimator selection procedure, which we refer to as *cvCovEst*, using a 5-fold CV scheme. The library of candidate estimators is provided in Table 2.1, which includes details on these estimators' possible hyperparameters. Seventy-four estimators make up the library of candidates. We note that no penalty is attributed to estimators generating rank-deficient estimates, like the sample covariance matrix when $J > n$, though this limitation is generally of practical importance. When the situation dictates that the resulting estimate

must be positive-definite, the library of candidates should be assembled accordingly.

Simulation Study Results

To examine empirically the optimality results of Theorem 2.1, we computed analytically, for each replication, the cross-validated conditional risk differences of the cross-validated selection, $\hat{k}_{p_n,n}$, and the cross-validated oracle selection, $\tilde{k}_{p_n,n}$. The Monte Carlo expectations of the risk differences stratified by n, J/n, and the covariance model were computed from the cross-validated conditional risk differences. The ratios of the expected risk differences are presented in Figure 2.1. These results make clear that, for virtually all models considered, the estimator chosen by our proposed cross-validated selection procedure has a risk difference asymptotically identical on average to that of the cross-validated oracle.

A stronger result, corresponding to Equation (2.8) of Theorem 2.1, is presented in Figure 2.2. For all but Models 1 and 8, we find that our algorithm's selection is virtually equivalent to the cross-validated oracle selection for $n \geq 200$ and $J/n \geq 0.5$. Even for Model 8, in which the covariance matrices are more difficult to estimate due to their dense structures, we find that our selector identifies the optimal estimator with probability tending to 1 for $n \geq 200$ and $J/n = 5$.

More impressive still are the results presented in Figure 2.3 that characterize the full-dataset conditional risk difference ratios. For all covariance matrix models considered, with the exception of Model 1, our procedure's selections attain near asymptotic optimality for moderate values of n and J/n. This suggests that our loss-based estimator selection approach's theoretical guarantee, as outlined in Corollary 2.1, is achievable in many practical settings.

In addition to verifying our method's asymptotic behavior, we compared the estimates generated by our method against those of the individual candidate procedures using the simulated data sets. This was accomplished by computing the Frobenius norm of each estimate against the corresponding true covariance matrix. The mean norms over all simulations were then computed for each covariance matrix estimation procedure, again stratified by n, J/n, and the covariance matrix model (Figure 2.4). Our CV scheme was used to select hyperparameters of these competing approaches where necessary. As stated previously, our CV approach is generally equivalent to these estimators' hyperparameter selection procedures. The hyperparameters considered are provided in Table 2.2. Where appropriate, the competing methods' hyperparameters are more varied than those used by cvCovEst, reflecting more aggressive estimation procedures that one might employ when only using a single family of estimators.

We repeated this benchmarking experiment using the spectral norm to assess the accuracy of our estimation procedure with respect to the leading eigenvalue of the covariance matrix. Recall that the spectral norm of a square matrix is defined as it's largest absolute eigenvalue. Though our theoretical results do not relate to to this norm, outcomes similar to those in Figure 2.4 are expected given the relationship between these two norms for reasons previously described.

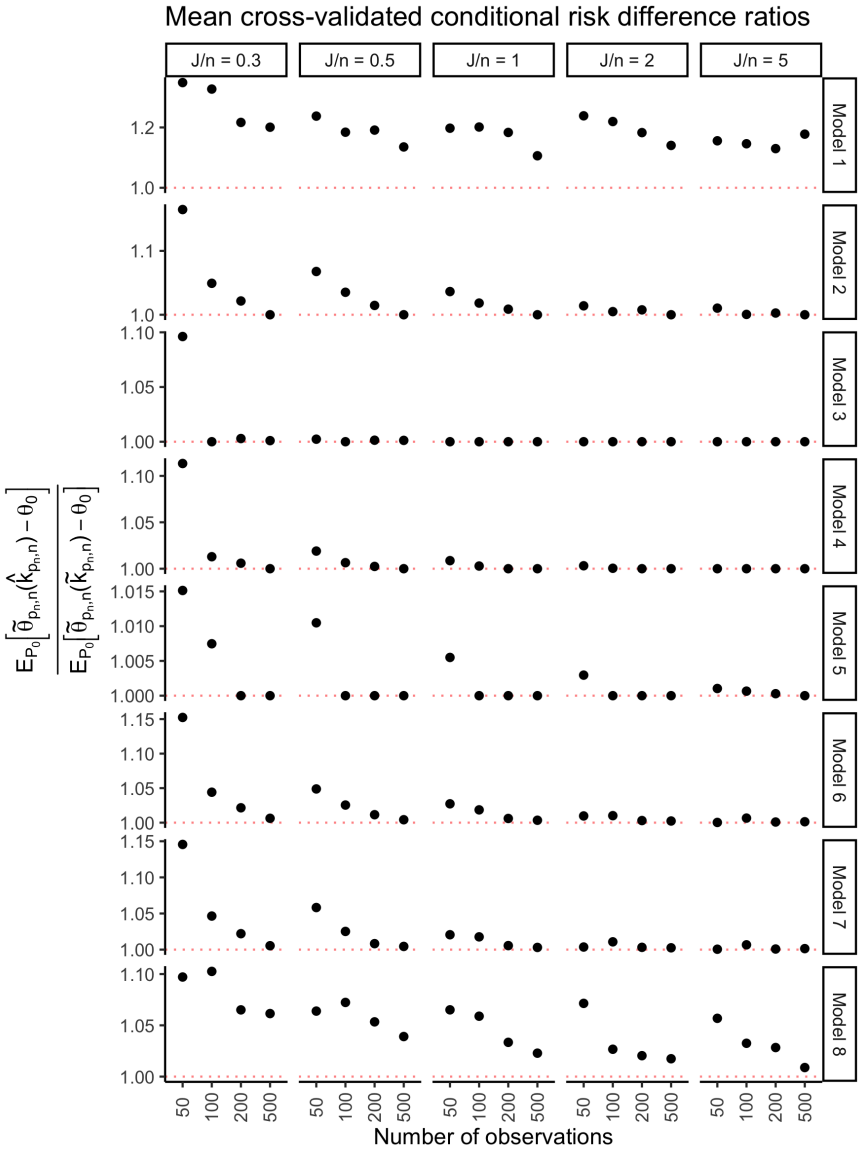


Figure 2.1: Comparison of the cross-validated selection and cross-validated oracle selection's mean cross-validated conditional risk differences. Note the differing y-axis scales for the different models.

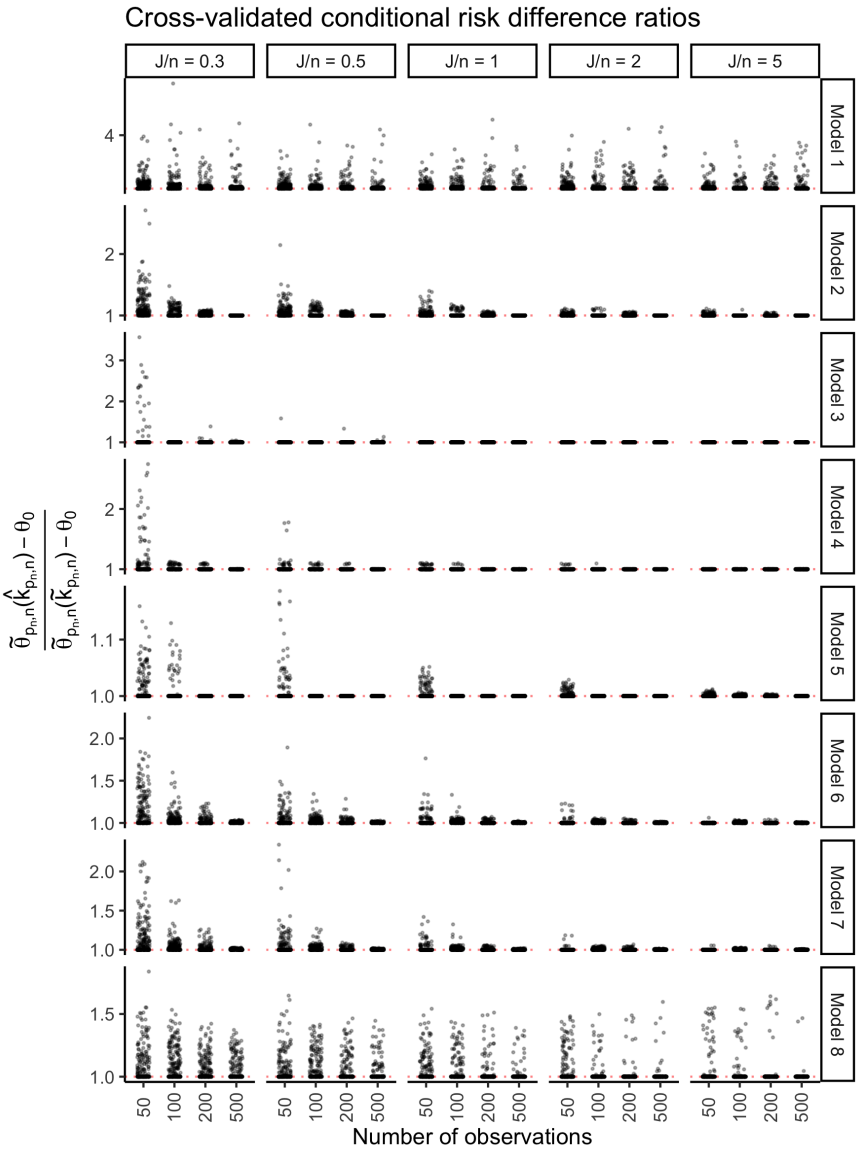


Figure 2.2: Comparison of the cross-validated ($\hat{k}_{p_n,n}$) and cross-validated oracle ($\tilde{k}_{p_n,n}$) selections' cross-validated conditional risk differences. The proposed cross-validated selection procedure achieves asymptotic equivalence in most settings for relatively small sample sizes and numbers of features.

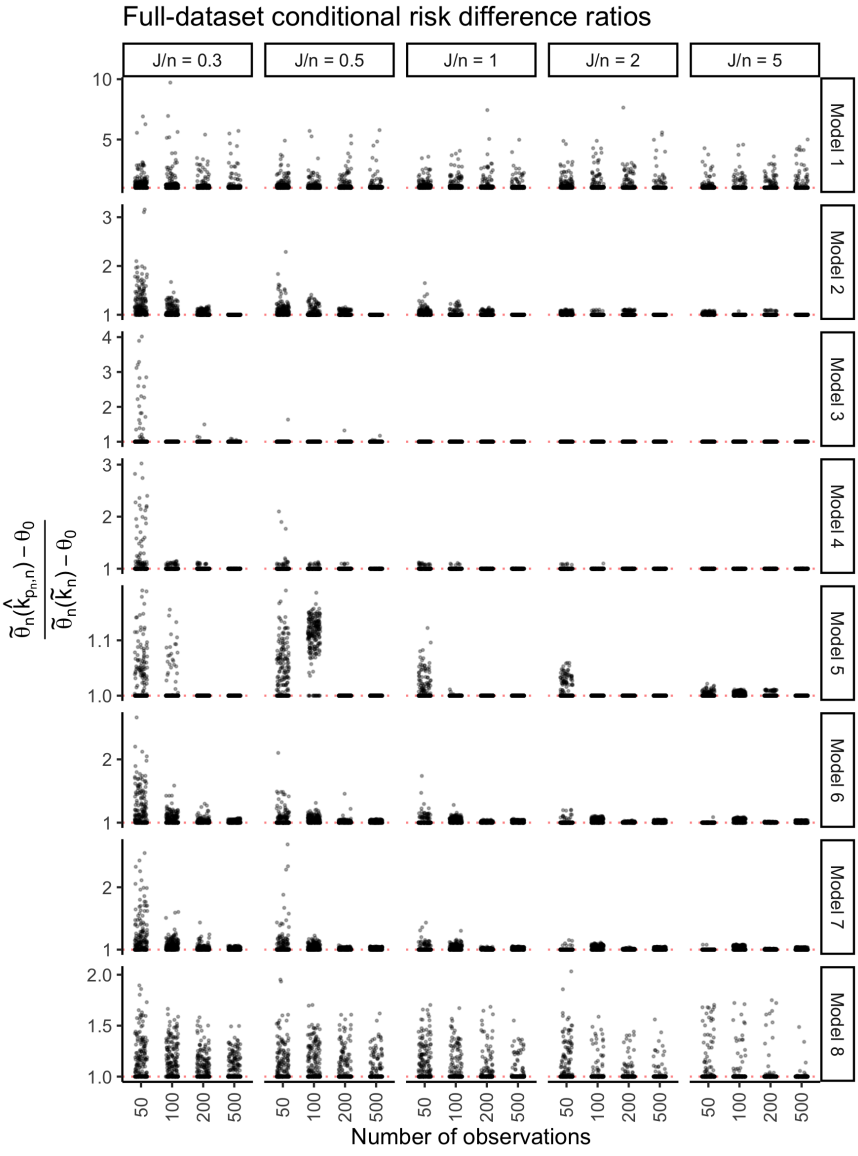


Figure 2.3: Comparison of the cross-validated selection and oracle selection's full-dataset conditional risk differences.

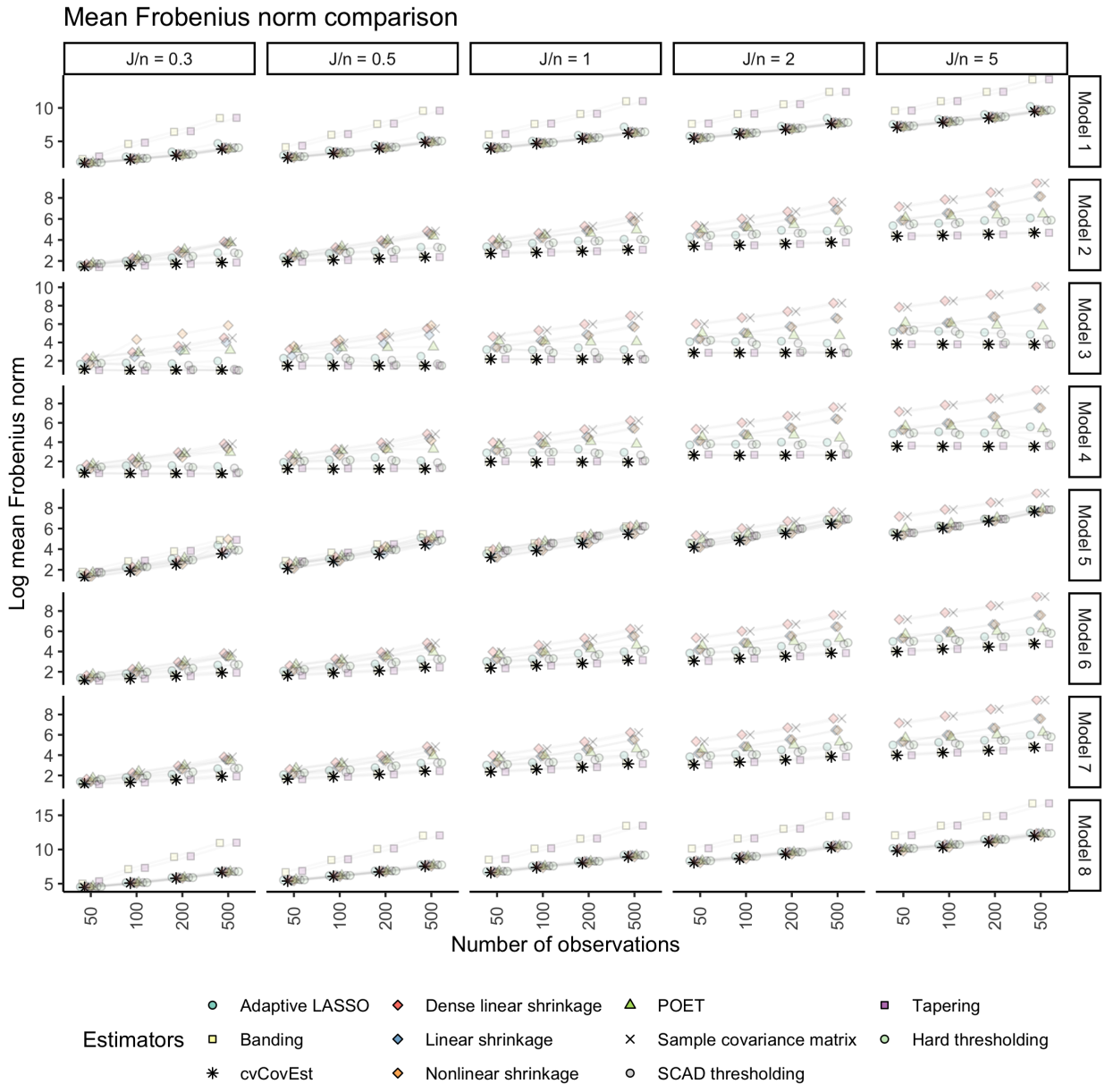


Figure 2.4: Comparison of competing, bespoke covariance matrix estimation procedures to our cross-validated selection approach in terms of the Monte Carlo mean Frobenius norm under a variety of data-generating processes. Note that the scales of the y-axis are tailored to the covariance matrix model.

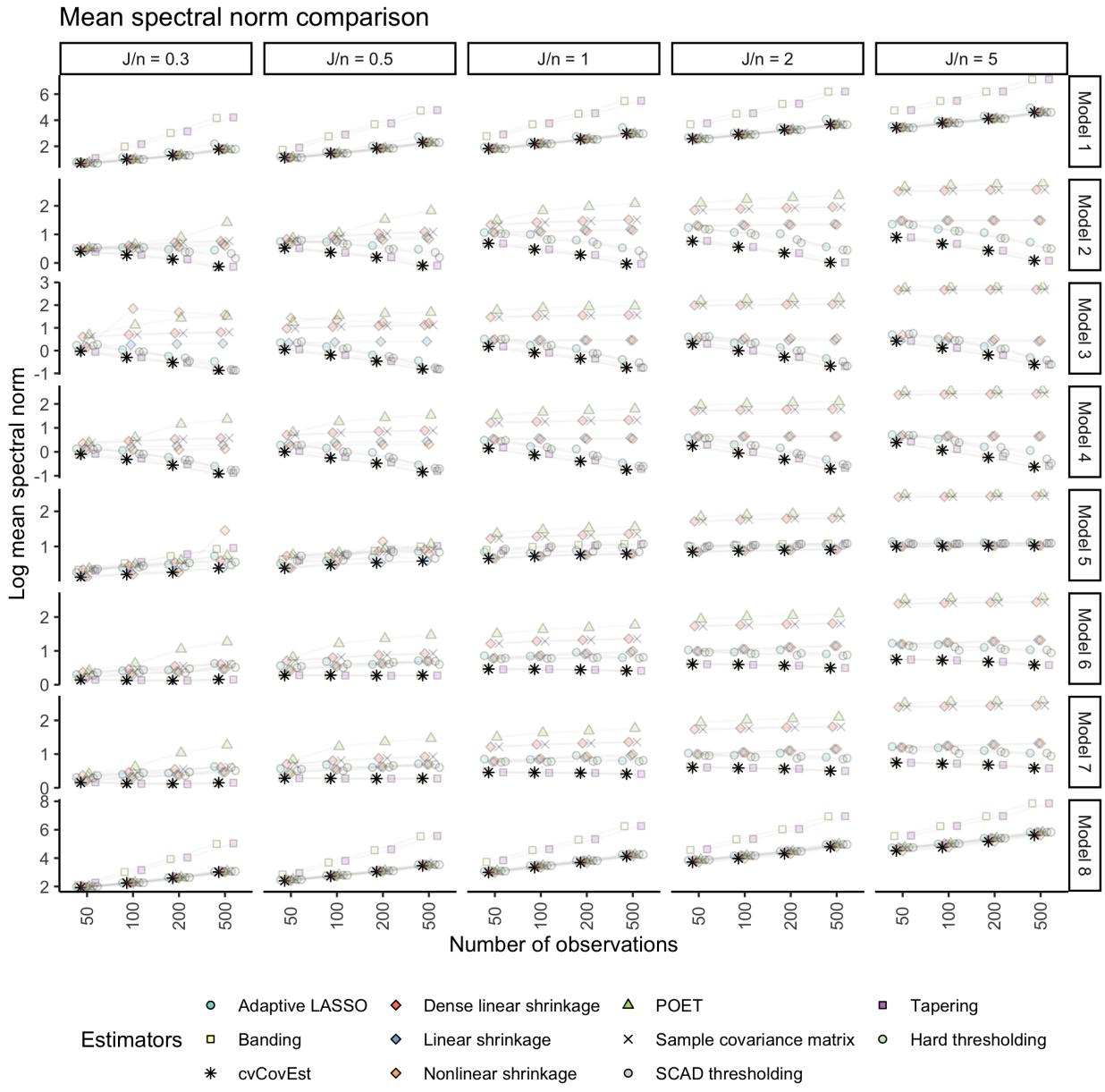


Figure 2.5: Comparison of competing, bespoke covariance matrix estimation procedures to our cross-validated selection approach in terms of the Monte Carlo mean spectral norm under a variety of data-generating processes. Note that the scales of the y-axis are tailored to the covariance matrix model.

Estimator	Hyperparameters
Sample covariance matrix	NA
Hard thresholding	Thresholds = $\{0.05, 0.10, \ldots, 1.00\}$
SCAD thresholding	Thresholds = $\{0.05, 0.10, \ldots, 1.00\}$
Adaptive LASSO	Thresholds = $\{0.1, 0.2, \ldots, 0.5\}$; exponential weights = $\{0.1, 0.2, \ldots, 0.5\}$
Banding	Bands = $\{1, 2, \ldots, 10\}$
Tapering	Bands = $\{2, 4, \ldots, 10\}$
Linear shrinkage	NA
Dense linear shrinkage	NA
Nonlinear shrinkage	NA
POET using hard thresholding	Latent factors = $\{1, 2, \ldots, 10\}$; thresholds = $\{0.1, 0.2, \ldots, 1.0\}$

Table 2.2: Families of candidate estimators compared against the cross-validated loss-based estimator selection procedure. Note that the library of candidate estimators used by the proposed method is provided in Table 2.1

The results, presented in Figures 2.4 and 2.5, demonstrate that our estimator selection procedure performs at least as well as the best alternative estimation strategy. This suggests that procedures dedicated to or relying upon the accurate estimation of leading eigenvalues and eigenvectors, like principal component analysis and latent variable estimation, might benefit from the integration of our cross-validated covariance matrix estimation framework.

2.6 Real Data Examples

Single-cell transcriptome sequencing (scRNA-seq) allows researchers to study the gene expression profiles of individual cells. The fine-grained transcriptomic data that it provides have been used to identify rare cell populations and to elucidate the developmental relationships between diverse cellular states.

Given that a typical scRNA-seq data set possesses tens of thousands of features (genes), most workflows prescribe a dimensionality reduction step. In addition to reducing the amount of computational resources needed to analyze the data, reducing the dimensions mitigates the effect of corrupting noise on interesting biological signal. The lower-dimensional embedding is then used in downstream analyses, like novel cell-type discovery via clustering.

One of the most popular methods used to reduce the dimensionality of scRNA-seq data is uniform manifold approximation and projection (UMAP) [McInnes et al., 2018]. This method captures the most salient non-linear relationships among a high-dimensional data set's features and projects them onto a reduced-dimensional space. Instead of applying

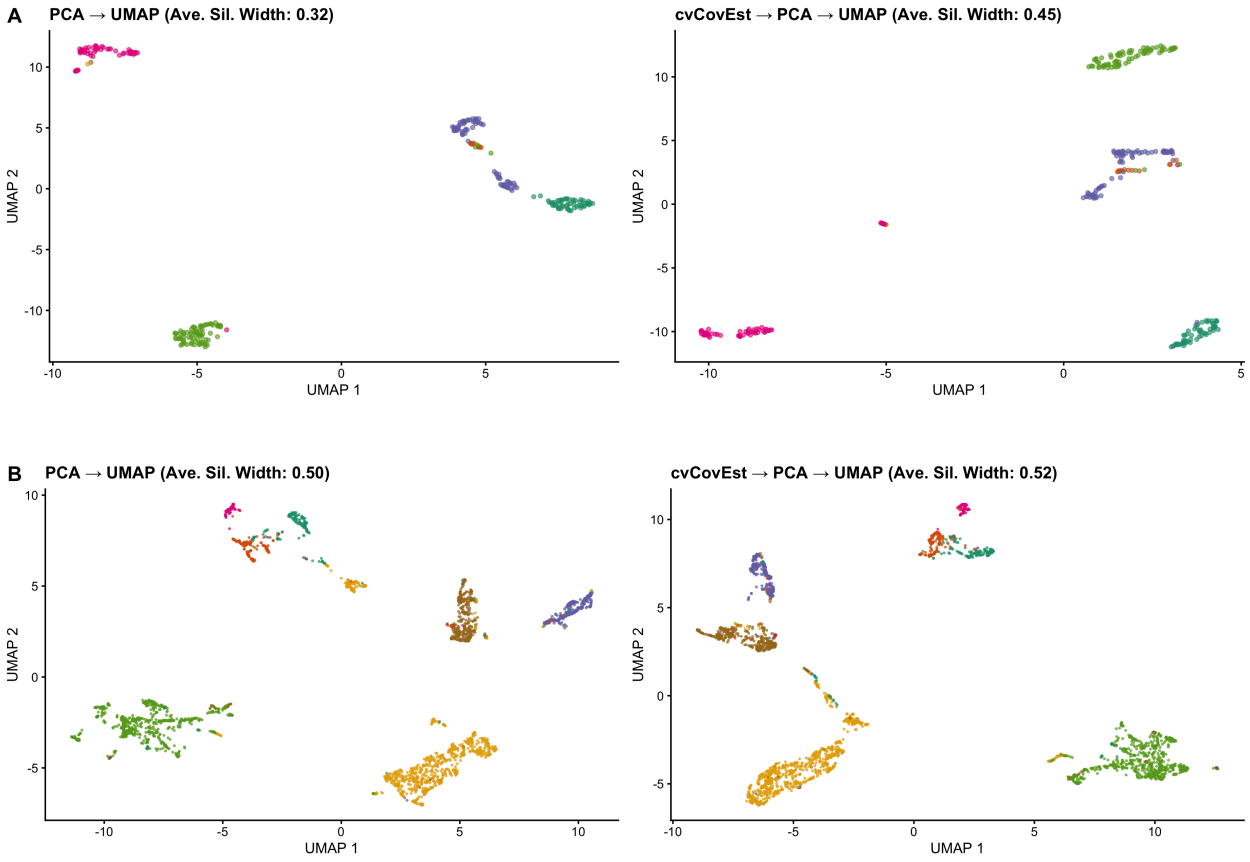


Figure 2.6: Comparisons of scRNA-seq data sets' UMAP embeddings based on vanilla PCA or PCA with the cross-validated selection's covariance matrix estimate. The data sets consist of **(A)** 285 cells collected from the visual cortex of mice and **(B)** 2,816 mouse brain cells. Distinct cell types are indicated by color.

UMAP directly, the scRNA-seq data set's leading principal components (PCs) are often used as an initialization.

This initial dimensionality reduction by PCA is believed to play a helpful role in denoising. However, PCA typically relies on the sample covariance matrix, and so when the data set is high-dimensional, the resulting principal components are known to be poor estimates of those of the population [Johnstone and Lu, 2009]. We hence posit that our cross-validated estimator selection procedure could form a basis for an improved PCA. That is, we hope that the eigenvectors resulting from the eigendecomposition of our estimated covariance matrix could be used to generate a set of estimates closer to the true PCs in terms of risk. These PCs could then be fed to UMAP to produce an enhanced embedding. Indeed, simulation results provided in Figure 2.5 suggest that cvCovEst produces estimates of the leading eigenvalue at least as well as those produced by the sample covariance matrix, in terms of the spectral

Estimator	Hyperparameters
Sample covariance matrix	NA
Hard thresholding	Thresholds = $\{0.05, 0.10, \ldots, 0.30\}$
SCAD thresholding	Thresholds = $\{0.05, 0.10, \ldots, 0.50\}$
Adaptive LASSO	Thresholds = $\{0.1, 0.2, \ldots, 0.5\}$; exponential weights = $\{0.1, 0.2, \ldots, 0.5\}$
Linear shrinkage	NA
Dense linear shrinkage	NA
POET using hard thresholding	Latent factors= $\{5, 6, \ldots, 10\}$; thresholds = $\{0.05, 0.10, \ldots, 0.3\}$

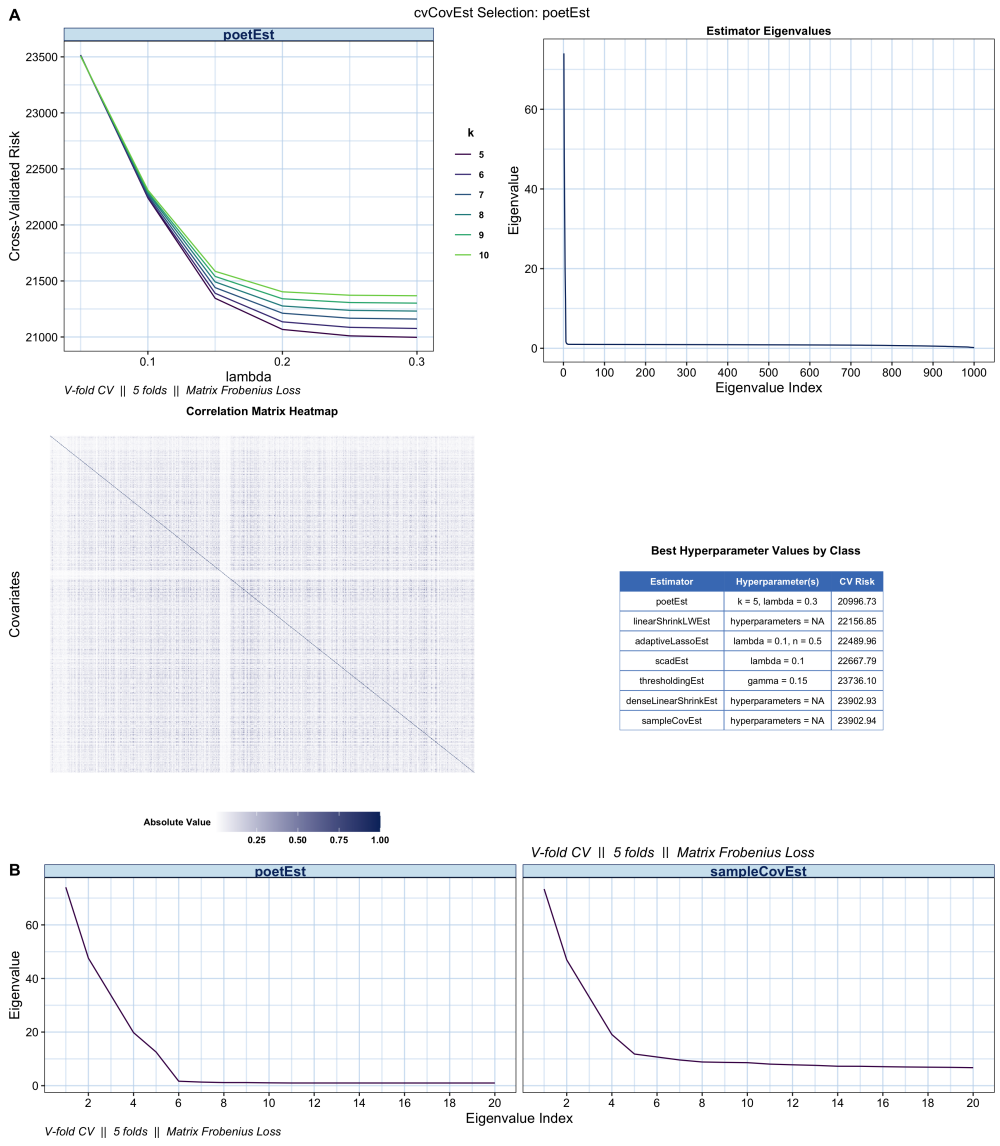
Table 2.3: Families of candidate estimators used in single-cell transcriptomic data analyses

norm.

We applied our procedure to two scRNA-seq data sets for which the cell types are known *a priori*. These data were obtained from the `scRNAseq` Bioconductor `R` package [Risso and Cole, 2020], and prepared for analysis using a workflow outlined in Amezquita et al. [2020]. A 5-fold CV scheme was used; the library of candidate estimators is provided in Table 2.3. We expect that cells of the same class will form tight, distinct clusters within the low-dimensional representations. The resulting embeddings, which we refer to as the *cvCovEst-based* embeddings, were then compared to those produced by UMAP using traditional PCA for initialization, which we refer to as the *PCA-based* embeddings. For each embedding, the 20 leading PCs were fed to UMAP. The first data set is a collection of 285 mouse visual cortex cells [Tasic et al., 2016], and the second data set consists of 2,816 mouse brain cells [Zeisel et al., 2015]. The 1,000 most variable genes of each data set were used to compute the PCs of both embeddings.

The resulting UMAP plots are presented in Figure 2.6. Though the two embeddings generated for each data set are qualitatively similar, the low-dimensional representation relying on our loss-based approach is more refined in Figure 2.6A. A number of cells erroneously clustered in the PCA-based embedding are correctly represented in the cvCovEst-based embedding. This explains the 41% increase in average silhouette width of our method relative to the traditional approach. Further insight is gleaned from the diagnostic plots of Figure 2.7. Figure 2.7A indicates that cvCovEst selected the POET estimator [Fan et al., 2013] with 5 latent factors and a thresholding hyperparameter of 0.3. It that the selected estimator significantly improves upon the sample covariance matrix in terms of the cross-validated Frobenius risk. Figure 2.7B provides further insight into the discrepancies between the UMAP results of Figure 2.6A: the sample covariance matrix likely over-estimates many of the leading eigenvalues.

The embeddings in Figure 2.6B qualitatively identical, and so too are their average silhouette widths. This is expected, the Zeisel et al. [2015] is not truly high-dimensional.



Estimator	Hyperparameter(s)	CV Risk
poetEst	k = 5, lambda = 0.3	20996.73
linearShrinkLWEst	hyperparameters = NA	22156.85
adaptiveLassoEst	lambda = 0.1, n = 0.5	22489.96
scadEst	lambda = 0.1	22667.79
thresholdingEst	gamma = 0.15	23736.10
denseLinearShrinkEst	hyperparameters = NA	23902.93
sampleCovEst	hyperparameters = NA	23902.94

Figure 2.7: Tasic dataset: Diagnostic plots and tables generated using the cvCovEst R package. **(A)** The top-left plot presents the cross-validated Frobenius risk of the estimator selected by our method. k represents the number of potential latent factors, and `lambda` the thresholding value used. The top-right panel contains a line plot of the selected estimator's eigenvalues. The bottom-left plot displays the absolute values of the estimated correlation matrix output by the cvCovEst selection, and the bottom-right table lists the best performing estimators from all classes of estimators considered. **(B)** Side-by-side line plots of the estimated leading eigenvalues of the cvCovEst selection and the sample covariance matrix.

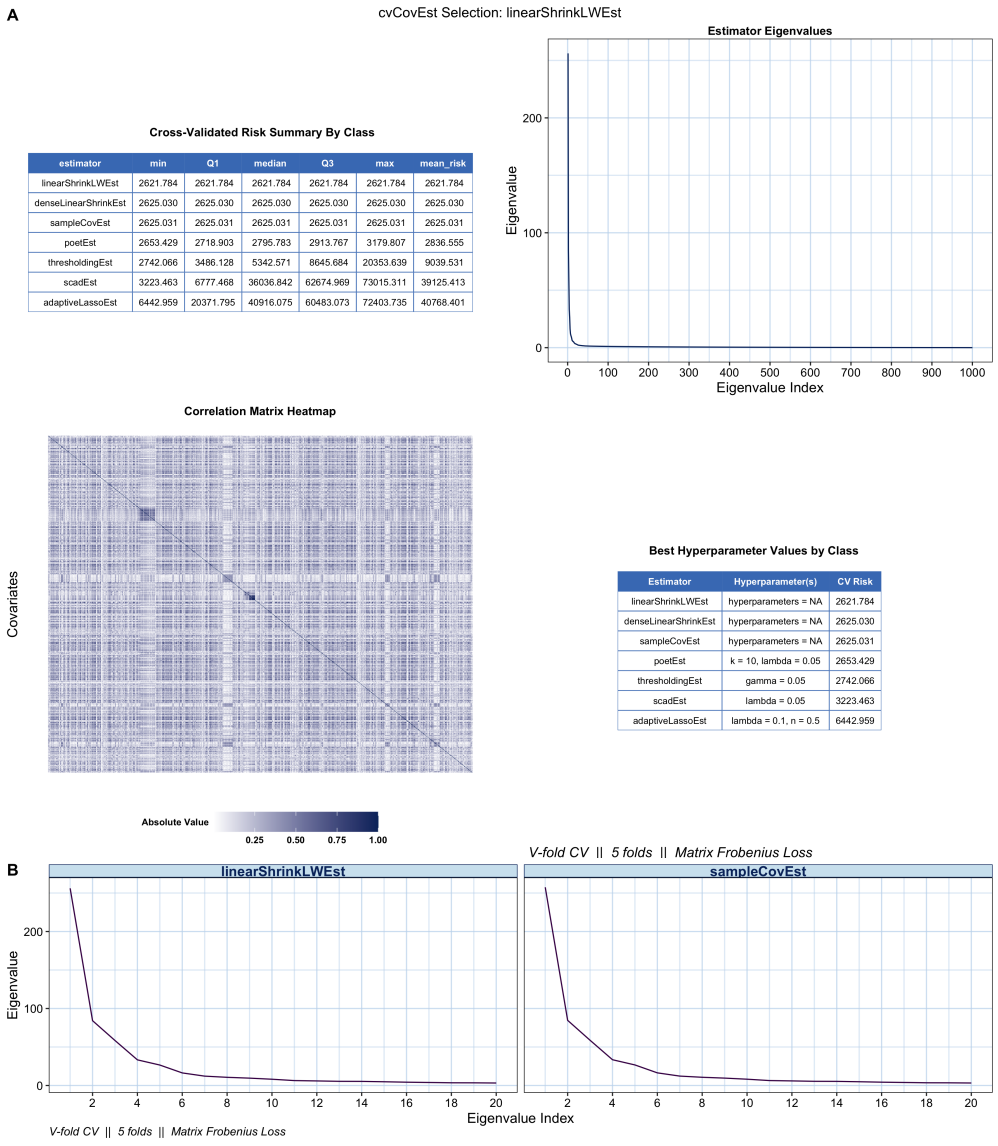


Figure 2.8: Zeisel dataset: Diagnostic plots and tables generated using the cvCovEst R package. The components in **(A)** can be interpreted in the same manner as the previous figure, with the exception of the top-left panel. In this table, a five number summary of the cross-validated Frobenius risk is given for each class of estimators considered across all possible combinations of hyperparameters, if any. It is clear from the tables in **(A)** that the cvCovEst selection is essentially equivalent in terms of cross-validated risk to the sample covariance matrix. The plots in **(B)** further highlight that the 20 leading eigenvalue of the cvCovEst estimate and the sample covariance matrix are indistinguishable.

The sample covariance matrix likely is a reasonable estimator in this setting. Ideally, data-adaptive selection procedures should be cognizant of this. Indeed, cvCovEst, when applied to the Zeisel et al. [2015] data set, selects an estimator whose cross-validated empirical risk is only slightly smaller than that of the sample covariance matrix, and whose leading PCs are virtually identical (Figure 2.8).

