Aggregating treatment effects across multiple outcomes

(Job market paper: click here for the latest draft)

Chun Pong Lau*

November 3, 2025

Abstract

Empirical researchers commonly observe multiple outcomes intended to measure an underlying abstract variable. For example, the abstract variable "crime" can be measured using crime rates for different types of offenses, and "wealth" can be measured using different asset ownerships. How should one aggregate these multiple outcomes into a single quantity? In this paper, I show the shortcomings of common approaches and propose a new approach to aggregate outcomes. First, I document that three methods are commonly used in the empirical literature: principal component analysis (PCA), inverse-variance matrix (IVM) weighting, and standardized averaging (SA). I show that PCA has several unattractive properties: it is sensitive to arbitrary choices of normalization, it can lead to non-standard limiting distributions, it can produce negative weights on some outcomes, and it does not even necessarily maximize precision. IVM does not suffer from the first two problems, but also has the negative weighting problem. SA is more attractive, but need not maximize precision. I use statistical decision theory to develop an approach to aggregating outcomes that minimizes mean-squared error while ensuring interpretable weights. The framework allows the researcher to flexibly incorporate prior information about the relative quality of different outcomes. It also allows for valid inference that takes the prior information into account. I apply the decision-theoretic procedure to two recent empirical applications.

*Kenneth C. Griffin Department of Economics, The University of Chicago, ccplau@uchicago.edu.
I am grateful to Alex Torgovitsky, Stéphane Bonhomme, Azeem Shaikh, and Max Tabord-Meehan for their continued guidance, support, and encouragement throughout the project. I also thank Dylan Balla-Elliott, Xinyue Bei, Debopam Bhattacharya, Chris Blattman, Yong Cai, Tim Christensen, Joshua Dean, Deniz Dutz, Max Farrell, Joseph Hardwick, Thomas Hierons, Damon Jones, Omkar Katta, Jacob Leshno, Xinran Li, José Luis Montiel Olea, Ian Pitman, Kirill Ponomarev, Guillaume Pouliot, Doron Ravid, Frank Schorfheide, Thomas Wiemann, and Davide Viviano for helpful feedback and discussions, as well as participants of the Econometrics Advising Group for valuable comments. All errors are my own.

Contents

1 Introduction

Researchers commonly observe multiple related outcomes that measure an underlying abstract variable in order to evaluate the effect of a treatment. For instance, Bruhn et al. (2018) measures entrepreneurial confidence and goal setting based on entrepreneurs' responses to questions on attitudes and beliefs. Jones et al. (2019) measures productivity using different employment and workplace-related outcomes. Bau (2022) measures wealth using various asset ownerships. Bhatt et al. (2023) measures crime involvement using different types of offenses and arrests. In order to learn the treatment effect on the abstract variable, to summarize the effect on the outcomes, or to improve interpretability, researchers often aggregate the related outcomes into one index.

Some commonly used approaches to aggregate outcomes are principal component analysis (PCA) which was popularized by Filmer and Pritchett (2001), the equally-weighted standardized averaging (SA) approach by Kling et al. (2007), and the inverse-variance matrix (IVM) weighting approach by Anderson (2008). To show the popularity of these methods, Table 1 counts the number of articles from "top-five" economics journals that applied these methods in the last five years.

In this paper, I analyze the properties and shortcomings of these approaches and propose a new method to aggregate outcomes optimally. In the first part of the paper, I show how using PCA, SA, and IVM affects precision and interpretability. Precision is measured by the asymptotic variance of the treatment effect on the aggregated outcome. Interpretability requires the weights on the outcomes to be nonnegative and sum to one. This is because negative weights can cause sign reversal in the aggregated treatment effect. The negative weights issue here is similar to but different from the negative weights issue in subpopulations, such as in difference-in-differences (e.g., de Chaisemartin and D'HaultfŒuille (2017); Goodman-Bacon (2021); Sun and Abraham (2021); Borusyak et al. (2024)) and local average treatment effects (e.g., Blandhol et al. (2025); Słoczyński (2024)).

I show that PCA-aggregated treatment effects do not maximize precision, can be difficult to interpret due to negative weights, is sensitive to an arbitrary sign normalization, and can potentially lead to a nonstandard asymptotic distribution. This is in contrast with how PCA is sometimes motivated as a way to increase power or improve interpretability. IVM can also have negative weights and does not always maximize precision. SA puts equal and positive weights on the outcomes, so it is not subject to interpretation concerns for negative weights. However, SA does not utilize the correlation structure.

In the second part of the paper, I develop a new approach to aggregate treatment

effects using statistical decision theory. The approach takes precision into account using the mean-squared error of the aggregated treatment effect as the objective, and enforces nonnegative weights to avoid interpretation issues. Since the abstract variable of interest is unobserved, I allow for two optimality criteria. The first one is the minimax criterion, which minimizes the worst-case risk. The other is the recently proposed adaptive regret concept (Armstrong et al., 2024) that measures the ratio of the cost of deviating from the optimal weight. The second approach has an advantage that it does not require specifying the level of misspecification as in the minimax criterion. In either case, I show that the weights can be computed using convex optimization.

My statistical decision framework can allow for various economic specifications. It can incorporate shape restrictions, such as some treatment effects being more important than others. I show that the decision-theoretic framework can be motivated via a communication model, where the reader has subjective weights on the treatment effects.

I illustrate the tools proposed in this paper by considering two empirical examples. First, I revisit an analysis in Campante and Do (2014) that study how isolated cities affect public good provision. The authors created a public good provision index using PCA. They found that state isolation has a negative effect on the PCA index. However, one of the outcomes in the PCA index has a negative weight. I apply my decision-theoretic approach and find that a negative effect can still be found after making one of the outcomes less important via shape restrictions. However, the effect is not significant at the original 10% level being used. Next, I revisit Bruhn et al. (2018) that studies the impact of management consulting on small and medium enterprises. To study how the consulting program affects entrepreneurs' goal setting and confidence, they use PCA and SA to create entrepreneurial spirit indices. I revisit their analysis using my decision-theoretic approach on the part where they found a significant effect on the PCA index at the 10% level but not on the SA index. I find the consulting program has a positive effect on entrepreneurial spirit, but it is not significant.

1.1 Survey on the use of PCA, SA, and IVM

Table 1 documents the practice of using PCA, SA, and IVM to aggregate outcomes. The table shows they are prominent in applied work. I restricted the search to the "top-five" economics journals between 2020 and 2024. The counts for PCA are based on searching the papers that contain the phrase "principal component" on Google Scholar and journal websites. The counts for SA and IVM are based on searching articles that cited Kling et al. (2007) and Anderson (2008), respectively, on Google Scholar. The counts in Table 1

Table 1: Usage of PCA, SA, and IVM in "top-five" economics journals in 2020–2024.

Journal \ Count	PCA	SA	IVM
American Economic Review	10	13	14
Econometrica	4	2	2
Journal of Political Economy	4	6	0
Quarterly Journal of Economics	7	7	5
Review of Economic Studies	5	5	2
Total	30	33	23

Notes: PCA refers to principal component analysis, SA refers to equally-weighted standardized averaging (Kling et al., 2007), and IVM refers to inverse-variance matrix weighting (Anderson, 2008). The above table focuses on using PCA, SA, and IVM to aggregate outcomes.

only include the articles that use these methods to aggregate outcomes.

1.2 Related literature

This paper is related to two different strands of literature in econometrics.

First, my paper is related to the literature that studies how to aggregate outcomes. O'Brien (1984) is an early work in biostatistics that studied the use of generalized least squares (GLS) to aggregate outcomes. Recently, Anderson and Magruder (2023) studied the power of SA under the assumption that treatment effects are homogeneous. Gómez (2024) conducted a simulation study on the power properties for PCA, SA, and IVM, but did not discuss negative weights or derive theoretical properties on precision for all three methods. Allee et al. (2022) reviewed the use of PCA and factor analysis in accounting research. Similar to my survey on economic papers in Table 1, they found that 219 articles used PCA or factor analysis in ten major accounting journals from 2015 to 2019. They did not discuss asymptotic variance, power, the problem of negative weights, sensitivity to normalization, and nonstandard asymptotic distribution for PCA and factor analysis. They also did not discuss SA and IVM in their review.

Apart from PCA, SA, and IVM, other approaches have been recently proposed to aggregate outcomes, but they involve a different objective, or require additional assumptions or data, such as Anderson and Magruder (2023), Hu et al. (2024), Fu and Green (2025), and Stoetzer et al. (2025). I review them in Section 3.4.2 ahead after introducing the problem formally.

Second, this paper is related to the literature on statistical decision theory and optimal estimation by Donoho et al. (1990), Donoho (1994), Cai and Low (2004), Armstrong and

Kolesár (2018, 2021a,b), and others. The idea of using adaptive regret to adapt over a range of misspecification was recently proposed by Armstrong et al. (2024), which is related to Bickel (1984) that adapts over granular sets. While I use the adaptive concept in Armstrong et al. (2024), my setting is different in that I allow all estimators to be biased, whereas they assume there is one unbiased estimator. I also restrict attention to a convex combination of the estimators to ensure interpretability on the treatment effect of the aggregated outcome. My paper is also related to the use of statistical decision theory in site selection and evidence aggregation problems, such as the recent work by Gechter et al. (2024), Ishihara and Kitagawa (2024), and Montiel Olea et al. (2025). These papers have a different goal from reporting an aggregated treatment effect. See also Wald (1950), Savage (1951), and Manski (2004) for statistical decision theory.

In concurrent and independent work, Fedchenko (2025) studies treatment effect estimation with summary indices. Similar to my paper, Fedchenko (2025) points out that negative weights can cause interpretation issues, shows how to conduct valid inference due to the data-dependent weights, and points out that the claims on power improvement do not necessarily hold. Despite the above similarities, there are several notable differences between our work. First, I develop a new statistical decision approach to aggregate, while Fedchenko (2025) recommends using SA or simple averaging. Second, I consider the treatment effect on the latent outcome as the target parameter in addition to the treatment effect on the summary index, whereas Fedchenko (2025) considers the latter parameter. Third, Fedchenko (2025) mainly focuses on SA and IVM. I also study PCA due to its popularity as documented in Table 1. Fourth, Fedchenko (2025) only points out the negative weights problem for PCA, while I show that there are three other issues with the PCA approach.

1.3 Organization of the paper

Section 2 introduces the setup and describes criteria for aggregating treatment effects. Section 3 studies commonly used methods and their shortcomings. Section 4 presents the decision-theoretic approach. Sections 5 presents two empirical applications. Section 6 concludes. All proofs can be found in the appendix.

2 Framework

This section introduces the setup and discusses some natural criteria for aggregation.

2.1 Notations

Consider a researcher who observes multiple outcomes $\boldsymbol{Y}_i \equiv (Y_{i,1}, \ldots, Y_{i,q})' \in \mathbb{R}^q$ to measure the effect of a treatment $D_i \in \mathbb{R}$, where $i = 1, \ldots, n$ indexes observation. Let $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_q)' \in \mathbb{R}^q$ be the vector of treatment effects on these outcomes (possibly using additional covariates). Researchers are often interested in learning the treatment effect on a latent outcome of interest $L_i \in \mathbb{R}$, in addition to (or instead of) the treatment effect on each of the outcomes. Let $\theta \in \mathbb{R}$ be the treatment effect of D_i on L_i. For exposition purposes, I will frequently refer to the running example below on aggregating multiple asset/wealth-related outcomes, which is a common setting in empirical work (such as Blattman et al. (2020), Banerjee et al. (2021), Bau (2022), and many others).

Example 2.1 (Running example)**.** A researcher is interested in learning the effect of a cash transfer program on wealth. The researcher collects multiple asset-related outcomes $\boldsymbol{Y}_i$, such as the ownership of goats, cows, cars, televisions, and cooling devices. Let L_i be wealth and $D_i \in \{0, 1\}$ be the indicator that equals 1 if treatment is received, and equals 0 otherwise. Here, θ is the treatment effect of D_i on L_i. △

Section 2.2 shows additional empirical applications to facilitate the interpretation of θ.

The outcomes are required to be oriented in the same direction and standardized in a way that the researcher wants to interpret the effect (e.g., using the full sample, or the control sample). In terms of the running example, the orientation requirement is satisfied when each $Y_{i,j}$ is increasing in the number or ownership of assets. This assumption is summarized below and is maintained throughout the paper.

Assumption 2.2. *The outcomes $\boldsymbol{Y}_i$ have been oriented so that a higher value means a better outcome, and they have been suitably normalized.*

For a given weight $\boldsymbol{w} \equiv (w_1, \ldots, w_q) \in \mathbb{R}^q$, the treatment effect estimator of D_i on $\boldsymbol{w}'\boldsymbol{Y}_i$ is the same as linear aggregating the treatment effects of D_i on each outcome $Y_{i,j}$ using $\boldsymbol{w}$. Lemma A.2 in the appendix formally describes this equivalence for linear models with covariates and potentially data-dependent weights. Hence, I focus on the following representation of weighted average of treatment effects to estimate θ:

$$\widehat{\tau} \equiv \boldsymbol{w}'\widehat{\boldsymbol{\beta}} = w_1\widehat{\beta}_1 + \cdots + w_q\widehat{\beta}_q, \tag{1}$$

for an estimator $\widehat{\boldsymbol{\beta}} \equiv (\widehat{\beta}_1, \ldots, \widehat{\beta}_q)'$ of $\boldsymbol{\beta}$. Without additional assumptions, $\tau \equiv \mathbb{E}[\widehat{\tau}]$ may not be equal to θ. In terms of running example (Example 2.1), this means the treatment effect of the cash transfer program on the aggregated asset $\boldsymbol{w}'\boldsymbol{Y}_i$ may not be exactly

equal to θ (i.e., the treatment effect on wealth). Thus, I use τ to distinguish it from θ. As discussed in Section 3, many approaches (including PCA, SA, and IVM) have the above representation.

The above setup assumes that the treatment effects are related to the same θ for exposition purposes. This is not necessary when one believes there is one latent outcome for each domain. Researchers can create a summary index for each domain of outcomes (e.g., as suggested in Anderson (2008)). In terms of the running example on wealth above, researchers could create indices on livestock and durable assets separately.

2.2 Additional empirical examples

I discuss additional empirical examples below apart from the running example of wealth to explain the setup.

Example 2.3 (Public good provision)**.** Campante and Do (2014) study the impact of isolated capital cities on corruption, accountability, and public good provision.

In this example, let θ be the effect of state isolation on public good provision. They observe three outcome variables related to public good provision $\boldsymbol{Y}_i$, namely, smart state index, the percentage with health insurance, and the log number of hospital beds. $\widehat{\boldsymbol{\beta}}$ is the regression estimator of how isolation affects each of the three outcomes. They aggregate the three variables using PCA to create a public goods provision index, so $\boldsymbol{w}$ is the corresponding PCA weights. △

Example 2.4 (Entrepreneurial spirit index)**.** Bruhn et al. (2018) study the impact of offering management consulting services to small and medium enterprises in Mexico. They study how consulting affects productivity, returns on assets, and "entrepreneurial spirit." Entrepreneurial spirit measures the confidence of entrepreneurs and goal setting.

Let θ be the impact of consulting on entrepreneurial spirit. They create an entrepreneurial spirit index using the response to eight outcomes $\boldsymbol{Y}_i$. Each outcome is the entrepreneur's response (coded as 1 to 5) to a survey question, such as "I have professional goals." Thus, $\widehat{\boldsymbol{\beta}}$ is the treatment effect of consulting on the responses. They create an index by PCA and SA, so $\boldsymbol{w}$ is the corresponding PCA or SA weights. △

Example 2.5 (Violence involvement)**.** Bhatt et al. (2023) study the impact of community programs on serious violence involvement θ. To measure crime involvement, they use related outcomes $\boldsymbol{Y}_i$, i.e., shooting and homicide victimizations, arrests, and other serious violent-crime arrests. $\widehat{\boldsymbol{\beta}}$ is the treatment effect of each outcome. They create a standardized index that averages the three outcomes. They also create an index using

social costs and PCA. Hence, $\boldsymbol{w}$ depends on the approach used. △

2.3 Criteria for choosing the weights

Even if one focuses on using a linear combination in (1), there are many choices for the weights. In this subsection, I discuss two criteria for choosing the weights $\boldsymbol{w} \in \mathbb{R}^q$.

2.3.1 Interpretability

The interpretability criterion requires the weights to be nonnegative and sum to one. Negative weights can cause the overall effect to be negative even though the treatment effect on each outcome is positive. For example, in the running example on wealth (Example 2.1), suppose the cash transfer program has a positive treatment effect on each asset. With negative weights, it is possible that the treatment effect on the aggregated outcome is negative, which is undesirable. In addition, negative weights can also cause reverse ordering of the aggregated outcome. For instance, suppose the weight on cows is negative in the running example. This means holding other assets fixed, an individual with more cows has worse wealth. Hence, the above shows that having negative weights in the aggregated outcome is an unattractive property.

The sign reversal issue described above is related to, but different from, the negative weights issues pointed out in the recent econometrics literature, such as Blandhol et al. (2025) and Słoczyński (2024) on instrumental variables and de Chaisemartin and D'HaultfŒuille (2017); Goodman-Bacon (2021), Sun and Abraham (2021), Borusyak et al. (2024) on difference-in-differences. In the aforementioned papers, they are concerned with the negative weights from the treatment effect of the subpopulations. In the context of aggregating outcomes, the negative weights are coming from the treatment effect of different outcomes on the same population. Regarding the issue of reversed ordering of the aggregated outcome, Vyas and Kumaranayake (2006) and Kolenikov and Angeles (2009) pointed out that PCA socio-economic status indices can have this problem as some variables can receive negative weights. Anderson and Magruder (2023) also mentions that GLS-weighted indices can have negative weights.

Based on the above reasons, it is reasonable to focus on the class of convex weights that are nonnegative and sum to one:

$$\mathcal{W}_{\text{cvx}} \equiv \{\boldsymbol{w} \in \mathbb{R}^q : \boldsymbol{w}'\mathbf{1}_q = 1, \boldsymbol{w} \geq \mathbf{0}_q\}, \tag{2}$$

where $\mathbf{1}_q$ is a vector of q ones and $\mathbf{0}_q$ is a vector of q zeros. Requiring the weights to

sum to one has two purposes. First, this restricts the scale of the weights to avoid having infinitely large weights. Second, this nests the homogeneous treatment effects case, i.e., if $\widehat{\beta}_j = \overline{\beta}$ for any $j = 1, \ldots, q$, then $\widehat{\tau} = \overline{\beta}$ for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$.

Remark 2.6. After computing the weights, researchers may "post-process" the aggregated outcome $\boldsymbol{w}'\boldsymbol{Y}_i$ by standardizing it (e.g., Parker and Vogl (2023)) or rescaling it to $[0, 1]$ (e.g., Ajzenman (2021)). I discuss the implications of this post-processing step in Appendix Section A.3.1. ■

2.3.2 Precision

A natural criterion for $\widehat{\tau}$ to be precise is to choose the weights to minimize the mean-squared error (MSE). This amounts to using the squared loss function $(\widehat{\tau} - \theta)^2$. The MSE can be written as a function of $\boldsymbol{w}$ as follows:

$$\text{MSE}(\boldsymbol{w}; \theta) \equiv \mathbb{E}[(\widehat{\tau} - \theta)^2] = \text{Var}[\widehat{\tau}] + \mathbb{E}[\widehat{\tau} - \theta]^2. \tag{3}$$

The MSE criterion evaluates the quality of the estimator by the variance and bias. In terms of the running example, the bias component measures how well the treatment effect on the weighted average of assets captures the treatment effect on wealth. The variance component evaluates the noisiness of the aggregated treatment effect. This criterion nests the special case where the treatment effect on each asset β_j equals the treatment effect on wealth θ, i.e., $\beta_j = \theta$ for each $j = 1, \ldots, q$. In this special case, the MSE becomes the variance. My decision-theoretic approach to be introduced in Section 4 shows how to handle the bias formally without making this assumption.

3 Common aggregating methods and their shortcomings

In this section, I study common methods for aggregating outcomes and show their shortcomings. Each method is mainly evaluated using the precision and interpretation criteria described in Section 2.3. I also evaluate precision in terms of the asymptotic variance of $\widehat{\tau}_n$. I show that there are additional issues for PCA.

3.1 Principal component analysis (PCA)

PCA is a dimension-reduction tool that was developed more than a century ago (Pearson, 1901; Hotelling, 1933). PCA transforms a set of correlated variables into a new set of

uncorrelated variables called principal components. Refer to Table 1 for the popularity of using PCA to aggregate treatment effects.

PCA is often performed on the correlation matrix of the outcomes, denoted as $\mathbf{\Sigma}_Y$. Using a correlation matrix is preferred to a covariance matrix because the outcomes have different units, so PCA on the covariance matrix is not scale invariant (Jolliffe, 2002). In this subsection, I maintain Assumption 2.2 such that $\boldsymbol{Y}_i$ represents the standardized outcomes and $\boldsymbol{\beta}$ is the treatment effect on such standardized outcomes.

The PCA approach aggregates outcomes using the first principal component (PC1) of $\mathbf{\Sigma}_Y$, i.e., it finds the vector $\boldsymbol{w} \equiv (w_1, \ldots, w_q) \in \mathbb{R}^q$ that maximizes the variance of the linear combination of the outcomes $\text{Var}[\boldsymbol{w}'\boldsymbol{Y}_i] = \boldsymbol{w}'\mathbf{\Sigma}_Y\boldsymbol{w}$. For textbook expositions on PCA, see, for instance, Jolliffe (2002) and Jackson (2005).

Section 3.1.1 reviews the relevant theory of PCA. Sections 3.1.2 and 3.1.3 evaluate PCA using the interpretation and precision criteria, respectively. Section 3.1.4 explains that PCA-aggregated treatment effects suffer from an arbitrary sign normalization. Section 3.1.5 shows that the PCA-aggregated treatment effect can potentially have a nonstandard asymptotic distribution.

3.1.1 The PCA problem

The problem of finding the PC1 of $\mathbf{\Sigma}_Y$ can be written as

$$\boldsymbol{w}_{\text{pca}} = \underset{\boldsymbol{w} \in \mathcal{W}_{\text{unit}}}{\arg\max} \quad \boldsymbol{w}'\mathbf{\Sigma}_Y\boldsymbol{w}, \tag{4}$$

where the class of weights is

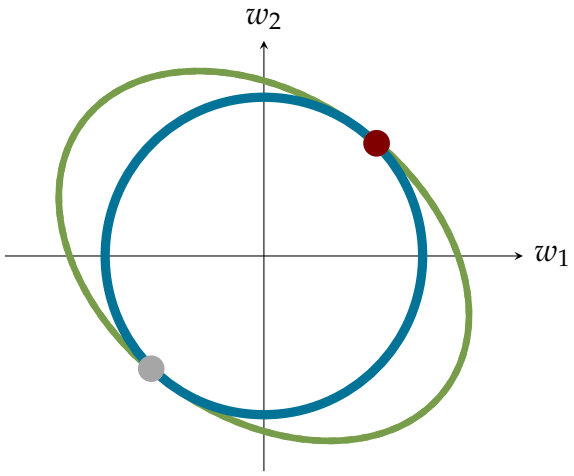
$$\mathcal{W}_{\text{unit}} \equiv \{\boldsymbol{w} \in \mathbb{R}^q : \boldsymbol{w}'\boldsymbol{w} = 1, \boldsymbol{c}'\boldsymbol{w} \geq 0\}, \tag{5}$$

for a given $\boldsymbol{c} \in \mathbb{R}^q$. I explain the above choices and the role of $\boldsymbol{c}$ in (5) below.

First, the PC1 of $\mathbf{\Sigma}_Y$ is the leading eigenvector of $\mathbf{\Sigma}_Y$. The PC1 of $\mathbf{\Sigma}_Y$ is unique when the largest eigenvalue is unique. To formalize this, let $\{(\boldsymbol{\nu}_j, \lambda_j)\}_{j=1}^q$ be the eigenpairs of $\mathbf{\Sigma}_Y$, where $\{\lambda_j\}_{j=1}^q$ are the eigenvalues of the matrix $\mathbf{\Sigma}_Y$ and $\boldsymbol{\nu}_j$ is a unit-length eigenvector corresponding to λ_j for each $j = 1, \ldots, q$. I assume the eigenvalues are ordered as $\lambda_1 \geq \cdots \geq \lambda_q$. The assumption below ensures the leading eigenvector of $\mathbf{\Sigma}_Y$ is unique. The problem of having repeated eigenvalues is briefly discussed in Remark 3.7.

Assumption 3.1 (Unique leading eigenvector for $\mathbf{\Sigma}_Y$)**.** $\lambda_1 > \lambda_2$.

Figure 1: Sign normalization for the PCA problem.



Notes: The green ellipse represents the objective of the PCA problem. The blue unit-length circle represents the unit-length constraint. Both the grey and red dots are on the largest ellipse that contains the unit-length circle. The red dot is identified by the $w_1 + w_2 \geq 0$ constraint.

Second, normalization on the length of $\boldsymbol{w}$ is required because the variance $\text{Var}[\boldsymbol{w}'\boldsymbol{Y}_i]$ is unbounded without any restrictions on $\boldsymbol{w}$. The unit-length restriction

$$\boldsymbol{w}'\boldsymbol{w} = w_1^2 + \cdots + w_q^2 = 1 \tag{6}$$

in (5) is commonly used. Other length normalizations are possible, although the results would change based on the choice of normalization and they could lead to a more difficult optimization problem (see, for instance, Jolliffe (2002)).

Third, the inequality $\boldsymbol{c}'\boldsymbol{w} \geq 0$ in (5) is an identification condition used to obtain a unique solution. This identification condition is needed because the leading eigenvector is unique up to a sign change even though (6) is imposed (i.e., both $\boldsymbol{\nu}_1$ and $-\boldsymbol{\nu}_1$ are leading eigenvectors of $\boldsymbol{\Sigma}_Y$). Figure 1 explains this via the geometry of the PCA problem, where the PC1 problem finds the largest ellipse (i.e., the objective $\boldsymbol{w}'\boldsymbol{\Sigma}_Y\boldsymbol{w}$ in (4)) subject to the unit circle (i.e., the unit-length constraint (6)). The unit circle alone is not sufficient to pin down a unique solution because both the grey and red dots are valid solutions. The inequality constraint $w_1 + w_2 \geq 0$ can be used to rule out the grey point and leads to a unique solution. However, the choice of the identification constraint is arbitrary. Table 2 shows how several common programming languages implement the additional sign constraints to identify the unique solution to the PCA problem.

Based on Table 2, the constraint $\boldsymbol{c}'\boldsymbol{w} > 0$ covers the Stata implementation by setting $\boldsymbol{c}$ as a vector of ones (note that Stata imposes a strict inequality based on the manual). This

Table 2: Sign constraints for PCA in common implementations.

Programming language	**Sign constraints**
Stata	The `pca` command chooses the sign of the loadings such that the sum of each principal component is positive (StataCorp, 2025).
Matlab	The `pca` command chooses the sign of the loadings such that the largest component of each principal component is positive (The MathWorks Inc., 2025).
R	The `princomp` command sets the signs of the loadings such that the first entry of each principal component is nonnegative as the default (R Core Team, 2025).

also covers the R implementation by setting $\boldsymbol{c}$ as a vector where the first entry equals 1 and the remaining entries are 0. This normalization means the ordering of the variables matters. The normalization used by Matlab cannot be written as one linear constraint, but can be analyzed similarly.

I conclude this subsection with the following remarks.

Remark 3.2. It can be shown that the PCA problem in (4) using an estimator for $\boldsymbol{\Sigma}_Y$ is equivalent to the following problem that finds the best-fitted hyperplane to the vector of outcomes

$$\begin{aligned} \min_{\boldsymbol{F}\in\mathbb{R}^n, \boldsymbol{w}\in\mathbb{R}^q} \quad & \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{w}F_i)'(\boldsymbol{Y}_i - \boldsymbol{w}F_i), \\ \text{s.t.} \quad & \boldsymbol{w}'\boldsymbol{w} = 1, \end{aligned}$$

where $\boldsymbol{F} \equiv (F_1, \ldots, F_n) \in \mathbb{R}^n$. See, for instance, Bai and Ng (2002, Section 3) or James et al. (2021, Section 12.2.2), for related discussions. The length normalization constraint above is chosen to be the same as (6). ■

Remark 3.3. The PCA problem in (4) can be written as

$$\min_{\boldsymbol{w}\in\mathcal{W}_{\text{unit}}} \quad \boldsymbol{w}'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{w}$$

when $\boldsymbol{\Sigma}_Y$ is positive definite. Thus, the PCA problem can be viewed as finding the smallest eigenvector of the precision matrix $\boldsymbol{\Sigma}_Y^{-1}$. ■

3.1.2 Interpretation

The class of weights (5) used by PCA does not restrict the weights to be positive. Whether the weights can be all positive depends on the correlation. It is known that if $\mathbf{\Sigma}_Y$ has only positive elements, then all the terms in the leading eigenvector of $\mathbf{\Sigma}_Y$ have the same sign using the Perron-Frobenius theorem (see, for instance, Exercise 8.7.1 of Mardia et al. (1979)). Hence, unless $\mathbf{\Sigma}_Y$ has only positive entries, it is possible for the PC1 of $\mathbf{\Sigma}_Y$ to have different signs. Outcomes can be negatively correlated with each other even if they are all positively affected by the treatment. In the following, I discuss two possibilities where negative correlations can occur.

The first possibility where this may occur is when the set of outcomes includes goods that are substitutes. In terms of the running example on wealth, some outcomes that are substitutes for each other might be included in the wealth index. For instance, the asset index created by Bau (2022) contains assets that are likely to be substitutes (air conditioner, air cooler, and fan). In the data, air conditioners are negatively correlated to air coolers and fans. The PCA asset index gave negative weights to air coolers and fans (see Figure A.3 in the appendix for the weights). To understand how negative correlation arises with substitutes, I consider a stylized two-good example below.

Example 3.4 (Negative weights in PCA index due to substitutes)**.** This example considers a consumer optimization problem with two goods, where the consumer's utility is increasing in both goods. Suppose the researcher creates a summary index by PCA using the two goods. I show that negative weights arise in this example due to a negative correlation between the two outcomes.

Consider the problem below where the goods are substitutes (e.g., cows and goats):

$$\begin{aligned}(Y_{i,1}^\star, Y_{i,2}^\star) = \arg\max_{y_1, y_2} \quad & y_1^{A_i} y_2^{1-A_i}, \\ \text{s.t.} \quad & y_1 + p_2 y_2 \leq \text{Income}_i(D_i),\end{aligned} \tag{7}$$

where the price of good 1 is 1, p_2 is the price of good 2, $\text{Income}_i(D_i) \geq 0$ is the income of individual i depending on $D_i \in \{0, 1\}$, and $A_i \in (0, 1)$ is a random utility parameter.

The optimal solution to problem (7) is $Y_{i,1}^\star = A_i \text{Income}_i(D_i)$ and $Y_{i,2}^\star = \frac{(1-A_i)\text{Income}_i(D_i)}{p_2}$. I consider the general scenario in Appendix A.1.1. For concreteness, suppose $p_2 = 2$, A_i equals 0.2 or 0.8 with equal probabilities, D_i equals 0 or 1 with equal probabilities, $\text{Income}_i(D_i) = 10 + 5D_i + V_i$, where $V_i \sim \text{Uniform}[0, 5]$ and (V_i, D_i, A_i) are mutually independent. Then, $\text{Corr}[Y_{i,1}^\star, Y_{i,2}^\star] \approx -0.82 < 0$.

The negative correlation in the two outcomes is related to the two goods being substitutes for each other. As a result, running PCA on these two goods would lead to an index with negative weights. However, D_i has a positive effect on both outcomes. Hence, using PCA to create an index to represent utility here is unreasonable.

In Appendix A.1.2, I further analyze a case with perfect substitutes. △

Another possibility for negative correlation to arise is when one discretizes a continuous variable into mutually exclusive binary indicators (see Kolenikov and Angeles (2009) for more discussion). To see this in terms of the running example of wealth index in Example 2.1, let Y_i be a categorical variable with support $\{y_0, y_1, \ldots, y_L\}$ that indicates the type of cooling device owned by a household. If the following binary indicators are created to represent ownership of the types of cooling device $Z_{i,l} = \mathbb{1}[Y_i = y_l]$ for $l = 1, \ldots, L$, then $\text{Cov}[Z_{i,l}, Z_{i,k}] = -\mathbb{E}[Z_{i,l}]\mathbb{E}[Z_{i,k}] \leq 0$ for $l \neq k$.

Finally, the class of weights $\mathcal{W}_{\text{unit}}$ in (5) requires the weights to be of unit length. In addition to the possibility for PCA to have negative weights, the weights do not typically sum to one. Hence, even if $\beta_j = \overline{\beta}$ for all $j = 1, \ldots, q$, $\boldsymbol{w}'\boldsymbol{\beta}$ can be different from $\overline{\beta}$.

3.1.3 Precision

This subsection studies how using PCA to aggregate outcomes affects the precision of the aggregated treatment effect. It is often mentioned that PCA is used because it maximizes the variance of the linear combination of the outcomes, or it is a tool for dimension reduction. While these statements follow from the definition of PCA, I show that these properties do not necessarily translate to a precise aggregated treatment effect.

Using PC1 to weight the outcomes does not necessarily minimize the asymptotic variance of the aggregated treatment effect estimator. This is because PCA is a maximization problem involving $\boldsymbol{\Sigma}_Y$ and not directly related to the minimization of variance of the aggregated treatment effect. I show the details in the appendix. To illustrate this, I consider an example with binary treatment below.

The analysis below studies how running PCA on the correlation matrix of the outcomes does not lead to an aggregated treatment effect with low variance. Thus, the analysis below assumes the variance matrices are known. In practice, such matrices have to be estimated, and the weights are computed by running PCA on such estimated matrices. Computing the correct standard errors of the aggregated treatment effect requires taking these estimated weights into account (details are in Appendix A.7).

Example 3.5. Suppose the researcher observes q asset outcomes to evaluate the effect of

a cash transfer program $D_i \in \{0,1\}$ as in Example 2.1. Let $\boldsymbol{Y}_i$ be the vector of suitably standardized outcomes (as maintained in Assumption 2.2) such that

$$Y_{i,j} = \xi_j + \beta_j D_i + U_{i,j}, \tag{8}$$

where $\xi_j, \beta, U_{i,j} \in \mathbb{R}$. Let $\widehat{\boldsymbol{\beta}}_n \equiv (\widehat{\beta}_{n,1}, \ldots, \widehat{\beta}_{n,q})'$ be the estimator for $\boldsymbol{\beta}$. Let the treatment effect on the wealth index be $\widehat{\tau}_n \equiv \boldsymbol{w}'\widehat{\boldsymbol{\beta}}_n = \sum_{j=1}^{q} w_j \widehat{\beta}_{n,j}$ for a fixed $\boldsymbol{w} \in \mathbb{R}^q$. Under standard assumptions, the asymptotic variance of $\widehat{\tau}_n$ is given by

$$\sigma_\tau^2(\boldsymbol{w}) \equiv \boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}, \tag{9}$$

where $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ is the asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$. Note that computing $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ has to take into account that each outcome is standardized by the sample standard deviation. The expression of $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ is shown in (A.22) in the appendix. Requiring the PC1 of $\boldsymbol{\Sigma}_Y$ to minimize $\sigma_\tau^2(\boldsymbol{w})$ means the smallest eigenvector of $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ has to be equal to the leading eigenvector of $\boldsymbol{\Sigma}_Y$. This is generally a strong condition. For exposition purposes, I present a simplified analysis below and defer the details to Appendix A.4.

Assume homoskedastic errors, $\mathrm{Var}[Y_{i,j}] = 1$ and $\beta_j = 0$ for $j = 1, \ldots, q$. Then,

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \frac{1}{p_D(1-p_D)}\boldsymbol{\Sigma}_Y, \tag{10}$$

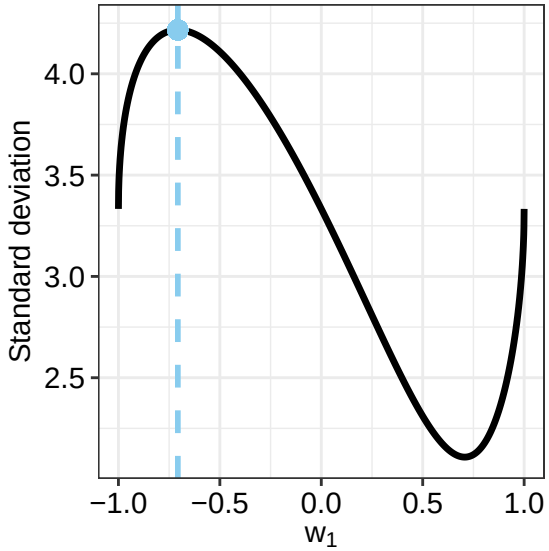
where $p_D \equiv \mathbb{P}[D_i = 1]$. Here, the PCA problem that maximizes $\boldsymbol{w}'\boldsymbol{\Sigma}_Y\boldsymbol{w}$ is the same as maximizing the asymptotic variance of the aggregated treatment effect $\boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}$.

For concreteness, Figure 2 shows a numerical example using the above model with $q = 2$, $\beta_1 = \beta_2 = 0$, $p_D = 0.1$, and $\mathrm{Cov}[Y_{i,1}, Y_{i,2}] = -0.6$. The figure plots the asymptotic standard deviation of $\widehat{\tau}_n$, i.e., $\sigma_\tau(\boldsymbol{w})$ in (9), against w_1 (where $w_2 = \sqrt{1 - w_1^2}$ is required to be nonnegative). Since $\mathrm{Var}[Y_{i,1}, Y_{i,2}] < 0$, a leading eigenvector of $\boldsymbol{\Sigma}_Y$ is $(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. But this vector maximizes $\sigma_\tau(\boldsymbol{w})$. △

Next, consider the MSE criterion in (3). PCA does not minimize this in general as well following a similar argument as in Example 3.5. In addition, even if the variance is known, it does not provide information about θ.

Apart from variance, PCA also requires strong conditions to maximize the t-statistic. This is in contrast with how PCA is sometimes motivated as a way to improve power. The analysis on t-statistic requires additional notations, so I summarize the main idea here and show the details in Appendix A.5. In terms of the setting in Example 3.5, the

Figure 2: A two-outcome numerical example that evaluates the precision of PCA.



Notes: The above plots the asymptotic standard deviation of the aggregated treatment effect against w_1. See Example 3.5 for the data-generating process. The vertical dashed line represents the choice made by PCA.

t-statistic squared has a one-to-one relationship with R-squared. Thus, to maximize R-squared, the weights should be chosen to "separate" the means of $\boldsymbol{w}'\boldsymbol{Y}_i$ for the treated and control groups as much as possible. This means $\boldsymbol{w}'\boldsymbol{\beta}$ is also important, but PCA on $\boldsymbol{\Sigma}_Y$ does not achieve this goal without additional conditions because $\boldsymbol{\Sigma}_Y$ pools the information from treated and control groups together. I also discuss the more general cases in the appendix.

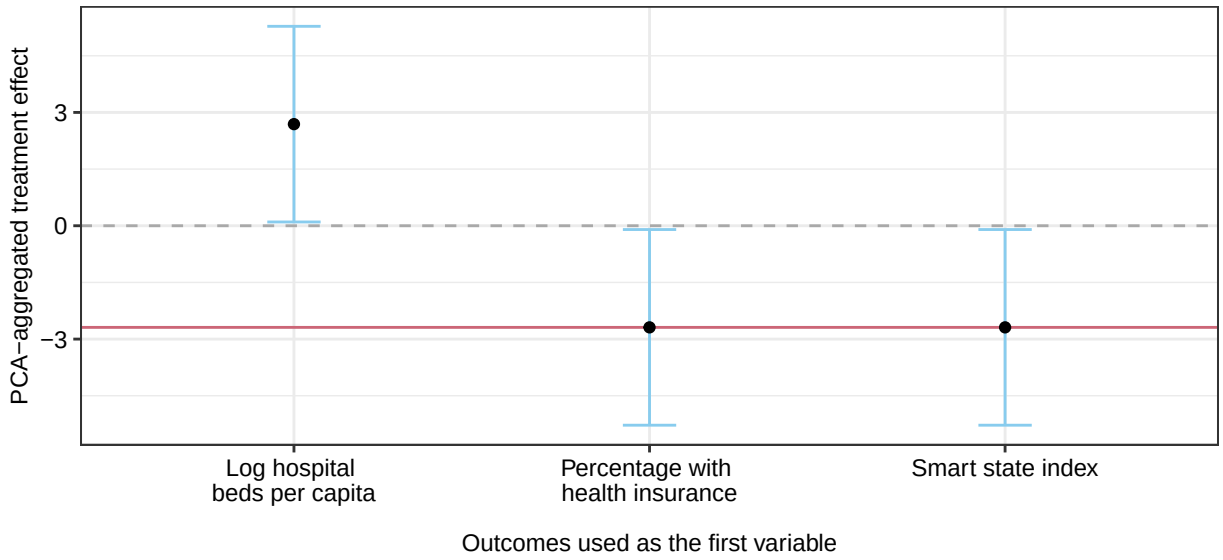
3.1.4 Sensitivity to the sign (identification) constraint

The uniqueness of the solution to (4) is related to the inequality constraint $\boldsymbol{c}'\boldsymbol{w} \geq 0$ in (5) to rule out one solution. However, there is no unique method of imposing the identification constraint as reviewed in Table 2. This causes the results to be sensitive to the identification condition. In particular, changing the identification condition can potentially lead to an aggregated treatment effect to be of different sign. I illustrate this via the following empirical example.

Example 2.3 (revisited). This example revisits one of the analyses in Campante and Do (2014) that generates a public good provision index using PCA on three outcomes. It shows that the treatment effect on the PCA index is sensitive to the sign constraint.

This example replicates the OLS analysis in R to generate a PCA index and then re-

Figure 3: The PCA-aggregated treatment effects replicated using R.



Notes: The x-axis shows which variable is the first variable passed to the function `princomp` in R. The points show the PCA-aggregated treatment effect. The blue bars show the 90% confidence interval that uses the robust option as in the original Stata replication code. The red horizontal line shows the PCA-aggregated treatment effect obtained from Stata.

gresses it on average distance from the state capital and other controls. As reviewed in Table 2, the `princomp` function depends on what the "first variable" is. Figure 3 uses different outcomes as the "first variable" and shows the coefficient on the average distance away from the capital using the `princomp` function in R. It shows that positive or negative effects are possible when different outcomes are used as the "first variable," while being significant at the 10% level. Standard errors are computed as in the authors' replication code. See Appendix A.7 on computing standard errors with the generated outcomes. The red line shows the treatment effect estimated using the authors' replication code in Stata. △

When all treatment effects and PCA weights have the same sign, it might be clear what the "correct" choice of the sign identification constraint is. However, it can create an ambiguity when the signs of the treatment effects and loadings are mixed.

3.1.5 Nonstandard asymptotic distribution

In this subsection, I show that PCA-aggregated treatment effects can have a nonstandard asymptotic distribution due to weak identification. This is due to the $\boldsymbol{c}'\boldsymbol{w} \geq 0$ constraint in (5) that is used to achieve identification.

If $\boldsymbol{c}'\boldsymbol{w}_{\text{pca}}$ is "near" 0, this creates a weak identification issue. Intuitively, this is because it becomes more difficult to pin down the unique solution when compared to the strong identification case that $\boldsymbol{c}'\boldsymbol{w}_{\text{pca}}$ is "far away" from 0. This observation and its implication on inference is summarized in the proposition below, where I use a drifting sequence to model the behavior that $\boldsymbol{c}'\boldsymbol{w}_{\text{pca}}$ is "near" 0. This is related to the approach used to study the issue of weak instruments by Staiger and Stock (1997). To analyze the large-sample properties, Assumption A.1 (stated in the appendix) requires that the estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_Y$ are consistent and a suitable central limit theorem applies.

Proposition 3.6. *Let Assumptions 2.2, 3.1 and A.1 hold. Suppose* $\boldsymbol{c}'\boldsymbol{w}_{\text{pca},n} = \frac{\delta}{\sqrt{n}}$ *for some* $\delta > 0$. *Let* $\widehat{\boldsymbol{w}}_{\text{pca},n}$ *be the solution to* (4) *using a consistent estimator of* $\boldsymbol{\Sigma}_Y$, *and* $\tau_{\text{pca},n} \equiv \boldsymbol{w}'_{\text{pca},n}\boldsymbol{\beta}$. *Consider* $\mathcal{C} = [t_{\text{lb}}, t_{\text{ub}}]$ *where* $-\infty < t_{\text{lb}} < t_{\text{ub}} < \infty$. *If* $\tau_{\text{pca}} \neq 0$, *then*

$$\lim_{\delta \downarrow 0} \lim_{n \to \infty} \mathbb{P}[\sqrt{n}(\widehat{\tau}_{\text{pca},n} - \tau_{\text{pca},n}) \in \mathcal{C}] \leq 0.5.$$

The analysis uses matrix/eigenvector perturbation (see, for instance, Stewart and Sun (1990), for a comprehensive reference) because the PCA problem is not convex (the equality constraint is not affine), and eigenvectors do not generally have a closed-form unless in special cases. With strong identification, it is still possible to obtain asymptotic normality. I show the details for the strong identification case in Appendix A.7.1. I illustrate the issue with weak identification by a calibrated simulation exercise in Appendix A.8.

Finally, the following remark states that the above is not the only possibility to have weak identification.

Remark 3.7. Having repeated leading eigenvalues (i.e., relaxing Assumption 3.1) can lead to a similar identification issue. This is because in this case, the leading eigenvector is no longer unique. ■

3.2 Inverse variance matrix weighting (IVM)

Anderson (2008) creates a summary index by a weighted average of standardized outcomes using the inverse of the covariance matrix. Following pages 1485 and 1949 of Anderson (2008), Assumption 2.2 is imposed such that each outcome is demeaned and divided by its control group standard deviation.

The weights are given by

$$\boldsymbol{w}_{\text{ivm}} \equiv \frac{\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q}, \tag{11}$$

where the covariance matrix $\boldsymbol{\Sigma}_Y$ is obtained from the outcomes standardized as described above. The weight in (11) can be obtained by solving

$$\min_{\boldsymbol{w}\in\mathcal{W}_{\text{sto}}} \ \boldsymbol{w}'\boldsymbol{\Sigma}_Y\boldsymbol{w}, \tag{12}$$

where the class of weights requires the entries to sum-to-one ("sto"):

$$\mathcal{W}_{\text{sto}} \equiv \{\boldsymbol{w} \in \mathbb{R}^q : \boldsymbol{w}'\mathbf{1}_q = 1\}. \tag{13}$$

Anderson (2008, page 1485) discussed that using the covariance matrix in weighting ensures highly-correlated outcomes receive less weight, and that the resulting weight is the efficient generalized least squares (GLS) estimator (O'Brien, 1984). The connection with GLS can be found in the theorem on page 1082 of O'Brien (1984), which states that if $\boldsymbol{Y}$ is a vector of q unbiased estimator of a scalar parameter with covariance matrix $\boldsymbol{\Sigma}_Y$, then the corresponding is best linear unbiased estimate is given by $\boldsymbol{w}_{\text{ivm}}'\boldsymbol{Y}$.

3.2.1 Interpretation

The class of weights (13) used by IVM only imposes a length normalization, with the signs being unrestricted. The denominator of the weights (11) is positive because $\boldsymbol{\Sigma}_Y$ is positive definite. However, the numerator can contain negative entries because it involves summing rows of the inverse of $\boldsymbol{\Sigma}_Y$. Hence, the weights can be negative.

3.2.2 Precision

The objective in (12) is not necessarily the same as minimizing the asymptotic variance on the aggregated treatment effect because it focuses on the variance on outcomes. It also does not equal the MSE in (3) without additional assumptions on θ.

In the following, I revisit the binary treatment example. Similar to Section 3.1.3, the below analysis assumes the variance matrices are known. In practice, such matrices have to be estimated and (11) is computed using such estimated matrix. I defer the discussion of generated outcomes and how to correctly compute standard errors to Appendix A.7.

Example 3.8. Consider the binary treatment setup as in Example 3.5 with the exception that each outcome is divided by the standard deviation from the control group as in Anderson (2008). Thus, the derivation for the asymptotic variance is similar as Example 3.5 with the difference on how the outcomes are standardized.

Minimizing $\boldsymbol{w}'\boldsymbol{\Sigma}_Y\boldsymbol{w}$ and $\boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}$ over $\boldsymbol{w} \in \mathcal{W}_{\text{sto}}$ do not necessarily lead to the same solution. One possibility that they give the same solution is with homoskedasticity, $\boldsymbol{\beta} = \mathbf{0}_q$, and each outcome has a variance of 1 in the control group. This is because in this specification, $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ is a scalar multiple of $\boldsymbol{\Sigma}_Y$ as in (9). However, this may not be true with heteroskedasticity or more general setup. I show the details in Appendix A.4.

Next, consider the MSE in (3). Even if the variance of $\widehat{\boldsymbol{\beta}}_n$ is known, it still requires the knowledge of θ, so IVM does not minimize MSE without additional assumptions. △

Similar to the PCA analysis, $\boldsymbol{w}_{\text{ivm}}$ does not necessarily maximize t-statistic or improve power. The details on the connection between the t-statistic and IVM can be found in Appendix A.5.

3.3 Equally-weighted standardized averaging (SA)

Kling et al. (2007) creates a summary index by taking an equally-weighted average of treatment effects after each treatment effect is demeaned by the control group's mean and divided by the standard deviation of the corresponding outcome in the control group. Hence, their weight is

$$\boldsymbol{w}_{\text{sa}} \equiv \left(\frac{1}{q}, \ldots, \frac{1}{q}\right)' = \frac{1}{q}\mathbf{1}_q.$$

Since the above weights are positive, there is no interpretation issues from negative weights as in the two previous methods.

Kling and Liebman (2004) and the appendix of Kling et al. (2007) discussed why they divide by the control variances. From footnote 13 of Kling and Liebman (2004): this is for comparison on effect size relative to the control group, which is related to Glass's delta (see, for instance, Glass et al. (1981)). They compute the sample variance of the resulting weighted average of treatment effects via seemingly unrelated regression in order to account for the correlation in $\widehat{\boldsymbol{\beta}}_n$.

3.3.1 Precision

Equal weighting minimizes the asymptotic variance of $\widehat{\tau}_n$ if the asymptotic variance of the vector of treatment effects is an equicorrelation matrix. This is summarized in the proposition below, where the result follows from a constrained optimization calculation.

Proposition 3.9. *Let $\overline{\mathbf{\Sigma}}$ be an equicorrelation matrix, i.e., $\overline{\mathbf{\Sigma}}$ has diagonal entries equal 1 and off-diagonal entries all equal to $\overline{\rho} \in (-\frac{1}{q-1}, 1)$. Then, $\boldsymbol{w}_{\text{sa}} = \arg\min_{\boldsymbol{w} \in \mathcal{W}_{\text{sto}}} \boldsymbol{w}' \overline{\mathbf{\Sigma}} \boldsymbol{w}$.*

The length normalization choice that $\boldsymbol{w}$ sums to one in Proposition 3.9 to restrict the length of $\boldsymbol{w}$ is a natural minimal assumption such that $\boldsymbol{w}_{\text{sa}} \in \mathcal{W}_{\text{sto}}$ in (13). The restriction on the correlation is to maintain positive definiteness of $\overline{\mathbf{\Sigma}}$ (see, for instance, Abadir and Magnus (2005, page 241)).

While the above shows under what conditions equal weighting can be optimal, using the variance matrix to weight the outcomes can increase precision. I demonstrate this using a stylized example with duplicated variables in Appendix A.6 as a limiting case of having highly correlated outcomes.

Similar to the previous analysis, $\boldsymbol{w}_{\text{sa}}$ does not necessarily maximize t-statistic or improve power. The details on the connection between the t-statistic and SA can be found in Appendix A.5.

3.4 Discussion

The analysis above shows that PCA, SA, and IVM do not necessarily minimize the variance of the aggregated treatment effects. PCA and IVM can also have interpretation issues due to negative weights. In Section 3.4.1, I explain how to choose the weights to minimize the variance and show the large-sample properties. In Section 3.4.2, I explain other related methods.

3.4.1 Variance-minimizing weights

The previous subsections showed that PCA, SA, and IVM do not necessarily minimize the asymptotic variance of the aggregated treatment effect. This subsection explains how to choose the weights and conduct inference on the aggregated treatment effect if the goal is to minimize the asymptotic variance. The weights are required to be nonnegative and sum to one to ensure interpretability as defined in Section 2.3.1. A caveat to the analysis is that this ignores bias. This can be thought of as computing the optimal weights with $\beta_j = \theta$ for $j = 1, \ldots, q$. I will discuss how to take bias into account using

statistical decision theory in Section 4.

Let $\widehat{\boldsymbol{\beta}}_n$ be the estimator for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ be the asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$, where $\boldsymbol{\Sigma}$ is positive definite. Let $\boldsymbol{w}_{\text{vmc}}$ be the solution to the following problem

$$\min_{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}} \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}, \tag{14}$$

where $\mathcal{W}_{\text{cvx}}$ is defined in (2) and "vmc" is the acronym for variance minimization subject to convex weights. Let $\widehat{\boldsymbol{w}}_{\text{vmc},n}$ be the solution to (14) when $\boldsymbol{\Sigma}$ is replaced by its consistent estimator $\widehat{\boldsymbol{\Sigma}}_n$. The following proposition characterizes the limiting distribution of $\widehat{\boldsymbol{w}}'_{\text{vmc},n}\widehat{\boldsymbol{\beta}}_n$. Assumption A.25 requires consistency, and a central limit theorem applies for $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$.

Proposition 3.10. *Let Assumptions 2.2 and A.25 hold. Then,*

$$\sqrt{n}(\widehat{\boldsymbol{w}}'_{\text{vmc},n}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}'_{\text{vmc}}\boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{w}'_{\text{vmc}}\boldsymbol{Z}_{\boldsymbol{\beta}} + \boldsymbol{\beta}'\boldsymbol{h}^{\star}(\widetilde{\boldsymbol{Z}}),$$

where $\boldsymbol{Z}_{\boldsymbol{\beta}}$ and $\boldsymbol{h}^{\star}(\widetilde{\boldsymbol{Z}})$ are given in Theorem A.27.

The term $\boldsymbol{h}^{\star}(\widetilde{\boldsymbol{Z}})$ captures the limiting distribution of the estimated weights, and can be nonnormal due to the nonnegativitiy constraints in (2). Nevertheless, asymptotic normality can be restored when $\boldsymbol{w}_{\text{vmc}} > \boldsymbol{0}_q$. I show the details in Appendix A.7.4.

3.4.2 Other methods and discussion

Apart from PCA, IVM, and SA, other methods have been proposed to aggregate outcomes and treatment effects. These methods typically require additional assumptions, or data, or tuning parameters. I briefly review some of these methods below.

Recently, Hu et al. (2024) takes a measurement error approach and creates a linear index for the latent variable. Their index require identification assumptions related to nonlinear models with measurement errors (Hu and Schennach, 2008) to identify the joint distribution of the latent variable and the observed outcomes. Fu and Green (2025) uses an instrumental variables strategy to identify the treatment effect on the latent outcome of interest. Their strategy requires exclusion restrictions and independence assumptions and uses either the treatment or other measurements as an instrumental variable. My approach is different in that I do not require a specific number of outcomes or extra identification assumptions.

Anderson and Magruder (2023) is related in that they propose machine learning approaches to choose the weights to maximize power for aggregating outcomes in pre-

analysis plans. They penalize their power function using a tuning parameter on the Herfindahl-Hirschman index to avoid putting all weights on one outcome. Their approach requires cross-validation and choosing different possible outcomes. I take a statistical decision approach, focus on mean-squared error as the objective, and allow researchers to flexibly model the parameter space. My approach with adaptive regret (to be introduced in Section 4) does not require the researcher to use a specific level of misspecification.

Stoetzer et al. (2025) proposes a hierarchical item response approach to estimate the treatment effect on the latent outcome using observed indicators. They require parametric assumptions on the latent variables and related indicators. Item response theory has also been used in economics and psychometrics. In the psychometrics literature, Douglas (2001) shows that when binary proxies are used, point identification of the latent variable would require a large number of binary proxies. My approach does not require a specific number of outcomes. See Williams (2019) for related identification results of using item response theory to identify a latent variable that is used as an independent variable. Lubotsky and Wittenberg (2006) also focuses on the scenario where there are available proxies for a latent variable as an independent variable. Their strategy depends on whether the error components in the proxies are independent, and they recommend reporting both the instrumental variables estimator and the least squares estimator without additional assumptions on the error terms.

Using additional data on costs and benefits, researchers have reported an aggregated measure in terms of costs and benefits. For instance, Heckman et al. (2010) estimates the benefit-cost ratio and the rate of return for the Perry Preschool Program, and Bhatt et al. (2023) constructs an index based on social costs. I do not assume that researchers have access to such data and focus on the statistical problem of aggregating outcomes.

Finally, researchers sometimes motivate the use of aggregation due to concerns about multiple testing. Whether multiple hypothesis testing should be used depends on the decision the researcher wants to make. See Romano et al. (2010) on a survey of multiple testing, and Viviano et al. (2025) for an economic model on when multiple hypothesis testing is appropriate.

4 A statistical decision approach

In this section, I develop a statistical decision framework to aggregate treatment effects. The framework provides a principled approach using the mean-squared error in Section

2.3 as the objective, while ensuring interpretable weights. It also allows researchers to flexibly incorporate their information on the relative quality of the outcomes.

To develop the theory, I introduce the decision problem in Section 4.1. Section 4.2 discusses the parameter space. Two approaches to estimate the weights are then considered. Section 4.3 considers the minimax approach and Section 4.4 considers the adaptive approach. Section 4.5 shows how to conduct inference. Section 4.6 provides an implementation algorithm and summarizes.

4.1 The decision problem

This subsection introduces the researcher's problem and defines the risk function. Assume the researcher observes a vector of treatment effects $\widehat{\boldsymbol{\beta}}$ that follows

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{15}$$

where $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_q) \in \mathbb{R}^q$ is the vector of population means and the variance matrix $\boldsymbol{\Sigma}$ is assumed to be a known positive definite matrix. The above is the treatment effects on the standardized outcomes (Assumption 2.2). The normal distribution in (15) can be justified by a large-sample approximation and is mainly for exposition purposes.

The Gaussian assumption on treatment effects above has also been used in other contexts, such as recently in site selection and evidence aggregation problems (e.g., Gechter et al. (2024), Ishihara and Kitagawa (2024), and Montiel Olea et al. (2025)), although they have a different goal. In the setup of this paper, (15) models the distribution of multiple correlated treatment effects. I also assume that the researcher reports a linear average of treatment effects on the observed outcomes $\widehat{\tau} = \boldsymbol{w}'\widehat{\boldsymbol{\beta}}$ as in (1) with $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ in (2). Focusing on the class of weights $\mathcal{W}_{\text{cvx}}$ that are nonnegative and sum to one is to avoid the problem of having interpretation issues as discussed in Section 2.3.1.

The assumption below ensures that the covariance matrix $\boldsymbol{\Sigma}$ in (15) is positive definite.

Assumption 4.1. *$\boldsymbol{\Sigma}$ has real-valued eigenvalues bounded below by $\lambda_{\text{lb}} > 0$ and above by $\lambda_{\text{ub}} < \infty$ where $\lambda_{\text{ub}} > \lambda_{\text{lb}}$.*

To measure the quality of $\boldsymbol{\beta}$ in capturing θ, define

$$\boldsymbol{b} \equiv \boldsymbol{\beta} - \theta \mathbf{1}_q \in \mathbb{R}^q. \tag{16}$$

In terms of the running example (Example 2.1), $\boldsymbol{b}$ can be interpreted as how well the treatment effects of various assets measures the treatment effect of wealth.

The squared loss function is used to model the performance of the estimator $\widehat{\tau}$ relative to θ as in Section 2.3.2. Since $\widehat{\tau}$ is a linear combination of $\widehat{\boldsymbol{\beta}}$ in (1), I can write the loss function as follows

$$L(\boldsymbol{w}, \widehat{\boldsymbol{\beta}}, \theta) \equiv (\widehat{\tau} - \theta)^2 = (\boldsymbol{w}'\widehat{\boldsymbol{\beta}} - \theta)^2. \tag{17}$$

The loss function is written as a function of $\boldsymbol{w}$ since the researcher's action is to choose the weights $\boldsymbol{w}$ in forming the estimator $\widehat{\tau}$.

The risk function of the estimator $\widehat{\tau}$ is the expectation of the loss function (17) with respect to the distribution of $\widehat{\boldsymbol{\beta}}$ in (15). For any $\boldsymbol{b} \in \mathbb{R}^q$ and $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, the risk function is defined as

$$R(\boldsymbol{w}, \boldsymbol{b}) \equiv \mathbb{E}[L(\boldsymbol{w}, \widehat{\boldsymbol{\beta}}, \theta)] = \mathbb{E}[(\boldsymbol{w}'\widehat{\boldsymbol{\beta}} - \boldsymbol{w}'\boldsymbol{\beta})^2 + (\boldsymbol{w}'\boldsymbol{\beta} - \theta)^2] = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + (\boldsymbol{w}'\boldsymbol{b})^2, \tag{18}$$

where the last equality follows from (15), the definition of $\boldsymbol{b}$ in (16), and that $\boldsymbol{w}$ sums to one.

4.2 Parameter space and maximum risk

In this subsection, I introduce various options to model and motivate the parameter space of $\boldsymbol{b}$. In terms of the running example on wealth, this is needed because the quality of the treatment effects on different assets in capturing the effect on wealth is unknown although the variance in $\boldsymbol{\Sigma}$ in (18) is known. Researchers may also have prior information on the relative quality of the treatment effects on various assets.

Let $\mathcal{S}(B)$ be a nonempty parameter space on $\boldsymbol{b}$ and $B \geq 0$ be a maximum level of misspecification chosen by the researcher. In the following, I describe three examples to show that $\mathcal{S}(B)$ can be flexible to reflect researcher's restriction or prior information on treatment effects. More details can be found in Appendix B.4.

Example 4.2 (Bounds on the overall magnitude)**.** I start with the restriction in which the researcher places an upper bound on the vector $\boldsymbol{b}$ in $\mathcal{S}(B)$ using the ℓ_p-norm of $\boldsymbol{b}$ for $p \geq 1$ (such as in Armstrong and Kolesár (2021b)). In this case, the parameter space can be written as

$$\mathcal{S}(B) = \{\boldsymbol{b} \in \mathbb{R}^q : \|\boldsymbol{b}\|_p \leq B\}. \tag{19}$$

In terms of the running example on wealth, the above places an upper bound B on the quality of treatment effect on different assets. If $B = 0$, this means $\beta_j = \theta$ for each $j = 1, \ldots, q$ (e.g., the treatment effect on each asset equals the treatment effect on wealth). If $B > 0$, choosing $p = \infty$ can be interpreted as setting the same bound on each b_j (i.e.,

b_j for each asset is bounded in $[-B, B]$). Choosing $p \in [1, \infty)$ allows for some treatment effects to be more important or having a better quality in capturing θ than others. △

Example 4.3 (Shape restrictions)**.** Researchers can impose shape constraints when they believe the treatment effects on some outcomes are more important than others.

In the entrepreneurial spirit index from Example 2.4, Bruhn et al. (2018) mentions that two of the outcomes might not be directly affected by their consulting program, but are driven by an improvement in business. Thus, they construct another entrepreneurial spirit index that excludes these two outcomes.

In this empirical example, shape restrictions that make these two outcomes less important can be an informative alternative to dropping them directly. Let $\mathcal{J}_0$ be the set that indices the two outcomes they exclude, and $\mathcal{J}_1$ be the set that indices the remaining outcomes. A shape constraint that can be reasonable in this setting is $\kappa |b_j| \leq |b_\ell|$ for $j \in \mathcal{J}_1$ and $\ell \in \mathcal{J}_0$ where $\kappa \in \mathbb{R}$ controls the relative importance between outcomes in $\mathcal{J}_0$ and $\mathcal{J}_1$. Setting $\kappa > 1$ can be interpreted as the outcomes in $\mathcal{J}_1$ being more "important."

Together with a bound on $\boldsymbol{b}$ as in Example 4.3, the above can be described as

$$\mathcal{S}(B) = \{\boldsymbol{b} \in \mathbb{R}^q : \|\boldsymbol{b}\|_p \leq B, \boldsymbol{Q}|\boldsymbol{b}| \leq \mathbf{0}_l\},$$

where $\boldsymbol{Q} \in \mathbb{R}^{l \times q}$ and $|\boldsymbol{b}|$ refers to taking absolute value on $\boldsymbol{b}$ componentwise. Other types of constraints can be used, such as $\boldsymbol{Q}\boldsymbol{b} \leq \mathbf{0}_l$. △

Example 4.4 (Communication and subjective weights)**.** Another concern that researchers may have is related to the communication of the weighted average of treatment effects $\widehat{\tau}$ to the reader. In terms of the running example (Example 2.1), readers may think the treatment effect of the cash transfer program on wealth θ is the treatment effect on the weighted average of assets, but have subjective views on what the weights should be for each asset. The researcher's decision has to take this communication issue into account.

This communication problem can be described through a parameter space on $\boldsymbol{b}$. For exposition purposes, consider the case that $\theta = \boldsymbol{\gamma}'\boldsymbol{\beta}$ for some known $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ and $\|\boldsymbol{b}\|_p \leq B$ as in Example 4.2. In this case, the vector of misspecification can be written as $\boldsymbol{b} = \boldsymbol{\beta} - (\boldsymbol{\gamma}'\boldsymbol{\beta})\mathbf{1}_q$. In Appendix B.4, I show that this can be studied via a shape constraint in the parameter space. In addition to a known $\boldsymbol{\gamma}$, I show how to allow for an unknown $\boldsymbol{\gamma}$ or ambiguity around a given $\boldsymbol{\gamma}$. △

After defining the parameter space $\mathcal{S}(B)$, the maximum risk can be evaluated as the

maximum of (18) over $\boldsymbol{b} \in \mathcal{S}(B)$, i.e.,

$$R_{\max}(B, \boldsymbol{w}) \equiv \max_{\boldsymbol{b} \in \mathcal{S}(B)} R(\boldsymbol{w}, \boldsymbol{b}) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + \max_{\boldsymbol{b} \in \mathcal{S}(B)} (\boldsymbol{w}'\boldsymbol{b})^2, \tag{20}$$

where the second equality holds because the variance term is independent of $\boldsymbol{b}$. By Assumption 4.1, the variance component $V(\boldsymbol{w}) \equiv \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$ is strictly convex in $\boldsymbol{w}$ because $\boldsymbol{\Sigma}$ is positive definite. Since $\mathcal{S}(B)$ is nonempty, the maximum bias $\overline{M}(B, \boldsymbol{w}) \equiv \max_{\boldsymbol{b} \in \mathcal{S}(B)} (\boldsymbol{w}'\boldsymbol{b})^2$ is convex in $\boldsymbol{w}$ for a given $B \geq 0$ because it is a maximum of convex functions (Boyd and Vandenberghe, 2004, Page 81).

4.3 Estimating weights using the minimax approach

This subsection takes a minimax approach to estimate the weights by minimizing the worst-case risk over the parameter space. In terms of the running example on wealth, the minimax approach chooses the weight on each asset by minimizing the worst possible MSE over all possible quality of treatment effects on the assets specified in $\boldsymbol{b} \in \mathcal{S}(B)$.

The minimax problem is

$$R^\star(B) \equiv \min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} R_{\max}(B, \boldsymbol{w}) = \min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} \left[V(\boldsymbol{w}) + \overline{M}(B, \boldsymbol{w}) \right], \tag{21}$$

where the maximum risk is from (20). $R^\star(B)$ is referred to as the minimax risk. Suppose $\mathcal{S}(B)$ is bounded and nonempty. Following the discussion in Section 4.2, $V(\boldsymbol{w})$ is strictly convex in $\boldsymbol{w}$ and $\overline{M}(B, \boldsymbol{w})$ is convex in $\boldsymbol{w}$. Hence, the objective of (21) is strictly convex. It follows that (21) has a unique optimal solution. I denote the optimal solution to the minimax problem (21) as follows:

$$\boldsymbol{w}^\star(B) = \arg\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} R_{\max}(B, \boldsymbol{w}). \tag{22}$$

The expression for $\overline{M}(B, \boldsymbol{w})$ depends on the parameter space. Computing the optimal solution can be done by convex optimization and the details are in Appendix C.1.

In the following, I discuss three remarks on some special cases of the minimax problem (21) and an example with two outcomes to build more intuition.

Remark 4.5 (Variance minimization when $B = 0$)**.** Consider $B = 0$ so that $\beta_j = \theta$ for each $j = 1, \ldots, q$. This leads to $\overline{M}(B, \boldsymbol{w}) = 0$ and the objective of the minimax problem (21) reduces to $V(\boldsymbol{w})$. Hence, the optimal weight is the variance-minimizing weight. ■

Remark 4.6 (Connection with matrix regularization)**.** Assume the parameter space is as described in Example 4.2 that restricts $\boldsymbol{b}$ with the ℓ_2-norm, i.e., $\mathcal{S}(B) = \{\boldsymbol{b} \in \mathbb{R}^q : \|\boldsymbol{b}\|_2 \leq B\}$. Then, $\overline{M}(B, \boldsymbol{w}) = B^2\|\boldsymbol{w}\|_2^2$. The objective of (21) becomes $R_{\max}(B, \boldsymbol{w}) = \boldsymbol{w}'(\boldsymbol{\Sigma} + B^2\boldsymbol{I}_q)\boldsymbol{w}$, where $\boldsymbol{I}_q$ is an identity matrix of size $q \times q$. Thus, the optimal weight for the minimax problem can be viewed as minimizing the quadratic form on the regularized matrix $\boldsymbol{\Sigma} + B^2\boldsymbol{I}_q$ subject to $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. ■

Remark 4.7 (Connection with standardized weighting)**.** Suppose $B \longrightarrow \infty$ and the parameter space is as described in Example 4.2 that restricts $\boldsymbol{b}$ with the ℓ_p-norm and $p \in (1, \infty)$. Then, the solution is (see Appendix Corollary B.11):

$$\lim_{B\to\infty} \boldsymbol{w}^\star(B) = \frac{1}{q}\mathbf{1}_q.$$

Thus, the above is similar to SA which puts equal weights on each treatment effect (although SA divides outcomes by the standard deviation from the control sample as in Section 3.3). This provides a statistical decision-theoretic justification of using equal weights as the optimal solution when $B \longrightarrow \infty$. ■

Example 4.8. Suppose the researcher observes two assets as in Example 2.1, so that $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2)'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. Assume $\rho \in (-1, 1)$ and $\sigma_2 > 0$, so (15) can be written as

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_2 \\ \rho\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

I assume $\sigma_2 \neq 1$ so the two treatment effects do not have the same variances. Suppose the Euclidean norm is used to bound $\boldsymbol{b} \equiv (b_1, b_2)'$ in $\mathcal{S}(B)$ as in Example 4.2. Since $w_1 + w_2 = 1$, I focus on characterizing the optimal solution for the first weight. The optimal solution $w_1^\star(B)$ for the minimax problem with two outcomes is given as follows:

$$w_1^\star(B) = \begin{cases} 0 & \text{if } \sigma_2^2 - \rho\sigma_2 \leq -B^2, \\ 1 & \text{if } 1 - \rho\sigma_2 \leq -B^2, \\ \frac{\sigma_2^2 - \rho\sigma_2 + B^2}{1 + \sigma_2^2 - 2\rho\sigma_2 + 2B^2} & \text{otherwise.} \end{cases} \tag{23}$$

The optimal weights in (23) depend on B, σ_2^2, and ρ. First, consider $B = 0$. The optimal weight focuses on minimizing the variance of the estimator as discussed in Remark 4.5. If the second treatment effect is more precise (i.e., $\sigma_2 < \rho < 1$ here), then it is optimal to set $w_1^\star(0) = 0$. This is reasonable because the treatment effect on both assets are unbiased relative to the treatment effect on wealth when $B = 0$.

As B increases, (23) suggests that it may not be optimal to put all the weights on the second treatment effect even if $\sigma_2 < \rho$. This is related to the correlation of the two treatment effects and also because the second treatment effect might have a poorer quality in capturing the treatment effect on wealth when $B > 0$.

As $B \longrightarrow \infty$, the optimal weights become $w_1^\star(B) = \frac{1}{2}$ to reflect that researchers want to be agnostic and put equal weights on both treatment effects as in Remark 4.7. △

4.4 Estimating weights using the adaptive approach

Solving the minimax problem in Section 4.3 requires the researcher to choose the value $B \geq 0$. However, researchers may not want to commit to a specific value of B, but are willing to specify a set $\mathcal{B} \subseteq \mathbb{R}$ such that $B \in \mathcal{B}$. In the context of the running example on wealth, researchers may want to choose the weights on assets that is optimal over different upper bounds on the quality of different treatment effects. In particular, researcher might want to find the weights that is a "middle ground" between variance minimization ($B = 0$) and a large B. To avoid computing the weights that are optimal for a specific value of B, I consider a recent adaptive framework by Armstrong et al. (2024).

4.4.1 Definitions

I briefly review the relevant definitions of the methodology proposed by Armstrong et al. (2024). For a given $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ and $B \in \mathcal{B}$, the adaptive regret is defined as the following ratio

$$A(B, \boldsymbol{w}) \equiv \frac{R_{\max}(B, \boldsymbol{w})}{R^\star(B)} = \frac{R_{\max}(B, \boldsymbol{w})}{R_{\max}(B, \boldsymbol{w}^\star(B))}, \tag{24}$$

where the maximum risk $R_{\max}(B, \boldsymbol{w})$ and the minimax risk $R^\star(B)$ have been defined in (20) and (21) respectively. For a given $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, the adaptive regret can be interpreted as how much worse $\boldsymbol{w}$ performs relative to an oracle that knows the optimal weight is $\boldsymbol{w}^\star(B)$. In particular, $100[A(B, \boldsymbol{w}) - 1]\%$ equals the percentage increase in the worst-case MSE.

They define

$$A_{\max}(\mathcal{B}, \boldsymbol{w}) \equiv \sup_{B \in \mathcal{B}} A(B, \boldsymbol{w})$$

as the worst-case adaptive regret, and

$$A^\star(\mathcal{B}) \equiv \inf_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} \sup_{B \in \mathcal{B}} A(B, \boldsymbol{w}). \tag{25}$$

An estimator $\boldsymbol{w}_A$ is optimally adaptive if

$$A^\star(\mathcal{B}) = A_{\max}(\mathcal{B}, \boldsymbol{w}_A). \tag{26}$$

4.4.2 Solving for the optimally adaptive weights

Solving the optimally adaptive weights using (25) directly involves two steps. First, the inner problem finds the worst-case adaptive regret over $B \in \mathcal{B}$ for each $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. Second, the outer problem chooses the weight that minimizes the worst-case adaptive regret.

In this section, I derive new properties of the adaptive regret for the setup in Section 4.1. Before proceeding, I discuss some differences between my model and Armstrong et al. (2024). This is because the properties developed below are based on a different setup from Armstrong et al. (2024). First, I assumed that all treatment effects $\widehat{\boldsymbol{\beta}}$ in (15) can be misspecified relative to θ, whereas Armstrong et al. (2024) assumes that an unbiased estimator is available. Allowing for all components in $\widehat{\boldsymbol{\beta}}$ to be potentially misspecified relative to θ is important under the context of aggregating outcomes. For instance, in running example 2.1, it is unlikely that the treatment effect on each asset captures the treatment effect on wealth perfectly. Second, I focus on the linear estimator (1) where the weights $\boldsymbol{w}$ are restricted to the convex class of weights as in (2). Armstrong et al. (2024) does not restrict their estimators to this class, and their solution approach is different. The motivation for my restriction is to prevent the interpretation issues on the resulting aggregated treatment effects as explained in Section 2.3.

First, I show a useful property on the adaptive regret under the assumption that is satisfied by the parameter space introduced in Section 4.2.

Assumption 4.9. *For any $B \geq 0$, $\overline{M}(B, \boldsymbol{w}) = B^2 m(\boldsymbol{w})$ for some function $m : \mathbb{R}^q \longrightarrow \mathbb{R}$ where $m(\boldsymbol{w}) \geq 0$ and is convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$.*

For instance, in Example 4.2, $\overline{M}(B, \boldsymbol{w}) = B^2 \|\boldsymbol{w}\|_{p^\star}^2$ where the $\ell_{p^\star}$-norm is the dual norm for the ℓ_p-norm such that $\frac{1}{p} + \frac{1}{p^\star} = 1$ (see, for instance, Boyd and Vandenberghe (2004, Chapter A.1.6)). In this case, $m(\boldsymbol{w}) = \|\boldsymbol{w}\|_{p^\star}^2$. I show the other cases in Appendix B.4. The following proposition shows a useful property on the adaptive regret.

Proposition 4.10. *Let Assumptions 2.2, 4.1, and 4.9 hold. Consider the minimax and adaptive problems defined in (21) and (24), respectively. Let $\mathcal{B} = [\underline{B}, \overline{B}]$ where $\overline{B} \geq \underline{B} \geq 0$ and $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$.*

(a) $A(B, \boldsymbol{w})$ has one of the following shapes:

(i) $A(B, \boldsymbol{w})$ is monotone in $B \in \mathcal{B}$.

(ii) There exists $B_0 \in (\underline{B}, \overline{B})$ *such that* $A(B, \boldsymbol{w})$ *is nonincreasing in* B *for any* $B \in [\underline{B}, B_0)$ *and nondecreasing in* B *for any* $B \in [B_0, \overline{B}]$.

(b) The worst-case adaptive regret is given by

$$A_{max}(\mathcal{B}, \boldsymbol{w}) = \sup_{B \in \mathcal{B}} A(B, \boldsymbol{w}) = \max\{A(\underline{B}, \boldsymbol{w}), A(\overline{B}, \boldsymbol{w})\}.$$

If $\overline{B} = \infty$, *then* $A(\overline{B}, \boldsymbol{w})$ *refers to* $\lim_{B \to \infty} A(B, \boldsymbol{w})$.

In the above, Proposition 4.10(a) shows the shape of the adaptive regret over $B \in \mathcal{B}$ for each $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. It implies that the supremum over $B \in \mathcal{B}$ is achieved at the endpoints under Assumption 4.9. This simplifies the computation of the inner problem in (25).

Suppose the parameter space is as described in Example 4.2 that restricts $\boldsymbol{b}$ using the ℓ_p-norm with $p \in (1, \infty)$. Then, the optimally adaptive weight can be interpreted as a middle ground between the regret against the variance-minimizing weight (for $\underline{B} = 0$) and equal weighting (for $\overline{B} = \infty$ in Remark 4.7).

The following proposition is another useful property on the optimally adaptive weight $\boldsymbol{w}_A$. It shows that at the optimum, it cannot be the case where $A(\underline{B}, \boldsymbol{w}_A) \neq A(\overline{B}, \boldsymbol{w}_A)$ when $A(B, \boldsymbol{w})$ is strictly convex and continuous in $\boldsymbol{w}$.

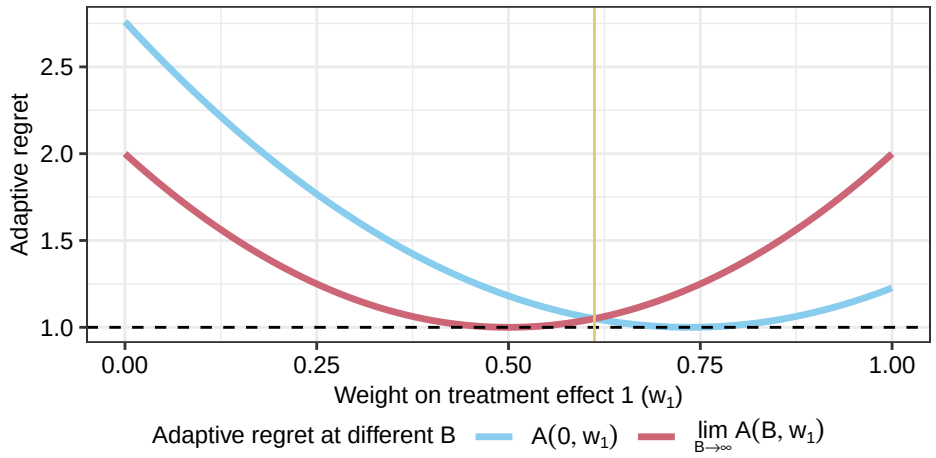
Proposition 4.11. *Consider the same assumptions and definitions as in Proposition 4.10 with* $\mathcal{B} = [\underline{B}, \overline{B}]$ *and* $\underline{B} < \overline{B}$. *Suppose that* $A(\underline{B}, \boldsymbol{w})$ *and* $A(\overline{B}, \boldsymbol{w})$ *are strictly convex and continuous in* $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. *Let* $\boldsymbol{w}_A \in \mathcal{W}_{\text{cvx}}$ *be the optimally adaptive estimator as defined in (26). Then,*

$$A(\underline{B}, \boldsymbol{w}_A) = A(\overline{B}, \boldsymbol{w}_A).$$

The assumptions on strictly convexity and continuity of the adaptive regret over $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ are satisfied by the functions and parameter spaces discussed in the previous subsections. Strict convexity is satisfied when B is finite. This is because $R^\star(B) > 0$ is finite, and from the definition in (24) that $A(B, \boldsymbol{w}) = \frac{R_{\max}(B, \boldsymbol{w})}{R^\star(B)}$, where $R_{\max}(B, \boldsymbol{w})$ is a sum of strictly convex and convex functions in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ as defined in (21). In addition, $R^\star(B)$ is independent of $\boldsymbol{w}$. Therefore, strictly convexity in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ follows. The adaptive regret $A(B, \boldsymbol{w})$ is also continuous in $\boldsymbol{w}$ for a finite B and compact parameter spaces $\mathcal{S}(B)$. To see this, recall again that $R_{\max}(B, \boldsymbol{w}) = V(\boldsymbol{w}) + \overline{M}(B, \boldsymbol{w})$ as defined in (21). $V(\boldsymbol{w})$ is continuous in $\boldsymbol{w}$. $\overline{M}(B, \boldsymbol{w}) = \max_{\boldsymbol{b} \in \mathcal{S}(B)} (\boldsymbol{w}'\boldsymbol{b})^2$ is continuous due to the Berge maximum theorem (see, for instance, Aliprantis and Border (2006, Theorem 17.31)).

Remark 4.12. In the case where $A(B, \boldsymbol{w})$ is only convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ instead of strictly

Figure 4: Illustration of $A(0, w_1)$ and $\lim_{B\to\infty} A(B, w_1)$ for the two-outcome example.



Notes: The above shows the adaptive regret for the two-outcome example in Example 4.8. The data generating process uses $\sigma_1^2 = 1$, $\sigma_2^2 = 2.25$, and $\rho = 0.2$. The yellow vertical line shows the point chosen by the adaptive approach.

convex, the worst-case adaptive regret is not necessarily strictly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. Nevertheless, strict convexity can be restored by adding a small penalty term to the function, i.e., replace $A^\star([0,\infty], \boldsymbol{w})$ by $A^\star_\kappa([0,\infty], \boldsymbol{w}) \equiv A^\star([0,\infty], \boldsymbol{w}) + \kappa\|\boldsymbol{w}\|_2^2$ for a small $\kappa > 0$. This is because $A^\star([0,\infty], \boldsymbol{w})$ is convex in $\boldsymbol{w}$ and $\kappa\|\boldsymbol{w}\|_2^2$ is strictly convex in $\boldsymbol{w}$ for $\kappa > 0$. This approach of adding a strictly convex penalty term is also used by Gafarov (2025) for moment inequality models. ■

Computing the optimally adaptive weight can be done by convex optimization. I show the details and explain how computing the optimal weight can be written as one optimization problem in Appendix C.2.

Example 4.8 (continued). Assume that $\mathcal{B} = [0,\infty]$ and $\rho\sigma_2 < 1$. The optimally adaptive weight on $\widehat{\beta}_1$ is given by

$$w_1^\star(\mu_1) \equiv \frac{\mu_1[R^\star(0)^{-1}(\sigma_2^2 - \rho\sigma_2) - 2] + 2}{\mu_1[R^\star(0)^{-1}(\sigma_2^2 - 2\rho\sigma_2 + 1) - 4] + 4}, \tag{27}$$

where $\mu_1 \in (0,1)$ satisfies the optimality condition given in Appendix B.5.3 and the minimax risk $R^\star(0)$ is obtained by evaluating (B.40) at $B = 0$. The proof of (27) is in Appendix B.5.3. I discuss the observations and intuitions below.

For exposition purposes, I present a numerical example with $\sigma_2^2 = 2.25$ and $\rho = 0.2$ in Figure 4. The blue curve is the adaptive regret for $B = 0$. It is the regret relative to the variance-minimizing weight (see Remark 4.5). The red curve is the adaptive regret

for $B = \infty$. It is the regret relative to equal weighting that minimizes bias (see Remark 4.7). The optimally adaptive weight is achieved at the point where both curves intersect, which is consistent with Proposition 4.11. This is also reflected in (27) that the optimally adaptive weight is like a weighted average related to the Lagrange multiplier μ_1.

Here are some intuitions that the optimally adaptive weight is not on the boundary in this two-outcome example. The optimally adaptive weight minimizes the higher of the two "costs" (bias squared and variance). If the optimally adaptive weight is on the boundary, it has to be more costly to shift the weights to allow for positive weights on both treatment effects. Assume without loss of generality that the solution is $w_1 = 1$. But when $B = \infty$, the potential bias can be enormous when $w_1 = 1$ and the cost reduces when the weights move to $w_1 < 1$. Hence, for the optimally adaptive weight to be boundary, it has to be that moving to the interior causes a higher cost in term of variance. This can only be true if the variance is minimized when $w_1 = 1$. But if $w_1 = 1$ minimizes variance, one can shift a small amount of weight to the second outcome that increases variance by a bit, but has a larger reduction in bias squared, thereby giving a smaller adaptive regret. It follows that $w_1 = 1$ is not optimal. A similar analysis can be applied for $w_1 = 0$. Therefore, boundary weight is not optimal in this two-outcome example with $\mathcal{B} = [0, \infty]$. △

4.5 Inference

This section describes how to perform valid inference based on the minimax or adaptive procedures. When $B > 0$, standard Wald confidence intervals for θ will not be valid because the aggregated treatment effect is biased relative to θ. As a result, I construct fixed-length confidence intervals (FLCI) that take the bias into account (Donoho, 1994; Armstrong and Kolesár, 2018, 2021a,b).

4.5.1 Minimax approach

Let $\boldsymbol{w}^\star(B)$ be the solution to the minimax problem in (22). Then,

$$\frac{\boldsymbol{w}^\star(B)'\widehat{\boldsymbol{\beta}} - \theta}{\sqrt{\boldsymbol{w}^\star(B)'\boldsymbol{\Sigma}\boldsymbol{w}^\star(B)}} \sim \mathcal{N}\left(\frac{\boldsymbol{w}^\star(B)'\boldsymbol{b}}{\sqrt{\boldsymbol{w}^\star(B)'\boldsymbol{\Sigma}\boldsymbol{w}^\star(B)}}, 1\right). \tag{28}$$

The above holds due to the distributional assumption in (15).

A $100(1-\alpha)\%$ fixed-length confidence interval centered around $\boldsymbol{w}^\star(B)'\widehat{\boldsymbol{\beta}}$ can be

formed as

$$\mathcal{I} \equiv \left[\boldsymbol{w}^\star(B)'\widehat{\boldsymbol{\beta}} \pm c_\alpha \sqrt{\boldsymbol{w}^\star(B)'\boldsymbol{\Sigma}\boldsymbol{w}^\star(B)} \right], \tag{29}$$

where the critical value $c_\alpha \in \mathbb{R}$ is chosen to be the smallest value such that it controls size, i.e., it is the smallest c_α such that the following holds

$$\max_{\boldsymbol{b} \in \mathcal{S}(B)} \mathbb{P}\left[\left| \frac{\boldsymbol{w}^\star(B)'\widehat{\boldsymbol{\beta}} - \theta}{\sqrt{\boldsymbol{w}^\star(B)'\boldsymbol{\Sigma}\boldsymbol{w}^\star(B)}} \right| > c_\alpha \right] \leq \alpha. \tag{30}$$

4.5.2 Adaptive approach

Similar to the discussion in Section 4.5.1, one can construct a $100(1-\alpha)\%$ confidence interval for the adaptive weights chosen in Section 4.4 centered at the adaptive estimator.

Suppose the researcher adapts over $\mathcal{B} = [\underline{B}, \overline{B}]$. Similar to the recommendation in Armstrong et al. (2024), one choice is to construct FLCIs as in (29) but with $\boldsymbol{w}^\star(B)$ replaced by the optimally adaptive weight and $\mathcal{S}(B)$ replaced with $\mathcal{S}(\underline{B})$ and $\mathcal{S}(\overline{B})$ in order to summarize the range of critical values needed to guarantee coverage under different assumptions.

4.5.3 Asymptotic validity

In the discussion so far, I assumed that $\boldsymbol{\Sigma}$ is known. In Appendix B.1, I show the details for the asymptotic validity of FLCI using a consistent estimator for $\boldsymbol{\Sigma}$. The FLCI constructs a confidence interval around θ. I provide additional procedures on inference around $\boldsymbol{w}^\star(B)'\boldsymbol{\beta}$ in Appendix B.7. This can be another useful approach in assessing the uncertainty of the estimated weight.

4.6 Implementation and summary

This subsection summarizes the statistical decision-theoretic procedures.

Step 1. Estimates the treatment effects $\widehat{\beta}_n$ and asymptotic variance $\widehat{\boldsymbol{\Sigma}}_n$.

Step 2. Specify the parameter space (see Section 4.2).

Step 3. Estimate the weights using the minimax or adaptive approach.

(a) For the minimax approach, choose $B \geq 0$ and solve (22).

(b) For the adaptive approach, choose $\mathcal{B} = [\underline{B}, \overline{B}]$ and solve (25).

Step 4. Conduct inference using FLCI (see Section 4.5). See Appendix B.7 for additional procedures for inference.

5 Empirical applications

In this section, I illustrate the methodology proposed in this paper by revisiting two empirical examples.

5.1 Campante and Do (2014)

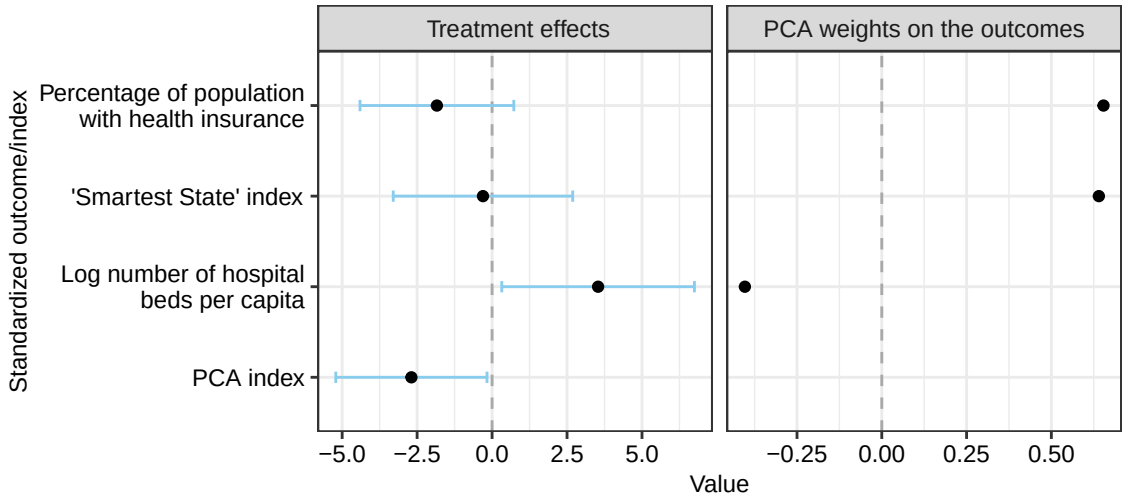
In the analysis of how state capital being isolated from population affects public good provision, Campante and Do (2014, "CD" in the following) creates a PCA index of public good provision. CD finds that state isolation has a negative effect on public good expenditure. But CD points out that expenditure is not related to the effectiveness of how resources are used. Hence, they create an index of public good provision using PCA. As introduced in Example 2.3, the index aggregates the three outcomes "smartest state" index (an index that measures educational outcomes), the percentage with health insurance, and the log number of hospital beds using PCA.

Figure 5 presents the regression results on each standardized outcome and the corresponding PCA weights. The regressions control for log area, log income, log population, etc. The outcome of the log number of hospital beds received a negative weight. The figure reports the 90% confidence interval that treats the weights as "fixed" as originally implemented without accounting for the statistical uncertainty due to the estimated PCA weights.

Next, I study the effect with my decision-theoretic approach. The results are presented in Figure 6. Panel (a) of the figure shows the results from the minimax approach, and panel (b) shows the results from the adaptive approach. As discussed in Section 4.3, $B = 0$ corresponds to the variance-minimizing weight. Setting large values of B leads to the weights converge to having equal weights.

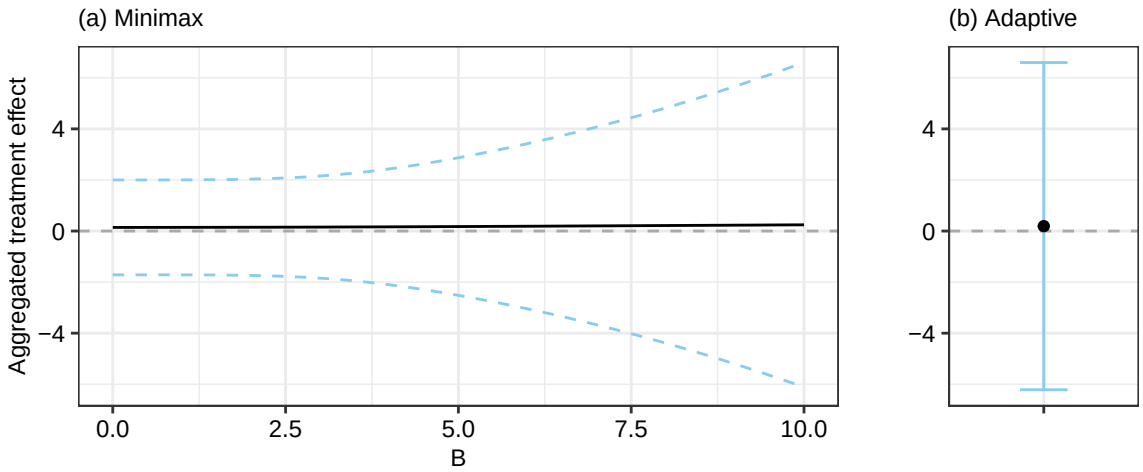
The aggregated treatment effects obtained from the minimax or adaptive approach in Figure 6 show that the effect is positive and close to 0. The effect is also not significant at the 10% level. This is at contrast with the treatment effect on the PCA index in Figure 5 that gives a negative effect. The difference in the sign and significance of the aggregated treatment effect is related to the decision-theoretic approach that requires the weights to be nonnegative. The PCA approach puts a negative weight on the outcome "log number

Figure 5: Teatment effects and PCA weights for the CD application.



Notes: Panel (a) shows the treatment effects on the individual outcomes and the PCA index, as well as the 90% confidence intervals. Panel (b) shows the PCA weights on each outcome.

Figure 6: Results for the statistical decision approach (without shape restrictions) for the CD application.

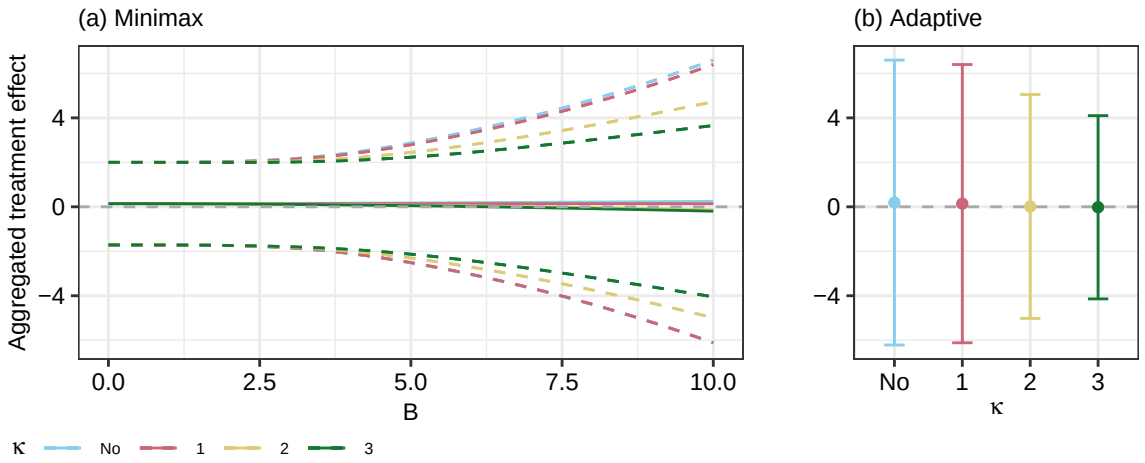


Notes: In panel (a), the solid line shows the treatment effect using the minimax approach for different values of B, and the dashed lines show the 90% FLCI. In panel (b), the dot represents the treatment effect when adapting over $[0, 10]$, and the bar represents the 90% FLCI.

of hospital beds per capita," which has a positive treatment effect.

Since one might expect state isolation has a negative impact on public goods, the effect of state isolation on the log number of hospital beds per capita might be biased, so that there is a positive effect. To investigate what assumptions could support the

Figure 7: Results for the statistical decision approach (with shape restrictions) for the CD application.



Notes: Each color corresponds to a specific value of κ on the shape constraint. See the notes under Figure 6 for the meaning of the lines.

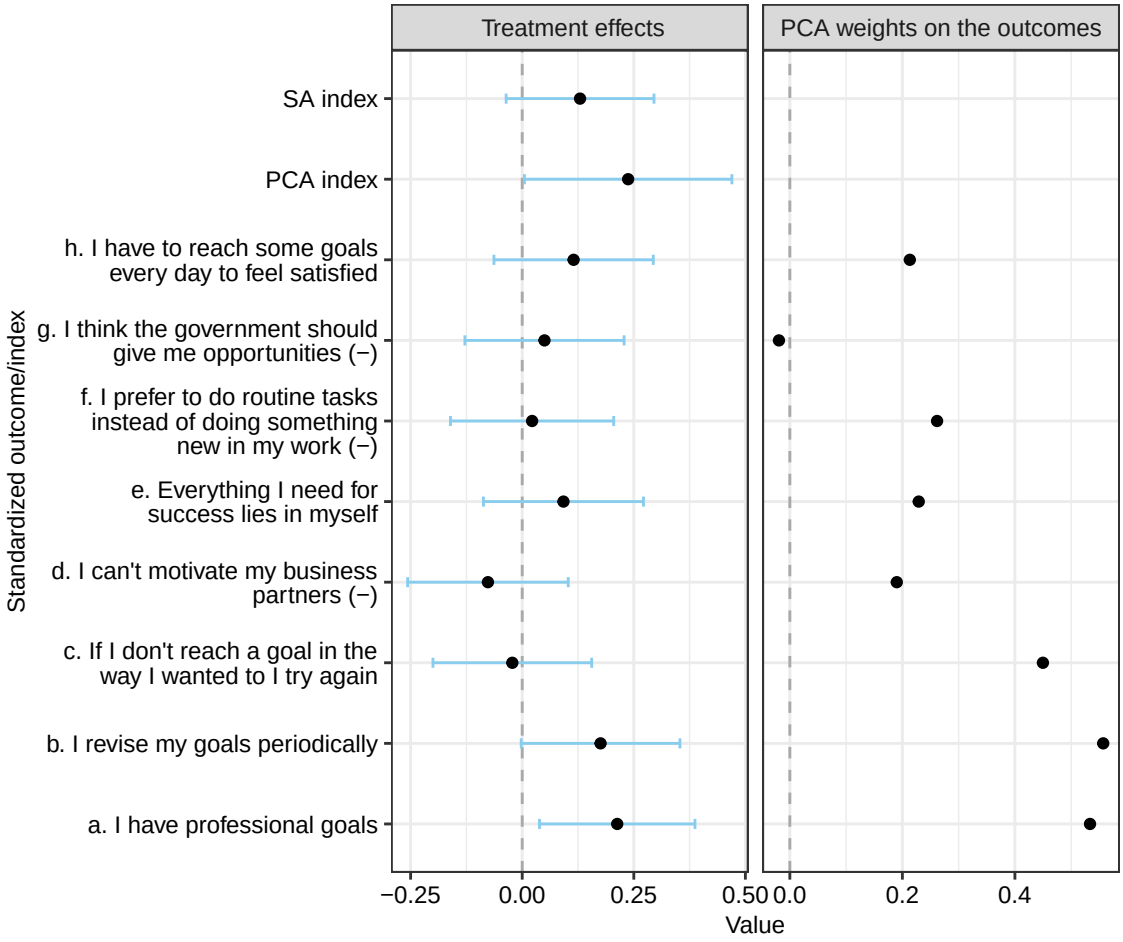
authors' finding that there is a negative effect of capital isolation on the PCA index of public good provision, I consider imposing shape restrictions that constrain the relative importance of the outcomes. Let $\boldsymbol{b} \equiv (b_1, b_2, b_3) \in \mathbb{R}^3$ be as defined in (16), where b_1 is the component on the outcome "log number of hospital beds per capita." Then, I explore whether a negative impact can be found by making the treatment effect on hospital beds "less important." More specifically, I impose a shape restriction as $\kappa|b_1| \leq |b_j|$ for $j \in \{2, 3\}$ where κ can be interpreted as varying the relative importance of hospital beds against the two other outcomes. A larger value of κ means the treatment effect on hospital beds is less important than the treatment effect on other outcomes.

Figure 7 reports the results with shape restrictions under different values of κ. As κ increases, the emphasis on the number of hospital beds in the treatment effect of the aggregated outcome reduces. The figure suggests that to support the claim that there is a negative impact of state isolation on public good provision, one has to allow for $B > 0$ and put less emphasis on the number of hospital beds. Despite a negative effect can be found when B is large for $\kappa = 3$, they are not significant at the 10% level.

5.2 Bruhn et al. (2018)

Bruhn et al. (2018, "BKS" in the following) studies the impact of management consulting services on small and medium enterprises. To understand whether firm growth is

Figure 8: Treatment effects and PCA weights for the BKS application.



Notes: See notes under Figure 5.

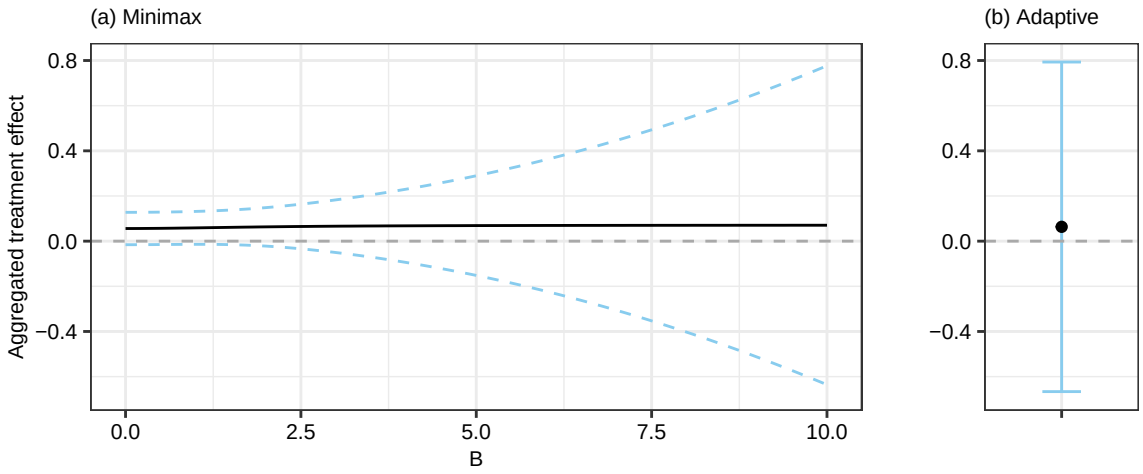
affected by managerial skills, BKS ran a randomized controlled trial in Mexico with 432 small and medium enterprises. 150 out of the 432 enterprises were randomly chosen to receive subsidized consulting services that lasted for one year. The consultants were asked to examine the problems that prevented firm growth, to suggest solutions, and to help implement them. BKS studies how such consulting programs affect the enterprises' productivity, return on assets, and "entrepreneurial spirit."

Entrepreneurial spirit is an index that aggregates outcomes on entrepreneurial attitudes, confidence, and goal setting. BKS constructs entrepreneurial spirit indices using two sets of outcomes as shown below.

Set 1. This includes all eight variables shown in Figure 8.

Set 2. This excludes "I can't motivate my business partners" and "everything I need for success lies in myself."

Figure 9: Results for the statistical decision approach (without shape restrictions) for the BKS application.