2.7 Discussion

This work extends Dudoit and van der Laan [2005]'s framework for asymptotically optimal, data-adaptive estimator selection to the problem of covariance matrix estimation in high-dimensional settings. We provide sufficient conditions under which our cross-validated procedure is asymptotically optimal in terms of risk, and show that it generalizes the cross-validated hyperparameter selection procedures employed by existing estimation approaches. Future work might derive analogous results for other loss functions, or perhaps even for other parameters like the precision matrix.

The simulation study provides evidence that near-optimal results are achieved in data sets with relatively modest numbers of observations and many features across models indexed by diverse covariance matrix structures. These results also establish that our cross-validated procedure performs as well as the best bespoke estimation procedure in a variety of settings. Our scRNA-seq data examples further illustrate the utility of our approach in fields where high-dimensional data are collected routinely.

Practitioners need no longer rely upon intuition alone when deciding which candidate estimator is best from among a library of diverse estimators. We expect that a variety of computational procedures relying upon the accurate estimation of the covariance matrix beyond the exploratory analyses considered here, like clustering and latent variable estimation, stand to benefit from the application of this framework.

2.8 Proofs

Proof. **Proposition 2.1.**

Assume without loss of generality that $\mathbb{E}_{P_0}[X] = 0$ and $\eta_0 = 1$. Then,

$$
\begin{aligned}
\hat{\theta}_{p_n,n}(k,1) &= \mathbb{E}_{B_n}\left[\frac{1}{np_n}\sum_{i=1}^{n}\mathbb{I}(B_n(i)=1)\|X_iX_i^\top - \hat{\Psi}_k(P^0_{n,B_n})\|^2_{F,1}\right] \\
&= \mathbb{E}_{B_n}\left[\frac{1}{np_n}\sum_{i=1}^{n}\mathbb{I}(B_n(i)=1)\sum_{j=1}^{J}\sum_{l=1}^{J}(X_i^{(j)}X_i^{(l)} - \hat{\Psi}_k(P^0_{n,B_n})^{(jl)})^2\right] \\
&= \mathbb{E}_{B_n}\left[\frac{1}{np_n}\sum_{j=1}^{J}\sum_{l=1}^{J}\left(\sum_{\{i:B_n(i)=1\}}\left((X_i^{(j)}X_i^{(l)})^2 - 2X_i^{(j)}X_i^{(l)}\hat{\Psi}_k(P^0_{n,B_n})^{(jl)}\right)\right.\right. \\
&\qquad\qquad \left.\left. + (\hat{\Psi}_k(P^0_{n,B_n})^{(jl)})^2\right)\right] \\
&= \mathbb{E}_{B_n}\left[\sum_{j=1}^{J}\sum_{l=1}^{J}\left((\hat{\Psi}_k(P^0_{n,B_n})^{(jl)})^2 - 2S(P^1_{n,B_n})^{(jl)}\hat{\Psi}_k(P^0_{n,B_n})^{(jl)}\right.\right. \\
&\qquad\qquad \left.\left. + \frac{1}{np_n}\sum_{\{i:B_n(i)=1\}}(X_i^{(j)}X_i^{(l)})^2\right)\right] \\
&= \mathbb{E}_{B_n}\left[\sum_{j=1}^{J}\sum_{l=1}^{J}\left(\left(\hat{\Psi}_k(P^0_{n,B_n})^{(jl)}\right)^2 - 2S(P^1_{n,B_n})^{(jl)}\hat{\Psi}_k(P^0_{n,B_n})^{(jl)}\right)\right] + C_1,
\end{aligned}
$$

where C_1 is constant with respect to $\hat{\Psi}_k(P^0_{n,B_n})$.

From Equation (2.5), notice that

$$
\begin{aligned}
\hat{R}_n(\hat{\Psi}_k,1) &= \mathbb{E}_{B_n}\left[\|\hat{\Psi}_k(P^0_{n,B_n}) - S_n(P^1_{n,B_n})\|^2_{F,1}\right] \\
&= \mathbb{E}_{B_n}\left[\sum_{j=1}^{J}\sum_{l=1}^{J}\left((\hat{\Psi}_k(P^0_{n,B_n})^{(jl)})^2 - 2S(P^1_{n,B_n})^{(jl)}\hat{\Psi}_k(P^0_{n,B_n})^{(jl)} + (S(P^1_{n,B_n})^{(jl)})^2\right)\right] \\
&= \mathbb{E}_{B_n}\left[\sum_{j=1}^{J}\sum_{l=1}^{J}\left((\hat{\Psi}_k(P^0_{n,B_n})^{(jl)})^2 - 2S(P^1_{n,B_n})^{(jl)}\hat{\Psi}_k(P^0_{n,B_n})^{(jl)}\right)\right] + C_2,
\end{aligned}
$$

where C_2 is constant with respect to $\hat{\Psi}_k(P^0_{n,B_n})$.
Thus,

$$
\begin{aligned}
\hat{k}_{p_n,n} &= \arg\min_{k\in\{1,\ldots,K\}}\hat{\theta}_{p_n,n}(k,1) \\
&= \arg\min_{k\in\{1,\ldots,K\}}\hat{R}_n(\hat{\Psi}_k,1).
\end{aligned}
$$

□

Proof. **Proposition 2.2.**

For any $a = 1, \ldots, J$ and $b = 1, \ldots, J$, we find that

$$\begin{aligned}
0 &= \frac{\delta}{\delta\psi^{(ab)}}\Theta(\psi, \eta, P_0) \\
&= \frac{\delta}{\delta\psi^{(ab)}}\mathbb{E}_{P_0}\left[L(X; \psi, \eta)\right] \\
&= \frac{\delta}{\delta\psi^{(ab)}}\mathbb{E}_{P_0}\left[\sum_{j=1}^{J}\sum_{l=1}^{J}\eta^{(jl)}\left(X^{(j)}X^{(l)} - \psi^{(jl)}\right)^2\right] \\
&= \mathbb{E}_{P_0}\left[\frac{\delta}{\delta\psi^{(ab)}}\sum_{j=1}^{J}\sum_{l=1}^{J}\eta^{(jl)}\left(X^{(j)}X^{(l)} - \psi^{(jl)}\right)^2\right] \\
&\propto \mathbb{E}_{P_0}\left[X^{(a)}X^{(b)} - \psi^{(ab)}\right] \\
\Rightarrow \psi^{(ab)} &= \psi_0^{(ab)}.
\end{aligned}$$

It follows that ψ_0 is a risk minimizer. □

Proof. **Proposition 2.3.**

Let the hard-thresholding estimator minimizing the expected cross-validated conditional risk under P_0 be indexed by k_0. Then

$$\begin{aligned}
\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0] &\leq \mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(k_0) - \theta_0] \\
&= \mathbb{E}_{P_0}\left[\mathbb{E}_{B_n}\left[\mathbb{E}_{P_0}\left[\|\hat{\Psi}_{k_0}(P^0_{n,B_n}) - XX^\top\|^2_{F,1}\right.\right.\right. \\
&\qquad\qquad \left.\left.\left. - \|\psi_0 - XX^\top\|^2_{F,1} \mid P^0_{n,B_n}\right]\right]\right] \\
&= \mathbb{E}_{P_0}\left[\mathbb{E}_{B_n}\left[\mathbb{E}_{P_0}\left[\|\hat{\Psi}_{k_0}(P^0_{n,B_n}) - \psi_0\|^2_{F,1} \mid P^0_{n,B_n}\right]\right]\right] \\
&= O(s(J)\log J).
\end{aligned}$$

The first inequality follows from the definition of the cross-validated conditional risk under P_0; the last equality follows from Theorem 2 (and its subsequent discussion) of Bickel and Levina [2008c], since X is sub-Gaussian by the boundedness condition of Assumption 1 and $J/n \to m$ as $n, J \to \infty$. Then,

$$\frac{c(\delta, \overline{M}(J))(1 + \log(K)}{np_n\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0]} = \Omega\left(\frac{J}{s(J)\log J}\right).$$

□

Lemma 2.1. *Let* $Z_k \equiv L(X; \hat{\Psi}_k(P^0_{n,B_n})) - L(X; \psi_0)$. *Then,*

$$Var_{P_0}\left[Z_k \mid P^0_{n,B_n}, B_n\right] \leq \overline{M}(J)\mathbb{E}_{P_0}\left[Z_k \mid P^0_{n,B_n}, B_n\right],$$

where $\overline{M}(J) \equiv 4(M_1 + M_2)^2J^2$.

Proof.

$$\begin{aligned}
\mathbb{E}_{P_0}\left[Z_k \mid P^0_{n,B_n}, B_n\right] &= \mathbb{E}_{P_0}\left[L(X;\hat{\Psi}_k(P^0_{n,B_n})) - L(X;\psi_0)\middle| P^0_{n,B_n}, B_n\right] \\
&= \mathbb{E}_{P_0}\left[\sum_{j=1}^{J}\sum_{l=1}^{J}(X^{(j)}X^{(l)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))^2 - (X^{(j)}X^{(l)} - \psi_0^{(jl)})^2\middle| P^0_{n,B_n}, B_n\right] \\
&= \sum_{j=1}^{J}\sum_{l=1}^{J}\mathbb{E}_{P_0}\Big[(\psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))(2X^{(j)}X^{(l)} - \psi_0^{(jl)} \\
&\qquad\qquad - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))\Big| P^0_{n,B_n}, B_n\Big] \\
&= \sum_{j=1}^{J}\sum_{l=1}^{J}(\psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n})) \\
&\qquad\qquad \mathbb{E}_{P_0}\left[2X^{(j)}X^{(l)} - \psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n})\middle| P^0_{n,B_n}, B_n\right] \\
&= \sum_{j=1}^{J}\sum_{l=1}^{J}(\psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))^2 \\
&= \sum_{j=1}^{J}\sum_{l=1}^{J}\mathbb{E}_{P_0}\left[(\psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))^2\middle| P^0_{n,B_n}, B_n\right],
\end{aligned}$$

where the second to last equality follows by noting that $\mathbb{E}_{P_0}[X^{(j)}X^{(l)}|P^0_{n,B_n}, B_n] = \mathbb{E}_{P_0}[X^{(j)}X^{(l)}]$ and that, by definition, $\mathbb{E}_{P_0}[X^{(j)}X^{(l)}] = \psi_0^{(jl)}$. Then,

$$
\begin{aligned}
\mathrm{Var}_{P_0}\left[Z_k \mid P^0_{n,B_n}, B_n\right] &\leq \mathbb{E}_{P_0}\left[Z_k^2 \mid P^0_{n,B_n}, B_n\right] \\
&= \mathbb{E}_{P_0}\Bigg[\Bigg(\sum_{j=1}^{J}\sum_{l=1}^{J}(\psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))(2X^{(j)}X^{(l)} - \psi_0^{(jl)} \\
&\qquad\qquad - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))\Bigg)^2 \Bigg| P^0_{n,B_n}, B_n\Bigg] \\
&= J^4\mathbb{E}_{P_0}\Bigg[\Bigg(\frac{1}{J^2}\sum_{j=1}^{J}\sum_{l=1}^{J}(\psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))(2X^{(j)}X^{(l)} - \psi_0^{(jl)} \\
&\qquad\qquad - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))\Bigg)^2 \Bigg| P^0_{n,B_n}, B_n\Bigg] \\
&\leq J^4\mathbb{E}_{P_0}\Bigg[\frac{1}{J^2}\sum_{j=1}^{J}\sum_{l=1}^{J}(\psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))^2(2X^{(j)}X^{(l)} - \psi_0^{(jl)} \\
&\qquad\qquad - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))^2 \Bigg| P^0_{n,B_n}, B_n\Bigg] \\
&\leq J^2 4(M_1+M_2)^2\mathbb{E}_{P_0}\left[\sum_{j=1}^{J}\sum_{l=1}^{J}(\psi_0^{(jl)} - \hat{\Psi}_k^{(jl)}(P^0_{n,B_n}))^2 \Bigg| P^0_{n,B_n}, B_n\right] \\
&= 4(M_1+M_2)^2J^2\mathbb{E}_{P_0}\left[Z_k \mid P^0_{n,B_n}, B_n\right] \\
&= \overline{M}(J)\mathbb{E}_{P_0}\left[Z_k \mid P^0_{n,B_n}, B_n\right].
\end{aligned}
$$

Here, the second inequality holds from the application of Jensen's Inequality to the square of the double sum, which is effectively an expectation of a discrete uniform random variable when scaled by J^2. The final inequality results from Assumptions 1 and 2, and concludes the proof. □

The following lemma is a result taken directly from Dudoit and van der Laan [2005]. It is restated here for convenience.

Lemma 2.2. *Let $Y_1, Y_2, \ldots$ be a sequence of random variables. If $\mathbb{E}[|Y_n|] = O(g(n))$ for some positive function $g(\cdot)$, then $Y_n = O_P(g(n))$.*

Proof. We must show that, for each $\epsilon > 0$, there exists an N and $B > 0$ such that $\mathbb{P}(|Y_n|/g(n) > B) < \epsilon$ for all $n > N$. By assumption, there exists an N and a $C > 0$ such that $\mathbb{E}[|Y_n|]/g(n) < C$ for all $n > N$. By defining $C/B = \epsilon$ and making use of Markov's Inequality, we find that

$$
\mathbb{P}\left(\frac{|Y_n|}{g(n)} > B\right) \leq \frac{\mathbb{E}\left[|Y_n|\right]}{g(n)B} \leq \frac{C}{B} = \epsilon.
$$

□

Having derived Lemma 2.1, the remainder of the proof for Theorem 2.1 closely follows the proof of the first theorem in Dudoit and van der Laan [2005].

Proof. **Theorem 2.1, Finite-Sample Result.**

$$
\begin{aligned}
0 &\leq \tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta_0 \\
&= \mathbb{E}_{B_n}\Bigg[\int (L(x;\hat{\Psi}_{\hat{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP_0(x) \\
&\qquad - (1+\delta)\int L((x;\hat{\Psi}_{\hat{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP^1_{n,B_n}(x) \\
&\qquad + (1+\delta)\int (L(x;\hat{\Psi}_{\hat{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP^1_{n,B_n}(x)\Bigg] \\
&\leq \mathbb{E}_{B_n}\Bigg[\int (L(x;\hat{\Psi}_{\hat{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP_0(x) \\
&\qquad - (1+\delta)\int (L(x;\hat{\Psi}_{\hat{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP^1_{n,B_n}(x) \\
&\qquad + (1+\delta)\int (L(x;\hat{\Psi}_{\tilde{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP^1_{n,B_n}(x)\Bigg] \\
&= \mathbb{E}_{B_n}\Bigg[\int (L(x;\hat{\Psi}_{\hat{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP_0(x) \\
&\qquad - (1+\delta)\int (L(x;\hat{\Psi}_{\hat{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP^1_{n,B_n}(x) \\
&\qquad + (1+\delta)\int (L(x;\hat{\Psi}_{\tilde{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP^1_{n,B_n}(x) \\
&\qquad - (1+2\delta)\int (L(x;\hat{\Psi}_{\tilde{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP_0(x) \\
&\qquad + (1+2\delta)\int (L(x;\hat{\Psi}_{\tilde{k}_{p_n,n}}(P^0_{n,B_n})) - L(x;\psi_0))dP_0(x)\Bigg].
\end{aligned}
\tag{2.18}
$$

The first inequality is by assumption, and the second is by definition of $\hat{k}_{p_n,n}$ such that $\hat{\theta}_{p_n,n}(\hat{k}_{p_n,n}) \leq \hat{\theta}_{p_n,n}(k)\ \forall k$. For simplicity in the remainder of the proof, we replace $\hat{k}_{p_n,n}$ and $\tilde{k}_{p_n,n}$ with $\hat{k}$ and $\tilde{k}$, respectively, in a slight abuse of notation.

Now, let the first two terms of the last expression in Equation (2.18) be denoted by $R_{\hat{k},n}$, and the third and fourth terms by $T_{\tilde{k},n}$. The last term is the cross-validated oracle's risk difference: $(1+2\delta)(\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta_0)$. Thus,

$$
0 \leq \tilde{\theta}_{p_n,n}(\hat{k}) - \theta_0 \leq (1+2\delta)(\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta_0) + R_{\hat{k},n} + T_{\tilde{k},n}. \tag{2.19}
$$

We next show that $\mathbb{E}_{P_0}[R_{\hat{k},n}+T_{\tilde{k},n}] \leq 2c(\delta, \overline{M}(J))(1+\log(K))/(np_n)$, where $c(\delta, \overline{M}(J)) = 2(1+\delta)^2\overline{M}(J)(1/\delta + 1/3)$ for some $\delta > 0$. For convenience, let

$$\begin{aligned}
\hat{H}_k &\equiv \int (L(x; \hat{\Psi}_k(P^0_{n,B_n})) - L(x; \psi_0))dP^1_{n,B_n}(x) \\
\tilde{H}_k &\equiv \int (L(x; \hat{\Psi}_k(P^0_{n,B_n})) - L(x; \psi_0))dP_0(x) \\
R_{k,n}(B_n) &\equiv (1+\delta)(\tilde{H}_k - \hat{H}_k) - \delta\tilde{H}_k \\
T_{k,n}(B_n) &\equiv (1+\delta)(\hat{H}_k - \tilde{H}_k) - \delta\tilde{H}_k,
\end{aligned}$$

so that $R_{k,n} = \mathbb{E}_{B_n}[R_{k,n}(B_n)]$ and $T_{k,n} = \mathbb{E}_{B_n}[T_{k,n}(B_n)]$.

Given B_n and P^0_{n,B_n}, let $Z_{k,i},\ 1 \leq i \leq np_n$, denote the np_n i.i.d. copies of Z_k corresponding with the validation set, that is, with $\{X_i : B_n(i) = 1\}$ (as defined in Lemma 2.1). Then, $\hat{H}_k = \sum_i Z_{k,i}/np_n$ and $\tilde{H}_k = \mathbb{E}_{P_0}[Z_{k,i} \mid P^0_{n,B_n}, B_n]$. Hence, $\tilde{H}_k - \hat{H}_k$ is an empirical mean of np_n i.i.d. centered random variables. Further, by Assumptions 1 and 2, $|Z_{k,i}| < 2(M_1+M_2)^2J^2$ a.s.. Next, we apply Bernstein's Inequality to the centered empirical mean $\tilde{H}_k - \hat{H}_k$, using the property of the $Z_{k,i}$'s derived in Lemma 2.1, to obtain a bound for the tail probabilities of $R_{k,n}(B_n)$ and $T_{k,n}(B_n)$:

$$\sigma_k^2 \equiv \mathrm{Var}_{P_0}\left[Z_k \mid P^0_{n,B_n}, B_n\right] \leq \overline{M}(J)\mathbb{E}\left[Z_k \mid P^0_{n,B_n}, B_n\right] = \overline{M}(J)\tilde{H}_k.$$

Then, for $s > 0$, Bernstein's Inequality yields

$$\begin{aligned}
\mathbb{P}_{P_0}\left(R_{k,n}(B_n) > s \mid P^0_{n,B_n}, B_n\right) &= \mathbb{P}_{P_0}\left(\tilde{H}_k - \hat{H}_k > \frac{s + \delta\tilde{H}_k}{1+\delta}\middle| P^0_{n,B_n}, B_n\right) \\
&\leq \mathbb{P}_{P_0}\left(\tilde{H}_k - \hat{H}_k > \frac{s + \delta\sigma_k^2/\overline{M}(J)}{1+\delta}\middle| P^0_{n,B_n}, B_n\right) \\
&\leq \exp\left\{-\frac{np_n}{2(1+\delta)^2}\frac{\left(s + \delta\sigma_k^2/\overline{M}(J)\right)^2}{\sigma_k^2 + \frac{\overline{M}(J)}{3(1+\delta)}\left(s + \delta\sigma_k^2/\overline{M}(J)\right)}\right\}.
\end{aligned}$$

Note that

$$\frac{\left(s + \delta\sigma_k^2/\overline{M}(J)\right)^2}{\sigma_k^2 + \frac{\overline{M}(J)}{3(1+\delta)}\left(s + \delta\sigma_k^2/\overline{M}(J)\right)} = \frac{s + \delta\sigma_k^2/\overline{M}(J)}{\frac{\sigma_k^2}{s+\delta\sigma_k^2/\overline{M}(J)} + \frac{\overline{M}(J)}{3(1+\delta)}} \geq \frac{s + \delta\sigma_k^2/\overline{M}(J)}{\overline{M}(J)\left(\frac{1}{\delta} + \frac{1}{3}\right)} \geq \frac{s}{\overline{M}(J)\left(\frac{1}{\delta} + \frac{1}{3}\right)}.$$

And so, for $s > 0$,

$$\mathbb{P}_{P_0}\left(R_{k,n}(B_n) > s \mid P^0_{n,B_n}, B_n\right) \leq \exp\left\{-\frac{np_n}{c\left(\delta, \overline{M}(J)\right)}s\right\} \leq K\exp\left\{-\frac{np_n}{c\left(\delta, \overline{M}(J)\right)}s\right\},$$

where $c(\delta, \overline{M}(J)) = 2(1+\delta)^2\overline{M}(J)(\frac{1}{\delta} + \frac{1}{3})$. The same bound applies for the marginal probabilities of $\mathbb{P}_{P_0}(R_{k,n}(B_n) > s)$ since they hold for arbitrary choices of B_n and P^0_{n,B_n}. The second inequality follows from K being larger than or equal to 1 by definition.

Finally, for any $u > 0$, we have by the properties of expectations and the previously derived result that

$$\begin{aligned}
\mathbb{E}_{P_0}[R_{\hat{k},n}] &= \int_0^\infty \mathbb{P}_{P_0}(R_{\hat{k},n} > s)ds - \int_{-\infty}^0 \mathbb{P}_{P_0}(R_{\hat{k},n} \leq s)ds \\
&\leq \int_0^\infty \mathbb{P}_{P_0}(R_{\hat{k},n} > s)ds \\
&\leq u + \int_u^\infty \mathbb{P}_{P_0}(R_{\hat{k},n} > s)ds \\
&\leq u + \int_u^\infty K\exp\left\{-\frac{np_n}{c\left(\delta, \overline{M}(J)\right)}s\right\}ds.
\end{aligned}$$

Since the expression on the right-hand side of the inequality above achieves its minimum value of $c(\delta, \overline{M}(J))(1 + \log(K)/(np_n)$ at $u_n = c(\delta, \overline{M}(J))\log(K)/(np_n)$, then

$$\mathbb{E}_{P_0}[R_{\hat{k},n}] \leq c\left(\delta, \overline{M}(J)\right)\frac{1+\log(K)}{np_n}.$$

The same bound applies to $\mathbb{E}_{P_0}[T_{\tilde{k},n}]$. Therefore, taking the expected values of the inequality in Equation (2.19), we produce the desired finite-sample result in Equation (2.6). □

Proof. **Theorem 2.1, High-Dimensional Asymptotic Result.**

The expected risk differences ratio's convergence follows directly from Equation (2.6) for some $\delta > 0$, so long as $c(\delta, \overline{M}(J))(1 + \log(K))/(np_n\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0]) \to 0$ as $n, J \to \infty$. Given the assumption in Kolmogorov asymptotics that $J/n \to m < \infty$ as $J, n \to \infty$, an equivalent condition is that $m(M_1 + M_2)^2 J(1 + \log(K))/(p_n\mathbb{E}_{P_0}[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0]) \to 0$ as $n, J \to \infty$. Convergence in probability then follows from Lemma 2.2.

□

Though there are minor adaptations to the assumptions to reflect the use of high-dimensional asymptotics, the proof of Corollary 2.1 follows that of Corollary 2.1 in Dudoit and van der Laan [2005].

Proof. **Corollary 2.1.**

The asymptotic statement of Equation (2.10) is an immediate result of Theorem 2.1's Equation (2.8).

$$\frac{\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta_0}{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0}\frac{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0}{\tilde{\theta}_n(\tilde{k}_n) - \theta_0} \xrightarrow{P} 1.$$

Letting $Z_{1,n} \equiv (n(1-p_n))^\gamma(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0)$ and $Z_{2,n} \equiv n^\gamma(\tilde{\theta}_n(\tilde{k}_n) - \theta_0)$, and assuming that the sufficient condition in Equation (2.11) holds, we find that $Z_{1,n}/Z_{2,n} \xrightarrow{d} 1$ by the Continuous Mapping Theorem. Then, notice that

$$\frac{Z_{1,n}}{Z_{2,n}} = \frac{(1-p_n)^\gamma(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_0)}{\tilde{\theta}_n(\tilde{k}_n) - \theta_0},$$

which yields the desired sufficient condition when $p_n \to 0$. In the case of single-split validation, $Z_{2,n} \stackrel{d}{=} Z_{1,n/(1-p_n)}$, and so $Z_{2,n} \xrightarrow{d} Z$ implies that $(Z_{1,n}, Z_{2,n}) \xrightarrow{d} (Z, Z)$.

□

Chapter 3

Treatment Effect Modifier Discovery in Clinical Trials

3.1 Introduction

Precision medicine is now a chief focus of the biomedical establishment. Its promise of tailored interventions and therapies is impossible to overlook, potentially spelling major improvements in patient outcomes [Kraus, 2018, Ginsburg and Phillips, 2018]. Much effort has therefore been invested in the development of quantitative methods capable of uncovering patient sub-populations which benefit more, or less, from novel therapies than the standard of care.

These groups of patients are distinguished from one another based upon diverse biometric measurements referred to as predictive biomarkers [Royston and Sauerbrei, 2008, Kraus, 2018]. Examples include age, sex at birth, ethnicity, and gene expression data taken from various tissue samples. Once identified, these biomarkers may provide clinicians and biologists with mechanistic insight about the disease or therapy, and spur the development of diagnostic tools for targeted treatment regimes.

The statistical discovery of predictive biomarkers has, to date, largely been a byproduct of conditional average treatment effect (CATE) estimation. This typically unknown parameter contrasts the expected outcomes of patients under different treatments as a function of their characteristics, thereby defining the optimal treatment rule. Employing an estimate of the CATE, clinicians can identify a subgroup of patients that draws most benefit from a therapy. When estimated using sparse modelling or otherwise interpretable methods, interpretable machine learning algorithms can be used to find potentially predictive biomarkers [Robins et al., 2008, Tian et al., 2014, Luedtke and van der Laan, 2016, Chen et al., 2017, Zhao et al., 2018, Wager and Athey, 2018, Fan et al., 2020, Bahamyirou et al., 2022, Hines et al., 2022a].

While CATE estimation procedures are demonstrably successful at predictive biomarker discovery in settings where the number of features is small relative to the sample size, it is not so in modern clinical trial in which the number of features frequently exceeds the number

of enrolled patients. The high-dimensional nature of trial data make the CATE estimation problem particularly difficult. Methods proposed for this setting must rely on convenient — and sometimes unverifiable — assumptions about the underlying data-generating process. Examples are sparsity, linear associations, and negligible dependence structures [Tian et al., 2014, Chen et al., 2017, Zhao et al., 2018, Fan et al., 2020, Bahamyirou et al., 2022]. When these assumptions are violated, as is the case when, for example, the set of biomarkers is comprised of gene expression data, the CATE estimate will be biased but may be viable. The biomarkers designated as predictive, however, will likely be false positives (as demonstrated in Section 3.4).

Hines et al. [2022a] recently proposed a collection of variable importance parameters that assess the impact of variables, either individually or in predefined sets, on the variance of the CATE. These parameters are based on popular variable-dropout procedures and on previous work about the variance of the conditional treatment effect [Levy et al., 2021]. While the proposed estimators of these parameters are consistent and asymptotically linear under non-restrictive assumptions about the data-generating process, Hines et al. [2022a] note that quantifying treatment effect modification in this way is misleading when variables are highly correlated. Dropout-based importance metrics may also be deceiving when there are many variables; other features may act as surrogates for the omitted covariate(s) [Hastie et al., 2009, Chap. 15]. This framework is therefore inappropriate for the discovery of predictive biomarkers in high dimensions.

Still other procedures not relying on CATE estimation have recently been proposed. Sechidis et al. [2018] developed an information-theoretic approach for identifying these treatment effect modifiers, though the statistical properties of the procedure are not established and the simulations do not consider high-dimensional data. Zhu et al. [2022] recently developed a penalized linear modelling method for the identification of predictive biomarkers in high dimensions that accounts for the biomarker correlation structure. Like the previous method, however, no formal statistical guarantees are provided.

Myriad methods attempting to identify high-dimensional interactions more generally might also be considered for our task [for example, Hao and Zhang, 2014, Jiang and Liu, 2014, Tang et al., 2020]. They too generally rely on untenable simplifying assumptions about the data-generating process. These include, but are not limited to, assumptions of normality, sparsity of the main effects, sparsity of interaction effects, and bounds on the condition number of the biomarker covariance matrix. Large sets of biomarkers are again unlikely to satisfy such conditions, barring these methods' use for predictive biomarker discovery in high-dimensions.

A simpler alternative is to fit individual (generalized) linear models of the outcome for each biomarker. Each model is comprised of the biomarker's main effect and a treatment-biomarker interaction term. The effect size estimate of the latter serves as a measure importance; larger magnitudes equate to increased treatment effect modification. Hypothesis testing about these treatment-biomarker interaction effects is also possible. As with CATE estimation methods, however, this simple approach imposes stringent parametric conditions on the data-generating process. When the outcome is continuous, for example, inference is

only possible when all marginal biomarker-outcome relationships are truly linear. In small samples, an additional assumption of Gaussian error terms is needed for valid hypothesis testing. Violation of these unrealistic conditions again produces unreliable predictive biomarker identification.

A lack of Type-I error control has marked repercussions in many biomedical applications. In drug target discovery, limited resources are wasted by performing biological follow-up experiments on false positives. In diagnostic development, the inclusion of non-predictive biomarkers may dilute the signal from truly informative ones. In a sequencing-based diagnostic, invalid biomarkers will compete with others for sequencing reads, reducing the sequencing depth and, thereby, the quantification accuracy of predictive biomarkers. These failings have direct, detrimental effects on patient health outcomes.

Motivated by these drawbacks, we present in this work a flexible approach for directly assessing the predictive potential of individual biomarkers. That is, we estimate (a transformation of) each biomarker's *univariate* CATE, a novel variable importance parameter for treatment effect modification. What is more, our procedure permits the formal statistical testing of these biomarkers' predictive effects under non-restrictive assumptions about the underlying data-generating process, and we find that it controls the false discovery rate (FDR) at the nominal level for realistic sample sizes. We also demonstrate on real-world data that our method provides reasonable sub-population identification results when combined with standard clustering approaches.

We emphasize that our framework is not a competitor of treatment rule estimation procedures, it is complementary. The estimation of the CATE and the identification of predictive biomarkers are related but distinct pursuits. To highlight this, we might consider a two-step procedure wherein the full set of biomarkers is filtered using our method, and then the CATE is estimated using the remaining features. The benefits of such a strategy are numerous. The results of the initial stage can help assess whether the assumption of sparsity used by existing methods is tenable, and therefore whether estimating the CATE is feasible. If not, then the ranking of biomarkers might still provide biological or clinical insight, or motivate further study. If so, the CATE may be estimated more accurately, thanks to the reduced number of features considered, using flexible methods like those of Tian et al. [2014], Luedtke and van der Laan [2016], or Wager and Athey [2018]. Further, the rankings generated in the initial stage can impart intuition about the otherwise uninterpretable treatment rule produced by "black-box" methods.

The remainder of the chapter is organized as follows: In Section 3.2, the estimation setting and problem are detailed in statistical terms. Section 3.3 then describes the proposed inferential procedures. The asymptotic behavior of our method is then verified empirically through a comprehensive simulation study in Section 3.4. Application of the proposed approach to clinical trial data then follows in Section 3.5. We end with a brief discussion of the method in Section 3.6. Throughout, we emphasize inference about the univariate CATEs in a randomized control trial setting, though some remarks on its application to observational data are also provided.

3.2 Variable Importance Parameters

Consider n identically and independently distributed (*i.i.d.*) random vectors $X_i = (W_i, A_i, Y_i^{(1)}, Y_i^{(0)}) \sim P_{X,0}$, $i = 1, \ldots, n$, corresponding to complete but unobserved data generated by participants in a randomized control trial or observational study. We drop the indices for notational convenience where possible throughout the remainder of the article. Here, $W = (V, B)$ is a $(q + p)$-length random vector of q pre-treatment covariates, V, like location and income, and p pre-treatment biomarkers, B, such as gene expression data, A is a binary random variable representing a treatment assignment, and $Y^{(1)}$ and $Y^{(0)}$ are random variables corresponding to the potential outcomes of clinical interest under both treatment and control conditions, respectively [Rubin, 1974]. The number of biomarkers p is assumed to be approximately equal to or larger than n. Generally, only one potential outcome is observed per unit. We ignore this point for now, and return to it in the next section.

Clinically relevant predictive biomarkers are often those that have a strong influence on the outcome of interest on the absolute scale. As such, an ideal target of inference when these outcomes are continuous and the number of covariates small is the CATE conditioning on the set of biomarkers:

$$\mathbb{E}_{P_{X,0}}\left[Y^{(1)} - Y^{(0)} \middle| B\right].$$

For reasons previously discussed, however, accurate and interpretable estimation of this parameter is generally challenging when p is large, preventing the accurate recovery of predictive biomarkers.

Indexing the biomarkers of by $j = 1, \ldots, p$, such that $B = (B_1, \ldots, B_p)$, centering them such that $\mathbb{E}_{P_{X,0}}[B_j] = 0$, and assuming that $\mathbb{E}_{P_{X,0}}[B_j^2] > 0$, we instead target the full-data variable importance parameter $\Psi^F(P_{X,0}) = (\Psi_1^F(P_{X,0}), \ldots, \Psi_p^F(P_{X,0}))$ where

$$\Psi_j^F(P_{X,0}) \equiv \frac{\mathbb{E}_{P_{X,0}}\left[\left(Y^{(1)} - Y^{(0)}\right) B_j\right]}{\mathbb{E}_{P_{X,0}}\left[B_j^2\right]}. \tag{3.1}$$

Let $\bar{Q}_{X,0}(A, W) \equiv \mathbb{E}_{P_{X,0}}[Y^{(A)}|W]$. Then this parameter can be represented as

$$\Psi_j^F(P_{X,0}) = \frac{\mathbb{E}_{P_{X,0}}\left[\left(\bar{Q}_{X,0}(1, W) - \bar{Q}_{X,0}(0, W)\right) B_j\right]}{\mathbb{E}_{P_{X,0}}\left[B_j^2\right]}.$$

Assuming the expectation of $\bar{Q}_{X,0}(1, W) - \bar{Q}_{X,0}(0, W)$ conditional on any given B_j is linear in B_j, $\Psi^F(P_{X,0})$ is the vector of simple linear regression coefficients generated by regressing the differences in expected potential outcomes against the individual elements of B. That is, let $f(W) = \bar{Q}_{X,0}(1, W) - \bar{Q}_{X,0}(0, W)$, and assume that $\mathbb{E}_{P_{X,0}}[f(W)|B_j] = \beta_j B_j$. Then,

for $j = 1, \ldots, p$,

$$\begin{aligned}
\Psi_j^F(P_{X,0}) &= \frac{\mathbb{E}_{P_{X,0}}\left[f(W)B_j\right]}{\mathbb{E}_{P_{X,0}}\left[B_j^2\right]} \\
&= \frac{\mathbb{E}_{P_{X,0}}\left[\mathbb{E}_{P_{X,0}}\left[f(W)B_j|B_j\right]\right]}{\mathbb{E}_{P_{X,0}}\left[B_j^2\right]} \\
&= \frac{\mathbb{E}_{P_{X,0}}\left[\beta_j B_j^2\right]}{\mathbb{E}_{P_{X,0}}\left[B_j^2\right]} \\
&= \beta_j \ .
\end{aligned}$$

While the true relationship between the difference of potential outcomes and a predictive biomarker is almost surely nonlinear, $\Psi^F(P_{X,0})$ remains a well-defined parameter and is a generally informative target of inference. Biomarkers with the largest absolute values in $\Psi^F(P_{X,0})$ generally modify the effect of treatment the most.

Analogous simplifications of the high-dimensional regression problem are applicable to other types of outcome variables. For binary outcomes, we might similarly wish to quantify the importance of biomarkers on the absolute risk scale using a slightly modified univariate CATE parameter, $\Psi^{F(\text{binary})}(P_{X,0}) = (\Psi_1^{F(\text{binary})}(P_{X,0}), \ldots, \Psi_p^{F(\text{binary})}(P_{X,0}))$, where:

$$\Psi_j^{F(\text{binary})}(P_{X,0}) \equiv \frac{\mathbb{E}_{P_x}\left[\left(\mathbb{P}_{P_{X,0}}\left[Y^{(1)} = 1 \middle| W\right] - \mathbb{P}_{P_{X,0}}\left[Y^{(0)} = 1 \middle| W\right]\right) B_j\right]}{\mathbb{E}_{P_{X,0}}\left[B_j^2\right]} \tag{3.2}$$

for centered biomarkers $j = 1, \ldots, p$. This is, in fact, the same parameter as $\Psi^F(P_{X,0})$ but presented in a more intuitive form for the binary outcome context: assuming a linear relationship between the difference of the potential outcomes' probability of success and the covariates, this variable importance parameter consists of the simple linear regression coefficients of the difference in the conditional potential outcome success probabilities regressed on each biomarker. Again, the true relationship between the difference of potential outcome probabilities and covariates is unlikely to be linear. Nevertheless, this parameter is telling of biomarkers' predictive capacities.

We stress that the parameters in Equations (3.1) and (3.2) are reasonable approximations of all but pathological treatment effect modification relationships; they summarize the true, marginal functional parameters using interpretable linear models. A case in which $\Psi_j^F(P_{X,0})$ will fail to capture treatment effect modification due to biomarker j is when the $\mathbb{E}_{P_0}[Y^{(1)} - Y^{(0)}|B_j]$ is parabolic: the orthogonal projection of $Y^{(1)} - Y^{(0)}$ onto B_j produces a variable importance parameter value of zero. If such relationships are suspected, however, it suffices to target the corresponding variable importance parameters of the squared biomarkers. Analogous parameters based on transformations of the biomarkers should be considered when the data-generating process is assumed to possess other similarly troublesome nonlinearities.

3.3 Inference

As previously mentioned, only one of the potential outcomes, $Y^{(0)}$ or $Y^{(1)}$, is observed per unit. Instead of $\{X_i\}_{i=1}^n$, we have access to n i.i.d. random observations $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where W and and A are defined as before, and $Y = AY^{(1)} + (1-A)Y^{(0)}$ is a continuous or binary random outcome variable. P_0 is the unknown data-generating distribution of the observed data that is fully determined by $P_{X,0}$ and the (conditional) treatment assignment distribution $g_{A|W}$. That is, P_0 is an element of the nonparametric statistical model $\mathcal{M} = \{P_{P_{X,0},g_{A|W}} : P_{X,0} \in \mathcal{M}_X, g_{A|W}\}$. In a perfect RCT, $g_{A|W} = g_A = \text{Bernoulli}(0.5)$. The challenge therefore lies in estimating the full-data, causal parameter of Equations (3.1) and (3.2) with the observed data; it is generally impossible without making additional assumptions about P_0. We begin by providing such identification conditions.

Throughout the remainder of the chapter, we represent the empirical distribution of P_0 by P_n, the conditional outcome regression function by $\bar{Q}_0(A, W) \equiv \mathbb{E}_{P_0}[Y|A, W]$, and the treatment assignment mechanism by $g_0(W) \equiv \mathbb{P}_{P_0}[A = 1|W]$. Where possible, we simplify notation further by writing $\bar{Q}_0(A, W)$ and $g_0(W)$ as $\bar{Q}_0$ and g_0, respectively. All proofs are provided in Section 3.7.