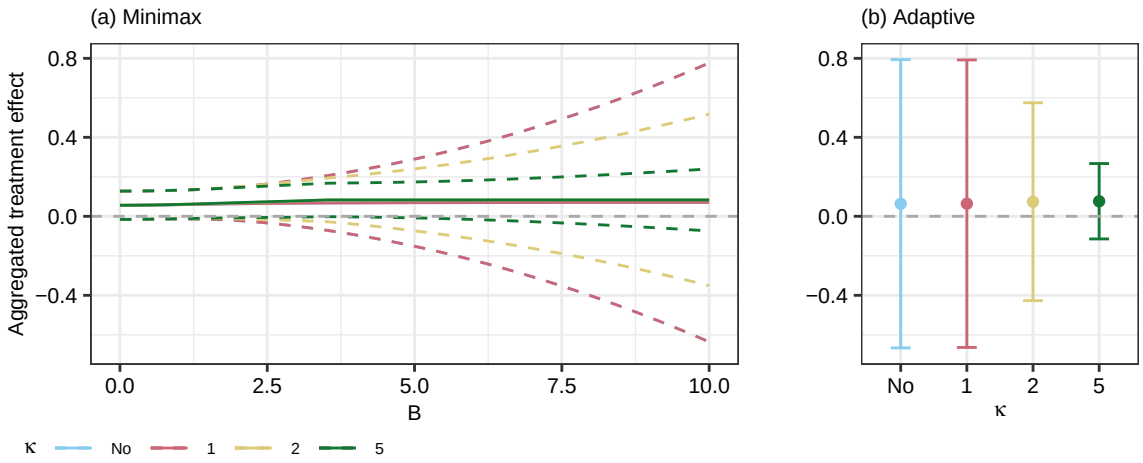
Notes: See the notes under Figure 6.

BKS also considers **Set 2** in addition to **Set 1** because they are concerned that improvement in some outcomes are due to improvement in business instead of the consulting program. They pointed out that the two outcomes excluded in **Set 2** are particularly subject to this interpretation. Thus, they examine the impact of the consulting program on the entrepreneurial index defined via these two sets of outcomes. For each set of outcomes, they construct one index via PCA and another index using SA. I focus on the regression specification without controlling for baseline outcomes because BKS found a significant effect using the PCA index but not for the SA index at the 10% level in this specification. Figure 8 shows the treatment effects on standardized outcomes, PCA index, and SA index, as well as the PCA weights. The regressions control for strata dummies and re-randomization variables (such as the principal decision maker's years of schooling, business age, and whether the principal decision maker is male). It also shows the 90% confidence intervals as in the original analysis. See the appendix on how to adjust for the standard errors.

First, I apply my decision-theoretic approach on **Set 1** of the outcomes. Figure 9 summarizes the results for the minimax and adaptive approach. Panel (a) of Figure 9 shows the treatment effects using the minimax optimal weights for $B \in [0, 10]$. I am able to find a positive effect of the consulting program on entrepreneurial spirit. However, the results are not significant at the 10% level.

Next, I examine what can be learned from data if there is a concern that some out-

Figure 10: Results for the statistical decision approach (with shape restrictions) for the BKS application.



Notes: See the notes under Figure 7.

comes are not directly affected by the consulting program. BKS reran the analysis by considering **Set 2** of the outcomes that dropped two outcomes. Instead of dropping them, I study the effect by imposing shape restrictions as in Example 4.3, where I make the two outcomes less important. The relative importance is controlled by the parameter $\kappa > 0$. The results are summarized in Figure 10. I can still find a positive effect, but they are not significant at the 10% level.

6 Conclusion

Researchers often observe multiple related outcomes that are related to an underlying abstract concept, such as crime and wealth. The outcomes are often aggregated into an index in order to evaluate the treatment effect on these abstract concepts. In this paper, I first studied the properties and issues of the three most popular approaches, namely PCA, SA, and IVM. I show that PCA has several unattractive properties. PCA can have negative weights, does not necessarily maximize precision, is sensitive to arbitrary choices of normalization, and can lead to non-standard limiting distributions. IVM does not suffer from the last two issues, but also has the negative weighting problem. Although PCA and IVM use the correlation information in computing the weights, they do not use the variance matrix of the treatment effects. SA does not have negative weights, but it does not use the correlation structure of the outcomes.

I proposed a statistical decision-theoretic approach to aggregate outcomes that minimizes the mean-squared error of the aggregated treatment effect while ensuring interpretable weights. The weights can be computed by the minimax or the adaptive regret criterion. The adaptive regret criterion has the advantage that it does not require the researcher to commit to a specific level of misspecification. I show that convex optimization can be used to compute the weights. I illustrated my approach through two empirical applications.

References

ABADIR, K. M. AND J. R. MAGNUS (2005): *Matrix algebra*, vol. 1, Cambridge University Press.

AJZENMAN, N. (2021): "The Power of Example: Corruption Spurs Corruption," *American Economic Journal: Applied Economics*, 13, 230–57.

ALIPRANTIS, C. D. AND K. C. BORDER (2006): *Infinite dimensional analysis: a hitchhiker's guide*, Springer, 3 ed.

ALLEE, K. D., C. DO, AND F. G. RAYMUNDO (2022): "Principal component analysis and factor analysis in accounting research," *Journal of Financial Reporting*, 7, 1–39.

ANDERSON, M. AND J. MAGRUDER (2023): "Highly Powered Analysis Plans," *Working Paper*.

ANDERSON, M. L. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103, 1481–1495.

ANDERSON, T. AND H. RUBIN (1956): "Statistical Inference in Factor Analysis," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 111.

ANDREWS, I., M. GENTZKOW, AND J. M. SHAPIRO (2020): "Transparency in Structural Research," *Journal of Business & Economic Statistics*, 38, 711–722.

ANDREWS, I. AND J. M. SHAPIRO (2021): "A Model of Scientific Communication," *Econometrica*, 89, 2117–2142.

ARAGÓN, F. J., M. A. GOBERNA, M. A. LÓPEZ, AND M. M. RODRÍGUEZ (2019): *Nonlinear optimization*, Springer.

ARMSTRONG, T., P. M. KLINE, AND L. SUN (2024): "Adapting to misspecification," Tech. rep., Working Paper.

ARMSTRONG, T. B. AND M. KOLESÁR (2018): "Optimal inference in a class of regression models," *Econometrica*, 86, 655–683.

——— (2021a): "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness," *Econometrica*, 89, 1141–1177.

——— (2021b): "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, 12, 77–108.

BAI, J. AND S. NG (2002): "Determining the number of factors in approximate factor models," *Econometrica*, 70, 191–221.

BAI, J. AND P. WANG (2016): "Econometric analysis of large factor models," *Annual Review of Economics*, 8, 53–80.

BANERJEE, A., E. DUFLO, AND G. SHARMA (2021): "Long-term effects of the targeting the ultra poor program," *American Economic Review: Insights*, 3, 471–486.

BATTAGLIA, L., T. CHRISTENSEN, S. HANSEN, AND S. SACHER (2024): "Inference for Regression with Variables Generated from Unstructured Data," *Working Paper*.

BAU, N. (2022): "Estimating an Equilibrium Model of Horizontal Competition in Education," *Journal of Political Economy*, 130, 1717–1764.

BERTSEKAS, D. (2009): *Convex optimization theory*, vol. 1, Athena Scientific.

BERTSIMAS, D. AND J. TSITSIKLIS (1997): *Introduction to Linear Optimization*, vol. 6, Athena Scientific, Belmont (Mass.).

BHATT, M. P., S. B. HELLER, M. KAPUSTIN, M. BERTRAND, AND C. BLATTMAN (2023): "Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago*," *The Quarterly Journal of Economics*, 139, 1–56.

BICKEL, P. (1984): "Parametric robustness: small biases can be worthwhile," *The Annals of Statistics*, 12, 864–879.

BLANDHOL, C., J. BONNEY, M. MOGSTAD, AND A. TORGOVITSKY (2025): "When is TSLS Actually LATE?" *Working paper*.

BLATTMAN, C., N. FIALA, AND S. MARTINEZ (2020): "The Long-Term Impacts of Grants on Poverty: Nine-Year Evidence from Uganda's Youth Opportunities Program," *American Economic Review: Insights*, 2, 287–304.

BONHOMME, S. (2020): "Discussion of "Transparency in Structural Research" by Isaiah Andrews, Matthew Gentzkow, and Jesse Shapiro," *Journal of Business & Economic Statistics*, 38, 723–725.

BORUSYAK, K., X. JARAVEL, AND J. SPIESS (2024): "Revisiting Event-Study Designs: Robust and Efficient Estimation," *The Review of Economic Studies*, 91, 3253–3285.

BOYD, S. P. AND L. VANDENBERGHE (2004): *Convex optimization*, Cambridge university press.

BRUHN, M., D. KARLAN, AND A. SCHOAR (2018): "The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico," *Journal of Political Economy*, 126, 635–687.

CAI, T. T. AND M. G. LOW (2004): "An Adaptation Theory for Nonparametric Confidence Intervals," *Annals of Statistics*, 1805–1840.

CAMPANTE, F. R. AND Q.-A. DO (2014): "Isolated Capital Cities, Accountability, and Corruption: Evidence from US States," *American Economic Review*, 104, 2456–81.

DE CHAISEMARTIN, C. AND X. D'HAULTFŒUILLE (2017): "Fuzzy Differences-in-Differences," *The Review of Economic Studies*, 85, 999–1028.

DONOHO, D. L. (1994): "Statistical estimation and optimal recovery," *The Annals of Statistics*, 22, 238–270.

DONOHO, D. L., R. C. LIU, AND B. MACGIBBON (1990): "Minimax risk over hyperrectangles, and implications," *The Annals of Statistics*, 1416–1437.

DOUGLAS, J. A. (2001): "Asymptotic identifiability of nonparametric item response models," *Psychometrika*, 66, 531–540.

FEDCHENKO, D. (2025): "Summary indices in empirical research: practices, pitfalls, and proposals," *Working Paper*.

FILMER, D. AND L. H. PRITCHETT (2001): "Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India," *Demography*, 38, 115–132.

FRANKEL, A. AND M. KASY (2022): "Which Findings Should Be Published?" *American Economic Journal: Microeconomics*, 14, 1–38.

FU, J. AND D. P. GREEN (2025): "Causal Inference for Experiments with Latent Outcomes: Key Results and Their Implications for Design and Analysis," Tech. rep., Working Paper.

GAFAROV, B. (2025): "Simple subvector inference on sharp identified set in affine models," *Journal of Econometrics*, 249, 105952.

GECHTER, M., K. HIRANO, J. LEE, M. MAHMUD, O. MONDAL, J. MORDUCH, S. RAVINDRAN, AND A. S. SHONCHOY (2024): "Selecting Experimental Sites for External Validity," .

GHOJOGH, B., F. KARRAY, AND M. CROWLEY (2023): "Eigenvalue and Generalized Eigenvalue Problems: Tutorial," .

GLASS, G. V., B. MCGAW, AND M. L. SMITH (1981): *Meta-analysis in social research*, SAGE Publications.

GÓMEZ, M. (2024): "Indexes and Multiple Hypothesis Testing," .

GOODMAN-BACON, A. (2021): "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 225, 254–277, themed Issue: Treatment Effect 1.

HASTIE, T., R. TIBSHIRANI, J. FRIEDMAN, ET AL. (2009): "The elements of statistical learning," .

HECKMAN, J., R. PINTO, AND P. SAVELYEV (2013): "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103, 2052–86.

HECKMAN, J. J., S. H. MOON, R. PINTO, P. A. SAVELYEV, AND A. YAVITZ (2010): "The rate of return to the HighScope Perry Preschool Program," *Journal of public Economics*, 94, 114–128.

HOTELLING, H. (1933): "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, 24, 417.

HU, Y. AND S. M. SCHENNACH (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

HU, Y., J.-L. SHIU, Y. XIN, AND J. YAO (2024): "Optimal Linear Rank Indexes for Latent Variables," *Working Paper*.

ISHIHARA, T. AND T. KITAGAWA (2024): "Evidence Aggregation for Treatment Choice," Tech. rep., Working Paper.

JACKSON, J. E. (2005): *A user's guide to principal components*, John Wiley & Sons.

JAMES, G., D. WITTEN, T. HASTIE, AND R. TIBSHIRANI (2021): *An introduction to statistical learning: with applications in R*, vol. 103, Springer.

JOLLIFFE, I. T. (2002): *Principal component analysis, Second edition*, Springer.

JONES, D., D. MOLITOR, AND J. REIF (2019): "What do Workplace Wellness Programs do? Evidence from the Illinois Workplace Wellness Study*," *The Quarterly Journal of Economics*, 134, 1747–1791.

KASY, M. AND J. SPIESS (2024): "Optimal Pre-Analysis Plans: Statistical Decisions Subject to Implementability," .

KLING, J. R. AND J. B. LIEBMAN (2004): "Experimental analysis of neighborhood effects on youth," *SSRN Electronic Journal*.

KLING, J. R., J. B. LIEBMAN, AND L. F. KATZ (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75, 83–119.

KOLENIKOV, S. AND G. ANGELES (2009): "SOCIOECONOMIC STATUS MEASUREMENT WITH DISCRETE PROXY VARIABLES: IS PRINCIPAL COMPONENT ANALYSIS A RELIABLE ANSWER?" *Review of Income and Wealth*, 55, 128–165.

KOSOROK, M. R. (2008): *Introduction to empirical processes and semiparametric inference*, Springer.

LUBOTSKY, D. AND M. WITTENBERG (2006): "Interpretation of regressions with multiple proxies," *The Review of Economics and Statistics*, 88, 549–562.

MAGNUS, J. R. AND H. NEUDECKER (2019): *Matrix differential calculus with applications in statistics and econometrics*, John Wiley & Sons, 3 ed.

MANSKI, C. F. (2004): "Statistical treatment rules for heterogeneous populations," *Econometrica*, 72, 1221–1246.

MARDIA, K. V., J. T. KENT, AND C. C. TAYLOR (1979): *Multivariate analysis*, John Wiley & Sons.

——— (2024): *Multivariate analysis*, John Wiley & Sons.

MEYER, C. D. (2023): *Matrix analysis and applied linear algebra*, Society for Industrial and Applied Mathematics, 2 ed.

MONTIEL OLEA, J. L., B. PRALLON, C. QIU, J. STOYE, AND Y. SUN (2025): "Externally Valid Selection of Experimental Sites via the k-Median Problem," .

MURPHY, K. M. AND R. H. TOPEL (1985): "Estimation and inference in two-step econometric models," *Journal of Business & Economic Statistics*, 3, 88–97.

NICULESCU, C. P. AND L.-E. PERSSON (2018): *Convex Functions and Their Applications: A Contemporary Approach*, Springer., 1 ed.

NOCEDAL, J. AND S. J. WRIGHT (2006): *Numerical optimization*, Springer.

O'BRIEN, P. C. (1984): "Procedures for comparing samples with multiple endpoints," *Biometrics*, 1079–1087.

PAGAN, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25, 221–47.

PARKER, S. W. AND T. VOGL (2023): "Do Conditional Cash Transfers Improve Economic Outcomes in the Next Generation? Evidence from Mexico," *The Economic Journal*, 133, 2775–2806.

PEARSON, K. (1901): "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2, 559–572.

R CORE TEAM (2025): "Principal Components Analysis," https://stat.ethz.ch/R-manual/R-devel/library/stats/html/princomp.html, accessed: September 2, 2025.

ROMANO, J. P., A. M. SHAIKH, AND M. WOLF (2010): "Hypothesis testing in econometrics," *Annual Review of Economics*, 2, 75–104.

SAVAGE, L. J. (1951): "The theory of statistical decision," *Journal of the American Statistical association*, 46, 55–67.

SHAPIRO, A. (1993): "Asymptotic behavior of optimal solutions in stochastic programming," *Mathematics of Operations Research*, 18, 829–845.

SHAPIRO, A., D. DENTCHEVA, AND A. RUSZCZYNSKI (2021): *Lectures on Stochastic Programming: Modeling and Theory, Third Edition*, Philadelphia, PA: Society for Industrial and Applied Mathematics.

STAIGER, D. AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

STATACORP (2025): "Stata 19 Base Reference Manual," https://www.stata.com/manuals/mvpca.pdf, accessed: September 2, 2025.

STEWART, G. W. (2001): *Matrix Algorithms: Volume II: Eigensystems*, SIAM.

STEWART, G. W. AND J.-G. SUN (1990): *Matrix perturbation theory*, Academic Press, Inc.

STOETZER, L. F., X. ZHOU, AND M. STEENBERGEN (2025): "Causal inference with latent outcomes," *American Journal of Political Science*, 69, 624–640.

SUN, L. AND S. ABRAHAM (2021): "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 225, 175–199, themed Issue: Treatment Effect 1.

SŁOCZYŃSKI, T. (2024): "When Should We (Not) Interpret Linear IV Estimands as LATE?" Tech. rep.

TABER, C. (2020): "Thoughts on "Transparency in Structural Research"," *Journal of Business & Economic Statistics*, 38, 726–727.

TAMER, E. (2020): "Discussion on "Transparency in Structural Research" by I. Andrews, M. Gentkow and J. Shapiro," *Journal of Business & Economic Statistics*, 38, 728–730.

THE MATHWORKS INC. (2025): "pca Principal component analysis of raw data," https://www.mathworks.com/help/stats/pca.html, accessed: September 2, 2025.

VARIAN, H. R. (2014): *Intermediate microeconomics: a modern approach*, W. W. Norton & Company, 9 ed.

VIVIANO, D., K. WÜTHRICH, AND P. NIEHAUS (2025): "A model of multiple hypothesis testing," Tech. rep., Working Paper.

VYAS, S. AND L. KUMARANAYAKE (2006): "Constructing socio-economic status indices: how to use principal components analysis," *Health Policy and Planning*, 21, 459–468.

WACHSMUTH, G. (2013): "On LICQ and the uniqueness of Lagrange multipliers," *Operations Research Letters*, 41, 78–80.

WALD, A. (1950): "Statistical decision functions," in *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer, 342–357.

WILLIAMS, B. (2019): "Identification of a nonseparable model under endogeneity using binary proxies for unobserved heterogeneity," *Quantitative Economics*, 10, 527–563.

WOOLDRIDGE, J. M. (2013): *Introductory econometrics a modern approach*, South-Western cengage learning, 5 ed.

Appendix

A Appendix for Section 3

In the appendix, I assume that $\{(D_i, \boldsymbol{X}_i', \boldsymbol{Y}_i')\}_{i=1}^n$ are i.i.d. across i.

A.1 Additional results for the examples on substitutes

A.1.1 Details for Example 3.4

Consider the setup and the optimal solution (7) in Example 3.4 and recall that (V_i, D_i, A_i) are assumed to be mutually independent.

To compute the correlation terms, note that

$$\begin{aligned}
\mathbb{E}[Y_{i,1}^\star Y_{i,2}^\star] &= \frac{1}{p_2}\mathbb{E}[A_i(1-A_i)\text{Income}_i(D_i)^2] = \frac{1}{p_2}\mathbb{E}[A_i(1-A_i)]\mathbb{E}[\text{Income}_i(D_i)^2], \\
\mathbb{E}[Y_{i,1}^\star] &= \mathbb{E}[A_i\text{Income}_i(D_i)] = \mathbb{E}[A_i]\mathbb{E}[\text{Income}_i(D_i)], \\
\mathbb{E}[Y_{i,2}^\star] &= \frac{1}{p_2}\mathbb{E}[(1-A_i)\text{Income}_i(D_i)] = \frac{1}{p_2}\mathbb{E}[(1-A_i)]\mathbb{E}[\text{Income}_i(D_i)],
\end{aligned}$$

because $A_i \perp\!\!\!\perp D_i$ by assumption.

Write $\mu_A \equiv \mathbb{E}[A_i]$, $\sigma_A^2 \equiv \text{Var}[A_i] = \mathbb{E}[A_i^2] - \mathbb{E}[A_i]^2$, $\mu_I \equiv \mathbb{E}[\text{Income}_i(D_i)]$, and $\sigma_I^2 \equiv \text{Var}[\text{Income}_i(D_i)] = \mathbb{E}[\text{Income}_i(D_i)^2] - \mathbb{E}[\text{Income}_i(D_i)]^2$. Hence, the covariance between $Y_{i,1}^\star$ and $Y_{i,2}^\star$ can be computed as

$$\text{Cov}[Y_{i,1}^\star, Y_{i,2}^\star] = \frac{1}{p_2}\left[\mu_A(1-\mu_A)\sigma_I^2 - \sigma_A^2(\mu_I^2 + \sigma_I^2)\right].$$

It follows that $\text{Cov}[Y_{i,1}^\star, Y_{i,2}^\star] < 0$ if and only if

$$\mu_A(1-\mu_A)\sigma_I^2 < \sigma_A^2(\mu_I^2 + \sigma_I^2).$$

Now, I return to the numerical specification in Example 3.4 where $p_2 = 2$, $\mathbb{P}[A_i = 0.2] = \mathbb{P}[A_i = 0.8] = 0.5$, $\mathbb{P}[D_i = 0] = \mathbb{P}[D_i = 1] = 0.5$, $\text{Income}_i(D_i) = 10 + 5D_i + V_i$, where $V_i \sim \text{Uniform}[0, 5]$ and (V_i, D_i, A_i) are mutually independent. Hence,

$$\begin{aligned}
\mathbb{E}[A_i] &= 0.5, \\
\text{Var}[A_i] &= \mathbb{E}[A_i^2] - \mathbb{E}[A_i]^2 = 0.5(0.8^2 + 0.2^2) - 0.5^2 = 0.09,
\end{aligned}$$

$$\mathbb{E}[\text{Income}_i(D_i)] = 15,$$
$$\text{Var}[\text{Income}_i(D_i)] = \frac{25}{3}.$$

As a result,

$$\mu_A(1-\mu_A)\sigma_I^2 = 0.5(1-0.5)\frac{25}{3} = \frac{25}{12},$$

and

$$\sigma_A^2(\mu_I^2 + \sigma_I^2) = 21.$$

Hence, $\text{Cov}[Y_{i,1}^\star, Y_{i,2}^\star] = \frac{\frac{25}{12}-21}{p_2} = -9.4583$. To compute the correlation, note that

$$\begin{aligned}
\text{Var}[Y_{i,1}^\star] &= \text{Var}[A_i \text{Income}_i(D_i)] \\
&= \mathbb{E}[A_i^2]\mathbb{E}[\text{Income}_i(D_i)^2] - \{\mathbb{E}[A_i]\mathbb{E}[\text{Income}_i(D_i)]\}^2 \\
&= 0.5(0.8^2 + 0.2^2)\left(15^2 + \frac{25}{3}\right) - (7.5)^2 \\
&= \frac{277}{12},
\end{aligned}$$

and

$$\begin{aligned}
\text{Var}[Y_{i,2}^\star] &= \frac{1}{p_2^2}\text{Var}[(1-A_i)\text{Income}_i(D_i)] \\
&= \frac{\mathbb{E}[(1-A_i)^2]\mathbb{E}[\text{Income}_i(D_i)^2] - \{\mathbb{E}[(1-A_i)]\mathbb{E}[\text{Income}_i(D_i)]\}^2}{p_2^2} \\
&= \frac{\mathbb{E}[A_i^2]\mathbb{E}[\text{Income}_i(D_i)^2] - \{\mathbb{E}[A_i]\mathbb{E}[\text{Income}_i(D_i)]\}^2}{p_2^2} \\
&= \frac{277}{48}.
\end{aligned}$$

Therefore, $\text{Corr}[Y_{i,1}^\star, Y_{i,2}^\star] \approx -0.82$.

A.1.2 Additional example with perfect substitutes

Consider a stylized two-good example below where $Y_{i,1}$ and $Y_{i,2}$ represent two assets (e.g., cows and sheep). Suppose the agents view them as perfect substitutes (see, e.g., Chapter 5 of Varian (2014)). More precisely, consider the following consumer optimiza-

tion problem

$$(Y_{i,1}, Y_{i,2}) = \underset{y_1, y_2}{\arg\max} \quad y_1 + A_i y_2 \qquad \text{s.t.} \quad y_1 + p_2 y_2 \leq \text{Income}_i(D_i), \tag{A.1}$$

where the price of good 1 is normalized to 1, p_2 is the price of good 2, $\text{Income}_i(D_i) \geq 0$ is an income specific to individual i that is affected by $D_i \in \{0, 1\}$, and A_i is the utility parameter for individual i that follows a truncated normal distribution with mean μ, variance 1, and bounded in $[\underline{A}, \overline{A}]$. Assume that A_i is independent of $\text{Income}_i(D_i)$ and that $0 < \underline{A} < p_2 < \overline{A}$.

The optimal solution to (A.1) is given by

$$(Y_{i,1}, Y_{i,2}) = \begin{cases} (0, \frac{\text{Income}_i(D_i)}{p_2}) & \frac{A_i}{p_2} > 1, \\ \{(t, \frac{\text{Income}_i(D_i) - t}{p_2}) : t \in [0, \text{Income}_i(D_i)]\} & \frac{A_i}{p_2} = 1, \\ (\text{Income}_i(D_i), 0) & \frac{A_i}{p_2} < 1. \end{cases} \tag{A.2}$$

Using the optimal consumption bundle (A.2), I have

$$\begin{aligned} \mathbb{E}[Y_{i,1}] &= \mathbb{E}[Y_{i,1} | A_i > p_2] \mathbb{P}[A_i > p_2] + \mathbb{E}[Y_{i,1} | A_i = p_2] \mathbb{P}[A_i = p_2] \\ &\quad + \mathbb{E}[Y_{i,1} | A_i < p_2] \mathbb{P}[A_i < p_2] \\ &= \mathbb{E}[\text{Income}_i(D_i)] \mathbb{P}[A_i < p_2] \end{aligned}$$

where the first equality follows from the law of total probability, the second equality follows from $\mathbb{P}[A_i = p_2] = 0$ and $Y_{i,1} = 0$ for $A_i > p_2$, and the last equality follows from the independence of A_i and $\text{Income}_i(D_i)$. The term $\mathbb{P}[A_i < p_2]$ has a closed-form expression as by the properties of truncated normal distribution.

Similarly,

$$\begin{aligned} \mathbb{E}[Y_{i,2}] &= \mathbb{E}[Y_{i,2} | A_i > p_2] \mathbb{P}[A_i > p_2] + \mathbb{E}[Y_{i,2} | A_i = p_2] \mathbb{P}[A_i = p_2] \\ &\quad + \mathbb{E}[Y_{i,2} | A_i < p_2] \mathbb{P}[A_i < p_2] \\ &= \frac{1}{p_2} \mathbb{E}[\text{Income}_i(D_i)] \mathbb{P}[A_i > p_2]. \end{aligned}$$

I also have

$$\begin{aligned}\mathbb{E}[Y_{i,1}Y_{i,2}] &= \mathbb{E}[Y_{i,1}Y_{i,2}|A_i > p_2]\mathbb{P}[A_i > p_2] + \mathbb{E}[Y_{i,1}Y_{i,2}|A_i = p_2]\mathbb{P}[A_i = p_2] \\ &\quad + \mathbb{E}[Y_{i,1}Y_{i,2}|A_i < p_2]\mathbb{P}[A_i < p_2] \\ &= 0,\end{aligned}$$

where the equality follows because $Y_{i,1}Y_{i,2} = 0$ for $A_i \neq p_2$ and $\mathbb{P}[A_i = p_2] = 0$.

Combining the above results, the covariance of the pair of goods is

$$\begin{aligned}\text{Cov}[Y_{i,1}, Y_{i,2}] &= \mathbb{E}[Y_{i,1}Y_{i,2}] - \mathbb{E}[Y_{i,1}]\mathbb{E}[Y_{i,2}] \\ &= -\frac{1}{p_2}\mathbb{E}[\text{Income}_i(D_i)]^2\mathbb{P}[A_i < p_2]\mathbb{P}[A_i > p_2] \\ &< 0, \end{aligned} \tag{A.3}$$

where the inequality follows because $\underline{A} < p_2 < \overline{A}$ and $\text{Income}_i(D_i) > 0$.

It follows that the two outcomes are negatively correlated with each other. Thus, the PCA approach is going to put weights with opposite signs on the two outcomes. This PCA index is again counterintuitive as in Section A.1.1.

A.2 Proofs for propositions in the main text

To begin with, I impose the following high-level assumption in the large-sample analysis. It requires the treatment effects and variance estimators to be consistent and that a suitable central limit theorem applies to characterize the limiting distribution. As Assumption 2.2 is imposed, $\widehat{\boldsymbol{\beta}}_n$ is estimated using the standardized outcome and $\boldsymbol{\Sigma}$ is the corresponding asymptotic variance matrix of the treatment effects using the standardized outcomes. To allow for possibly general choices of the matrix used to compute the PC1, I use $\boldsymbol{\Omega}$ and $\widehat{\boldsymbol{\Omega}}_n$ below. In the case that the correlation matrix of $\boldsymbol{Y}_i$ is used, $\boldsymbol{\Omega}$ is $\boldsymbol{\Sigma}_Y$ and $\widehat{\boldsymbol{\Omega}}_n$ is the corresponding consistent estimator.

Assumption A.1.

(a) *$\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\Omega}}_n$ are consistent estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ respectively.*

(b) $$\sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \\ \text{vech}[\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega}]\end{pmatrix} \xrightarrow{d} \begin{pmatrix}\boldsymbol{Z}_{\boldsymbol{\beta}} \\ \boldsymbol{Z}_{\text{vech}[\boldsymbol{\Omega}]}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{0}_q \\ \boldsymbol{0}_\ell\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{\Psi}'_{\boldsymbol{\Omega},\boldsymbol{\beta}} \\ \boldsymbol{\Psi}_{\boldsymbol{\Omega},\boldsymbol{\beta}} & \boldsymbol{\Psi}_{\boldsymbol{\Omega}}\end{pmatrix}\right).$$

In the above, $\ell \equiv \frac{q(q+1)}{2}$ is the number of entries in the lower triangular portion of the symmetric variance matrix $\boldsymbol{\Omega}$. $\text{vech}(\cdot)$ is the half-vectorization notation that stacks the

columns of the lower triangular portion of the matrix into one vector of length ℓ. For example, $\text{vech}[\left(\begin{smallmatrix} x_1 & x_2 \\ x_2 & x_3 \end{smallmatrix}\right)] = (x_1, x_2, x_3)'$.

In light of the above assumption, in the following, I show a slightly more general result to Proposition 3.6, where PCA uses $\boldsymbol{\Omega}$ instead of $\boldsymbol{\Sigma}_Y$. I use $\{(\boldsymbol{\nu}_j, \lambda_j)\}_{j=1}^q$ in Section 3.1.1 be the eigenpairs of $\boldsymbol{\Omega}$ instead and that Assumption 3.1 is applied to $\boldsymbol{\Omega}$ instead of $\boldsymbol{\Sigma}_Y$.

The proof proceeds as follows. Let $\boldsymbol{l} \in \mathbb{R}^q$ such that $\boldsymbol{c} \neq \boldsymbol{l}$ and $\boldsymbol{l}'\boldsymbol{w}_{\text{pca}} > 0$. Let $\widehat{\boldsymbol{\nu}}_{n,1} = \arg\max_{\boldsymbol{w}:\boldsymbol{w}'\boldsymbol{w}=1, \boldsymbol{l}'\boldsymbol{w}\geq 0} \boldsymbol{w}'\widehat{\boldsymbol{\Omega}}_n\boldsymbol{w}$ be an estimator of $\boldsymbol{w}_{\text{pca}}$ and $\widehat{\varsigma}_n \equiv \text{sign}(\boldsymbol{c}'\widehat{\boldsymbol{\nu}}_{n,1})$. Hence, $\widehat{\boldsymbol{w}}_{\text{pca},n} = \widehat{\varsigma}_n\widehat{\boldsymbol{\nu}}_{n,1}$. First, by a similar reasoning as in Proposition A.19, I have

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1}'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\text{pca},n}'\boldsymbol{\beta}) &= \sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1}'\widehat{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\nu}}_{n,1}\boldsymbol{\beta} + \widehat{\boldsymbol{\nu}}_{n,1}\boldsymbol{\beta} - \boldsymbol{w}_{\text{pca},n}'\boldsymbol{\beta}) \\
&= \widehat{\boldsymbol{\nu}}_{n,1}'[\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})] + \boldsymbol{\beta}'[\sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1} - \boldsymbol{w}_{\text{pca},n})] \\
&\xrightarrow{d} \begin{pmatrix} \boldsymbol{w}_{\text{pca}}' & \boldsymbol{\beta}'\boldsymbol{B}_{\boldsymbol{\nu}} \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_{\boldsymbol{\beta}} \\ \boldsymbol{Z}_{\text{vech}[\boldsymbol{\Omega}]} \end{pmatrix} \\
&\equiv \boldsymbol{Z}_\tau. \qquad \text{(A.4)}
\end{aligned}$$

Next, note that

$$\begin{aligned}
\{\widehat{\varsigma}_n = 1\} &= \{\boldsymbol{c}'\widehat{\boldsymbol{\nu}}_{n,1} \geq 0\} \\
&= \{\sqrt{n}(\boldsymbol{c}'\widehat{\boldsymbol{\nu}}_{n,1} - \boldsymbol{c}'\boldsymbol{w}_{\text{pca},n}) \geq -\sqrt{n}\boldsymbol{c}'\boldsymbol{w}_{\text{pca},n}\} \\
&= \{\sqrt{n}\boldsymbol{c}'(\widehat{\boldsymbol{\nu}}_{n,1} - \boldsymbol{w}_{\text{pca},n}) \geq -\delta\} \qquad \text{(A.5)}
\end{aligned}$$

where the first line uses the definition of $\widehat{\varsigma}_n$, the second line subtracts $\sqrt{n}\boldsymbol{c}'\boldsymbol{w}_{\text{pca},n}$ on both sides, and the third line uses the assumption on $\boldsymbol{c}'\boldsymbol{w}_{\text{pca},n}$. Using the notations of Assumption A.1 and (A.4), I have

$$\begin{aligned}
\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\nu}}_{n,1}'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\text{pca},n}'\boldsymbol{\beta} \\ \boldsymbol{c}'(\widehat{\boldsymbol{\nu}}_{n,1} - \boldsymbol{w}_{\text{pca},n}) \end{pmatrix} &= \begin{pmatrix} \widehat{\boldsymbol{\nu}}_{n,1}' & \boldsymbol{\beta}' \\ \boldsymbol{0}_q' & \boldsymbol{c}' \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1} - \boldsymbol{w}_{\text{pca},n}) \end{pmatrix} \\
&\xrightarrow{d} \begin{pmatrix} \boldsymbol{w}_{\text{pca}}' & \boldsymbol{\beta}' \\ \boldsymbol{0}_q' & \boldsymbol{c}' \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_{\boldsymbol{\beta}} \\ \boldsymbol{B}_{\boldsymbol{\nu}}\boldsymbol{Z}_{\text{vech}[\boldsymbol{\Omega}]} \end{pmatrix}. \qquad \text{(A.6)}
\end{aligned}$$

Next, write

$$\begin{aligned}
\sqrt{n}(\widehat{\tau}_n - \tau_n) &= \sqrt{n}(\widehat{\boldsymbol{w}}_{\text{pca},n}'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\text{pca},n}'\boldsymbol{\beta}) \\
&= \sqrt{n}(\widehat{\varsigma}_n\widehat{\boldsymbol{\nu}}_{n,1}'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\text{pca},n}'\boldsymbol{\beta})
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{n}(\widehat{\varsigma}_n \widehat{\boldsymbol{\nu}}_{n,1}' \widehat{\boldsymbol{\beta}}_n - \widehat{\varsigma}_n \boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta} + \widehat{\varsigma}_n \boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta} - \boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta}) \\
&= \widehat{\varsigma}_n \sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1}' \widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta}) + (\widehat{\varsigma}_n - 1)\sqrt{n}(\boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta}), \qquad \text{(A.7)}
\end{aligned}
$$

where the first line uses the definition of $\widehat{\tau}_n$ and τ_n, the second line uses the definition of $\widehat{\boldsymbol{w}}_{\mathrm{pca},n}$ in the beginning of the proof, and the third line adds and subtracts.

Hence, for any $t \in \mathbb{R}$,

$$
\begin{aligned}
&\mathbb{P}[\sqrt{n}(\widehat{\tau}_n - \tau_n) \leq t] \\
&\quad = \mathbb{P}[\sqrt{n}(\widehat{\tau}_n - \tau_n) \leq t, \widehat{\varsigma}_n = 1] + \mathbb{P}[\sqrt{n}(\widehat{\tau}_n - \tau_n) \leq t, \widehat{\varsigma}_n = -1] \\
&\quad = \mathbb{P}[\sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1}' \widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta}) \leq t, \widehat{\varsigma}_n = 1] \\
&\qquad + \mathbb{P}[-\sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1}' \widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta}) \leq t + 2\sqrt{n}(\boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta}), \widehat{\varsigma}_n = -1]. \qquad \text{(A.8)}
\end{aligned}
$$

Let $\mathcal{C} \equiv [t_{\mathrm{lb}}, t_{\mathrm{ub}}]$ be the interval as defined in the statement of the proposition. Since $\boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta} \neq 0$ by the hypothesis of the proposition, $2\sqrt{n}(\boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta})$ diverges to ∞ if $\boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta} > 0$ and to $-\infty$ if $\boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta} < 0$. By (A.4), $\sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1}' \widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta})$ is tight. In addition, $\{\widehat{\varsigma}_n = -1\} = \{\sqrt{n}\boldsymbol{c}'(\widehat{\boldsymbol{\nu}}_{n,1} - \boldsymbol{w}_{\mathrm{pca},n}) < -\delta\}$ by (A.5). The above facts and with (A.6), I have

$$
\lim_{n\to\infty} \mathbb{P}[-\sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1}' \widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta}) - 2\sqrt{n}(\boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta}) \in \mathcal{C}, \widehat{\varsigma}_n = -1] = 0. \qquad \text{(A.9)}
$$

Using (A.8) and (A.9), I have

$$
\begin{aligned}
\lim_{n\to\infty} \mathbb{P}[\sqrt{n}(\widehat{\tau}_n - \tau_n) \in \mathcal{C}] &= \lim_{n\to\infty} \mathbb{P}[\sqrt{n}(\widehat{\boldsymbol{\nu}}_{n,1}' \widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\mathrm{pca},n}' \boldsymbol{\beta}) \in \mathcal{C}, \widehat{\varsigma}_n = 1] \\
&\leq \lim_{n\to\infty} \mathbb{P}[\widehat{\varsigma}_n = 1] \\
&= \lim_{n\to\infty} \mathbb{P}[\sqrt{n}\boldsymbol{c}'(\widehat{\boldsymbol{\nu}}_{n,1} - \boldsymbol{w}_{\mathrm{pca},n}) \geq -\delta] \\
&= \mathbb{P}[\boldsymbol{c}' \boldsymbol{B}_{\boldsymbol{\nu}} \boldsymbol{Z}_{\mathrm{vech}[\boldsymbol{\Omega}]} \geq -\delta], \qquad \text{(A.10)}
\end{aligned}
$$

where the first line follows from (A.8) and (A.9), the third line follows from (A.5), and the fourth line follows from (A.6). The result of the proposition follows from taking the limit as $\delta \downarrow 0$ and that $\boldsymbol{c}' \boldsymbol{B}_{\boldsymbol{\nu}} \boldsymbol{Z}_{\mathrm{vech}[\boldsymbol{\Omega}]}$ is normal.

Proof of Proposition 3.9. The problem is a strictly convex problem because $\overline{\boldsymbol{\Sigma}}$ is positive definite and the constraint is affine. Let the Lagrangian of the optimization problem be

$$
\mathcal{L} = \boldsymbol{w}' \overline{\boldsymbol{\Sigma}} \boldsymbol{w} + \mu(1 - \boldsymbol{w}' \mathbf{1}_q).
$$

The first-order condition with respect to $\boldsymbol{w}$ is $2\overline{\boldsymbol{\Sigma}}\boldsymbol{w} - \mu \mathbf{1}_q = 0$. Solving,

$$\boldsymbol{w} = \frac{1}{2}\mu \overline{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_q. \tag{A.11}$$

Substituting this to the constraint gives $\frac{1}{2}\mu \mathbf{1}_q' \overline{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_q = 1$. This means $\mu = \frac{2}{\mathbf{1}_q' \overline{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_q}$. This is well-defined because $\overline{\boldsymbol{\Sigma}}$ is positive definite, so $\mathbf{1}_q' \overline{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_q > 0$. Substituting this back to (A.11) gives

$$\boldsymbol{w} = \frac{\overline{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_q}{\mathbf{1}_q' \overline{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_q}. \tag{A.12}$$

When $\overline{\boldsymbol{\Sigma}}$ is an equicorrelation matrix, it can be written as $\overline{\boldsymbol{\Sigma}} = (1-\overline{\rho})\boldsymbol{I}_q + \overline{\rho}\mathbf{1}_q\mathbf{1}_q'$ for some $\overline{\rho} \in (0,1)$. Then,

$$\overline{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_q = \frac{1}{1+(q-1)\overline{\rho}}\mathbf{1}_q \qquad \text{and} \qquad \mathbf{1}_q' \overline{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_q = \frac{q}{1+(q-1)\overline{\rho}}.$$

Substituting the above into (A.12) gives $\frac{1}{q}\mathbf{1}_q$ as the optimal solution. □

A.3 Supplemental results and proofs

A.3.1 Supplemental results on the linear model

Lemma A.2. *Suppose the researcher observes $\{(D_i, \boldsymbol{X}_i', \boldsymbol{Y}_i')\}_{i=1}^n$, where $D_i \in \mathbb{R}$, $\boldsymbol{X}_i \in \mathbb{R}^{d_x}$, $\boldsymbol{Y}_i \equiv (Y_{i,1}, \ldots, Y_{i,j}) \in \mathbb{R}^q$, and the intercept is contained in $\boldsymbol{X}_i$. Consider the following estimators and linear models.*

(a) For each outcome $j = 1, \ldots, q$, consider the linear model

$$Y_{i,j} = \beta_j D_i + \boldsymbol{\zeta}_j' \boldsymbol{X}_i + U_{i,j}, \tag{A.13}$$

where $\beta_j \in \mathbb{R}$, $\boldsymbol{\zeta}_j \in \mathbb{R}^{d_x}$, $U_{i,j} \in \mathbb{R}$, and $i = 1, \ldots, n$. Let $\widehat{\boldsymbol{\beta}}_n \equiv (\widehat{\beta}_{n,1}, \ldots, \widehat{\beta}_{n,q})' \in \mathbb{R}^q$ where $\widehat{\beta}_{n,j}$ is the estimator of β_j in (A.13).

(b) Let $\boldsymbol{w}_n \in \mathbb{R}^q$ be a vector of weights that is potentially data-dependent and consider the linear model

$$\boldsymbol{w}_n' \boldsymbol{Y}_i = \tau D_i + \overline{\boldsymbol{\zeta}}' \boldsymbol{X}_i + \overline{U}_i, \tag{A.14}$$

where $\tau \in \mathbb{R}$, $\overline{\boldsymbol{\zeta}} \in \mathbb{R}^{d_x}$, and $\overline{U}_i \in \mathbb{R}$ for $i = 1, \ldots, n$. Let $\widehat{\tau}_n$ be the estimator of τ in

(A.14).

Write $\boldsymbol{D} \equiv (D_1, \ldots, D_n)'$, $\boldsymbol{X} \equiv (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)'$, $\boldsymbol{P} \equiv \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ *and* $\boldsymbol{M} \equiv \boldsymbol{I}_n - \boldsymbol{P}$ *where* $\boldsymbol{I}_n$ *is the identity matrix. Assume* $\boldsymbol{D}'\boldsymbol{M}\boldsymbol{D}$ *is invertible. Then,* $\widehat{\tau}_n = \boldsymbol{w}_n'\widehat{\boldsymbol{\beta}}_n$.

Proof of Lemma A.2. Write $\widetilde{\boldsymbol{Y}}_j \equiv (Y_{1,j}, \ldots, Y_{n,q})'$ for $j = 1, \ldots, q$, and $\boldsymbol{Y} \equiv (\widetilde{\boldsymbol{Y}}_1, \ldots, \widetilde{\boldsymbol{Y}}_q)'$. Then, from (A.13) and the Frisch-Waugh-Lovell (FWL) theorem, I have

$$\widehat{\beta}_{n,j} = (\boldsymbol{D}'\boldsymbol{M}\boldsymbol{D})^{-1}(\boldsymbol{D}'\boldsymbol{M}\widetilde{\boldsymbol{Y}}_j), \tag{A.15}$$

for $j = 1, \ldots, n$.

Next, for (A.14), I have

$$\widehat{\tau}_n = (\boldsymbol{D}'\boldsymbol{M}\boldsymbol{D})^{-1}[\boldsymbol{D}'\boldsymbol{M}(\widetilde{\boldsymbol{Y}}\boldsymbol{w}_n)] = \boldsymbol{w}_n'\widehat{\boldsymbol{\beta}}_n, \tag{A.16}$$

using (A.15). Hence, the result follows. □

The following lemma discusses the implications of Remark 2.6. In the case of standardizing $\boldsymbol{w}_n'\boldsymbol{Y}_i$, it means setting a_n as the sample mean of $\boldsymbol{w}_n'\boldsymbol{Y}_i$ and b_n as the sample standard deviation of $\boldsymbol{w}_n'\boldsymbol{Y}_i$. In the case of rescaling $\boldsymbol{w}_n'\boldsymbol{Y}_i$ to $[0,1]$, it means setting $a_n = \min_{i=1,\ldots,n} \boldsymbol{w}_n'\boldsymbol{Y}_i$ and $b_n = \max_{i=1,\ldots,n} \boldsymbol{w}_n'\boldsymbol{Y}_i$.

Lemma A.3. *Consider the same assumptions and notations as in Lemma A.2. Let* $a_n, b_n \in \mathbb{R}$ *be potentially data-dependent parameters. Define* $Z_i \equiv \frac{\boldsymbol{w}_n'\boldsymbol{Y}_i - a_n}{b_n}$ *for* $i = 1, \ldots, n$ *where* $b_n \neq 0$. *Consider the model*

$$Z_i = \tau_z D_i + \boldsymbol{\zeta}_z'\boldsymbol{X}_i + V_i, \tag{A.17}$$

where $\tau_z \in \mathbb{R}$, $\boldsymbol{\zeta}_z \in \mathbb{R}^{d_x}$, *and* $V_i \in \mathbb{R}$. *Let* $\widehat{\tau}_{z,n} = \frac{\widehat{\tau}_n}{b_n}$.

Proof of Lemma A.3. Using the definition of Z_j, define $\boldsymbol{Z} \equiv (Z_1, \ldots, Z_n)' = \frac{1}{b_n}(\boldsymbol{w}_n'\boldsymbol{Y}_1 - a_n, \ldots, \ldots, \boldsymbol{w}_n'\boldsymbol{Y}_n - a_n) = \frac{1}{b_n}(\widetilde{\boldsymbol{Y}}\boldsymbol{w}_n - a_n\boldsymbol{1}_n)$. Using the FWL theorem on (A.17), I have

$$\begin{aligned}
\widehat{\tau}_{z,n} &= (\boldsymbol{D}'\boldsymbol{M}\boldsymbol{D})^{-1}(\boldsymbol{D}'\boldsymbol{M}\boldsymbol{Z}) \\
&= \frac{1}{b_n}(\boldsymbol{D}'\boldsymbol{M}\boldsymbol{D})^{-1}(\boldsymbol{D}'\boldsymbol{M}\widetilde{\boldsymbol{Y}}\boldsymbol{w}_n) - \frac{a_n}{b_n}(\boldsymbol{D}'\boldsymbol{M}\boldsymbol{D})^{-1}(\boldsymbol{D}'\boldsymbol{M}\boldsymbol{1}_n) \\
&= \frac{1}{b_n}\widehat{\tau}_n,
\end{aligned}$$

by Lemma A.2. □

Note that whenever the post-processing step is such that $b_n \neq 1$, then even if $\widehat{\beta}_{n,j} = \overline{\beta}$ for $j = 1, \ldots, n$ and $\boldsymbol{w}'\mathbf{1}_q = 1$, $\widehat{\tau}_{z,n}$ will not be equal to $\overline{\beta}$.

A.3.2 Supplemental results on eigenvectors

Lemma A.4. *Let Assumption 3.1 hold. Suppose that $\boldsymbol{c}'\boldsymbol{\nu}_1 \neq 0$. Let $\boldsymbol{w}_{\text{pca}}$ be the optimal solution to the problem (4). Then, $\boldsymbol{w}_{\text{pca}} = \text{sign}(\boldsymbol{c}'\boldsymbol{\nu}_1)\boldsymbol{\nu}_1$.*

Proof of Lemma A.4. For notational simplicity, denote $\widetilde{\boldsymbol{\nu}}_1 \equiv \text{sign}(\boldsymbol{c}'\boldsymbol{\nu}_1)\boldsymbol{\nu}_1$. Assume to the contrary that $\boldsymbol{w}_{\text{pca}} \neq \widetilde{\boldsymbol{\nu}}_1$. First, I verify that $\widetilde{\boldsymbol{\nu}}_1$ is an optimal solution to (4). Note that $\boldsymbol{c}'\widetilde{\boldsymbol{\nu}}_1 = \text{sign}(\boldsymbol{c}'\boldsymbol{\nu}_1)\boldsymbol{c}'\boldsymbol{\nu}_1 \geq 0$ by definition and $\widetilde{\boldsymbol{\nu}}_1'\widetilde{\boldsymbol{\nu}}_1 = \boldsymbol{\nu}_1'\boldsymbol{\nu}_1 = 1$ since $\boldsymbol{\nu}_1$ has unit length. Hence, $\widetilde{\boldsymbol{\nu}}_1$ is a feasible solution to (4). On the other hand,

$$\widetilde{\boldsymbol{\nu}}_1'\boldsymbol{\Sigma}_Y\widetilde{\boldsymbol{\nu}}_1 = \text{sign}(\boldsymbol{c}'\boldsymbol{\nu}_1)^2\boldsymbol{\nu}_1'\boldsymbol{\Sigma}_Y\boldsymbol{\nu}_1 = \boldsymbol{\nu}_1'\boldsymbol{\Sigma}_Y\boldsymbol{\nu}_1 = \lambda_1,$$

because $(\boldsymbol{\nu}_1, \lambda_1)$ is an eigenpair of $\boldsymbol{\Sigma}_Y$. Hence, $\widetilde{\boldsymbol{\nu}}_1$ is an optimal solution to (4).

Since the leading eigenvalue is unique from Assumption 3.1, it follows that $\boldsymbol{w}_{\text{pca}}$ and $\widetilde{\boldsymbol{\nu}}_1$ are parallel to each other. If $\boldsymbol{w}_{\text{pca}} \neq \widetilde{\boldsymbol{\nu}}_1$, it can only be that $\boldsymbol{w}_{\text{pca}} = -\widetilde{\boldsymbol{\nu}}_1$ since both vectors have unit length. But this implies

$$\boldsymbol{c}'\boldsymbol{w}_{\text{pca}} = -\boldsymbol{c}'\widetilde{\boldsymbol{\nu}}_1 = -\text{sign}(\boldsymbol{c}'\widetilde{\boldsymbol{\nu}}_1)\boldsymbol{c}'\widetilde{\boldsymbol{\nu}}_1 < 0,$$

where the inequality follows because $\text{sign}(\boldsymbol{c}'\widetilde{\boldsymbol{\nu}}_1)\boldsymbol{c}'\widetilde{\boldsymbol{\nu}}_1 \geq 0$ and $\boldsymbol{c}'\widetilde{\boldsymbol{\nu}}_1 \neq 0$ by the hypothesis. But this contradicts that $\boldsymbol{w}_{\text{pca}}$ solves (4) because it satisfies $\boldsymbol{c}'\boldsymbol{w}_{\text{pca}} \geq 0$. This completes the proof. □

A.4 Precision analysis with standardized outcomes

This section studies how standardizing outcomes using the sample standard deviation affects the limiting distribution. Under Assumption 2.2, $\boldsymbol{Y}_i$ represents the standardized and oriented outcome. The following vectorized representation that stacks (8) over $j = 1, \ldots, q$ will be helpful in latter analysis:

$$\boldsymbol{Y}_i = \boldsymbol{\xi} + \boldsymbol{\beta}D_i + \boldsymbol{U}_i, \tag{A.18}$$

where $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_q)'$, $\boldsymbol{\xi} \equiv (\xi_1, \ldots, \xi_q)'$, and $\boldsymbol{U}_i \equiv (U_{i,1}, \ldots, U_{i,q})'$.

Let $\widetilde{Y}_{i,j}$ be the jth outcome before standardization for $j = 1, \ldots, q$. To keep the analysis general, I will allow for a general choice of how the outcomes are standardized, and

assume they are divided by $\widehat{s}_{n,j}$ for $j = 1, \ldots, q$. In the following, $\widehat{s}_{n,j}$ can be the sample standard deviation of outcome j using the full sample, control sample, or other choices. Hence,

$$Y_{i,j} = \widehat{s}_{n,j}^{-1} \widetilde{Y}_{i,j}. \tag{A.19}$$

Let $\widehat{\boldsymbol{\beta}}_n$ be the treatment effect of D_i on $\boldsymbol{Y}_i$ and $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ be the treatment effect of D_i on $\widetilde{\boldsymbol{Y}}_i$ (potentially using additional covariates). By Lemma A.2, it means that

$$\widehat{\boldsymbol{\beta}}_n = \operatorname{diag}[\widehat{\boldsymbol{s}}_n]^{-1} \widehat{\widetilde{\boldsymbol{\beta}}}_n \quad \text{and} \quad \boldsymbol{\beta} = \operatorname{diag}[\boldsymbol{s}]^{-1} \widetilde{\boldsymbol{\beta}}, \tag{A.20}$$

where $\boldsymbol{s} \equiv (s_1, \ldots, s_q)'$.

To study the large-sample properties, I consider the following assumptions. The assumptions are standard in that it requires that each outcome's standard deviation (suitably defined depending on Assumption 2.2 and the method, e.g., on the full sample or using the control sample) is positive, the $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ and $\widehat{\boldsymbol{s}}_n$ are consistent, and that a suitable central limit theorem can be applied.

Assumption A.5.

1. $s_j > 0$ *for each* $j = 1, \ldots, q$.
2. $\widehat{\widetilde{\boldsymbol{\beta}}}_n \xrightarrow{p} \widetilde{\boldsymbol{\beta}}$ *and* $\widehat{\boldsymbol{s}}_n \xrightarrow{p} \boldsymbol{s}$.
3. *Suppose that*

$$\sqrt{n} \begin{pmatrix} \widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{s}}_n - \boldsymbol{s} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \boldsymbol{Z}_{\widetilde{\boldsymbol{\beta}}} \\ \boldsymbol{Z}_{\boldsymbol{s}} \end{pmatrix} \equiv \mathcal{N}\left(\begin{pmatrix} \mathbf{0}_q \\ \mathbf{0}_q \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Psi}_{\widetilde{\boldsymbol{\beta}}} & \boldsymbol{\Psi}_{\boldsymbol{s},\widetilde{\boldsymbol{\beta}}}' \\ \boldsymbol{\Psi}_{\boldsymbol{s},\widetilde{\boldsymbol{\beta}}} & \boldsymbol{\Psi}_{\boldsymbol{s}} \end{pmatrix} \right).$$

Under Assumption A.5, I have $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}$ by the continuous mapping theorem and using (A.20).

For the limiting distribution of $\widehat{\boldsymbol{\beta}}_n$, I have

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) &= \sqrt{n}(\operatorname{diag}[\widehat{\boldsymbol{s}}_n]^{-1} \widehat{\widetilde{\boldsymbol{\beta}}}_n - \operatorname{diag}[\boldsymbol{s}]^{-1} \widetilde{\boldsymbol{\beta}}) \\
&= \operatorname{diag}[\widehat{\boldsymbol{s}}_n]^{-1} \sqrt{n}(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}) + \sqrt{n}(\operatorname{diag}[\widehat{\boldsymbol{s}}_n]^{-1} - \operatorname{diag}[\boldsymbol{s}]^{-1}) \widetilde{\boldsymbol{\beta}} \\
&= \operatorname{diag}[\widehat{\boldsymbol{s}}_n]^{-1} \sqrt{n}(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}) + \operatorname{diag}[\widehat{\boldsymbol{s}}_n]^{-1} \operatorname{diag}[\boldsymbol{s}]^{-1} \operatorname{diag}[\widetilde{\boldsymbol{\beta}}] \sqrt{n}(\boldsymbol{s} - \widehat{\boldsymbol{s}}_n) \\
&\xrightarrow{d} \operatorname{diag}[\boldsymbol{s}]^{-1} \boldsymbol{Z}_{\widetilde{\boldsymbol{\beta}}} - \operatorname{diag}[\boldsymbol{s}]^{-2} \operatorname{diag}[\widetilde{\boldsymbol{\beta}}] \boldsymbol{Z}_{\boldsymbol{s}} \\
&\equiv \operatorname{diag}[\boldsymbol{s}]^{-1} \boldsymbol{Z}_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{H} \boldsymbol{Z}_{\boldsymbol{s}}
\end{aligned}$$

$$\sim \mathcal{N}(\mathbf{0}_q, \mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}), \tag{A.21}$$

where the third line used

$$\begin{aligned}
\left(\text{diag}[\widehat{\boldsymbol{s}}_n]^{-1} - \text{diag}[\boldsymbol{s}]^{-1}\right)\widetilde{\boldsymbol{\beta}} &= \begin{pmatrix} (\widehat{s}_{n,1}^{-1} - s_1^{-1})\widetilde{\beta}_1 \\ \vdots \\ (\widehat{s}_{n,q}^{-1} - s_q^{-1})\widetilde{\beta}_q \end{pmatrix} \\
&= \begin{pmatrix} \frac{s_1 - \widehat{s}_{n,1}}{\widehat{s}_{n,1}s_1}\widetilde{\beta}_1 \\ \vdots \\ \frac{s_q - \widehat{s}_{n,q}}{\widehat{s}_{n,q}s_q}\widetilde{\beta}_q \end{pmatrix} \\
&= \text{diag}\left[\begin{pmatrix} \frac{\widetilde{\beta}_1}{\widehat{s}_{n,1}s_1} \\ \vdots \\ \frac{\widetilde{\beta}_q}{\widehat{s}_{n,q}s_q} \end{pmatrix}\right] \begin{pmatrix} s_1 - \widehat{s}_{n,1} \\ \vdots \\ s_q - \widehat{s}_{n,q} \end{pmatrix},
\end{aligned}$$

the fourth line used Slutsky's theorem, the fifth line defined $\boldsymbol{H} \equiv \text{diag}[\boldsymbol{s}]^{-2}\,\text{diag}[\widetilde{\boldsymbol{\beta}}]$, and the last line defined

$$\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}} \equiv \text{diag}[\boldsymbol{s}]^{-1}\mathbf{\Psi}_{\widetilde{\boldsymbol{\beta}}}\,\text{diag}[\boldsymbol{s}]^{-1} - \text{diag}[\boldsymbol{s}]^{-1}\mathbf{\Psi}'_{\boldsymbol{s},\widetilde{\boldsymbol{\beta}}}\boldsymbol{H}' - \boldsymbol{H}\mathbf{\Psi}_{\boldsymbol{s},\widetilde{\boldsymbol{\beta}}}\,\text{diag}[\boldsymbol{s}]^{-1} + \boldsymbol{H}\mathbf{\Psi}_{\boldsymbol{s}}\boldsymbol{H}'. \tag{A.22}$$

A.4.1 Precision analysis for PCA

To begin with, the analog of (A.18) using the nonstandardized outcomes $\widetilde{\boldsymbol{Y}}_i$ can be written as

$$\widetilde{\boldsymbol{Y}}_i = \widetilde{\boldsymbol{\xi}} + \widetilde{\boldsymbol{\beta}}D_i + \widetilde{\boldsymbol{U}}_i, \tag{A.23}$$

where $\widetilde{\boldsymbol{\beta}} \equiv (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_q)'$, $\widetilde{\boldsymbol{\xi}} \equiv (\widetilde{\xi}_1, \ldots, \widetilde{\xi}_q)'$, and $\widetilde{\boldsymbol{U}}_i \equiv (\widetilde{U}_{i,1}, \ldots, \widetilde{U}_{i,q})'$. I show the analysis on the above model in Appendix A.4.4 after presenting the main results.

Using the linear model (A.23) above and Assumption A.8, I have

$$\text{Var}[\widetilde{\boldsymbol{Y}}_i] = \widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}' p_D(1 - p_D) + \text{Var}[\widetilde{\boldsymbol{U}}_i]. \tag{A.24}$$

Suppose that $\boldsymbol{s} = \mathbf{1}_q$ and $\widetilde{\boldsymbol{\beta}} = \mathbf{0}_q$. Then, (A.22) becomes

$$\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}} \equiv \mathbf{\Psi}_{\widetilde{\boldsymbol{\beta}}}. \tag{A.25}$$

Using the limiting distribution derived in Section A.4.4 without standardizing the outcomes, I have

$$\mathbf{\Psi}_{\widetilde{\boldsymbol{\beta}}} = \frac{\text{Var}[D_i\widetilde{\boldsymbol{U}}_i]}{p_D^2} + \frac{\text{Var}[(1-D_i)\widetilde{\boldsymbol{U}}_i]}{(1-p_D)^2}. \tag{A.26}$$

Define $\mathbf{\Sigma}_{\widetilde{U},1} \equiv \text{Var}[\widetilde{\boldsymbol{U}}_i|D_i=1]$ and $\mathbf{\Sigma}_{\widetilde{U},0} \equiv \text{Var}[\widetilde{\boldsymbol{U}}_i|D_i=0]$. Using D_i is binary, I have

$$\text{Var}[D_i\widetilde{\boldsymbol{U}}_i] = \mathbb{E}[D_i^2\widetilde{\boldsymbol{U}}_i\widetilde{\boldsymbol{U}}_i'] - \mathbb{E}[D_i\widetilde{\boldsymbol{U}}_i]\mathbb{E}[D_i\widetilde{\boldsymbol{U}}_i]' = \mathbb{E}[D_i\widetilde{\boldsymbol{U}}_i\widetilde{\boldsymbol{U}}_i'] = p_D\mathbf{\Sigma}_{\widetilde{U},1},$$

and

$$\text{Var}[(1-D_i)\widetilde{\boldsymbol{U}}_i] = \mathbb{E}[(1-D_i)^2\widetilde{\boldsymbol{U}}_i\widetilde{\boldsymbol{U}}_i'] - \mathbf{0}_q\mathbf{0}_q' = \mathbb{E}[(1-D_i)\widetilde{\boldsymbol{U}}_i\widetilde{\boldsymbol{U}}_i'] = (1-p_D)\mathbf{\Sigma}_{\widetilde{U},0}.$$

Under homoskedasticity so that $\mathbf{\Sigma}_{\widetilde{U},0} = \mathbf{\Sigma}_{\widetilde{U},1} \equiv \mathbf{\Sigma}_{\widetilde{U}}$, I have

$$\begin{aligned}
\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}} &= \mathbf{\Psi}_{\widetilde{\boldsymbol{\beta}}} \\
&= \frac{\mathbf{\Sigma}_{\widetilde{U}}}{p_D} + \frac{\mathbf{\Sigma}_{\widetilde{U}}}{1-p_D} \\
&= \frac{1}{p_D(1-p_D)}\mathbf{\Sigma}_{\widetilde{U}} \\
&= \frac{1}{p_D(1-p_D)}\text{Var}[\widetilde{\boldsymbol{Y}}_i] \\
&= \frac{1}{p_D(1-p_D)}\text{Var}[\boldsymbol{Y}_i],
\end{aligned} \tag{A.27}$$

where the first line uses (A.25), the second line uses the homoskedastic assumption, the fourth line uses (A.24), and the last line uses $\boldsymbol{s}$ is the standard deviations of $\boldsymbol{Y}_i$, so it is the same as correlation matrix. This shows that PCA is even trying to maximize the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ instead of minimizing it because $\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ is proportional to $\text{Var}[\boldsymbol{Y}_i]$.

The assumptions that $\widetilde{\boldsymbol{\beta}} = \mathbf{0}_q$, $\boldsymbol{s} = \mathbf{1}_q$, and homoskedasticity are used to simplify the analysis. Without such analysis, one has to consider the other covariance terms in Assumption A.5. But the asymptotic variance (A.22) can be used to analyze the variance.

Next, I consider the heteroskedastic case. Here, (A.24) can be written as

$$\mathbf{\Sigma}_Y = \widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}'p_D(1-p_D) + \mathbf{\Sigma}_{\widetilde{U}} = \widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}'p_D(1-p_D) + p_D\mathbf{\Sigma}_{\widetilde{U},1} + (1-p_D)\mathbf{\Sigma}_{\widetilde{U},0}. \tag{A.28}$$

On the other hand, (A.26) becomes $\frac{\mathbf{\Sigma}_{\widetilde{U},1}}{p_D} + \frac{\mathbf{\Sigma}_{\widetilde{U},0}}{1-p_D}$, so that (A.27) becomes

$$\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \mathbf{\Psi}_{\widetilde{\boldsymbol{\beta}}} = \frac{\mathbf{\Sigma}_{\widetilde{U},1}}{p_D} + \frac{\mathbf{\Sigma}_{\widetilde{U},0}}{1-p_D} \tag{A.29}$$

under $\boldsymbol{\beta} = \mathbf{0}_q$ and $\boldsymbol{s} = \mathbf{1}_q$.

Combining the above, I have

$$p_D(1-p_D)\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}} - \mathbf{\Sigma}_Y = (1-2p_D)\mathbf{\Sigma}_{\widetilde{U},1} - (1-2p_D)\mathbf{\Sigma}_{\widetilde{U},0} - \widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}' p_D(1-p_D).$$

If $\widetilde{\boldsymbol{\beta}} = \mathbf{0}_q$, then

$$p_D(1-p_D)\boldsymbol{w}'\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w} - \boldsymbol{w}'\mathbf{\Sigma}_Y\boldsymbol{w} = (1-2p_D)(\boldsymbol{w}'\mathbf{\Sigma}_{\widetilde{U},1}\boldsymbol{w} - \boldsymbol{w}'\mathbf{\Sigma}_{\widetilde{U},0}\boldsymbol{w}). \tag{A.30}$$

The difference between $\boldsymbol{w}'\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}$ and $\boldsymbol{w}'\mathbf{\Sigma}_Y\boldsymbol{w}$ now depends on p_D, $\mathbf{\Sigma}_{\widetilde{U},1}$ and $\mathbf{\Sigma}_{\widetilde{U},0}$.

A.4.2 Precision analysis for IVM

The analysis in subsection A.4.1 can be used to analyze the IVM weights, although the interpretations of the variables are different. This is because Anderson (2008) standardizes the outcomes using the control group's standard deviations. Therefore, $\boldsymbol{s}$ represents the standard deviation of the outcomes using the control sample and $\widehat{\boldsymbol{s}}_n$ is the estimator of $\boldsymbol{s}$. The terms in (A.19) and (A.20) in this subsection are under this interpretation.

Under homoskedasticity, $\boldsymbol{s} = \mathbf{1}_q$, and $\widetilde{\boldsymbol{\beta}} = \mathbf{0}_q$, the analysis in (A.27) can be applied. It provides a way to justify that minimizing $\boldsymbol{w}'\mathbf{\Sigma}_Y\boldsymbol{w}$ and $\boldsymbol{w}'\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}$ lead to the same solution over $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$.

Under heteroskedasticity, $\boldsymbol{s} = \mathbf{1}_q$, and $\widetilde{\boldsymbol{\beta}} = \mathbf{0}_q$, the analysis in (A.30) can be applied. Since the RHS of (A.30) depends on $\boldsymbol{w}$, minimizing $\boldsymbol{w}'\mathbf{\Sigma}_Y\boldsymbol{w}$ and $\boldsymbol{w}'\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}$ can lead to different solutions. Nevertheless, when $p_D = \frac{1}{2}$, both can still lead to the same solution.

In the following, I compare the IVM weights with the weights that maximize precision.

Proposition A.6. *Let Assumption A.5 hold and $\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ as defined in (A.22). Write $\boldsymbol{w}_{\text{ivm}}$ as in (12) and $\boldsymbol{w}_{\text{b}} \equiv \arg\min_{\boldsymbol{w}\in\mathcal{W}_{\text{sto}}} \boldsymbol{w}'\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}$. Assume that $\mathbf{\Sigma}_Y$ and $\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ are positive definite matrices. Then, $\boldsymbol{w}_{\text{ivm}} = \boldsymbol{w}_{\text{b}}$ if and only if $\mathbf{\Sigma}_Y^{-1}\mathbf{1}_q = \kappa\mathbf{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q$ for some $\kappa \neq 0$.*

Proof of Proposition A.6. $\boldsymbol{w}_{\text{ivm}}$ has been given in (11). By a similar computation, $\boldsymbol{w}_{\text{b}} =$

$\frac{\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}$. If $\boldsymbol{w}_{\text{ivm}} = \boldsymbol{w}_{\text{b}}$, it means that

$$\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q = \left(\frac{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}\right)\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q.$$

The scalar $\frac{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}$ is well-defined and nonzero because $\boldsymbol{\Sigma}_Y$ and $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ are positive definite.

On the other hand, if $\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q = \kappa\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q$ for some $\kappa \neq 0$, then

$$\boldsymbol{w}_{\text{ivm}} = \frac{\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q} = \frac{\kappa\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}{\kappa\mathbf{1}_q'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q} = \frac{\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q} = \boldsymbol{w}_{\text{b}}.$$

Hence, the proof is complete. □

A.4.3 Precision analysis for SA

The following proposition characterizes when SA maximizes precision. For the same reason as in Section 3.3.1, I consider $\boldsymbol{w} \in \mathcal{W}_{\text{sto}}$.

Proposition A.7. *Consider the same notations and assumptions as in Proposition A.6. Write $\boldsymbol{w}_{\text{sa}}$ as in Section 3.3. and $\boldsymbol{w}_{\text{b}} \equiv \arg\min_{\boldsymbol{w}\in\mathcal{W}_{\text{sto}}} \boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}$. Assume that $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ is a positive definite matrix. Then, $\boldsymbol{w}_{\text{sa}} = \boldsymbol{w}_{\text{b}}$ if and only if $\mathbf{1}_q = \kappa\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q$ for some $\kappa \neq 0$.*

Proof of Proposition A.7. $\boldsymbol{w}_{\text{b}}$ has been given in the proof of Proposition A.6. If $\boldsymbol{w}_{\text{sa}} = \boldsymbol{w}_{\text{b}}$, it means that

$$\mathbf{1}_q = q\frac{\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}.$$

The scalar $\frac{q}{\mathbf{1}_q'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}$ is well-defined and nonzero because $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ is positive definite.

On the other hand, if $\mathbf{1}_q = \kappa\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q$ for some $\kappa \neq 0$, then

$$\boldsymbol{w}_{\text{sa}} = \frac{1}{q}\mathbf{1}_q = \frac{1}{\mathbf{1}_q'\mathbf{1}_q}\mathbf{1}_q = \frac{\kappa\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}{\kappa\mathbf{1}_q'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q} = \frac{\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{1}_q} = \boldsymbol{w}_{\text{b}}.$$

Hence, the proof is complete. □

A.4.4 Precision analysis without standardization

I assume that the following hold for the linear model (A.23).

Assumption A.8.

(a) $p_D \equiv \mathbb{P}[D_i = 1] \in (0,1)$.

(b) $\mathbb{E}[\widetilde{\boldsymbol{U}}_i | D_i] = \mathbf{0}_q$.

(c) $\mathrm{Var}[\widetilde{\boldsymbol{U}}_i] < \infty$.

(d) $\mathrm{Var}[\widetilde{\boldsymbol{Y}}_i]$ *and* $\mathrm{Var}[\widetilde{\boldsymbol{U}}_i]$ *are symmetric positive definite matrices with bounded eigenvalues.*

In the above, (a) requires D_i to have variation in data. Part (b) imposes the standard independence assumption. Part (c) ensures that $\mathrm{Var}[\widetilde{\boldsymbol{Y}}_i]$ is finite. Part (d) requires the matrices to be positive definite.

Proof of equation (A.26). Consider the linear model (A.23). Note that from (A.23), the estimator on the treatment effect for the jth outcome can be written as

$$\begin{aligned}\widehat{\widetilde{\beta}}_{n,j} &= \frac{1}{n_1}\sum_{i=1}^{n} D_i(\widetilde{\xi}_j + \widetilde{\beta}_j D_i + \widetilde{U}_{i,j}) - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)(\widetilde{\xi}_j + \widetilde{\beta}_j D_i + \widetilde{U}_{i,j}) \\ &= \widetilde{\beta}_j + \frac{1}{n_1}\sum_{i=1}^{n} D_i \widetilde{U}_{i,j} - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)\widetilde{U}_{i,j},\end{aligned}$$

for $j = 1, \ldots, q$, where $n_1 \equiv \sum_{i=1}^{n} D_i$ and $n_0 = \sum_{i=1}^{n}(1-D_i)$.

I have $\widehat{\widetilde{\tau}}_n = \boldsymbol{w}'\widehat{\widetilde{\boldsymbol{\beta}}}_n$ by Lemma A.2. The estimator $\widehat{\widetilde{\tau}}_n$ for $\widetilde{\tau}$ defined in Example 3.5 can be written as

$$\begin{aligned}\widehat{\widetilde{\tau}}_n &= \sum_{j=1}^{q} w_j \widetilde{\beta}_j + \sum_{j=1}^{q} w_j \left[\frac{1}{n_1}\sum_{i=1}^{n} D_i \widetilde{U}_{i,j} - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)\widetilde{U}_{i,j}\right] \\ &= \sum_{j=1}^{q} w_j \widetilde{\beta}_j + \frac{1}{n_1}\sum_{i=1}^{n} D_i \sum_{j=1}^{q} w_j \widetilde{U}_{i,j} - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)\sum_{j=1}^{q} w_j \widetilde{U}_{i,j}. \qquad \text{(A.31)}\end{aligned}$$

Using (A.31), recentering $\widehat{\widetilde{\tau}}_n$ around $\sum_{j=1}^{q} w_j \widetilde{\beta}_j$ and scaling by $\sqrt{n}$ gives

$$\begin{aligned}\sqrt{n}\left(\widehat{\widetilde{\tau}}_n - \sum_{j=1}^{q} w_j \widetilde{\beta}_j\right) &= \frac{n}{n_1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \sum_{j=1}^{q} w_j \widetilde{U}_{i,j} - \frac{n}{n_0}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-D_i)\sum_{j=1}^{q} w_j \widetilde{U}_{i,j} \\ &= \begin{pmatrix} \frac{n}{n_1} & -\frac{n}{n_0}\end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \sum_{j=1}^{q} w_j \widetilde{U}_{i,j} \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-D_i)\sum_{j=1}^{q} w_j \widetilde{U}_{i,j}\end{pmatrix}\end{aligned}$$

$$= \begin{pmatrix} \frac{n}{n_1} & -\frac{n}{n_0} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^n D_i(\boldsymbol{w}'\widetilde{\boldsymbol{U}}_i) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n (1-D_i)(\boldsymbol{w}'\widetilde{\boldsymbol{U}}_i) \end{pmatrix}$$
$$\xrightarrow{d} \begin{pmatrix} \frac{1}{p_D} & -\frac{1}{1-p_D} \end{pmatrix} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega_{U,1} & \omega_{U,10} \\ \omega_{U,10} & \omega_{U,0} \end{pmatrix} \right), \tag{A.32}$$

where the last line follows from the Slutsky's theorem, the central limit theorem, Assumption A.8, and defined

$$\begin{aligned} \omega_{U,1} &\equiv \operatorname{Var}[D_i(\boldsymbol{w}'\widetilde{\boldsymbol{U}}_i)] = \boldsymbol{w}' \operatorname{Var}[D_i\widetilde{\boldsymbol{U}}_i]\boldsymbol{w}, \\ \omega_{U,0} &\equiv \operatorname{Var}[(1-D_i)(\boldsymbol{w}'\widetilde{\boldsymbol{U}}_i)] = \boldsymbol{w}' \operatorname{Var}[(1-D_i)\widetilde{\boldsymbol{U}}_i]\boldsymbol{w}, \\ \omega_{U,10} &\equiv \operatorname{Cov}[D_i(\boldsymbol{w}'\widetilde{\boldsymbol{U}}_i), (1-D_i)(\boldsymbol{w}'\widetilde{\boldsymbol{U}}_i)] = 0, \end{aligned}$$

where the computation of ω_{10} used $D_i(1-D_i) = 0$ since D_i is binary and that $\mathbb{E}[\widetilde{\boldsymbol{U}}_i|D_i] = 0$ from Assumption A.8(c).

It follows that the asymptotic variance in (A.32) can be written as

$$\sigma_\tau^2 = \frac{1}{p_D^2}\omega_{U,1} + \frac{1}{(1-p_D)^2}\omega_{U,0} = \boldsymbol{w}' \left\{ \frac{\operatorname{Var}[D_i\widetilde{\boldsymbol{U}}_i]}{p_D^2} + \frac{\operatorname{Var}[(1-D_i)\widetilde{\boldsymbol{U}}_i]}{(1-p_D)^2} \right\} \boldsymbol{w}.$$

□

A.5 Analysis on common aggregation methods and the *t*-statistic

A.5.1 Setup

This section begins by studying the t-statistic under the setup in Example 3.5. As in Assumption 2.2 and Appendix A.4, the vector $\boldsymbol{Y}_i$ represents the suitably standardized outcomes and $\widetilde{\boldsymbol{Y}}_i$ represents the nonstandardized outcomes. The linear model for $\widetilde{\boldsymbol{Y}}_i$ is given in (A.23) and the linear model for $\boldsymbol{Y}_i$ is given in (A.18). Let $\widehat{\widetilde{\boldsymbol{\xi}}}_n$ and $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ be the estimators of $\widetilde{\boldsymbol{\xi}}$ and $\widetilde{\boldsymbol{\beta}}$ respectively.

To analyze the t-statistic, I need to specify the linear model and the hypothesis. I specialize the analysis to the following as in Example 3.5, where the outcome is $\boldsymbol{w}'\boldsymbol{Y}_i$:

$$\boldsymbol{w}'\boldsymbol{Y}_i = \overline{\xi} + \tau D_i + \overline{U}_i, \tag{A.33}$$

where $\boldsymbol{Y}_i = \widehat{\mathbf{S}}_n\widetilde{\boldsymbol{Y}}_i$ and $\widehat{\mathbf{S}}_n \equiv \operatorname{diag}[\widehat{s}_n]^{-1}$. In addition, I write $\mathbf{S} \equiv \operatorname{diag}[s]^{-1}$. The above equation is related to (A.18) by writing $\overline{\xi} = \boldsymbol{w}'\boldsymbol{\xi}$, $\tau = \boldsymbol{w}'\boldsymbol{\beta}$, and $\overline{U}_i = \boldsymbol{w}'\boldsymbol{U}_i$. In the

discussion below, the definitions of $\widehat{\mathbf{S}}_n$ and $\mathbf{S}$ depend on the method used (as described by Assumption 2.2) as discussed in the main text.

Consider testing the following hypothesis

$$H_0 : \tau = 0 \quad \text{vs.} \quad H_1 : \tau \neq 0. \tag{A.34}$$

Let $\widehat{T}_n(\boldsymbol{w})$ be the sample t-statistic used to test the above hypothesis. Under the setup in Example 3.5 with a binary treatment and homoskedastic errors, the t-statistic squared can be written as

$$[\widehat{T}_n(\boldsymbol{w})]^2 \equiv (n-2)\frac{\widehat{R}_n^2(\boldsymbol{w})}{1-\widehat{R}_n^2(\boldsymbol{w})}, \tag{A.35}$$

where $\widehat{R}_n^2(\boldsymbol{w})$ is the sample analog of the population R-squared (Wooldridge, 2013, equation (4.41)). The t-statistic and R-squared are functions of $\boldsymbol{w}$ because the outcome in (A.33) depends on $\boldsymbol{w}$.

A.5.2 Comparison with PCA, SA, and IVM

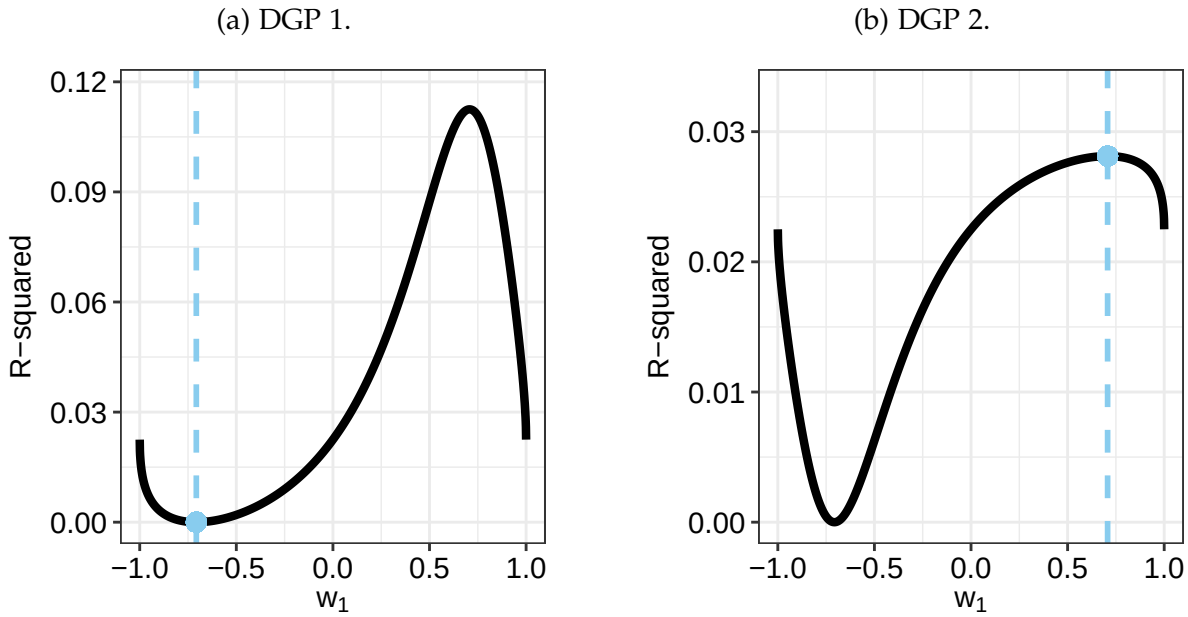
The following proposition shows that the t-statistic squared is maximized by the PC1 of $\boldsymbol{\Sigma}_Y$ under a specific condition. Since (A.35) has a factor of n in the expression, I consider the following ratio of t-statistic for a fixed $\boldsymbol{w}_0$ to have a well-defined probability limit:

$$\Upsilon(\boldsymbol{w};\boldsymbol{w}_0) \equiv \operatorname*{plim}_{n\to\infty} \frac{[\widehat{T}_n(\boldsymbol{w})]^2}{[\widehat{T}_n(\boldsymbol{w}_0)]^2} = \frac{R^2(\boldsymbol{w})}{1-R^2(\boldsymbol{w})}\frac{1}{\frac{R^2(\boldsymbol{w}_0)}{1-R^2(\boldsymbol{w}_0)}}. \tag{A.36}$$

Proposition A.9. *Let Assumptions 3.1, A.5, and A.8 hold. Suppose that $\boldsymbol{\beta} \neq \mathbf{0}_q$. Consider testing (A.34) in the linear model (A.33). Let $\Upsilon(\boldsymbol{w};\boldsymbol{w}_0)$ be the ratio defined in (A.36) for any $\boldsymbol{w}_0 \in \mathbb{R}^q$ such that $\|\boldsymbol{w}_0\|_2 = 1$ and $R^2(\boldsymbol{w}_0) \in (0,1)$. Then, the PC1 of $\boldsymbol{\Sigma}_Y$ solves $\max_{\boldsymbol{w}:\|\boldsymbol{w}\|_2=1} \Upsilon(\boldsymbol{w};\boldsymbol{w}_0)$ if and only if it is parallel to $\boldsymbol{\beta}$.*

In the above, choosing $\boldsymbol{w}_0$ such that $R^2(\boldsymbol{w}_0) \neq 1$ is to ensure the ratio $\frac{R^2(\boldsymbol{w}_0)}{1-R^2(\boldsymbol{w}_0)}$ in (A.36) is well-defined in the probability limit. Proposition A.9 implies that finding the weights $\boldsymbol{w}$ to maximize the square of t-statistic depends on both $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_Y$. This can be seen from the one-to-one relationship between $[\widehat{T}_n(\boldsymbol{w})]^2$ and $\widehat{R}_n^2(\boldsymbol{w})$ in (A.35) such that maximizing t-statistic squared is the same as maximizing R-squared. In the binary treatment case, R-squared is maximized if the means of $\boldsymbol{w}'\boldsymbol{Y}_i$ for the $D_i = 1$ and $D_i = 0$ populations are as separate as possible. Thus, this means that $\boldsymbol{\beta}$ matters because a larger $(\boldsymbol{w}'\boldsymbol{\beta})^2$ increases separation.

Figure A.1: A two-outcome numerical example that evaluates the R-squared of PCA.



Notes: The above plots the R-squared of the aggregated treatment effect against w_1. See Example A.10 for the data generating process. The vertical dashed lines represent the choice made by running PCA on the outcomes in both DGPs.

Example A.10. This numerical example demonstrates Proposition A.9 and shows that PCA can maximize or minimize the t-statistic squared (or equivalently, the R-squared).

Following the same notations in Example 3.5, consider the following DGPs:

DGP 1. Same DGP as Example 3.5, i.e., $\text{Var}[Y_{i,1}] = \text{Var}[Y_{i,2}] = 1$, $\beta_1 = \beta_2 = 0.5$, $p_D = 0.1$, and $\text{Cov}[Y_{i,1}, Y_{i,2}] = -0.6$.

DGP 2. Same as DGP 1, except that $\text{Cov}[Y_{i,1}, Y_{i,2}] = 0.6$.

Figure A.1 shows the (population) R-squared when $w_1 Y_{i,1} + w_2 Y_{i,2}$ is used as the outcome. As noted in (A.35), there is a one-to-one relationship between t-statistic squared and R-squared. PCA chooses a different w_1 under the two DGPs because the leading eigenvector is determined by $\text{Cov}[Y_{i,1}, Y_{i,2}]$ and the two DGPs differ by the sign of $\text{Cov}[Y_{i,1}, Y_{i,2}]$. The PC1 is $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ when $\text{Cov}[Y_{i,1}, Y_{i,2}] > 0$ and $(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ when $\text{Cov}[Y_{i,1}, Y_{i,2}] < 0$.

Thus, PCA can maximize or minimize the t-statistic (or equivalently, the R-squared) under the DGPs. △

Remark A.11. Maximizing $R^2(\boldsymbol{w})$ (or t-statistic squared in the homoskedastic case) is related to linear discriminant analysis (LDA). LDA is a technique that classifies observa-

tions into different groups using a linear combination of the features (see, for instance, Hastie et al. (2009) or Mardia et al. (2024)). This is related to the context of aggregating treatment effects when $\boldsymbol{Y}_i$ is viewed as "features" and D_i is viewed as "labels." Then, $R^2(\boldsymbol{w})$ is related to Fisher's linear discriminant (with the caveat on the difference in the definitions of the variance matrices in both quantities). The connection and similarity of LDA and $R^2(\boldsymbol{w})$ highlights the importance of using D_i if one wants to maximize t-statistic (improves power). Here, running PCA on $\boldsymbol{Y}_i$ is like unsupervised learning, and maximizing t-statistic is like supervised learning. ■

Studying when the IVM and SA weights maximize (A.36) can be analyzed in a manner similar to the PCA approach. The result is summarized in the propositions as follows. Similar to the PCA results, they show that IVM and SA do not automatically maximize the t-statistic squared. The assumption $\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta} \neq 0$ is used below to assume a nonzero denominator.

Proposition A.12. *Let Assumptions A.5 and A.8 hold. Suppose that $\boldsymbol{\beta} \neq \mathbf{0}_q$ and $\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta} \neq 0$. Consider testing (A.34) in the linear model (A.33). Let $\Upsilon(\boldsymbol{w};\boldsymbol{w}_0)$ be the ratio defined in (A.36) for any $\boldsymbol{w}_0 \in \mathcal{W}_{\text{sto}}$ and $R^2(\boldsymbol{w}_0) \in (0,1)$. Then, $\boldsymbol{w}_{\text{ivm}}$ solves $\max_{\boldsymbol{w}\in\mathcal{W}_{\text{sto}}} \Upsilon(\boldsymbol{w};\boldsymbol{w}_0)$ if and only if $\boldsymbol{\beta}$ is parallel to $\mathbf{1}_q$.*

Proposition A.13. *Consider the same assumptions, the testing problem, and the ratio $\Upsilon(\boldsymbol{w};\boldsymbol{w}_0)$ as in Proposition A.12. Then, $\boldsymbol{w}_{\text{sa}}$ solves $\max_{\boldsymbol{w}\in\mathcal{W}_{\text{sto}}} \Upsilon(\boldsymbol{w};\boldsymbol{w}_0)$ if and only if it is parallel to $\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}$.*

A.5.3 Proofs

Lemma A.14. *Let Assumptions A.5 and A.8 hold. Consider the linear model (A.33). Let $\widehat{R}_n^2(\boldsymbol{w})$ be as defined in (A.35). For any $\boldsymbol{w} \neq \mathbf{0}_q$,*

$$\widehat{R}_n^2(\boldsymbol{w}) \xrightarrow{p} R^2(\boldsymbol{w}) \equiv \frac{\text{Var}[D_i](\boldsymbol{w}'\boldsymbol{\beta})^2}{\boldsymbol{w}'\boldsymbol{\Sigma}_Y\boldsymbol{w}}.$$

Proof of Lemma A.14. Let $\overline{y}_n(\boldsymbol{w}) \equiv \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}'\boldsymbol{Y}_i$ be the average of the weighted outcomes, $\widehat{\text{TSS}}_n(\boldsymbol{w}) \equiv \sum_{i=1}^n (\boldsymbol{w}'\boldsymbol{Y}_i - \overline{y}_n(\boldsymbol{w}))^2$ be the total sum of squares, and $\widehat{\text{ESS}}_n(\boldsymbol{w}) \equiv \sum_{i=1}^n (\widehat{\xi}_n + \widehat{\tau}_n D_i - \overline{y}_n(\boldsymbol{w}))^2$ be the explained sum of squares where $\widehat{\xi}_n$ is the estimator of ξ. Hence,

$$\widehat{R}_n^2(\boldsymbol{w}) = \frac{\widehat{\text{ESS}}_n(\boldsymbol{w})}{\widehat{\text{TSS}}_n(\boldsymbol{w})}.$$

But from the linear model, $\widehat{\xi}_n + \widehat{\tau}_n D_i = \overline{y}_n(\boldsymbol{w}) - \widehat{\tau}_n \overline{D}_n + \widehat{\tau}_n D_i = \overline{y}_n(\boldsymbol{w}) + \widehat{\tau}_n(D_i - \overline{D}_n)$

where $\overline{D}_n \equiv \frac{1}{n}\sum_{i=1}^n D_i$. Hence, $\widehat{\text{ESS}}_n(\boldsymbol{w}) = \widehat{\tau}_n^2 \sum_{i=1}^n (D_i - \overline{D}_n)^2$. Note that

$$\widehat{\tau}_n = \frac{\sum_{i=1}^n (D_i - \overline{D}_n)(\boldsymbol{w}'\boldsymbol{Y}_i - \overline{y}_n(\boldsymbol{w}))}{\sum_{i=1}^n (D_i - \overline{D}_n)^2}.$$

This gives

$$\widehat{R}_n^2(\boldsymbol{w}) \equiv \frac{[\sum_{i=1}^n (D_i - \overline{D}_n)(\boldsymbol{w}'\boldsymbol{Y}_i - \overline{y}_n(\boldsymbol{w}))]^2}{[\sum_{i=1}^n (D_i - \overline{D}_n)^2]\widehat{\text{TSS}}_n(\boldsymbol{w})}. \tag{A.37}$$

Using the notations defined in the beginning of Section A.5, I can write $\overline{\widetilde{\boldsymbol{Y}}}_n \equiv \frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{Y}}_i$, $\overline{y}_n(\boldsymbol{w}) = \boldsymbol{w}'\widehat{\mathbf{S}}_n(\frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{Y}}_i) = \boldsymbol{w}'\widehat{\mathbf{S}}_n\overline{\widetilde{\boldsymbol{Y}}}_n$,

$$\widehat{\text{TSS}}_n(\boldsymbol{w}) \equiv \sum_{i=1}^n (\boldsymbol{w}'\widehat{\mathbf{S}}_n\widetilde{\boldsymbol{Y}}_i - \overline{y}_n(\boldsymbol{w}))^2 = \boldsymbol{w}'\widehat{\mathbf{S}}_n\left[\sum_{i=1}^n (\widetilde{\boldsymbol{Y}}_i - \overline{\widetilde{\boldsymbol{Y}}}_n)(\widetilde{\boldsymbol{Y}}_i - \overline{\widetilde{\boldsymbol{Y}}}_n)'\right]\widehat{\mathbf{S}}_n'\boldsymbol{w}, \tag{A.38}$$

and

$$\sum_{i=1}^n (D_i - \overline{D}_n)(\boldsymbol{w}'\widehat{\mathbf{S}}_n\widetilde{\boldsymbol{Y}}_i - \overline{y}_n(\boldsymbol{w})) = \boldsymbol{w}'\widehat{\mathbf{S}}_n\sum_{i=1}^n (D_i - \overline{D}_n)(\widetilde{\boldsymbol{Y}}_i - \overline{\widetilde{\boldsymbol{Y}}}_n). \tag{A.39}$$

Let $\boldsymbol{B}_{n,\widetilde{\boldsymbol{Y}}} \equiv \sum_{i=1}^n (\widetilde{\boldsymbol{Y}}_i - \overline{\widetilde{\boldsymbol{Y}}}_n)(\widetilde{\boldsymbol{Y}}_i - \overline{\widetilde{\boldsymbol{Y}}}_n)'$, $\boldsymbol{B}_{n,\widetilde{\boldsymbol{Y}},D} \equiv \sum_{i=1}^n (D_i - \overline{D}_n)(\widetilde{\boldsymbol{Y}}_i - \overline{\widetilde{\boldsymbol{Y}}}_n)$, and $B_{n,D} \equiv \sum_{i=1}^n (D_i - \overline{D}_n)^2$. Together with (A.37) to (A.39),

$$\widehat{R}_n^2(\boldsymbol{w}) = \frac{\boldsymbol{w}'\widehat{\mathbf{S}}_n\boldsymbol{B}_{n,\widetilde{\boldsymbol{Y}},D}\boldsymbol{B}_{n,\widetilde{\boldsymbol{Y}},D}'\widehat{\mathbf{S}}_n\boldsymbol{w}}{B_{n,D}(\boldsymbol{w}'\widehat{\mathbf{S}}_n\boldsymbol{B}_{n,\widetilde{\boldsymbol{Y}}}\widehat{\mathbf{S}}_n\boldsymbol{w})}. \tag{A.40}$$

Under the given assumptions of this lemma, I have $\widehat{\mathbf{S}}_n \xrightarrow{p} \mathbf{S}$, $\frac{1}{n}B_{n,D} \xrightarrow{p} \text{Var}[D_i]$, $\frac{1}{n}\boldsymbol{B}_{n,\widetilde{\boldsymbol{Y}}} \xrightarrow{p} \text{Var}[\widetilde{\boldsymbol{Y}}_i]$, and $\frac{1}{n}\boldsymbol{B}_{n,\widetilde{\boldsymbol{Y}},D} \xrightarrow{p} \text{Cov}[D_i, \widetilde{\boldsymbol{Y}}_i] = \widetilde{\boldsymbol{\beta}}\,\text{Var}[D_i]$ by the continuous mapping theorem. Dividing the numerator and denominator of (A.40) by n^2, I have

$$\widehat{R}_n^2(\boldsymbol{w}) \xrightarrow{p} \frac{\text{Var}[D_i]^2\boldsymbol{w}'\mathbf{S}\widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\beta}}\mathbf{S}'\boldsymbol{w}}{\text{Var}[D_i](\boldsymbol{w}'\mathbf{S}\,\text{Var}[\widetilde{\boldsymbol{Y}}_i]\mathbf{S}\boldsymbol{w})} = \frac{\text{Var}[D_i](\boldsymbol{w}'\boldsymbol{\beta})^2}{\boldsymbol{w}'\boldsymbol{\Sigma}_Y\boldsymbol{w}} \tag{A.41}$$

by (A.20), continuous mapping theorem, and by the definition of $\boldsymbol{\Sigma}_Y$. □

Proposition A.15. *Let Assumptions 3.1, A.5 and A.8 hold. In addition, suppose $\boldsymbol{\beta} \neq \mathbf{0}_q$. Consider the linear model in (A.33) and the $R^2(\boldsymbol{w})$ defined in Lemma A.14.*

(a) The optimal solution to

$$\max_{\boldsymbol{w}:\|\boldsymbol{w}\|_2^2=1} R^2(\boldsymbol{w})$$

is given by $\pm \boldsymbol{w}_{\mathrm{R}}$*, where*

$$\boldsymbol{w}_{\mathrm{R}} \equiv \frac{\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta}}}.$$

(b) $\boldsymbol{w}_{\mathrm{R}} = \pm\boldsymbol{\nu}_1$ *if and only if* $\boldsymbol{\nu}_1$ *is parallel to* $\boldsymbol{\beta}$*, where* $\boldsymbol{\nu}_1$ *is the unit-length leading eigenvector of* $\boldsymbol{\Sigma}_Y$ *defined in the beginning of Section 3.1.1.*

Proof of Proposition A.15(a). Write $\sigma_D^2 \equiv \mathrm{Var}[D_i]$, then by Lemma A.14:

$$R^2(\boldsymbol{w}) = \sigma_D^2 \frac{\boldsymbol{w}'(\boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{w}}{\boldsymbol{w}'\boldsymbol{\Sigma}_Y\boldsymbol{w}}. \tag{A.42}$$

The above is a generalized Rayleigh quotient and is homogeneous of degree 0. Since $p_D \in (0,1)$ from Assumption A.8, it follows that $\sigma_D^2 = p_D(1-p_D) > 0$. It is equivalent to a generalized eigenvalue problem, and the optimal solution satisfies (Ghojogh et al., 2023, Section 4.3):

$$[\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{\beta}\boldsymbol{\beta}')]\boldsymbol{w} = \lambda\boldsymbol{w}, \tag{A.43}$$

because $\boldsymbol{\Sigma}_Y$ is invertible and that $\boldsymbol{\beta} \neq \mathbf{0}_q$ as assumed in the proposition. This amounts to finding the leading eigenvector of the matrix $\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{\beta}\boldsymbol{\beta}')$.

Now returning to (A.43), suppose that $\lambda \neq 0$. This means $\boldsymbol{w}$ is parallel to $\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}$ because $\boldsymbol{\beta}'\boldsymbol{w}$ is a scalar. Let $\boldsymbol{w} = t\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}$ for some nonzero $t \in \mathbb{R}$. Since $\boldsymbol{w}$ is restricted to have unit length, this means $t^2\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta} = \|\boldsymbol{w}\|_2^2 = 1$. Thus, $t = \pm(\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta})^{-\frac{1}{2}}$. Note that $t \neq 0$ because $\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta} = \|\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}\|_2^2$ and by positive definiteness of $\boldsymbol{\Sigma}_Y$, $\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta} \neq \mathbf{0}_q$. This means $\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta} > 0$. Recall that the generalized Rayleigh quotient is homogeneous of degree 0. Therefore, $\boldsymbol{w}^\star = \pm\frac{\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta}}}$.

Since $\boldsymbol{w}$ is assumed to have unit length, multiplying $\boldsymbol{w}'$ on both sides of (A.43) gives

$$\lambda = \boldsymbol{w}'[\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{\beta}\boldsymbol{\beta}')]\boldsymbol{w}.$$

Evaluating the above at $\boldsymbol{w} = \boldsymbol{w}^\star$ gives the corresponding eigenvalue $\lambda^\star \equiv \boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}$. Note that $\lambda^\star > 0$ because $\boldsymbol{\beta}$ is a nonzero vector by the hypothesis of this proposition and $\boldsymbol{\Sigma}_Y$ is positive definite by Assumptions A.5 and A.8(d). Therefore, the eigenvalue associated with the eigenvector $\boldsymbol{w}^\star$ is positive.

Note that $\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{\beta}\boldsymbol{\beta}')$ has rank 1. This follows by first noting that $\boldsymbol{\Sigma}_Y$ is a full rank matrix by Assumptions A.5 and A.8(d) and $\boldsymbol{\beta}\boldsymbol{\beta}'$ has rank 1. The rank of $\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{\beta}\boldsymbol{\beta}')$ is

bounded above by the following inequality (Meyer, 2023, Chapter 4):

$$\text{rank}[\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{\beta}\boldsymbol{\beta}')] \leq \min\{\text{rank}[\boldsymbol{\Sigma}_Y^{-1}], \text{rank}[\boldsymbol{\beta}\boldsymbol{\beta}']\} = 1. \tag{A.44}$$

Since $\boldsymbol{\beta} \neq \mathbf{0}_q$ is assumed in the proposition, $\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta} \neq \mathbf{0}_q$ because $\boldsymbol{\Sigma}_Y^{-1}$ has full rank. Therefore, $\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{\beta}\boldsymbol{\beta}')$ cannot be a zero matrix and hence, it cannot have rank 0. It follows that $\text{rank}[\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{\beta}\boldsymbol{\beta}')] = 1$ by (A.44), so it can have at most one positive eigenvalue. But a positive eigenvalue has been found above, so $\lambda^\star$ is the largest eigenvalue.

□

Proof of Proposition A.15(b). First, assume that $\boldsymbol{w}_{\text{R}} = \pm\boldsymbol{\nu}_1$. Using Proposition A.15(a), I have $\boldsymbol{\beta} = \pm\sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta}}\boldsymbol{\Sigma}_Y\boldsymbol{\nu}_1$. This holds because $\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta} > 0$ as discussed in the proof of Proposition A.15(a) and $\boldsymbol{\Sigma}_Y$ is a full rank matrix by Assumptions A.5 and A.8(d). But $\boldsymbol{\Sigma}_Y\boldsymbol{\nu}_1 = \lambda_1\boldsymbol{\nu}_1$ because $(\lambda_1, \boldsymbol{\nu}_1)$ is an eigenpair of $\boldsymbol{\Sigma}_Y$ (using the notations defined in the beginning of Section 3.1.1). It follows that

$$\boldsymbol{\beta} = \pm\sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta}}\boldsymbol{\Sigma}_Y\boldsymbol{\nu}_1 = \pm\lambda\sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta}}\boldsymbol{\nu}_1,$$

which means $\boldsymbol{\beta}$ is parallel to $\boldsymbol{\nu}_1$ because $\lambda\sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta}}$ is a scalar.

Conversely, assume that $\boldsymbol{\beta}$ is parallel to $\boldsymbol{\nu}_1$. This means $\boldsymbol{\beta} = t\boldsymbol{\nu}_1$ for some $t \in \mathbb{R}\backslash\{0\}$. Using Proposition A.15(a),

$$\begin{aligned}
\boldsymbol{w}_{\text{R}} &= \frac{\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}_Y^{-2}\boldsymbol{\beta}}} \\
&= \frac{\boldsymbol{\Sigma}_Y^{-1}(t\boldsymbol{\nu}_1)}{\sqrt{(t\boldsymbol{\nu}_1)'\boldsymbol{\Sigma}_Y^{-2}(t\boldsymbol{\nu}_1)}} \\
&= \frac{t\lambda_1^{-1}\boldsymbol{\nu}_1}{\sqrt{t^2\lambda_1^{-2}\boldsymbol{\nu}_1'\boldsymbol{\nu}_1}} \\
&= \frac{t\lambda_1^{-1}\boldsymbol{\nu}_1}{\sqrt{t^2\lambda_1^{-2}}} \\
&= \pm\boldsymbol{\nu}_1.
\end{aligned}$$

In the above, the second equality follows from $\boldsymbol{\beta} = t\boldsymbol{\nu}_1$. The third equality uses that $(\boldsymbol{\nu}_1, \lambda_1)$ is an eigenpair of $\boldsymbol{\Sigma}_Y$ so $(\boldsymbol{\nu}_1, \lambda_1^{-1})$ is also an eigenpair of $\boldsymbol{\Sigma}_Y^{-1}$ because $\boldsymbol{\Sigma}_Y$ is pos-

itive definite by Assumptions A.5 and A.8(d), so it is invertible and $\lambda_1 > 0$. The fourth equality uses $\boldsymbol{\nu}_1$ is a unit-length eigenvector as assumed in the beginning of Section 3.1.1. □

Proposition A.16. *Let Assumptions 3.1, A.5 and A.8 hold. Consider the linear model in* (A.33) *and the problem of testing* (A.34) *using* (A.35). *For any* $\boldsymbol{w}_0 \in \mathbb{R}^q$ *such that* $\|\boldsymbol{w}_0\|_2 = 1$ *and* $R^2(\boldsymbol{w}_0) \neq 1$, *the solution to the following problem*

$$\max_{\boldsymbol{w}:\|\boldsymbol{w}\|_2=1} \operatorname*{plim}_{n\to\infty} \frac{[\widehat{T}_n(\boldsymbol{w})]^2}{[\widehat{T}_n(\boldsymbol{w}_0)]^2}.$$

is given by $\pm\boldsymbol{w}_{\mathrm{R}}$.

Proof of Proposition A.16. To begin with, note that $\widehat{R}^2_n(\boldsymbol{w})$ is the sample analog of $R^2(\boldsymbol{w})$. For any $\boldsymbol{w} \in \mathbb{R}^q$ such that $\|\boldsymbol{w}\|_2^2 = 1$, $\widehat{R}^2_n(\boldsymbol{w}) \xrightarrow{p} R^2(\boldsymbol{w})$ holds by Lemma A.14. Next, for any $\boldsymbol{w}, \boldsymbol{w}_0 \in \mathbb{R}^q$ such that $\|\boldsymbol{w}\|_2 = \|\boldsymbol{w}_0\|_2 = 1$, I have

$$\frac{[\widehat{T}_n(\boldsymbol{w})]^2}{[\widehat{T}_n(\boldsymbol{w}_0)]^2} = \frac{\frac{\widehat{R}^2_n(\boldsymbol{w})}{1-\widehat{R}^2_n(\boldsymbol{w})}}{\frac{\widehat{R}^2_n(\boldsymbol{w}_0)}{1-\widehat{R}^2_n(\boldsymbol{w}_0)}} \xrightarrow{p} \frac{\frac{R^2(\boldsymbol{w})}{1-R^2(\boldsymbol{w})}}{\frac{R^2(\boldsymbol{w}_0)}{1-R^2(\boldsymbol{w}_0)}}, \tag{A.45}$$

by Lemma A.14, the continuous mapping theorem and the definition of $[\widehat{T}_n(\boldsymbol{w})]^2$ in (A.35).