Assumption 3.1. *No unmeasured confounding:* $Y^{(a)} \perp A|W$ *for* $a = \{0, 1\}$.

Assumption 3.2. *Positivity: There exists some constant* $\epsilon > 0$ *such that* $\mathbb{P}_{P_0}[\epsilon < g_0(W) < 1 - \epsilon] = 1$.

A3.1 assures that there are no unmeasured confounders of treatment and outcome, allowing for treatment allocation to be viewed as the product of a randomized experiment. A3.2 is an overlapping support condition stating that all observations may be assigned to either treatment condition regardless of covariates. These conditions, regularly cited in the causal inference literature, are generally satisfied in randomized control trials. Altogether, they lead to the following result:

Theorem 3.1. *Under the conditions of A3.1 and A3.2, letting* $\mathbb{E}_{P_0}[B_j] = 0$*, and assuming that* $\mathbb{E}_{P_0}[B_j^2] > 0$*,*

$$\begin{aligned} \Psi_j(P_0) &\equiv \frac{\mathbb{E}_{P_0}\left[\left(\bar{Q}_0(1, W) - \bar{Q}_0(0, W)\right) B_j\right]}{\mathbb{E}_{P_0}\left[B_j^2\right]} \\ &= \Psi_j^F(P_{X,0}) \end{aligned} \tag{3.3}$$

for $j = 1, \ldots, p$ *such that* $\Psi(P_0) = (\Psi_1(P_0), \ldots, \Psi_p(P_0)$.

Having established the conditions under which $\Psi^F(P_{X,0})$ can be estimated from the observed data, we now focus on inference about $\Psi(P_0)$. Define the Augmented Inverse Probability Weight (AIPW) [Robins et al., 1995] transform as

$$T_a(O; \bar{Q}_0, g_0) = \frac{I(A = a)}{I(a = 1)g_0(W) + I(a = 0)(1 - g_0(W))}(Y - \bar{Q}_0(A, W)) + \bar{Q}_0(a, W), \tag{3.4}$$

and let $\tilde{T}(O; P_0) = T_1(O; \bar{Q}_0, g_0) - T_0(O; \bar{Q}_0, g_0)$.

Theorem 3.2. *The efficient influence function (EIF) of $\Psi_j(P)$ for $P \in \mathcal{M}$ and $j = 1, \ldots, p$ is given by*

$$D_j(O; P) \equiv \frac{\left(\tilde{T}(O; P) - \Psi_j(P) B_j\right) B_j}{\mathbb{E}_P\left[B_j^2\right]}. \tag{3.5}$$

The EIF of Equation (3.5) informs the construction of nonparametric efficient estimators of $\Psi_j(P_0)$ under non-restrictive assumptions about the data-generating process [Bickel et al., 1993a, Hines et al., 2022b]. Many approaches exist for deriving these efficient estimators, such as one-step estimation [Pfanzagl and Wefelmeyer, 1985, Bickel et al., 1993a], estimating equations [van der Laan and Robins, 2003a, Chernozhukov et al., 2017, 2018], or targeted maximum likelihood estimation [van der Laan and Rubin, 2006a, van der Laan and Rose, 2011a, 2018a]. We use the, in this case, straightforward method of estimating equations. The resulting estimator is intuitive: it corresponds to the estimator of the simple linear regression coefficient of centered biomarker j regressed on the adjusted predicted differences in potential outcomes. Further, it is identical to the one-step estimator.

Corollary 3.1. *Let P_m be the empirical distribution of another dataset of m random observations distributed according to P_0 and distinct of P_n. If such a dataset is not available, it might be generated using sample-splitting techniques. We require that the size of P_m grows linearly with the size of P_n. That is, $O(m) = O(n)$. This is trivially accomplished when using most sample-splitting frameworks, like K-fold cross-validation. Then define $\bar{Q}_m$ and g_m as estimates of the nuisance parameters $\bar{Q}_0$ and g_0 fit to P_m. The estimating equation estimator of $\Psi_j(P_0)$ is then given by:*

$$\Psi_j^{(ee)}(P_n; P_m) = \frac{\sum_{i=1}^n \tilde{T}(O_i; P_m) B_{ij}}{\sum_{i=1}^n B_{ij}^2}, \tag{3.6}$$

where we again assume that the biomarkers are centered such that $\sum B_{ij} = 0$. This estimator is double robust.

The double-robustness property signifies that $\Psi_j^{(\text{ee})}(P_n; P_m)$ is a consistent estimator of $\Psi_j(P_0)$ so long as either the estimator of the conditional expectation or the estimator of the propensity score are consistent. In particular, when g_0 is known, as in most clinical trials, it is guaranteed to be consistent.

Under the following conditions, we can detail this estimator's limiting distribution.

Assumption 3.3. *Known treatment assignment mechanism: g_0 is known.*

Assumption 3.4. *Nuisance parameter estimator convergence: Let*

$$\|\bar{Q}_m(A, W) - \bar{Q}_0(A, W)\|_2^2 \equiv \int (\bar{Q}_m(a, w) - \bar{Q}_0(a, w))^2 dP_0(a, w) \ .$$

and

$$\|g_m(W) - g_0(W)\|_2^2 \equiv \int (g_m(w) - g_0(w))^2 dP_0(w) \; .$$

Then it must be that $\|\bar{Q}_m - \bar{Q}_0\|_2 \; \|g_m - g_0\|_2 = o_P(n^{-1/2})$*, where* $\|\cdot\|_2$ *denotes the* $L_2(P_0)$ *norm.*

Theorem 3.3. *If A3.3 or A3.4 are satisfied and* $\mathbb{E}_{P_0}[B_j^2] > 0$ *for* $j = 1, \ldots, p$*, then*

$$\sqrt{n}\left(\Psi_j^{(ee)}(P_n; P_m) - \Psi_j(P_0)\right) \xrightarrow{D} N\left(0, \mathbb{V}_{P_0}\left[D_j(O; P_0)\right]\right). \tag{3.7}$$

Again, A3.3 is generally satisfied in clinical trials, implying that the estimating equation estimator of Equation (3.6) is asymptotically linear. Valid hypothesis testing is possible even when the conditional outcome regression is biased. This results from the form of the EIF, and is discussed in the proof (Section 3.7).

In observational settings, A3.4 requires that the conditional outcome regression estimates and the treatment assignment rule estimates converge in probability to their respective true parameters at a rate faster than $n^{-1/4}$. When the number of biomarkers and covariates is moderate relative to sample size, these conditions are typically satisfied by estimating these parameters using flexible machine learning algorithms [van der Laan and Rose, 2011a] like the Super Learner of van der Laan et al. [2007a]. Relying on the general asymptotic theory of cross-validated loss-based estimation [van der Laan and Dudoit, 2003a], the Super Learner method constructs a convex combination of estimators from a pre-specified library that minimizes the cross-validated risk of a pre-defined loss function. Even in a high-dimensional setting where the number of biomarkers is far larger than n, recent results about Random Forests [Wager and Athey, 2018] and deep neural networks [Farrell et al., 2021] suggest conditions for which A3.4 is satisfied. Generally, fast convergence of these estimators in high dimensions requires strong smoothness and sparsity assumptions about the underlying parameters [Hines et al., 2022b].

Under A3.1, A3.2, and either A3.3 or A3.4, Theorem 3.3 delivers the means by which to construct α-level Wald-type confidence intervals for $\Psi_j(P_0)$. However, the estimator of Equation (3.6) and any accompanying testing procedure require that the nuisance parameters be estimated on a separate dataset.

Since practitioners rarely have access to two datasets from the same data-generating process, we propose a cross-validated estimator that uses all available data. Begin by randomly partitioning the n observations of P_n into K independent validation sets $P_{n,1}^1, \ldots, P_{n,K}^1$ of approximately equal size. For $k = 1, \ldots, K$, define the training set as, in a slight abuse of notation, $P_{n,k}^0 = P_n \setminus P_{n,k}^1$. Then the cross-validated estimator of $\Psi_j(P_0)$ is defined as:

$$\Psi_j^{(\text{CV})}(P_n) = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{i=1}^n I(O_i \in P_{n,k}^1)\tilde{T}(O_i; P_{n,k}^0)B_{ij}}{\sum_{i=1}^n I(O_i \in P_{n,k}^1)B_{ij}^2}, \tag{3.8}$$

and has the same limiting distribution as $\Psi_j^{(\text{ee})}(P_n; P_m)$ under conditions consistent with those of either A3.3 or A3.3. The accompanying cross-validated estimator of the EIF's standard

deviation for biomarker j is given by

$$\sigma_j^{(\mathrm{CV})}(P_n) = \left(\frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{\sum_{i=1}^{n} I(O_i \in P_{n,k}^1)} \sum_{i=1}^{n} I(O_i \in P_{n,k}^1) \left(D_j(O_i; P_{n,k}^0) \right)^2 \right] \right)^{1/2}.$$

The α-level Wald-type confidence intervals for $\Psi_j(P_0)$ are then constructed as

$$\Psi_j^{(\mathrm{CV})}(P_n) \pm \frac{z_{(1-\alpha/2)} \sigma_j^{(\mathrm{CV})}(P_n)}{\sqrt{n}},$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)^{\text{th}}$ quantile of a standard Normal distribution. Inference about $\Psi_j(P_0)$ is therefore made possible under non-restrictive assumptions about the data-generating process when using data-adaptive methods and cross-validation to the estimate nuisance parameters. Even in small samples where the limiting properties of $\Psi_j^{(\mathrm{ee})}(P_n; P_m)$ might not be attained, the generalized variance moderation technique of Hejazi et al. [2023] can be used for Type-I error control.

In summary, we take as target of inference the simple linear regression slopes of the difference of predicted outcomes under treatment and control conditions regressed on each biomarker. We suggest that these parameters be estimated by learning the conditional outcome regression and treatment assignment rule (if necessary), using them to predict the potential outcomes, and then fitting simple linear regressions to the difference in predicted potential outcomes as a function of each centered biomarker. Under the conditions outlined in A3.1, A3.2, and A3.3 or A3.4, we prove that our estimator targets the causal parameter of interest and is asymptotically linear, providing a straightforward statistical test to assess whether a biomarker modifies the treatment effect. Even when the causal inference conditions of A3.1 and A3.2 are not satisfied, $\Psi(P_0)$ remains an interpretable statistical parameter. It captures the strength of treatment-biomarker interactions in high dimensions, and inference about it can be performed using the same cross-validated procedure.

3.4 Simulation Study

Details

This work is motivated by the need to identify predictive biomarkers in clinical trials. Of particular interest is their detection for drug target discovery and diagnostic assay development. The former requires the identification of biomarkers causally related to the outcome of interest, whereas the latter seeks a small set of strongly predictive biomarkers. We therefore focus on these applications throughout the simulation study.

A varied collection of data-generating processes, defined below, are considered to demonstrate that the theoretical guarantees outlined in the previous section are achieved for a range of functional forms of the conditional outcome regression. Recall that Y corresponds to the outcome, A the treatment assignment, W the covariates, and B the biomarkers, a subset of the covariates. The treatment assignment rules, g_0, are treated as known.

- Class 1: Moderate dimensions, non-sparse treatment-biomarker effects vector with independent biomarkers

  - Linear conditional outcome regression:

$$\begin{aligned} W = B &\sim N(0, I_{100\times 100}) \\ A|W = A &\sim \text{Bernoulli}(1/2) \\ Y|A, W &\sim N\left(W^\top\left(\beta + I(A=1)\gamma^{(1)} + I(A=0)\gamma^{(0)}\right), 1/2\right). \end{aligned}$$

    Here, $\beta = (\beta_1, \ldots, \beta_{100})^\top$ such that $\beta_1 = \ldots = \beta_{20} = 2$ and $\beta_{21} = \ldots = \beta_{100} = 0$, and $\gamma^{(a)} = (\gamma_1^{(a)}, \ldots, \gamma_{100}^{(a)})^\top$ where $\gamma_1^{(1)} = \ldots = \gamma_{50}^{(1)} = 5$, $\gamma_1^{(0)} = \ldots = \gamma_{50}^{(0)} = -5$ and $\gamma_{51}^{(1)} = \ldots = \gamma_{100}^{(1)} = 0$ for $a = \{0, 1\}$.

  - Kinked conditional outcome regression: W and A are distributed as above. The conditional outcome is defined as

$$Y|A, W \sim N\left(W^\top\left(I(A=1)\gamma + I(A=0)\,\text{diag}(I(W>0))\,\gamma\right), 1/2\right),$$

    where $\gamma = (\gamma_1, \ldots, \gamma_{100})$, $\gamma_1 = \ldots = \gamma_{50} = 10$, $\gamma_{51} = \ldots = \gamma_{100} = 0$, and $\text{diag}(\cdot)$ is a diagonal matrix whose diagonal equals the input vector.

  - Nonlinear conditional outcome regression: W and A are distributed as above. Then,

$$Y|A, W \sim N\left(\exp\left\{|W^\top\beta|\right\} + I(A=1)W^\top\gamma, 1/2\right),$$

    where $\beta_1 = \ldots = \beta_{20} = 1$ and $\beta_{21} = \ldots = \beta_{100} = 0$, and where $\gamma_1 = \ldots = \gamma_{50} = 5$ and $\gamma_{51} = \ldots = \gamma_{100} = 0$.

- Class 2: High dimensions, sparse treatment-biomarker effects vector with correlated biomarkers

  - Linear conditional outcome regression:

$$\begin{aligned} C &\sim \text{Bernoulli}(1/2) \\ W|C = B|C &\sim N(-I(C=0) + I(C=1), \Sigma_{500\times 500}) \\ A|W = A &\sim \text{Bernoulli}(1/2) \\ Y|A, W &\sim N\left(W^\top\left(\beta + I(A=1)\gamma\right), 1/2\right) \end{aligned}$$

    Here, C is an unobserved subgroup indicator, $\beta = (2, 2, 2, 2, 2, 0, \ldots, 0)$, and $\gamma = (5, 5, 5, 5, 0, \ldots, 0)$. The biomarker covariance matrix, Σ, is the estimated gene expression correlation matrix of the 500 most variable genes taken from the tumours of patients with metastatic or recurrent colorectal cancer [Watanabe et al., 2011]. These genes were first clustered using hierarchical clustering based on their Euclidean distance with complete linkage, and the correlation matrix

was then estimated using the cross-validated estimation procedure of Chapter 2 [Boileau et al., 2021b] implemented in the `cvCovEst R` package [Boileau et al., 2021a, R Core Team, 2022], relying on the banding and tapering estimators of Bickel and Levina [2008a] and Cai et al. [2010a], respectively. The gene expression data has been made available by the Bioconductor [Huber et al., 2015] experiment package `curatedCRCdata` [Parsana et al., 2021].

– Kinked conditional outcome regression: C, W and A are distributed as above. The conditional outcome distribution is as follows:

$$Y|A,W \sim N\left(W^\top \left(I(A=1)\gamma + I(A=0)\,\text{diag}(I(W>0))\,\gamma\right), 1/2\right),$$

where $\gamma = (10, 10, 10, 10, 0, \ldots, 0)$.

– Nonlinear conditional outcome regression: C, W and A are distributed as above. Then,

$$Y|A,W \sim N\left(\exp\left\{|W^\top\beta|\right\} + I(A=1)W^\top\gamma, 1/2\right),$$

where $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)$ and $\gamma = (5, 5, 5, 5, 0, \ldots, 0)$.

The first class of data-generating processes reflects the scenario in which a set of biomarkers known to be associated with the outcome, perhaps based on prior clinical or biological investigations, are assessed for potential treatment-biomarker interactions. Since they have been cherry-picked, a reasonable assumption is that a non-negligible proportion of these biomarkers modify the effect of treatment on the outcome of interest. The second set of data-generating processes is representative of exploratory scenarios wherein a vast number of biomarkers, like tumor gene expression data collected prior to the start of treatment, are explored for strong effect modification. Further, these data-generating processes contain two subgroups, representing, for example, unknown patient subpopulations in a clinical trial. These models each possess four non-zero treatment-biomarker interactions in the leading entries of γ or $\gamma^{(0)}$ and $\gamma^{(1)}$. The biomarkers that modify the treatment effect are correlated mimicking a small gene set.

The collections of moderate and high-dimensional data-generating processes each contain three outcome regression models. Their sketches are provided in Figure 3.1. The simplest "linear" models correspond to the functional form assumed by many existing high-dimensional CATE estimation procedures for a continuous outcome [Tian et al., 2014, Chen et al., 2017, Zhao et al., 2018, Ning et al., 2020]. The "kinked" data-generating processes are named so for the kink in the marginal conditional outcome regression of its predictive biomarkers. These marginal relationships are representative of predictive biomarkers in clinical trials assessing the efficacy of the standard of care against combinations of the standard of care and another drug, and where the treatment group outperforms the control group in all biomarker defined subpopulations, but with different treatment effect sizes. Finally, the "nonlinear" data-generating mechanisms represent those whose conditional outcome regressions deviate most from assumptions of linearity. We expect these to pose the greatest

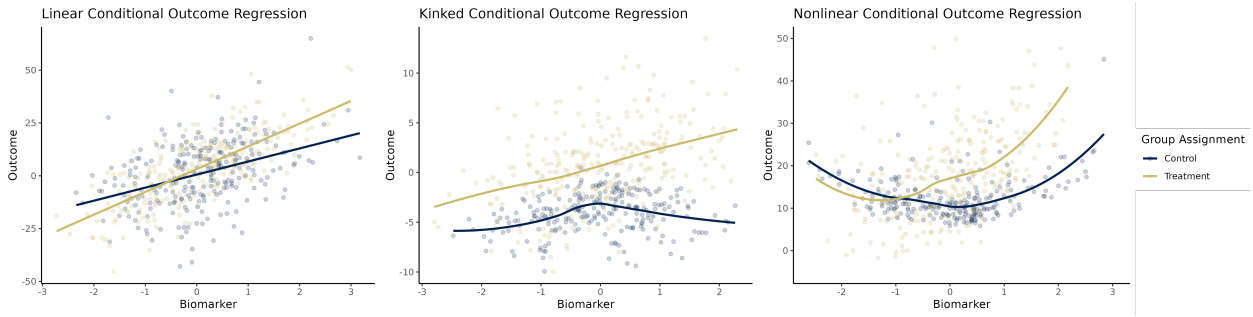


Figure 3.1: Sketches of predictive biomarkers' marginal relationships with the outcome variable for the considered conditional outcome regression models.

challenge with respect to identifying predictive biomarkers. We note that the linear conditional outcome regression models are not identifiable, but this is not a concern for generative purposes.

Two hundred datasets of 125, 250, and 500 observations were generated for each of these data-generating processes — 3,600 in all — by sampling without replacement from simulated populations of 100,000 observations. Each model's $\Psi(P_0)$ was computed from its respective population. These random samples and estimands are used in the following subsections to assess the finite sample performance of our proposed procedure and to benchmark its ability to discover predictive biomarkers against that of popular CATE estimation methods.

The cross-validated estimator of Equation (3.8) is used to estimate the vector of univariate CATE simple linear regression coefficients in the simulated datasets using 5-fold cross-validation. Throughout the remainder of the chapter, we refer to our proposed method as *uniCATE*. van der Laan et al. [2007a]'s Super Learner procedure is used to estimate the conditional outcome regressions. The library of candidate algorithms is made up of ordinary linear, LASSO, and elastic net regressions [Tibshirani, 1996, Zou and Hastie, 2005], polynomial splines [Stone et al., 1997], XGboost [Chen and Guestrin, 2016], Random Forests [Breiman, 2001], and the mean model.

Bias and Variance of Univariate CATE Estimator

The theoretical results of Section 3.3 are asymptotic, yet many clinical trials are made up of a small to moderate numbers of participants. We therefore verify that uniCATE's estimates, metrics of biomarkers' predictive importance, are accurate when computed under realistic sample sizes. We computed the empirical bias and variance of the cross-validated estimator when applied to each data-generating process and at each simulated sample size. The results of our analysis of the nonlinear models are presented in Figure 3.2. Those of the linear and kinked models, presented in Figures 3.3 and 3.4, respectively, are virtually identical.

We find that uniCATE is approximately unbiased across sample sizes regardless of the conditional outcome regressions' complexities, as suggested by Theorem 3.3. However, the

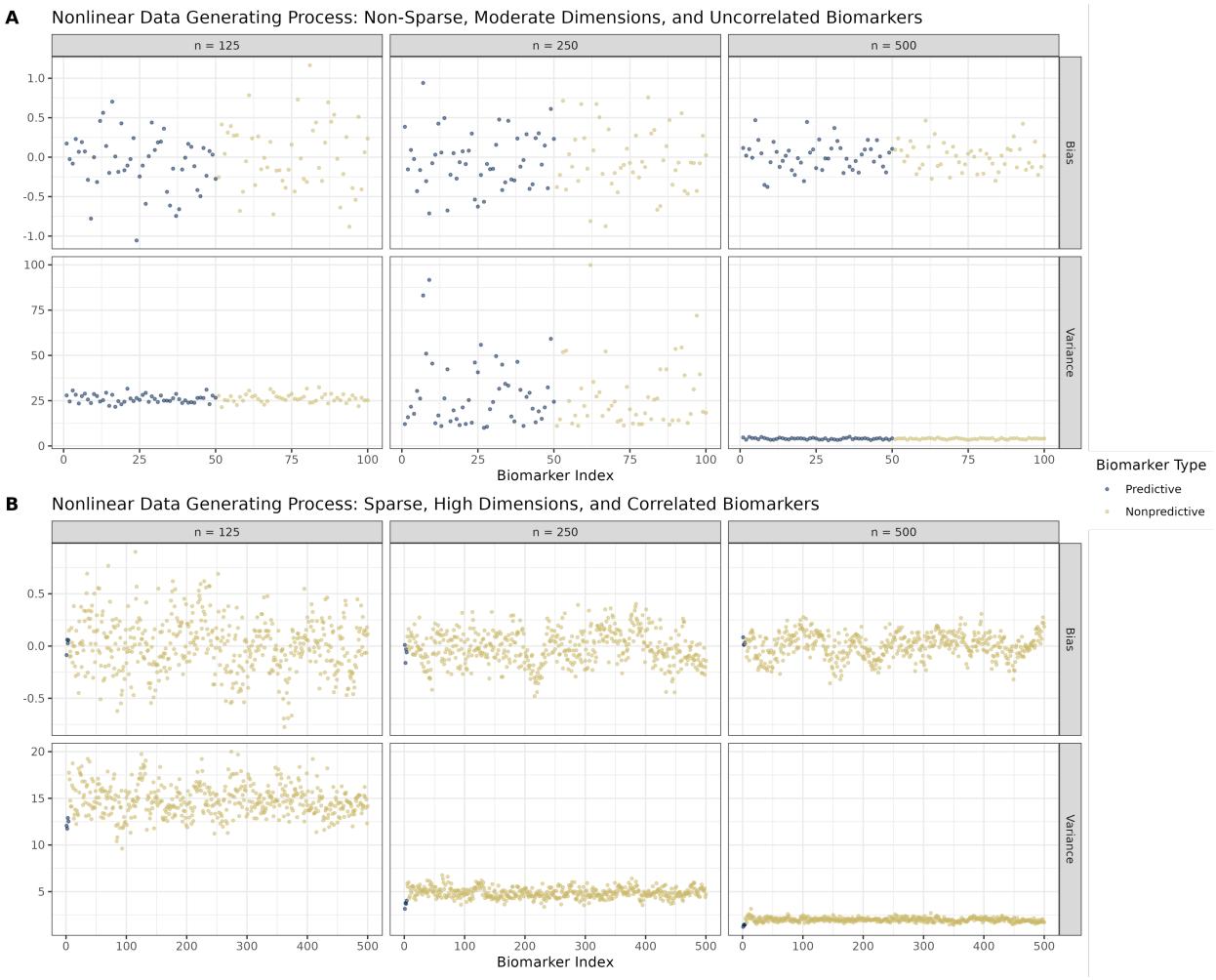


Figure 3.2: The empirical biases and variances of uniCATE estimates for all biomarkers across all simulation scenarios with a nonlinear conditional outcome regression. Biomarkers colored blue are truly predictive, and those colored gold are nonpredictive.

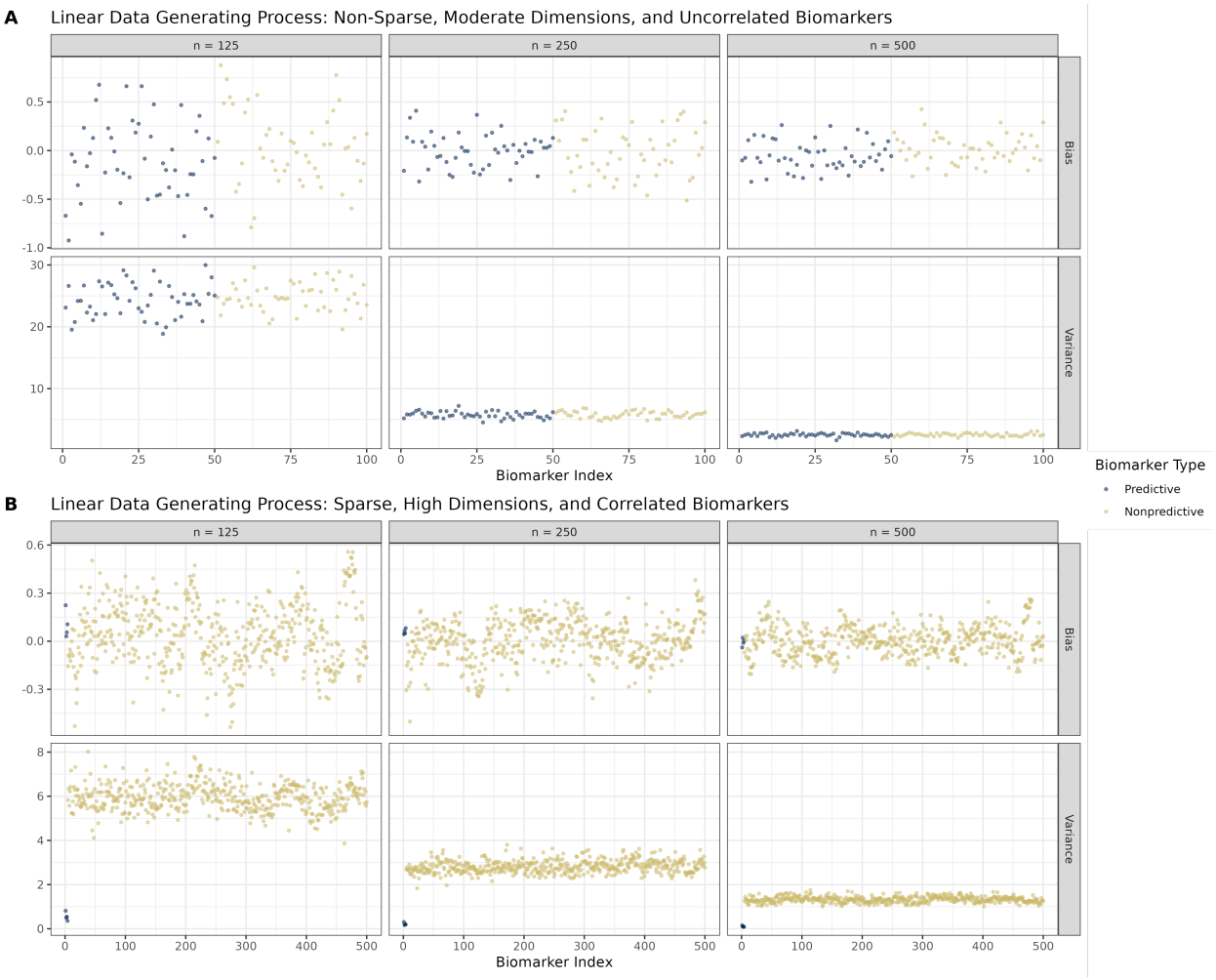


Figure 3.3: The empirical biases and variances of uniCATE estimates for all biomarkers across all simulation scenarios with a linear conditional outcome regression. Biomarkers coloured blue are truly predictive and those coloured gold are nonpredictive.

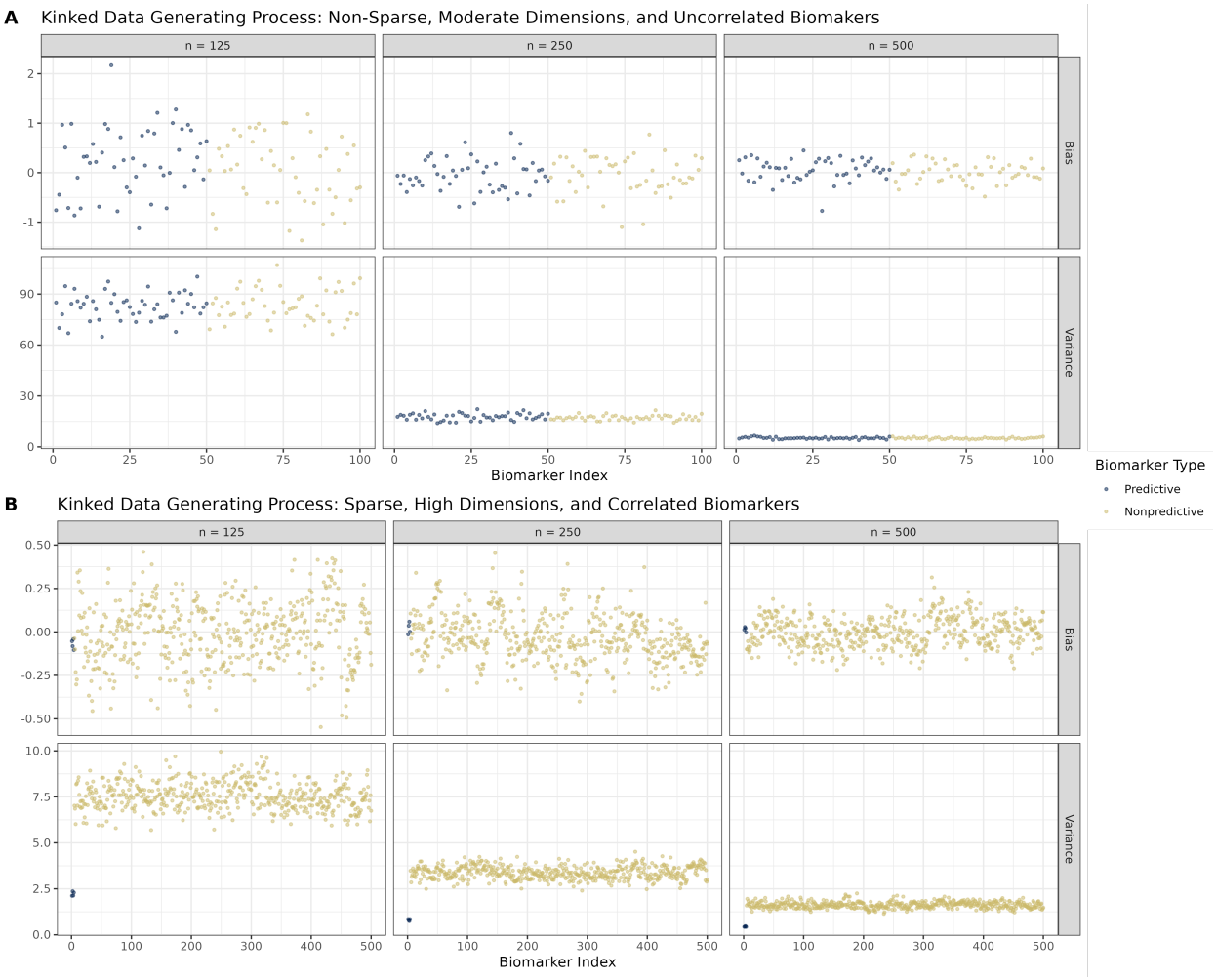


Figure 3.4: The empirical biases and variances of uniCATE estimates for all biomarkers across all simulation scenarios with a kinked conditional outcome regression. Biomarkers coloured blue are truly predictive and those coloured gold are nonpredictive.

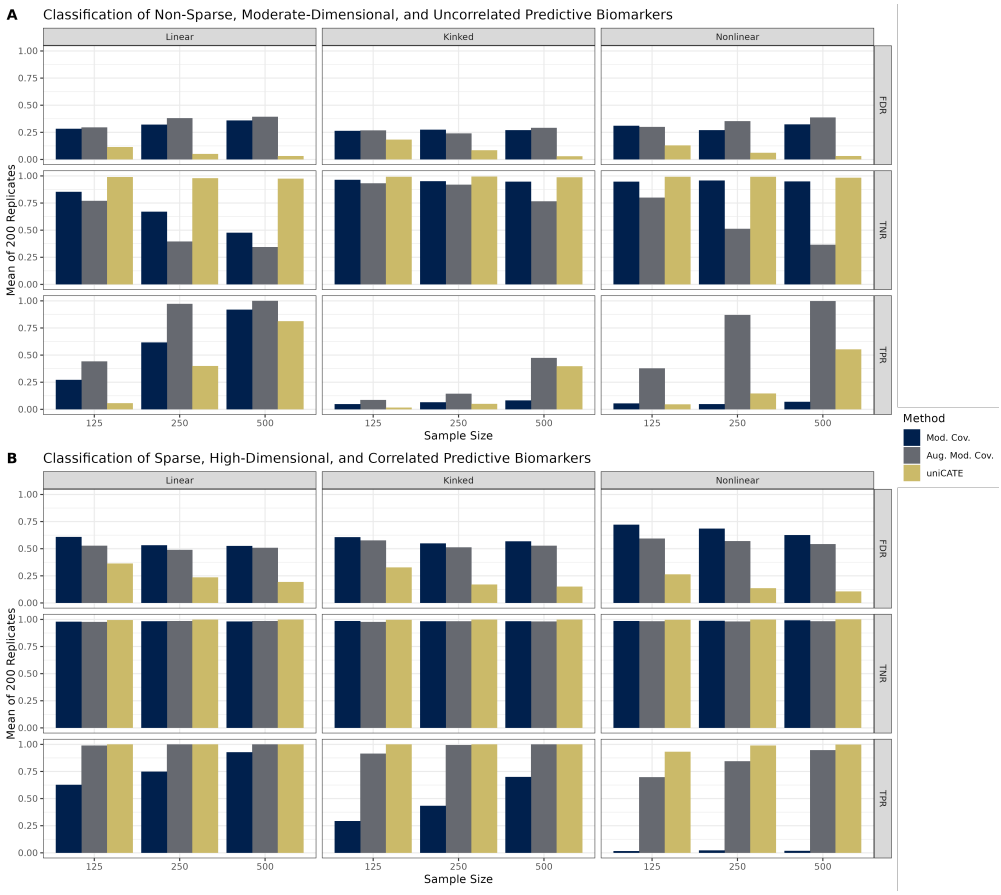


Figure 3.5: The empirical predictive biomarker classification results for the moderate dimensions, non-sparse treatment-biomarker interaction settings with uncorrelated biomarkers (A) and the high-dimension, sparse treatment-biomarker interaction settings with correlated biomarkers(B).

estimator is highly variable in the moderate dimension, non-sparse (e.g. Figure 3.2A) scenarios when $n = 125$ and 250, and somewhat variable when $n = 125$ in the high dimension, sparse data-generating processes (e.g. Figure 3.2B). As expected, the empirical variance of the estimator decreases drastically in all simulation settings as sample sizes increase.

This is encouraging for diagnostic biomarker assay development: the ranking of predictive biomarkers reported by uniCATE is reliable under realistic sample sizes and data-generating processes. These results suggest that our method accurately and precisely evaluates biomarkers with respect to their predictive abilities when the number of truly predictive biomarkers is small in samples possessing as few as 250 observations. Similar behavior is observed when there are a large number of predictive biomarkers in trials of 500 subjects or more.

Type-I Error Control

In addition to evaluating the accuracy of uniCATE's estimates, we assess the method's ability to distinguish predictive biomarkers from non-predictive biomarkers. This is of particular importance in applications requiring the reduction of the pool of potential predictive biomarkers, as in the development of diagnostic assays, or generating hypotheses for biological and clinical validation in drug target discovery. We therefore evaluate uniCATE's Type-I error rate control across the simulation scenarios using a target FDR [Benjamini and Hochberg, 1995] of 5%. The inferential procedure described in Section 3.3 is used to test whether predictive biomarkers' linear approximations of the univariate CATE are significantly different from zero. Nominal p-values are adjusted using the FDR-controlling procedure of Benjamini and Hochberg [1995]. We note that nominal FDR control is not guaranteed by this adjustment method in the high-dimensional simulations because of the biomarkers' correlation structure. The results are presented in Figure 3.5.

Our method's capacity to identify predictive biomarkers was compared to that of popular CATE estimation methods: the modified covariates approach and its augmented counterpart [Tian et al., 2014, Chen et al., 2017]. Briefly, the former directly estimates the linear model coefficients of the treatment-biomarker interactions, using a linear working model for these terms, without having to model or estimate the main effects. While Tian et al. [2014]'s method is flexible since it avoids making any assumptions about the functional form of the main biomarker effects, it can lack precision in small-sample, high-dimensional settings. Tian et al. [2014] and Chen et al. [2017] therefore proposed "augmented" versions of this method that explicitly account for this source of variation. While Tian et al.'s [2014] and Chen et al.'s [2017] augmentation procedures differ, they are identical in the randomized control trials with continuous outcome variables [Chen et al., 2017]: they are equivalent to fitting a (penalized) multivariate linear regression with treatment-biomarker interaction terms.

We again emphasize that these methods are not true competitors of our procedure. Their primary goal, CATE estimation, differs from that of uniCATE. However, Tian et al. [2014] and Chen et al. [2017] demonstrated that the (augmented) modified covariates approach could identify potentially predictive biomarkers when fit using regularized linear regressions like the LASSO. Biomarkers with non-zero treatment-biomarker interaction coefficient estimates are classified as predictive. We therefore applied these approaches, using 10-fold cross-validation to select the LASSO hyperparameters, to all simulated datasets. The implementations of these estimators provided by the `personalized` `R` software package [Huling and Yu, 2021] was used.

The results pertaining to the moderate dimension simulations ($p = 100$) with 50 predictive biomarkers are presented in Figure 3.5A. Only uniCATE is capable of controlling the Type-I error rate; it approximately achieves the nominal FDR of 5% in all settings with samples sizes of 250 and above. The modified covariates approach and its augmented counterpart possess FDRs no lower than 25% across all scenarios. Indeed, their control of Type-I error generally worsens as sample size increases. Our method's superior performance with respect to FDR control is likely due to its conservativeness: many of the predictive biomark-

ers are not recognized in smaller sample-size settings. As sample size grows, however, so too does its true positive rate (TPR) while maintaining a near perfect true negative rate (TNR). When $n = 500$, uniCATE generally identifies close to or more predictive biomarkers than the modified covariates approach, and nearly as many as the augmented modified covariates method.

Our procedure's performance with respect to FDR control is again superior to that of the CATE estimation approaches in the high dimensional simulation scenarios with 500 biomarkers (Figure 3.5B). While the adjustment procedure of Benjamini and Hochberg [1995] does not guarantee FDR control at the desired rate in these scenarios due to the correlation structure of the tests, it is nearly achieved in larger sample sizes. uniCATE also marginally outperforms other approaches in terms of the TNR. Unlike in the moderate sample-size simulations, however, our method identifies predictive biomarkers more efficiently than the other procedures considered.