Fixing $\boldsymbol{w}_0$ as in the statement of the proposition, the probability limit in (A.45) is an increasing function in $R^2(\boldsymbol{w})$ for $R^2(\boldsymbol{w}) \in [0, 1)$. This follows because $\frac{\mathrm{d}(\frac{x}{1-x})}{\mathrm{d}x} = \frac{1}{(1-x)^2} > 0$ for any $x \neq 1$. Hence, the probability limit of (A.45) is maximized when $R^2(\boldsymbol{w})$ is maximized. This optimal weight $\boldsymbol{w}$ is given by $\boldsymbol{w}^\star$ from Proposition A.15(a). □

Proof of Proposition A.9. The result follows from Propositions A.15 and A.16. □

Proof of Proposition A.12. Following the proof to Proposition A.15(a), the optimal $\boldsymbol{w}$ that maximizes $R^2(\boldsymbol{w})$ is parallel to $\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}$. Since $R^2(\boldsymbol{w})$ is homogeneous of degree 0, the optimal solution is $\boldsymbol{w}_{\mathrm{I}} \equiv \frac{\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}}{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}}$ by the given assumptions of the proposition.

Suppose $\boldsymbol{\beta}$ is parallel to $\mathbf{1}_q$, i.e., $\boldsymbol{\beta} = t\mathbf{1}_q$ for some $t \in \mathbb{R}\backslash\{0\}$. Thus,

$$\boldsymbol{w}_{\mathrm{I}} = \frac{t\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q}{t\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q} = \boldsymbol{w}_{\mathrm{ivm}}.$$

Next, suppose $\boldsymbol{w}_{\mathrm{I}} = \boldsymbol{w}_{\mathrm{ivm}}$. This means

$$\frac{\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}}{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}} = \frac{\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q},$$

or equivalently, $\boldsymbol{\beta} = \frac{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}}{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\mathbf{1}_q}\mathbf{1}_q$, because $\boldsymbol{\Sigma}_Y$ is invertible and $\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta} \neq 0$, i.e., $\boldsymbol{\beta}$ is parallel to $\mathbf{1}_q$. □

Proof of Proposition A.13. The expression for $\boldsymbol{w}_{\mathrm{I}}$ has been derived in the proof of Proposition A.12. Suppose $\boldsymbol{w}_{\mathrm{sa}}$ is parallel to $\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}$, then $t\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta} = \boldsymbol{w}_{\mathrm{sa}} = \frac{1}{q}\mathbf{1}_q$ for some $t \in \mathbb{R}\backslash\{0\}$. Thus,

$$\boldsymbol{w}_{\mathrm{I}} = \frac{\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}}{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}} = \frac{\frac{1}{tq}\mathbf{1}_q}{\frac{1}{tq}\mathbf{1}_q'\mathbf{1}_q} = \frac{1}{q}\mathbf{1}_q = \boldsymbol{w}_{\mathrm{sa}}.$$

Next, suppose $\boldsymbol{w}_{\mathrm{I}} = \boldsymbol{w}_{\mathrm{sa}}$. This means

$$\frac{\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}}{\mathbf{1}_q'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}} = \frac{1}{q}\mathbf{1}_q = \boldsymbol{w}_{\mathrm{sa}},$$

i.e., $\boldsymbol{w}_{\mathrm{sa}}$ is parallel to $\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\beta}$. □

A.5.4 Extensions

The analysis at the beginning of this section has assumed that there are no covariates to follow Example 3.5, and for exposition purposes. The analysis has shown that each of the methods does not necessarily maximize the t-statistic squared. In this subsection, I discuss various extensions to the earlier analysis and to show that the similar intuition and conclusion continue to apply with appropriate adjustments.

First, a similar analysis can be performed when covariates are included using the FWL theorem by replacing the outcome and D_i with the residualized variables and adjusting for the degrees of freedom in the t-statistic to account for the number of covariates.

Second, where homoskedasticity is not assumed, the t-statistic squared does not have the representation in terms of $\widehat{R}_n^2(\boldsymbol{w})$. Nevertheless, the t-statistic can still be analyzed under Assumption A.5. In particular, testing (A.34) can be performed by the test statistic $\widehat{T}_n(\boldsymbol{w}) = \frac{\sqrt{n}(\boldsymbol{w}'\widehat{\boldsymbol{\beta}}_n)}{\sqrt{\boldsymbol{w}'\widehat{\boldsymbol{\Sigma}}_{n,\widehat{\boldsymbol{\beta}}}\boldsymbol{w}}}$ where $\widehat{\boldsymbol{\Sigma}}_{n,\widehat{\boldsymbol{\beta}}}$ is a consistent estimator of $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$. In this case, one can conduct a similar analysis as in Section A.5.2 by considering a suitable $\boldsymbol{w}_0$ and the ratio $\frac{[\widehat{T}_n(\boldsymbol{w})]^2}{[\widehat{T}_n(\boldsymbol{w}_0)]^2}$ where $\boldsymbol{w}$ and $\boldsymbol{w}_0$ are in the corresponding class of weights that are suitable PCA,

SA, or IVM. For a fixed $\boldsymbol{w}_0$, the ratio can still be analyzed via the Rayleigh quotient in $\boldsymbol{w}$ below

$$\frac{\boldsymbol{w}'(\boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{w}}{\boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}}. \tag{A.46}$$

Thus, the analysis in Section A.5.2 can still be applied by redefining the matrices and corresponding assumption appropriately and the weights would be required to parallel to $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1}\boldsymbol{\beta}$ to maximize the quotient.

Finally, the analysis can be used to study the probability of rejection. Let $\text{cv} \in \mathbb{R}$ be a given critical value that depends on the significance level. Then, $\mathbb{P}[|\widehat{T}_n(\boldsymbol{w})| > \text{cv}] = \mathbb{P}[|\frac{\sqrt{n}(\boldsymbol{w}'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}'\boldsymbol{\beta})}{\sqrt{\boldsymbol{w}'\widehat{\boldsymbol{\Sigma}}_{n,\widehat{\boldsymbol{\beta}}}\boldsymbol{w}}} + \frac{\sqrt{n}(\boldsymbol{w}'\boldsymbol{\beta})}{\sqrt{\boldsymbol{w}'\widehat{\boldsymbol{\Sigma}}_{n,\widehat{\boldsymbol{\beta}}}\boldsymbol{w}}}| > \text{cv}]$. To proceed, one could consider the local alternative where $\boldsymbol{\beta} = n^{-1/2}\boldsymbol{b}$ and $\boldsymbol{b} \in \mathbb{R}^q$ to ensure that $\sqrt{n}(\boldsymbol{w}'\boldsymbol{\beta})$ is finite. Thus, this converges to $\mathbb{P}[|Z + \frac{\boldsymbol{w}'\boldsymbol{b}}{\sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}}}| > \text{cv}]$, where $Z \sim \mathcal{N}(0,1)$ by Assumption A.5 and Slutsky's theorem. This is the same as evaluating the probability of a folded normal distribution $|\mathcal{N}(\frac{\boldsymbol{w}'\boldsymbol{b}}{\sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}}}, 1)|$ and is increasing in the Rayleigh quotient $\frac{\boldsymbol{w}'\boldsymbol{b}\boldsymbol{b}'\boldsymbol{w}}{\boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}}$. Hence, the earlier analysis in Section A.5.2 and the previous paragraph can again be applied to show what choice of $\boldsymbol{w}$ maximizes the probability of rejection. In this case, the Rayleigh quotient (A.46) can be used but with $\boldsymbol{\beta}$ replaced by $\boldsymbol{b}$ and the statements have to be updated accordingly.

A.6 An example with duplicated outcomes

In this subsection, I consider a stylized example with duplicated outcomes and show that SA can lead to overcounting. This does not suggest that researchers include duplicated outcomes in practice. The stylized example tries to model the limiting case where some highly correlated outcomes are included in the index.

Example A.17 (Duplicated outcomes and simple averaging)**.** Consider the setting described by Example 3.5. Let $Y_{i,1}$ and $Y_{i,2}$ be two independent outcomes with $\text{Var}[Y_{i,1}] = \text{Var}[Y_{i,2}] = 1$. Set $Y_{i,2} = \cdots = Y_{i,q}$ (i.e., outcomes 2 to q are duplicated copies). In addition, assume that treatment effects are homogeneous $\beta_1 = \cdots = \beta_q = \overline{\beta}$ and the errors are homoskedastic. In this case, it seems unreasonable to assign equal weights to all outcomes. This is because the second treatment effect is effectively weighted by $\frac{q-1}{q}$ and the first treatment effect is weighted by $\frac{1}{q}$, but both has the same effect $\overline{\beta}$ and unit variance.

On the other hand, the variance-minimizing solution to $\boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w}$ subject to $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$

(i.e., the class of convex weights as defined in (2)) is

$$\left\{(w_1^\star, w_2^\star, \ldots, w_q^\star) : w_1^\star = \frac{1}{2},\ \sum_{j=2}^{q} w_j^\star = \frac{1}{2},\ w_j^\star \geq 0 \text{ for } j = 2, \ldots, q\right\}. \tag{A.47}$$

Since $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ is positive semidefinite, the solution is not unique. But this is as expected because $Y_{i,2}, \ldots, Y_{i,q}$ are all the same, so only the sum of weights on these $(q-1)$ random variables matters. (A.47) shows that computing the weights using $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ is helpful and leads to more reasonable weights. The proof for (A.47) can be found below. △

Proof of equation (A.47). Following the notations in Example 3.5, write the error term as $\boldsymbol{U}_i = (U_{i,1}, U_{i,2}, \ldots, U_{i,2})'$ where $U_{i,2}$ is repeated for $(q-1)$ times. Let $\sigma_{1,j}^2 \equiv \text{Var}[D_i U_{i,j}]$ and $\sigma_{0,j}^2 \equiv \text{Var}[(1-D_i)U_{i,j}]$ for $j = 1, 2$. Then,

$$\text{Var}[D_i \boldsymbol{U}_i] = \text{Var}\left[\begin{pmatrix} D_i U_{i,1} \\ D_i U_{i,2} \\ \vdots \\ D_i U_{i,q} \end{pmatrix}\right] = \text{Var}\left[\begin{pmatrix} D_i U_{i,1} \\ D_i U_{i,2} \\ \vdots \\ D_i U_{i,q} \end{pmatrix}\right] = \begin{pmatrix} \sigma_{1,1}^2 & \mathbf{0}_{q-1}' \\ \mathbf{0}_{q-1} & \sigma_{1,2}^2 \mathbf{1}_{(q-1)\times(q-1)} \end{pmatrix},$$

and similarly,

$$\text{Var}[(1-D_i)\boldsymbol{U}_i] = \text{Var}\left[\begin{pmatrix} (1-D_i) U_{i,1} \\ (1-D_i) U_{i,2} \\ \vdots \\ (1-D_i) U_{i,q} \end{pmatrix}\right] = \begin{pmatrix} \sigma_{0,1}^2 & \mathbf{0}_{q-1}' \\ \mathbf{0}_{q-1} & \sigma_{0,2}^2 \mathbf{1}_{(q-1)\times(q-1)} \end{pmatrix}.$$

Thus, the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} &= \frac{\text{Var}[D_i \boldsymbol{U}_i]}{p_D^2} + \frac{\text{Var}[(1-D_i)\boldsymbol{U}_i]}{(1-p_D)^2} \\ &= \frac{1}{p_D^2}\begin{pmatrix} \sigma_{1,1}^2 & \mathbf{0}_{q-1}' \\ \mathbf{0}_{q-1} & \sigma_{1,2}^2 \mathbf{1}_{(q-1)\times(q-1)} \end{pmatrix} + \frac{1}{(1-p_D)^2}\begin{pmatrix} \sigma_{0,1}^2 & \mathbf{0}_{q-1}' \\ \mathbf{0}_{q-1} & \sigma_{0,2}^2 \mathbf{1}_{(q-1)\times(q-1)} \end{pmatrix} \\ &= \begin{pmatrix} \varsigma_{1,\widehat{\boldsymbol{\beta}}}^2 & \mathbf{0}_{q-1}' \\ \mathbf{0}_{q-1} & \varsigma_{2,\widehat{\boldsymbol{\beta}}}^2 \mathbf{1}_{(q-1)\times(q-1)} \end{pmatrix}, \end{aligned}$$

where

$$\varsigma^2_{j,\widehat{\boldsymbol{\beta}}} \equiv \frac{\sigma^2_{1,j}}{p_D^2} + \frac{\sigma^2_{0,j}}{(1-p_D)^2},$$

for $j = 1, 2$.

Therefore,

$$\boldsymbol{w}'\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{w} = \varsigma^2_{1,\widehat{\boldsymbol{\beta}}} w_1^2 + \varsigma^2_{2,\widehat{\boldsymbol{\beta}}} \left(\sum_{j=2}^{q} w_j \right)^2 = \varsigma^2_{1,\widehat{\boldsymbol{\beta}}} w_1^2 + \varsigma^2_{2,\widehat{\boldsymbol{\beta}}} (1 - w_1)^2, \tag{A.48}$$

where the last equality follows because $\boldsymbol{w}'\mathbf{1}_q = 1$ as $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. The first-order condition of the above is

$$2\varsigma^2_{1,\widehat{\boldsymbol{\beta}}} w_1 - 2(1 - w_1)\varsigma^2_{2,\widehat{\boldsymbol{\beta}}} = 0,$$

or equivalently,

$$w_1 = \frac{\varsigma^2_{2,\widehat{\boldsymbol{\beta}}}}{\varsigma^2_{1,\widehat{\boldsymbol{\beta}}} + \varsigma^2_{2,\widehat{\boldsymbol{\beta}}}}. \tag{A.49}$$

This leads to the optimal solution that minimizes (A.48).

In the example, I assumed that $\beta_j = \overline{\beta}$ for any $j = 1, \ldots, q$, and that errors are homoskedastic. Since $Y_{i,j} = \xi_j + \overline{\beta} D_i + U_{i,j}$ and $\text{Var}[Y_{i,j}] = 1$ for any $j = 1, \ldots, q$, this means $1 = \text{Var}[Y_{i,j}] = \overline{\beta}^2 \text{Var}[D_i] + \text{Var}[U_{i,j}]$, or equivalently,

$$\text{Var}[U_{i,j}] = 1 - \overline{\beta}^2 p_D (1 - p_D), \tag{A.50}$$

for $j = 1, \ldots, q$. Hence,

$$\sigma^2_{1,j} = \mathbb{E}[D_i^2 U_{i,j}^2] - \mathbb{E}[D_i U_{i,j}]^2 = \mathbb{E}[D_i^2]\mathbb{E}[U_{i,j}^2] = p_D[1 - \overline{\beta}^2 p_D(1 - p_D)]$$

and

$$\sigma^2_{0,j} = \mathbb{E}[(1 - D_i)^2 U_{i,j}^2] - \mathbb{E}[(1 - D_i) U_{i,j}]^2 = (1 - p_D)[1 - \overline{\beta}^2 p_D(1 - p_D)].$$

for $j = 1, \ldots, q$. As a result, this implies that $\varsigma^2_{1,\widehat{\boldsymbol{\beta}}} = \varsigma^2_{2,\widehat{\boldsymbol{\beta}}}$. It follows from (A.49) that $w_1^\star = \frac{1}{2}$ and $\sum_{j=2}^{q} w_j^\star = \frac{1}{2}$. □

A.7 Additional details on inference for generated outcomes

This subsection shows the limiting distribution when one studies generated outcomes. Recall that it has been assumed throughout the appendix that $\{(D_i, \boldsymbol{X}_i', \boldsymbol{Y}_i')\}_{i=1}^n$ are i.i.d. across i. In addition, I use Assumption A.1 in the next two subsections. This assumption characterizes the limiting distribution of $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\Omega}$. Note that the standardization step has been included in $\widehat{\boldsymbol{\beta}}_n$ as discussed in the previous section on precision (see, for instance, (A.22)). Hence, the asymptotic variance matrix in Assumption A.1 has already accounted for the correct calculation for the variance matrices that account for the standardization step as in Assumption 2.2.

A.7.1 Inference for PCA

Lemma A.18. *Let Assumptions 3.1 hold on $\boldsymbol{\Omega}$ and A.1 hold. Write $\{(\boldsymbol{\nu}_j, \lambda_j)\}_{j=1}^q$ as the eigenpairs of $\boldsymbol{\Omega}$ such that $\{\boldsymbol{\nu}_j\}_{j=1}^q$ are unit-length eigenvectors and $\lambda_1 \geq \cdots \geq \lambda_q$. Let $\widehat{\boldsymbol{w}}_{\text{pca},n} = \arg\max_{\boldsymbol{w} \in \mathcal{W}_{\text{unit}}} \boldsymbol{w}' \widehat{\boldsymbol{\Omega}}_n \boldsymbol{w}$ where $\mathcal{W}_{\text{unit}}$ is as defined in (5). Suppose $\boldsymbol{c}' \boldsymbol{w}_{\text{pca}} > 0$. Then, $\widehat{\boldsymbol{w}}_{\text{pca},n} \xrightarrow{p} \boldsymbol{w}_{\text{pca}}$.*

Proof of Lemma A.18. By Assumption 3.1, $\widehat{\boldsymbol{w}}_{\text{pca},n}' \widehat{\boldsymbol{\Omega}}_n \widehat{\boldsymbol{w}}_{\text{pca},n} = \sup_{\boldsymbol{w} \in \mathcal{W}_{\text{unit}}} \boldsymbol{w}' \widehat{\boldsymbol{\Omega}}_n \boldsymbol{w}$ because the leading eigenvalue is unique. In addition, for any $\boldsymbol{w} \in \mathcal{W}_{\text{unit}}$, $|\boldsymbol{w}' \widehat{\boldsymbol{\Omega}}_n \boldsymbol{w} - \boldsymbol{w} \boldsymbol{\Omega} \boldsymbol{w}| \leq \|\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega}\| \|\boldsymbol{w}\|_2^2$. Hence, $\sup_{\boldsymbol{w} \in \mathcal{W}_{\text{unit}}} |\boldsymbol{w}' \widehat{\boldsymbol{\Omega}}_n \boldsymbol{w} - \boldsymbol{w}' \boldsymbol{\Omega} \boldsymbol{w}| \xrightarrow{p} 0$ by Assumption A.1. Therefore, $\widehat{\boldsymbol{w}}_{\text{pca},n} \xrightarrow{p} \boldsymbol{w}_{\text{pca}}$ by Theorem 2.12(i) of Kosorok (2008). □

Proposition A.19. *Consider the same assumptions as in Lemma A.18. Write $\widehat{\tau}_n \equiv \widehat{\boldsymbol{w}}_{\text{pca},n}' \widehat{\boldsymbol{\beta}}_n$ and $\tau \equiv \boldsymbol{w}_{\text{pca}}' \boldsymbol{\beta}$. Then,*

$$\sqrt{n}(\widehat{\tau}_n - \tau) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{w}_{\text{pca}}' \boldsymbol{\Sigma} \boldsymbol{w}_{\text{pca}} + 2 \boldsymbol{w}_{\text{pca}}' \boldsymbol{\Psi}_{\boldsymbol{\Omega}, \boldsymbol{\beta}}' \boldsymbol{B}_{\boldsymbol{\nu}}' \boldsymbol{\beta} + \boldsymbol{\beta}' \boldsymbol{B}_{\boldsymbol{\nu}} \boldsymbol{\Psi}_{\boldsymbol{\Omega}} \boldsymbol{B}_{\boldsymbol{\nu}}' \boldsymbol{\beta}),$$

where $\boldsymbol{B}_{\boldsymbol{\nu}} \equiv \sum_{j=2}^q \boldsymbol{\nu}_j \frac{\text{vec}[\boldsymbol{\nu}_j \boldsymbol{\nu}_1']' \mathbf{D}}{\lambda_1 - \lambda_j}$, $\mathbf{D}$ is the duplication matrix such that $\mathbf{D}\, \text{vech}[\boldsymbol{\Omega}] = \text{vec}[\boldsymbol{\Omega}]$, and $\{(\boldsymbol{\nu}_j, \lambda_j)\}_{j=1}^q$ are the eigenpairs of $\boldsymbol{\Omega}$ such that $\boldsymbol{c}' \boldsymbol{\nu}_j \geq 0$ for $j = 1, \ldots, q$.

Proof of Proposition A.19. Let $\widehat{\tau}_n$ and τ be as defined in the statement of the proposition. Then,

$$\begin{aligned}
\sqrt{n}(\widehat{\tau}_n - \tau) &= \sqrt{n}(\widehat{\boldsymbol{w}}_{\text{pca},n}' \widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\text{pca}}' \boldsymbol{\beta}) \\
&= \sqrt{n}(\widehat{\boldsymbol{w}}_{\text{pca},n}' \widehat{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{w}}_{\text{pca},n}' \boldsymbol{\beta} + \widehat{\boldsymbol{w}}_{\text{pca},n}' \boldsymbol{\beta} - \boldsymbol{w}_{\text{pca}}' \boldsymbol{\beta}) \\
&= \widehat{\boldsymbol{w}}_{\text{pca},n}' [\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})] + \boldsymbol{\beta}' [\sqrt{n}(\widehat{\boldsymbol{w}}_{\text{pca},n} - \boldsymbol{w}_{\text{pca}})]
\end{aligned}$$

$$
\begin{aligned}
&= \begin{pmatrix} \widehat{\boldsymbol{w}}'_{\text{pca},n} & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}(\widehat{\boldsymbol{w}}_{\text{pca},n} - \boldsymbol{w}_{\text{pca}}) \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{w}'_{\text{pca}} & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}\sum_{j=2}^{q} \frac{\boldsymbol{\nu}'_j(\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega})\boldsymbol{\nu}_1}{\lambda_1 - \lambda_j}\boldsymbol{\nu}_j \end{pmatrix} + o_{\mathbb{P}}(1) \\
&= \begin{pmatrix} \boldsymbol{w}'_{\text{pca}} & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}\sum_{j=2}^{q} \frac{\text{vec}[\boldsymbol{\nu}_j\boldsymbol{\nu}'_1]' \,\text{vec}[\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega}]}{\lambda_1 - \lambda_j}\boldsymbol{\nu}_j \end{pmatrix} + o_{\mathbb{P}}(1) \\
&= \begin{pmatrix} \boldsymbol{w}'_{\text{pca}} & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_q & \boldsymbol{0} \\ \boldsymbol{0} & \sum_{j=2}^{q} \boldsymbol{\nu}_j \frac{\text{vec}[\boldsymbol{\nu}_j\boldsymbol{\nu}'_1]'\mathbf{D}}{\lambda_1 - \lambda_j} \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}[\text{vech}(\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega})] \end{pmatrix} + o_{\mathbb{P}}(1) \\
&\equiv \begin{pmatrix} \boldsymbol{w}'_{\text{pca}} & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_q & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_{\boldsymbol{\nu}} \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}[\text{vech}(\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega})] \end{pmatrix} + o_{\mathbb{P}}(1) \\
&\xrightarrow{d} \begin{pmatrix} \boldsymbol{w}'_{\text{pca}} & \boldsymbol{\beta}'\boldsymbol{B}_{\boldsymbol{\nu}} \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_{\boldsymbol{\beta}} \\ \boldsymbol{Z}_{\text{vech}[\boldsymbol{\Omega}]} \end{pmatrix} \\
&\sim \mathcal{N}(0, \boldsymbol{w}'_{\text{pca}}\boldsymbol{\Sigma}\boldsymbol{w}_{\text{pca}} + 2\boldsymbol{w}'_{\text{pca}}\boldsymbol{\Psi}'_{\boldsymbol{\Omega},\boldsymbol{\beta}}\boldsymbol{B}'_{\boldsymbol{\nu}}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{B}_{\boldsymbol{\nu}}\boldsymbol{\Psi}_{\boldsymbol{\Omega}}\boldsymbol{B}'_{\boldsymbol{\nu}}\boldsymbol{\beta}),
\end{aligned}
$$

where the first line uses the definition of the estimators, the second line adds and subtracts, the fifth line uses Lemma A.18 and uses eigenvector perturbation (see, for instance, Stewart (2001, Theorem 3.11 of Chapter 1)), the seventh line defines $\boldsymbol{B}_{\boldsymbol{\nu}}$ as in the statement of the proposition, the eighth and last lines use the notations in Assumption A.1. □

A.7.2 Inference for IVM

Lemma A.20. *Let $\boldsymbol{\Sigma}$ be a positive definite matrix. Define $\boldsymbol{a}(\boldsymbol{\Sigma}) \equiv \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_q}{\mathbf{1}'_q\boldsymbol{\Sigma}^{-1}\mathbf{1}_q}$. Then,*

$$
\frac{d\boldsymbol{a}(\boldsymbol{\Sigma})}{d\,\text{vec}[\boldsymbol{\Sigma}]} = \frac{-(\mathbf{1}'_q\boldsymbol{\Sigma}^{-1}) \otimes \boldsymbol{\Sigma}^{-1}(\mathbf{1}'_q\boldsymbol{\Sigma}^{-1}\mathbf{1}_q) + (\boldsymbol{\Sigma}^{-1}\mathbf{1}_q)[(\mathbf{1}'_q\boldsymbol{\Sigma}^{-1}) \otimes (\mathbf{1}'_q\boldsymbol{\Sigma}^{-1})]}{(\mathbf{1}'_q\boldsymbol{\Sigma}^{-1}\mathbf{1}_q)^2}.
$$

Proof of Lemma A.20. By matrix derivative equalities (see, for instance, Example 18.8a of Magnus and Neudecker (2019)),

$$
\begin{aligned}
\text{d}(\boldsymbol{\Sigma}^{-1}\mathbf{1}_q) &= \text{d}(\boldsymbol{I}_q\boldsymbol{\Sigma}^{-1}\mathbf{1}_q) = -\boldsymbol{\Sigma}^{-1}(\text{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{1}_q, && \text{(A.51)} \\
\text{d}(\mathbf{1}'_q\boldsymbol{\Sigma}^{-1}\mathbf{1}_q) &= -\mathbf{1}'_q\boldsymbol{\Sigma}^{-1}(\text{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{1}_q. && \text{(A.52)}
\end{aligned}
$$

Since $\boldsymbol{\Sigma}^{-1}\mathbf{1}_q$ is a column vector and $\mathbf{1}'_q\boldsymbol{\Sigma}^{-1}\mathbf{1}_q$ is a scalar, I have $\text{vec}[\boldsymbol{\Sigma}^{-1}\mathbf{1}_q] = \boldsymbol{\Sigma}^{-1}\mathbf{1}_q$ and $\text{vec}[\mathbf{1}'_q\boldsymbol{\Sigma}^{-1}\mathbf{1}_q] = \mathbf{1}'_q\boldsymbol{\Sigma}^{-1}\mathbf{1}_q$.

Thus, (A.51) and (A.52) can be written as

$$\mathrm{d}(\boldsymbol{\Sigma}^{-1}\mathbf{1}_q) = -\operatorname{vec}[\boldsymbol{I}_q\boldsymbol{\Sigma}^{-1}(\mathrm{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{1}_q] = -\left[(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1})\otimes\boldsymbol{\Sigma}^{-1}\right]\operatorname{vec}[\mathrm{d}\boldsymbol{\Sigma}], \tag{A.53}$$

$$\mathrm{d}(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1}\mathbf{1}_q) = -\operatorname{vec}[\mathbf{1}_q'\boldsymbol{\Sigma}^{-1}(\mathrm{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{1}_q] = -\left[(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1})\otimes(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1})\right]\operatorname{vec}[\mathrm{d}\boldsymbol{\Sigma}], \tag{A.54}$$

using properties on the Kronecker product (see, for instance, Example 18.5 of Magnus and Neudecker (2019)). It follows that

$$\frac{\mathrm{d}(\boldsymbol{\Sigma}^{-1}\mathbf{1}_q)}{\mathrm{d}\operatorname{vec}[\boldsymbol{\Sigma}]} = -(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1})\otimes\boldsymbol{\Sigma}^{-1}, \tag{A.55}$$

$$\frac{\mathrm{d}(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1}\mathbf{1}_q)}{\mathrm{d}\operatorname{vec}[\boldsymbol{\Sigma}]} = -(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1})\otimes(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1}). \tag{A.56}$$

Hence, the gradient of $\boldsymbol{a}(\boldsymbol{\Sigma})$ with respect to $\operatorname{vec}[\boldsymbol{\Sigma}]$ is given by

$$\frac{\mathrm{d}\boldsymbol{a}(\boldsymbol{\Sigma})}{\mathrm{d}\operatorname{vec}[\boldsymbol{\Sigma}]} = \frac{-(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1})\otimes\boldsymbol{\Sigma}^{-1}(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1}\mathbf{1}_q) + (\boldsymbol{\Sigma}^{-1}\mathbf{1}_q)[(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1})\otimes(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1})]}{(\mathbf{1}_q'\boldsymbol{\Sigma}^{-1}\mathbf{1}_q)^2},$$

using the chain rule, (A.55) and (A.56). Since $\boldsymbol{\Sigma}$ is positive definite, $\mathbf{1}_q'\boldsymbol{\Sigma}^{-1}\mathbf{1}_q > 0$ in the above derivative. □

The following shows the limiting distribution of the aggregated treatment effect using $\widehat{\boldsymbol{w}}_{\mathrm{ivm},n}$.

Proposition A.21. *Let Assumption A.1 hold and $\boldsymbol{\Omega}$ is positive definite. Let $\boldsymbol{w}_{\mathrm{ivm}}$ be as defined in (11) and $\widehat{\boldsymbol{w}}_{\mathrm{ivm},n} \equiv \frac{\widehat{\boldsymbol{\Omega}}_n^{-1}\mathbf{1}_q}{\mathbf{1}_q'\widehat{\boldsymbol{\Omega}}_n^{-1}\mathbf{1}_q}$. Define $\widehat{\tau}_{\mathrm{ivm},n} \equiv \widehat{\boldsymbol{w}}_{\mathrm{ivm},n}'\widehat{\boldsymbol{\beta}}_n$, and $\tau_{\mathrm{ivm}} \equiv \boldsymbol{w}_{\mathrm{ivm}}'\boldsymbol{\beta}$. Then,*

$$\sqrt{n}(\widehat{\tau}_{\mathrm{ivm},n} - \tau_{\mathrm{ivm}}) \xrightarrow{d} \mathcal{N}(0, \sigma_{\mathrm{ivm}}^2),$$

where

$$\sigma_{\mathrm{ivm}}^2 \equiv \boldsymbol{w}_{\mathrm{ivm}}'\boldsymbol{\Sigma}\boldsymbol{w}_{\mathrm{ivm}} + 2\boldsymbol{w}_{\mathrm{ivm}}'\boldsymbol{\Psi}_{\boldsymbol{\Omega},\boldsymbol{\beta}}'\mathbf{D}_{\mathsf{a}}'\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{D}_{\mathsf{a}}\boldsymbol{\Psi}_{\boldsymbol{\Omega}}\mathbf{D}_{\mathsf{a}}'\boldsymbol{\beta}.$$

Proof of Proposition A.21. To begin with, write $\boldsymbol{a}(\boldsymbol{\Omega}) \equiv \frac{\boldsymbol{\Omega}^{-1}\mathbf{1}_q}{\mathbf{1}_q'\boldsymbol{\Omega}^{-1}\mathbf{1}_q}$ where $\boldsymbol{\Omega}$ is a positive definite matrix. The expression $\frac{\mathrm{d}\boldsymbol{a}(\boldsymbol{\Omega})}{\mathrm{d}\operatorname{vec}[\boldsymbol{\Omega}]}$ is given in Lemma A.20. Since $\boldsymbol{\Omega}$ is positive definite, $\mathbf{1}_q'\boldsymbol{\Omega}^{-1}\mathbf{1}_q > 0$.

Let $\mathbf{G} \equiv \frac{\mathrm{d}\boldsymbol{a}(\boldsymbol{\Omega})}{\mathrm{d}\operatorname{vec}[\boldsymbol{\Omega}]}$ and $\mathbf{D}$ be the duplication matrix such that $\operatorname{vec}[\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega}] = \mathbf{D}\operatorname{vech}[\widehat{\boldsymbol{\Omega}}_n -$

$\boldsymbol{\Omega}$]. Hence, using the Delta method,

$$\sqrt{n}[\boldsymbol{a}(\widehat{\boldsymbol{\Omega}}_n) - \boldsymbol{a}(\boldsymbol{\Omega})] = \mathbf{D}_{\mathsf{a}}\sqrt{n}\,\mathrm{vech}[\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega}] + o_{\mathbb{P}}(1), \tag{A.57}$$

where $\mathbf{D}_{\mathsf{a}} \equiv \mathbf{GD}$. Hence,

$$\begin{aligned}
\sqrt{n}(\widehat{\tau}_{\mathrm{ivm},n} - \tau_{\mathrm{ivm}}) &= \sqrt{n}[\boldsymbol{a}(\widehat{\boldsymbol{\Omega}}_n)'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{a}(\boldsymbol{\Omega})'\boldsymbol{\beta}] \\
&= \sqrt{n}[\boldsymbol{a}(\widehat{\boldsymbol{\Omega}}_n)'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{a}(\widehat{\boldsymbol{\Omega}}_n)'\boldsymbol{\beta} + \boldsymbol{a}(\widehat{\boldsymbol{\Omega}}_n)'\boldsymbol{\beta} - \boldsymbol{a}(\boldsymbol{\Omega})'\boldsymbol{\beta}] \\
&= \begin{pmatrix} \boldsymbol{a}(\widehat{\boldsymbol{\Omega}}_n)' & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}[\boldsymbol{a}(\widehat{\boldsymbol{\Omega}}_n) - \boldsymbol{a}(\boldsymbol{\Omega})] \end{pmatrix} \\
&\xrightarrow{d} \begin{pmatrix} \boldsymbol{a}(\boldsymbol{\Omega})' & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_{\boldsymbol{\beta}} \\ \mathbf{D}_{\mathsf{a}}\boldsymbol{Z}_{\mathrm{vech}[\boldsymbol{\Omega}]} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{w}_{\mathrm{ivm}}' & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_{\boldsymbol{\beta}} \\ \mathbf{D}_{\mathsf{a}}\boldsymbol{Z}_{\mathrm{vech}[\boldsymbol{\Omega}]} \end{pmatrix} \\
&\sim \mathcal{N}(0, \sigma^2_{\mathrm{ivm}}),
\end{aligned}$$

where the first line follows from the definition of $\boldsymbol{a}(\cdot)$ and the estimators, the second line follows from adding and subtracting, the fourth line follows from (A.57) and that $\boldsymbol{a}(\widehat{\boldsymbol{\Omega}}_n) \xrightarrow{p} \boldsymbol{a}(\boldsymbol{\Omega})$ by the continuous mapping theorem and the given assumptions, the fifth line follows from Assumption A.1, and the last line follows from defining

$$\begin{aligned}
\sigma^2_{\mathrm{ivm}} &\equiv \begin{pmatrix} \boldsymbol{w}_{\mathrm{ivm}}' & \boldsymbol{\beta}'\mathbf{D}_{\mathsf{a}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Psi}_{\boldsymbol{\Omega},\boldsymbol{\beta}}' \\ \boldsymbol{\Psi}_{\boldsymbol{\Omega},\boldsymbol{\beta}} & \boldsymbol{\Psi}_{\boldsymbol{\Omega}} \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_{\mathrm{ivm}} \\ \mathbf{D}_{\mathsf{a}}'\boldsymbol{\beta} \end{pmatrix} \\
&= \boldsymbol{w}_{\mathrm{ivm}}'\boldsymbol{\Sigma}\boldsymbol{w}_{\mathrm{ivm}} + 2\boldsymbol{w}_{\mathrm{ivm}}'\boldsymbol{\Psi}_{\boldsymbol{\Omega},\boldsymbol{\beta}}'\mathbf{D}_{\mathsf{a}}'\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{D}_{\mathsf{a}}\boldsymbol{\Psi}_{\boldsymbol{\Omega}}\mathbf{D}_{\mathsf{a}}'\boldsymbol{\beta}.
\end{aligned}$$

□

A.7.3 General result

This subsection concludes with a general result that shows one has to account for the "generated outcome" in computing the correct standard error.

Assumption A.22. *Suppose $\widehat{\boldsymbol{w}}_n \in \mathbb{R}^q$ and $\widehat{\boldsymbol{\beta}}_n \in \mathbb{R}^q$ have the following representation:*

$$\sqrt{n}(\widehat{\boldsymbol{w}}_n - \boldsymbol{w}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i) + o_{\mathbb{P}}(1),$$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i) + o_{\mathbb{P}}(1).$$

where $\boldsymbol{\varphi}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i) \equiv (\boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)', \boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)')'$ *is the corresponding influence function such that* $\mathbb{E}[\boldsymbol{\varphi}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)] = \mathbf{0}_{2q}$ *and* $\mathbb{E}[\boldsymbol{\varphi}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)\boldsymbol{\varphi}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)']$ *exists.*

The above assumption allows the definition of the estimator to be general to allow for flexible choices. This is because researchers may choose to divide outcomes from the outcomes' standard deviation using the full sample or control subsample depending on how they want to interpret effect size. Under Assumption 2.2, $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_n$ are the treatment effects on the (suitably) standardized outcomes. Hence, the terms in Assumption A.22 have taken the standardization into account and the influence function has been adjusted for the estimated weights. See Example A.24 below for a binary treatment example. The proposition below shows that one has to take the generated outcome using the data-dependent weights into account in conducting inference unless the additional variance terms equal 0. This generated outcome issue is related to the literature on generated regressors studied since Pagan (1984) and Murphy and Topel (1985) although I focus on the dependent variable. This is also related to conducting inference on generated variables from unstructured data studied by Battaglia et al. (2024) in which they discuss how variables are generated and used as regressors via a "two-step approach."

Proposition A.23. *Let Assumption A.22 hold. Write* $\widehat{\tau}_n \equiv \widehat{\boldsymbol{w}}_n'\widehat{\boldsymbol{\beta}}_n$ *and* $\tau \equiv \boldsymbol{w}_0'\boldsymbol{\beta}$*. Then,*

$$\sqrt{n}(\widehat{\boldsymbol{w}}_n'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_0'\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \sigma_\tau^2),$$

where

$$\sigma_\tau^2 \equiv \sigma_{\tau,\boldsymbol{\beta}}^2 + 2\sigma_{\tau,\boldsymbol{\beta},\boldsymbol{w}}^2 + \sigma_{\tau,\boldsymbol{w}}^2,$$

$\sigma_{\tau,\boldsymbol{\beta}}^2 \equiv \boldsymbol{w}_0' \operatorname{Var}[\boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)]\boldsymbol{w}_0$, $\sigma_{\tau,\boldsymbol{\beta},\boldsymbol{w}}^2 \equiv \boldsymbol{w}_0' \operatorname{Cov}[\boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i), \boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)]\boldsymbol{\beta}$*, and* $\sigma_{\tau,\boldsymbol{w}}^2 \equiv \boldsymbol{\beta}' \operatorname{Var}[\boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)]\boldsymbol{\beta}$*.*

The above proposition shows that the uncertainty in the weights also contribute to the asymptotic variance and should not be treated as "fixed."

Proof of Proposition A.23. Using the notations in Assumption A.22, I can write

$$\begin{aligned}
\sqrt{n}(\widehat{\tau}_n - \tau) &= \sqrt{n}(\widehat{\boldsymbol{w}}_n'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_0'\boldsymbol{\beta}) \\
&= \sqrt{n}(\widehat{\boldsymbol{w}}_n'\widehat{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{w}}_n'\boldsymbol{\beta} + \widehat{\boldsymbol{w}}_n'\boldsymbol{\beta} - \boldsymbol{w}_0'\boldsymbol{\beta})
\end{aligned}$$

$$
\begin{aligned}
&= \widehat{\boldsymbol{w}}_n' \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) + \boldsymbol{\beta}' \sqrt{n}(\widehat{\boldsymbol{w}}_n - \boldsymbol{w}_0) \\
&= \begin{pmatrix} \widehat{\boldsymbol{w}}_n' & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}(\widehat{\boldsymbol{w}}_n - \boldsymbol{w}_0) \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{w}_0' + o_{\mathbb{P}}(1) & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i) + o_{\mathbb{P}}(1) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i) + o_{\mathbb{P}}(1) \end{pmatrix} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\boldsymbol{w}_0' \boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i) + \boldsymbol{\beta}' \boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i) \right] + o_{\mathbb{P}}(1), \qquad \text{(A.58)}
\end{aligned}
$$

where the fifth line uses $\widehat{\boldsymbol{w}}_n - \boldsymbol{w}_0 = o_{\mathbb{P}}(1)$ and Assumption A.22.

Continuing from the last line above,

$$
\begin{aligned}
\sqrt{n}(\widehat{\tau}_n - \tau) &= \begin{pmatrix} \boldsymbol{w}_0' & \boldsymbol{\beta}' \end{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i) \\ \boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i) \end{pmatrix} + o_{\mathbb{P}}(1) \\
&\xrightarrow{d} \begin{pmatrix} \boldsymbol{w}_0' & \boldsymbol{\beta}' \end{pmatrix} \mathcal{N}\left(\begin{pmatrix} \mathbf{0}_q \\ \mathbf{0}_q \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\beta}}' \\ \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{w}} \end{pmatrix} \right) \\
&= \mathcal{N}(0, \boldsymbol{w}_0' \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{w}_0 + 2\boldsymbol{w}_0' \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\beta}} \boldsymbol{\beta} + \boldsymbol{\beta}' \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\beta}),
\end{aligned}
$$

where the second line uses the Central Limit Theorem because the second moment of $(\boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)', \boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)')'$ exists and defines $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} \equiv \text{Var}[\boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)]$, $\boldsymbol{\Sigma}_{\boldsymbol{w}} \equiv \text{Var}[\boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)]$, and $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\beta}} \equiv \text{Cov}[\boldsymbol{\varphi}_{\boldsymbol{w}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i), \boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \boldsymbol{X}_i, \boldsymbol{Y}_i)]$, and the last line follows from the continuous mapping theorem. The proposition holds from defining $\sigma^2_{\tau,\boldsymbol{\beta}} \equiv \boldsymbol{w}_0' \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{w}_0$, $\sigma^2_{\tau,\boldsymbol{\beta},\boldsymbol{w}} \equiv \boldsymbol{w}_0' \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\beta}} \boldsymbol{\beta}$, and $\sigma^2_{\tau,\boldsymbol{w}} \equiv \boldsymbol{\beta}' \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\beta}$. □

Example A.24. Suppose $D_i \in \{0, 1\}$, there is no other controls, and that Assumption A.8 holds. Then, the estimator for the treatment effects as follows for each $j = 1, \ldots, q$:

$$
\begin{aligned}
\widetilde{\widehat{\beta}}_{n,j} &= \frac{1}{n_1} \sum_{i=1}^n D_i \widetilde{Y}_{i,j} - \frac{1}{n_0} \sum_{i=1}^n (1 - D_i) \widetilde{Y}_{i,j} \\
&= \mathbb{E}[\widetilde{Y}_{i,j} | D_i = 1] - \mathbb{E}[\widetilde{Y}_{i,j} | D_i = 0] \\
&\quad + \frac{1}{n_1} \sum_{i=1}^n D_i (\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j} | D_i = 1]) - \frac{1}{n_0} \sum_{i=1}^n (1 - D_i)(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j} | D_i = 0]) \\
&= \widetilde{\beta}_j + \frac{1}{n_1} \sum_{i=1}^n D_i (\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j} | D_i = 1]) - \frac{1}{n_0} \sum_{i=1}^n (1 - D_i)(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j} | D_i = 0]), \qquad \text{(A.59)}
\end{aligned}
$$

where the first line follows from the difference-in-means estimator, the second line adds and subtracts $\mathbb{E}[\widetilde{Y}_{i,j} | D_i = 1]$ and $\mathbb{E}[\widetilde{Y}_{i,j} | D_i = 0]$, and the third line uses the definition that

$\widetilde{\beta}_j \equiv \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 1] - \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 0]$.

Hence, recentering and rescaling $\widehat{\widetilde{\beta}}_{n,j}$ for each $j = 1, \ldots, q$ gives

$$\begin{aligned}
\sqrt{n}(\widehat{\widetilde{\beta}}_{n,j} - \widetilde{\beta}_j) &= \frac{\sqrt{n}}{n_1}\sum_{i=1}^{n} D_i(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 1]) - \frac{\sqrt{n}}{n_0}\sum_{i=1}^{n}(1 - D_i)(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 0]) \\
&= \frac{n}{n_1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 1]) \\
&\quad - \frac{n}{n_0}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1 - D_i)(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 0]) \\
&= \frac{1}{p_D}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 1]) \\
&\quad - \frac{1}{1 - p_D}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1 - D_i)(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 0]) + o_{\mathbb{P}}(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varphi_j(D_i, \widetilde{Y}_{i,j}) + o_{\mathbb{P}}(1), \qquad \text{(A.60)}
\end{aligned}$$

where the first line follows from (A.60), the third line uses $\frac{n_1}{n} - p_D = \frac{n_1}{n} - \mathbb{E}[D_i] = o_{\mathbb{P}}(1)$ and $\frac{n_0}{n} - (1 - p_D) = \frac{n_0}{n} - \mathbb{E}[1 - D_i] = o_{\mathbb{P}}(1)$ by the weak law of large numbers, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 1]) = O_{\mathbb{P}}(1)$, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1 - D_i)(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i = 1]) = O_{\mathbb{P}}(1)$, the seconds moments are finite by Assumption A.8, and that $o_{\mathbb{P}}(1)O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$, the last line uses $\varphi_j(D_i, \widetilde{Y}_{i,j}) \equiv \frac{D_i(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i=1])}{p_D} - \frac{(1-D_i)(\widetilde{Y}_{i,j} - \mathbb{E}[\widetilde{Y}_{i,j}|D_i=0])}{1-p_D}$.

Next, define $\boldsymbol{\varphi}_{\beta}(D_i, \widetilde{\boldsymbol{Y}}_i) \equiv (\varphi_1(D_i, \widetilde{Y}_{i,1}), \ldots, \varphi_q(D_i, \widetilde{Y}_{i,q}))'$. I can stack (A.60) across $j = 1, \ldots, q$ to get the following asymptotic linear representation:

$$\sqrt{n}(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{\varphi}_{\boldsymbol{\beta}}(D_i, \widetilde{\boldsymbol{Y}}_i) + o_{\mathbb{P}}(1).$$

△

A.7.4 Large-sample properties for the variance-minimizing weight

In this subsection, I show the large-sample properties of the optimal weights for the variance-minimization problem in (14).

Assumption A.25.

(a) Let $\widehat{\boldsymbol{\Sigma}}_n$ be a consistent estimator of $\boldsymbol{\Sigma}$.

(b) $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}_n$ are positive definite matrices.

(c) Suppose that

$$\sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\\ \text{vech}[\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}]\end{pmatrix}\xrightarrow{d}\begin{pmatrix}\boldsymbol{Z}_{\boldsymbol{\beta}}\\ \boldsymbol{Z}_{\text{vech}[\boldsymbol{\Sigma}]}\end{pmatrix}=\mathcal{N}\left(\begin{pmatrix}\mathbf{0}_q\\ \mathbf{0}_\ell\end{pmatrix},\begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{\Psi}'_{\boldsymbol{\Sigma},\boldsymbol{\beta}}\\ \boldsymbol{\Psi}_{\boldsymbol{\Sigma},\boldsymbol{\beta}} & \boldsymbol{\Psi}_{\boldsymbol{\Sigma}}\end{pmatrix}\right).$$

Let $\widehat{\boldsymbol{w}}_{\text{vmc},n}$ be the solution to (14) when $\boldsymbol{\Sigma}$ is replaced by the sample analog $\widehat{\boldsymbol{\Sigma}}_n$, i.e.,

$$\min_{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}}\ \boldsymbol{w}'\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{w}, \tag{A.61}$$

The class of weights $\mathcal{W}_{\text{cvx}}$ creates an additional complication for statistical inference as the weights are constrained to be nonnegative.

First, the following proposition shows that $\widehat{\boldsymbol{w}}_{\text{vmc},n}$ is consistent.

Proposition A.26. *Let Assumption A.25 hold. Let $\boldsymbol{w}_{\text{vmc}}$ and $\widehat{\boldsymbol{w}}_{\text{vmc},n}$ be as defined in (14) and (A.61), respectively. Then, $\widehat{\boldsymbol{w}}_{\text{vmc},n}\xrightarrow{p}\boldsymbol{w}_{\text{vmc}}$.*

Next, I derive the limiting distribution of the weights in the following proposition.

Theorem A.27. *Consider the same notations as in Proposition A.26. Define the random variable $\widetilde{\boldsymbol{Z}}\equiv 2(\boldsymbol{w}'_{\text{vmc}}\otimes\boldsymbol{I})\mathbf{B}\boldsymbol{Z}_{\text{vech}[\boldsymbol{\Sigma}]}$ where $\mathbf{B}$ is a duplication matrix such that $\mathbf{B}\,\text{vech}[\sqrt{n}(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_n)]=\text{vec}[\sqrt{n}(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_n)]$. Let $\boldsymbol{\mu}^\star\equiv(\mu_1^\star,\ldots,\mu_q^\star)'$ be the Lagrange multipliers to the problem (14). Denote $\mathcal{J}_{\text{active}}\subseteq\{1,\ldots,q\}$ as the set of indices for the active inequality constraints for $\mathcal{W}_{\text{cvx}}$ at $\boldsymbol{w}_{\text{vmc}}$ and $\mathcal{J}_{\text{active},+}(\boldsymbol{\mu}^\star)\equiv\{j\in\mathcal{J}_{\text{active}}:\mu_j^\star>0\}$. Let $\boldsymbol{h}^\star(\boldsymbol{\zeta})$ be the optimal solution to the following program:*

$$\begin{aligned}\min_{\boldsymbol{h}\in\mathbb{R}^q}\quad & \boldsymbol{h}'\boldsymbol{\zeta}+\boldsymbol{h}'\boldsymbol{\Sigma}\boldsymbol{h},\\ s.t.\quad & \boldsymbol{h}'\mathbf{1}_q=0,\\ & h_j=0\ \text{for}\ j\in\mathcal{J}_{\text{active},+}(\boldsymbol{\mu}^\star),\\ & h_j\geq 0\ \text{for}\ j\in\mathcal{J}_{\text{active}}\backslash\mathcal{J}_{\text{active},+}(\boldsymbol{\mu}^\star).\end{aligned} \tag{A.62}$$

Then,

$$\sqrt{n}(\widehat{\boldsymbol{w}}_{\text{vmc},n}-\boldsymbol{w}_{\text{vmc}})\xrightarrow{d}\boldsymbol{h}^\star(\widetilde{\boldsymbol{Z}}). \tag{A.63}$$

The limiting distribution is expressed in terms of a stochastic program (A.62). The following corollary shows the limiting distribution of $\widehat{\boldsymbol{w}}'_{\text{vmc},n}\widehat{\boldsymbol{\beta}}_n$.

Corollary A.28. *Consider the same notations and assumptions as in Theorem A.27. Then,*

$$\sqrt{n}(\widehat{\boldsymbol{w}}'_{\text{vmc},n}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}'_{\text{vmc}}\boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{w}'_{\text{vmc}}\boldsymbol{Z}_{\boldsymbol{\beta}} + \boldsymbol{\beta}'\boldsymbol{h}^{\star}(\widetilde{\boldsymbol{Z}}).$$

Finally, I show that a normal limiting distribution can be restored if the population solution $\boldsymbol{w}_{\text{vmc}}$ is all positive.

Corollary A.29. *Consider the same notations and assumptions as in Theorem A.27. Let* $\boldsymbol{w}_{\text{vmc}} > \boldsymbol{0}_q$.