These results demonstrate that uniCATE recovers truly predictive biomarkers more reliably than interpretable treatment rule estimators. In most simulation scenarios, uniCATE provides well controlled Type-I error rates while its TNR and TPR are comparable or superior to other methods. However, when the number of truly predictive biomarkers is large and the sample size small, uniCATE's biomarker classification will be conservative. In this setting, our method still provides good Type-I error control, limiting the waste of resources on the investigation of false positives, as would be the result if using existing methods. If the investigator prefers a less conservative approach, since, for example, the cost of follow-up experiments is low, the (augmented) modified covariate approach may be considered instead.

3.5 Application to IMmotion Trials

Until recently, Tyrosine kinase inhibitors targeting vascular endothelial growth factor (VEGF) were the standard of care for patients with metastatic renal cell carcinoma (mRCC) [Rini et al., 2019]. Unfortunately, many patients with mRCC find these treatments, like sunitinib, ineffective, and most develop a resistance over time [Rini and Atkins, 2009]. Immune checkpoint inhibitors like atezolizumab can produce more durable results and improve overall survival in pre-treated patients with mRCC [Motzer et al., 2015a,b, McDermott et al., 2018]. A combination of atezolizumab and bevacizumab, the latter of which also binds to VEGF, was shown to improve the objective response rate (ORR) in a Phase 1b study [Wallin et al., 2016]. Objective response is a binary indicator of clinically meaningful response to treatment. These findings were supported by a Phase 2 study, IMmotion 150, which compared atezolizumab alone and in combination with bevacizumab against sunitinib [McDermott et al., 2018]. In a subsequent Phase 3 study, IMmotion 151 [Rini et al., 2019], the atezolizumab and bevacizumab combination improved progression free survival and objective response over sunitinib in patients whose cancer cells expressed the programmed death-1 ligand 1 (PD-L1), but not all of these patients showed benefit. These results motivate the search for biomarkers that are more predictive of this treatment's clinical benefit than PD-L1

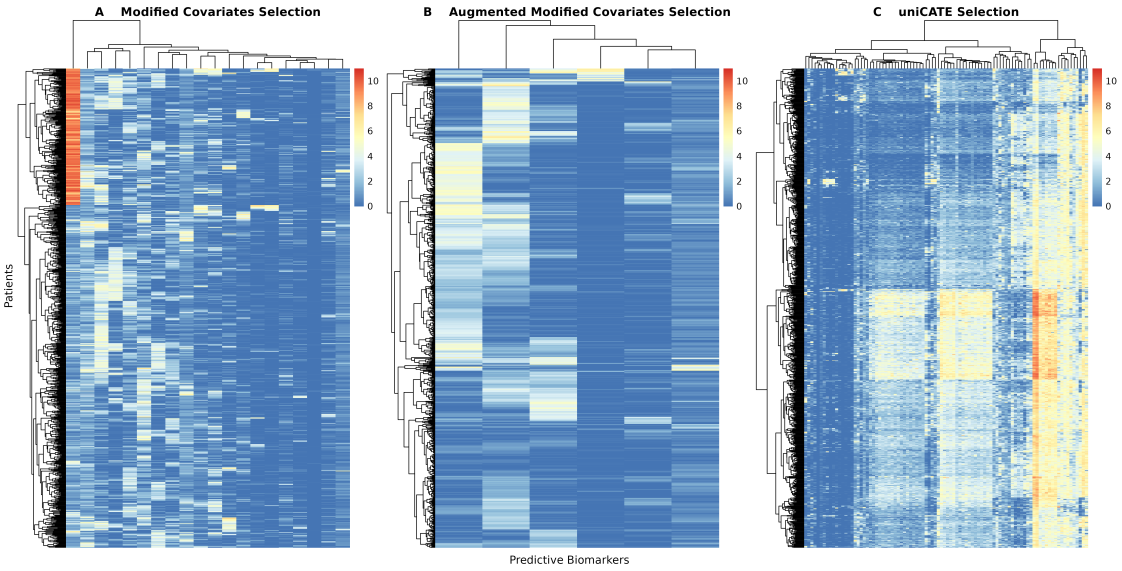


Figure 3.6: Heatmaps of the modified covariates approach's (A), augmented modified covariates approach's (B), and uniCATE procedure's (C) predictive biomarkers' log-transformed gene expression data from the IMmotion151 trial. Rows and columns are ordered via hierarchical clustering with complete linkage and Euclidean distance.

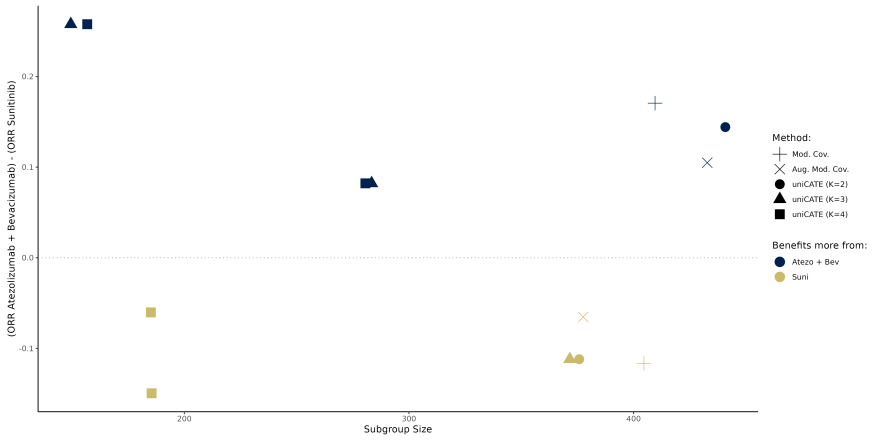


Figure 3.7: Comparison of the ORR across the methods' predicted subgroups in the IMmotion151 trial. The hierarchical clustering with complete linkage and Euclidean distance applied to uniCATE's predictive biomarkers was used to iteratively define two, three, and four clusters (K). The points are slightly horizontally jittered along the x-axis to avoid overplotting.

Method	Predictive Biomarkers
Modified Covariates	ADCY8, CDH17, COL6A6, CSMD3, CXCL5, EEF1A2, GJB6, GRIA4, H19, IGKV1-9, KLK4, MMP3, MUC17, PZP, TCHH, TEX15, TRIM63, VIL1, WFIKKN2, XIST
Augmented Modified Covariates	EEF1A2, IGKV1-9, MMP3, PZP, TEX15, TRIM63
uniCATE	WFIKKN2, NMRK2, KLK1, TRIM63, IGKV1-9, HHATL, UCHL1, CLDN1, EEF1A2, C8A, KCNJ3, ITIH2, IGLV3-21, TCHH, ATP1A3, IGLL5, ENPP3, IGKV3-15, IGLC3, SAA1, TEX15, IGKV1-16, IGKV1-5, IGHG1, GRIN2A, IGHV2-5, SERPIND1, IGHV1-18, DEFB1, CYP2J2, IGHV1-24, CES3, IGKV3-11, IGLV1-40, IGHV1-2, SLC17A4, KLK4, MMP7, ANKRD36BP2, IGHV3-11, IGHV4-31, IGHV4-34, IGLV3-19, HAMP, CSMD3, PDZK1IP1, IGHG3, MUC17, ALPK2, IGLV2-14, FRAS1, DNAH11, IGHGP, SAA2, BMPER, IGLV1-47, MMP3, FOSB, HPD, SYT13, IGHV4-59, SLC38A5, IGHA1, CYP2C9, IGKC, IGLC2, PGF, IGHV3-21, H19, FCRL5, PVALB, IGHV3-74, SLC6A3, IGHV1-46, IGLV2-23, IGLV3-1, HBA1, IGLV1-44, IGKV3-20, IGKV4-1, LAMA1, IGHV3-48, IGHV5-51, IGHG2, HBA2, KNG1, IGKV1-27, IGHM, IGLV2-11, FGL1, CYP4F22, IGLV1-51

Table 3.1: The list of genes classified as predictive biomarkers by the considered methods.

expression.

Potentially predictive biomarkers were found by applying uniCATE to subsets of the sunitinib ($n = 71$) and atezolizumab-bevacizumab ($n = 77$) treatment arms of the IMmotion 150 trial. Only patients with pre-treatment tumor RNA-seq samples were included. The 500 most variable genes based on this log-transformed RNA-seq data comprised the collection of potentially predictive biomarkers. Details of the gene expression data collection and preparation have previously been described by McDermott et al. [2018]. Objective response was used as the outcome variable. The conditional outcome regression model was fit with a Super Learner whose library contained (penalized) GLMs with treatment-biomarker interaction terms, XGboost models, Random Forest models, and the mean model. A nominal FDR cutoff of 5% was employed. The modified covariates and augmented modified covariates approach for binary outcomes of Tian et al. [2014] were also applied to these data.

The uniCATE method identified 92 genes as predictive biomarkers, whereas the modified covariates approach and its augmented counterpart identified 20 and 6, respectively. All results are listed in Table 3.1. That the approaches of Tian et al. [2014] are more conservative

than ours is a reversal of Section 3.4's simulation results, but may be explained by the more complex correlation structure of these data. Indeed, the former rely on sparse linear models which are known to select but a few features from any given highly correlated set. Our procedure, however, uncovers sets of correlated predictive biomarkers since their individual hypothesis testing results will also be associated. This property is desirable when analyzing genomic data as large gene sets permit more thorough biological exploration and improved interpretation than do single, uncorrelated genes [Subramanian et al., 2005]. The reporting of gene sets also improves the reproducibility of findings [Subramanian et al., 2005].

Rank	Gene Set Name	Genes in Set (K)	Description	Genes in Overlap (k)	k/K	p-value	FDR q-value
1	GOCC_IMMUNOGLOBULIN_COMPLEX	157	A protein complex that in its canonical form is composed of two identical immunoglobulin heavy chains and two identical immunoglobulin light chains, held together by disulfide bonds and sometimes complexed with additional proteins. An immunoglobulin complex may be embedded in the plasma membrane or present in the extracellular space, in mucosal areas or other tissues, or circulating in the blood or lymph. [GOC:add, GOC:jl, ISBN:0781765196]	37	0.24	0	0
2	GOBP_HUMORAL_IMMUNE_RESPONSE_MEDIATED_BY_CIRCULATING_IMMUNOGLOBULIN	149	An immune response dependent upon secreted immunoglobulin. An example of this process is found in Mus musculus. [GO_REF:0000022, GOC:add, ISBN:0781735149]	36	0.24	0	0
3	GOBP_COMPLEMENT_ACTIVATION	171	Any process involved in the activation of any of the steps of the complement cascade, which allows for the direct killing of microbes, the disposal of immune complexes, and the regulation of other immune processes; the initial steps of complement activation involve one of three pathways, the classical pathway, the alternative pathway, and the lectin pathway, all of which lead to the terminal complement pathway. [GO_REF:0000022, GOC:add, ISBN:0781735149]	36	0.21	0	0
4	GOMF_ANTIGEN_BINDING	158	Interacting selectively and non-covalently with an antigen, any substance which is capable of inducing a specific immune response and of reacting with the products of that response, the specific antibody or specifically sensitized T-lymphocytes, or both. Binding may counteract the biological activity of the antigen. [GOC:jl, ISBN:0198506732, ISBN:0721662544]	35	0.22	0	0
5	GOBP_B_CELL_MEDIATED_IMMUNITY	219	Any process involved with the carrying out of an immune response by a B cell, through, for instance, the production of antibodies or cytokines, or antigen presentation to T cells. [GO_REF:0000022, GOC:add, ISBN:0781735149]	36	0.16	0	0
6	GOBP_HUMORAL_IMMUNE_RESPONSE	373	An immune response mediated through a body fluid. [GOC:hb, ISBN:0198506732]	37	0.10	0	0
7	GOBP_LYMPHOCYTE_MEDIATED_IMMUNITY	351	Any process involved in the carrying out of an immune response by a lymphocyte. [GO_REF:0000022, GOC:add, ISBN:0781735149]	36	0.10	0	0
8	GOBP_ADAPTIVE_IMMUNE_RESPONSE_BASED_ON_SOMATIC_RECOMBINATION_OF_IMMUNE_RECEPTORS_BUILT_FROM_IMMUNOGLOBULIN_SUPERFAMILY_DOMAINS	358	An immune response mediated by lymphocytes expressing specific receptors for antigen produced through a somatic diversification process that includes somatic recombination of germline gene segments encoding immunoglobulin superfamily domains. Recombined receptors for antigen encoded by immunoglobulin superfamily domains include T cell receptors and immunoglobulins (antibodies) produced by B cells. The first encounter with antigen elicits a primary immune response that is slow and not of great magnitude. T and B cells selected by antigen become activated and undergo clonal expansion. A fraction of antigen-reactive T and B cells become memory cells, whereas others differentiate into effector cells. The memory cells generated during the primary response enable a much faster and stronger secondary immune response upon subsequent exposures to the same antigen (immunological memory). An example of this is the adaptive immune response found in Mus musculus. [GOC:add, GOC:mtg_sensu, ISBN:0781735149, ISBN:1405196831]	36	0.10	0	0
9	GOBP_REGULATION_OF_COMPLEMENT_ACTIVATION	114	Any process that modulates the frequency, rate or extent of complement activation. [GOC:go_curators]	27	0.24	0	0
10	GOBP_PHAGOCYTOSIS	374	A vesicle-mediated transport process that results in the engulfment of external particulate material by phagocytes and their delivery to the lysosome. The particles are initially contained within phagocytic vacuoles (phagosomes), which then fuse with primary lysosomes to effect digestion of the particles. [ISBN:0198506732]	35	0.09	0	0

Table 3.2: GSEA of GO terms for uniCATE's selected predictive biomarkers using IMmotion 150 data.

We performed a gene set enrichment analysis (GSEA) of gene ontology (GO) terms with uniCATE's 92 predictive biomarkers using MSigDB [Subramanian et al., 2005, Liberzon et al., 2011]. The top results are presented in Table 3.2. We found that these genes are generally associated with immune responses, including those mediated by B cells and lymphocytes. Similar findings have been reported by Au et al. [2021] in the context of clear cell renal cell carcinoma patients' therapeutic responses to nivolumab, another immune checkpoint inhibitor.

Now, having learned of potentially predictive biomarkers, we assessed how well they delineate patient sub-populations in the IMmotion 151 study. This study's subjects are believed to be drawn from the same population as those enrolled in IMmotion 150. Eight hundred and ten subjects possessed baseline tumor gene expression data for the 500 genes considered in our IMmotion 150 analysis: 406 in the atezolizumab-bevacizumab combination arm, and 404 in the sunitinib arm. Figure 3.6 presents the heatmaps of log-transformed gene expression data for each methods' set of predictive genes. Subgroups are easily discerned in uniCATE's heatmap, but not so much in the other procedures'. This further emphasizes the benefits of uniCATE's capacity to identify sets of correlated predictive biomarkers. Note that the most prominent cluster of patients in the modified covariates method's heatmap is driven by the *XIST* gene. It is not selected by the augmented modified covariates procedure or uniCATE. Upon further inspection, it does not appear to have a strong predictive effect (Figure 3.8).

It is unclear from these heatmaps alone whether these subgroups correspond to clusters of patients that benefit more from one therapy than another. We therefore established subgroups by performing hierarchical clustering with complete linkage using Euclidean distance on the methods' selections of IMmotion 151's log-transformed gene expression data. The difference in ORR was then computed between patients receiving the atezolizumab-bevacizumab combination and the sunitinib regiment within each of these subgroups. The subgroups identified by the (augmented) modified covariates methods' biomarkers had negligible differences in ORR (not shown). Instead, we used their estimated treatment rules to predict whether each patient would benefit more from the atezolizumab-bevacizumab combination or sunitinib, and then computed the difference in ORR within these groups. The findings are presented in Figure 3.7.

Figure 3.7 evaluates the patient populations classified by each method. When classifying patients into subgroups, the modified covariates approach and the augmented modified covariates approach subset patients into two groups. Ideally, one of the patient groups should produce a large, positive ORR difference representing an increased benefit from the novel drug combination. Figure 3.7 shows that when two patient groups are of interest, uniCATE's biomarkers-defined subgroups are comparable to the groups identified by the other two methods in terms of the effect size and the group size. The unsupervised clustering approach used in uniCATE also permits the definition of multiple clusters, providing a more refined investigation of patient sub-populations. When considering three or four clusters ($K = 3, 4$), one subgroup is found to respond much better on average to the atezolizumab-bevacizumab combination than to sunitinib. This difference in ORR is greater than that of any subgroup

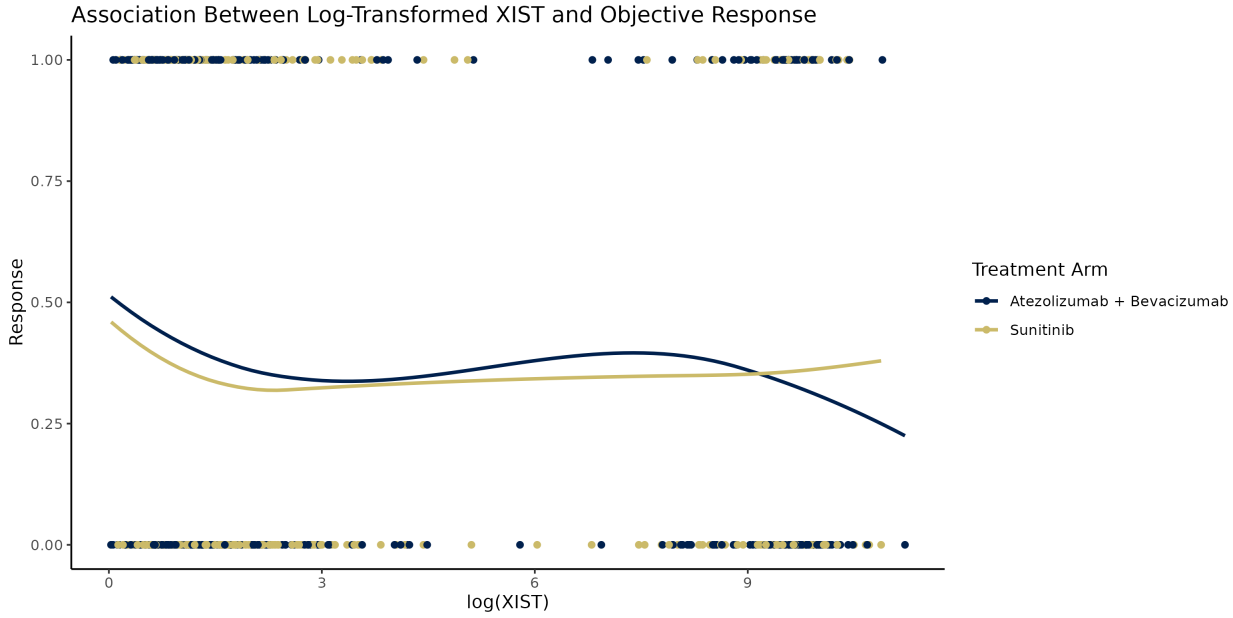


Figure 3.8: While the log-transformed *XIST* gene expression data can be used to define two patient subpopulations within the IMmotion 151 study, it does not appear to have a strong predictive effect like the simulated biomarkers of Figure 3.1.

defined using treatment assignment rules.

These results suggest that uniCATE uncovered biomarkers that influence whether mRCC patients are more likely to respond to tyrosyne kinase inhibitors alone or in combination with immune checkpoint inhibitors. More work is necessary to validate these findings, and to determine whether these biomarkers could form the basis of assays that inform treatment decisions. Demonstrating that these biomarkers are predictive of other clinical endpoints, like overall survival, progression-free survival and safety, would make for compelling evidence. Recovering these biomarkers in a comparison of this drug combination to the current standard of care would be more convincing still.

3.6 Discussion

In this work, we demonstrate how predictive biomarker discovery, typically a byproduct of treatment rule estimation, is better addressed as a standalone variable importance estimation problem. We derive a novel nonparametric estimator for a causal parameter which we argue is generally useful and interpretable, and show that this estimator is consistent and asymptotically linear under non-restrictive assumptions. We then verify that our proposed procedure's asymptotic guarantees are approximately achieved across diverse data-generating distributions in a thorough simulation study of moderate to high-dimensional randomized

control trials. Our method is then used in an exploratory analysis of real clinical trial data, producing biologically meaningful results that identify patient subgroups with greater treatment effect heterogeneity than procedures not explicitly developed for predictive biomarker discovery.

While we derive theory for uniCATE's application to observational data, we benchmarked it exclusively in randomized control trial settings since they constitutes our primary application area of interest. Evaluating our method in quasi-experimental settings, however, offers an interesting avenue of future research. Subsequent work may also explore analogous (causal) variable importance parameters based on, for example, the relative CATE, or adapt the univariate CATE for time-to-event outcomes. The study of other non and semiparametric estimators of these parameters, such as one-step estimators or targeted maximum likelihood estimators, might also prove fruitful. Finally, future work might assess whether treatment effect variable importance parameter inference procedures could be coupled with novel multiple testing adjustment approaches, like that of Fithian and Lei [2020], to better account for the complex correlation structures often found among biomarkers.

3.7 Proofs

Theorem 3.1: Identification

Proof. Standard results ensure that $\Psi^F(P_{X,0})$ is identified by $\Psi(P_0)$: By the law of double expectation, we find that $\mathbb{E}_{P_{X,0}}[Y^{(A)}B_j] = \mathbb{E}_{P_{X,0}}[\mathbb{E}_{P_{X,0}}[Y^{(A)}|W]B_j]$, and by A3.1, A3.2 that $\bar{Q}_{X,0}(A,W) = \bar{Q}_0(A,W)$. □

Theorem 3.2: Efficient Influence Function of $\Psi_j(P)$

Proof. We follow the general guidelines in the review of Hines et al. [2022b] to derive the EIF of $\Psi_j(P_0)$. Define the fixed distribution P whose support is contained in the support of P_0. We define the parametric submodel of P_0 for $t \in [0,1]$ as

$$P_t = tP + (1-t)P_0.$$

Then,

$$
\begin{aligned}
D_j(O, P_0) &= \frac{d}{dt}\Psi_j(P_t)\Big|_{t=0} \\
&= \frac{d}{dt}\frac{\mathbb{E}_{P_t}\left[\left(\bar{Q}_t(1,W) - \bar{Q}_t(0,W)\right)B_j\right]}{\mathbb{E}_{P_t}\left[B_j^2\right]}\Bigg|_{t=0} \\
&= \frac{1}{\mathbb{E}_{P_t}\left[B_j^2\right]^2}\Bigg(\frac{d}{dt}\left\{\mathbb{E}_{P_t}\left[\left(\bar{Q}_t(1,W) - \bar{Q}_t(0,W)\right)B_j\right]\right\}\mathbb{E}_{P_t}\left[B_j^2\right] - \\
&\qquad\qquad \mathbb{E}_{P_t}\left[\left(\bar{Q}_t(1,W) - \bar{Q}_t(0,W)\right)B_j\right]\frac{d}{dt}\left\{\mathbb{E}_{P_t}\left[B_j^2\right]\right\}\Bigg)\Bigg|_{t=0} \\
&= \frac{1}{\mathbb{E}_{P_0}\left[B_j^2\right]^2}\Bigg(\Big(\tilde{T}(O,P_0)B_j - \mathbb{E}_{P_0}\left[\left(\bar{Q}_0(1,W) - \bar{Q}_0(0,W)\right)B_j\right]\Big)\mathbb{E}_{P_0}\left[B_j^2\right] \\
&\qquad\qquad -\mathbb{E}_{P_0}\left[\left(\bar{Q}_0(1,W) - \bar{Q}_0(0,W)\right)B_j\right]\left(B_j^2 - \mathbb{E}_{P_0}\left[B_j^2\right]\right)\Bigg) \\
&= \frac{\left(\tilde{T}(O,P_0) - \Psi_j(P_0)B_j\right)B_j}{\mathbb{E}_{P_0}\left[B_j^2\right]}
\end{aligned}
$$

□

Corollary 3.1: Estimating equation estimator derivation and double robustness.

Proof. The estimating equation estimator for the j^{th} biomarker is given by:

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} D_j(O_i; P_m) \\
&= \frac{\sum_{i=1}^{n}\left(\tilde{T}(O_i; P_m) - \Psi B_{ij}\right)B_{ij}}{\sum_{i=1}^{n} B_{ij}^2} \\
\implies \Psi_j^{(ee)}(P_n; P_m) &= \frac{\sum_{i=1}^{n}\tilde{T}(O_i; P_m)B_{ij}}{\sum_{i=1}^{n} B_{ij}^2}.
\end{aligned}
$$

Then, by the Weak Law of Large Numbers,

$$\begin{aligned}\Psi_j^{(ee)}(P_n; P_m) - \Psi_j(P_0) \to & \frac{\mathbb{E}_{P_0}\left[\tilde{T}(O; P_m)B_j\right]}{\mathbb{E}_{P_0}\left[B_j^2\right]} - \frac{\mathbb{E}_{P_0}\left[\left(\bar{Q}_0(1,W) - \bar{Q}_0(0,W)\right)B_j\right]}{\mathbb{E}_{P_0}\left[B_j^2\right]} \\ & \propto \mathbb{E}_{P_0}\left[B_j\left(\frac{g_0(W)}{g_m(W)} - 1\right)\left(\bar{Q}_0(1,W) - \bar{Q}_m(1,W)\right)\right. \\ & \qquad \left. -B_j\left(\frac{1-g_0(W)}{1-g_m(W)} - 1\right)\left(\bar{Q}_0(0,W) - \bar{Q}_m(0,W)\right)\right].\end{aligned}$$

If $g_m = g_0$, then this estimator is consistent. The same is true if either $\|g_m - g_0\|_2^2 = o_P(1)$ or $\|\bar{Q}_m - \bar{Q}_0\|_2^2 = o_P(1)$. □

Theorem 3.3: Limiting distribution of the estimating equation estimator.

Proof. Define the plug-in estimator for the univariate CATE of biomarker j with nuisance parameters estimated using P_m as $\Psi_j(P_n; P_m)$. Then we have through the von Mises expansion of $\Psi_j(\cdot)$ about P_0 that

$$\begin{aligned}\sqrt{n}\left(\Psi_j(P_n; P_m) - \Psi_j(P_0)\right) = & \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_j(O; P_0) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_j(O; P_m) \\ & + \sqrt{n}\left(\mathbb{E}_{P_n} - \mathbb{E}_{P_0}\right)\left[D_j(O; P_m) - D_j(O; P_0)\right] - \sqrt{n}R(P_0, P_m).\end{aligned} \tag{3.9}$$

The first term is the sum of mean-zero random variables, and so it converges to a Normal with variance equal to that of the EIF, scaled by n, as $n \to \infty$. The second term is the bias term that is accounted for by the estimating equation estimator $\Psi_j^{(ee)}(P_n; P_m)$. The third and fourth terms are the empirical process and remainder terms, respectively, and we must show that they converge to zero in probability.

The analysis of empirical process term is identical to that of the average treatment effect presented in Zheng and van der Laan [2011] due to the similarity of these parameters. Essentially, so long as the conditional outcome regression and propensity score estimators converge in probability to some function under the L2 norm, the empirical process term is $o_P(1)$.

We now study the remainder term:

$$
\begin{aligned}
-R(P_0, P_m) &= \frac{\mathbb{E}_{P_0}\left[\left(\tilde{T}(O; P_m) - \Psi(P_m)B_j\right)B_j\right]}{\mathbb{E}_{P_0}\left[B_j^2\right]} + (\Psi_j(P_m) - \Psi_j(P_0)) \\
&= \frac{1}{\mathbb{E}_{P_0}\left[B_j^2\right]}\mathbb{E}_{P_0}\left[\tilde{T}(O; P_m)B_j - \mathbb{E}_{P_0}\left[B_j^2\right]\Psi_j(P_0)\right] \\
&= \frac{1}{\mathbb{E}_{P_0}\left[B_j^2\right]}\mathbb{E}_{P_0}\left[B_j\left(T_1(O; P_m) - \bar{Q}_0(1, W) - T_0(O; P_m) + \bar{Q}_0(0, W)\right)\right] \\
&= \frac{1}{\mathbb{E}_{P_0}\left[B_j^2\right]}\mathbb{E}_{P_0}\left[B_j\left(\frac{g_0(W)}{g_m(W)} - 1\right)\left(\bar{Q}_0(1, W) - \bar{Q}_m(1, W)\right)\right. \\
&\qquad\qquad \left. -B_j\left(\frac{1 - g_0(W)}{1 - g_m(W)} - 1\right)\left(\bar{Q}_0(0, W) - \bar{Q}_m(0, W)\right)\right] \\
&\leq \frac{1}{\mathbb{E}_{P_0}\left[B_j^2\right]}\left(\left|\mathbb{E}_{P_0}\left[B_j\left(\frac{g_0(W)}{g_m(W)} - 1\right)\left(\bar{Q}_0(1, W) - \bar{Q}_m(1, W)\right)\right]\right|\right. \\
&\qquad\qquad \left. + \left|\mathbb{E}_{P_0}\left[B_j\left(\frac{1 - g_0(W)}{1 - g_m(W)} - 1\right)\left(\bar{Q}_0(0, W) - \bar{Q}_m(0, W)\right)\right]\right|\right) \\
&\leq \frac{1}{\mathbb{E}_{P_0}\left[B_j^2\right]}\left(\mathbb{E}_{P_0}\left[B_j^2\left(\frac{g_0(W) - g_m(W)}{g_m(W)}\right)^2\right]^{1/2}\right. \\
&\qquad\qquad \mathbb{E}_{P_0}\left[\left(\bar{Q}_0(1, W) - \bar{Q}_m(1, W)\right)^2\right]^{1/2} \\
&\qquad + \mathbb{E}_{P_0}\left[B_j^2\left(\frac{g_m(W) - g_0(W)}{1 - g_m(W)}\right)^2\right]^{1/2} \\
&\qquad\qquad \left.\mathbb{E}_{P_0}\left[\left(\bar{Q}_0(0, W) - \bar{Q}_m(0, W)\right)^2\right]^{1/2}\right)
\end{aligned}
\tag{3.10}
$$

If g_0 is known, as in a randomized control trial, then the remainder term is exactly zero. When neither g_0 or $\bar{Q}_0$ is now known, then the remainder term of Equation (3.9) is $o_P(1)$ under the conditions of A3.4. The conditions on convergence rates can be relaxed even further: The remainder term converges to zero in probability so long as the last line of Equation (3.10) is $o_P(n^{-1/2})$. That is, we may obtain our desired result even if, say, $\bar{Q}_m$ converges at slower rate to $\bar{Q}_0$ than $n^{-1/4}$ in probability so long as g_m converges more quickly to g_0. □

Chapter 4

A Nonparametric Framework for Treatment Effect Modifier Variable Importance Parameters

4.1 Introduction

The detection and quantification of heterogeneous treatment effects are central to numerous areas of study in the medical and social sciences. Examples include precision medicine, as seen in Chapter 3, where practitioners seek patient subgroups exhibiting differing benefits from a given therapy, and economics, where policymakers assess the impact of government interventions on diverse population strata. This heterogeneity is generally linked to treatment effect modifiers (TEM). TEMs are pre-treatment covariates which, as their name suggests, modify the effect of a treatment, alternatively referred to as an exposure, on the outcome. In precision medicine, the response of patients with a shared disease to a common therapy may be a function of, for example, sex-at-birth, age, genetic mutations, and environmental exposures. Uncovering TEMs is therefore of great importance when investigating or attempting to account for disparate effects of treatment in a population.

Some parametric modeling techniques can accomplish just that in traditional asymptotic settings under stringent conditions about the data-generating process (DGP). When including treatment-covariate interaction terms in addition to main effect terms in a linear model for a continuous outcome, TEMs are generally defined as the features with non-zero interaction coefficients. Consistent estimation and valid hypothesis testing of the TEMs are possible when the DGP admits a linear relationship between the outcome, treatment, and covariates. Generalized linear models (GLM) might be used for TEM discovery in more general settings, such as when the outcome is binary or a non-negative integer. With time-to-event outcomes, the Cox proportional hazards model with treatment-covariate interactions might be used. If the posited functional relationship does not correspond to reality, however, inference is invalid [see, for example, Hernán, 2010].

Furthermore, the parameters corresponding to the aforementioned models, like the odds ratio of a logistic regression model or the hazards ratio of a proportional hazards model, depend on the other covariates included in the model. These conditional parameters are said to be *noncollapsible* [for a discussion and worked example, see Greenland et al., 1999], in the sense that marginalizing over the other covariates in the model may produce a marginal parameter whose value differs from the marginal parameter directly obtained by omitting these covariates (to see this, recall that for two random variables X and Y and an arbitrary function $g(\cdot)$, we generally find that $\mathbb{E}[g(\mathbb{E}[Y \mid X])] \neq g(\mathbb{E}[\mathbb{E}[Y \mid X]]) = g(\mathbb{E}[Y]))$. Noncollapsible parameters lack a causal interpretation that unambiguously relates them to marginal treatment effects.

More flexible approaches targeting the conditional average treatment effect (CATE) may be employed to address these issues. In inferring the expected difference in potential outcomes — that is, the difference in outcomes that could be computed if each observation's outcomes under treatment and control were measured — as a function of the covariates [Rubin, 1974], CATE estimators are uniquely suited for TEM discovery. The double-robust estimators of Zhao et al. [2018], Semenova and Chernozhukov [2020], and Bahamyirou et al. [2022], which model the CATE using a linear model, permit valid statistical inference about features' ability to modify the effect of treatment under less restrictive assumptions about the DGP than traditional parametric methods. Others, like the Super-Learner-based [van der Laan et al., 2007b] estimator of Luedtke and van der Laan [2016] or the Random-Forests-inspired [Breiman, 2001] estimators of Wager and Athey [2018] and Cui et al. [2022], rely on nonparametric supervised statistical learning algorithms to identify potential TEMs under even fewer constraints on the DGP.

When the number of potential TEMs is commensurate with the number of observations, or indeed much larger, the above parametric and CATE estimators' capacity to reliably uncover TEMs diminishes. Estimation of the linear model coefficients requires penalized regression methods like the LASSO [Tibshirani, 1996, Tian et al., 2014, Chen and Guestrin, 2016], rendering hypothesis testing of treatment-covariate coefficients difficult. Practitioners might instead rely on the asymptotic feature selection properties of the LASSO, but these hold only under restrictive and unverifiable conditions on sparsity and covariate correlation structures [Zhao and Yu, 2006]. Similar limitations plague the CATE estimators relying on method-specific variable importance measures. In particular, the causal forests of Wager and Athey [2018] and Cui et al. [2022] can assess the importance of variables, employing a permutation-based approach analogous to those of traditional Random Forests. In high dimensions, however, this metric can produce unreliable rankings of covariates' treatment modification abilities: correlated features are likely to act as surrogates for one another, leading to deflated importance scores [Hastie et al., 2009, Chap. 15].

Instead of depending on algorithm-specific modeling strategies that treat TEM discovery as a byproduct of conditional outcome or CATE estimation, Williamson et al. [2022], Hines et al. [2022a] and Boileau et al. [2022]—the latter of which makes up the contents

of Chapter 3—recently proposed TEM variable importance parameters (TEM-VIP[1]) that directly assess the strength of covariates' capacity to modify the effect of treatment. These algorithm-agnostic parameters, defined within nonparametric statistical models and which may be augmented with causal interpretations, permit formal statistical inference about TEMs.

Combining popular variable dropout procedures and previous work about the variance of the conditional treatment effect estimator, Levy et al. [2021], Williamson et al. [2022], and Hines et al. [2022a] proposed analogous TEM-VIPs measuring individual or predefined sets of variables' influence on the CATE variance. For instance, Hines et al. [2022a] define the TEM-VIP of a set of covariates as one minus the ratio of the variance of the CATE conditioning on all but these covariates and the variance of the CATE conditioning on all available covariates. The accompanying, nonparametric estimators are consistent and asymptotically linear under nonrestrictive assumptions about the DGP. However, these TEM-VIPs might produce misleading assessments of TEM impact when the covariates are correlated: TEM-VIPs will generally possess values that do not reflect covariates' capacity for treatment effect modification, like the Random-Forests-based CATE variable importance measure of Wager and Athey [2018], Cui et al. [2022]. We expect this issue to be exacerbated in high dimensions due to the increased chance of complex correlation structures. Additionally, repeatedly omitting variables and estimating nuisance parameters is computationally expensive — and perhaps intractable — when the number of potential TEMs is large.

We derived in Chapter 3 a marginal TEM-VIP expressly for high-dimensional DGPs with continuous or binary outcomes. Assuming that the expected difference in potential outcomes is linear in any given covariate when conditioning on said covariate, the proposed TEM-VIP is the simple linear regression coefficient obtained by regressing this difference on the potential TEM. We argued that this parameter provides a meaningful summary in all but pathological DGPs: the larger it is, the larger the variables capacity for treatment effect modification. Further, it does not suffer from the previously mentioned issues associated with dropout- and permutation-based variable importance measures. A nonparametric estimator of this TEM-VIP was proposed, and shown to be double-robust and asymptotically linear under mild conditions on the DGP. A simulation study demonstrated that these asymptotic properties were approximately achieved in finite-sample, high-dimensional randomized control trials.

This TEM-VIP is limited, however, to absolute effect modification for continuous and binary responses. Expanding on this work and taking inspiration from previous research on non- and semiparametric approaches [for example, Rosenblum and van der Laan, 2010, Tchetgen Tchetgen et al., 2009, Tuglus et al., 2011, Chambaz et al., 2012, Yadlowsky et al., 2021], we present a general framework for defining and performing inference about marginal model-agnostic TEM-VIPs. Our approach is demonstrated through the creation of a new absolute TEM-VIP for DGPs with right-censored time-to-event outcomes, and of new relative

[1]Previous work has referred to TEM-VIPs as (treatment effect) variable importance measures (TE-)VIMs [for example, Williamson et al., 2022, Hines et al., 2022a], which we believe blurs the distinction between parameters and estimators and fails to emphasize that these measures of variable importance are well-defined parameters of a statistical model.

TEM-VIPs for DGPs with continuous, binary, and right-censored time-to-event outcomes. We derive one-step, estimating equation and targeted maximum likelihood (TML) estimators based on these parameters' efficient influence functions, study their asymptotic behavior, and investigate their finite sample properties in simulation experiments. This general framework equips practitioners with the tools to define bespoke TEM-VIPs, readily derive nonparametric estimators, and establish sufficient conditions for which these estimators permit reliable inference.

The remainder of the article is organized as follows: Section 4.2 presents TEM-VIPs and related inference procedures in data-generating processes with binary treatment variables and continuous outcomes. The CATE-based TEM-VIPs of the previous chapter are re-framed in terms of treatment effect modification discovery for continuous outcomes in Section 4.2. Sufficient identifiability conditions for the estimation of TEM-VIPs using observational data are also presented, as are nonparametric estimators of this estimand. The asymptotic properties of these estimators are then studied. A proposal for a relative TEM-VIP follows in Section 4.2. Accompanying causal identifiability conditions, nonparametric estimators, and sufficient conditions for the desirable asymptotic behavior of these estimators are given. Sections 4.3 and 4.4 introduce analogous developments for data-generating processes with binary and time-to-event outcomes, respectively. We discuss, in Section 4.5, the general procedure for defining model-agnostic TEM-VIPs, deriving accompanying estimators, and studying their asymptotic characteristics in nonparametric models. Simulation studies and a real data application are then presented in Sections 4.6 and 4.7, respectively, and we end with a discussion of our contributions in Section 4.8. Proofs are relegated to the Section 4.9 for clarity of exposition.

4.2 Continuous Outcomes

Problem Setting

Let there be n independent and identically distributed (i.i.d.) random vectors $\{X_i\}_{i=1}^n$, such that $X_i = (W_i, A_i, Y_i^{(0)}, Y_i^{(1)}) \sim P_{X,0} \in \mathcal{M}_X$, where W_i is a set of p covariates that are possibly treatment-outcome confounders, A_i is a binary variable indicating treatment assignment (0 for control, 1 for treatment), and $Y_i^{(1)}$ and $Y_i^{(0)}$ are continuous potential outcomes [Rubin, 1974] that are assumed to be bounded between 0 and 1 without loss of generality. The potential outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$ are the outcomes one would observe for the i^{th} observation had it been assigned to the treatment and control conditions, respectively. Here, p is of similar magnitude as, or larger than, n. Finally, $\mathcal{M}_X$ is a nonparametric model of possible DGPs. We omit the subscript i where possible throughout the remainder of the chapter to ease notational burden.

The true DGP, $P_{X,0}$, is generally unknown, and realizations of its random vectors are typically unmeasurable, as only one potential outcome is observed. Nevertheless, $P_{X,0}$ allows for the definition of causal parameters on which statistical inference may subsequently be

performed. An example of such a parameter is the conditional average treatment effect (CATE):

$$\mathbb{E}_{P_{X,0}}\left[Y^{(1)} - Y^{(0)}|W\right].$$

As discussed in Section 4.1, however, the CATE poses a challenging estimation problem — even if somehow provided with the complete data generated according to $P_{X,0}$ — due to the dimension of W. Likewise, the recovery of treatment effect modifiers using traditional variable importance techniques based on CATE estimates, like penalized linear models or Random Forests [Tian et al., 2014, Chen et al., 2017, Zhao et al., 2018, Wager and Athey, 2018, Ning et al., 2020, Bahamyirou et al., 2022], is generally unreliable in high dimensions. We instead consider the causal TEM-VIP proposed Chapter 3, which we re-introduce here in greater generality.

Absolute Treatment Effect Modification Variable Importance Parameter

Causal Parameter

Indexing W by $j = 1, \ldots, p$, assuming without loss of generality that $\mathbb{E}_{P_{X,0}}[W_j] = 0$, and requiring that $\mathbb{E}_{P_{X,0}}[W_j^2] > 0$, an absolute TEM-VIP of the j^{th} covariate can be defined as a mapping

$$\Psi_j^F(P_{X,0}) \equiv \frac{\mathbb{E}_{P_{X,0}}\left[\left(Y^{(1)} - Y^{(0)}\right) W_j\right]}{\mathbb{E}_{P_{X,0}}\left[W_j^2\right]}. \tag{4.1}$$

Letting $\bar{Q}_{P_{X,0}}(a, W) \equiv \mathbb{E}_{P_{X,0}}[Y^{(a)}|W]$, it is straightforward to show that

$$\Psi_j^F(P_{X,0}) = \frac{\mathbb{E}_{P_{X,0}}\left[\left(\bar{Q}_{P_{X,0}}(1, W) - \bar{Q}_{P_{X,0}}(0, W)\right) W_j\right]}{\mathbb{E}_{P_{X,0}}\left[W_j^2\right]}.$$

The estimand is then given by $\Psi^F : \mathcal{M}_X \to \mathbb{R}^p$, $\Psi^F(P_{X,0}) = (\Psi_1^F(P_{X,0}), \ldots, \Psi_p^F(P_{X,0}))$.

Just like the estimand presented in Chapter 3, $\Psi^F(P_{X,0})$ is the vector of simple linear regression coefficients generated by regressing the differences in expected potential outcomes against the individual elements of W. When the relationship between $f(W)$ and the W_j's is nonlinear, as is almost surely the case in most applications, this $\Psi^F(P_{X,0})$ can instead be viewed as assessing the correlation between the difference in potential outcomes and each potential TEM, re-normalized to be on the same scale as Y.

Though TEM-VIPs provide a continuous measure of the strength of the treatment effect modifications, some applications may call for the dichotomization of covariates into TEMs and non-TEMs based on $\Psi^F(P_{X,0})$. By default, we classify the j^{th} covariate W_j as a TEM if $|\Psi_j^F(P_{X,0})| > 0$ and note that, in some settings, it may make sense to impose a non-zero threshold for this classification. We emphasize that TEMs need not be treatment–outcome confounders.

Identifiability Through Observed-Data Parameter

The full data $\{X_i\}_{i=1}^n = \{(W_i, A_i, Y_i^{(0)}, Y_i^{(1)})\}_{i=1}^n$ are generally censored through the treatment assignment mechanism. We instead have access to n i.i.d. random variables $O = (W, A, Y) \sim P_0 \in \mathcal{M}$. The statistical model, $\mathcal{M}$, is fully determined by $\mathcal{M}_X$: for each $P_X \in \mathcal{M}_X$, there exists a unique $P \in \mathcal{M}$, where W and A are defined as in the full-data DGP and Y, the observed outcome variable, is given by $AY^{(1)} + (1 - A)Y^{(0)}$. Here, P_0 is the unknown DGP of the observed data. Throughout the remainder of the chapter, we denote the empirical distribution by P_n, the expected conditional outcome $\mathbb{E}_{P_0}[Y|A, W]$ by $\bar{Q}_0(A, W)$, and the propensity score $\mathbb{P}_{P_0}[A = 1|W]$ by $g_0(W)$. $\bar{Q}_0(A, W)$ and $g_0(W)$ are written as $\bar{Q}_0$ and g_0 where possible for notational convenience.

Recall too that we place the following moment conditions on W:

Assumption 4.1. *Centered covariates:* $\mathbb{E}_{P_{X,0}}[W] = 0$*, without loss of generality.*

Assumption 4.2. *Non-zero variance:* $\mathbb{E}_{P_{X,0}}[W_j^2] > 0$*, for* $j = 1, \ldots, p$.

Assumption A4.1 lightens notation; it has no practical implications. A4.2 is easily satisfied in practice by filtering variables exhibiting no variability. Note too that pre-treatment covariates with zero variance cannot possibly modify the effect of the treatment.

Now, the challenge lies in establishing an equivalence between a parameter of the DGP for the observed data O and $\Psi^F(P_{X,0})$. We repeat sufficient identifiability conditions outlined in the previous chapter for just that for completeness of presentation:

Assumption 4.3. *No unmeasured confounding:* $Y^{(a)} \perp A|W$*, for* $a \in \{0, 1\}$.

Assumption 4.4. *Positivity: there exists some constant* $\epsilon > 0$ *such that* $\mathbb{P}_{P_0}[\epsilon < g_0(W) < 1 - \epsilon] = 1$.

Theorem 4.1. *Assuming that A4.1, A4.2, A4.3, and A4.4 hold, we find that*

$$\begin{aligned} \Psi_j(P_0) &\equiv \frac{\mathbb{E}_{P_0}\left[\left(\bar{Q}_0(1, W) - \bar{Q}_0(0, W)\right) W_j\right]}{\mathbb{E}_{P_0}\left[W_j^2\right]} \\ &= \Psi_j^F(P_{X,0}), \end{aligned} \tag{4.2}$$

for $j = 1, \ldots, p$*. The parameter* $\Psi : \mathcal{M} \to \mathbb{R}^p$ *defined as* $\Psi(P_0) = (\Psi_1(P_0), \ldots, \Psi_p(P_0))$ *is therefore equal to the full-data estimand* $\Psi^F(P_{X,0})$.

The two latest assumptions are ingrained in the causal inference literature. Assumption A4.3 ensures that treatment assignment is regarded as if performed in a randomized experiment. It is more easily satisfied by considering many pre-treatment covariates as potential confounders. Assumption A4.4 requires all observations to have a non-zero probability of receiving either treatment condition, guaranteeing that $\bar{Q}_0(1, W)$ and $\bar{Q}_0(0, W)$ are equal to $\bar{Q}_{P_{X,0}}(1, W)$ and $\bar{Q}_{P_{X,0}}(0, W)$, respectively.

Inference

Having established an identifiable parameter, we detail procedures for performing inference about it. We first, however, briefly review the basics of nonparametric asymptotic theory.

Preliminaries Consider a degenerate distribution $\tilde{P}$ that places all support of its random observations $\tilde{O}$ on $\tilde{o}$. Further assume that $\tilde{o}$ is contained in the support of $P \in \mathcal{M}$ and define a two-component mixture model $P_\epsilon = \epsilon\tilde{P} + (1-\epsilon)P$. Appealing to Riesz's representation theorem [Fisher and Kennedy, 2021, Hines et al., 2022b], the efficient influence function of a parameter $\Theta(P)$ is defined through the following Gateaux — that is, functional — derivative:

$$\left.\frac{d\Theta(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \lim_{\epsilon\to 0}\frac{\Theta(P_\epsilon)-\Theta(P)}{\epsilon} = \int \theta(o)(d\tilde{P}(o) - dP(o)) = \int D(o,P)d\tilde{P}(o) = D(\tilde{o},P)\ .$$

Here, θ is defined such that $\mathbb{E}_P[\theta(O)] = \Theta(P)$, and $D(o,P) = \theta(o) - \int \theta(o)dP(o) = \theta(o) - \Theta(P)$. $D(O,P)$ therefore generalizes the concept of directional derivatives to functionals, measuring Θ's sensitivity to perturbations of P. Intuitively, then, $\mathbb{E}_P[D(O,P)] = 0$ for $O \sim P$. When the variance of $D(O,P)$ is bounded under all $P \in \mathcal{M}$, we say that the Gateaux derivative is well-defined and that Θ is pathwise differentiable.

Now, similar to how asymptotic approximations of mean-based parameters are studied through Taylor expansions, the asymptotic behavior of the plug-in estimator $\Theta(P_n)$, a functional, is studied by way of a von Mises expansion [von Mises, 1947, Bickel et al., 1993b, van der Laan and Robins, 2003b, Hines et al., 2022b]. This functional equivalent to the Taylor expansion is defined in terms of the efficient influence function (EIF):

$$\begin{aligned}\sqrt{n}\left(\Theta(P_n) - \Theta(P_0)\right) = \sqrt{n}\mathbb{E}_{P_n}\left[D(O,P_0)\right] - \sqrt{n}\mathbb{E}_{P_n}\left[D(O,P_n)\right] \\ + \sqrt{n}\left(\mathbb{E}_{P_n} - \mathbb{E}_{P_0}\right)\left(D(O,P_n) - D(O,P_0)\right) - \sqrt{n}R(P_0,P_n)\ .\end{aligned} \tag{4.3}$$

$R(P_0,P_n)$ is a second-order remainder term in n. Since $\mathbb{E}_{P_0}[D(O,P_0)] = 0$, the first term of Equation (4.3) converges to a Gaussian random variable with mean zero and variance equal to $\mathbb{E}_{P_0}[D(O,P_0)^2]$ by the central limit theorem. The second term is a bias term that generally does not vanish asymptotically. The third term of the von Mises expansion can generally be shown to converge to zero in probability under sufficient empirical process conditions. Alternatively, sample-splitting procedures, often referred to as *cross-fitting*, can be used to relax these conditions, as suggested by Pfanzagl and Wefelmeyer [1985], Klaassen [1987], Zheng and van der Laan [2011], Chernozhukov et al. [2017]. The remainder term can generally be shown to converge to zero in probability under convergence rate assumptions about nuisance parameter estimators.

Nonparametric estimators, like the one-step [Pfanzagl and Wefelmeyer, 1985, Bickel et al., 1993b], estimating equation [van der Laan and Robins, 2003b, Chernozhukov et al., 2017], and targeted maximum likelihood (TML) estimators [van der Laan and Rubin, 2006b, van der Laan and Rose, 2011b, 2018b], correct the asymptotic bias term. They are constructed from the efficient influence function.

One-Step This estimator is derived by subtracting the asymptotic bias term from the plug-in estimator: $\Theta^{(\mathrm{OS})}(P_n) \equiv \Theta(P_n) + \mathbb{E}_{P_n}[D(O, P_n)]$.

Estimating Equation $\Theta^{(\mathrm{EE})}(P_n)$ is the solution to the following estimating equation: $0 = \mathbb{E}_{P_n}[D(O, P_n)]$.

TML $\Theta^{(\mathrm{TML})}(P_n)$ is obtained by tilting P_n to generate a $P_n^\star$ such that $\mathbb{E}_{P_n}[D(O, P_n^\star)] \approx 0$. There are many ways to achieve this. Examples are provided later in the chapter, as well as in van der Laan and Rubin [2006b], van der Laan and Rose [2011b, 2018b]. The estimator is then defined as $\Theta^{(\mathrm{TML})}(P_n) \equiv \Theta(P_n^\star)$, the plug-in estimator using $P_n^\star$. Unlike one-step and estimating equation estimators, TML estimators constrain estimates to the parameter space.

Provided the required conditions ensuring the third and fourth terms of Equation (4.3) converge in probability to zero are met, these estimators are asymptotically linear and efficient. That is, they are asymptotically normally distributed with mean $\Theta(P_0)$ and variance $\mathbb{E}_{P_0}[D(O, P_0)^2/n]$ and have the smallest asymptotic variance among regular and asymptotically linear (RAL) estimators in a nonparametric model [Bickel et al., 1993b, Tsiatis, 2006, van der Laan and Rose, 2011b].

Inference about $\Theta(P_0)$ can then be based on the asymptotically normal distribution of its one-step, estimating equation, and TML estimators. In particular, the α-level Wald-type confidence interval for $\Theta(P_0)$ can be constructed identically for each of the three estimators, $\Theta^{(\cdot)}(P_n)$, as follows:

$$\Theta^{(\cdot)}(P_n) \pm z_{1-\alpha/2}\sqrt{\frac{\mathbb{E}_{P_0}[D(O, P_0)^2]}{n}}, \tag{4.4}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{\text{th}}$ quantile of the standard Normal distribution. Of course, $\mathbb{E}_{P_0}[D(O, P_0)^2]$ is generally unknown; an estimator, $\mathbb{E}_{P_n}[D(O, P_n)^2]$, is used instead. When there are many tests to perform in small-to-moderate sample sizes, the empirical Bayes approach to variance estimation proposed by Hejazi et al. [2023] might also be employed for improved Type I error rate control.

Efficient Influence Function We return to our study of the TEM-VIP $\Psi(P_0)$, defined in Equation (4.2). The efficient influence function of $\Psi_j(P)$ for $P \in \mathcal{M}$ was previously derived by Chapter 3. We restate it here for convenience.

Proposition 4.1. *Assume A4.1 and A4.2. Define $\Psi_j(P)$ as in Equation (4.2) for some $P \in \mathcal{M}$. The efficient influence function at $P \in \mathcal{M}$ of this parameter is given by*

$$\begin{aligned} D_j(O, P) \equiv \frac{W_j}{\mathbb{E}_P\left[W_j^2\right]}\Bigg(&\frac{2A-1}{Ag(W) + (1-A)(1-g(W))}\left(Y - \bar{Q}(A, W)\right) \\ &+ \bar{Q}(1, W) - \bar{Q}(0, W) - \Psi_j(P)W_j\Bigg). \end{aligned} \tag{4.5}$$

Estimators We now present nonparametric estimators of the TEM-VIP of Equation (4.2).

One-step and estimating equation estimators. As seen in the previous chapter, the one-step estimator of $\Psi_j(P_0)$, for $j = 1, \ldots, p$, is identical to the estimating equation estimator $\Psi_j^{(\mathrm{EE})}(P_n)$. Again, we present this estimator here for completeness. Let $\bar{Q}_n$ and g_n be, respectively, estimators of $\bar{Q}_0$ and g_0 trained on P_n. Then

$$\Psi_j^{(\mathrm{OS})}(P_n) = \Psi_j^{(\mathrm{EE})}(P_n) \equiv \sum_{i=1}^{n} \frac{W_{ij}}{\sum_{i=1}^{n} W_{ij}^2} \Bigg(\frac{2A_i - 1}{A_i g_n(W_i) + (1 - A_i)(1 - g_n(W_i))} \left(Y_i - \bar{Q}_n(A_i, W_i)\right) + \bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i) \Bigg) .$$

Targeted maximum likelihood estimator. The TML estimator's derivation is slightly more involved. Define the negative log-likelihood loss function for $\bar{Q}$ as

$$L(O; \bar{Q}) \equiv -\log \left\{ \bar{Q}(A, W)^Y \left(1 - \bar{Q}(A, W)\right)^{(1-Y)} \right\} ,$$

and a parametric working submodel for $\bar{Q}$ as

$$\bar{Q}_j(\epsilon)(A, W) \equiv \mathrm{logit}^{-1} \left\{ \mathrm{logit}\, \bar{Q}(A, W) + \epsilon H_j(A, W) \right\} ,$$

where

$$H_j(A, W) \equiv \frac{W_j}{\mathbb{E}_P[W_j^2]} \frac{2A - 1}{Ag(W) + (1 - A)(1 - g(W))} .$$

Now, denoting an initial estimator of $\bar{Q}_0$ trained on P_n by $\bar{Q}_n^0$, we update $\bar{Q}_n^0$ by computing $\epsilon_{n,j}^1$ such that

$$\epsilon_{n,j}^1 = \arg\min_\epsilon \mathbb{E}_{P_n} \left[L(O; \bar{Q}_{n,j}^0(\epsilon)) \right] ,$$

where, though not immediately clear in the notation, $\bar{Q}_{n,j}^0(\epsilon)$ depends directly on $\bar{Q}_n^0$ and indirectly (through $H_j(A, W)$) on g_n and $\sum_i W_{ij}^2/n$, an estimator of $\mathbb{E}_P[W_j^2]$. A tilted conditional outcome estimator is then computed as $\bar{Q}_{n,j}^1 \equiv \bar{Q}_{n,j}^0(\epsilon_{n,j}^1)$. The solution to the above equation, $\epsilon_{n,j}^1$, is the maximum likelihood estimator (MLE) of a univariate logistic regression's slope coefficient obtained by regressing Y on $H_{n,j}(A, W)$ while taking $\bar{Q}_n^0(A, W)$ as an offset. Here, $H_{n,j}$ is the empirical version of H_j, using g_n and $\sum_i W_{ij}^2/n$ in place of g and $\mathbb{E}_P[W_j^2]$, respectively.

We define $P_n^\star$ as the tilted P_n, where $\bar{Q}_n^0$ is replaced by $\bar{Q}_{n,j}^1$. Exploiting a classical result of logistic regression in parametric statistical models [van der Laan and Rubin, 2006b], it follows that $\mathbb{E}_{P_n}[D(O, P_n^\star)] \approx 0$. The TML estimator of the j^{th} TEM-VIP is therefore given by $\Psi_j^{(\mathrm{TML})}(P_n) \equiv \Psi_j(P_n^\star)$.

We highlight that this estimator is appropriate even when the outcome is not restricted to $(0, 1)$. It suffices to shift each of the observed outcomes Y_i, $i = 1, \ldots, n$, by $-\min_i\{Y_i\}$ and to scale them by $\max_i\{Y_i\} - \min_i\{Y_i\}$ prior to computing the TML estimate, and then rescaling the TML estimate by the same quantities. Note too that other loss functions might be used to tilt P_n, like the squared error loss. See Gruber and van der Laan [2010] for a comparison and discussion.

Asymptotic Behavior Next, we study the asymptotic behavior of these absolute TEM-VIP estimators. Note that the asymptotic distributions of $\Psi^{(\text{OS})}(P_n)$, $\Psi^{(\text{EE})}(P_n)$, and $\Psi^{(\text{TML})}(P_n)$ are identical through their dependence on $D_j(O, P_0)$, $j = 1, \ldots, p$. In particular, they are double-robust, meaning they are consistent even when one of the nuisance parameters is inconsistently estimated.

Assumption 4.5. *Conditional outcome estimator consistency:*

$$\|\bar{Q}_n(A, W) - \bar{Q}_0(A, W)\|_2^2 = \int (\bar{Q}_n(a, w) - \bar{Q}_0(a, w))^2 dP_0(a, w) = o_P(1).$$

Assumption 4.6. *Propensity score estimator consistency:*

$$\|g_n(W) - g_0(W)\|_2^2 = \int (g_n(w) - g_0(w))^2 dP_0(w) = o_P(1).$$

Proposition 4.2. *Under A4.1 and A4.2, and either A4.5 or A4.6, $\mathbb{E}_{P_0}[D_j(O, P_n)|P_n] = o_P(1)$ for $j = 1, \ldots, p$. That is, $\Psi^{(OS)}(P_n) \xrightarrow{P} \Psi(P_0)$, and the same is true of the estimating equation and TML estimators.*

Further, these estimators' sampling distribution can be specified under the following assumptions:

Assumption 4.7. *Donsker conditions: There exists a P_0-Donsker class[2] $\mathbb{G}_0$ such that $\mathbb{P}_{P_0}[D_j(O, P_n) \in \mathbb{G}_0] \to 1$ and $\mathbb{E}_{P_0}[(D_j(O, P_n) - D_j(O, P_0))^2|P_n] = o_P(1)$ for each j.*

Assumption 4.8. *Shared rate convergence: $\|\bar{Q}_n(A, W) - \bar{Q}_0(A, W)\|_2 \, \|g_n(W) - g_0(W)\|_2 = o_P(n^{-1/2})$.*

Assumption 4.9. *Bounded covariates: There exists $C \in \mathbb{R}_+$, such that $|W_j| \leq C$ for $j = 1, \ldots, p$.*

Theorem 4.2. *Assuming A4.1, A4.2, A4.7, A4.8, and A4.9, $\sqrt{n}(\Psi_j^{(OS)}(P_n) - \Psi_j(P_0)) = (1/\sqrt{n}) \sum_i D_j(O_i, P_0) + o_P(1)$ for $j = 1, \ldots, p$. This implies that $\sqrt{n}(\Psi_j^{(OS)}(P_n) - \Psi_j(P_0)) \xrightarrow{D} N(0, \mathbb{E}_{P_0}[D_j(O, P_0)^2])$. This result is true of the estimating equation and TML estimators, too.*

A4.7 is a generally unverifiable entropy assumption guaranteeing that the third term on the right-hand side of Equation (4.3) converges to zero in probability. However, it is equivalent to placing weak regularity conditions on $\bar{Q}$ and g that are more transparent. If these nuisance parameters are càdlàg (*continue à droite, limite à gauche*: right continuous, left limits) [Neuhaus, 1971] and have finite supremum and (sectional) variation norms [Gill

[2] A class $\mathcal{F}$ with bounded suprepmum norm is P-Donsker if $\mathcal{F}$ is pre-Gaussian and the empirical process $\mathbb{G}(\mathcal{F})$ converges weakly under $L_\infty(P)$ to the Gaussian process $\mathbb{G}_P(\mathcal{F})$ in n. Here, $\mathbb{G}(\mathcal{F}) = \{\mathbb{G}(f), f \in \mathcal{F}\}$ [Bickel et al., 1993b, van der Laan and Rose, 2011b, Bickel and Doksum, 2015].

et al., 1995], for example, then A4.7 is met [see van der Laan and Gruber, 2016, van der Laan, 2017, for a discussion of these conditions]. Alternatively, nonparametric estimators may be extended to perform a sample-splitting procedure such that this requirement is replaced by even milder conditions [Bickel, 1982, Schick, 1986, Klaassen, 1987, Zheng and van der Laan, 2011, Chernozhukov et al., 2017], like the convergence of $\bar{Q}_n$ or g_n to fixed functionals that are not necessarily equal to $\bar{Q}_0$ and g_0, respectively [Zheng and van der Laan, 2011]. In Chapter 3 we derive such a cross-fitted estimator based on the one-step approach.

A4.8 is satisfied when the nuisance parameter estimators jointly converge at the standard semiparametric rate of $n^{-1/2}$. This so-called "shared rate convergence" condition also allows for one of the nuisance parameters to converge more slowly if the other estimator converges more quickly. When p is small relative to n, A4.8 is typically satisfied by estimating $\bar{Q}_0$ and g_0 with flexible machine learning methods that place few, if any, assumptions on the functional form of these parameters. One such approach is the Super Learner framework [van der Laan et al., 2007b]. In high-dimensional observational settings, however, this convergence property is only met by appealing to smoothness and sparsity assumptions about the nuisance parameters. Examples of these conditions for Random Forests and deep neural networks are outlined in Wager and Athey [2018] and Farrell et al. [2021], respectively. Of note, A4.8 is satisfied regardless of p's size relative to n when either of the nuisance parameters are known, as is the case with g_0 in randomized control trials (RCT). This follows from inspection of the second-order remainder term of Equation (4.3), which equals to zero in this scenario.

We note that A4.7 and A4.8 are satisfied when estimating the nuisance parameters with the Highly Adaptive LASSO under the condition that these parameters are càdlàg and have bounded sectional variation norm [van der Laan, 2017, Bibaut and van der Laan, 2019]. Current implementations of this estimator are currently too computationally demanding for use wit the high-dimensional DGPs considered, however.

The final assumption, A4.9, is a sufficient, technical condition required to bound the second-order remainder of Equation (4.3). While it may appear stringent, and is, for example, not satisfied by covariates generated according to a multivariate Gaussian distribution, we believe that it is generally applicable. Many — if not most — of the random variables studied in the biological, physical, and social sciences are bounded by their very nature or by limitations of measurement instruments. We demonstrate Theorem 4.2's practical robustness to A4.9 in the simulation experiments of Section 4.6 by generating covariates with Gaussian distributions.

We remark that while Theorem 4.2 states that $\Psi^{(\mathrm{OS})}(P_n)$ and $\Psi^{(\mathrm{TML})}(P_n)$ — as well as their cross-fitted counterparts — are asymptotically identical, noticeable differences in behavior are possible in finite samples. This is explored in the simulation study of Section 4.6.

Relative Treatment Effect Modification Variable Importance Parameter

Causal Parameter

While the TEM-VIP of Equation (4.1) is a generally informative assessment of treatment effect modification, situations may arise where a relative TEM-VIP is of greater interest. Examples include scenarios with non-negative outcome variables. A parameter based on the ratio of conditional expected potential outcomes might provide a more expressive metric of treatment effect modification:

$$\frac{\mathbb{E}_{P_{X,0}}\left[Y^{(1)}\middle|W\right]}{\mathbb{E}_{P_{X,0}}\left[Y^{(0)}\middle|W\right]} = \frac{\bar{Q}_{P_{X,0}}(1,W)}{\bar{Q}_{P_{X,0}}(0,W)} \, .$$

The full-data model, $\mathcal{M}_X$, is identical to the one presented in Section 4.2, save that $Y^{(0)}, Y^{(1)} \in \mathbb{R}_+$.

As with the CATE, estimating the conditional parameter above is challenging in high dimensions, making TEM discovery difficult. Again assuming A4.1 and A4.2, we instead propose a TEM-VIP inspired by a GLM of the outcome with a log link function:

$$\Gamma_j^F(P_{X,0}) \equiv \frac{\mathbb{E}_{P_{X,0}}\left[\left(\log \bar{Q}_{P_{X,0}}(1,W) - \log \bar{Q}_{P_{X,0}}(0,W)\right) W_j\right]}{\mathbb{E}_{P_{X,0}}\left[W_j^2\right]} \, . \tag{4.6}$$

Then $\Gamma^F : \mathcal{M}_X \to \mathbb{R}^p$, $\Gamma^F(P_{X,0}) = (\Gamma_1^F(P_{X,0}), \ldots, \Gamma_p^F(P_{X,0}))$ is the target of inference.

Assuming that the expectation of $\log \bar{Q}_{P_{X,0}}(1,W) - \log \bar{Q}_{P_{X,0}}(0,W)$ conditional on any given W_j is linear in W_j, $\Gamma^F(P_{X,0}) = (\Gamma_1^F(P_{X,0}), \ldots, \Gamma_p^F(P_{X,0}))$ is the vector of simple linear regression coefficients produced by regressing the log-ratio of expected conditional potential outcomes against individual covariates. As with $\Psi^F(P_{X,0})$, $\Gamma^F(P_{X,0})$ remains an informative estimand under violations of this linearity assumption in all but pathological scenarios, and can be viewed as assessing the correlation between the log-ratio of potential outcomes and each covariates. As in the absolute TEM-VIP case, W_j is said to be a TEM under this relative VIP if $|\Gamma_j^F(P_{X,0})| > 0$.

Identifiability Through Observed-Data Parameter

Relating $\Gamma^F(P_{X,0})$ to some parameter of P_0 follows directly from the result of Theorem 4.1.

Corollary 4.1. *Under the conditions outlined in Theorem 4.1,*

$$\begin{aligned} \Gamma_j(P_0) &\equiv \frac{\mathbb{E}_{P_0}\left[\left(\log \bar{Q}_0(1,W) - \log \bar{Q}_0(0,W)\right) W_j\right]}{\mathbb{E}_{P_0}\left[W_j^2\right]} \\ &= \Gamma_j^F(P_{X,0}), \end{aligned} \tag{4.7}$$

for $j = 1, \ldots, p$. The observed-data parameter $\Gamma : \mathcal{M} \to \mathbb{R}^p$ defined as $\Gamma(P_0) = (\Gamma_1(P_0), \ldots, \Gamma_p(P_0))$ is therefore equal to the full-data estimand $\Gamma^F(P_{X,0})$.

Inference

Efficient Influence Function To lighten notation, $D_j(O, P)$ is recycled to represent the efficient influence function of $\Gamma_j(P)$ and all other parameters throughout the remainder of the chapter.

Proposition 4.3. *Assume A4.1 and A4.2, and define* $\Gamma_j(P)$ *as in Equation* (4.7) *for* $P \in \mathcal{M}$. *The efficient influence function of this parameter is*

$$D_j(O,P) \equiv \frac{W_j}{\mathbb{E}_P\left[W_j^2\right]}\left(\frac{2A-1}{Ag(W)+(1-A)(1-g(W))}\frac{Y-\bar{Q}(A,W)}{\bar{Q}(A,W)} + \log\frac{\bar{Q}(1,W)}{\bar{Q}(0,W)} - \Gamma_j(P)W_j\right). \tag{4.8}$$

Estimators Nonparametric estimators of $\Gamma(P_0)$ are given next.

One-step and estimating equation estimators. From the von Mises expansion of Equation (4.3), we find that the one-step TEM-VIP estimator for the j^{th} potential TEM is given by

$$\Gamma_j^{(\text{OS})}(P_n) \equiv \sum_{i=1}^{n}\frac{W_{ij}}{\sum_{i=1}^{n}W_{ij}^2}\left(\frac{2A_i-1}{A_ig_n(W_i)+(1-A_i)(1-g_n(W_i))}\frac{Y_i-\bar{Q}_n(A_i,W_i)}{\bar{Q}_n(A_i,W_i)} + \log\frac{\bar{Q}_n(1,W_i)}{\bar{Q}_n(0,W_i)}\right).$$

As with the absolute TEM-VIP, the estimating equation estimator of $\Gamma(P_0)$, $\Gamma_j^{(\text{EE})}(P_n)$, is identical to $\Gamma_j^{(\text{OS})}(P_n)$.

Targeted maximum likelihood estimator. The TML estimator of $\Gamma_j(P_0)$, $\Gamma_j^{(\text{TML})}(P_n)$, is computed using a targeting strategy that is almost identical to that of $\Psi_j^{(\text{TML})}(P_n)$. The only departure from the previously presented procedure occurs in the definition of $H_j(A, W)$. For the relative TEM-VIP, we let

$$H_j(A,W) \equiv \frac{W_j}{\mathbb{E}_P[W_j^2]\bar{Q}(A,W)}\frac{2A-1}{Ag(W)+(1-A)(1-g(W))}.$$

The calculation of the tilted estimator $\bar{Q}_{n,j}^1$ is otherwise unchanged. It then follows that $\Gamma_j^{(\text{TML})}(P_n) \equiv \Gamma_j(P_n^\star)$, where we again stress through notation that $\Gamma_j^{(\text{TML})}(P_n)$ is a plug-in estimator relying on the tilted empirical distribution $P_n^\star$. $P_n^\star$ is identical to P_n save that $\bar{Q}_{n,j}^1$ is used in place of $\bar{Q}_n$.

Asymptotic Behavior As before, we begin the study of $\Gamma^{(\text{OS})}(P_n)$, $\Gamma^{(\text{EE})}(P_n)$, and $\Gamma^{(\text{TML})}(P_n)$'s identical asymptotic behavior with sufficient conditions for consistency.

Proposition 4.4. *If A4.1, A4.2, and A4.5 are satisfied,* $\mathbb{E}_{P_0}[D_j(O, P_n)|P_n] = o_P(1)$ *for* $j = 1, \ldots, p$. *That is,* $\Gamma^{(OS)}(P_n) \xrightarrow{P} \Gamma(P_0)$. *This result holds for the estimating equation and TML estimators as well.*

We contrast Propositions 4.2 and 4.4. Unlike $\Psi^{(\text{OS})}$ and $\Psi^{(\text{TML})}$, $\Gamma^{(\text{OS})}$ and $\Gamma^{(\text{TML})}$ are not doubly robust. Consistent estimation of $\bar{Q}_0$ is required and, baring practical positivity violations, estimation of g_0 has no impact.

Next, the asymptotic linearity of these estimators is established.

Assumption 4.10. *Convergence rate of conditional outcome estimator:* $\|\bar{Q}_n(A, W) - \bar{Q}_0(A, W)\|_2 = o_P(n^{-1/4})$.

Theorem 4.3. *Under A4.1, A4.2, A4.7, A4.8, A4.9, and A4.10,* $\sqrt{n}(\Gamma_j^{(OS)}(P_n) - \Gamma_j(P_0)) = (1/\sqrt{n}) \sum_i D_j(O_i, P_0) + o_P(1)$. *Again, this result applies to the estimating equation and TML estimators, and implies that* $\sqrt{n}(\Gamma_j^{(OS)}(P_n) - \Gamma_j(P_0)) \xrightarrow{D} N(0, \mathbb{E}_{P_0}[D_j(O, P_0)^2])$.

The conditions required for the asymptotic linearity of $\Gamma^{(\text{OS})}(P_n)$ and $\Gamma^{(\text{TML})}(P_n)$ are largely similar to those of $\Psi^{(\text{OS})}(P_n)$ and $\Psi^{(\text{TML})}(P_n)$. The sole difference is that candidate estimators of the conditional expected outcome must converge at a rate no slower than $o_P(n^{-1/4})$. The propensity score estimator, however, may converge at a slower rate so long as A4.8 is satisfied. While this distinction has little impact in observational study settings, the same cannot be said in RCTs. Knowing g_0 does not guarantee the asymptotic linearity of $\Gamma^{(\text{OS})}(P_n)$ and $\Gamma^{(\text{TML})}(P_n)$ — an accurate estimator of $\bar{Q}_0$ is essential.

4.3 Binary Outcomes

Consider the setting identical to that described in the previous section, save that the outcome, Y, is a binary random variable. Noting that $\bar{Q}(A, W) = \mathbb{P}[Y = 1|A, W]$, it follows that all results of Section 4.2 apply to these DGPs. That is, the absolute and relative TEM-VIPs, as well as their respective asymptotically linear estimators, can just as readily be used to detect treatment effect modifiers when the outcome is binary.

4.4 Right-Censored Time-to-Event Outcomes

Returning to the motivating example of the introduction, the discovery of treatment effect modifiers is essential to precision medicine: they delineate patient subgroups, allowing for tailored care. They can also provide mechanistic insight on experimental therapies and improve the success rate of clinical trials. However, the data generated and collected in many therapeutic areas, like oncology, are characterized censored time-to-event outcomes

like time to death or disease recurrence. The TEM-VIPs presented thus far are not readily applicable to this setting.

Problem Setting

Consider n i.i.d. random vectors $\{X_i\}_{i=1}^n$, where $X = (W, A, C^{(0)}, C^{(1)}, T^{(0)}, T^{(1)}) \sim P_{X,0} \in \mathcal{M}_X$. We again define $\mathcal{M}_X$ as a nonparametric statistical model of possible full-data DGPs and denote the true DGP by $P_{X,0}$. As before, W and A are, respectively, the vector of pre-treatment covariates and the binary treatment indicator. Here, $C^{(a)}$ and $T^{(a)}$ correspond, respectively, to the (discrete or continuous) censoring and event times, from which we define the right-censored time-to-event $\tilde{T}^{(a)} = \min\{T^{(a)}, C^{(a)}\}$ and the censoring indicator $\Delta^{(a)} = I(T^{(a)} > C^{(a)})$, under condition $a \in \{0, 1\}$.

Causal parameters of interest in this setting often build upon the conditional survival function $S_{P_{X,0}}(t|a, W) \equiv \mathbb{P}_{P_{X,0}}[T^{(a)} > t|W]$, $a \in \{0, 1\}$. Consider the CATE of the survival probability at time t:

$$\mathbb{E}_{P_{X,0}}\left[S_{P_{X,0}}(t|1, W) - S_{P_{X,0}}(t|0, W)|W\right] .$$

The difference in conditional restricted mean survival times (RMST) [Chen and Tsiatis, 2001, Royston and Parmar, 2011] for time t might be a meaningful target causal parameter too:

$$\begin{aligned}&\mathbb{E}_{P_{X,0}}\left[\min\{T^{(1)}, t\} - \min\{T^{(0)}, t\}\Big|W\right] \\ &\qquad = \mathbb{E}_{P_{X,0}}\left[\int_0^t \left\{S_{P_{X,0}}(u|1, W) - S_{P_{X,0}}(u|0, W)\right\} du\Big|W\right] .\end{aligned}$$

A derivation of the above equality is found in Díaz et al. [2019].

As with the CATE in DGPs with continuous and binary outcomes, however, the recovery of treatment effect modifiers from these parameters is unreliable in high dimensions. We suggest using the TEM-VIPs described in the subsequent subsections instead.

Absolute Treatment Effect Modification Variable Importance Parameter

Causal Parameter

Under A4.1 and A4.2, the following measure of absolute treatment effect modification for time-to-event outcomes can be used:

$$\Psi_j^F(P_{X,0}; t) \equiv \frac{\mathbb{E}_{P_{X,0}}\left[W_j \int_0^t \left\{S_{P_{X,0}}(u|1, W) - S_{P_{X,0}}(u|0, W)\right\} du\right]}{\mathbb{E}_{P_{X,0}}\left[W_j^2\right]} . \tag{4.9}$$

The estimand is then given by $\Psi^F : \mathcal{M}_X \times \mathbb{R}_+ \to \mathbb{R}^p$, $\Psi^F(P_{X,0}; t) = (\Psi_1^F(P_{X,o}; t), \ldots, \Psi_p^F(P_{X,0}; t))$. We reuse "$\Psi$" to emphasize that this is an absolute effect parameter.

Similar to the continuous outcome scenario, $\Psi_j^F(P_{X,0}; t)$ captures the correlation of the difference in conditional RMSTs and the j^{th} covariate, standardized to be on the outcome's scale. $\Psi^F(P_{X,0}; t) = (\Psi_1^F(P_{X,0}; t), \ldots, \Psi_p^F(P_{X,0}; t))$ therefore generally identifies the pre-treatment covariates responsible for the largest differences in expected truncated survival times.

Identifiability Through Observed-Data Parameter

As before, the full data $\{X_i\}_{i=1}^n$ are typically not observable. Define $T = AT^{(1)} + (1-A)T^{(0)}$ and $C = AC^{(1)} + (1-A)C^{(0)}$. We instead have access to $\{O_i\}_{i=1}^n$, a set of n random variables $O = (W, A, \tilde{T}, \Delta) \sim P_0 \in \mathcal{M}$, where W and A are defined as in the full-data model, $\tilde{T} = \min\{T, C\} = A\min\{T^{(1)}, C^{(1)}\} + (1-A)\min\{T^{(0)}, C^{(0)}\}$ is the right-censored time-to-event, and $\Delta = I(T > C)$ is the censoring indicator. Again, P_0 is the true unknown DGP for the observed data O and is fully specified by $P_{X,0}$, and $\mathcal{M}$ is the model of possible observed-data DGPs. Further, let $S_0(t|A, W) \equiv \mathbb{P}_{P_0}[T > t|A, W]$ and $\mathbb{P}_{P_0}[C > t|A, W] \equiv c_0(t|A, W)$ represent the observed conditional survival and censoring functions, respectively.