(a) Define $\boldsymbol{B}_{\text{vmc}} \equiv -\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{I}_q - \boldsymbol{1}_q \frac{\boldsymbol{1}_q'\boldsymbol{\Sigma}^{-1}}{\boldsymbol{1}_q'\boldsymbol{\Sigma}^{-1}\boldsymbol{1}_q}\right)(\boldsymbol{w}'_{\text{vmc}} \otimes \boldsymbol{I})\mathbf{B}$. *Then,*

$$\sqrt{n}(\widehat{\boldsymbol{w}}_{\text{vmc},n} - \boldsymbol{w}_{\text{vmc}}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{B}_{\text{vmc}}\boldsymbol{\Psi}_{\boldsymbol{\Sigma}}\boldsymbol{B}'_{\text{vmc}}).$$

(b) Define $\sigma^2_{\text{vmc}} \equiv \boldsymbol{w}'_{\text{vmc}}\boldsymbol{\Sigma}\boldsymbol{w}_{\text{vmc}} + 2\boldsymbol{w}'_{\text{vmc}}\boldsymbol{\Psi}'_{\boldsymbol{\Sigma},\boldsymbol{\beta}}\boldsymbol{B}'_{\text{vmc}}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{B}_{\text{vmc}}\boldsymbol{\Psi}_{\boldsymbol{\Sigma}}\boldsymbol{B}'_{\text{vmc}}\boldsymbol{\beta}$. *Then,*

$$\sqrt{n}(\widehat{\boldsymbol{w}}'_{\text{vmc},n}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}'_{\text{vmc}}\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\text{vmc}}).$$

A.7.4.1 Proofs

Proof of Proposition A.26. This follows from Proposition B.12 with $\overline{M}(B, \boldsymbol{w}) = 0$. □

Proof of Theorem A.27. The problems (14) and (A.61) can be viewed as setting $\overline{M}(B, \boldsymbol{w}) = 0$ for the minimax problem (21) (or setting $B = 0$). In this case, I have $\boldsymbol{w}_{\text{vmc}} = \boldsymbol{w}^{\star}(0) = \overline{\boldsymbol{w}}(\boldsymbol{0}_q)$, where $\overline{\boldsymbol{w}}(\cdot)$ is defined as a solution to (B.75). Applying Proposition B.13 applies with $\overline{M}(B, \boldsymbol{w}) = 0$ gives

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{w}}_{\text{vmc},n} - \boldsymbol{w}_{\text{vmc}}) &= \sqrt{n}[\overline{\boldsymbol{w}}(\nabla \mathsf{d}_n(\boldsymbol{w}_{\text{vmc}})) - \overline{\boldsymbol{w}}(\boldsymbol{0}_q)] + o_{\mathbb{P}}(1) \\ &= \boldsymbol{D}\overline{\boldsymbol{w}}_0(\sqrt{n}\nabla \mathsf{d}_n(\boldsymbol{w}_{\text{vmc}})) + o_{\mathbb{P}}(1),\end{aligned}$$

where $\boldsymbol{D}\overline{\boldsymbol{w}}_0$ is the directional derivative at 0, and $\mathsf{d}_n(\cdot)$ is given in (B.76).

To compute the directional derivative $\boldsymbol{D}\overline{\boldsymbol{w}}_0(\cdot)$, I utilize the extra structure for this $B = 0$ case. In particular, problem (14) reduces to $\min_{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}} \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$. In this problem, the equality constraint in $\mathcal{W}_{\text{cvx}}$ is $\boldsymbol{w}'\boldsymbol{1}_q = 1$ (see (2)), so its gradient is $\boldsymbol{1}_q$. At any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, at least one component of $\boldsymbol{w}$ is nonzero. Hence, the gradient of the active inequality constraints (i.e., those such that $w_j = 0$) cannot be linearly dependent with the equality constraint in $\mathcal{W}_{\text{cvx}}$. Hence, linear independence constraint qualification (LICQ) is satisfied at the optimal solution (see, for instance, Nocedal and Wright (2006, Definitions 12.1

and 12.4)). Hence, the Lagrange multipliers are unique by Wachsmuth (2013).

Let $f(\boldsymbol{w}) \equiv \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$ and $\widehat{f}_n(\boldsymbol{w}) \equiv \boldsymbol{w}'\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{w}$ for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. Then,

$$\begin{aligned}
\sqrt{n}\nabla \mathsf{d}_n(\boldsymbol{w}) &= \sqrt{n}(\nabla \widehat{f}_n(\boldsymbol{w}) - \nabla f(\boldsymbol{w})) \\
&= 2\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{w} \\
&= 2(\boldsymbol{w}' \otimes \boldsymbol{I})\operatorname{vec}[\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})] \\
&= 2(\boldsymbol{w}' \otimes \boldsymbol{I})\mathbf{B}\operatorname{vech}[\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})] \\
&\xrightarrow{d} 2(\boldsymbol{w}' \otimes \boldsymbol{I})\mathbf{B}\boldsymbol{Z}_{\operatorname{vech}[\boldsymbol{\Sigma}]} \equiv \widetilde{\boldsymbol{Z}},
\end{aligned}$$

where the third line used a duplication matrix $\mathbf{B}$ such that $\mathbf{B}\operatorname{vech}[\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})] = \operatorname{vec}[\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})]$, and the derivation uses properties of the Kronecker product and vectorizations (see, for instance, Example 18.5 of Magnus and Neudecker (2019)) and the continuous mapping theorem. The definition of $\widetilde{\boldsymbol{Z}}$ follows from the one in the statement of the theorem. The directional derivative $\boldsymbol{D}\overline{\boldsymbol{w}}_0(\boldsymbol{\zeta})$ is given by $\boldsymbol{h}^\star(\boldsymbol{\zeta})$ in (A.62). It is also unique because $\boldsymbol{\Sigma}$ is assumed to be positive definite. Then, the convergence in distribution follows page 163 of Shapiro et al. (2021). □

Proof of Corollary A.28. Using Proposition B.13 and Theorem A.27,

$$\begin{aligned}
\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \\ (\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{w}_{\text{vmc}} \end{pmatrix} &= \sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \\ (\boldsymbol{w}_{\text{vmc}}' \otimes \boldsymbol{I}_q)\operatorname{vec}[\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}] \end{pmatrix} \\
&= \sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \\ (\boldsymbol{w}_{\text{vmc}}' \otimes \boldsymbol{I}_q)\mathbf{B}\operatorname{vech}[\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}] \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{I}_q & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{w}_{\text{vmc}}' \otimes \boldsymbol{I}_q)\mathbf{B} \end{pmatrix}\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \\ \operatorname{vech}[\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}] \end{pmatrix} \\
&\xrightarrow{d} \begin{pmatrix} \boldsymbol{Z}_{\boldsymbol{\beta}} \\ (\boldsymbol{w}_{\text{vmc}}' \otimes \boldsymbol{I}_q)\mathbf{B}\boldsymbol{Z}_{\operatorname{vech}[\boldsymbol{\Sigma}]} \end{pmatrix},
\end{aligned} \tag{A.64}$$

where the first line uses the properties of the Kronecker product and vectorizations (see, for instance, Example 18.5 of Magnus and Neudecker (2019)), the second line uses the duplication matrix $\mathbf{B}$ as in Theorem A.27, and the last line uses the convergence of distribution in Assumption A.25. Thus,

$$\sqrt{n}(\widehat{\boldsymbol{w}}_{\text{vmc},n}'\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}_{\text{vmc}}'\boldsymbol{\beta}) = \sqrt{n}[\widehat{\boldsymbol{w}}_{\text{vmc},n}'(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) + \boldsymbol{\beta}'(\widehat{\boldsymbol{w}}_{\text{vmc},n} - \boldsymbol{w}_{\text{vmc}})]$$

$$
\begin{aligned}
&= \begin{pmatrix} \widehat{\boldsymbol{w}}'_{\text{vmc},n} & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}(\widehat{\boldsymbol{w}}_{\text{vmc},n} - \boldsymbol{w}_{\text{vmc}}) \end{pmatrix} \\
&\xrightarrow{d} \boldsymbol{w}'_{\text{vmc}} \boldsymbol{Z}_{\boldsymbol{\beta}} + \boldsymbol{\beta}' \boldsymbol{h}^\star \left(\widetilde{\boldsymbol{Z}} \right),
\end{aligned}
$$

where the last line follows from (A.64). □

Proof of Corollary A.29(a). In this case, none of the inequality constraints are active. Hence, (A.62) reduces to

$$
\begin{aligned}
\min_{\boldsymbol{h} \in \mathbb{R}^q} \quad & \boldsymbol{h}' \boldsymbol{\zeta} + \boldsymbol{h}' \boldsymbol{\Sigma} \boldsymbol{h}, \\
\text{s.t.} \quad & \boldsymbol{h}' \mathbf{1}_q = 0.
\end{aligned}
\tag{A.65}
$$

Let the Lagrangian of (A.65) be

$$
\mathcal{L} = \boldsymbol{h}' \boldsymbol{\zeta} + \boldsymbol{h}' \boldsymbol{\Sigma} \boldsymbol{h} + \kappa (\mathbf{1}_q' \boldsymbol{h}),
$$

where κ is the Lagrangian multiplier. The first-order condition with respect to $\boldsymbol{h}$ leads to

$$
0 = \frac{\partial \mathcal{L}}{\partial \boldsymbol{h}} = \boldsymbol{\zeta} + 2 \boldsymbol{\Sigma} \boldsymbol{h} + \kappa \mathbf{1}_q.
$$

Rearranging gives the solution $\boldsymbol{h}_0^\star(\boldsymbol{\zeta}) = -\frac{1}{2} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\zeta} + \kappa \mathbf{1}_q)$. But $\boldsymbol{h}_0^\star(\boldsymbol{\zeta})' \mathbf{1}_q = 0$, so this gives $-\boldsymbol{\zeta}' \boldsymbol{\Sigma}^{-1} \mathbf{1}_q - \kappa \mathbf{1}_q' \boldsymbol{\Sigma}^{-1} \mathbf{1}_q = 0$, or equivalently,

$$
\kappa = -\frac{\mathbf{1}_q' \boldsymbol{\Sigma}^{-1} \boldsymbol{\zeta}}{\mathbf{1}_q' \boldsymbol{\Sigma}^{-1} \mathbf{1}_q}.
$$

Hence,

$$
\boldsymbol{h}_0^\star(\boldsymbol{\zeta}) = -\frac{1}{2} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\zeta} - \frac{\mathbf{1}_q' \boldsymbol{\Sigma}^{-1} \boldsymbol{\zeta}}{\mathbf{1}_q' \boldsymbol{\Sigma}^{-1} \mathbf{1}_q} \mathbf{1}_q \right) = -\frac{1}{2} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{I}_q - \mathbf{1}_q \frac{\mathbf{1}_q' \boldsymbol{\Sigma}^{-1}}{\mathbf{1}_q' \boldsymbol{\Sigma}^{-1} \mathbf{1}_q} \right) \boldsymbol{\zeta}.
$$

Recall from the statement of (A.27) that $\widetilde{\boldsymbol{Z}} \equiv 2 (\boldsymbol{w}'_{\text{vmc}} \otimes \boldsymbol{I}) \mathbf{B} \boldsymbol{Z}_{\text{vech}[\boldsymbol{\Sigma}]}$. Then,

$$
\sqrt{n}(\widehat{\boldsymbol{w}}_{\text{vmc},n} - \boldsymbol{w}_{\text{vmc}}) \xrightarrow{d} \boldsymbol{h}_0^\star(\widetilde{\boldsymbol{Z}}) = -\frac{1}{2} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{I}_q - \mathbf{1}_q \frac{\mathbf{1}_q' \boldsymbol{\Sigma}^{-1}}{\mathbf{1}_q' \boldsymbol{\Sigma}^{-1} \mathbf{1}_q} \right) \widetilde{\boldsymbol{Z}}.
$$

The result follows from defining the $\boldsymbol{B}_{\text{vmc}}$ as in the statement of this corollary. □

Proof of Corollary A.29(b). Using Proposition B.13, Corollary A.28, and Corollary A.29(a), I can write

$$\begin{aligned}
&\sqrt{n}(\widehat{\boldsymbol{w}}'_{\text{vmc},n}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{w}'_{\text{vmc}}\boldsymbol{\beta}) \\
&\quad= \sqrt{n}[\widehat{\boldsymbol{w}}'_{\text{vmc},n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) + \boldsymbol{\beta}'(\widehat{\boldsymbol{w}}_{\text{vmc},n} - \boldsymbol{w}_{\text{vmc}})] \\
&\quad= \begin{pmatrix} \widehat{\boldsymbol{w}}'_{\text{vmc},n} & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \sqrt{n}(\widehat{\boldsymbol{w}}_{\text{vmc},n} - \boldsymbol{w}_{\text{vmc}}) \end{pmatrix} \\
&\quad= \begin{pmatrix} \widehat{\boldsymbol{w}}'_{\text{vmc},n} & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \\ \boldsymbol{B}_{\text{vmc}}\sqrt{n}\,\text{vech}[\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}] \end{pmatrix} + o_{\mathbb{P}}(1) \\
&\quad\xrightarrow{d} \begin{pmatrix} \boldsymbol{w}'_{\text{vmc}} & \boldsymbol{\beta}' \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_{\boldsymbol{\beta}} \\ \boldsymbol{B}_{\text{vmc}}\boldsymbol{Z}_{\text{vech}[\boldsymbol{\Sigma}]} \end{pmatrix} \\
&\quad\sim \mathcal{N}(0, \boldsymbol{w}'_{\text{vmc}}\boldsymbol{\Sigma}\boldsymbol{w}_{\text{vmc}} + 2\boldsymbol{w}'_{\text{vmc}}\boldsymbol{\Psi}'_{\boldsymbol{\Sigma},\boldsymbol{\beta}}\boldsymbol{B}'_{\text{vmc}}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{B}_{\text{vmc}}\boldsymbol{\Psi}_{\boldsymbol{\Sigma}}\boldsymbol{B}'_{\text{vmc}}\boldsymbol{\beta}),
\end{aligned} \tag{A.66}$$

by Assumption A.25. □

A.8 Additional simulations related to Proposition 3.6

The goal of this section is to examine the impact of $\boldsymbol{c}'\boldsymbol{w}_{\text{pca}}$ being close to binding. The DGP is as follows. I assume that there are $q = 25$ outcomes that follow the linear model

$$Y_{i,j} = \xi_j + \beta_j D_i + U_{i,j},$$

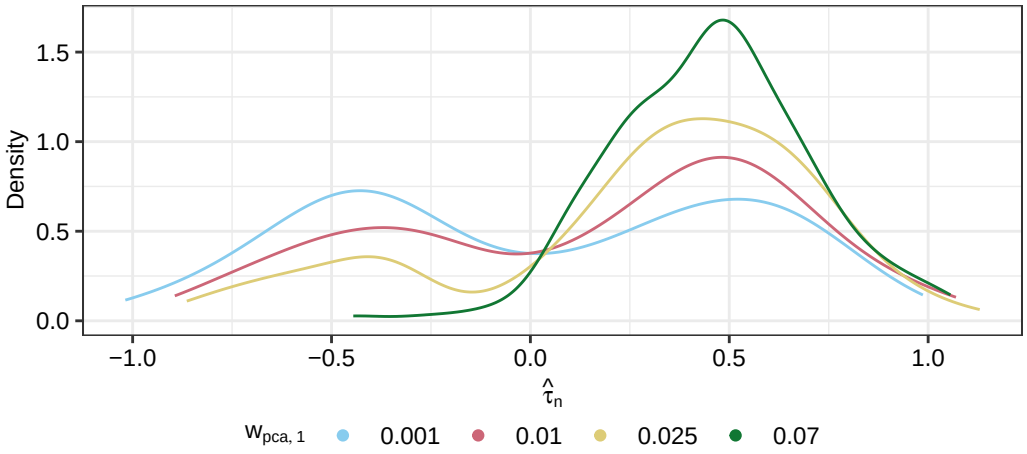
as in (8). The treatment variable D_i is binary such that $\mathbb{P}[D_i = 1] = 0.3$. The error terms follow $\boldsymbol{U}_i \sim \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{\Sigma}_U)$ and $\boldsymbol{U}_i \perp\!\!\!\perp D_i$. To have a more realistic DGP, I calibrate $\boldsymbol{\Sigma}_U$ using the data from Bau (2022). I set the variance matrix of $\boldsymbol{Y}_i$ such that it matches the correlation matrix for the outcomes from the asset index in Bau (2022).

I assume that PCA is used to aggregate the outcomes, and that $\boldsymbol{c} = \boldsymbol{e}_1$ (i.e., unit vector where the first entry equals 1 and the rest equals 0) is used in the sign normalization constraint for the PCA problem. I set $\boldsymbol{\beta}$ such that $\beta_j = 0.5$ for $j = 16, \ldots, 20$ and $\beta_j = 0$ otherwise. In addition, I consider different DGPs on the correlation matrix of $\boldsymbol{Y}_i$ that replaces $\text{Corr}[Y_{i,1}, Y_{i,2}]$ with $\text{Corr}[Y_{i,1}, Y_{i,2}] + \omega$ using the values shown in Table 3. I show the value of $w_{\text{pca},1}$ corresponding to each value of ω. Such changes allow me to obtain different values of the first entry in the leading eigenvector. In the following, I write $\tau_0 = \boldsymbol{w}'_{\text{pca}}\boldsymbol{\beta}$ as the value of the target parameter implied by the DGP and $\widehat{\tau}_n = \widehat{\boldsymbol{w}}'_{\text{pca},n}\widehat{\boldsymbol{\beta}}_n$ as the corresponding estimator.

Table 3: The values of ω used in the DGPs.

$\boldsymbol{\omega}$	-0.518	-0.100	0.042	0.127
$\boldsymbol{w}_{\text{pca},\mathbf{1}}$	0.070	0.025	0.010	0.001

Figure A.2: Distribution of $\widehat{\tau}_n$ under different DGPs.

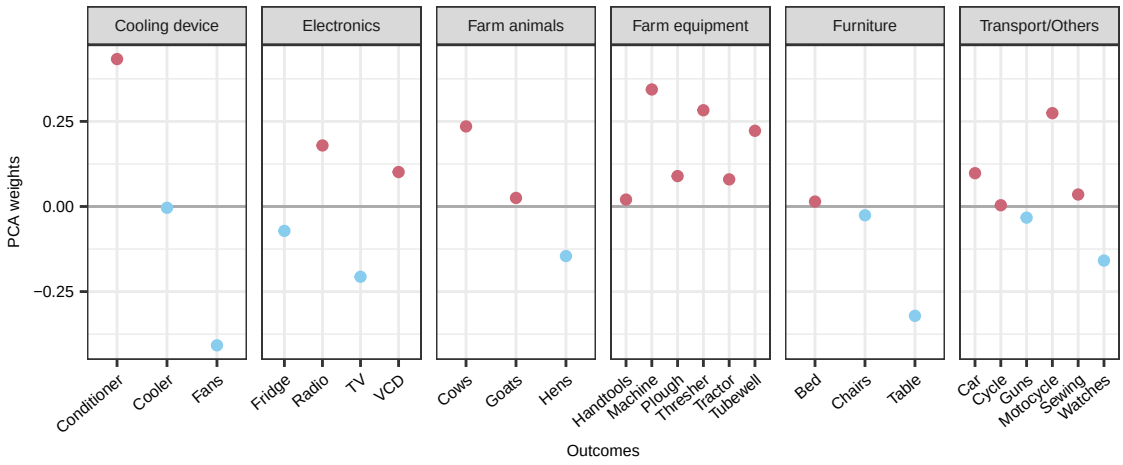


The simulations are based on 1,000 replications. Figure A.2 shows the distribution of $\widehat{\boldsymbol{w}}'_{\text{pca},n}\widehat{\boldsymbol{\beta}}_n$ for each of the DGPs in which the distribution is nonstandard when $w_{\text{pca},1}$ is close to 0. In particular, as $w_{\text{pca},1}$ becomes closer to 0, the distribution of $\widehat{\tau}_n$ becomes nonstandard (and bimodal).

A.9 Supplemental details on empirical examples

Figure A.3 shows the weights on the asset variables in Bau (2022).

Figure A.3: Weights on the asset variables in Bau (2022).



B Appendix for Section 4

B.1 Details for the asymptotic validity for FLCI

In this appendix, I provide the main results regarding the asymptotic validity for FLCI. Let $\mathcal{P}$ be the class of distributions on $\widehat{\boldsymbol{\beta}}_n$. In addition, let $\mathfrak{S}_n(B) \equiv \{(\theta, P) \in \Theta \times \mathcal{P} : \sqrt{n}(\boldsymbol{\beta}_P - \theta \mathbf{1}_q) \in \mathcal{S}(B)\}$. I will require the following assumptions on uniformity.

Assumption B.1.

(a) $\widehat{\boldsymbol{\Sigma}}_n$ *is uniformly consistent for* $\boldsymbol{\Sigma}_{\theta,P}$ *for any* $(\theta, P) \in \mathfrak{S}_n(B)$.

(b) For any $(\theta, P) \in \mathfrak{S}_n(B)$, $\boldsymbol{\Sigma}_{\theta,P} \in \mathcal{M}$, *where* $\mathcal{M}$ *is a compact set of positive definite matrices with eigenvalues bounded below by* $\lambda_{\text{lb}} > 0$ *and above by* $\lambda_{\text{ub}} < \infty$ *where* $\lambda_{\text{ub}} > \lambda_{\text{lb}}$.

(c) $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_P)$ *converges in distribution to* $\mathcal{N}(0, \boldsymbol{\Sigma}_{\theta,P})$ *uniformly under* $(\theta, P) \in \mathfrak{S}_n(B)$.

I maintain the following assumption about the amount of misspecification. It is an asymptotic device to control the misspecification and should not be interpreted as $\boldsymbol{\beta}_P$ being equal to θ as $n \longrightarrow \infty$.

Assumption B.2. *Let* $P \in \mathcal{P}$ *and* $\mathcal{S}(B)$ *be the parameter space as in Section 4.2 for* $B \geq 0$. *Then,* $\boldsymbol{\beta}_P = \theta \mathbf{1}_q + \boldsymbol{b}_n$ *where* $\boldsymbol{b}_n = n^{-1/2}\widetilde{\boldsymbol{b}}$ *and* $\widetilde{\boldsymbol{b}} \in \mathcal{S}(B)$ *holds.*

The following theorem establishes the validity of the FLCI.

Proposition B.3. *Let* $\alpha \in (0, 1)$. *Let Assumptions 2.2, 4.9, B.1, and B.2 hold, and* $\widehat{\boldsymbol{\Sigma}}_n$ *be an estimator of* $\boldsymbol{\Sigma}$ *that satisfies the preceding assumption.*

For a given finite $B \geq 0$, *let* $\widehat{\boldsymbol{w}}_n$ *be the weight estimated from the minimax problem using* B *or the adaptive problem over* $\mathcal{B} = [\underline{B}, B]$, *with* $\boldsymbol{\Sigma}$ *replaced by* $\widehat{\boldsymbol{\Sigma}}_n$. *Let* $\widehat{c}_{\alpha,n}$ *be the smallest critical*

value that satisfies (30) *that uses* $\widehat{\mathbf{\Sigma}}_n$ *instead of* $\mathbf{\Sigma}$.

Write $\mathcal{I}_n \equiv \left[\widehat{\boldsymbol{w}}_n'\widehat{\boldsymbol{\beta}}_n \pm \widehat{c}_{\alpha,n}\sqrt{\frac{\widehat{\boldsymbol{w}}_n'\widehat{\mathbf{\Sigma}}_n\widehat{\boldsymbol{w}}_n}{n}}\right]$ *as the sample analog of the confidence interval in* (29). *Then,* $\mathcal{I}_n$ *is asymptotically valid, i.e.,*

$$\liminf_{n\to\infty} \inf_{(\theta,P)\in\mathfrak{S}_n(B)} \mathbb{P}[\theta \in \mathcal{I}_n] \geq 1-\alpha.$$

B.2 Proofs for propositions in the main text

Proof of Proposition 4.10(a). This part of the proposition characterizes the shape of $A(B,\boldsymbol{w})$ over $B \in \mathcal{B}$ for a given $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. By direct computation, the derivative of $A(B,\boldsymbol{w})$ with respect to B^2 is given by

$$\frac{\partial A(B,\boldsymbol{w})}{\partial(B^2)} = \frac{R^\star(B)\frac{\partial R_{\max}(B,\boldsymbol{w})}{\partial(B^2)} - R_{\max}(B,\boldsymbol{w})\frac{\partial R^\star(B)}{\partial(B^2)}}{[R^\star(B)]^2} \equiv \frac{N(B,\boldsymbol{w})}{[R^\star(B)]^2}, \tag{B.1}$$

for any given $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ where

$$N(B,\boldsymbol{w}) \equiv R^\star(B)\frac{\partial R_{\max}(B,\boldsymbol{w})}{\partial(B^2)} - R_{\max}(B,\boldsymbol{w})\frac{\partial R^\star(B)}{\partial(B^2)}, \tag{B.2}$$

is the numerator of $\frac{\partial A(B,\boldsymbol{w})}{\partial(B^2)}$ in (B.1).

From Proposition B.6, there are only three possibilities on the shape of $A(B,\boldsymbol{w})$ against $B \in [\underline{B},\overline{B}]$ for a given $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$.

First, if $N(0,\boldsymbol{w}) \geq 0$, then $N(B,\boldsymbol{w}) \geq 0$ for any $B \geq 0$. Thus, $\frac{\partial A(B,\boldsymbol{w})}{\partial(B^2)} \geq 0$, i.e., $A(B,\boldsymbol{w})$ is nondecreasing in B.

Second, if $N(0,\boldsymbol{w}) \leq 0$, then there are two scenarios to consider. The first scenario is that $N(B,\boldsymbol{w}) \leq 0$ for any $B \geq 0$. Hence, $\frac{\partial A(B,\boldsymbol{w})}{\partial(B^2)} \leq 0$, i.e., $A(B,\boldsymbol{w})$ is nonincreasing in B.

The remaining scenario is that there exists B_0 such that $N(B,\boldsymbol{w}) \leq 0$ for any $B \in [0,B_0)$ and $N(B,\boldsymbol{w}) \geq 0$ for any $B \in [B_0,\infty)$. This means $A(B,\boldsymbol{w})$ is nonincreasing in B for any $B \in [0,B_0)$ and nondecreasing in B for any $B \in [B_0,\infty)$. This completes the proof. □

Proof of Proposition 4.10(b). In all the possible cases discussed in Proposition 4.10(a), the maximum of $A(B,\boldsymbol{w})$ over $B \in [\underline{B},\overline{B}]$ where $\overline{B} \geq \underline{B} \geq 0$ must be achieved at the endpoints. □

Proof of Proposition 4.11. To begin with, since $\underline{B}$ and $\overline{B}$ are held fixed in the proof, I de-

fine $\underline{A}(\boldsymbol{w}) \equiv A(\underline{B}, \boldsymbol{w})$ and $\overline{A}(\boldsymbol{w}) \equiv A(\overline{B}, \boldsymbol{w})$ for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ in this proof. I also define $A_{\max}(\boldsymbol{w}) \equiv A_{\max}(\mathcal{B}, \boldsymbol{w}) = \max\{\underline{A}(\boldsymbol{w}), \overline{A}(\boldsymbol{w})\}$ in this proof. Using the assumptions on the adaptive regret and the definition of the optimally adaptive weight in (26), I have $\boldsymbol{w}_A = \arg\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} A_{\max}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} \max\{\underline{A}(\boldsymbol{w}), \overline{A}(\boldsymbol{w})\}$.

Assume to the contrary that $\underline{A}(\boldsymbol{w}_A) \neq \overline{A}(\boldsymbol{w}_A)$. Set

$$\underline{A}(\boldsymbol{w}_A) > \overline{A}(\boldsymbol{w}_A) \tag{B.3}$$

without loss of generality. This means $A_{\max}(\boldsymbol{w}_A) = \underline{A}(\boldsymbol{w}_A)$. Define

$$\Delta_A \equiv \underline{A}(\boldsymbol{w}_A) - \overline{A}(\boldsymbol{w}_A) > 0. \tag{B.4}$$

Let $\boldsymbol{w}_1 \equiv \arg\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} \underline{A}(\boldsymbol{w})$. Note that $\boldsymbol{w}_1$ is unique since $\underline{A}(\boldsymbol{w})$ is strictly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. In addition, $\underline{A}(\boldsymbol{w}_1) = 1$ from the definition of adaptive regret. This follows because $\underline{A}(\boldsymbol{w}) \equiv \frac{R_{\max}(\underline{B}, \boldsymbol{w})}{\min_{\boldsymbol{u} \in \mathcal{W}_{\text{cvx}}} R_{\max}(\underline{B}, \boldsymbol{u})}$. Thus, $\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} \underline{A}(\boldsymbol{w}) = \frac{\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} R_{\max}(\underline{B}, \boldsymbol{w})}{\min_{\boldsymbol{u} \in \mathcal{W}_{\text{cvx}}} R_{\max}(\underline{B}, \boldsymbol{u})} = 1$. For the same argument, $\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} \overline{A}(\boldsymbol{w}) = 1$ as well. Since $\boldsymbol{w}_A \in \mathcal{W}_{\text{cvx}}$, it follows that $\underline{A}(\boldsymbol{w}_A) \geq \min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} \underline{A}(\boldsymbol{w})$ and $\overline{A}(\boldsymbol{w}_A) \geq \min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} \overline{A}(\boldsymbol{w})$. Together with $\underline{A}(\boldsymbol{w}_A) > \overline{A}(\boldsymbol{w}_A)$ in (B.3), it follows that

$$\underline{A}(\boldsymbol{w}_A) > \overline{A}(\boldsymbol{w}_A) \geq \min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} \overline{A}(\boldsymbol{w}) = 1. \tag{B.5}$$

From the above, it follows that $\boldsymbol{w}_A \neq \boldsymbol{w}_1$ because $\underline{A}(\boldsymbol{w}_1) = 1$ and $\underline{A}(\boldsymbol{w}_A) > 1$.

Since $\underline{A}(\boldsymbol{w})$ is assumed to be strictly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ in the proposition, it follows that for any $t \in (0, 1)$, the following holds

$$\underline{A}((1-t)\boldsymbol{w}_A + t\boldsymbol{w}_1) < (1-t)\underline{A}(\boldsymbol{w}_A) + t\underline{A}(\boldsymbol{w}_1) = (1-t)\underline{A}(\boldsymbol{w}_A) + t \leq \underline{A}(\boldsymbol{w}_A), \tag{B.6}$$

where the last inequality follows from (B.5).

Next, $\overline{A}(\boldsymbol{w})$ is continuous in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. In particular, it is continuous at $\boldsymbol{w}_A$. This means that for any $\varepsilon > 0$, there exists $\varphi_\epsilon > 0$ such that for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, $\|\boldsymbol{w} - \boldsymbol{w}_A\| < \varphi_\epsilon$ implies $|\overline{A}(\boldsymbol{w}) - \overline{A}(\boldsymbol{w}_A)| < \varepsilon$. Set $t_\varepsilon = \frac{\varphi_\epsilon}{2\|\boldsymbol{w}_1 - \boldsymbol{w}_A\|} > 0$. From the discussion two paragraphs ago, $\boldsymbol{w}_1 \neq \boldsymbol{w}_A$, so this choice of t_ε is well-defined. If such t_ε leads to $t_\varepsilon > 1$, divide it by a large enough number such that $t_\varepsilon \in (0, 1)$. This value of t_ε will satisfy

$$\|[(1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1] - \boldsymbol{w}_A\| = \|-t_\varepsilon \boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1\| = t_\varepsilon \|\boldsymbol{w}_1 - \boldsymbol{w}_A\| = \frac{\varphi_\varepsilon}{2} < \varphi_\varepsilon.$$

Hence, for such t_ε, $|\overline{A}((1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1) - \overline{A}(\boldsymbol{w}_A)| < \epsilon$. In addition, $(1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1 \in \mathcal{W}_{\text{cvx}}$. This follows because $t_\varepsilon > 0$, $\boldsymbol{w}_A, \boldsymbol{w}_1 \in \mathcal{W}_{\text{cvx}}$, so $(1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1 \geq \mathbf{0}_q$, and that $[(1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1]'\mathbf{1}_q = (1-t_\varepsilon)\boldsymbol{w}_A'\mathbf{1}_q + t_\varepsilon \boldsymbol{w}_1'\mathbf{1}_q = (1-t_\varepsilon) + t_\varepsilon = 1$. This means

$$\overline{A}(\boldsymbol{w}_A) - \varepsilon < \overline{A}((1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1) < \overline{A}(\boldsymbol{w}_A) + \varepsilon. \tag{B.7}$$

Set $\varepsilon = \frac{\Delta_A}{4} > 0$, so that

$$\overline{A}((1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1) < \overline{A}(\boldsymbol{w}_A) + \varepsilon = \underline{A}(\boldsymbol{w}_A) - \Delta_A + \varepsilon = \underline{A}(\boldsymbol{w}_A) - \frac{3}{4}\Delta_A, \tag{B.8}$$

where the first inequality follows from (B.7), the first equality follows from (B.4), and the last equality follows from the choice of ε.

Combining the results in (B.6) and (B.8), I have

$$\begin{aligned} A_{\max}((1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1) &= \max\{\underline{A}((1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1), \overline{A}((1-t_\varepsilon)\boldsymbol{w}_A + t_\varepsilon \boldsymbol{w}_1)\} \\ &< \max\left\{\underline{A}(\boldsymbol{w}_A), \underline{A}(\boldsymbol{w}_A) - \frac{3}{4}\Delta_A\right\} \\ &\leq \underline{A}(\boldsymbol{w}_A), \end{aligned}$$

which shows that $\boldsymbol{w}_A$ cannot minimize $A_{\max}(\boldsymbol{w})$. The argument for assuming $\underline{A}(\boldsymbol{w}_A) < \overline{A}(\boldsymbol{w}_A)$ is similar in (B.3). This completes the proof. □

Proof of Proposition B.3. Let $\boldsymbol{w}^\star(B, \boldsymbol{\Sigma})$ be the solution to the minimax problem or $\boldsymbol{w}_A(\boldsymbol{\Sigma})$ be the solution to the adaptive problem over $\mathcal{B} = [\underline{B}, B]$. Lemma B.7 showed that $\boldsymbol{w}^\star(B, \boldsymbol{\Sigma})$ is uniformly continuous in $\boldsymbol{\Sigma}$. Lemma B.8 showed that $\boldsymbol{w}_A(\boldsymbol{\Sigma})$ is uniformly continuous in $\boldsymbol{\Sigma}$. Here, $\widehat{\boldsymbol{w}}_n$ is either $\boldsymbol{w}^\star(B, \widehat{\boldsymbol{\Sigma}}_n)$ or $\boldsymbol{w}_A(\widehat{\boldsymbol{\Sigma}}_n)$. Then, I can write

$$\begin{aligned} \sqrt{n}(\widehat{\tau}_n - \theta) &= \sqrt{n}(\widehat{\boldsymbol{w}}'\widehat{\boldsymbol{\beta}}_n - \theta) \\ &= \widehat{\boldsymbol{w}}_n'\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_P) + \widehat{\boldsymbol{w}}_n'\sqrt{n}(\boldsymbol{\beta}_P - \theta\mathbf{1}_q) \\ &= \widehat{\boldsymbol{w}}_n'\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_P) + \widehat{\boldsymbol{w}}_n'\widetilde{\boldsymbol{b}}, \end{aligned}$$

where the first line follows from the definition of the estimator $\widehat{\tau}_n$, the second line follows from adding and subtracting and that $\widehat{\boldsymbol{w}}_n \in \mathcal{W}_{\text{cvx}}$, and the third line follows form Assumption B.2. In the above, the uniform consistency of $\widehat{\boldsymbol{w}}_n$ follows from the beginning of the proof. The convergence of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_P)$ follows from Assumption B.1 and the asymptotic variance is also bounded away from zero. The proof then follows from applying Appendix C of Armstrong and Kolesár (2021b) in the current context. □

B.3 Supplemental results and proofs

Lemma B.4. *Let Assumption 4.9 hold. Let $\boldsymbol{w}^\star(B)$ be the solution to the minimax problem (21). Then, for any $B_1, B_2 \geq 0$,*

$$R^\star(B_2) \leq R^\star(B_1) + (B_2^2 - B_1^2) m(\boldsymbol{w}^\star(B_1)).$$

Proof of Lemma B.4. Recall that the minimax problem (21) has a unique solution for each $B \geq 0$. Then, for any $B_1, B_2 \geq 0$, I have

$$\begin{aligned} R^\star(B_2) &= V(\boldsymbol{w}^\star(B_2)) + B_2^2 m(\boldsymbol{w}^\star(B_2)) \\ &\leq V(\boldsymbol{w}^\star(B_1)) + B_2^2 m(\boldsymbol{w}^\star(B_1)) \\ &= V(\boldsymbol{w}^\star(B_1)) + B_1^2 m(\boldsymbol{w}^\star(B_1)) + (B_2^2 - B_1^2) m(\boldsymbol{w}^\star(B_1)) \\ &= R^\star(B_1) + (B_2^2 - B_1^2) m(\boldsymbol{w}^\star(B_1)), \end{aligned}$$

where the first line uses the definition of minimax risk from (21), the second line follows because $\boldsymbol{w}^\star(B_1)$ is not an optimal solution to the minimax problem for $B = B_2$, the third line adds and subtracts $B_1^2 m(\boldsymbol{w}^\star(B_1))$, and the last line follows from the definition of the minimax risk from (21) again. □

Lemma B.5. *Consider the same assumptions and notations as in Lemma B.4. Then, $m(\boldsymbol{w}^\star(B))$ is nonincreasing in B.*

Proof of Lemma B.5. For any $B_1 > B_2 \geq 0$ (I consider strict inequality to make sure the division below is well-defined), I have

$$R^\star(B_1) \leq R^\star(B_2) + (B_1^2 - B_2^2) m(\boldsymbol{w}^\star(B_2)), \tag{B.9}$$

$$R^\star(B_2) \leq R^\star(B_1) + (B_2^2 - B_1^2) m(\boldsymbol{w}^\star(B_1)). \tag{B.10}$$

from Lemma B.4. Since $B_1^2 - B_2^2 > 0$, the above inequalities imply that

$$m(\boldsymbol{w}^\star(B_2)) \geq \frac{R^\star(B_1) - R^\star(B_2)}{B_1^2 - B_2^2} \geq m(\boldsymbol{w}^\star(B_1)),$$

where the first inequality uses (B.9) and the second inequality uses (B.10). □

Proposition B.6. *Let Assumption 4.9 hold. Consider the minimax and adaptive problems defined in (21) and (24) respectively. Then, for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, $N(B, \boldsymbol{w})$ defined in (B.2) is nondecreasing in $B \geq 0$.*

Proof of Proposition B.6. To begin with, note that

$$\left.\frac{\partial R^\star(B)}{\partial(B^2)}\right|_{\boldsymbol{w}=\boldsymbol{w}^\star(B)} = m(\boldsymbol{w}^\star(B)),$$

by Assumption 4.9 and Danskin's theorem (see, for instance, Bertsekas (2009)) by noting that for each given $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, $R_{\max}(\boldsymbol{w}, B) = V(\boldsymbol{w}) + (B^2)m(\boldsymbol{w})$ is affine in (B^2), $\mathcal{W}_{\text{cvx}}$ is compact, and that $\boldsymbol{w}^\star(B)$ is unique.

Since $\overline{M}(B, \boldsymbol{u}) = B^2 m(\boldsymbol{u})$ for any $B \geq 0$ and $\boldsymbol{u} \in \mathcal{W}_{\text{cvx}}$, then $\frac{\partial \overline{M}(B,\boldsymbol{u})}{\partial(B^2)} = m(\boldsymbol{u})$ and $\frac{\partial \overline{M}(B,\boldsymbol{w}^\star(B))}{\partial(B^2)} = m(\boldsymbol{w}^\star(B))$. Hence, $N(B, \boldsymbol{u})$ in (B.2) can be written as

$$N(B, \boldsymbol{u}) = R^\star(B)m(\boldsymbol{u}) - R_{\max}(B, \boldsymbol{u})m(\boldsymbol{w}^\star(B)). \tag{B.11}$$

Consider any $B_1 > B_2 \geq 0$, note that

$$\begin{aligned} B_1^2 m(\boldsymbol{w}^\star(B_1)) - B_2^2 m(\boldsymbol{w}^\star(B_2)) &= [B_2^2 + (B_1^2 - B_2^2)]m(\boldsymbol{w}^\star(B_1)) - B_2^2 m(\boldsymbol{w}^\star(B_2)) \\ &= B_2^2[m(\boldsymbol{w}^\star(B_1)) - m(\boldsymbol{w}^\star(B_2))] \\ &\quad + (B_1^2 - B_2^2)m(\boldsymbol{w}^\star(B_1)). \end{aligned} \tag{B.12}$$

For any $\boldsymbol{u} \in \mathcal{W}_{\text{cvx}}$ and $B_1 \geq B_2 \geq 0$, I have

$$\begin{aligned} N(B_1, \boldsymbol{u}) - N(B_2, \boldsymbol{u}) &= [R^\star(B_1)m(\boldsymbol{u}) - R_{\max}(B_1, \boldsymbol{u})m(\boldsymbol{w}^\star(B_1))] \\ &\quad - [R^\star(B_2)m(\boldsymbol{u}) - R_{\max}(B_2, \boldsymbol{u})m(\boldsymbol{w}^\star(B_2))] \\ &= m(\boldsymbol{u})[R^\star(B_1) - R^\star(B_2)] - R_{\max}(B_1, \boldsymbol{u})m(\boldsymbol{w}^\star(B_1)) \\ &\quad + R_{\max}(B_2, \boldsymbol{u})m(\boldsymbol{w}^\star(B_2)) \\ &= m(\boldsymbol{u})[R^\star(B_1) - R^\star(B_2)] - [V(\boldsymbol{u}) + B_1^2 m(\boldsymbol{u})]m(\boldsymbol{w}^\star(B_1)) \\ &\quad + [V(\boldsymbol{u}) + B_2^2 m(\boldsymbol{u})]m(\boldsymbol{w}^\star(B_2)) \\ &= m(\boldsymbol{u})[R^\star(B_1) - R^\star(B_2)] - V(\boldsymbol{u})[m(\boldsymbol{w}^\star(B_1)) - m(\boldsymbol{w}^\star(B_2))] \\ &\quad - m(\boldsymbol{u})[B_1^2 m(\boldsymbol{w}^\star(B_1)) - B_2^2 m(\boldsymbol{w}^\star(B_2))] \\ &= m(\boldsymbol{u})[R^\star(B_1) - R^\star(B_2)] - V(\boldsymbol{u})[m(\boldsymbol{w}^\star(B_1)) - m(\boldsymbol{w}^\star(B_2))] \\ &\quad - m(\boldsymbol{u})B_2^2[m(\boldsymbol{w}^\star(B_1)) - m(\boldsymbol{w}^\star(B_2))] \\ &\quad - m(\boldsymbol{u})(B_1^2 - B_2^2)m(\boldsymbol{w}^\star(B_1)) \\ &= m(\boldsymbol{u})[R^\star(B_1) - R^\star(B_2) + (B_2^2 - B_1^2)m(\boldsymbol{w}^\star(B_1))] \\ &\quad - R_{\max}(B_2, \boldsymbol{u})[m(\boldsymbol{w}^\star(B_1)) - m(\boldsymbol{w}^\star(B_2))] \end{aligned}$$

$$\geq 0,$$

the first equality uses (B.11), the second equality groups the terms associated with $m(\boldsymbol{u})$, the third equality uses the definition of the maximum risk in (21) and the assumption that $\overline{M}(B, \boldsymbol{u}) = B^2 m(\boldsymbol{u})$, the fourth equality groups the terms by $V(\boldsymbol{u})$ and $m(\boldsymbol{u})$, the fifth equality uses (B.12), and the sixth equality groups the terms by $m(\boldsymbol{u})$ and uses the definition of $R_{\max}(B_2, \boldsymbol{u})$. The last line follows from $m(\boldsymbol{u}) \geq 0$, $R_{\max}(B_2, \boldsymbol{u}) \geq 0$, Lemmas B.4 and B.5. □

Lemma B.7. *Let Assumption 4.9 hold. Assume that $\mathcal{M}$ is a compact set of positive definite matrices with eigenvalues bounded below by $\lambda_{\text{lb}} > 0$ and above by $\lambda_{\text{ub}} < \infty$ where $\lambda_{\text{ub}} > \lambda_{\text{lb}}$. For a given finite $B \geq 0$, the optimal solution to the minimax problem in (21) is uniformly continuous in $\boldsymbol{\Sigma} \in \mathcal{M}$.*

Proof of Lemma B.7. Let $\boldsymbol{w}^\star(B, \boldsymbol{\Sigma})$ be the optimal solution to the minimax problem in (21). This notation is to emphasize $\boldsymbol{\Sigma}$ as an argument. Under Assumption 4.9, the maximum risk function can be written as

$$R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}) \equiv V(\boldsymbol{w}, \boldsymbol{\Sigma}) + B^2 m(\boldsymbol{w}) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + B^2 m(\boldsymbol{w}), \tag{B.13}$$

where I have further denoted the functions $R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma})$ and $V(\boldsymbol{w}, \boldsymbol{\Sigma})$ as a function of $\boldsymbol{\Sigma}$ to emphasize the dependence on $\boldsymbol{\Sigma}$. In the above, $V(\boldsymbol{w}, \boldsymbol{\Sigma})$ is strongly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ for a given $\boldsymbol{\Sigma} \in \mathcal{M}$. The function $m(\boldsymbol{w})$ is convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. By the definitions of strong convexity and convexity (see, for instance, Aragón et al. (2019)), it follows from (B.13) that $R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma})$ is strongly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. Therefore, for any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}_{\text{cvx}}$ and $\boldsymbol{\Sigma} \in \mathcal{M}$,

$$R_{\max}(\boldsymbol{w}_1, B, \boldsymbol{\Sigma}) \geq R_{\max}(\boldsymbol{w}_2, B, \boldsymbol{\Sigma}) + \boldsymbol{g}'(\boldsymbol{w}_1 - \boldsymbol{w}_2) + \frac{C_R}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 \tag{B.14}$$

for any $\boldsymbol{g} \in \partial R_{\max}(\boldsymbol{w}_2, B, \boldsymbol{\Sigma})$ ($\boldsymbol{g}$ is a subdifferential of $R_{\max}(\boldsymbol{w}_2, B, \boldsymbol{\Sigma})$ and $\partial R_{\max}(\boldsymbol{w}_2, B, \boldsymbol{\Sigma})$ is the set of all subdifferentials) and for some $C_R > 0$.

Next, for any $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{M}$ and $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$,

$$\begin{aligned} R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_1) - R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_2) &= V(\boldsymbol{w}, \boldsymbol{\Sigma}_1) + B^2 m(\boldsymbol{w}) - V(\boldsymbol{w}, \boldsymbol{\Sigma}_2) - B^2 m(\boldsymbol{w}) \\ &= \boldsymbol{w}'\boldsymbol{\Sigma}_1\boldsymbol{w} - \boldsymbol{w}'\boldsymbol{\Sigma}_2\boldsymbol{w} \\ &= \boldsymbol{w}'(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\boldsymbol{w}. \end{aligned} \tag{B.15}$$

Hence, by the triangle inequality and (B.15),

$$|R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_1) - R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_2)| \leq \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2 \|\boldsymbol{w}\|_2^2 \leq \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2, \tag{B.16}$$

since $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$.

Recall the maximum risk (B.13) is strictly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ for any $\boldsymbol{\Sigma} \in \mathcal{M}$, so the optimal solution is unique. For any $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{M}$, let $\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2)$ be the optimal solution to the minimax problem. By the optimality of the solutions, I have

$$R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1), B, \boldsymbol{\Sigma}_1) \leq R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_1), \tag{B.17}$$

$$R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_2) \leq R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1), B, \boldsymbol{\Sigma}_2). \tag{B.18}$$

Note that $\mathcal{W}_{\text{cvx}}$ is a polyhedral set because it is a collection of finite inequalities. For any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, $R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_2) < \infty$ for any finite $B \geq 0$. Denote ri as relative interior and dom as effective domain (see, for instance, Bertsekas (2009), for definitions). In addition, $\text{ri}(\text{dom}(R_{\max})) \cap \mathcal{W} \neq \emptyset$. Recall that $\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2)$ minimizes the maximum risk $R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_2)$ over $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. By Proposition 5.4.7 of Bertsekas (2009), there exists $\boldsymbol{g}_2 \in \partial R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_2)$, I have

$$\boldsymbol{g}_2'(\boldsymbol{w} - \boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2)) \geq 0, \tag{B.19}$$

for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$.

Applying (B.14) with $\boldsymbol{w}_1 = \boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1)$, $\boldsymbol{w}_2 = \boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2)$, $\boldsymbol{g} = \boldsymbol{g}_2$, and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_2$, I have

$$\begin{aligned}
\frac{C_R}{2}\|\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1) - \boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2)\|_2^2 &\leq R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1), B, \boldsymbol{\Sigma}_2) - R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_2) \\
&\quad - \boldsymbol{g}_2'[\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1) - \boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2)], \\
&\leq R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1), B, \boldsymbol{\Sigma}_2) - R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_2) \\
&= [R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1), B, \boldsymbol{\Sigma}_2) - R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1), B, \boldsymbol{\Sigma}_1)] \\
&\quad + [R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1), B, \boldsymbol{\Sigma}_1) - R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_1)] \\
&\quad + [R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_1) - R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_2)] \\
&\leq [R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1), B, \boldsymbol{\Sigma}_2) - R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_1), B, \boldsymbol{\Sigma}_1)] \\
&\quad + [R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_1) - R_{\max}(\boldsymbol{w}^\star(B, \boldsymbol{\Sigma}_2), B, \boldsymbol{\Sigma}_2)] \\
&\leq 2\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2
\end{aligned} \tag{B.20}$$

where the second inequality follows from (B.19), the third inequality follows from (B.17),

and the fourth inequality follows from (B.16).

Therefore, the above shows that for any $\varepsilon > 0$, there exists $\delta = \frac{C_R}{4}\varepsilon^2 > 0$ such that for any $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathcal{M}$ and $\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_2 < \delta$, $\|\boldsymbol{w}^\star(B, \mathbf{\Sigma}_1) - \boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)\|_2 < \varepsilon$ holds. □

Lemma B.8. *Consider the same assumptions and notations as in Lemma B.7. Write* $m(\boldsymbol{w}) = \max_{\boldsymbol{b}\in\mathcal{S}(1)}(\boldsymbol{w}'\boldsymbol{b})^2$. *The optimal solution to the adaptive problem in* (26) *is uniformly continuous in* $\mathbf{\Sigma} \in \mathcal{M}$.

Proof of Lemma B.8. Similar to the proof of Lemma B.7, I also write the maximum risk also as a function of $\mathbf{\Sigma}$ as in (B.13). As a result, the adaptive regret for a given $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, $B \in [0, \infty)$ and $\mathbf{\Sigma} \in \mathcal{M}$ is $A(\boldsymbol{w}, B, \mathbf{\Sigma}) \equiv \frac{R_{\max}(\boldsymbol{w}, B, \mathbf{\Sigma})}{R^\star(B, \mathbf{\Sigma})}$ where $R^\star(B, \mathbf{\Sigma}) \equiv \min_{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}} R_{\max}(\boldsymbol{w}, B, \mathbf{\Sigma})$. Let $\mathcal{B} = [\underline{B}, \overline{B}]$. Since the solution is unique, I also denote the optimally adaptive estimator as

$$\boldsymbol{w}_A(\mathbf{\Sigma}) = \underset{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}}{\arg\min}\, A_{\max}(\boldsymbol{w}, \mathbf{\Sigma}), \tag{B.21}$$

and

$$A_{\max}(\boldsymbol{w}, \mathbf{\Sigma}) \equiv \max\{A(\boldsymbol{w}, \underline{B}, \mathbf{\Sigma}), A(\boldsymbol{w}, \overline{B}, \mathbf{\Sigma})\}$$

in this proof.

I have showed that $R_{\max}(\boldsymbol{w}, B, \mathbf{\Sigma})$ is uniformly continuous in $\mathbf{\Sigma} \in \mathcal{M}$ in Lemma B.7. Next, for any $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathcal{M}$, let $\boldsymbol{w}^\star(B, \mathbf{\Sigma}_1)$ and $\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)$ be the optimal solution to the minimax problem in (21) as in Lemma B.7. Then, $R^\star(B, \mathbf{\Sigma}_1) = R_{\max}(\boldsymbol{w}^\star(B, \mathbf{\Sigma}_1), B, \mathbf{\Sigma}_1)$ and $R^\star(B, \mathbf{\Sigma}_2) = R_{\max}(\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2), B, \mathbf{\Sigma}_2)$. Thus,

$$\begin{aligned}
R^\star(B, \mathbf{\Sigma}_1) - R^\star(B, \mathbf{\Sigma}_2) &= R_{\max}(\boldsymbol{w}^\star(B, \mathbf{\Sigma}_1), B, \mathbf{\Sigma}_1) - R_{\max}(\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2), B, \mathbf{\Sigma}_2) \\
&\leq R_{\max}(\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2), B, \mathbf{\Sigma}_1) - R_{\max}(\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2), B, \mathbf{\Sigma}_2), \\
&= \boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)'\mathbf{\Sigma}_1\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2) + B^2 m(\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)) \\
&\quad - \boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)'\mathbf{\Sigma}_2\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2) - B^2 m(\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)) \\
&= \boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)'(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2) \\
&\leq |\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)'(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)| \\
&\leq \|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_2 \|\boldsymbol{w}^\star(B, \mathbf{\Sigma}_2)\|_2^2 \\
&\leq \|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_2,
\end{aligned} \tag{B.22}$$

where the second line follows from $R_{\max}(\boldsymbol{w}^\star(B, \mathbf{\Sigma}_1), B, \mathbf{\Sigma}_1) \leq R_{\max}(\boldsymbol{w}, B, \mathbf{\Sigma}_1)$ for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, the third line follows from the definition of the maximum risk function,

and the last line uses that $\|\boldsymbol{w}^\star(B,\boldsymbol{\Sigma}_2)\|_2^2 \leq 1$ because $\boldsymbol{w}^\star(B,\boldsymbol{\Sigma}_2) \in \mathcal{W}_{\text{cvx}}$. By a similar reasoning, I have

$$\begin{aligned} R^\star(B,\boldsymbol{\Sigma}_2) - R^\star(B,\boldsymbol{\Sigma}_1) &= R_{\max}(\boldsymbol{w}^\star(B,\boldsymbol{\Sigma}_2),B,\boldsymbol{\Sigma}_2) - R_{\max}(\boldsymbol{w}^\star(B,\boldsymbol{\Sigma}_1),B,\boldsymbol{\Sigma}_1) \\ &\leq R_{\max}(\boldsymbol{w}^\star(B,\boldsymbol{\Sigma}_1),B,\boldsymbol{\Sigma}_2) - R_{\max}(\boldsymbol{w}^\star(B,\boldsymbol{\Sigma}_1),B,\boldsymbol{\Sigma}_1), \\ &\leq \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2. \end{aligned} \tag{B.23}$$

Combining (B.22) and (B.23), I obtain

$$|R^\star(B,\boldsymbol{\Sigma}_1) - R^\star(B,\boldsymbol{\Sigma}_2)| \leq \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2. \tag{B.24}$$

Next, I bound $R^\star(B,\boldsymbol{\Sigma})$. For any $\boldsymbol{\Sigma} \in \mathcal{M}$ and finite $B \geq 0$, I have

$$R^\star(B,\boldsymbol{\Sigma}) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + B^2 m(\boldsymbol{w}) \geq \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \geq \lambda_{\text{lb}}\|\boldsymbol{w}\|_2^2 \geq \frac{\lambda_{\text{lb}}}{q}, \tag{B.25}$$

because $m(\boldsymbol{w}) \geq 0$, $\boldsymbol{\Sigma} - \lambda_{\text{lb}}\boldsymbol{I}_q$ is positive semidefinite, $\|\boldsymbol{w}\|_2^2 \geq \frac{1}{q}$ by the Cauchy-Schwarz inequality. For the upper bound, note that $m(\boldsymbol{w}) \leq 1$ because $m(\boldsymbol{w}) = \max_{\boldsymbol{b}\in\mathcal{S}(1)}(\boldsymbol{w}'\boldsymbol{b})^2$, $(\boldsymbol{w}'\boldsymbol{b})^2 \leq \|\boldsymbol{w}\|_{p^\star}^2\|\boldsymbol{b}\|_p^2 \leq 1$, $\|\boldsymbol{b}\|_p^2 \leq 1$ because $\boldsymbol{b} \in \mathcal{S}(1)$ and $\|\boldsymbol{w}\|_{p^\star}^2 \leq 1$. Hence,

$$R^\star(B,\boldsymbol{\Sigma}) = \max_{\boldsymbol{w}\in\mathcal{W}} \left[\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + B^2 m(\boldsymbol{w})\right] \leq \lambda_{\text{ub}}\|\boldsymbol{w}\|_2^2 + B^2 \leq \lambda_{\text{ub}} + B^2, \tag{B.26}$$

is an upper bound on $R^\star(B,\boldsymbol{\Sigma})$ because $\lambda_{\text{ub}}\boldsymbol{I}_q - \boldsymbol{\Sigma}$ is positive semidefinite.

Since $R^\star(B,\boldsymbol{\Sigma}) > 0$ for any $\boldsymbol{\Sigma} \in \mathcal{M}$ from (B.25), it follows that $\frac{1}{R^\star(B,\boldsymbol{\Sigma})} > 0$. Thus,

$$\left|\frac{1}{R^\star(B,\boldsymbol{\Sigma}_1)} - \frac{1}{R^\star(B,\boldsymbol{\Sigma}_2)}\right| \leq \left|\frac{R^\star(B,\boldsymbol{\Sigma}_1) - R^\star(B,\boldsymbol{\Sigma}_2)}{R^\star(B,\boldsymbol{\Sigma}_1)R^\star(B,\boldsymbol{\Sigma}_2)}\right| \leq \frac{q^2\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2}{\lambda_{\text{lb}}^2} \tag{B.27}$$

using (B.24) and (B.25).

The function $R^\star(B,\boldsymbol{\Sigma})$ is positive from (B.25) and also uniformly continuous in $\boldsymbol{\Sigma} \in \mathcal{M}$ from (B.24) and by the definition. It follows that for any $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{M}$,

$$\begin{aligned} |A(\boldsymbol{w},B,\boldsymbol{\Sigma}_1) - A(\boldsymbol{w},B,\boldsymbol{\Sigma}_2)| &= \left|\frac{R_{\max}(\boldsymbol{w},B,\boldsymbol{\Sigma}_1)}{R^\star(B,\boldsymbol{\Sigma}_1)} - \frac{R_{\max}(\boldsymbol{w},B,\boldsymbol{\Sigma}_2)}{R^\star(B,\boldsymbol{\Sigma}_2)}\right| \\ &= \left|\frac{R_{\max}(\boldsymbol{w},B,\boldsymbol{\Sigma}_1) - R_{\max}(\boldsymbol{w},B,\boldsymbol{\Sigma}_2)}{R^\star(B,\boldsymbol{\Sigma}_1)}\right. \\ &\qquad \left. + R_{\max}(\boldsymbol{w},B,\boldsymbol{\Sigma}_2)\left[\frac{1}{R^\star(B,\boldsymbol{\Sigma}_1)} - \frac{1}{R^\star(B,\boldsymbol{\Sigma}_2)}\right]\right| \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{|R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_1) - R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_2)|}{|R^\star(B, \boldsymbol{\Sigma}_1)|} \\
&\quad + |R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_2)| \left| \frac{1}{R^\star(B, \boldsymbol{\Sigma}_1)} - \frac{1}{R^\star(B, \boldsymbol{\Sigma}_2)} \right| \\
&\leq \frac{q\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2}{\lambda_{\text{lb}}} + \frac{q^2(\lambda_{\text{ub}} + B^2)\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2}{\lambda_{\text{lb}}^2} \\
&\equiv \mathsf{K}_1(B)\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2, \qquad \text{(B.28)}
\end{aligned}
$$

where the second last line uses (B.16), (B.25), (B.26), (B.27), and $R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma}_2) > 0$, and the last line defined $\mathsf{K}_1(B) \equiv \frac{q}{\lambda_{\text{lb}}} + \frac{q^2(\lambda_{\text{ub}}+B^2)}{\lambda_{\text{lb}}^2} \geq 0$.

Next, since $|\max\{a_1, a_2\} - \max\{b_1, b_2\}| \leq \max\{|a_1 - b_1|, |a_2 - b_2|\}$, and $\mathsf{K}_1(B) \geq 0$ in (B.28), using the above derivation on the bound of $A(\boldsymbol{w}, B, \boldsymbol{\Sigma}_2)$, this shows that there exists $\mathsf{K}_2 > 0$ such that

$$
\begin{aligned}
&|A_{\max}(\boldsymbol{w}, \boldsymbol{\Sigma}_1) - A_{\max}(\boldsymbol{w}, \boldsymbol{\Sigma}_2)| \\
&\quad = \left|\max\{A(\boldsymbol{w}, \underline{B}, \boldsymbol{\Sigma}_1), A(\boldsymbol{w}, \overline{B}, \boldsymbol{\Sigma}_1)\} - \max\{A(\boldsymbol{w}, \underline{B}, \boldsymbol{\Sigma}_2), A(\boldsymbol{w}, \overline{B}, \boldsymbol{\Sigma}_2)\}\right| \\
&\quad \leq \max\{|A(\boldsymbol{w}, \underline{B}, \boldsymbol{\Sigma}_1) - A(\boldsymbol{w}, \underline{B}, \boldsymbol{\Sigma}_2)|, |A(\boldsymbol{w}, \overline{B}, \boldsymbol{\Sigma}_1) - A(\boldsymbol{w}, \overline{B}, \boldsymbol{\Sigma}_2)|\} \\
&\quad \leq \max\{\mathsf{K}_1(\underline{B})\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2, \mathsf{K}_1(\overline{B})\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2\} \\
&\quad \leq \max\{\mathsf{K}_1(\underline{B}), \mathsf{K}_1(\overline{B})\}\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2 \\
&\quad \equiv \mathsf{K}_2\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_2, \qquad \text{(B.29)}
\end{aligned}
$$

where the last line defined $\mathsf{K}_2 \equiv \max\{\mathsf{K}_1(\underline{B}), \mathsf{K}_1(\overline{B})\}$ from (B.28).

Now, for any $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{M}$, consider $\boldsymbol{w}_A(\boldsymbol{\Sigma}_1)$ and $\boldsymbol{w}_A(\boldsymbol{\Sigma}_2)$ that are optimal solution to the adaptive problem (B.21). Then,

$$
A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_1), \boldsymbol{\Sigma}_1) \leq A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2), \boldsymbol{\Sigma}_1), \qquad \text{(B.30)}
$$

$$
A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2), \boldsymbol{\Sigma}_2) \leq A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_1), \boldsymbol{\Sigma}_2). \qquad \text{(B.31)}
$$

By the definition of the adaptive regret,

$$
A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2), \boldsymbol{\Sigma}_1) - A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_1), \boldsymbol{\Sigma}_1) \geq \mathsf{K}_3\|\boldsymbol{w}_A(\boldsymbol{\Sigma}_2) - \boldsymbol{w}_A(\boldsymbol{\Sigma}_1)\|^2, \qquad \text{(B.32)}
$$

where $\mathsf{K}_3 > 0$ the last line follows because $A(\boldsymbol{w}, B, \boldsymbol{\Sigma})$ is strongly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ (in which the proof follows from Lemma B.7, particularly equations (B.14), (B.19), and (B.20), because $A(\boldsymbol{w}, B, \boldsymbol{\Sigma})$ is a (positive) scalar multiple of $R_{\max}(\boldsymbol{w}, B, \boldsymbol{\Sigma})$ when B and $\boldsymbol{\Sigma}$ are held fixed) and that $A_{\max}$ is the maximum of two strongly convex functions.

It follows that

$$\begin{aligned}
\mathsf{K}_3\|\boldsymbol{w}_A(\boldsymbol{\Sigma}_2)-\boldsymbol{w}_A(\boldsymbol{\Sigma}_1)\|^2 &\leq A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2),\boldsymbol{\Sigma}_1)-A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_1),\boldsymbol{\Sigma}_1)\\
&= [A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2),\boldsymbol{\Sigma}_1)-A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2),\boldsymbol{\Sigma}_2)]\\
&\quad + [A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2),\boldsymbol{\Sigma}_2)-A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_1),\boldsymbol{\Sigma}_1)]\\
&\leq |A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2),\boldsymbol{\Sigma}_1)-A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2),\boldsymbol{\Sigma}_2)|\\
&\quad + |A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_2),\boldsymbol{\Sigma}_2)-A_{\max}(\boldsymbol{w}_A(\boldsymbol{\Sigma}_1),\boldsymbol{\Sigma}_1)|\\
&\leq 2\mathsf{K}_2\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\|_2,
\end{aligned}$$

where the first line follows from (B.32) and the last line follows from (B.29).

Therefore, the above shows that for any $\varepsilon>0$, there exists $\delta=\frac{\mathsf{K}_3}{2\mathsf{K}_2}\varepsilon^2>0$ such that for any $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\in\mathcal{M}$ and $\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\|_2<\delta$, $\|\boldsymbol{w}_A(B,\boldsymbol{\Sigma}_1)-\boldsymbol{w}_A(B,\boldsymbol{\Sigma}_2)\|_2<\varepsilon$ holds. □

B.4 Additional details for the parameter space

B.4.1 Shape restrictions

In this subsection, I discuss details related to imposing shape restrictions in the parameter space. I have explained in Example 4.3 that shape restrictions are desirable in empirical practice. I show that evaluating the maximum risk subject to the parameter space that involves absolute values on $\boldsymbol{b}$ is the same as evaluating the parameter space without absolute values.

First, consider shape restrictions in the form of $\boldsymbol{Q}\boldsymbol{b}\leq\boldsymbol{0}_l$, where $\boldsymbol{Q}\in\mathbb{R}^{l\times q}$ and $\boldsymbol{0}_l$ is a vector of l zeros. The parameter space can be described as

$$\mathcal{S}(B)=\{\boldsymbol{b}\in\mathbb{R}^q:\|\boldsymbol{b}\|_p\leq B,\boldsymbol{Q}\boldsymbol{b}\leq\boldsymbol{0}_l\}. \tag{B.33}$$

For the purpose of computing the maximum risk (20), the set (B.33) on linear inequalities on $\boldsymbol{b}$ also describes inequality constraints on the magnitude of $\boldsymbol{b}$, i.e., inequality constraints on $|\boldsymbol{b}|$, as in the entrepreneurial spirit example in Example 4.3. More specifically, consider

$$\mathcal{S}_{\text{abs}}(B)=\{\boldsymbol{b}\in\mathbb{R}^q:\|\boldsymbol{b}\|_p\leq B,\boldsymbol{Q}|\boldsymbol{b}|\leq\boldsymbol{0}_l\}, \tag{B.34}$$

where $|\boldsymbol{b}|$ is the vector that takes the absolute value of each component in the vector $\boldsymbol{b}$. The following proposition shows the maximum risk over (B.34) is the same as the maximum risk over linear inequalities in the form of (B.33). This simplifies computation.

Proposition B.9. *Consider the notations as defined in* (20). *Let* $p \geq 1$, $B \geq 0$ *and* $\boldsymbol{Q} \in \mathbb{R}^{l \times q}$. *In addition, let* $\mathcal{S}_{\text{aug}}(B) \equiv \{\boldsymbol{b} \in \mathbb{R}^q : \|\boldsymbol{b}\|_p \leq B, \boldsymbol{Q}\boldsymbol{b} \leq \boldsymbol{0}_l, \boldsymbol{b} \geq \boldsymbol{0}_q\}$. *Suppose* $\mathcal{S}_{\text{aug}}(B) \neq \emptyset$. *Then,*

$$\max_{\boldsymbol{b} \in \mathcal{S}_{\text{abs}}(B)} (\boldsymbol{w}'\boldsymbol{b})^2 = \max_{\boldsymbol{b} \in \mathcal{S}_{\text{aug}}(B)} (\boldsymbol{w}'\boldsymbol{b})^2,$$

where $\mathcal{S}_{\text{abs}}(B)$ *is defined in* (B.34).

Proof of Proposition B.9. Let $m_{\text{abs}} \equiv \max_{\boldsymbol{b} \in \mathcal{S}_{\text{abs}}(B)} (\boldsymbol{w}'\boldsymbol{b})^2$ and $m_{\text{aug}} \equiv \max_{\boldsymbol{b} \in \mathcal{S}_{\text{aug}}(B)} (\boldsymbol{w}'\boldsymbol{b})^2$.

To begin with, note that for any $\boldsymbol{b} \in \mathcal{S}_{\text{aug}}(B)$, $|\boldsymbol{b}| = \boldsymbol{b}$ since $\boldsymbol{b} \geq \boldsymbol{0}_q$. Therefore, $\boldsymbol{Q}|\boldsymbol{b}| \leq \boldsymbol{0}_l$. It follows that $\boldsymbol{b} \in \mathcal{S}_{\text{abs}}(B)$ as well. Hence, $m_{\text{aug}} \leq m_{\text{abs}}$.

Next, for any $\boldsymbol{b} \in \mathcal{S}_{\text{abs}}(B)$, $\boldsymbol{Q}|\boldsymbol{b}| \leq \boldsymbol{0}_l$ by construction and $|\boldsymbol{b}| \geq \boldsymbol{0}_q$. Hence, $|\boldsymbol{b}| \in \mathcal{S}_{\text{aug}}(B)$. Note that the objective is bounded above as follows

$$(\boldsymbol{w}'\boldsymbol{b})^2 = |\boldsymbol{w}'\boldsymbol{b}|^2 \leq (|\boldsymbol{w}|'|\boldsymbol{b}|)^2 = (\boldsymbol{w}'|\boldsymbol{b}|)^2,$$

where the inequality follows from the triangle inequality and the second equality follows because $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, so $\boldsymbol{w} \geq \boldsymbol{0}_q$. As a result, this means for each $\boldsymbol{b} \in \mathcal{S}_{\text{abs}}(B)$, there exists $|\boldsymbol{b}| \in \mathcal{S}_{\text{aug}}(B)$ such that the objective is weakly larger. Hence, $m_{\text{abs}} \leq m_{\text{aug}}$.

Combining the two parts gives $m_{\text{abs}} = m_{\text{aug}}$. □

The above proposition shows that although the two sets $\mathcal{S}_{\text{abs}}(B)$ and $\mathcal{S}_{\text{aug}}(B)$ are different, they lead to the same maximum value. Hence, the representation in (B.33) can capture this case.

In general, the maximum misspecification further satisfies Assumption 4.9 because

$$\overline{M}(B, \boldsymbol{w}) = \max_{\boldsymbol{b} \in \mathcal{S}(B)} (\boldsymbol{w}'\boldsymbol{b})^2 = B^2 \max_{\widetilde{\boldsymbol{b}} \in \mathcal{S}(1)} (\boldsymbol{w}'\widetilde{\boldsymbol{b}})^2 \equiv B^2 m(\boldsymbol{w}), \tag{B.35}$$

where $m(\boldsymbol{w}) \equiv \max_{\widetilde{\boldsymbol{b}} \in \mathcal{S}(1)} (\boldsymbol{w}'\widetilde{\boldsymbol{b}})^2$ is convex in $\boldsymbol{w}$ and $\mathcal{S}(B)$ is as defined in (B.33).

B.4.2 A model of communication and subjective weights

Another concern that researchers may have is related to the communication of the weighted average of treatment effects $\widehat{\tau}$ to the reader. Readers may think that θ is a weighted average of the treatment effects $\boldsymbol{\beta}$, but have a different view on what the weights should be. Then, the researcher's decision has to take this communication issue into account.

The communication problem between the researcher and the reader can be studied using the decision framework above. Suppose that $\theta = \boldsymbol{\gamma}'\boldsymbol{\beta}$, where $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ is a vector of subjective weights of the readers that can be known or unknown to the researcher. Here, I assume that the reader also wants an interpretable estimator, so $\boldsymbol{\gamma}$ also belongs to the set $\mathcal{W}_{\text{cvx}}$.

The modeling choice on $\boldsymbol{\gamma}$ depends on the context or the goal of the researcher. I consider the following three cases on the restrictions on $\boldsymbol{\gamma}$:

1. A known $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$.
2. An unknown $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$.
3. There is a known reference weight $\boldsymbol{\eta} \in \mathcal{W}_{\text{cvx}}$ such that $\boldsymbol{\gamma} = \boldsymbol{\eta} + \boldsymbol{\delta}$, $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$, $\boldsymbol{\delta} \in \mathcal{W}_{\text{cvx}}$, $\|\boldsymbol{\delta}\|_p \leq D$, and $D \geq 0$.

In the above, Cases 1 and 2 can be viewed as special cases of Case 3. This is because setting $D = 0$ in Case 3 corresponds to setting $\boldsymbol{\eta} = \boldsymbol{\gamma}$ in Case 1. Setting a "sufficiently large" value in D in Case 3 corresponds to allowing for any $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ as in Case 2. This is formalized and discussed further in Proposition D.4 of the Appendix. In the following, I consider the first case where $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ is known. The other cases are considered in Appendix D.

When θ is a weighted average of $\boldsymbol{\beta}$, (16) becomes

$$\boldsymbol{b} \equiv \boldsymbol{\beta} - (\boldsymbol{\gamma}'\boldsymbol{\beta})\mathbf{1}_q. \tag{B.36}$$

Suppose the researcher is interested in bounding $\boldsymbol{b}$ using the ℓ_p-norm as in Example 4.2. The parameter space $\mathcal{S}(B)$ from (19) has to take (B.36) into account. Thus, I write the parameter space for $\boldsymbol{b}$ as $\mathcal{S}_0(B, \boldsymbol{\gamma})$ with $\boldsymbol{\gamma}$ also as an argument below, in light of (B.36):

$$\mathcal{S}_0(B, \boldsymbol{\gamma}) \equiv \{\boldsymbol{b} \in \mathbb{R}^q : \|\boldsymbol{b}\|_p \leq B, \boldsymbol{b} = \boldsymbol{\beta} - (\boldsymbol{\gamma}'\boldsymbol{\beta})\mathbf{1}_q \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^q\}. \tag{B.37}$$

Lemma D.1 shows that the parameter space (B.37) has an equivalent representation as

$$\mathcal{S}_1(B, \boldsymbol{\gamma}) \equiv \{\boldsymbol{b} \in \mathbb{R}^q : \|\boldsymbol{b}\|_p \leq B, \boldsymbol{\gamma}'\boldsymbol{b} = 0\}.$$

Thus, the communication problem can be modeled via the framework introduced in Section B.4.1 on shape restrictions. This is because $\boldsymbol{\gamma}'\boldsymbol{b} = 0$ is equivalent to the restrictions $\boldsymbol{\gamma}'\boldsymbol{b} \leq 0$ and $-\boldsymbol{\gamma}'\boldsymbol{b} \leq 0$. As a result, the maximum risk for a given $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$

can be written as follows using (B.35):

$$\overline{M}(B, \boldsymbol{w}, \boldsymbol{\gamma}) = B^2 \max_{\widetilde{\boldsymbol{b}} \in \mathcal{S}_1(1,\boldsymbol{\gamma})} (\boldsymbol{w}'\widetilde{\boldsymbol{b}})^2. \tag{B.38}$$

As before, (B.38) satisfies Assumption 4.9 on multiplicative separability as well. Shape constraints as in Section B.4.1 can also be added to the parameter space $\mathcal{S}_1(B, \boldsymbol{\gamma})$.

I provide further analysis and worked examples in Appendix D. The communication and subjective weights aspect in my statistical decision-theoretic approach is also related to models of scientific communication (such as Andrews and Shapiro (2021), Frankel and Kasy (2022), and Kasy and Spiess (2024)), transparency (Andrews et al. (2020) and the related discussion by Bonhomme (2020), Taber (2020), and Tamer (2020)).

B.5 Additional details for Example 4.8

Below, I return to the running example with the finite-sample model with two outcomes for illustration. Suppose the Euclidean norm in $\mathcal{S}(B)$ is used to bound $\boldsymbol{b} = (b_1, b_2)'$.

B.5.1 Maximum risk

Since $w_1 + w_2 = 1$, the variance and maximum misspecification can be written as a function of $w_1 \in [0, 1]$ as follows:

$$\begin{aligned} V(w_1) &= (1 + \sigma_2^2 - 2\rho\sigma_2)w_1^2 + 2(\rho\sigma_2 - \sigma_2^2)w_1 + \sigma_2^2, \\ \overline{M}(B, w_1) &= B^2 \|(w_1, 1 - w_1)\|_2^2 = B^2(2w_1^2 - 2w_1 + 1) \equiv B^2 m(w_1), \end{aligned} \tag{B.39}$$

where I have further defined $m(w_1) \equiv 2w_1^2 - 2w_1 + 1$ for further convenience later.

B.5.2 Minimax problem

Using the optimal weights in (23), the minimax risk is

$$R^\star(B) = R_{\max}(B, w_1^\star(B)) = \begin{cases} \sigma_2^2 + B^2 & \text{if } \sigma_2^2 - \rho\sigma_2 \leq -B^2, \\ 1 + B^2 & \text{if } 1 - \rho\sigma_2 \leq -B^2, \\ \frac{(1-\rho^2)\sigma_2^2 + (1+\sigma_2^2)B^2 + B^4}{1 + \sigma_2^2 - 2\rho\sigma_2 + 2B^2} & \text{otherwise.} \end{cases} \tag{B.40}$$

B.5.3 Adaptive problem

Assume that $\mathcal{B} = [0, \infty]$ and $\rho\sigma_2 < 1$. The optimization problem for computing the adaptive weight is

$$\begin{aligned} \min_{t, w_1 \in \mathbb{R}} \quad & t, \\ \text{s.t.} \quad & \frac{V(w_1)}{R^\star(0)} \le t, \\ & 2(2w_1^2 - 2w_1 + 1) \le t, \\ & w_1 \le 1, \\ & w_1 \ge 0, \end{aligned} \tag{B.41}$$

where $V(w_1)$ is defined in (B.39). The Lagrangian is

$$\mathcal{L} = t + \lambda_1[R^\star(0)^{-1}V(w_1) - t] + \lambda_2(4w_1^2 - 4w_1 + 2 - t) + \lambda_3(w_1 - 1) + \lambda_4(-w_1), \tag{B.42}$$

where $\{\lambda_l\}_{l=1}^4$ are the Lagrangian multipliers.

Recall from (B.39) that $V(w_1) = (1 + \sigma_2^2 - 2\rho\sigma_2)w_1^2 + 2(\rho\sigma_2 - \sigma_2^2)w_1 + \sigma_2^2$. Using (B.42), the KKT conditions are as follows.

- **Stationarity:**

$$0 = 1 - \lambda_1 - \lambda_2, \tag{B.43}$$

$$0 = \lambda_1 R^\star(0)^{-1}[2(1 + \sigma_2^2 - 2\rho\sigma_2)w_1 + 2(\rho\sigma_2 - \sigma_2^2)] + \lambda_2(8w_1 - 4) + \lambda_3 - \lambda_4. \tag{B.44}$$

- **Primal feasibility:**

$$\frac{(1 + \sigma_2^2 - 2\rho\sigma_2)w_1^2 + 2(\rho\sigma_2 - \sigma_2^2)w_1 + \sigma_2^2}{R^\star(0)} - t \le 0, \tag{B.45}$$

$$4w_1^2 - 4w_1 + 2 - t \le 0, \tag{B.46}$$

$$w_1 - 1 \le 0, \tag{B.47}$$

$$-w_1 \le 0. \tag{B.48}$$

- **Dual feasibility:**

$$\lambda_1,\ \lambda_2,\ \lambda_3,\ \lambda_4 \ge 0. \tag{B.49}$$

- **Complementary slackness:**

$$\lambda_1 \left[\frac{(1+\sigma_2^2-2\rho\sigma_2)w_1^2+2(\rho\sigma_2-\sigma_2^2)w_1+\sigma_2^2}{R^\star(0)} - t \right] = 0, \tag{B.50}$$

$$\lambda_2(4w_1^2-4w_1+2-t) = 0, \tag{B.51}$$

$$\lambda_3(w_1-1) = 0, \tag{B.52}$$

$$\lambda_4 w_1 = 0. \tag{B.53}$$

Note that $R^\star(0) > 0$ for each of the three cases shown in (B.40) for $B = 0$.

Proof of (27) for no corner solution. I start to analyze the case when $w_1^\star = 0$. In this case, the implications on the KKT conditions are as follows.

- **Implications on complementary slackness, i.e., (B.50) to (B.53):**

  At $w_1^\star = 0$, the conditions become

$$\lambda_1[\sigma_2^2 - R^\star(0)t] = 0, \quad \lambda_2(2-t) = 0 \quad \text{and} \quad \lambda_3 = 0. \tag{B.54}$$

  $\lambda_4 \geq 0$ is otherwise unrestricted.

- **Implications on primal feasibility, i.e., (B.45) to (B.49):**

  At $w_1^\star = 0$, the conditions (B.45) and (B.46) become

$$\sigma_2^2 \leq R^\star(0)t \quad \text{and} \quad 2 \leq t. \tag{B.55}$$

- **Implications on stationarity, i.e., (B.43) and (B.44):**

  At $w_1^\star = 0$ and (B.54), (B.43) remains unchanged, but (B.44) becomes

$$\begin{aligned} 0 &= 2\lambda_1 R^\star(0)^{-1}(\rho\sigma_2 - \sigma_2^2) - 4\lambda_2 - \lambda_4 \\ &= \lambda_1[2R^\star(0)^{-1}(\rho\sigma_2 - \sigma_2^2) + 4] - 4 - \lambda_4, \end{aligned} \tag{B.56}$$

  where the second equality uses (B.43) that $\lambda_2 = 1 - \lambda_1$.

Using the above implications on the KKT conditions, I consider different possibilities on the Lagrangian multipliers. Note that $\lambda_1 \in [0,1]$ by (B.43) and (B.49).

1. Suppose that $\lambda_1 = 0$. Then, (B.56) requires that $\lambda_4 = -4$, which violates (B.49).

2. Suppose that $\lambda_1 = 1$. The complementary slackness condition (B.54) requires

$$\sigma_2^2 = R^\star(0)t. \tag{B.57}$$

From (B.40), the expression for $R^\star(0)$ depends on whether $\sigma_2^2 - \rho\sigma_2 \leq 0$ holds or not.

(a) Assume $\sigma_2^2 - \rho\sigma_2 \leq 0$, so $R^\star(0) = \sigma_2^2$. Since (B.55) requires $t \geq 2$, it violates (B.57).

(b) Assume $\sigma_2^2 - \rho\sigma_2 > 0$, so $R^\star(0) = \frac{(1-\rho^2)\sigma_2^2}{1+\sigma_2^2-2\rho\sigma_2}$. But $\rho \in (-1, 1)$, (B.49) and (B.56) lead to $\lambda_4 = 2R^\star(0)^{-1}(\rho\sigma_2 - \sigma_2^2) < 0$, which is violated by the condition.

3. Suppose that $\lambda_1 \in (0, 1)$. Then, (B.43) gives $\lambda_2 \in (0, 1)$. The complementary slackness condition (B.54) requires

$$\sigma_2^2 = R^\star(0)t \quad \text{and} \quad t = 2. \tag{B.58}$$

Similar to the previous case, I analyze based on the expression for $R^\star(0)$ depending on whether $\sigma_2^2 - \rho\sigma_2 \leq 0$ holds or not.

(a) Assume $\sigma_2^2 - \rho\sigma_2 \leq 0$, so that $R^\star(0) = \sigma_2^2$. This means $\sigma_2^2 = \sigma_2^2 t$, so $t = 1$ Thus, (B.58) is violated.

(b) Asume $\sigma_2^2 - \rho\sigma_2 > 0$. Then (B.49) and (B.56) require that

$$\begin{aligned}\lambda_4 &= \lambda_1[2R^\star(0)^{-1}(\rho\sigma_2 - \sigma_2^2) + 4] - 4 \\ &= 2\lambda_1 R^\star(0)^{-1}(\underbrace{\rho\sigma_2 - \sigma_2^2}_{<0}) + 4(\underbrace{\lambda_1 - 1}_{<0}) \\ &< 0,\end{aligned}$$

which violates dual feasibility (B.49).

It follows that there is no $\lambda_1 \geq 0$ such that the KKT conditions can be satisfied so $w_1^\star = 0$ is not optimal.

Next, I analyze the case when $w_1^\star = 1$. In this case, the implications on the KKT conditions are as follows.

- **Implications on complementary slackness, i.e., (B.50) to (B.53):**

At $w_1^\star = 1$, the conditions become

$$\lambda_1[1 - R^\star(0)t] = 0, \quad \lambda_2(2 - t) = 0 \quad \text{and} \quad \lambda_4 = 0. \tag{B.59}$$

$\lambda_3 \geq 0$ is otherwise unrestricted.

- **Implications on primal feasibility, i.e., (B.45) to (B.49):**

  At $w_1^\star = 1$, the conditions (B.45) and (B.46) become

$$1 \leq R^\star(0)t \quad \text{and} \quad 2 \leq t. \tag{B.60}$$

- **Implications on stationarity, i.e., (B.43) and (B.44):**

  At $w_1^\star = 1$ and (B.59), (B.43) remains unchanged, but (B.44) becomes

$$\begin{aligned} 0 &= 2\lambda_1 R^\star(0)^{-1}(1 - \rho\sigma_2) + 4\lambda_2 + \lambda_3 \\ &= \lambda_1[2R^\star(0)^{-1}(1 - \rho\sigma_2) - 4] + 4 + \lambda_3, \end{aligned} \tag{B.61}$$

  where the second equality uses (B.43) that $\lambda_2 = 1 - \lambda_1$.

Using the above implications on the KKT conditions, I consider different possibilities on the Lagrangian multipliers.

1. Suppose that $\lambda_1 = 0$. Then, (B.61) requires that $\lambda_3 = -4$, which violates (B.49).
2. Suppose that $\lambda_1 = 1$. The complementary slackness condition (B.59) requires

$$1 = R^\star(0)t, \tag{B.62}$$

   and (B.61) becomes

$$\lambda_3 = -2R^\star(0)^{-1}(1 - \rho\sigma_2). \tag{B.63}$$

   Note that $R^\star(0) > 0$. But by the assumption that $\rho\sigma_2 < 1$, it follows from (B.63) that $\lambda_3 < 0$. This violates dual feasibility (B.49).
3. Suppose that $\lambda_1 \in (0, 1)$. Then, (B.43) gives $\lambda_2 \in (0, 1)$. The condition (B.61) can be further written as follows

$$2\lambda_1 R^\star(0)^{-1}(\underbrace{1 - \rho\sigma_2}_{>0}) + 4(\underbrace{1 - \lambda_1}_{>0}) + \lambda_3 = 0, \tag{B.64}$$

The above implies that $\lambda_3 < 0$ using the assumption that $\rho\sigma_2 < 1$. This violates dual feasibility (B.49).

It follows that there is no $\lambda_1 \geq 0$ such that the KKT conditions can be satisfied, so $w_1^\star = 1$ is not optimal. □

Proof of the interior solution in (27). Suppose that $w_1^\star \in (0,1)$. The KKT conditions are given in (B.43) to (B.53). Note that the implications on the KKT conditions when $w_1^\star \in (0,1)$ are as follows.

- **Implications on complementary slackness, i.e., (B.50) to (B.53):**

  The conditions related to the Lagrangian multiplier for $w_1 \in [0,1]$ become

  $$\lambda_3 = 0 \quad \text{and} \quad \lambda_4 = 0. \tag{B.65}$$

  $\lambda_1, \lambda_2 \geq 0$ are not further unrestricted.

- **Implications on primal feasibility, i.e., (B.45) to (B.49):**

  They impose restrictions conditions on t.

  $$\frac{(1+\sigma_2^2-2\rho\sigma_2)w_1^2 + 2(\rho\sigma_2-\sigma_2^2)w_1 + \sigma_2^2}{R^\star(0)} \leq t \quad \text{and} \quad 4w_1^2 - 4w_1 + 2 \leq t. \tag{B.66}$$

- **Implications on stationarity, i.e., (B.43) and (B.44):**

  (B.43) remains unchanged, but (B.44) becomes

  $$0 = \lambda_1 R^\star(0)^{-1}[2(1+\sigma_2^2-2\rho\sigma_2)w_1 + 2(\rho\sigma_2-\sigma_2^2)] + (1-\lambda_1)(8w_1-4) \tag{B.67}$$

  where the equality uses (B.43) that $\lambda_2 = 1-\lambda_1$. This can be rearranged as follows:

  $$w_1 = \frac{\lambda_1[R^\star(0)^{-1}(\sigma_2^2-\rho\sigma_2)-2]+2}{\lambda_1[R^\star(0)^{-1}(\sigma_2^2-2\rho\sigma_2+1)-4]+4} \equiv w_1^\star(\lambda_1). \tag{B.68}$$

  The result in the example follows by changing λ_1 to μ_1.

□

B.6 Additional details for Remark 4.7

This subsection provides additional results for Remark 4.7. The following lemma characterizes the minimax solution for the limiting case as $B \longrightarrow \infty$, and show that the optimal

solution will tend to focus on minimizing $m(\boldsymbol{w})$.

Lemma B.10. *Let Assumption 4.9 hold and $m(\boldsymbol{w})$ be strictly convex and continuous in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. Then, as $B \longrightarrow \infty$,*

$$\boldsymbol{w}^\star(B) \longrightarrow \arg\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} m(\boldsymbol{w}),$$

where $\boldsymbol{w}^\star(B)$ is the solution to the minimax problem as defined in (22).

Proof of Lemma B.10. The set $\mathcal{W}_{\text{cvx}}$ is compact. Note that $V(\boldsymbol{w})$ is continuous in $\boldsymbol{w}$. In addition, $\overline{M}(B, \boldsymbol{w}) = \max_{\boldsymbol{b} \in \mathcal{S}(B)} (\boldsymbol{w}'\boldsymbol{b})^2$ is continuous due to the Berge maximum theorem (see, for instance, Aliprantis and Border (2006, Theorem 17.31)). It follows that $V(\boldsymbol{w})$ and $m(\boldsymbol{w})$ are bounded on $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. In addition, let $\boldsymbol{w}_m \equiv \arg\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} m(\boldsymbol{w})$. As such, define $\mathsf{C}_V \equiv \max_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} |V(\boldsymbol{w}) - V(\boldsymbol{w}_m)| \geq 0$. Hence, for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, it must be that $|V(\boldsymbol{w}) - V(\boldsymbol{w}_m)| \leq \mathsf{C}_V$. In particular, the following must hold

$$V(\boldsymbol{w}) - V(\boldsymbol{w}_m) \geq -\mathsf{C}_V. \tag{B.69}$$

For any $\delta > 0$, let $\mathcal{W}_m(\delta) \equiv \{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}} : \|\boldsymbol{w} - \boldsymbol{w}_m\|_2 \geq \delta\}$. This is a set that is not in a small neighborhood of $\boldsymbol{w}_m$. By the continuity and strict convexity of $m(\boldsymbol{w})$, there must exist $\varepsilon_\delta > 0$ such that

$$m(\boldsymbol{w}) \geq m(\boldsymbol{w}_m) + \varepsilon_\delta, \tag{B.70}$$

for any $\boldsymbol{w} \in \mathcal{W}_m(\delta)$.

Recall the maximum risk function $R_{\max}(B, \boldsymbol{w})$ defined in (20). For any $\boldsymbol{w} \in \mathcal{W}_m(\delta)$,

$$\begin{aligned} R_{\max}(B, \boldsymbol{w}) - R_{\max}(B, \boldsymbol{w}_m) &= V(\boldsymbol{w}) + B^2 m(\boldsymbol{w}) - V(\boldsymbol{w}_m) - B^2 m(\boldsymbol{w}_m) \\ &= [V(\boldsymbol{w}) - V(\boldsymbol{w}_m)] + B^2 [m(\boldsymbol{w}) - m(\boldsymbol{w}_m)] \\ &\geq -\mathsf{C}_V + B^2 \varepsilon_\delta, \end{aligned} \tag{B.71}$$

where the inequality follows from (B.69) and (B.70).

Now, pick $B_0 = \sqrt{\frac{\mathsf{C}_V}{\varepsilon_\delta}} + 1 > 0$ such that $-\mathsf{C}_V + B_0^2 \varepsilon_\delta > 0$. Using (B.71), this means for any $B \geq B_0$, I have

$$R_{\max}(B, \boldsymbol{w}) - R_{\max}(B, \boldsymbol{w}_m) \geq -\mathsf{C}_V + B^2 \varepsilon_\delta \geq -\mathsf{C}_V + B_0^2 \varepsilon_\delta > 0, \tag{B.72}$$

for any $\boldsymbol{w} \in \mathcal{W}_m(\delta)$.

Recall that $\boldsymbol{w}^\star(B)$ is defined as the minimax solution as in (22). This implies that

$R_{\max}(B, \boldsymbol{w}^\star(B)) \leq R_{\max}(B, \boldsymbol{w})$ for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. In particular,

$$R_{\max}(B, \boldsymbol{w}^\star(B)) \leq R_{\max}(B, \boldsymbol{w}_m). \tag{B.73}$$

Then for any $B \geq B_0$, (B.72) implies that $\boldsymbol{w}^\star(B) \in \mathcal{W}_{\text{cvx}} \backslash \mathcal{W}_m(\delta)$. This is because if $\boldsymbol{w}^\star(B) \in \mathcal{W}_m(\delta)$, then $R_{\max}(B, \boldsymbol{w}^\star(B)) > R_{\max}(B, \boldsymbol{w}_m)$, contradicting (B.73).

The above have showed that for any $\delta > 0$, there exists B_0 such that for any $B \geq B_0$, $\boldsymbol{w}^\star(B) \in \mathcal{W}_{\text{cvx}} \backslash \mathcal{W}_m(\delta)$, or equivalently, $\|\boldsymbol{w}^\star(B) - \boldsymbol{w}_m\|_2 < \delta$. Therefore, the proof is complete. □

The intuition is that as B becomes so large, the effect of not putting emphasis to the information from the variance matrix is "negligible." Hence, the researcher should focus on minimizing $m(\boldsymbol{w})$ instead of hedging against variance reduction.

The following corollary is a consequence of the lemma above. It shows that when the ℓ_p-norm is used to restrict $\boldsymbol{b}$ for $p \in (1, \infty)$, then the optimal weight as $B \longrightarrow \infty$ is to put equal weights on the outcomes.

Corollary B.11. *Consider the same assumptions and notations as in Lemma B.10. Suppose the parameter space in (19) is used with the ℓ_p-norm with $p \in (1, \infty)$ as in Example 4.2. Then,*

$$\lim_{B \to \infty} \boldsymbol{w}^\star(B) = \frac{1}{q} \mathbf{1}_q.$$

Proof of Corollary B.11. Let $\ell_{p^\star}$-norm be the dual norm of the ℓ_p-norm. Then, $m(\boldsymbol{w}) = \|\boldsymbol{w}\|_{p^\star}^2$ where $p^\star = \frac{p}{p-1}$ from Example 4.2. By the definition of $\ell_{p^\star}$-norm, $m(\boldsymbol{w}) = [\sum_{j=1}^q h(w_j)]^{\frac{2}{p^\star}}$ where $h(w_j) \equiv |w_j|^{p^\star}$. Since $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, $h(w_j) = w_j^{p^\star}$ for $j = 1, \ldots, q$.

Let $a_j = 1$ for $j = 1, \ldots, q$. Then,

$$\frac{1}{q} \sum_{j=1}^q h(w_j) = \frac{\sum_{j=1}^q a_j h(w_j)}{\sum_{j=1}^q a_j} \geq h\left(\frac{\sum_{j=1}^q a_j w_j}{\sum_{j=1}^q a_j} \right) = h(q^{-1}) = q^{-p^\star},$$

where the first inequality follows from Jensen's inequality since h is convex (see, for instance, Proposition 3.8 of Aragón et al. (2019)), and the second equality holds because $\sum_{j=1}^q w_j = 1$ for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. Since $p^\star \in (1, \infty)$, equality holds if and only if w_j are the same for all $j = 1, \ldots, q$. Recall the support of $\boldsymbol{w}$ is $\mathcal{W}_{\text{cvx}}$, it follows that $\arg\min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} m(\boldsymbol{w}) = \frac{1}{q} \mathbf{1}_q$. □

B.7 Alternative procedures for inference

In this section, I discuss alternative approaches for conducting inference around τ. Section 4.5 focuses on conducting inference using FLCI around θ. This section examines additional large-sample properties of the estimated weights by viewing $\widehat{\tau}$ as estimators estimated from the constrained optimization problems introduced in Sections 4.

In this subsection, I establish the large-sample properties for the weights obtained from the minimax problem in (21) using an estimated weights. Let $\widehat{\mathbf{\Sigma}}_n$ be a consistent estimator of $\mathbf{\Sigma}$ and define

$$\widehat{\boldsymbol{w}}_n \equiv \underset{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}}{\arg\min}\ \widehat{R}_{\max,n}(B,\boldsymbol{w}) \quad \text{where} \quad \widehat{R}_{\max,n}(B,\boldsymbol{w}) \equiv \boldsymbol{w}'\widehat{\mathbf{\Sigma}}_n\boldsymbol{w} + \overline{M}(B,\boldsymbol{w}). \tag{B.74}$$

The minimax problem (21) has a similar structure to Section 3.4.1 in that both approaches consider the same class of weights $\mathcal{W}_{\text{cvx}}$. The difference arises from adding the extra term $\overline{M}(B,\boldsymbol{w})$ to the objective. The results in this section can be viewed as an extension to the results in Section A.7.4. I am going to impose Assumption A.25 for the analysis. First, I consider the probability limit.

Proposition B.12. *Let Assumption A.25 hold and consider a finite $B \geq 0$. Consider $\widehat{\boldsymbol{w}}_n$ as defined in (B.74). Let $\boldsymbol{w}^\star(B)$ be the optimal solution to the minimax problem (21) as defined in (22). Then, $\widehat{\boldsymbol{w}}_n \xrightarrow{p} \boldsymbol{w}^\star(B)$.*

Next, I show the limiting distribution below.

Proposition B.13. *Consider the same notations and assumptions as in Proposition B.12. Let $\overline{\boldsymbol{w}}(\boldsymbol{\zeta})$ be the optimal solution to the following program:*

$$\min_{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}}\ [R_{max}(\boldsymbol{w},B) + \boldsymbol{\zeta}'\boldsymbol{w}], \tag{B.75}$$

where $\boldsymbol{\zeta} \in \mathbb{R}^q$. Then,

$$\sqrt{n}(\widehat{\boldsymbol{w}}_n - \boldsymbol{w}^\star(B)) = \sqrt{n}\left[\overline{\boldsymbol{w}}\left(2(\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma})\boldsymbol{w}^\star(B)\right) - \boldsymbol{w}^\star(B)\right] + o_{\mathbb{P}}(1).$$

B.7.1 Supplemental lemmas and their proofs

Lemma B.14. *Let Assumption A.25 hold. Consider (B.74). Let $\boldsymbol{w}^\star(B)$ be the optimal solution to (21) as defined in (22). Define $\mathsf{d}_n(\boldsymbol{w}) \equiv \widehat{R}_{max,n}(B,\boldsymbol{w}) - R_{max}(B,\boldsymbol{w})$ for a given finite $B \geq 0$ and $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. Then, $\mathsf{d}_n(\boldsymbol{w})$ is Lipschitz continuous and differentiable at $\boldsymbol{w}^\star(B)$.*

Proof of Lemma B.14. To begin with, note that $\mathsf{d}_n(\boldsymbol{w})$ can be written as

$$\mathsf{d}_n(\boldsymbol{w}) \equiv \widehat{R}_{\max,n}(B,\boldsymbol{w}) - R_{\max}(B,\boldsymbol{w}) = \boldsymbol{w}'(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{w}. \tag{B.76}$$

The above is differentiable at $\boldsymbol{w}^\star(B)$. For Lipschitz continuity, note that for any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}_{\text{cvx}}$,

$$\begin{aligned}
|\mathsf{d}_n(\boldsymbol{w}_1) - \mathsf{d}_n(\boldsymbol{w}_2)| &= \left|[\boldsymbol{w}_1'(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{w}_1] - [\boldsymbol{w}_2'(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{w}_2]\right| \\
&= |(\boldsymbol{w}_1 + \boldsymbol{w}_2)'(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})(\boldsymbol{w}_1 - \boldsymbol{w}_2)| \\
&\leq |\boldsymbol{w}_1'(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})(\boldsymbol{w}_1 - \boldsymbol{w}_2)| + |\boldsymbol{w}_2'(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})(\boldsymbol{w}_1 - \boldsymbol{w}_2)| \\
&\leq \|\boldsymbol{w}_1\|_2 \|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + \|\boldsymbol{w}_2\|_2 \|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 \\
&\leq 2\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2,
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from the Cauchy-Schwarz inequality, the third inequality uses $\max_{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}} \|\boldsymbol{w}\|_2 \leq 1$. Hence, d_n is Lipschitz continuous. □

Lemma B.15. *Consider the same notations and assumptions as in Lemma B.14. Then,*

$$\|\nabla \mathsf{d}_n(\boldsymbol{w})\|_2 = O_{\mathbb{P}}(n^{-1/2}).$$

Proof of Lemma B.15. From (B.76),

$$\nabla \mathsf{d}_n(\boldsymbol{w}) = 2(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{w}. \tag{B.77}$$

Hence,

$$\|\nabla \mathsf{d}_n(\boldsymbol{w})\|_2 \leq 2\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F \tag{B.78}$$

where the inequality follows from the Cauchy-Schwarz inequality and that $\|\boldsymbol{w}\|_2 \leq 1$. Here, $\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma} = O_{\mathbb{P}}(n^{-1/2})$ by Assumption A.25(c). Thus, the result follows. □

Lemma B.16. *Consider the same notations and assumptions as in Lemma B.14. Then, there exists a neighborhood $\mathcal{W}$ of $\boldsymbol{w}^\star(B)$ such that*

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\|\nabla \mathsf{d}_n(\boldsymbol{w}) - \nabla \mathsf{d}_n(\boldsymbol{w}^\star(B))\|_2}{n^{-1/2} + \|\boldsymbol{w} - \boldsymbol{w}^\star(B)\|_2} = o_{\mathbb{P}}(1).$$

Proof of Lemma B.16. To begin with, note that

$$\begin{aligned}\|\nabla \mathsf{d}_n(\boldsymbol{w}) - \nabla \mathsf{d}_n(\boldsymbol{w}^\star(B))\|_2 &= \|2(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})(\boldsymbol{w} - \boldsymbol{w}^\star(B))\|_2 \\ &\leq 2\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F \|\boldsymbol{w} - \boldsymbol{w}^\star(B)\|_2, \end{aligned} \tag{B.79}$$

where the first equality uses (B.77) and the second line uses Cauchy-Schwarz inequality.

Hence,

$$\frac{\|\nabla \mathsf{d}_n(\boldsymbol{w}) - \nabla \mathsf{d}_n(\boldsymbol{w}^\star(B))\|_2}{n^{-1/2} + \|\boldsymbol{w} - \boldsymbol{w}^\star(B)\|_2} \leq 2\frac{\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F \|\boldsymbol{w} - \boldsymbol{w}^\star(B)\|_2}{n^{-1/2} + \|\boldsymbol{w} - \boldsymbol{w}^\star(B)\|_2} \leq 2\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F, \tag{B.80}$$

where the first inequality uses (B.79) and the second inequality uses $\frac{\|\boldsymbol{w}-\boldsymbol{w}^\star(B)\|_2}{n^{-1/2}+\|\boldsymbol{w}-\boldsymbol{w}^\star(B)\|_2} \leq 1$. The result follows from Assumption A.25 that $\widehat{\boldsymbol{\Sigma}}_n$ is a consistent estimator of $\boldsymbol{\Sigma}$. The neighborhood can be taken to be a small ball around $\boldsymbol{w}^\star(B)$. □

Lemma B.17. *Consider the same notations and assumptions as in Proposition B.12. Let $\overline{\boldsymbol{w}}(\boldsymbol{\zeta})$ be the optimal solution to (B.75). There exists a neighborhood $\mathcal{W}$ of $\boldsymbol{w}^\star(B)$ and a positive constant C_w such that for any $\boldsymbol{\zeta}$ in a neighborhood of $\mathbf{0}_q$, (B.75) has an optimal solution $\overline{\boldsymbol{w}}(\boldsymbol{\zeta}) \in \mathcal{W}$ and*

$$R_{max}(\boldsymbol{w}, B) + \boldsymbol{\zeta}'[\boldsymbol{w} - \overline{\boldsymbol{w}}(\boldsymbol{\zeta})] \geq R_{max}(\overline{\boldsymbol{w}}(\boldsymbol{\zeta}), B) + \mathsf{C}_w\|\boldsymbol{w} - \overline{\boldsymbol{w}}(\boldsymbol{\zeta})\|_2^2, \tag{B.81}$$

for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}} \cap \mathcal{W}$.

Proof of Lemma B.17. To begin with, the objective of (B.75) is strictly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. It follows that the optimal solution $\overline{\boldsymbol{w}}(\boldsymbol{\zeta})$ is unique. The minimum eigenvalue of $\boldsymbol{\Sigma}$ is also bounded below from 0 because $\boldsymbol{\Sigma}$ is assumed to be positive definite in Assumption A.25(b). Note that $R_{\max}(\boldsymbol{w}, B) = f_1(\boldsymbol{w}) + f_2(\boldsymbol{w})$ where $f_1(\boldsymbol{w}) = \overline{M}(B, \boldsymbol{w})$ is convex in $\boldsymbol{w}$, $f_2(\boldsymbol{w}) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$ is twice continuously differentiable, with $\nabla^2 f_2(\boldsymbol{w}) = \boldsymbol{\Sigma}$. It follows from pages 832 to 833 of Shapiro (1993) that the statements of this lemma hold. □

B.7.2 Proofs

Proof of Proposition B.12. Write $M_n(\boldsymbol{w}) \equiv -\widehat{R}_{\max,n}(\boldsymbol{w}, B)$ and $M(\boldsymbol{w}) \equiv -R_{\max}(\boldsymbol{w}, B)$ for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ and a finite $B \geq 0$. I also write $\boldsymbol{w}_0 \equiv \boldsymbol{w}^\star(B)$. I show that the proposition holds by verifying the assumptions in Theorem 2.12(i) of Kosorok (2008).