Sufficient identifiability conditions relating $\Psi^F(P_{X,0}; t)$ to a parameter of the observed-data DGP are provided next.

Assumption 4.11. *No unmeasured exposure-time-to-event confounding: $T^{(a)} \perp A|W$, for $a \in \{0, 1\}$. Unclear how to interpret A in this and the next assumption.*

Assumption 4.12. *No unmeasured time-to-event-censoring confounding: $T^{(a)} \perp C^{(a)}|A, W$, for $a \in \{0, 1\}$.*

Assumption 4.13. *Censoring mechanism positivity: There exists some $\epsilon > 0$ such that $\mathbb{P}_{P_0}[c_0(u|A, W) < 1 - \epsilon] = 1$ for all $u \in (0, t)$.*

Theorem 4.4. *Assuming A4.1, A4.2, A4.4, A4.11, A4.12, and A4.13 hold, we find that*

$$\begin{aligned} \Psi_j(P_0; t) &\equiv \frac{\mathbb{E}_{P_0}\left[W_j \int_0^t \{S_0(u|1, W) - S_0(u|0, W)\}\, du\right]}{\mathbb{E}_{P_0}\left[W_j^2\right]} \\ &= \Psi_j^F(P_{X,0}; t)\ , \end{aligned} \tag{4.10}$$

for $j = 1, \ldots, p$. The observed-data parameter $\Psi : \mathcal{M} \times \mathbb{R}_+ \to \mathbb{R}^p$, $\Psi(P_0; t) = (\Psi_1(P_0; t), \ldots, \Psi_p(P_0; t))$ is equal to $\Psi^F(P_{X,0}; t)$.

Beyond the condition that the covariates be centered and have non-zero variance, the assumptions required by Theorem 4.4 are standard in the causal inference literature for time-to-event parameters [see, for example, Moore and van der Laan, 2011, Benkeser et al.,

2019, Díaz et al., 2019]. A4.11 ensures that the treatment assignment mechanism can be viewed as random, conditional on the covariates. A4.12 requires that survival and censoring times are independent given treatment and covariates. Finally, A4.13 specifies that every random unit has a positive probability of being observed at every time up to and including t.

Inference

Efficient Influence Function The efficient influence function of the estimand in Equation (4.10) is provided below.

Proposition 4.5. *Define $\Psi_j(P;t)$ as in Equation (4.10) for some $P \in \mathcal{M}$ and assume A4.1 and A4.2. The uncentered efficient influence function of $S(t|a,W)$ is given by*

$$d(O,P;t,a) \equiv \frac{I(A=a)S(t|a,W)}{(Ag(W)+(1-A)(1-g(W)))} \int_0^t \frac{I(\tilde{T} \geq u)}{c(u_-|a,W)S(u|a,W)} \left(I(T=u) - \lambda(u|a,W)\right) du + S(t|a,W) \ ,$$

where $\lambda(u|A,W)$ is the conditional survival hazard at time u and u_- denotes the left-hand limit of u [Moore and van der Laan, 2011]. By the functional delta method, the efficient influence function of $\Psi_j(P;t)$ is

$$D_j(O,P;t) \equiv \frac{W_j}{\mathbb{E}_P\left[W_j^2\right]} \left(\int_0^t d(O,P;u,1) - d(O,P;u,0)\ du - \Psi_j(P;t) W_j \right) \ . \tag{4.11}$$

Estimators In practice, for numerical reasons, the integrals in the estimators presented next are approximated by weighted sums.

One-step and estimating equation estimators. It follows immediately from Proposition 4.5 that the one-step and estimating equation estimators of $\Psi_j(P_0;t)$ are defined as

$$\Psi_j^{(\mathrm{OS})}(P_n;t) = \Psi_j^{(\mathrm{EE})}(P_n;t) \equiv \frac{1}{\sum_{i=1}^n W_{ij}^2} \left(\sum_{i=1}^n W_{ij} \int_0^t d(O_i,P_n;u,1) - d(O_i,P_n;u,0) du \right) \ .$$

Targeted maximum likelihood estimator. Let the log-likelihood loss of $\lambda(u|A,W)$ be given by

$$L(O;\lambda,u) = -\log\left\{ \lambda(u|A,W)^{I(T=u)} (1-\lambda(u|A,W))^{1-I(T=u)} \right\} \ .$$

Define the parametric working submodel for $\lambda(u|A,W)$ as

$$\lambda(\epsilon)(u|A,W) = \mathrm{logit}^{-1}\{\mathrm{logit}\lambda(u|A,W) + \epsilon H_j(u|A,W)\} \ ,$$

where

$$H_j(u|A,W) \equiv \frac{W_j(2A-1)S(u|A,W)}{(Ag(W)+(1-A)(1-g(W)))\,\mathbb{E}_P\left[W_j^2\right]} \int_0^u \frac{1}{c(v_-|A,W)S(v|A,W)} dv \ .$$

Denoting the initial estimator of $\lambda_0(u|A,W)$ by $\lambda_n^0(u|A,W)$, we update $\lambda_n^0(u|A,W)$ by computing ϵ_n^1, where

$$\epsilon_{n,j}^1 = \arg\min_\epsilon \mathbb{E}_{P_n}\left[L(O;\lambda_n^0(\epsilon),u)\right] .$$

This empirical expectation is minimized using the MLE of the univariate logistic regression of the event indicators $(1-\Delta)I(\tilde{T}=v)$ on $H_{n,j}(u|A,W)$, for time v ranging from 0 to u and with the initial hazard estimates as an offset. $H_{n,j}$ is the empirical counterpart of H_j, using S_n, g_n, c_n, and $\sum_i W_{ij}^2/n$ in place of S, g, c, and $\mathbb{E}_P[W_j^2]$, respectively. The longitudinal structure of the data need not be considered [Moore and van der Laan, 2011]; the repeated measures are treated as independent when estimating $\epsilon_{n,j}^1$ [Moore and van der Laan, 2011]. $\lambda_{n,j}^1(u|A,W)$ is then defined as $\lambda_n(\epsilon_{n,j}^1)(u|A,W)$. Setting $\lambda_{n,j}^0(u|A,W) \leftarrow \lambda_{n,j}^1(u|A,W)$, this procedure is repeated until $\epsilon_{n,j}^1 \approx 0$.

This procedure for tilting the conditional hazard at time u is performed at each observed time point between 0 and t. These tilted hazards replace their initial counterparts in P_n to form the tilted empirical distribution $P_n^\star$. Noting that $S_{n,j}(t|A,W) = \Pi_{u=1}^t(1-\lambda_{n,j}^1(u|A,W))$, it follows that the TML estimator of $\Psi_j(P_0;t)$ is given by $\Psi_j^{(\text{TML})}(P_n;t) \equiv \Psi_j(P_n^\star,t)$.

Asymptotic Behavior We now consider the asymptotic properties of these estimators.

Assumption 4.14. *Conditional survival estimator consistency:*

$$\|S_n(u|A,W) - S_0(u|A,W)\|_2^2 = \int (S_n(u|a,w) - S_0(u|a,w))^2 dP_0(a,w) = o_P(1)$$

for all $u \in [0,t]$.

Assumption 4.15. *Conditional propensity score estimator and censoring estimator consistency:* $\|g_n(W) - g_0(W)\|_2^2 = o_P(1)$ *and*

$$\|c_n(u|A,W) - c_0(u|A,W)\|_2^2 = \int (c_n(u|a,w) - c_0(u|a,w))^2 dP_0(a,w) = o_P(1)$$

for all $u \in [0,t]$.

Proposition 4.6. $\Psi^{(OS)}(P_n;t) \xrightarrow{P} \Psi(P_0;t)$ *when A4.1, A4.2, and either A4.14 or A4.15 are satisfied. This result also applies to* $\Psi^{(EE)}(P_n;t)$ *and* $\Psi^{(TML)}(P_n;t)$.

Assumption 4.16. *Shared convergence rate:* $\|g_n(W)-g_0(W)\|_2\|S_n(u|A,W)-S_0(u|A,W)\|_2 = o_P(n^{-1/2})$ *for all* $u \in [0,t]$.

Assumption 4.17. *Convergence rate of conditional censoring estimator:* $\|c_n(u|A,W) - c_0(u|A,W)\|_2 = o_P(n^{-1/4})$ *for all* $u \in [0,t]$.

Theorem 4.5. *Assuming that A4.1, A4.2, A4.7, A4.9, A4.16, and A4.17 are met,* $\sqrt{n}(\Psi_j^{(OS)}(P_n;t) - \Psi_j(P_0;t)) = (1/\sqrt{n})\sum_i D_j(O_i,P_0;t) + o_P(1)$. *The same is true for the estimating equation and TML estimators. Again, this implies that* $\sqrt{n}(\Psi_j^{(OS)}(P_n;t) - \Psi_j(P_0;t)) \xrightarrow{D} N(0, \mathbb{E}_{P_0}[D_j(O,P_0;t)^2])$.

Proposition 4.6 states that consistent estimation of the TEM-VIPs is possible if either the conditional survival function is consistently estimated or if the treatment assignment mechanism and the censoring mechanism are consistently estimated. This implies that, in an RCT, consistent estimates of $\Psi(P_0; t)$ only require that the censoring mechanism be consistently estimated. When there is no censoring or censoring is known to be independent of covariates, consistency is guaranteed when $c_0(t|A, W) = c_0(t|A)$ is estimated with the Kaplan-Meier estimator.

Enforcing more stringent conditions on the DGP and the nuisance parameter estimators results in Theorem 4.5. That is, requiring that the entropy constraint of A4.7 is satisfied — or, alternatively, that nuisance parameters are estimated via cross-fitting — and that the nuisance parameters estimators are consistent at the rates given in A4.16 and A4.17 ensures asymptotically normal estimators that are centered around the true parameter value. When the treatment assignment mechanism is known, as in an RCT, then the only necessary consistency rate condition is that of the censoring mechanism. Valid inference is therefore possible even when the conditional survival function is misspecified.

Relative Treatment Effect Modification Variable Importance Parameter

Causal Parameter

As mentioned in Section 4.2, a relative TEM-VIP may be of greater relevance than an absolute TEM-VIP in some contexts. In particular, when treatment effect modification is assessed in terms of conditional probabilities, as is done in this time-to-event setting, a relative measure may be more sensitive. We propose a causal parameter analogous to that of Equation (4.6):

$$\Gamma_j^F(P_{X,0}; t) \equiv \frac{\mathbb{E}_{P_{X,0}}\left[\left(\log S_{P_{X,0}}(t|1, W) - \log S_{P_{X,0}}(t|0, W)\right) W_j\right]}{\mathbb{E}_{P_{X,0}}\left[W_j^2\right]} \,. \tag{4.12}$$

Again, we assume A4.1 and A4.2. Then $\Gamma^F : \mathcal{M}_X \times \mathbb{R}_+ \to \mathbb{R}$,
$\Gamma^F(P_{X,0}; t) = (\Gamma_1^F(P_{X,0}; t), \ldots, \Gamma_p^F(P_{X,0}; t))$ can be interpreted in a similar fashion to the relative TEM-VIP of the continuous outcome DGP. As in Section 4.4, "Γ" is reused to stress that this is a relative parameter.

Identifiability Through Observed-Data Parameter

The causal TEM-VIP $\Gamma^F(P_{X,0}; t)$ is identifiable in the observed data under the conditions outlined in Theorem 4.4. This follows immediately given that $S_0(t|A, W) = S_{P_{X,0}}(t|A, W)$.

Corollary 4.2. *Under the assumptions of Theorem 4.4, it follows that*

$$\begin{aligned}\Gamma_j(P_0;t) &\equiv \frac{\mathbb{E}_{P_0}\left[\left(\log S_0(t|1,W) - \log S_0(t|0,W)\right) W_j\right]}{\mathbb{E}_{P_0}\left[W_j^2\right]} \\ &= \Gamma_j^F(P_{X,0};t)\end{aligned} \tag{4.13}$$

such that $\Gamma : \mathcal{M} \times \mathbb{R}_+ \to \mathbb{R}^p$, $\Gamma(P_0;t) = (\Gamma_1(P_0;t), \ldots, \Gamma_p(P_0;t)) = \Gamma^F(P_{X,0},t)$.

Inference

Efficient Influence Function The efficient influence function of the observed-data parameter presented in Equation (4.13) is given next.

Proposition 4.7. *Assuming A4.1 and A4.2, the efficient influence function of* $\Gamma(P;t)$ *is*

$$\begin{aligned}D_j(O,P;t) \equiv \frac{W_j}{\mathbb{E}_P\left[W_j^2\right]} \Bigg(&\frac{2A-1}{Ag(W) + (1-A)(1-g(W))} \\ &\int_0^t \frac{I(\tilde{T} \geq u)}{c(u_-|A,W)S(u|A,W)} \left(I(T=u) - \lambda(u|A,W)\right) du \\ &+ \log \frac{S(t|1,W)}{S(t|0,W)} - \Gamma(P;t)W_j \Bigg) .\end{aligned} \tag{4.14}$$

Estimators *One-step and estimating equation estimators.* $\Gamma^{(\mathrm{OS})}(P_n;t)$ and $\Gamma^{(\mathrm{EE})}(P_n;t)$, are then given by

$$\begin{aligned}\Gamma^{(\mathrm{OS})}(P_n;t) = \Gamma^{(\mathrm{EE})}(P_n;t) \equiv \frac{1}{\sum_{i=1}^n W_{ij}^2} \sum_{i=1}^n W_{ij} \Bigg(&\frac{2A_i - 1}{A_i g(W_i) + (1-A_i)(1-g(W_i))} \\ &\int_0^t \frac{I(\tilde{T}_i \geq u)}{c(u_-|A_i,W_i)S(u|A_i,W_i)} \\ &\left(I(T_i=u) - \lambda(u|A_i,W_i)\right) du \\ &+ \log \frac{S(t|1,W_i)}{S(t|0,W_i)} \Bigg) .\end{aligned}$$

Targeted maximum likelihood estimator. This estimator employs a conditional hazard estimator tilting procedure similar to that of $\Psi_j^{(\mathrm{TML})}(P_n;t)$. The definition of $H_j(t|A,W)$ is slightly modified:

$$H_j(t|A,W) \equiv \frac{W_j(2A-1)}{\left(Ag(W) + (1-A)(1-g(W))\right)\mathbb{E}_P\left[W_j^2\right]} \int_0^t \frac{1}{c(u_-|A,W)S(u|A,W)} du .$$

Then, given the tilted empirical distribution $P_n^\star$, $\Gamma_j^{(\mathrm{TML})}(P_n;t) \equiv \Gamma_j(P_n^\star;t)$.

Asymptotic Behavior

Proposition 4.8. *If A4.1, A4.2, and A4.14 are satisfied,* $\Gamma^{(OS)}(P_n;t) \xrightarrow{P} \Gamma(P_0;t)$. *The estimating equation and TML estimators share this property, too.*

Assumption 4.18. *Convergence rate of the conditional survival estimator:* $\|S_n(t|A,W) - S_0(t|A,W)\|_2 = o_P(n^{-1/4})$.

Theorem 4.6. *Assuming that A4.1, A4.2, A4.7, A4.9, A4.16, A4.17, and A4.18, are met,* $\sqrt{n}(\Gamma_j^{(OS)}(P_n;t) - \Gamma_j(P_0;t)) = (1/\sqrt{n})\sum_i D_j(O_i,P_0) + o_P(1)$. *It follows that* $\sqrt{n}(\Gamma_j^{(OS)}(P_n;t) - \Gamma_j(P_0;t)) \xrightarrow{D} N(0, \mathbb{E}_{P_0}[D_j(O,P_0;t)^2])$. *This result applies to the estimating equation and TML estimators as well.*

As for the relative TEM-VIP introduced in Equation (4.6), the nonparametric estimators of the estimand in Equation (4.12) are not double-robust. Further, consistent estimation of all nuisance parameters at the typical nonparametric rate is required to ensure the asymptotic linearity of the estimators. For example, in an RCT where censoring is assumed to be completely at random, consistent estimation of the survival function is necessary to produce consistent estimates of $\Gamma(P_0;t)$. If the conditions of Theorem 4.6 are satisfied, however, then asymptotically valid hypothesis testing about the parameter is possible using the Gaussian null distribution.

4.5 Deriving New Treatment Effect Modification Variable Importance Parameters

Readers might find the previous sections repetitive. This is purposeful. Their contents provide a blueprint for defining pathwise differentiable TEM-VIPs based on causal parameters of treatment effects, deriving estimators of these TEM-VIPs, and establishing conditions under which these estimators are regular and asymptotically linear and efficient. We formalize this framework in the following workflow.

1. Select a full-data, pathwise differentiable parameter $\Phi^F(P_X)$ of some treatment effect that is relevant to the problem at hand. For example, we consider the average treatment effect $\mathbb{E}_{P_X}[Y^{(1)} - Y^{(0)}]$ in Section 4.2 for continuous outcomes, and the difference in RMSTs $\mathbb{E}_{P_X}[\min\{T^{(1)},t\} - \min\{T^{(0)},t\}]$ in Section 4.4 for right-censored time-to-event outcomes.

2. Define $f(W)$ such that $\mathbb{E}_{P_X}[f(W)] = \Phi^F(P_X)$. Under A4.1 and A4.2, the TEM-VIP of covariate j is given by $\Theta_j^F(P_X) = \mathbb{E}_{P_X}[f(W)W_j]/\mathbb{E}_{P_X}[W_j^2]$. In Section 4.2, $f(W) = \bar{Q}_{P_X}(1,W) - \bar{Q}_{P_X}(0,W)$, the CATE, and in Section 4.4, $f(W) = \int_0^t S_{P_X}(u|1,W) - S_{P_X}(u|0,W)\,du$, the conditional RMST.

3. Establish the identifiability of the TEM-VIP in the observed-data model. Denoting the observed-data counterparts of Θ_j^F and Φ^F as Θ_j and Φ, respectively, the conditions establishing that $\Theta_j^F(P_X) = \Theta_j(P)$ are virtually identical to the conditions needed for the equality of $\Phi^F(P_X)$ and $\Phi(P)$. The only additional assumption required is that W_j have bounded variance. See Theorems 4.1 and 4.4 for examples.

4. Derive the efficient influence function of the TEM-VIP. This derivation is straightforward, relying on the chain rule and the definition of the efficient influence function for $\Phi(P)$. If the uncentered efficient influence function of $\Phi(P)$ is given by $d(O, P)$ for $O \sim P$, then the efficient influence function of the TEM-VIP, $\Theta_j(P)$, based on $\Phi(P)$ is $W_j/\mathbb{E}_P[W_j^2](d(O, P) - W_j\Theta_j(P))$. Consider the average treatment effect, whose uncentered efficient influence function is $d(O, P) = (2A - 1)/(Ag(A) + (1 - A)(1 - g(A)))(Y - \bar{Q}(A, W)) + \bar{Q}(1, W) - \bar{Q}(0, W)$. Using the previous formula, the efficient influence function of the absolute TEM-VIP with a continuous or binary outcome is that of Proposition 4.1, where the generic Θ_j is replaced by Ψ_j. Similarly, the uncentered efficient influence function of the log-ratio of expected conditional potential outcomes is given by $d(O, P) = (2A-1)/(Ag(A)+(1-A)(1-g(A)))(Y-\bar{Q}(A, W))/\bar{Q}(A, W)+\log(\bar{Q}(1, W)/\bar{Q}(0, W))$. The efficient influence function of the relative TEM-VIP for continuous and binary outcome settings is given in Proposition 4.3, where $\Theta_j = \Gamma_j$.

5. The one-step, estimating equation, and TML estimators can then be derived from the TEM-VIPs efficient influence function. Examples are found in Sections 4.2 and 4.4.

6. These estimators' asymptotic properties are identical to those of the nonparametric efficient estimators of Φ, assuming that the potential treatment effect modifiers are bounded. Again, examples are provided in Sections 4.2 and 4.4.

4.6 Simulation Studies

Next, we investigate the finite-sample performance of the proposed one-step and TML estimators for a subset of the previously introduced estimands. Recall that the one-step estimator is obtained by subtracting the empirical EIF from the plug-in estimator — and is equal, in the settings considered here, to the estimating equation estimator — and that the TML estimator is derived by first tilting the nuisance parameter estimators to ensure that the mean of the empirical EIF is negligible, and then using these updated estimators in the plug-in estimator. The one-step and TML estimators are implemented in the `unihtee` `R` software package [R Core Team, 2023], available at github.com/insightsengineering/unihtee and to be submitted to the Comprehensive R Archive Network (CRAN). These estimators' empirical absolute bias, variance, and Type I error rates are evaluated in two observational study scenarios — one with a continuous outcome and another with a binary outcome — and one RCT setting with a time-to-event outcome. These simulation experiments rely on the `simChef` `R`

package's simulation study framework [Duncan and Tang, 2022]. Code for reproducing these simulations is made available at github.com/PhilBoileau/pub_temvip-framework.

The nonparametric estimators' capacity to recover treatment effect modifiers is compared to that of Tian et al. [2014] and Chen et al. [2017]'s (augmented) modified covariates methods. These methods are among the few that enable treatment effect modification discovery in high-dimensional data under a variety of DGPs — albeit requiring stringent assumptions like sparsity and negligible correlation structure among pre-treatment covariates. We stress, however, that their primary goal is not the recovery of these treatment effect modifiers, but CATE estimation. The modified covariates approach estimates the CATE by cleverly transforming the outcome such that only the treatment-covariate interactions in a GLM need be modeled. The augmented modified covariates procedure models this transformed outcome as a function of all covariates to improve efficiency. Both procedures can incorporate propensity score weights to improve estimation in observational study scenarios. TEM discovery is possible when employing modeling strategies with built-in feature selection capabilities; we model the transformed outcome-covariates relationships with a linear model and fit this model with the LASSO [Tibshirani, 1996]. Variables are classified as TEMs when their estimated treatment-covariate interaction coefficients are non-zero. These methods are implemented in the `personalized` R package [Huling and Yu, 2021].

Continuous Outcome, Observational Study

The first DGP we consider has a continuous outcome Y, high-dimensional covariates W, and mimics an observational study, in that treatment status A is an unknown function of W:

$$\begin{aligned} W &\sim N(0, I_{500\times 500}) \\ A|W &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\frac{1}{4}\left(W_1 - W_2 + W_3\right)\right)\right) \\ Y|A, W &\sim 1 + 2\left|\sum_{j=1}^{5} W_j\right| + (5A - 2)\sum_{j=1}^{5} W_j + \epsilon \ , \end{aligned}$$

where $\epsilon \sim N(0, 1/2)$. Note that the treatment assignment mechanisms used here and in the following subsections were chosen to respect Assumption A4.4. Indeed, the estimators — particularly the TML estimators — presented in Sections 4.2 and 4.4 exhibit extreme variability in the presence of *practical positivity violations*. Practical positivity violations materialize in finite samples when the estimated probability of receiving treatment is negligible, and can occur even when the positivity assumption of A4.4 is satisfied.

We take as target of inference the absolute TEM-VIPs of Equation (4.1). We consider five sample sizes: $n = 125, 250, 500, 1{,}000$, and 2,000. Two hundred replicates are simulated at each sample size.

We consider the one-step and TML estimators of this parameter where $\bar{Q}_0$ and g_0 are estimated using the Super Learner algorithm of van der Laan et al. [2007b] implemented in

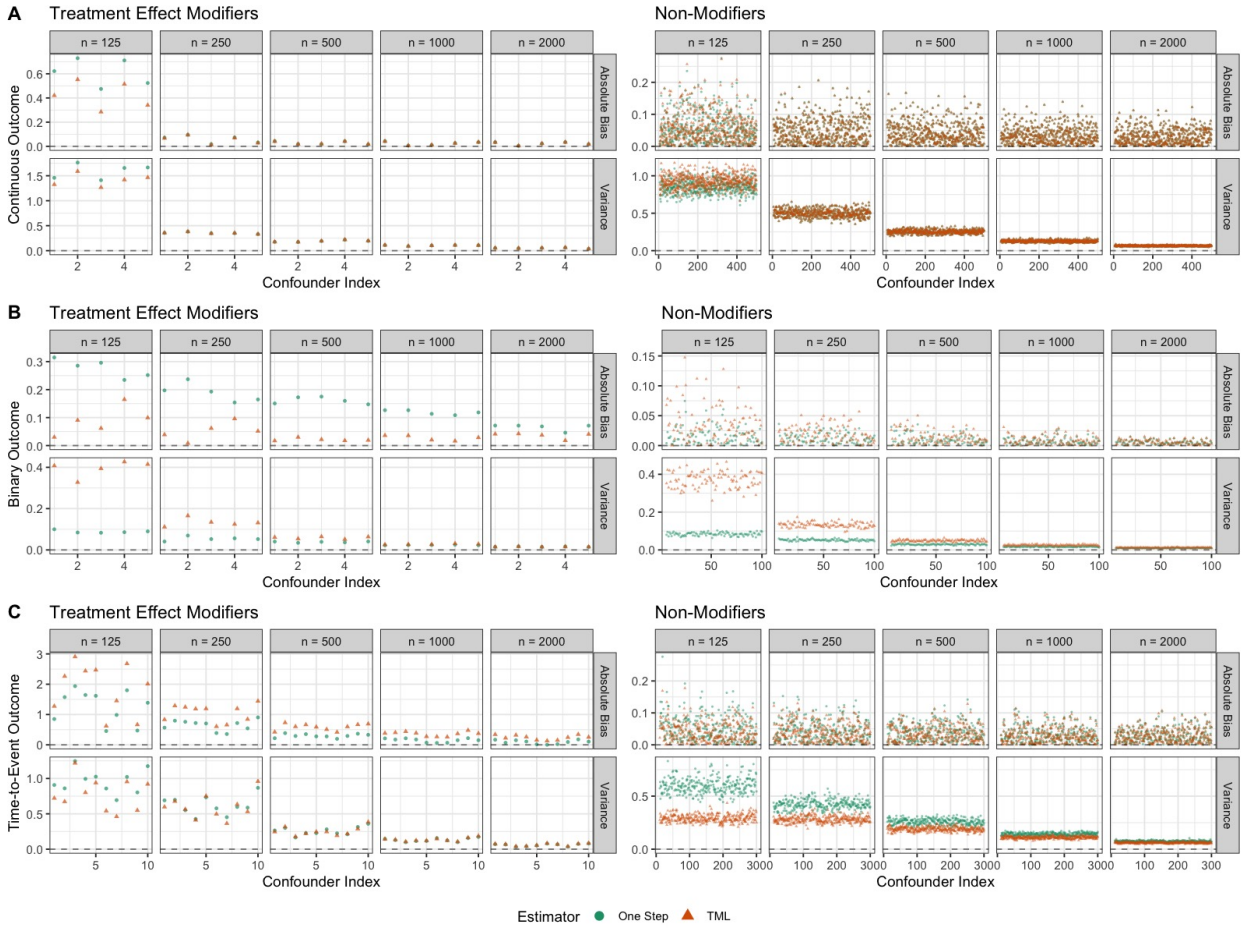


Figure 4.1: *Empirical bias and variance of one-step and TML estimators.* The empirical bias and variance of the one-step and TML estimators are stratified by DGP, treatment modifier status, and sample size (note the difference in y-axis scales between modifiers and non-modifiers). Two hundred replicates were simulated to compute the values in each scenario.

the `sl3` R package [Coyle et al., 2021]. This algorithm computes the convex combination of nuisance parameter estimators, referred to as base learners, that optimizes the cross-validated risk for the squared error and negative log-likelihood loss function for the conditional outcome and propensity score, respectively. For $\bar{Q}_0$, the base learners are comprised of the LASSO [Tibshirani, 1996], ridge regression [Hoerl and Kennard, 1970], elastic net [Zou and Hastie, 2005], and multivariate adaptive regression splines (MARS) [Friedman, 1991] estimators with main and treatment-covariate interaction terms, as well as Random Forests [Breiman, 2001]. For g_0, we consider the LASSO, ridge regression, elastic net, MARS, and Random Forests. The modified covariates method and its augmented counterpart estimate g_0 using LASSO, and employ the identity link function to estimate the association of the covariates and treatment on the outcome.

Figure 4.1A presents the empirical absolute bias and variance of the one-step and TML estimators. Both exhibit a small empirical bias for the TEMs for $n = 125$, but are otherwise approximately unbiased at all other sample sizes. These estimators' variances are virtually identical at all sample sizes, and rapidly decrease as sample size increases. The bias and variance for non-TEMs (covariates indices 6 to 500) are similarly negligible for both estimators in all sample sizes.

We next evaluate these estimators' ability to distinguish covariates that modify the effect of treatment from those that do not. The empirical false discovery rate (FDR), true negative rate (TNR), and true positive rate (TPR) are computed at each sample size. The FDR reports the proportion of incorrectly classified covariates among the set of predicted TEMs. The TNR and TPR measure the proportion of correctly classified non-TEMs and TEMs, respectively. Using nominal 5%-level, two-sided Wald-type hypothesis tests and accounting for multiple testing using the FDR-controlling approach of Benjamini and Hochberg [1995], we expect the one-step and TML estimators to achieve a 5% FDR in the largest sample sizes. The one-step and TML estimators' classification are compared to those of the modified covariates and augmented modified covariates methods. Again, variables with non-zero estimated treatment-covariate interaction coefficients are labeled as TEMs.

Of the four methods considered, only the one-step and TML estimators approximately control the FDR at the nominal level in all sample sizes (Figure 4.2A). The (augmented) modified covariates methods, on the other hand, maintain an FDR near 75%. Their performance does not improve as a function of n. Trends in the methods' FDRs are elucidated by their TNRs and TPRs. The one-step and TML estimators produce a near-perfect TNR while maintaining a competitive TPR. The augmented modified covariates procedure has a TPR near 100% in all sample sizes, yet has TNRs marginally lower than the one-step and TML estimators. The modified covariates method produces similar TNRs to its augmented counterpart, but has poorer TPRs. The parametric methods' inability to reliably classify TEMs might be due to the non-linearity of the expected conditional outcome or the number of features relative to the sample size.

Binary Outcome, Observational Study

We consider another observational DGP, this time with a binary outcome and a moderate number of correlated covariates:

$$
\begin{aligned}
W \sim N(0, \Sigma_{100\times 100}),\ \Sigma_{ij} &= \begin{cases} 1, & i = j \\ 0.1|i-j|^{-1.8}, & \text{otherwise} \end{cases} \\
A|W &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\frac{1}{4}\left(W_1 + W_2 + W_3\right)\right)\right) \\
Y|A, W &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left(1 - 2A + \sum_{j=1}^{5} W_j + \left(A - \frac{1}{2}\right)\sum_{j=1}^{5} W_j\right)\right) .
\end{aligned}
$$

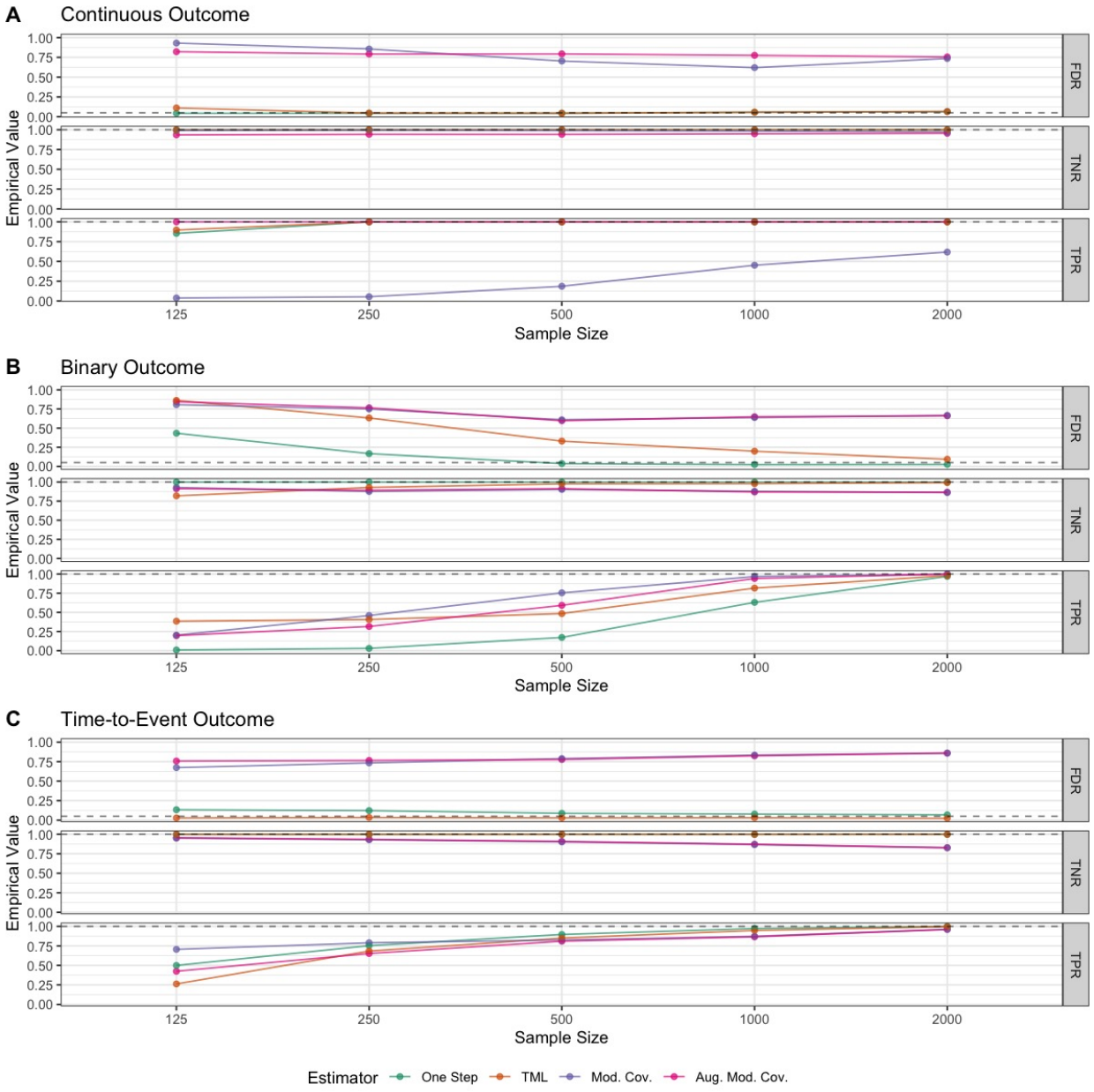


Figure 4.2: *TEM classification results.* The one-step, TML, modified covariates, and augmented modified covariates estimators' capacities to correctly identify TEMs from the set of covariates are measured in terms of the FDR, TPR, and TNR. These metrics are stratified by DGP and sample size. Two hundred replicates were simulated to compute the values in each scenario.

Here, Σ is a 100×100 Toeplitz matrix, so that the pre-treatment covariates' correlation structure imitates that of spatial or temporal data. Again, care is taken to avoid practical positivity violation issues.

We benchmark the estimation of the relative TEM-VIP presented in Equation (4.6). The true parameter values are approximated using Monte Carlo methods. Again, 200 replicates were simulated for each of $n = 125, 250, 500, 1{,}000$, and 2,000. The corresponding one-step and TML estimators are compared and their nuisance parameters are estimated using Super Learners, with the same base learners for $\bar{Q}_0$ and g_0 as in the continuous-outcome example. The (augmented) modified covariate methods again rely on the LASSO for propensity score estimation, and use the logistic link function to model the outcome conditional on the potential TEMs and treatment.

The empirical bias and variance of the one-step and TML estimators are provided in Figure 4.1B. Among the TEMs, the one-step estimator exhibits more finite-sample bias than the TML estimator, though this bias decreases as sample size increases. The TML estimator, however, has noticeably greater finite-sample variance than the one-step for $n = 125, 250$. Among the non-TEMs (pre-treatment covariates indexed 6 through 100), these estimators have similar bias. Again, however, the TML estimator has greater variance in the smaller sample sizes.

The empirical FDR, TNR, and TPR of the one-step and TML estimators, as well as those of the (augmented) modified covariates methods are presented in Figure 4.2B. Only the one-step estimator reliably controls the FDR at the 5% level at sample sizes of 500 and above. This is seemingly due to the estimator's conservative behavior: It achieves a near-perfect TNR at all sample sizes, but has the lowest TPR of all estimators regardless of sample size. The TML estimator fails to control the FDR at the desired levels in all sample sizes, though the FDR decreases with sample size and is nearly controlled at $n = 2{,}000$. The poor FDR of the TML estimator relative to the one-step estimator may be due to the latter's increased variability, exhibited in Figure 4.1B. The (augmented) modified covariates methods tend to perform similarly: their FDR hovers around 75% at all sample sizes, their TNR decreases marginally as n increases, and their TPRs are generally higher than those of the nonparametric estimators. Given that sparsity and linearity assumptions are satisfied, the lackluster FDR control of the (augmented) modified covariates procedures might be attributed to violations of the Irrepresentable Condition [Zhao and Yu, 2006] — the covariates' correlation structure is too complex.

Right-Censored Time-to-Event Outcome, Randomized Control Trial

Next, we simulate RCT data with known treatment assignment mechanism, a discrete right-censored time-to-event outcome, and a duration of 10 time units. Recall that $O = (W, A, \tilde{T}, \Delta)$, where W and A are defined as before, $\tilde{T}$ is the right-censored time-to-event,

and Δ is the censoring indicator. The simulation generative model is given by

$$\begin{aligned}
W &\sim N(0, \Sigma_{300\times 300}) \\
A &\sim \text{Bernoulli}(1/2) \\
C|A, W &\sim \min\left\{\text{Negative Binomial}\left(1, \text{logit}^{-1}\left(5 + A + W_1\right)\right), 10\right\} \\
T|A, W, C &\sim \text{Negative Binomial}\left(1, \text{logit}^{-1}\left(-2 - A + (10A - 5)\sum_{j=1}^{10} W_j\right)\right) \\
\tilde{T} &= \min\left\{T, C\right\} \\
\Delta &= I(T > c),
\end{aligned}$$

where the covariates' covariance matrix Σ is block-diagonal, with each block corresponding to ten moderately correlated features. This correlation structure loosely mimics the expression levels of a collection of genes.

The estimand is defined as the absolute TEM-VIP of Equation (4.9) at time $t = 9$. Again, the true parameter values are approximated through Monte Carlo methods. The one-step and TML estimators' conditional censoring hazard function is estimated by the LASSO and their conditional survival hazard function is estimated by the LASSO augmented with treatment-covariate interaction terms. The propensity scores of these nonparametric estimators and the (augmented) modified covariates methods are fixed at 1/2, as in a 1:1 RCT. Penalized Cox proportional hazards models are used by the parametric methods to model the conditional survival hazard. We highlight that our simulation DGP satisfies the proportional hazards and non-informative censoring assumptions, but that its covariates possess a complex correlation structure. This might worsen the (augmented) modified covariate methods' treatment effect modifier classification performance.

Figure 4.1C presents the one-step and TML estimators' empirical biases and variances. As for the binary DGP, both estimators are biased for the TEMs (indices 1–10) at all sample sizes, but approximately unbiased for all non-TEMs. As expected, however, the empirical bias associated with the TEMs decreases with sample size, and is negligible when $n = 2{,}000$. The empirical variances of these estimators behave as expected, too: they decrease with increasing sample size. The TML estimator's empirical variances are generally smaller than those of the one-step estimator.

The FDR, TNR, and TPR of all methods considered are reported in Figure 4.2C. The TML estimator is the only procedure to control the FDR at the nominal 5% level, while the one-step estimator possesses an FDR of approximately 10% for $n = 125, 250$, and which slowly decreases to the nominal rate by $n = 2{,}000$. The (augmented) modified covariates approaches result in empirical FDRs that grow with sample size, from approximately 70% for $n = 125$ to 90% for $n = 2{,}000$. The parametric methods' behavior with respect to the FDR might be explained by the relationship between their TNR and sample size: as sample size increases, they produce a greater amount of false positives. The nonparametric estimators, however, maintain a near-perfect TNR at all sample sizes. All procedures perform similarly with respect to the TPRs in all but the smallest sample size.

4.7 Application

We apply our framework to a clinical trial dataset with a right-censored time-to-event outcome. This analysis, as well as the results of the simulation studies, can be reproduced with the code found in this public repository: github.com/PhilBoileau/pub_temvip-framewor.

Trastuzumab is a monoclonal antibody targeting the *HER2* oncogene that demonstrably improves the clinical outcomes of breast cancer patients whose tumors over-express this gene. Improvement is not uniform, however: some patients are resistant to this therapy. Identifying biomarkers that predict response to trastuzumab is therefore of great interest [Loi et al., 2014].

Loi et al. [2014] make available a subset of patients enrolled in the FinHER clinical trial (GSE47994), a study comparing docetaxel and vinorelbine — chemotherapies — as adjuvant treatment for early-stage breast cancer [Joensuu et al., 2006]. Patients with over-expressed *HER2* disease were additionally randomized to receive either nine weekly trastuzumab infusions or no trastuzumab. Loi et al. [2014] provide the quality controlled, normalized gene expression data and relevant clinical information for 201 of these patients. Taking as outcome distant disease-free survival, defined as the time interval between the date of randomization and the date of first cancer recurrence or death, if prior to recurrence, we consider the 500 most variable genes for the purpose of TEM discovery.

Traditional approaches to this task rely on Cox proportional hazards models. For example, a penalized regression of the outcome on the treatment, genes, treatment-gene interactions, and pre-treatment covariates like age and chemotherapy could be fit, and the genes with non-zero estimated interaction coefficients would be classified as TEMs. This is equivalent to the augmented modified covariates approach of Tian et al. [2014]. Alternatively, individual regressions for each gene of the outcome conditioning on treatment, gene, pre-treatment covariates, and the treatment-gene interaction could be fitted. Genes with significant treatment-gene interactions would be reported as TEMs. However, both approaches perform inference about conditional parameters, the hazards ratio, while we aim to learn about parameters that reflect population-level information about treatment effect heterogeneity. Verifying the proportional hazards assumption is also impractical given the number of potential TEMs considered.

We instead use our framework, taking as estimand the RMST-based TEM-VIP of Equation (4.9). Patients' distant disease-free survival times are discretized into 6-month intervals for computational convenience. We use the TML estimator since the previous simulation experiments suggests that it controls the Type I error rate better than the one-step estimator at this sample size. Its element-wise variance is also likely lower. Given that previous evidence suggests possible higher-order interactions between patients' chemotherapy regimen, trastuzumab, and biomarkers [Loi et al., 2014], we estimate the conditional failure and censoring hazards using a Super Learner made up of the penalized generalized linear models using the logit link and possessing terms for the treatment, genes, and treatment-gene interactions, Random Forests, and XGBoost [Chen and Guestrin, 2016]. This procedures takes approximately 20 minutes to run on a personal computer with a single core of an Apple

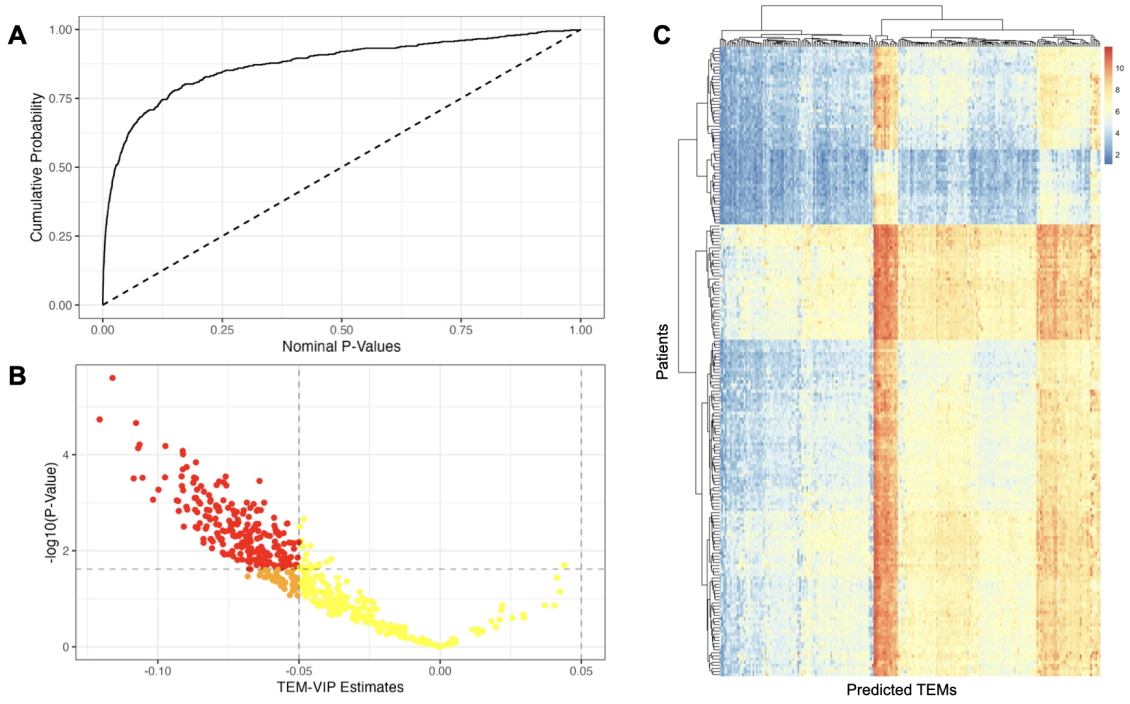


Figure 4.3: *FinHER clinical trial data analysis results.* **A** Empirical cumulative distribution function (eCDF) of nominal p-values. The dotted line corresponds to the eCDF under the null (a Uniform($[0, 1]$) distribution). **B** Volcano plot of the 500 most variable genes' TEM-VIP estimates and associated nominal p-values. Yellow genes are deemed unimportant due to their small estimated effect sizes and larger p-values; orange genes possess a meaningful estimated effect but fail to achieve the adjusted p-value cutoff; red genes are significant at the 5% FDR level and have large estimated TEM-VIPs. **C** The log-transformed gene expression data of genes with meaningful effect estimates are used to cluster patients. Hierarchical clustering with complete linkage is used for patients and identified TEMs alike.

M1 CPU. Parallelization can reduce this runtime further. We note that similar results are produced by directly estimating the nuisance parameters with Random Forests or XGBoost, though at the expense of an objective choice of nuisance estimators otherwise facilitated by the Super Learner estimator.

In this analysis, we have sought to dichotomize pre-treatment covariates into TEMs and non-TEMs based on the value of their estimated TEM-VIP. However, as expected, there seems to be a continuum in the biomarkers' capacity to influence the effect of treatment, in terms of both statistical significance and biological effect size. This can be seen in the empirical cumulative distribution function (eCDF) of the nominal p-values (Figure 4.3A) and in the volcano plot (Figure 4.3B). Hypothesis testing alone, with a null of $\Psi(P_0) = 0$, may therefore not be adequate. As in differential expression studies in transcriptomics, one can instead leverage the volcano plot and deem a biomarker of clinical interest if it is

Table 4.1: Top five selected TEMs

	Gene	Estimate	Std. Err.	Adj. p-Value
1	EPPK1	-0.116	0.025	0.001
2	NDUFB3	-0.121	0.028	0.004
3	BNIP3L	-0.108	0.025	0.004
4	PNKD	-0.106	0.027	0.006
5	DUSP4	-0.097	0.024	0.006

significant at the 5% FDR level *and* if its absolute estimated TEM-VIP is larger than 0.05 (for each unit increase in log2 gene expression, a TEM-VIP equal to 0.05 in this analysis approximately corresponds to an expected difference in RMST of about 18 days). There are 220 such biomarkers for the FinHER clinical trial. Alternatively, if one is interested only in modifications above a certain magnitude m, one could define the null hypothesis for the j^{th} biomarker as $|\Psi_j(P_0)| \leq m$. The (adjusted) p-values obtained from these tests could then be used to produce a ranked list of biomarkers for follow-up analyses. The above considerations highlight the importance of thinking carefully and critically about how to translate the biological question of interest into a statistical inference question, including defining what constitutes a meaningful effect size.

Now, the five genes with the smallest p-values from among the clinically meaningful biomarkers are presented in Table 4.1. All have previously been linked to breast cancer, and their estimated effects are generally in the direction expected by the literature. Increased *EPPK1* expression has been linked to estrogen-related receptor γ, which is associated with breast cancer growth suppression [Ariazi et al., 2002, Tiraby et al., 2011]. A meta-analysis of 11 genome-wide association studies found that a single nucleotide polymorphism in a *NDUFB3* promoter was significantly associated estrogen receptor negative breast cancer [Couch et al., 2016]. Moussay et al. [2011] found that *BNIP3L* upregulation is associated with TNFα stimulation, which is associated with trastuzumab resistance [Mercogliano et al., 2017]. Evidence suggests that overexpression of *MR-1S*, an isomer of *PNKD* associated with disordered cell differentiation, malignant transformation initiation, and accelerated metastasis, is therefore a potential therapeutic target of breast cancer [Wang et al., 2018]. Finally, Menyhart et al. [2017] found that increased expression of *DUSP4* correlates with increased resistance to trastuzumab.

We also present the log-transformed gene expression of the features with clinically meaningful TEM-VIP estimates in Figure 4.3C. We should expect them to define patient subgroups if these biomarkers truly modify the effect of treatment. Indeed, these genes' expression data produce multiple distinct patient clusters. We refrain from interpreting Figure 4.3C any further, however, considering it solely a diagnostic tool. Using patients' outcomes and biomarkers to compute TEM-VIP estimates, then relying on these estimates to data-adaptively define subgroups in the same data may cause overfitting. These results would ideally be validated on an external dataset, though, as is often the case with openly-

accessible clinical trial data, none are available. This might motivate extensions to this TEM discovery framework that support valid inference about both TEM-VIPs and patient subgroups using the same data.

4.8 Discussion

We propose several causally interpretable TEM-VIPs in full-data models, establish identifiability conditions to relate them to parameters of observed-data distributions, derive accompanying nonparametric estimators, and study these estimators' asymptotic behavior. Under non-stringent conditions on the DGPs and nuisance parameter estimators, we find that these estimators are consistent. Imposing a few additional assumptions results in efficient, asymptotically linear estimators that permit straightforward hypothesis testing about the corresponding TEM-VIPs. A general workflow for creating new TEM-VIPs and deriving associated nonparametric estimators is provided.

Simulation experiments demonstrate that the estimators' behavior approximates their established theoretical guarantees in realistic DGPs and for moderate sample sizes. As an additional validation of our methodology, we attempted to identify TEMs in a publicly available clinical trial dataset. These data were originally collected to assess the effect of a monoclonal antibody therapy, trastuzumab, on breast cancer patients. Many genes were classified as TEMs, and a literature review of the top-ranked genes suggests that they are associated with breast cancer. Indeed, a number of these TEMs are known biomarkers of trastuzumab resistance. A diagnostic plot of the predicted TEMs' expression data further suggests that they may be used to define patient subgroups, but this must be validated with external data.

This work gives rise to several research directions. The framework outlined in Section 4.5 permits the derivation of bespoke pathwise differentiable TEM-VIPs and accompanying nonparametric efficient estimators. In particular, researchers working in the biotechnology and pharmaceutical industries can perform inference about TEM-VIPs derived from estimands used in clinical trials. Such heterogeneous treatment effect analyses would closely track the statistical guidelines enforced by regulatory authorities, like those of the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use for clinical trials [International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2019]. This framework for TEM inference might also support statistically rigorous subgroup discovery. TEMs identified using our methodology could be used to cluster observations (i.e., patients), and, subsequently, treatment effects could be estimated within these groups. Whether there exists a sound approach that permits the application of this workflow to a single dataset, perhaps building on recent advances in post-selection inference, should be investigated. Future work might also determine whether these TEM-VIP estimators improve treatment rule estimation procedures by acting as variable filters. That is, only pre-treatment covariates with TEM-VIP estimates significantly different from zero would be used, along with known confounders, to learn the treatment

rule. Doing so would increase the interpretability of the rule and might improve estimation in high-dimensional regimes.

4.9 Proofs

Theorem 4.1

Proof. It follows immediately from A4.3 and A4.4 that $\bar{Q}_{P_{X,0}}(A,W) = \bar{Q}_0(A,W)$. Then $\Psi^F(P_{x,0}) = \Psi(P_0)$. □

Proposition 4.1

Proof. Using the generic definition provided in Equation (4.2), the efficient influence function of $\Psi_j(O,P)$ is

$$\begin{aligned} D_j(O,P) &= \frac{d}{d\epsilon}\Psi_j(P_\epsilon)\big|_{\epsilon=0} \\ &= \frac{W_j}{\mathbb{E}_P\left[W_j^2\right]}\left(\frac{I(A=1)}{g(W)}\left(Y-\bar{Q}(A,W)\right)+\bar{Q}(1,W)\right. \\ &\qquad \left. -\frac{I(A=0)}{1-g(W)}\left(Y-\bar{Q}(A,W)\right)+\bar{Q}(0,W)-\Psi_j(P)W_j\right) . \end{aligned}$$

□

Proposition 4.2

Proof. From the definition of D_j given by Equation (4.5), we find that

$$\begin{aligned} \mathbb{E}_{P_0}\left[D_j(O,P)\right] \propto \mathbb{E}_{P_0}\left\{W_j\left(\left(\frac{g_0(W)}{g(W)}-1\right)\left(\bar{Q}_0(1,W)-\bar{Q}(1,W)\right)\right.\right. \\ \left.\left. -\left(\frac{1-g_0(W)}{1-g(W)}-1\right)\left(\bar{Q}_0(0,W)-\bar{Q}(0,W)\right)\right)\right\} . \end{aligned}$$

It follows immediately that $\mathbb{E}_{P_0}[D_j(O,P)] = 0$ when $g = g_0$ or $\bar{Q} = \bar{Q}_0$ for $j = 1,\ldots,p$. □

Theorem 4.2

Proof. Asymptotic linearity of $\Psi^{(\mathrm{OS})}$ and $\Psi^{(\mathrm{TML})}$ are achieved when the third and fourth terms in the von Mises expansion of Equation (4.3) converge in probability to 0. Under A4.7, $D(O,P_n) \in \mathcal{G}_0$ with probability tending to one which implies that $\mathbb{E}_{P_0}[(D_j(O,P_n) - D_j(O,P_0))^2] = o_P(1)$. It follows that the third term of the von Mises expansion is $o_P(1)$.

What remains is to bound the error term. Similar to the proof for the cross-fitted estimator in the previous chapter, we find that

$$
\begin{aligned}
-R(P_0, P_n) = \frac{1}{\mathbb{E}_{P_0}\left[W_j^2\right]} \mathbb{E}_{P_0} \Bigg[& W_j \left(\frac{g_0(1,W)}{g_n(1,W)} - 1 \right) \left(\bar{Q}_0(1,W) - \bar{Q}_n(1,W) \right) \\
& - W_j \left(\frac{g_0(0,W)}{g_n(0,W)} - 1 \right) \left(\bar{Q}_0(0,W) - \bar{Q}_n(0,W) \right) \Bigg] \\
\leq \frac{1}{\mathbb{E}_{P_0}\left[W_j^2\right]} \Bigg(& \left| \mathbb{E}_{P_0} \left[W_j \left(\frac{g_0(1,W)}{g_n(1,W)} - 1 \right) \left(\bar{Q}_0(1,W) - \bar{Q}_n(1,W) \right) \right] \right| \\
& + \left| \mathbb{E}_{P_0} \left[W_j \left(\frac{g_0(0,W)}{g_n(0,W)} - 1 \right) \left(\bar{Q}_0(0,W) - \bar{Q}_n(0,W) \right) \right] \right| \Bigg) \\
\leq \frac{1}{\mathbb{E}_{P_0}\left[W_j^2\right]} \Bigg(& \mathbb{E}_{P_0} \left[W_j^2 \left(\frac{g_0(1,W) - g_n(1,W)}{g_n(1,W)} \right)^2 \right]^{1/2} \\
& \mathbb{E}_{P_0} \left[\left(\bar{Q}_0(1,W) - \bar{Q}_n(1,W) \right)^2 \right]^{1/2} \\
& + \mathbb{E}_{P_0} \left[W_j^2 \left(\frac{g_0(1,W) - g_n(1,W)}{g_n(0,W)} \right)^2 \right]^{1/2} \\
& \mathbb{E}_{P_0} \left[\left(\bar{Q}_0(0,W) - \bar{Q}_n(0,W) \right)^2 \right]^{1/2} \Bigg) \\
\overset{\text{a.s.}}{\leq} \frac{C^2}{\mathbb{E}_{P_0}\left[W_{j^2}\right]} \Bigg(& \mathbb{E}_{P_0} \left[\left(\frac{g_0(W)}{g_n(W)} - 1 \right)^2 \right]^{1/2} \\
& \mathbb{E}_{P_0} \left[\left(\bar{Q}_0(1,W) - \bar{Q}_n(1,W) \right)^2 \right]^{1/2} \\
& + \mathbb{E}_{P_0} \left[\left(\frac{1 - g_0(W)}{1 - g_n(W)} - 1 \right)^2 \right]^{1/2} \\
& \mathbb{E}_{P_0} \left[\left(\bar{Q}_0(0,W) - \bar{Q}_n(0,W) \right)^2 \right]^{1/2} \Bigg) .
\end{aligned}
$$

The last inequality follows from A4.9. A similar bound applies to $R(P_0, P_n)$. The remainder term of Equation (4.3) is therefore $o_P(1)$ under the conditions of A4.8.

It follows, applying the central limit theorem to the first term of the von Mises expansion, that $\sqrt{n}(\Psi_j^{(\text{OS})}(P_n) - \Psi_j(P_0) \xrightarrow{D} N(0, P_0 D_j(O, P_0))$. The same is true for $\Psi_j^{(\text{TMLE})}(P_n)$.

□

Corollary 4.1

Proof. The conditions outlined in Theorem 4.1 imply that $\bar{Q}_{P_{X,0}}(A,W) = \bar{Q}_0(A,W)$. It follows immediately that $\Gamma^F(P_{X,0})$ is equal to $\Gamma(P_0)$. □

Proposition 4.3

Proof. Using the same point mass contamination approach, we obtain the following efficient influence function for $\Gamma_j(O,P)$:

$$\begin{aligned} D_j(O,P) &= \frac{d}{d\epsilon}\Gamma_j(P_\epsilon)\big|_{\epsilon=0} \\ &= \frac{W_j}{\mathbb{E}_P\left[W_j^2\right]}\left(\frac{I(A=1)}{g(W)\bar{Q}(1,W)}\left(Y-\bar{Q}(A,W)\right)+\bar{Q}(1,W)\right. \\ &\qquad \left. -\frac{I(A=0)}{(1-g(W))\bar{Q}(0,W)}\left(Y-\bar{Q}(A,W)\right)+\bar{Q}(0,W)-\Psi_j(P)W_j\right). \end{aligned}$$

□

Proposition 4.4

Proof. From the definition of D_j given by Equation (4.8), we find that

$$\begin{aligned} \mathbb{E}_{P_0}\left[D_j(O,P)\right] \propto \mathbb{E}_{P_0}\Bigg\{ W_j\Bigg(&\frac{g_0(W)}{g(W)\bar{Q}(A,W)}\left(\bar{Q}_0(1,W)-\bar{Q}(1,W)\right) \\ &+\log\bar{Q}_0(1,W)-\log\bar{Q}(1,W) \\ &-\frac{1-g_0(W)}{(1-g(W))\,\bar{Q}(A,W)}\left(\bar{Q}_0(0,W)-\bar{Q}(0,W)\right) \\ &+\log\bar{Q}_0(0,W)-\log\bar{Q}(0,W)\Bigg)\Bigg\}. \end{aligned}$$

It follows immediately that $\mathbb{E}_{P_0}[D_j(O,P)] = 0$ when $\bar{Q} = \bar{Q}_0$. □

Theorem 4.3

Proof. The proof is analogous to Theorem 4.2. Again, the entropy constraint of A4.7 ensures that the third term of the von Mises expansion converges to zero in probability to 1. The

remainder term in the same von Mises expansion is shown to be $o_P(n^{-1/2})$:

$$
\begin{aligned}
-R(P_0, P_n) = \frac{1}{\mathbb{E}_{P_0}\left[W_j^2\right]} \mathbb{E}_{P_0}\Bigg[W_j \Bigg(& \frac{g_0(W)}{g_n(W)\bar{Q}_n(A,W)} \left(\bar{Q}_0(1,W) - \bar{Q}_n(1,W)\right) \\
& + \log \bar{Q}_0(1,W) - \log \bar{Q}_n(1,W) \\
& - \frac{1-g_0(W)}{(1-g_n(W))\,\bar{Q}_n(A,W)} \left(\bar{Q}_0(0,W) - \bar{Q}_n(0,W)\right) \\
& - \log \bar{Q}_0(0,W) + \log \bar{Q}_n(0,W) \Bigg) \Bigg] \\
\propto \mathbb{E}_{P_0}\left[W_j \frac{g_0(W)}{g_n(W)\bar{Q}_n(A,W)} \left(\bar{Q}_0(1,W) - \bar{Q}_n(1,W)\right) \right] & \\
+ \mathbb{E}_{P_0}\left[W_j \left(\log \bar{Q}_0(1,W) - \log \bar{Q}_n(1,W) \right) \right] & \\
- \mathbb{E}_{P_0}\left[W_j \frac{1-g_0(W)}{(1-g_n(W))\,\bar{Q}_n(A,W)} \left(\bar{Q}_0(0,W) - \bar{Q}_n(0,W)\right) \right] & \\
- \mathbb{E}_{P_0}\left[W_j \left(\log \bar{Q}_n(0,W) + \log \bar{Q}_0(0,W) \right) \right] & \\
= \mathbb{E}_{P_0}\left[W_j \frac{g_0(W)}{g_n(W)} \left(\frac{\bar{Q}_0(1,W)}{\bar{Q}_n(1,W)} - 1 \right) \right] - \mathbb{E}_{P_0}\left[W_j \left(\frac{\bar{Q}_0(1,W)}{\bar{Q}_n(1,W)} - 1 \right) \right] & \\
- \mathbb{E}_{P_0}\left[W_j \frac{1-g_0(W)}{(1-g_n(W))} \left(\frac{\bar{Q}_0(0,W)}{\bar{Q}_n(0,W)} - 1 \right) \right] & \\
+ \mathbb{E}_{P_0}\left[W_j \left(\frac{\bar{Q}_0(0,W)}{\bar{Q}_n(0,W)} - 1 \right) \right] + o_P(n^{-1/2}) & \\
\leq \left| \mathbb{E}_{P_0}\left[W_j \left(\frac{g_0(W)}{g_n(W)} - 1 \right) \left(\frac{\bar{Q}_0(1,W)}{\bar{Q}_n(1,W)} - 1 \right) \right] \right| & \\
+ \left| \mathbb{E}_{P_0}\left[W_j \left(\frac{1-g_0(W)}{1-g_n(W)} - 1 \right) \left(\frac{\bar{Q}_0(0,W)}{\bar{Q}_n(0,W)} - 1 \right) \right] \right| & \\
+ o_P(n^{-1/2}) & \\
\leq \mathbb{E}_{P_0}\left[\frac{W_j^2}{\bar{Q}_n(1,W)^2} \left(\frac{g_0(W)}{g(W)} - 1 \right)^2 \right]^{1/2} \mathbb{E}_{P_0}\left[\left(\bar{Q}_0(1,W) - \bar{Q}_n(1,W)\right)^2 \right]^{1/2} & \\
+ \mathbb{E}_{P_0}\left[\frac{W_j^2}{\bar{Q}_n(0,W)^2} \left(\frac{1-g_0(W)}{1-g_n(W)} - 1 \right)^2 \right]^{1/2} & \\
\mathbb{E}_{P_0}\left[\left(\bar{Q}_0(0,W) - \bar{Q}_n(0,W)\right)^2 \right]^{1/2} + o_P(n^{-1/2}) & \\
\overset{\text{a.s.}}{\leq} M\, \mathbb{E}_{P_0}\left[\left(\frac{g_0(W)}{g_n(W)} - 1 \right)^2 \right]^{1/2} \mathbb{E}_{P_0}\left[\left(\bar{Q}_0(1,W) - \bar{Q}_n(1,W)\right)^2 \right]^{1/2} & \\
+ M\, \mathbb{E}_{P_0}\left[\left(\frac{1-g_0(W)}{1-g_n(W)} - 1 \right)^2 \right]^{1/2} \mathbb{E}_{P_0}\left[\left(\bar{Q}_0(0,W) - \bar{Q}_n(0,W)\right)^2 \right]^{1/2} & \\
+ o_P(n^{-1/2}) &
\end{aligned}
$$

The second equality follows from A4.10 and the Maclaurin series of $\log(x+1)$. The final inequality follows from A4.9 and that Y is a positive random variable such that $W_j^2/\bar{Q}(A,W)^2 \leq M$ almost surely (a.s.). The reported result follows by applying the central limit theorem to the first term of the von Mises expansion.

□

Theorem 4.4

Proof. $S_{P_{X,0}}(t|A,W) \equiv S_0(t|A,W)$ is immediate from A4.4, A4.12 and A4.13. Then $\Psi^F(P_{X,0};t) = \Psi(P_0;t)$. □

Proposition 4.5

Proof. Using previous results from Moore and van der Laan [2011] and the functional delta method, we obtain:

$$\begin{aligned} D_j(O,P;t) &= \frac{d}{d\epsilon}\Psi_j(P_\epsilon;t)\big|_{\epsilon=0} \\ &= \frac{W_j}{\mathbb{E}_P\left[W_j^2\right]}\left(\int_0^t d(O,P;u,1) - d(O,P;u,0)\,du - \Psi_j(P;t)W_j\right) . \end{aligned}$$

□

Proposition 4.6

Proof. From the definition of $D_j(O,P;t)$ in Equation (4.11), we find that

$$\begin{aligned} &\mathbb{E}_{P_0}\left[D_j(O,P;t)\right] \propto \\ &\quad \mathbb{E}_{P_0}\left\{W_j\left(\int_0^t (d(O,P;u,1) - d(O,P_0;u,1)) - (d(O,P;u,0) - d(O,P_0;u,0))\,du\right)\right\} . \end{aligned}$$

□

Conditioning on W, it suffices to show that $\mathbb{E}_{P_0}[d_a(O,P;t) - d_a(O,P_0;t)] = 0$. It follows from previous results of van der Laan and Robins [2003b], Tsiatis [2006] and Cui et al. [2022] that this is achieved when A4.14 or A4.15 are satisfied.

Theorem 4.5

Proof. The proof is analogous Theorem 4.2's. From the results of Moore and van der Laan [2011] and the the functional delta method, the entropy condition of A4.7 implies that the

third term of the von Mises expansion for any given t is $o_P(1)$. Then,

$$\begin{aligned}
-R(P, P_0) &= \frac{1}{\mathbb{E}_{P_0}\left[W_j^2\right]} \mathbb{E}_{P_0}\Bigg\{ W_j \Bigg(\int_0^t (d_1(O, P; u) - d_1(O, P_0; u)) \\
&\qquad - (d_0(O, P; u) - d_0(O, P_0; u))\, du \Bigg) \Bigg\} \\
&\overset{\text{a.s.}}{\leq} \frac{C}{\mathbb{E}_{P_0}\left[W_j^2\right]} \Bigg| \mathbb{E}_{P_0}\Bigg\{ \int_0^t (d_1(O, P; u) - d_1(O, P_0; u)) \\
&\qquad - (d_0(O, P; u) - d_0(O, P_0; u))\, du \Bigg\} \Bigg|
\end{aligned}$$

Similar to the proof of Proposition 4.6, it suffices to show that the integrand is bounded by $o_P(n^{-1/2})$. Indeed, this has previously been established under conditions A4.7, A4.2, A4.9, A4.16 and A4.17. See, for example, van der Laan and Robins [2003b], [Tsiatis, 2006] and Cui et al. [2022]. □

Corollary 4.2

Proof. It follows from the conditions of Theorem 4.4 that $\Gamma^F(P_{X,0})$ is equal to $\Gamma(P_0)$. □

Proposition 4.7

Proof. Again relying on the point mass contamination approach, we obtain:

$$\begin{aligned}
D_j(O, P; t) &= \frac{d}{d\epsilon} \Gamma_j(P_\epsilon; t)\big|_{\epsilon=0} \\
&= \frac{W_j}{\mathbb{E}_P\left[W_j^2\right]} \Bigg(\frac{2A - 1}{Ag(W) + (1 - A)(1 - g(W))} \\
&\qquad \int_0^t \frac{I(\tilde{T} \geq u)}{c(u_-|A, W) S(u|A, W)} \left(I(T = u) - \lambda(u|A, W)\, du\right) \\
&\qquad + \log \frac{S(t|1, W)}{S(t|0, W)} - \Gamma(P; t) W_j \Bigg) .
\end{aligned}$$

□

Proposition 4.8

Proof. By the definition of Equation (4.14), we have that

$$\mathbb{E}_{P_0}\left[D_j(O,P;t)\right] \propto \mathbb{E}_{P_0}\Bigg\{W_j\Bigg(\left(\frac{g_0(W)}{g(W)} - \frac{1-g_0(W)}{1-g(W)}\right) \int_0^t \frac{c_0(u_-|A,W)}{c(u_-|A,W)S(u|A,W)}\left(\lambda_0(u|A,W) - \lambda(u|A,W)\right)\, du + \log\frac{S(t|1,W)}{S(t|0,W)} - \log\frac{S_0(t|1,W)}{S_0(t|0,W)}\Bigg)\Bigg\}$$

Then $\mathbb{E}_{P_0}\left[D_j(O,P;t)\right] = 0$ under A4.14. □

Theorem 4.6

Proof. A4.7 implies that the third term of the von Mises expansion is $o_P(1)$. We then find that

$$
\begin{aligned}
-R(P_0, P_n) = \frac{1}{\mathbb{E}_{P_0}\left[W_j^2\right]}\mathbb{E}_{P_0}\Bigg\{ & W_j\left(\frac{2A-1}{Ag_n(W)+(1-A)(1-g_n(W))}\right. \\
& \int_0^t \frac{I(\tilde{T}\geq u)}{c_n(u_-|A,W)S_n(u|A,W)}\left(I(T=u)-\lambda_n(u|A,W)\right)\,du \\
& \left.+\log\frac{S_n(t|1,W)}{S_n(t|0,W)}-\log\frac{S_0(t|1,W)}{S_0(t|0,W)}\right)\Bigg\} \\
= \frac{1}{\mathbb{E}_{P_0}\left[W_j^2\right]}\mathbb{E}_{P_0}\Bigg\{ & W_j\left(\left(\frac{g_0(W)}{g_n(W)}-\frac{1-g_0(W)}{1-g_n(W)}\right)\right. \\
& \int_0^t \frac{c_0(u_-|A,W)}{c_n(u_-|A,W)S_n(u|A,W)}\left(\lambda_0(u|A,W)-\lambda_n(u|A,W)\right)\,du \\
& \left.+\log\frac{S_n(t|1,W)}{S_n(t|0,W)}-\log\frac{S_0(t|1,W)}{S_0(t|0,W)}\right)\Bigg\} \\
= \frac{1}{\mathbb{E}_{P_0}\left[W_j^2\right]}\mathbb{E}_{P_0}\Bigg\{ & W_j\left(\left(\frac{g_0(W)}{g_n(W)}-\frac{1-g_0(W)}{1-g_n(W)}\right)\right. \\
& \int_0^t \frac{c_0(u_-|A,W)}{c_n(u_-|A,W)S_n(u|A,W)}\left(\lambda_0(u|A,W)-\lambda_n(u|A,W)\right)\,du \\
& \left.\left.+\left(\frac{S_n(t|1,W)}{S_0(t|1,W)}-1\right)-\left(\frac{S_0(t|0,W)}{S_n(t|0,W)}-1\right)\right)\right\}+o_P(n^{-1/2}) \\
= \frac{1}{\mathbb{E}_{P_0}\left[W_j^2\right]}\mathbb{E}_{P_0}\Bigg\{ & W_j\left(\left(\frac{g_0(W)}{g_n(W)}\frac{c_0(t|A,W)}{c_n(t|A,W)}-1\right)\left(\frac{S_0(t|1,W)}{S_n(t|1,W)}-1\right)\right. \\
& \left.\left.+\left(\frac{1-g_0(W)}{1-g_n(W)}\frac{c_0(t|A,W)}{c_n(t|A,W)}-1\right)\left(\frac{S_0(t|A=0,W)}{S_n(t|A=0,W)}-1\right)\right)\right\} \\
& +o_P(n^{-1/2}) \\
\overset{\text{a.s.}}{\leq} \frac{C}{\mathbb{E}_{P_0}\left[W_j^2\right]}\Bigg\{ & \mathbb{E}_{P_0}\left[\left(\frac{g_0(W)c_0(t|A,W)-g_n(W)c_n(t|A,W)}{g_n(W)c_n(t|A,W)}\right)^2\right]^{1/2} \\
& \mathbb{E}_{P_0}\left[\left(S_0(t|1,W)-S_n(t|1,W)\right)^2\right]^{1/2} \\
& +\mathbb{E}_{P_0}\left[\left(\frac{(1-g_0(W))c_0(t|A,W)-(1-g_n(W))c_n(t|A,W)}{(1-g_n(W))c_n(t|A,W)}\right)^2\right]^{1/2} \\
& \mathbb{E}_{P_0}\left[\left(S_0(t|0,W)-S_n(t|0,W)\right)^2\right]^{1/2}\Bigg\}+o_P(n^{-1/2})
\end{aligned}
$$

□

Bibliography

R. A. Amezquita, A. T. L. Lun, E. Becht, V. J. Carey, L. N. Carpp, L. Geistlinger, F. Marini, K. Rue-Albrecht, D. Risso, C. Soneson, L. Waldron, H. Pagès, M. L. Smith, W. Huber, M. Morgan, R. Gottardo, and S. C. Hicks. Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, 17(2):137–145, 2020. doi: 10.1038/s41592-019-0654-x. URL `https://doi.org/10.1038/s41592-019-0654-x`.

T. Anderson. *An introduction to multivariate statistical analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, third edition, 2003.

E. A. Ariazi, G. M. Clark, and J. E. Mertz. Estrogen-related Receptor αand Estrogen-related Receptor γAssociate with Unfavorable and Favorable Biomarkers, Respectively, in Human Breast Cancer1. *Cancer Research*, 62(22):6510–6518, 11 2002. ISSN 0008-5472.

L. Au, E. Hatipoglu, M. Robert de Massy, K. Litchfield, G. Beattie, A. Rowan, D. Schnidrig, R. Thompson, F. Byrne, S. Horswell, N. Fotiadis, S. Hazell, D. Nicol, S. T. Shepherd, A. Fendler, R. Mason, L. Del Rosario, K. Edmonds, K. Lingard, S. Sarker, M. Mangwende, E. Carlyle, J. Attig, K. Joshi, I. Uddin, P. D. Becker, M. W. Sunderland, A. Akarca, I. Puccio, W. W. Yang, T. Lund, K. Dhillon, M. D. Vasquez, E. Ghorani, H. Xu, C. Spencer, J. I. López, A. Green, U. Mahadeva, E. Borg, M. Mitchison, D. A. Moore, I. Proctor, M. Falzon, L. Pickering, A. J. Furness, J. L. Reading, R. Salgado, T. Marafioti, M. Jamal-Hanjani, C. Abbosh, K.-K. Shiu, J. Bridgewater, D. Hochhauser, M. Forster, S.-M. Lee, T. Ahmad, D. Papadatos-Pastos, S. Janes, P. Van Loo, K. Enfield, N. McGranahan, A. Huebner, S. Beck, P. Parker, H. Walczak, T. Enver, R. Hynds, R. Sinclair, C. wah Lok, Z. Rhodes, D. Moore, R. Khiroya, G. Trevisan, P. Ellery, M. Linch, S. Brandner, C. Hiley, S. Veeriah, M. Razaq, H. Shaw, G. Attard, M. A. Akther, C. Naceur-Lombardelli, L. Manzano, M. Al-Bakir, S. Summan, N. Kanu, S. Ward, U. Asghar, E. Lim, F. Gishen, A. Tookman, P. Stone, C. Stirling, N. Hunter, S. Vaughan, M. Mangwende, L. Spain, H. Yan, B. Shum, E. Carlyle, N. Yousaf, S. Popat, O. Curtis, G. Stamp, A. Toncheva, E. Nye, A. Murra, J. Korteweg, D. Josephs, A. Chandra, J. Spicer, R. Stewart, L.-R. Iredale, T. Mackay, B. Deakin, D. Enting, S. Rudman, S. Ghosh, L. Karapagniotou, E. Pintus, A. Tutt, S. Howlett, V. Michalarea, J. Brenton, C. Caldas, R. Fitzgerald, M. Jimenez-Linan, E. Provenzano, A. Cluroe, G. Stewart, C. Watts, R. Gilbertson, U. McDermott, S. Tavare, E. Beddowes, P. Roxburgh, A. Biankin, A. Chalmers, S. Fraser, K. Oien, A. Kidd, K. Blyth, M. Krebs, F. Blackhall, Y. Summers, C. Dive, R. Marais,

F. Gomes, M. Carter, J. Dransfield, J. Le Quesne, D. Fennell, J. Shaw, B. Naidu, S. Baijal, B. Tanchel, G. Langman, A. Robinson, M. Collard, P. Cockcroft, C. Ferris, H. Bancroft, A. Kerr, G. Middleton, J. Webb, S. Kadiri, P. Colloby, B. Olisemeke, R. Wilson, I. Tomlinson, S. Jogai, C. Ottensmeier, D. Harrison, M. Loda, A. Flanagan, M. McKenzie, A. Hackshaw, J. Ledermann, K. Chan, A. Sharp, L. Farrelly, H. Bridger, G. Kassiotis, B. Chain, J. Larkin, C. Swanton, S. A. Quezada, S. Turajlic, B. Challacombe, A. Chandra, S. Chowdhury, W. Drake, A. Fernando, K. Harrison-Phipps, S. Hazell, P. Hill, C. Horsfield, T. O'Brien, J. Olsburgh, A. Polson, S. Rudman, M. Varia, and H. Verma. Determinants of anti-pd-1 response and resistance in clear cell renal cell carcinoma. *Cancer Cell*, 39(11): 1497–1518.e11, 2021. ISSN 1535-6108. doi: https://doi.org/10.1016/j.ccell.2021.10.001. URL `https://www.sciencedirect.com/science/article/pii/S1535610821005432`.

A. Bahamyirou, M. E. Schnitzer, E. H. Kennedy, L. Blais, and Y. Yang. Doubly robust adaptive LASSO for effect modifier discovery. *The International Journal of Biostatistics*, 2022. doi: doi:10.1515/ijb-2020-0073. URL `https://doi.org/10.1515/ijb-2020-0073`.

J. Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003. doi: 10.1111/1468-0262.00392.

D. Bartz. Cross-validation based nonlinear shrinkage, 2016. URL `https://arxiv.org/abs/1611.00798`.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL `http://www.jstor.org/stable/2346101`.

D. Benkeser, P. B. Gilbert, and M. Carone. Estimating and testing vaccine sieve effects using machine learning. *Journal of the American Statistical Association*, 114(527):1038–1049, 2019.

D. Benkeser, I. Díaz, A. Luedtke, J. Segal, D. Scharfstein, and M. Rosenblum. Improving precision and power in randomized trials for covid-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*, 77(4):1467–1481, 2021a.

D. Benkeser, I. Díaz, and J. Ran. Inference for natural mediation effects under case-cohort sampling with applications in identifying covid-19 vaccine correlates of protection, 2021b.

G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

A. F. Bibaut and M. J. van der Laan. Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm, 2019. URL `https://arxiv.org/abs/1907.09244`.

P. J. Bickel. On Adaptive Estimation. *The Annals of Statistics*, 10(3):647 – 671, 1982.

P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume II. Chapman and and Hall, CRC Press, 2015. doi: 10.1201/9781315369266.

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199 – 227, 2008a. doi: 10.1214/009053607000000758. URL `https://doi.org/10.1214/009053607000000758`.

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008b. URL `http://www.jstor.org/stable/25464621`.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008c. URL `http://www.jstor.org/stable/25464728`.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993a.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993b.

P. Boileau, N. S. Hejazi, B. Collica, M. J. van der Laan, and S. Dudoit. cvCovEst: Cross-validated covariance matrix estimator selection and evaluation in R. *Journal of Open Source Software*, 6(63):3273, 2021a. doi: 10.21105/joss.03273. URL `https://doi.org/10.21105/joss.03273`.

P. Boileau, N. S. Hejazi, M. J. van der Laan, and S. Dudoit. Cross-validated loss-based covariance matrix estimator selection in high dimensions, 2021b. URL `https://arxiv.org/abs/2102.09715`.

P. Boileau, N. T. Qi, M. J. van der Laan, S. Dudoit, and N. Leng. A flexible approach for predictive biomarker discovery. *Biostatistics*, 07 2022. ISSN 1465-4644. doi: 10.1093/biostatistics/kxac029. URL `https://doi.org/10.1093/biostatistics/kxac029`. kxac029.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman and P. Spector. Submodel selection and evaluation in regression. The X-random case. *International Statistical Review*, pages 291–319, 1992.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011. doi: 10.1198/jasa.2011.tm10560. URL `https://doi.org/10.1198/jasa.2011.tm10560`.

T. T. Cai, C.-H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118 – 2144, 2010a. doi: 10.1214/09-AOS752. URL `https://doi.org/10.1214/09-AOS752`.

T. T. Cai, C.-H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, 38(4):2118–2144, 08 2010b. doi: 10.1214/09-AOS752.

A. Chambaz, P. Neuvial, and M. J. van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6(none): 1059 – 1099, 2012.

G. Chamie, T. D. Clark, J. Kabami, K. Kadede, E. Ssemmondo, R. Steinfeld, G. Lavoy, D. Kwarisiima, N. Sang, V. Jain, H. Thirumurthy, T. Liegler, L. B. Balzer, M. L. Petersen, C. R. Cohen, E. A. Bukusi, M. R. Kamya, D. V. Havlir, and E. D. Charlebois. A hybrid mobile approach for population-wide hiv testing in rural east africa: an observational study. *The Lancet HIV*, 3(3):e111–e119, 2016.

P.-Y. Chen and A. A. Tsiatis. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038, 2001.

S. Chen, L. Tian, T. Cai, and M. Yu. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4):1199–1209, 2017. doi: https://doi.org/10.1111/biom.12676. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12676`.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, May 2017. doi: 10.1257/aer.p20171038. URL `https://www.aeaweb.org/articles?id=10.1257/aer.p20171038`.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL `https://doi.org/10.1111/ectj.12097`.

F. J. Couch, K. B. Kuchenbaecker, K. Michailidou, G. A. Mendoza-Fandino, S. Nord, J. Lilyquist, C. Olswold, E. Hallberg, S. Agata, H. Ahsan, K. Aittomäki, C. Ambrosone, I. L. Andrulis, H. Anton-Culver, V. Arndt, B. K. Arun, B. Arver, M. Barile, R. B. Barkardottir, D. Barrowdale, L. Beckmann, M. W. Beckmann, J. Benitez, S. V. Blank, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, M. K. Bolla, B. Bonanni, H. Brauch, H. Brenner, B. Burwinkel, S. S. Buys, T. Caldes, M. A. Caligo, F. Canzian, J. Carpenter, J. Chang-Claude, S. J. Chanock, W. K. Chung, K. B. M. Claes, A. Cox, S. S. Cross, J. M. Cunningham, K. Czene, M. B. Daly, F. Damiola, H. Darabi, M. de la Hoya, P. Devilee,

O. Diez, Y. C. Ding, R. Dolcetti, S. M. Domchek, C. M. Dorfling, I. dos Santos-Silva, M. Dumont, A. M. Dunning, D. M. Eccles, H. Ehrencrona, A. B. Ekici, H. Eliassen, S. Ellis, P. A. Fasching, J. Figueroa, D. Flesch-Janys, A. Försti, F. Fostira, W. D. Foulkes, T. Friebel, E. Friedman, D. Frost, M. Gabrielson, M. D. Gammon, P. A. Ganz, S. M. Gapstur, J. Garber, M. M. Gaudet, S. A. Gayther, A.-M. Gerdes, M. Ghoussaini, G. G. Giles, G. Glendon, A. K. Godwin, M. S. Goldberg, D. E. Goldgar, A. González-Neira, M. H. Greene, J. Gronwald, P. Guénel, M. Gunter, L. Haeberle, C. A. Haiman, U. Hamann, T. V. O. Hansen, S. Hart, S. Healey, T. Heikkinen, B. E. Henderson, J. Herzog, F. B. L. Hogervorst, A. Hollestelle, M. J. Hooning, R. N. Hoover, J. L. Hopper, K. Humphreys, D. J. Hunter, T. Huzarski, E. N. Imyanitov, C. Isaacs, A. Jakubowska, P. James, R. Janavicius, U. B. Jensen, E. M. John, M. Jones, M. Kabisch, S. Kar, B. Y. Karlan, S. Khan, K.-T. Khaw, M. G. Kibriya, J. A. Knight, Y.-D. Ko, I. Konstantopoulou, V.-M. Kosma, V. Kristensen, A. Kwong, Y. Laitman, D. Lambrechts, C. Lazaro, E. Lee, L. Le Marchand, J. Lester, A. Lindblom, N. Lindor, S. Lindstrom, J. Liu, J. Long, J. Lubinski, P. L. Mai, E. Makalic, K. E. Malone, A. Mannermaa, S. Manoukian, S. Margolin, F. Marme, J. W. M. Martens, L. McGuffog, A. Meindl, A. Miller, R. L. Milne, P. Miron, M. Montagna, S. Mazoyer, A. M. Mulligan, T. A. Muranen, K. L. Nathanson, S. L. Neuhausen, H. Nevanlinna, B. G. Nordestgaard, R. L. Nussbaum, K. Offit, E. Olah, O. I. Olopade, J. E. Olson, A. Osorio, S. K. Park, P. H. Peeters, B. Peissel, P. Peterlongo, J. Peto, C. M. Phelan, R. Pilarski, B. Poppe, K. Pylkäs, P. Radice, N. Rahman, J. Rantala, C. Rappaport, G. Rennert, A. Richardson, M. Robson, I. Romieu, A. Rudolph, E. J. Rutgers, M.-J. Sanchez, R. M. Santella, E. J. Sawyer, D. F. Schmidt, M. K. Schmidt, R. K. Schmutzler, F. Schumacher, R. Scott, L. Senter, P. Sharma, J. Simard, C. F. Singer, O. M. Sinilnikova, P. Soucy, M. Southey, D. Steinemann, M. Stenmark-Askmalm, D. Stoppa-Lyonnet, A. Swerdlow, C. I. Szabo, R. Tamimi, W. Tapper, M. R. Teixeira, S.-H. Teo, M. B. Terry, M. Thomassen, D. Thompson, L. Tihomirova, A. E. Toland, R. A. E. M. Tollenaar, I. Tomlinson, T. Truong, H. Tsimiklis, A. Teulé, R. Tumino, N. Tung, C. Turnbull, G. Ursin, C. H. M. van Deurzen, E. J. van Rensburg, R. Varon-Mateeva, Z. Wang, S. Wang-Gohrke, E. Weiderpass, J. N. Weitzel, A. Whittemore, H. Wildiers, R. Winqvist, X. R. Yang, D. Yannoukakos, S. Yao, M. P. Zamora, W. Zheng, P. Hall, P. Kraft, C. Vachon, S. Slager, G. Chenevix-Trench, P. D. P. Pharoah, A. A. N. Monteiro, M. García-Closas, D. F. Easton, and A. C. Antoniou. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. *Nature Communications*, 7(1):11375, 2016.

J. R. Coyle, N. S. Hejazi, I. Malenica, R. V. Phillips, and O. Sofrygin. *sl3: Modern Pipelines for Machine Learning and Super Learning*, 2021. URL `https://doi.org/10.5281/zenodo.1342293`. R package version 1.4.2.

Y. Cui, M. R. Kosorok, E. Sverdrup, S. Wager, and R. Zhu. Estimating heterogeneous treatment effects with right-censored data via causal Survival Forests, 2022. URL `https://arxiv.org/abs/2001.09887`.

I. Díaz, E. Colantuoni, and M. Rosenblum. Enhanced precision in the analysis of randomized trials with ordinal outcomes. *Biometrics*, 72(2):422–431, 2016.

I. Díaz, E. Colantuoni, D. Hanley, and M. Rosenblum. Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime Data Analysis*, 25(3):439–468, July 2019. ISSN 1380-7870. doi: 10.1007/s10985-018-9428-5.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

J. Duncan and T. Tang. *simChef: Intensive Computational Experiments Made Easy*, 2022. URL `https://yu-group.github.io/simChef`. R package version 0.0.2.

B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2012.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi: 10.1198/016214501753382273.

J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186 – 197, 2008. doi: https://doi.org/10.1016/j.jeconom.2008.09.017. URL `http://www.sciencedirect.com/science/article/pii/S0304407608001346`. Econometric modelling in finance and risk management: An overview.

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 8 2013. doi: 10.1111/rssb.12016.

J. Fan, Y. Liao, and H. Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016a. doi: 10.1111/ectj.12061. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12061`.

J. Fan, Y. Liao, and W. Wang. Projected principal component analysis in factor models. *Annals of Statistics*, 44(1):219–254, 02 2016b. doi: 10.1214/15-AOS1364.

J. Fan, W. Wang, and Y. Zhong. Robust covariance estimation for approximate factor models. *Journal of Econometrics*, 208(1):5 – 22, 2019. doi: https://doi.org/10.1016/j.jeconom.2018.09.003. Special Issue on Financial Engineering and Risk Management.

Q. Fan, Y.-C. Hsu, R. P. Lieli, and Y. Zhang. Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 0(0):1–15, 2020. doi: 10.1080/07350015.2020.1811102. URL `https://doi.org/10.1080/07350015.2020.1811102`.

Y. Fang, B. Wang, and Y. Feng. Tuning-parameter selection in regularized estimations of large covariance matrices. *Journal of Statistical Computation and Simulation*, 86(3): 494–509, 2016. doi: 10.1080/00949655.2015.1017823. URL `https://doi.org/10.1080/00949655.2015.1017823`.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021. doi: https://doi.org/10.3982/ECTA16901. URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16901`.

A. Fisher and E. H. Kennedy. Visually communicating and teaching intuition for influence functions. *The American Statistician*, 75(2):162–172, 2021.

W. Fithian and L. Lei. Conditional calibration for false discovery rate control under dependence, 2020. URL `https://arxiv.org/abs/2007.10438`.

Y. Fong, A. B. McDermott, D. Benkeser, S. Roels, D. J. Stieh, A. Vandebosch, M. Le Gars, G. A. Van Roey, C. R. Houchens, K. Martins, L. Jayashankar, F. Castellino, O. Amoa-Awua, M. Basappa, B. Flach, B. C. Lin, C. Moore, M. Naisan, M. Naqvi, S. Narpala, S. O'Connell, A. Mueller, L. Serebryannyy, M. Castro, J. Wang, C. J. Petropoulos, A. Luedtke, O. Hyrien, Y. Lu, C. Yu, B. Borate, L. W. P. van der Laan, N. S. Hejazi, A. Kenny, M. Carone, D. N. Wolfe, J. Sadoff, G. E. Gray, B. Grinsztejn, P. A. Goepfert, S. J. Little, L. Paiva de Sousa, R. Maboa, A. K. Randhawa, M. P. Andrasik, J. Hendriks, C. Truyers, F. Struyf, H. Schuitemaker, M. Douoguih, J. G. Kublin, L. Corey, K. M. Neuzil, L. N. Carpp, D. Follmann, P. B. Gilbert, R. A. Koup, R. O. Donis, on behalf of the Immune Assays Team, the Coronavirus Vaccine Prevention Network (CoVPN)/ENSEMBLE Team, , and the United States Government (USG)/CoVPN Biostatistics Team. Immune correlates analysis of the ensemble single ad26.cov2.s dose vaccine efficacy clinical trial. *Nature Microbiology*, 7(12):1996–2010, 2022.

Y. Fong, Y. Huang, D. Benkeser, L. N. Carpp, G. Áñez, W. Woo, A. McGarry, L. M. Dunkle, I. Cho, C. R. Houchens, K. Martins, L. Jayashankar, F. Castellino, C. J. Petropoulos, A. Leith, D. Haugaard, B. Webb, Y. Lu, C. Yu, B. Borate, L. W. P. van der Laan, N. S. Hejazi, A. K. Randhawa, M. P. Andrasik, J. G. Kublin, J. Hutter, M. Keshtkar-Jahromi, T. H. Beresnev, L. Corey, K. M. Neuzil, D. Follmann, J. A. Ake, C. L. Gay, K. L. Kotloff, R. A. Koup, R. O. Donis, P. B. Gilbert, I. A. Team, C. V. P. N. C.-. P. Investigators, S. Team, and U. S. G. U. B. Team. Immune correlates analysis of the prevent-19 covid-19 vaccine efficacy clinical trial. *Nature Communications*, 14(1):331, 2023.

C. for Drug Evaluation and Research. Adjusting for covariates in randomized clinical trials for drugs and b, 2023. URL `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products`.

C. for Medicinal Products for Human Use. Guideline on adjustment for baseline covariates in clinical trials, 2023. URL

https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York, 2001.

J. H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19:1 – 67, 1991.

E. H. Geng, D. Nash, A. Kambugu, Y. Zhang, P. Braitstein, K. A. Christopoulos, W. Muyindike, M. B. Bwana, C. T. Yiannoutsos, M. L. Petersen, and J. N. Martin. Retention in care among hiv-infected patients in resource-limited settings: Emerging insights and new directions. *Current HIV/AIDS Reports*, 7(4):234–244, 2010. doi: 10.1007/s11904-010-0061-5. URL https://doi.org/10.1007/s11904-010-0061-5.

P. B. Gilbert, D. C. Montefiori, A. B. McDermott, Y. Fong, D. Benkeser, W. Deng, H. Zhou, C. R. Houchens, K. Martins, L. Jayashankar, F. Castellino, B. Flach, B. C. Lin, S. O'Connell, C. McDanal, A. Eaton, M. Sarzotti-Kelsoe, Y. Lu, C. Yu, B. Borate, L. W. P. van der Laan, N. S. Hejazi, C. Huynh, J. Miller, H. M. E. Sahly, L. R. Baden, M. Baron, L. D. L. Cruz, C. Gay, S. Kalams, C. F. Kelley, M. P. Andrasik, J. G. Kublin, L. Corey, K. M. Neuzil, L. N. Carpp, R. Pajon, D. Follmann, R. O. Donis, R. A. Koup, I. A. Team§, I. T. Moderna, C. V. P. N. C. E. C. Team§, and U. S. G. U. B. Team§. Immune correlates analysis of the mrna-1273 covid-19 vaccine efficacy clinical trial. *Science*, 375(6576):43–50, 2022. doi: 10.1126/science.abm3425. URL https://www.science.org/doi/abs/10.1126/science.abm3425.

R. D. Gill, M. J. Laan, and J. A. Wellner. Inefficient estimators of the bivariate survival function for three models. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 545–597, 1995.

G. S. Ginsburg and K. A. Phillips. Precision medicine: From science to value. *Health Affairs*, 37(5):694–701, 2018. doi: 10.1377/hlthaff.2017.1624. URL https://doi.org/10.1377/hlthaff.2017.1624. PMID: 29733705.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Universtiy Press, 3rd edition, 1996.

S. Greenland, J. Pearl, and J. M. Robins. Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29 – 46, 1999.

S. Gruber and M. J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1), 2010.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer series in statistics. Springer, 2002. ISBN 978-0-387-95441-7.

N. Hao and H. H. Zhang. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301, 2014. doi: 10.1080/01621459.2014.881741. URL https://doi.org/10.1080/01621459.2014.881741. PMID: 25386043.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference and prediction*. Springer, 2 edition, 2009. URL http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

D. V. Havlir, L. B. Balzer, E. D. Charlebois, T. D. Clark, D. Kwarisiima, J. Ayieko, J. Kabami, N. Sang, T. Liegler, G. Chamie, C. S. Camlin, V. Jain, K. Kadede, M. Atukunda, T. Ruel, S. B. Shade, E. Ssemmondo, D. M. Byonanebye, F. Mwangwa, A. Owaraganise, W. Olilo, D. Black, K. Snyman, R. Burger, M. Getahun, J. Achando, B. Awuonda, H. Nakato, J. Kironde, S. Okiror, H. Thirumurthy, C. Koss, L. Brown, C. Marquez, J. Schwab, G. Lavoy, A. Plenty, E. Mugoma Wafula, P. Omanya, Y.-H. Chen, J. F. Rooney, M. Bacon, M. van der Laan, C. R. Cohen, E. Bukusi, M. R. Kamya, and M. Petersen. Hiv testing and treatment with the use of a community health approach in rural africa. *New England Journal of Medicine*, 381(3):219–229, 2019.

N. S. Hejazi, M. J. van der Laan, H. E. Janes, P. B. Gilbert, and D. C. Benkeser. Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. *Biometrics*, 77(4):1241–1253, 2021.

N. S. Hejazi, P. Boileau, M. J. van der Laan, and A. E. Hubbard. A generalization of moderated statistics to data adaptive semiparametric estimation in high-dimensional biology, 2023.

M. A. Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1), 2010.

O. Hines, K. Diaz-Ordaz, and S. Vansteelandt. Variable importance measures for heterogeneous causal effects, 2022a. URL https://arxiv.org/abs/2204.06030.

O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 0(ja):1–48, 2022b. doi: 10.1080/00031305.2021.2021984. URL https://doi.org/10.1080/00031305.2021.2021984.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 1970. ISSN 00401706. URL http://www.jstor.org/stable/1271436.

W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, 12(2):115–121,

Feb 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3252. URL https://doi.org/10.1038/nmeth.3252.

J. D. Huling and M. Yu. Subgroup identification using the personalized package. *Journal of Statistical Software*, 98(5):1–60, 2021. doi: 10.18637/jss.v098.i05. URL https://www.jstatsoft.org/index.php/jss/article/view/v098i05.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R1). https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf, 2019. Accessed: 2023-01-24.

B. Jiang and J. S. Liu. Variable selection for general index models via sliced inverse regression. *The Annals of Statistics*, 42(5), oct 2014. doi: 10.1214/14-aos1233. URL https://doi.org/10.1214%2F14-aos1233.

H. Joensuu, P.-L. Kellokumpu-Lehtinen, P. Bono, T. Alanko, V. Kataja, R. Asola, T. Utriainen, R. Kokko, A. Hemminki, M. Tarkkanen, T. Turpeenniemi-Hujanen, S. Jyrkkiö, M. Flander, L. Helle, S. Ingalsuo, K. Johansson, A.-S. Jääskeläinen, M. Pajunen, M. Rauhala, J. Kaleva-Kerola, T. Salminen, M. Leinonen, I. Elomaa, and J. Isola. Adjuvant docetaxel or vinorelbine with or without trastuzumab for breast cancer. *New England Journal of Medicine*, 354(8):809–820, 2006.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, pages 295–327, 2001.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009. doi: 10.1198/jasa.2009.0121. PMID: 20617121.

C. A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, 15(4):1548–1562, 1987. ISSN 00905364. URL http://www.jstor.org/stable/2241690.

V. B. Kraus. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nature Reviews Rheumatology*, 14(6):354–362, Jun 2018. ISSN 1759-4804. doi: 10.1038/s41584-018-0005-9. URL https://doi.org/10.1038/s41584-018-0005-9.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254–4278, 12 2009. doi: 10.1214/09-AOS720.

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603 – 621, 2003. ISSN 0927-5398. doi: https://doi.org/10.1016/S0927-5398(03)00007-0.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365 – 411, 2004. doi: https://doi.org/10.1016/S0047-259X(03)00096-4.

O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40(2):1024–1060, 04 2012. doi: 10.1214/12-AOS989.

O. Ledoit and M. Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139:360 – 384, 2015. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2015.04.006.

O. Ledoit and M. Wolf. Analytical nonlinear shrinkage of large-dimensional covariance matrices. ECON - Working Papers 264, Department of Economics - University of Zurich, 2018. URL `https://EconPapers.repec.org/RePEc:zur:econwp:264`.

O. Ledoit and M. Wolf. The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation. *Journal of Financial Econometrics*, 06 2020. ISSN 1479-8409. doi: 10.1093/jjfinec/nbaa007.

J. Levy, M. van der Laan, A. Hubbard, and R. Pirracchio. A fundamental measure of treatment effect heterogeneity. *Journal of Causal Inference*, 9(1):83–108, 2021.

X. Li, S. Li, and A. Luedtke. Estimating the efficiency gain of covariate-adjusted analyses in future clinical trials using external data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):356–377, 03 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad007. URL `https://doi.org/10.1093/jrsssb/qkad007`.

A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 05 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr260. URL `https://doi.org/10.1093/bioinformatics/btr260`.

S. Loi, S. Michiels, R. Salgado, N. Sirtaine, V. Jose, D. Fumagalli, P.-L. Kellokumpu-Lehtinen, P. Bono, V. Kataja, C. Desmedt, M. Piccart, S. Loibl, C. Denkert, M. Smyth, H. Joensuu, and C. Sotiriou. Tumor infiltrating lymphocytes are prognostic in triple negative breast cancer and predictive for trastuzumab benefit in early breast cancer: Results from the FinHER trial. *Annals of Oncology*, 25(8):1544–1550, 2014.

A. R. Luedtke and M. J. van der Laan. Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics*, 12(1):305–332, 2016. doi: doi:10.1515/ijb-2015-0052. URL `https://doi.org/10.1515/ijb-2015-0052`.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, apr 1967.

D. F. McDermott, M. A. Huseni, M. B. Atkins, R. J. Motzer, B. I. Rini, B. Escudier, L. Fong, R. W. Joseph, S. K. Pal, J. A. Reeves, M. Sznol, J. Hainsworth, W. K. Rathmell, W. M. Stadler, T. Hutson, M. E. Gore, A. Ravaud, S. Bracarda, C. Suárez, R. Danielli, V. Gruenwald, T. K. Choueiri, D. Nickles, S. Jhunjhunwala, E. Piault-Louis, A. Thobhani, J. Qiu, D. S. Chen, P. S. Hegde, C. Schiff, G. D. Fine, and T. Powles. Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nature Medicine*, 24(6):749–757, 2018. doi: 10.1038/s41591-018-0053-3. URL `https://doi.org/10.1038/s41591-018-0053-3`.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 1802.03426, 2 2018. URL `http://arxiv.org/abs/1802.03426`.

O. Menyhart, J. Budczies, G. Munkácsy, F. J. Esteva, A. Szabó, T. Puig Miquel, and B. Győrffy. DUSP4 is associated with increased resistance against anti-HER2 therapy in breast cancer. *Oncotarget*, 8(44):77207–77218, 2017. ISSN 1949-2553. doi: https://doi.org/10.18632/oncotarget.20430. URL `https://www.oncotarget.com/article/20430/`.

M. F. Mercogliano, M. De Martino, L. Venturutti, M. A. Rivas, C. J. Proietti, G. Inurrigarro, I. Frahm, D. H. Allemand, E. G. Deza, S. Ares, F. G. Gercovich, P. Guzmán, J. C. Roa, P. V. Elizalde, and R. Schillaci. TNFα-Induced Mucin 4 Expression Elicits Trastuzumab Resistance in HER2-Positive Breast Cancer. *Clinical Cancer Research*, 23(3):636–648, 01 2017. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-16-0970. URL `https://doi.org/10.1158/1078-0432.CCR-16-0970`.

K. Moore and M. van der Laan. *RCTs with Time-to-Event Outcomes*, pages 259–269. Springer New York, NY, 06 2011. ISBN 978-1-4419-9781-4. doi: 10.1007/978-1-4419-9782-1_17.

R. J. Motzer, B. Escudier, D. F. McDermott, S. George, H. J. Hammers, S. Srinivas, S. S. Tykodi, J. A. Sosman, G. Procopio, E. R. Plimack, D. Castellano, T. K. Choueiri, H. Gurney, F. Donskov, P. Bono, J. Wagstaff, T. C. Gauler, T. Ueda, Y. Tomita, F. A. Schutz, C. Kollmannsberger, J. Larkin, A. Ravaud, J. S. Simon, L.-A. Xu, I. M. Waxman, and P. Sharma. Nivolumab versus everolimus in advanced renal-cell carcinoma. *New England Journal of Medicine*, 373(19):1803–1813, 2015a. doi: 10.1056/NEJMoa1510665. URL `https://doi.org/10.1056/NEJMoa1510665`. PMID: 26406148.

R. J. Motzer, B. I. Rini, D. F. McDermott, B. G. Redman, T. M. Kuzel, M. R. Harrison, U. N. Vaishampayan, H. A. Drabkin, S. George, T. F. Logan, K. A. Margolin, E. R. Plimack, A. M. Lambert, I. M. Waxman, and H. J. Hammers. Nivolumab for metastatic renal cell carcinoma: Results of a randomized phase ii trial. *Journal of Clinical Oncology*, 33(13):1430–1437, 2015b. doi: 10.1200/JCO.2014.59.0703. URL `https://doi.org/10.1200/JCO.2014.59.0703`. PMID: 25452452.

E. Moussay, T. Kaoma, J. Baginska, A. Muller, K. V. Moer, N. Nicot, P. V. Nazarov, L. Vallar, S. Chouaib, G. Berchem, and B. Janji. The acquisition of resistance to TNFαin breast cancer cells is associated with constitutive activation of autophagy as revealed by a transcriptome analysis using a custom microarray. *Autophagy*, 7(7):760–770, 2011.

G. Neuhaus. On Weak Convergence of Stochastic Processes with Multidimensional Time Parameter. *The Annals of Mathematical Statistics*, 42(4):1285 – 1295, 1971.

Y. Ning, P. Sida, and K. Imai. Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554, 06 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa020. URL `https://doi.org/10.1093/biomet/asaa020`.

A. Onatski. Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244 – 258, 2012. doi: https://doi.org/10.1016/j.jeconom.2012.01.034.

P. Parsana, M. Riester, and L. Waldron. *curatedCRCData: Clinically Annotated Data for the Colorectal Cancer Transcriptome*, 2021. This is a manually curated data collection for gene expression meta-analysis of patients with colorectal cancer. This resource provides uniformly prepared microarray data with curated and documented clinical metadata. It allows users to efficiently identify studies and patient subgroups of interest for analysis and to perform meta-analysis immediately without the challenges posed by harmonizing heterogeneous microarray technologies, study designs, expression data processing methods and clinical data formats.

M. Petersen, L. Balzer, D. Kwarsiima, N. Sang, G. Chamie, J. Ayieko, J. Kabami, A. Owaraganise, T. Liegler, F. Mwangwa, K. Kadede, V. Jain, A. Plenty, L. Brown, G. Lavoy, J. Schwab, D. Black, M. van der Laan, E. A. Bukusi, C. R. Cohen, T. D. Clark, E. Charlebois, M. Kamya, and D. Havlir. Association of Implementation of a Universal Testing and Treatment Intervention With HIV Diagnosis, Receipt of Antiretroviral Therapy, and Viral Suppression in East Africa. *JAMA*, 317(21):2196–2206, 06 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.5705. URL `https://doi.org/10.1001/jama.2017.5705`.

M. L. Petersen, Y. Wang, M. J. van der Laan, D. Guzman, E. Riley, and D. R. Bangsberg. Pillbox Organizers are Associated with Improved Adherence to HIV Antiretroviral Therapy and Viral Suppression: a Marginal Structural Model Analysis. *Clinical Infectious Diseases*, 45(7):908–915, 10 2007. ISSN 1058-4838. doi: 10.1086/521250. URL `https://doi.org/10.1086/521250`.

M. L. Petersen, M. J. van der Laan, S. Napravnik, J. J. Eron, R. D. Moore, and S. G. Deeks. Long-term consequences of the delay between virologic failure of highly active antiretroviral therapy and regimen modification. *AIDS*, 22(16), 2008.

J. Pfanzagl and W. Wefelmeyer. Contributions to a general asymptotic statistical theory. *Statistics & Risk Modeling*, 3(3-4):379–388, 1985.

H. Poincaré. *Calcul des probabilités*, volume 1. Gauthier-Villars, 1912.

M. Pourahmadi. *High-dimensional covariance estimation*. Wiley series in probability and statistics. Wiley, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

B. I. Rini and M. B. Atkins. Resistance to targeted therapy in renal-cell carcinoma. *The Lancet Oncology*, 10(10):992–1000, 2009. ISSN 1470-2045. doi: https://doi.org/10.1016/S1470-2045(09)70240-2. URL `https://www.sciencedirect.com/science/article/pii/S1470204509702402`.

B. I. Rini, T. Powles, M. B. Atkins, B. Escudier, D. F. McDermott, C. Suarez, S. Bracarda, W. M. Stadler, F. Donskov, J. L. Lee, R. Hawkins, A. Ravaud, B. Alekseev, M. Staehler, M. Uemura, U. De Giorgi, B. Mellado, C. Porta, B. Melichar, H. Gurney, J. Bedke, T. K. Choueiri, F. Parnis, T. Khaznadar, A. Thobhani, S. Li, E. Piault-Louis, G. Frantz, M. Huseni, C. Schiff, M. C. Green, and R. J. Motzer. Atezolizumab plus bevacizumab versus sunitinib in patients with previously untreated metastatic renal cell carcinoma (immotion151): a multicentre, open-label, phase 3, randomised controlled trial. *The Lancet*, 393(10189):2404–2415, 2019. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(19)30723-8. URL `https://www.sciencedirect.com/science/article/pii/S0140673619307238`.

D. Risso and M. Cole. *scRNAseq: Collection of Public Single-Cell RNA-Seq Datasets*, 2020. R package version 2.3.17.

H. Robbins. The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics*, 35(1):1–20, 1964.

J. Robins, L. Orellana, and A. Rotnitzky. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27(23):4678–4721, 2008. doi: https://doi.org/10.1002/sim.3301. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3301`.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

M. Rosenblum and M. J. van der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics*, 6(1), 2010.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009. doi: 10.1198/jasa.2009.0101. URL `https://doi.org/10.1198/jasa.2009.0101`.

P. Royston and M. K. B. Parmar. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, 30(19):2409–2421, 2011. doi: https://doi.org/10.1002/sim.4274. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4274`.

P. Royston and W. Sauerbrei. Interactions between treatment and continuous covariates: A step toward individualizing therapy. *Journal of Clinical Oncology*, 26(9):1397–1399, 2008. doi: 10.1200/JCO.2007.14.8981. URL `https://doi.org/10.1200/JCO.2007.14.8981`. PMID: 18349388.

D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

A. Schick. On Asymptotically Efficient Estimation in Semiparametric Models. *The Annals of Statistics*, 14(3):1139 – 1151, 1986.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. doi: https://doi.org/10.2202/1544-6115.1175.

K. Sechidis, K. Papangelou, P. D. Metcalfe, D. Svensson, J. Weatherall, and G. Brown. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics*, 34(19):3365–3376, 05 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty357. URL `https://doi.org/10.1093/bioinformatics/bty357`.

V. Semenova and V. Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 08 2020. ISSN 1368-4221. doi: 10.1093/ectj/utaa027. URL `https://doi.org/10.1093/ectj/utaa027`.

S. Smith. Covariance, subspace, and intrinsic cramer-rao bounds. *IEEE Transactions on Signal Processing*, 53(5):1610–1630, 2005. doi: 10.1109/TSP.2005.845428.

C. Stein. Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, 34(1):1373–1403, 1986. doi: 10.1007/BF01085007. URL `https://doi.org/10.1007/BF01085007`.

J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002. URL http://www.jstor.org/stable/3085839.

C. J. Stone, M. Hansen, C. Kooperberg, and Y. K. Truong. Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist*, 25:1371–1470, 1997.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506580102. URL https://www.pnas.org/content/102/43/15545.

C. Y. Tang, E. X. Fang, and Y. Dong. High-dimensional interactions detection with sparse principal hessian matrix. *J. Mach. Learn. Res.*, 21:19–1, 2020.

B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, D. Bertagnolli, J. Goldy, N. Shapovalova, S. Parry, C. Lee, K. Smith, A. Bernard, L. Madisen, S. M. Sunkin, M. Hawrylycz, C. Koch, and H. Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335–346, 2016. doi: 10.1038/nn.4216. URL https://doi.org/10.1038/nn.4216.

E. J. Tchetgen Tchetgen, J. M. Robins, and A. Rotnitzky. On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97(1):171–180, 12 2009. ISSN 0006-3444. doi: 10.1093/biomet/asp062. URL https://doi.org/10.1093/biomet/asp062.

L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014. doi: 10.1080/01621459.2014.951443. URL https://doi.org/10.1080/01621459.2014.951443. PMID: 25729117.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

C. Tiraby, B. C. Hazen, M. L. Gantner, and A. Kralli. Estrogen-Related Receptor γPromotes Mesenchymal-to-Epithelial Transition and Suppresses Breast Tumor Growth. *Cancer Research*, 71(7):2518–2528, 03 2011. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-10-1315. URL https://doi.org/10.1158/0008-5472.CAN-10-1315.

A. C. Tsai, S. D. Weiser, M. L. Petersen, K. Ragland, M. B. Kushel, and D. R. Bangsberg. A Marginal Structural Model to Estimate the Causal Effect of Antidepressant Medication Treatment on Viral Suppression Among Homeless and Marginally Housed

Persons With HIV. *Archives of General Psychiatry*, 67(12):1282–1290, 12 2010. ISSN 0003-990X. doi: 10.1001/archgenpsychiatry.2010.160. URL https://doi.org/10.1001/archgenpsychiatry.2010.160.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, New York, NY, 1st ed. 2006. edition, 2006.

C. Tuglus, K. E. Porter, and M. J. van der Laan. Targeted maximum likelihood estimation of conditional relative risk in a semi-parametric regression model. Working Paper 283, University of California, Berkeley, Berkeley, 2011. URL https://biostats.bepress.com/ucbbiostat/paper283/.

A. W. v. d. Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

M. van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The International Journal of Biostatistics*, 13(2):20150097, 2017. doi: doi:10.1515/ijb-2015-0097. URL https://doi.org/10.1515/ijb-2015-0097.

M. van der Laan and S. Gruber. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *The International Journal of Biostatistics*, 12(1):351–378, 2016.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Working Paper 130, University of California, Berkeley, Berkeley, 2003a. URL https://biostats.bepress.com/ucbbiostat/paper130/.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Working Paper 130, University of California, Berkeley, 2003b. URL https://biostats.bepress.com/ucbbiostat/paper130/.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media, 2003a.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media, 2003b.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011a.

M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media, 2011b.

M. J. van der Laan and S. Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Science & Business Media, 2018a.

M. J. van der Laan and S. Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Data*. Springer International Publishing, 2018b.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006a.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2(1), 2006b.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007a. doi: doi:10.2202/1544-6115.1309. URL `https://doi.org/10.2202/1544-6115.1309`.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007b. doi: doi:10.2202/1544-6115.1309. URL `https://doi.org/10.2202/1544-6115.1309`.

A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics and Decisions*, 24:351–371, 2006.

K. Vermeulen, O. Thas, and S. Vansteelandt. Increasing the power of the mann-whitney test in randomized experiments through flexible covariate adjustment. *Statistics in Medicine*, 34(6):1012–1030, 2015.

R. von Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL `https://doi.org/10.1080/01621459.2017.1319839`.

J. J. Wallin, J. C. Bendell, R. Funke, M. Sznol, K. Korski, S. Jones, G. Hernandez, J. Mier, X. He, F. S. Hodi, M. Denker, V. Leveque, M. Cañamero, G. Babitski, H. Koeppen, J. Ziai, N. Sharma, F. Gaire, D. S. Chen, D. Waterkamp, P. S. Hegde, and D. F. McDermott. Atezolizumab in combination with bevacizumab enhances antigen-specific t-cell migration in metastatic renal cell carcinoma. *Nature Communications*, 7(1):12624, 2016. doi: 10.1038/ncomms12624. URL `https://doi.org/10.1038/ncomms12624`.

J. Wang, W. Zhao, H. Liu, H. He, and R. Shao. Myofibrillogenesis regulator 1 (MR-1): a potential therapeutic target for cancer and PNKD. *Journal of Drug Targeting*, 26(8): 643–648, 2018.

T. Watanabe, T. Kobunai, Y. Yamamoto, K. Matsuda, S. Ishihara, K. Nozawa, H. Iinuma, T. Konishi, H. Horie, H. Ikeuchi, K. Eshima, and T. Muto. Gene expression signature and response to the use of leucovorin, fluorouracil and oxaliplatin in colorectal cancer patients. *Clinical and Translational Oncology*, 13(6):419–425, Jun 2011. ISSN 1699-3055. doi: 10.1007/s12094-011-0676-z. URL https://doi.org/10.1007/s12094-011-0676-z.

B. D. Williamson, P. B. Gilbert, N. R. Simon, and M. Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 0(0):1–14, 2022.

S. Yadlowsky, F. Pellegrini, F. Lionetto, S. Braune, and L. Tian. Estimation and validation of ratio-based conditional average treatment effects using observational data. *Journal of the American Statistical Association*, 116(533):335–352, 2021.

A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015. ISSN 0036-8075. doi: 10.1126/science.aaa1934.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7(90):2541–2563, 2006. URL http://jmlr.org/papers/v7/zhao06a.html.

Q. Zhao, D. S. Small, and A. Ertefaie. Selective inference for effect modification via the lasso, 2018.

W. Zheng and M. J. van der Laan. *Cross-Validated Targeted Minimum-Loss-Based Estimation*, pages 459–474. Springer New York, New York, NY, 2011. ISBN 978-1-4419-9782-1. doi: 10.1007/978-1-4419-9782-1_27. URL https://doi.org/10.1007/978-1-4419-9782-1_27.

W. Zheng and M. J. van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 459–474. Springer, 2011.

W. Zhu, C. Lévy-Leduc, and N. Ternès. Identification of prognostic and predictive biomarkers in high-dimensional data with pplasso, 2022. URL https://arxiv.org/abs/2202.01970.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x.