First, suppose that for some sequence in $\{\widehat{\boldsymbol{w}}_n\} \in \mathcal{W}_{\text{cvx}}$, $\liminf_{n\to\infty} M(\widehat{\boldsymbol{w}}_n) \geq M(\boldsymbol{w}_0)$ but $\|\widehat{\boldsymbol{w}}_n - \boldsymbol{w}_0\|_2 \longrightarrow 0$ does not hold. This means there exists $\varepsilon > 0$ and a subsequence

$\{\widehat{\boldsymbol{w}}_{n_k}\}$ such that

$$\|\widehat{\boldsymbol{w}}_{n_k} - \boldsymbol{w}_0\|_2 \geq \varepsilon. \tag{B.82}$$

Note that $R_{\max}(\boldsymbol{w}, B)$ is strictly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. As a result, using a similar argument as in (B.14) and (B.19), and noting that $\boldsymbol{w}_0$ is the optimal solution to the minimax problem, it follows that for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$, $R_{\max}(\boldsymbol{w}, B) \geq R_{\max}(\boldsymbol{w}_0, B) + \kappa\|\boldsymbol{w} - \boldsymbol{w}_0\|_2^2$ for some $\kappa > 0$. Such κ can be taken to be $\frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma})$ where $\lambda_{\min}(\boldsymbol{\Sigma})$ is the smallest eigenvalue of $\boldsymbol{\Sigma}$. Here, $\kappa > 0$ because $\boldsymbol{\Sigma}$ is positive definite. This means

$$-M(\widehat{\boldsymbol{w}}_{n_k}) \geq -M(\boldsymbol{w}_0) + \kappa\|\widehat{\boldsymbol{w}}_{n_k} - \boldsymbol{w}_0\|_2^2 \geq -M(\boldsymbol{w}_0) + \kappa\varepsilon^2, \tag{B.83}$$

where the first inequality follows from the preceding discussion on strong convexity and the second inequality uses (B.82). Rearranging (B.83) gives

$$M(\boldsymbol{w}_0) - \kappa\varepsilon^2 \geq M(\widehat{\boldsymbol{w}}_{n_k}).$$

This implies that $M(\boldsymbol{w}_0) - \kappa\varepsilon^2 \geq \liminf_{k\to\infty} M(\widehat{\boldsymbol{w}}_{n_k}) \geq \liminf_{n\to\infty} M(\widehat{\boldsymbol{w}}_n)$. This contradicts $\liminf_{n\to\infty} M(\widehat{\boldsymbol{w}}_n) \geq M(\boldsymbol{w}_0)$. Hence, the hypothesis in Theorem 2.12 of Kosorok (2008) holds.

Next, $M_n(\widehat{\boldsymbol{w}}_n) = \sup_{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}} M_n(\boldsymbol{w})$ by property of the minimax problem. In addition, for any $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$,

$$|M_n(\boldsymbol{w}) - M(\boldsymbol{w})| = |\boldsymbol{w}'(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{w}| \leq \|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F\|\boldsymbol{w}\|_2^2 \leq \|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F.$$

It is assumed that $\widehat{\boldsymbol{\Sigma}}_n \xrightarrow{p} \boldsymbol{\Sigma}$ by Assumption A.25. Hence, $\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_F = o_{\mathbb{P}}(1)$. As a result, this implies that $\sup_{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}}} |M_n(\boldsymbol{w}) - M(\boldsymbol{w})| \xrightarrow{p} 0$. The assumptions in Theorem 2.12(i) of Kosorok (2008) also hold. Hence, the consistency result of this proposition follows. □

Proof of Proposition B.13. With the given assumptions, (21) has a unique optimal solution following the discussion in (4.3). Hence, Assumption A1 of Shapiro (1993) holds. Next, Assumptions A3, B1 to B3 of Shapiro (1993) also hold from Lemmas B.14 to B.17. $\widehat{\boldsymbol{w}}_n$ is also consistent for $\boldsymbol{w}^\star(B)$ by Proposition B.12. It follows from Theorem 2.1 of Shapiro (1993) that

$$\widehat{\boldsymbol{w}}_n = \overline{\boldsymbol{w}}(\nabla \mathsf{d}_n(\boldsymbol{w}^\star(B))) + o_{\mathbb{P}}(n^{-1/2}). \tag{B.84}$$

Since $\boldsymbol{w}^\star(B) = \overline{\boldsymbol{w}}(\mathbf{0}_q)$, I obtain the following from (B.84):

$$\sqrt{n}(\widehat{\boldsymbol{w}}_n - \boldsymbol{w}^\star(B)) = \sqrt{n}\left[\overline{\boldsymbol{w}}(\nabla \mathsf{d}_n(\boldsymbol{w}^\star(B))) - \overline{\boldsymbol{w}}(\mathbf{0}_q)\right] + o_{\mathbb{P}}(1),$$

where $\nabla \mathsf{d}_n(\boldsymbol{w}^\star(B))$ is given in (B.77). $\square$

C Details on computation

This section studies how the minimax and adaptive approaches can be implemented in practice to compute optimal weights to average the treatment effects.

C.1 Minimax approach

The minimax problem in (21) is strictly convex in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$. Hence, convex optimization algorithms can be applied. With specific choices of norms and parameter spaces, the problem of finding the minimax optimal weights can be written as a quadratic program, so that modern solvers such as Gurobi, can be immediately applied.

In the following subsections, I discuss computation for the parameter spaces considered in Section 4.2 and how the computational problems can be written as one optimization problem. I consider computation with a finite $B > 0$.

C.1.1 Computation for parameter space that bounds misspecification

To begin with, suppose that the researcher places a bound on the misspecification as in Example 4.2. When the ℓ_1-norm is used in the parameter space (19), the maximum misspecification can be written as $\overline{M}(B, \boldsymbol{w}) = B^2 \|\boldsymbol{w}\|_\infty^2 = B^2 \max_{j=1,\dots,q} w_j^2$ because $w_j \in [0,1]$ for $j = 1, \dots, q$. Then, the optimization problem becomes the following quadratic program with linear constraints:

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^q,\ t \in \mathbb{R}} \quad & (\boldsymbol{w}' \boldsymbol{\Sigma} \boldsymbol{w} + B^2 t^2), \\ \text{s.t.} \quad & \boldsymbol{w} \leq t \mathbf{1}_q, \\ & \boldsymbol{w}' \mathbf{1}_q = 1, \\ & \boldsymbol{w} \geq \mathbf{0}_q, \end{aligned}$$

where the auxiliary variable t is used to constrain that $\|\boldsymbol{w}\|_\infty^2 = \max_{j=1,\dots,q} w_j^2$ equals the smallest t such that $t \geq w_j$ for $j = 1, \dots, q$ (see, for instance, Bertsimas and Tsitsiklis (1997, page 17)).

Similarly, when the ℓ_2-norm is used in the parameter space as in Example 4.2, then $\overline{M}(B, \boldsymbol{w}) = B^2 \|\boldsymbol{w}\|_2^2$. The optimization problem becomes the following quadratic program with linear constraints:

$$\min_{\boldsymbol{w} \in \mathbb{R}^q} \quad \boldsymbol{w}' (\boldsymbol{\Sigma} + B^2 \boldsymbol{I}_q) \boldsymbol{w},$$

$$\begin{aligned} \text{s.t.} \quad & \boldsymbol{w}'\mathbf{1}_q = 1, \\ & \boldsymbol{w} \geq \mathbf{0}_q. \end{aligned}$$

C.1.2 Computation for parameter space that has shape restrictions

Next, suppose shape restrictions are imposed as in Example 4.3 or Section B.4.1. Consider the shape constraints as described in (B.33). Then, the minimax problem can be written as

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathcal{W}_{\text{cvx}},\, t\in\mathbb{R}} \quad & (\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + t^2), \\ \text{s.t.} \quad & \max\left\{\left(\max_{\tilde{\boldsymbol{b}}\in\mathcal{S}(B)} \boldsymbol{w}'\boldsymbol{b}\right), \left(\max_{\tilde{\boldsymbol{b}}\in\mathcal{S}(B)} -\boldsymbol{w}'\boldsymbol{b}\right)\right\} \leq t. \end{aligned} \tag{C.1}$$

The inequality constraint can be further written as two constraints $\max_{\boldsymbol{b}\in\mathcal{S}(B)} \boldsymbol{w}'\boldsymbol{b} \leq t$ and $\max_{\boldsymbol{b}\in\mathcal{S}(B)}(-\boldsymbol{w}'\boldsymbol{b}) \leq t$. Since $\mathcal{S}(B) \neq \emptyset$ is assumed in Section 4.2, Slater's condition (Boyd and Vandenberghe, 2004, Chapter 5.2.3) is satisfied when there is a point such that strict inequality holds. Assuming that this is true, it follows that the program can be written as

$$\begin{aligned} \max_{\boldsymbol{b}\in\mathcal{S}(B)} \boldsymbol{w}'\boldsymbol{b} &= \min_{\boldsymbol{\mu}\geq\mathbf{0}_l} \left\{\max_{\|\boldsymbol{b}\|_p\leq B} [\boldsymbol{w}'\boldsymbol{b} + \boldsymbol{\mu}'(\boldsymbol{r} - \boldsymbol{Q}\boldsymbol{b})]\right\} \\ &= \min_{\boldsymbol{\mu}\geq\mathbf{0}_l} \left[\boldsymbol{\mu}'\boldsymbol{r} + \max_{\|\boldsymbol{b}\|_p\leq B} (\boldsymbol{w} - \boldsymbol{Q}'\boldsymbol{\mu})'\boldsymbol{b}\right] \\ &= \min_{\boldsymbol{\mu}\geq\mathbf{0}_l} \left(\boldsymbol{\mu}'\boldsymbol{r} + B\|\boldsymbol{w} - \boldsymbol{Q}'\boldsymbol{\mu}\|_{p^\star}\right), \end{aligned} \tag{C.2}$$

where the first equality follows from the corresponding Lagrange dual problem (see, for instance, Boyd and Vandenberghe (2004, Chapter 5)) and the third equality follows from properties of dual norm (see, for instance, Boyd and Vandenberghe (2004, Chapter A.1.6)) with the $\ell_{p^\star}$-norm being the dual norm for the ℓ_p-norm. Similarly, I have

$$\max_{\boldsymbol{b}\in\mathcal{S}(B)} (-\boldsymbol{w}'\boldsymbol{b}) = \min_{\boldsymbol{\mu}\geq\mathbf{0}_l} \left(\boldsymbol{\mu}'\boldsymbol{r} + B\|\boldsymbol{w} + \boldsymbol{Q}'\boldsymbol{\mu}\|_{p^\star}\right) \tag{C.3}$$

Since both (C.2) and (C.3) are optimization problems over their corresponding Lagrange multipliers, it follows that the minimax problem under the shape-constrained

parameter space (B.33) can be written as

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^q,\ t\in\mathbb{R},\ \boldsymbol{\mu}_1\in\mathbb{R}^l,\ \boldsymbol{\mu}_2\in\mathbb{R}^l} \quad & (\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + t^2), \\ \text{s.t.} \quad & \boldsymbol{\mu}_1'\boldsymbol{r} + B\|\boldsymbol{w} - \boldsymbol{Q}'\boldsymbol{\mu}_1\|_{p^\star} \leq t, \\ & \boldsymbol{\mu}_2'\boldsymbol{r} + B\|\boldsymbol{w} + \boldsymbol{Q}'\boldsymbol{\mu}_2\|_{p^\star} \leq t, \\ & \boldsymbol{w}'\mathbf{1}_q = 1, \\ & \boldsymbol{w} \geq \mathbf{0}_q, \\ & \boldsymbol{\mu}_1 \geq \mathbf{0}_l, \\ & \boldsymbol{\mu}_2 \geq \mathbf{0}_l. \end{aligned} \tag{C.4}$$

When the ℓ_2-norm is used in the parameter space in (B.33), problem (C.4) again becomes a quadratic problem.

C.1.3 Computation for parameter space on the communication model

The exact computation procedure depends on modeling assumption on $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$. Suppose that $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ is known (i.e., Case 1 of Section B.4.2). This can be formulated as a shape-constrained problem as in Section B.4.1. Therefore, the optimization program (C.4) can still be applied by setting $\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{\gamma}' \\ -\boldsymbol{\gamma}' \end{pmatrix}$.

C.2 Adaptive approach

As in Section C.1, computing the optimally adaptive weight for (25) can be formulated as a single convex optimization problem. This is because the adaptive regret is a maximum of two convex (or strictly convex) functions in $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ over a compact set. In addition, the optimization problem for the adaptive weights has the following epigraph form (see, for instance, Boyd and Vandenberghe (2004, Chapter 4)):

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^q,\ t\in\mathbb{R}} \quad & t, \\ \text{s.t.} \quad & A(\underline{B}, \boldsymbol{w}) \leq t, \\ & A(\overline{B}, \boldsymbol{w}) \leq t, \\ & \boldsymbol{w}'\mathbf{1}_q = 1, \\ & \boldsymbol{w} \geq \mathbf{0}_q. \end{aligned} \tag{C.5}$$

Similar to Section C.1, I discuss computation for different parameter spaces considered in Section 4.2, and how formulating them as a single optimization problem that can readily apply modern solvers is possible.

Remark C.1. With the penalized formulation discussed in Remark 4.12, the optimization problem can be modified slightly from (C.5) as follows:

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^q,\ t\in\mathbb{R}} \quad & t, \\ \text{s.t.} \quad & A(\underline{B},\boldsymbol{w}) + \kappa\|\boldsymbol{w}\|_2^2 \le t, \\ & A(\overline{B},\boldsymbol{w}) + \kappa\|\boldsymbol{w}\|_2^2 \le t, \\ & \boldsymbol{w}'\mathbf{1}_q = 1, \\ & \boldsymbol{w} \ge \mathbf{0}_q, \end{aligned}$$

for $\kappa > 0$. ■

C.2.1 Computation for parameter space that bounds misspecification

Suppose that the researcher places a bound on the misspecification as Example 4.2. Then, the adaptive regret at a specific $B \ge 0$ and $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ can be written as $A(B,\boldsymbol{w}) = \frac{\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}+B^2\|\boldsymbol{w}\|_{p^\star}^2}{R^\star(B)}$ using (20) and (24). Therefore, problem (C.5) can be further written as

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^q,\ t\in\mathbb{R}} \quad & t, \\ \text{s.t.} \quad & \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + \underline{B}^2\|\boldsymbol{w}\|_{p^\star}^2 \le tR^\star(\underline{B}), \\ & \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + \overline{B}^2\|\boldsymbol{w}\|_{p^\star}^2 \le tR^\star(\overline{B}), \\ & \boldsymbol{w}'\mathbf{1}_q = 1, \\ & \boldsymbol{w} \ge \mathbf{0}_q. \end{aligned} \tag{C.6}$$

In the above, note that $R^\star(\underline{B})$ and $R^\star(\overline{B})$ are the minimax risks that are inputs to the program. When the ℓ_2-norm is used in (19), then $p^\star = 2$, so (C.6) becomes a quadratic program. This is because the objective and the constraint for $\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}$ are linear, the two inequality constraints are convex quadratic functions.

C.2.2 Computation for parameter space that has shape restrictions

Suppose shape restrictions are imposed as in Section B.4.1 with the parameter space as in (B.33). A similar technique involving epigraph and Lagrangian dual formulation as in

Section C.1.2 can be applied. Therefore, problem (C.5) can be further written as follows

$$
\begin{aligned}
\min_{\substack{\boldsymbol{w}\in\mathbb{R}^q,\ t\in\mathbb{R},\ \boldsymbol{\mu}_{1,1},\boldsymbol{\mu}_{1,2},\boldsymbol{\mu}_{2,1},\boldsymbol{\mu}_{2,2}\in\mathbb{R}^l \\ u_{1,1},u_{1,2},u_{2,1},u_{2,2}\in\mathbb{R}_+}} \quad & t, \\
\text{s.t.} \quad & \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + u_{1,1}^2 \leq tR^\star(\underline{B}), \\
& \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + u_{1,2}^2 \leq tR^\star(\underline{B}), \\
& \boldsymbol{\mu}_{1,1}'\boldsymbol{r} + \underline{B}\|\boldsymbol{w} - \boldsymbol{Q}'\boldsymbol{\mu}_{1,1}\|_{p^\star} \leq u_{1,1}, \\
& \boldsymbol{\mu}_{1,2}'\boldsymbol{r} + \underline{B}\|\boldsymbol{w} + \boldsymbol{Q}'\boldsymbol{\mu}_{1,2}\|_{p^\star} \leq u_{1,2}, \\
& \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + u_{2,1}^2 \leq tR^\star(\overline{B}), \\
& \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + u_{2,2}^2 \leq tR^\star(\overline{B}), \\
& \boldsymbol{\mu}_{2,1}'\boldsymbol{r} + \overline{B}\|\boldsymbol{w} - \boldsymbol{Q}'\boldsymbol{\mu}_{2,1}\|_{p^\star} \leq u_{2,1}, \\
& \boldsymbol{\mu}_{2,2}'\boldsymbol{r} + \overline{B}\|\boldsymbol{w} + \boldsymbol{Q}'\boldsymbol{\mu}_{2,2}\|_{p^\star} \leq u_{2,2}, \\
& \boldsymbol{w}'\mathbf{1}_q = 1, \\
& \boldsymbol{w} \geq \mathbf{0}_q, \\
& \boldsymbol{\mu}_{1,1}, \boldsymbol{\mu}_{1,2}, \boldsymbol{\mu}_{2,1}, \boldsymbol{\mu}_{2,2} \geq \mathbf{0}_l.
\end{aligned}
\tag{C.7}
$$

Similar to (C.6), $R^\star(\underline{B})$ and $R^\star(\overline{B})$ are the minimax risks that are inputs to the program. When the ℓ_2-norm is used in (B.33), then $p^\star = 2$, so (C.7) becomes a quadratic program.

C.2.3 Computation for parameter space on ambiguity in communication

The exact computation procedure depends on the modeling assumption of $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$. Suppose that $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ is known (i.e., Case 1 of Section B.4.2). This can be formulated as a shape-constrained problem as in Section B.4.1. Therefore, the optimization program (C.7) can still be applied by setting $\boldsymbol{Q}$ as in Section C.1.3.

D Communication and subjective weights

D.1 Some preliminary results for the finite-sample model

Recall from (15) that $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is known. The misspecification is captured by $\boldsymbol{b} = \boldsymbol{\beta} - \theta \mathbf{1}_q$. The additional restriction in this section is that $\theta = \boldsymbol{\gamma}'\boldsymbol{\beta}$, where $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$. Thus, the vector of bias can be written as

$$\boldsymbol{b} = \boldsymbol{\beta} - (\boldsymbol{\gamma}'\boldsymbol{\beta})\mathbf{1}_q. \tag{D.1}$$

I explore various restrictions on $\boldsymbol{\gamma}$ in this section.

The following lemma is helpful in connecting with the previous results.

Lemma D.1. *Consider $B \geq 0$, $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$, and the following sets*

$$\begin{aligned}
\mathcal{S}_0(B, \boldsymbol{\gamma}) &\equiv \{\boldsymbol{b} \in \mathbb{R}^q : \|\boldsymbol{b}\|_p \leq B, \boldsymbol{b} = \boldsymbol{\beta} - (\boldsymbol{\gamma}'\boldsymbol{\beta})\mathbf{1}_q \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^q\}, \\
\mathcal{S}_1(B, \boldsymbol{\gamma}) &\equiv \{\boldsymbol{b} \in \mathbb{R}^q : \|\boldsymbol{b}\|_p \leq B, \boldsymbol{\gamma}'\boldsymbol{b} = 0\}.
\end{aligned}$$

Then, $\mathcal{S}_0(B, \boldsymbol{\gamma}) = \mathcal{S}_1(B, \boldsymbol{\gamma})$.

Proof of Lemma D.1. For any $\boldsymbol{x} \in \mathcal{S}_0(B, \boldsymbol{\gamma})$, there exists $\boldsymbol{\beta} \in \mathbb{R}^q$ such that $\boldsymbol{x} = \boldsymbol{\beta} - (\boldsymbol{\gamma}'\boldsymbol{\beta})\mathbf{1}_q$. Hence,

$$\boldsymbol{\gamma}'\boldsymbol{x} = \boldsymbol{\gamma}'\boldsymbol{\beta} - \boldsymbol{\gamma}'\boldsymbol{\beta}(\boldsymbol{\gamma}'\mathbf{1}_q) = \boldsymbol{\gamma}'\boldsymbol{\beta} - \boldsymbol{\gamma}'\boldsymbol{\beta} = 0,$$

because $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$. In addition, $\|\boldsymbol{x}\|_p \leq B$ since $\boldsymbol{x} \in \mathcal{S}_0(B, \boldsymbol{\gamma})$. It follows that $\boldsymbol{x} \in \mathcal{S}_1(B, \boldsymbol{\gamma})$. This means $\mathcal{S}_0(B, \boldsymbol{\gamma}) \subseteq \mathcal{S}_1(B, \boldsymbol{\gamma})$.

Now consider any $\boldsymbol{x} \in \mathcal{S}_1(B, \boldsymbol{\gamma})$ and $t \in \mathbb{R}$. Let $\boldsymbol{y} \equiv \boldsymbol{x} + t\mathbf{1}_q$. Then,

$$\boldsymbol{y} - (\boldsymbol{\gamma}'\boldsymbol{y})\mathbf{1}_q = \boldsymbol{x} + t\mathbf{1}_q - (\boldsymbol{\gamma}'\boldsymbol{x} + t\boldsymbol{\gamma}'\mathbf{1}_q)\mathbf{1}_q = \boldsymbol{x} + t\mathbf{1}_q - t\mathbf{1}_q = \boldsymbol{x},$$

where the second equality holds because $\boldsymbol{\gamma}'\boldsymbol{x} = 0$ as $\boldsymbol{x} \in \mathcal{S}_1(B, \boldsymbol{\gamma})$. Since $\|\boldsymbol{x}\|_p \leq B$ also holds, it follows that $\boldsymbol{x} \in \mathcal{S}_0(B, \boldsymbol{\gamma})$. This means $\mathcal{S}_1(B, \boldsymbol{\gamma}) \subseteq \mathcal{S}_0(B, \boldsymbol{\gamma})$. Therefore, the proof is complete. □

The above lemma is useful in that under the model (15), (D.1) and that $\|\boldsymbol{b}\|_p \leq B$, finding the "worst-case $\boldsymbol{b}$" does not require searching over $\boldsymbol{\beta} \in \mathbb{R}^q$ as well. It only requires an orthogonality condition $\boldsymbol{\gamma}'\boldsymbol{b} = 0$.

D.2 Known γ

This corresponds to Case 1 of Section B.4.2. Suppose $\gamma \in \mathcal{W}_{\text{cvx}}$ is known to the researcher. In this case, the maximum risk is a function of $w \in \mathcal{W}_{\text{cvx}}$ (to be chosen) by the minimax/adaptive problem and a known $\gamma \in \mathcal{W}_{\text{cvx}}$.

In this and the later subsection, I focus on using the ℓ_p norm to restrict b. In addition, I will be using the orthogonal restriction version of the parameter space (i.e., the space $\mathcal{S}_1(B, \gamma)$) in Lemma D.1. In this subsection of a known $\gamma \in \mathcal{W}_{\text{cvx}}$, I will write the parameter space as

$$\mathcal{S}_1(B, \gamma) = \{b \in \mathbb{R}^q : \|b\|_p \leq B, \gamma' b = 0\}, \tag{D.2}$$

for some $p \geq 1$. Shape restrictions, as in Examples 4.3, can be incorporated if needed.

Using the risk function in (18), the maximum risk is given by

$$\begin{aligned} R_{\max}(B, w, \gamma) &= \max_{b \in \mathcal{S}_1(B,\gamma)} [w'\Sigma w + (w'b)^2] \\ &= w'\Sigma w + \max_{b \in \mathcal{S}_1(B,\gamma)} (w'b)^2 \\ &\equiv V(w) + \overline{M}(B, w, \gamma), \end{aligned} \tag{D.3}$$

where $V(w) \equiv w'\Sigma w$ and $\overline{M}(B, w, \gamma) \equiv \max_{b \in \mathcal{S}_1(B,\gamma)} (w'b)^2$.

Next, let $\widetilde{w} \equiv w - \frac{w'\gamma}{\|\gamma\|_2^2}\gamma$. Then, for any $b \in \mathcal{S}_1(B, \gamma)$,

$$w'b = \widetilde{w}'b + \frac{(w'\gamma)(\gamma'b)}{\|\gamma\|_2^2} = \widetilde{w}'b, \tag{D.4}$$

where the second equality follows because $\gamma' b = 0$ for any $b \in \mathcal{S}_1(B, \gamma)$.

Let the $\ell_{p^\star}$-norm be the dual norm of the ℓ_p-norm that satisfies $\frac{1}{p} + \frac{1}{p^\star} = 1$. Then,

$$\overline{M}(B, w, \gamma) = \max_{b \in \mathcal{S}_1(B,\gamma)} (w'b)^2 = \max_{b \in \mathcal{S}_1(B,\gamma)} (\widetilde{w}'b)^2 = B^2 \max_{\widetilde{b} \in \mathcal{S}_1(1,\gamma)} (\widetilde{w}'\widetilde{b})^2 \tag{D.5}$$

where the first equality uses the definition of the maximum risk in (D.3), the second equality uses (D.4), and the third equality reparameterized b by defining $\widetilde{b}$ such that $B\widetilde{b} = b$ and that $\widetilde{b} \in \mathcal{S}_1(1, \gamma)$. In the above, the objective $(\widetilde{w}'\widetilde{b})^2 = \widetilde{b}'(\widetilde{w}\widetilde{w}')\widetilde{b}$ is convex in $\widetilde{b}$. In addition, the set $\mathcal{S}_1(1, \gamma)$ is nonempty and compact. Hence, by Bauer's maximum principle (see, for instance, Niculescu and Persson (2018, Corollary A.3.3)), the

supremum is attained at an extreme point of the set $\mathcal{S}_1(1,\boldsymbol{\gamma})$.

Note that from (D.5), the maximum bias function is multiplicatively separable in that it can be written as

$$\overline{M}(B,\boldsymbol{w},\boldsymbol{\gamma}) = B^2 m(\boldsymbol{w},\boldsymbol{\gamma}), \tag{D.6}$$

where $m(\boldsymbol{w},\boldsymbol{\gamma}) \equiv \max_{\widetilde{\boldsymbol{b}}\in\mathcal{S}_1(1,\boldsymbol{\gamma})}(\widetilde{\boldsymbol{w}}'\widetilde{\boldsymbol{b}})^2 = \max_{\widetilde{\boldsymbol{b}}\in\mathcal{S}_1(1,\boldsymbol{\gamma})}(\boldsymbol{w}'\widetilde{\boldsymbol{b}})^2$ using (D.4).

To illustrate the solution to the minimax problem, I consider an example with two outcomes below.

Example D.2 (Known $\boldsymbol{\gamma}$)**.** Suppose $q = 2$ and that $(\widehat{\beta}_1, \widehat{\beta}_2)$ are normally distributed as in Example 4.8. Let $\boldsymbol{w} \equiv (w_1, w_2)'$ and $\boldsymbol{\gamma} \equiv (\gamma_1, \gamma_2)'$. Since $\boldsymbol{w}, \boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$, the weights can be written as $\boldsymbol{w} = (w_1, 1-w_1)'$ and $\boldsymbol{\gamma} = (\gamma_1, 1-\gamma_1)'$. Assume that the ℓ_2-norm is used in $\mathcal{S}(B,\boldsymbol{\gamma})$ so that the dual norm is also the ℓ_2-norm. Then,

$$\widetilde{\boldsymbol{w}} = \boldsymbol{w} - \frac{\boldsymbol{w}'\boldsymbol{\gamma}}{\|\boldsymbol{\gamma}\|_2^2}\boldsymbol{\gamma} = \frac{\gamma_1 - w_1}{\gamma_1^2 + (1-\gamma_1)^2}\begin{pmatrix}\gamma_1 - 1\\ \gamma_1\end{pmatrix}. \tag{D.7}$$

Using the expression of $\widetilde{\boldsymbol{w}}$ in (D.7), the ℓ_2-norm is given by

$$\|\widetilde{\boldsymbol{w}}\|_2^2 = \frac{(\gamma_1 - w_1)^2}{\gamma_1^2 + (1-\gamma_1)^2} = \frac{w_1^2 - 2w_1\gamma_1 + \gamma_1^2}{\gamma_1^2 + (1-\gamma_1)^2}. \tag{D.8}$$

Using (D.6) and (D.8),

$$\overline{M}(B, w_1, \gamma_1) = B^2\|\widetilde{\boldsymbol{w}}\|_2^2 = \frac{B^2(w_1^2 - 2w_1\gamma_1 + \gamma_1^2)}{\gamma_1^2 + (1-\gamma_1)^2}. \tag{D.9}$$

Together with the expression of $V(w_1)$ given in (B.39), the minimax problem under maximum risk (D.3) can be written as follows

$$R^\star(B,\gamma_1) = \min_{w_1\in[0,1]} R_{\max}(B, w_1, \gamma_1),$$

where

$$\begin{aligned} &R_{\max}(B, w_1, \gamma_1) \\ &\quad = (1 + \sigma_2^2 - 2\rho\sigma_2)w_1^2 + 2(\rho\sigma_2 - \sigma_2^2)w_1 + \sigma_2^2 + \frac{B^2(w_1^2 - 2w_1\gamma_1 + \gamma_1^2)}{\gamma_1^2 + (1-\gamma_1)^2}. \end{aligned} \tag{D.10}$$

The first-order condition for (D.10) with respect to w_1 leads to

$$w^\star_{1,\text{int}}(B,\gamma_1) = \frac{\sigma_2^2 - \rho\sigma_2 + \frac{B^2\gamma_1}{\gamma_1^2+(1-\gamma_1)^2}}{1+\sigma_2^2-2\rho\sigma_2+\frac{B^2}{\gamma_1^2+(1-\gamma_1)^2}}. \tag{D.11}$$

The optimal solution $w_1^\star(B,\gamma_1)$ to the minimax problem above is

$$w_1^\star(B,\gamma_1) = \begin{cases} 0 & \text{if } B^2\gamma_1 \leq [\gamma_1^2+(1-\gamma_1)^2](\rho\sigma_2-\sigma_2^2), \\ 1 & \text{if } B^2(1-\gamma_1) \leq (\rho\sigma_2-1)[\gamma_1^2+(1-\gamma_1)^2], \\ w^\star_{1,\text{int}}(B,\gamma_1) & \text{otherwise.} \end{cases} \tag{D.12}$$

Below are some observations from the optimal solution $w_1^\star(B,\gamma_1)$ in (D.12). First, consider $B=0$. This corresponds to the case that $b_1 = b_2 = 0$. The interior weight in (D.11) becomes $w^\star_{1,\text{int}}(0,\boldsymbol{\gamma}) = \frac{\sigma_2^2-\rho\sigma_2}{1+\sigma_2^2-2\rho\sigma_2}$. This coincides with the optimal weight in (23) when $B=0$ in minimizing variances. If the DGP satisfies $\sigma^2-\rho\sigma_2 \leq 0$, then it is optimal to set $w_1^\star(0,\gamma_1)=0$ because the second treatment effect is more precise.

As B increases, it may not be optimal to place all weights on one of the treatment effects even if one is noisier than the other. When $B>0$, the optimal weight considers both variances and the bias due to w_1 not matching γ_1.

As $B \longrightarrow \infty$, the optimal weights are such that $\lim_{B\to\infty} w_1^\star(B,\gamma_1) = \gamma_1$. This means choosing $\boldsymbol{w}$ to match $\boldsymbol{\gamma}$. This follows because, as the bias of the treatment effects can grow to infinity, the loss of not matching the true $\boldsymbol{\gamma}$ also grows to infinity. △

The above example demonstrates that even if θ is a known weighted average of treatment effects, it is not always optimal to choose $\boldsymbol{w}$ to match $\boldsymbol{\gamma}$ because the true mean $\boldsymbol{\beta}$ is unknown.

D.3 Unknown γ

This corresponds to Case 2 of Section B.4.2 that assumes $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ is unknown to the researcher. As discussed in Lemma D.1. I will write the parameter space as

$$\mathcal{S}_2(B) = \{(\boldsymbol{b},\boldsymbol{\gamma}) \in \mathbb{R}^{2q} : \|\boldsymbol{b}\|_p \leq B, \boldsymbol{\gamma}'\boldsymbol{b} = 0, \boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}\}, \tag{D.13}$$

for some $p \geq 1$. Here, the maximum risk is maximizing over $\boldsymbol{b} \in \mathcal{S}(B)$ and $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ as follows:

$$\begin{aligned} R_{\max}(B, \boldsymbol{w}) &= \max_{(\boldsymbol{b},\boldsymbol{\gamma}) \in \mathcal{S}_2(B)} R(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \\ &= \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + \max_{(\boldsymbol{b},\boldsymbol{\gamma}) \in \mathcal{S}_2(B)} (\boldsymbol{w}'\boldsymbol{\beta})^2 \\ &\equiv V(\boldsymbol{w}) + \overline{M}(B, \boldsymbol{w}), \end{aligned} \tag{D.14}$$

where $V(\boldsymbol{w}) \equiv \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$ and $\overline{M}(B, \boldsymbol{w}) \equiv \max_{(\boldsymbol{b},\boldsymbol{\gamma}) \in \mathcal{S}_2(B)} (\boldsymbol{w}'\boldsymbol{b})^2$.

Similar to Section D.2, I can reparameterize $\boldsymbol{b} = B\widetilde{\boldsymbol{b}}$ so that the maximum bias squared can be written as

$$\overline{M}(B, \boldsymbol{w}) = B^2 \max_{(\widetilde{\boldsymbol{b}},\boldsymbol{\gamma}) \in \mathcal{S}_2(1)} (\boldsymbol{w}'\widetilde{\boldsymbol{b}})^2 \equiv B^2 m(\boldsymbol{w}), \tag{D.15}$$

where $m(\boldsymbol{w}) = \max_{(\widetilde{\boldsymbol{b}},\boldsymbol{\gamma}) \in \mathcal{S}_2(1)} (\boldsymbol{w}'\widetilde{\boldsymbol{b}})^2$. Numerical methods can be used to compute $m(\boldsymbol{w})$.

Using (D.15), the minimax problem is

$$\min_{\boldsymbol{w} \in \mathcal{W}} R_{\max}(B, \boldsymbol{w}). \tag{D.16}$$

To illustrate the solution to the minimax problem (D.16), I consider an example with two outcomes below.

Example D.3 (Unknown γ)**.** Consider the same setup and notations as in Example D.2, except that the minimax problem (D.16) is considered.

Recall that (D.9) gives the maximum misspecification for a given $B \geq 0$, $w_1 \in [0,1]$ and $\gamma_1 \in [0,1]$. The maximum misspecification in (D.15) when γ_1 is unknown becomes

$$\overline{M}(B, w_1) \equiv \max_{\gamma_1 \in [0,1]} \overline{M}(B, w_1, \gamma_1) = B^2 \max_{\gamma_1 \in [0,1]} \frac{w_1^2 - 2w_1\gamma_1 + \gamma_1^2}{\gamma_1^2 + (1-\gamma_1)^2}. \tag{D.17}$$

Since $m(w_1, \gamma_1) \equiv \frac{w_1^2 - 2w_1\gamma_1 + \gamma_1^2}{\gamma_1^2 + (1-\gamma_1)^2}$ is continuous in γ_1 for any given $w_1 \in [0,1]$, the extremum value theorem states that the maximum is achieved in $[0,1]$. In the following, I show that

$$\overline{M}(B, w_1) = B^2 \max\{w_1^2, (1-w_1)^2\}. \tag{D.18}$$

This means I want to show that the maximum cannot be achieved in $(0,1)$. Note that

$$\frac{\partial m(w_1,\gamma_1)}{\partial \gamma_1} = \frac{2(\gamma_1 - w_1)[(2w_1-1)\gamma_1 + (1-w_1))}{[\gamma_1^2 + (1-\gamma_1)^2]^2}.$$

Hence, the critical points can be $\gamma_1 = w_1$ or $\gamma_1 = \frac{w_1-1}{2w_1-1}$. When $\gamma_1 = w_1$, then $m(w_1,w_1) = 0$. But $m(w_1,\gamma_1) \geq 0$, so $m(w_1,0), m(w_1,1) \geq m(w_1,w_1)$. Now consider $\gamma_1 = \frac{w_1-1}{2w_1-1}$. Then, for any $w_1 \in [0,1]$, the critical point $\gamma_{1,c}(w_1) \equiv \frac{w_1-1}{2w_1-1}$ can only be inside $[0,1]$ when $w_1 = 0$ and $w_1 = 1$. To see this, first note that $\gamma'_{1,c}(w_1) = \frac{2w_1-1-2(w_1-1)}{(2w_1-1)^2} = \frac{1}{(2w_1-1)^2}$, so that $\gamma'_{1,c}(w_1) > 0$ for $w_1 \in [0,1]$. In addition, $\lim_{w_1 \to 0.5^+} \gamma_{1,c}(w_1) = -\infty$ and $\lim_{w_1 \to 0.5^-} \gamma_{1,c}(w_1) = +\infty$. At $w_1 = 0$, $\gamma_{1,c}(0) = 1$. Then, $\gamma_{1,c}(w_1) \geq 1$ for $w_1 \in [0, 0.5]$. At $w_1 = 1$, $\gamma_{1,c}(1) = 0$. Then, $\gamma_{1,c}(w_1) \leq 0$ for $w_1 \in [0.5, 1]$. Therefore, (D.19) holds.

It follows that the maximum risk is

$$R_{\max}(B, w_1) \equiv V(w_1) + B^2 \max\{w_1^2, (1-w_1)^2\}. \tag{D.19}$$

To compute the optimal weights, there are two cases to consider depending on the value of w_1. First, suppose that $w_1 \in [0, \frac{1}{2})$, so that $\max\{w_1^2, (1-w_1)^2\} = (1-w_1)^2$. Using (D.19) with the expression of $V(w_1)$ given in (B.39), the minimax problem can be written as follows

$$R^\star(B) = \min_{w_1 \in [0,\frac{1}{2}]} [(1 + \sigma_2^2 - 2\rho\sigma_2)w_1^2 + 2(\rho\sigma_2 - \sigma_2^2)w_1 + \sigma_2^2 + B^2(1-w_1)^2]. \tag{D.20}$$

The optimal solution to (D.16) is given by

$$w^\star_{1,1}(B) = \begin{cases} 0 & \text{if } \sigma_2^2 - \rho\sigma_2 \leq -B^2, \\ \frac{1}{2} & \text{if } \sigma_2^2 - 1 \geq -B^2, \\ \frac{\sigma_2^2 - \rho\sigma_2 + B^2}{1 + \sigma_2^2 - 2\rho\sigma_2 + B^2} & \text{otherwise.} \end{cases} \tag{D.21}$$

In the above, the interior solution is obtained from taking first-order conditions. The boundary solution follows from analyzing when the boundaries are hit and from noting that $1 + \sigma_2^2 - 2\rho\sigma_2 + 2B^2 \geq (1-\sigma_2)^2 + 2B^2 \geq 0$.

The way how B affects $w^\star_{1,1}(B)$ is similar to Example D.2. When $B = 0$, the solution $w^\star_{1,1}(0)$ focuses on variance minimization as long as w_1 is inside the domain $[0, \frac{1}{2}]$. As B increases, the optimal weight takes the bias into account.

Now, suppose $w_1 \in (\frac{1}{2}, 1]$, so that $\max\{w_1^2, (1-w_1)^2\} = w_1^2$. Using (D.19) with the

expression of $V(w_1)$ given in (B.39), the minimax problem can be written as follows

$$R^\star(B) = \min_{w_1 \in [\frac{1}{2}, 1]} [(1 + \sigma_2^2 - 2\rho\sigma_2)w_1^2 + 2(\rho\sigma_2 - \sigma_2^2)w_1 + \sigma_2^2 + B^2 w_1^2]. \tag{D.22}$$

The optimal solution to (D.16) is given by

$$w_{1,2}^\star(B) = \begin{cases} \frac{1}{2} & \text{if } 1 - \sigma_2^2 \geq -B^2, \\ 1 & \text{if } 1 - \rho\sigma_2 \leq -B^2, \\ \frac{\sigma_2^2 - \rho\sigma_2}{1 + \sigma_2^2 - 2\rho\sigma_2 + B^2} & \text{otherwise.} \end{cases} \tag{D.23}$$

Based on the optimal solutions in (D.21) and (D.23), the optimal solution to the minimax problem $\min_{w_1 \in \mathcal{W}} R_{\max}(B, w_1)$ is $w_{1,j^\star}^\star(B)$ where $j^\star = \arg\min_{j=1,2} R_{\max}(B, w_{1,j}^\star(B))$. △

D.4 Reference weight

This corresponds to Case 3 of Section B.4.2, where I assume that there exists a known reference weight $\boldsymbol{\eta} \in \mathcal{W}_{\text{cvx}}$ such that $\boldsymbol{\gamma} = \boldsymbol{\eta} + \boldsymbol{\delta}$, $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$, $\|\boldsymbol{\delta}\|_p \leq D$, and $D \geq 0$. This can be interpreted as a researcher who believes that a certain vector of weights $\boldsymbol{\eta}$ is likely to be the true weights on $\boldsymbol{\beta}$ for θ, but there is some ambiguity around $\boldsymbol{\eta}$. The vector $\boldsymbol{\delta}$ represents such ambiguity, and the norm of $\boldsymbol{\delta}$ is bounded above by some $D \geq 0$. Therefore, the parameter space can be written as

$$\begin{aligned} &\mathcal{S}_3(B, D, \boldsymbol{\eta}) \\ &\quad = \{(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathbb{R}^{2q} : \|\boldsymbol{b}\|_p \leq B, \|\boldsymbol{\delta}\|_p \leq D, \boldsymbol{\gamma} = \boldsymbol{\eta} + \boldsymbol{\delta} \in \mathcal{W}_{\text{cvx}}, (\boldsymbol{\eta} + \boldsymbol{\delta})'\boldsymbol{b} = 0\}, \end{aligned} \tag{D.24}$$

where $B, D \geq 0$ and $\boldsymbol{\eta} \in \mathcal{W}_{\text{cvx}}$.

The following proposition summarizes that the two cases discussed in Sections D.2 and D.3 can be viewed as special cases of the general "reference weight" case.

Proposition D.4. *Let $B \geq 0$ and $\boldsymbol{\eta}_0 \in \mathcal{W}_{\text{cvx}}$ be given. Consider the notations and parameter spaces described in (D.2), (D.13), and (D.24).*

(a) If $D = 0$, then

$$\mathcal{S}_3(B, 0, \boldsymbol{\eta}_0) = \{(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathbb{R}^{2q} : \boldsymbol{b} \in \mathcal{S}_1(B, \boldsymbol{\eta}_0), \boldsymbol{\gamma} = \boldsymbol{\eta}_0\}.$$

(b) If $D \geq \underline{D}$ where $\underline{D} \equiv \max_{\boldsymbol{y} \in \mathcal{W}_{\text{cvx}}} \|\boldsymbol{y} - \boldsymbol{\eta}_0\|_p$, then

$$\mathcal{S}_3(B, D, \boldsymbol{\eta}_0) = \mathcal{S}_2(B).$$

Proof of Proposition D.4(a). Suppose that $D = 0$, this implies that $\boldsymbol{\delta} = \mathbf{0}_q$. Hence, $\boldsymbol{\gamma} + \boldsymbol{\delta} = \boldsymbol{\gamma} = \boldsymbol{\eta}_0$. This means $\mathcal{S}_3(B, D, \boldsymbol{\eta})$ in (D.24) becomes

$$\begin{aligned}
\mathcal{S}_3(B, 0, \boldsymbol{\eta}_0) &= \{(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathbb{R}^{2q} : \|\boldsymbol{b}\|_p \leq B, \|\boldsymbol{\delta}\|_p \leq 0, \boldsymbol{\gamma} = \boldsymbol{\eta}_0 + \boldsymbol{\delta} \in \mathcal{W}_{\text{cvx}}, (\boldsymbol{\eta}_0 + \boldsymbol{\delta})'\boldsymbol{b} = 0\} \\
&= \{(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathbb{R}^{2q} : \|\boldsymbol{b}\|_p \leq B, \boldsymbol{\gamma} = \boldsymbol{\eta}_0 \in \mathcal{W}_{\text{cvx}}, \boldsymbol{\eta}_0'\boldsymbol{b} = 0\} \\
&= \{(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathbb{R}^{2q} : \boldsymbol{b} \in \mathcal{S}_1(B, \boldsymbol{\eta}_0), \boldsymbol{\gamma} = \boldsymbol{\eta}_0\},
\end{aligned}$$

as desired, where $\mathcal{S}_1(B, \boldsymbol{\eta}_0)$ is given in (D.2). □

Proof of Proposition D.4(b). Let $\mathcal{S}_2(B)$ be a given in (D.13). For any $(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathcal{S}_3(B, D, \boldsymbol{\eta}_0)$, I have $\boldsymbol{\gamma}'\boldsymbol{b} = 0$, $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ and $\|\boldsymbol{b}\|_p \leq B$ by construction. Hence, $(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathcal{S}_2(B)$. This shows that $\mathcal{S}_3(B, D, \boldsymbol{\eta}_0) \subseteq \mathcal{S}_2(B)$.

Next, for any $(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathcal{S}_2(B)$, write $\boldsymbol{\delta}_\gamma \equiv \boldsymbol{\gamma} - \boldsymbol{\eta}_0$. Then, by definition, it must be that $\|\boldsymbol{\delta}_\gamma\|_p = \|\boldsymbol{\gamma} - \boldsymbol{\eta}_0\|_p \leq \max_{\boldsymbol{y} \in \mathcal{W}_{\text{cvx}}} \|\boldsymbol{y} - \boldsymbol{\eta}_0\|_p \leq \underline{D}$. In addition, $\|\boldsymbol{b}\|_p \leq B$, $\boldsymbol{\gamma} \in \mathcal{W}_{\text{cvx}}$ and $\boldsymbol{\gamma}'\boldsymbol{b} = 0$ by the definition of $\mathcal{S}_2(B)$. It follows that $(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathcal{S}_3(B, D, \boldsymbol{\eta}_0)$. This shows that $\mathcal{S}_2(B) \subseteq \mathcal{S}_3(B, D, \boldsymbol{\eta}_0)$. Hence, the proof is complete. □

Using the parameter space given in (D.24), the minimax problem is

$$R^\star(B, D, \boldsymbol{\eta}) \equiv \min_{\boldsymbol{w} \in \mathcal{W}_{\text{cvx}}} R_{\max}(B, D, \boldsymbol{w}, \boldsymbol{\eta}), \tag{D.25}$$

where the maximum risk is given by

$$\begin{aligned}
R_{\max}(B, D, \boldsymbol{w}, \boldsymbol{\eta}) &= \max_{(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathcal{S}_3(B, D, \boldsymbol{\eta})} \left[\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + (\boldsymbol{w}'\boldsymbol{b})^2\right] \\
&= \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} + \max_{(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathcal{S}_3(B, D, \boldsymbol{\eta})} (\boldsymbol{w}'\boldsymbol{b})^2 \\
&\equiv V(\boldsymbol{w}) + \overline{M}(B, D, \boldsymbol{w}, \boldsymbol{\eta}),
\end{aligned} \tag{D.26}$$

where $V(\boldsymbol{w}) \equiv \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$ and $\overline{M}(B, \boldsymbol{w}, \boldsymbol{\gamma}) \equiv \max_{(\boldsymbol{b}, \boldsymbol{\gamma}) \in \mathcal{S}_3(B, D, \boldsymbol{\eta})} (\boldsymbol{w}'\boldsymbol{b})^2$.

Similar to Sections D.2 and D.3, I can reparameterize $\boldsymbol{b} = B\widetilde{\boldsymbol{b}}$ so that the maximum bias squared can be written as

$$\overline{M}(B, D, \boldsymbol{w}, \boldsymbol{\eta}) = B^2 \max_{(\widetilde{\boldsymbol{b}}, \boldsymbol{\gamma}) \in \mathcal{S}_3(1, D, \boldsymbol{\eta})} (\boldsymbol{w}'\boldsymbol{\gamma})^2 \equiv B^2 m(\boldsymbol{w}, D, \boldsymbol{\eta}), \tag{D.27}$$

where $m(\boldsymbol{w}, D, \boldsymbol{\eta}) = \max_{(\widetilde{\boldsymbol{b}},\boldsymbol{\gamma})\in\mathcal{S}_3(1,D,\boldsymbol{\eta})}(\boldsymbol{w}'\widetilde{\boldsymbol{b}})^2$.

To illustrate the solution to the minimax problem, I consider an example with two outcomes below.

Example D.5 (Reference weight)**.** Consider the same setup and notations as in Examples D.2 and D.3, except that the minimax problem (D.25) is considered. In addition, let $D > 0$ in this example.

Let $\boldsymbol{\eta} = (\eta_1, \eta_2)'$ and $\boldsymbol{\delta} = (\delta_1, \delta_2)'$. Since it is required that $\boldsymbol{\eta} + \boldsymbol{\delta} \in \mathcal{W}_{\text{cvx}}$ and $\boldsymbol{\eta} \in \mathcal{W}_{\text{cvx}}$, it must be that $\eta_2 = 1 - \eta_1$ and that $1 + (\delta_1 + \delta_2) = 1$. Thus, $\delta_2 = -\delta_1$. Hence, the restriction that $D \geq \|\boldsymbol{\delta}\|_2$ can be rewritten as $D \geq (\delta_1^2 + \delta_2^2)^{\frac{1}{2}}$. Since $D \geq 0$, this restriction is equivalent to $|\delta_1| \leq \frac{D}{\sqrt{2}} \equiv \widetilde{D}$. Hence, the parameter space $\mathcal{S}_3(B, D, \boldsymbol{\eta})$ can be specialized in the following form in the current example:

$$\begin{aligned}\widetilde{\mathcal{S}}_3(B, \widetilde{D}, \eta_1) = \{(b_1, b_2, \delta_1) \in \mathbb{R}^3 : & \, b_1^2 + b_2^2 \leq B^2, \\ & |\delta_1| \leq \widetilde{D}, \\ & \gamma_1 = \delta_1 + \eta_1 \in [0, 1], \\ & \gamma_1 b_1 + (1 - \gamma_1) b_2 = 0\}.\end{aligned} \tag{D.28}$$

Here, $\delta_1 + \eta_1 \in [0, 1]$ is the same as $\boldsymbol{\eta} + \boldsymbol{\delta} \in \mathcal{W}_{\text{cvx}}$ in this example. To see this, note that $\boldsymbol{\eta} + \boldsymbol{\delta} \in \mathcal{W}_{\text{cvx}}$ requires $\boldsymbol{\eta} + \boldsymbol{\delta} \geq \mathbf{0}_q$ and $(\boldsymbol{\eta} + \boldsymbol{\delta})'\mathbf{1}_q = 1$. $(\boldsymbol{\eta} + \boldsymbol{\delta})'\mathbf{1}_q = 1$ is always satisfied by the parameterization at the beginning of this example because $(\boldsymbol{\eta} + \boldsymbol{\delta})'\mathbf{1}_q = (\eta_1 + 1 - \eta_1) + (\delta_1 - \delta_1) = 1$. $\boldsymbol{\eta} + \boldsymbol{\delta} \geq \mathbf{0}_q$ requires $\eta_1 + \delta_1 \geq 0$ and $1 - \eta_1 - \delta_1 \geq 0$. Combining these two inequality constraints gives the second last condition in the parameter space (D.28). For the last condition, it follows directly from $\boldsymbol{\gamma}'\boldsymbol{b} = 0$.

Note that (D.28) can be further simplified as follows. This is because $|\delta_1| \leq \widetilde{D}$ can be written as $-\widetilde{D} \leq \delta_1 \leq \widetilde{D}$. $\delta_1 + \eta_1 \in [0, 1]$ can be written as $-\eta_1 \leq \delta_1 \leq 1 - \eta_1$. Combining both inequalities give $\max\{-\eta_1, -\widetilde{D}\} \leq \delta_1 \leq \min\{\widetilde{D}, 1 - \eta_1\}$.

Since $\widetilde{D}, \eta_1 \geq 0$, let

$$\widetilde{\mathcal{S}}_3(\widetilde{D}, \eta_1) = \left\{\delta_1 \in \mathbb{R} : -\min\{\eta_1, \widetilde{D}\} \leq \delta_1 \leq \min\{\widetilde{D}, 1 - \eta_1\}\right\}. \tag{D.29}$$

Using the derivation in (D.9) and (D.17) but with γ_1 replaced by $\delta_1 + \eta_1$, the maximum misspecification can be written as

$$\overline{M}(B, \widetilde{D}, w_1, \eta_1) = B^2 \max_{\delta_1 \in \widetilde{\mathcal{S}}_3(\widetilde{D}, \eta_1)} \frac{w_1^2 - 2w_1(\delta_1 + \eta_1) + (\delta_1 + \eta_1)^2}{(\delta_1 + \eta_1)^2 + [1 - (\delta_1 + \eta_1)]^2}. \tag{D.30}$$

Let $t \equiv \delta_1 + \eta_1$, $\underline{\gamma}_1 \equiv -\min\{\eta_1, \widetilde{D}\} + \eta_1$ and $\overline{\gamma}_1 \equiv \min\{\widetilde{D}, 1 - \eta_1\} + \eta_1$. Then, (D.30) can be written as

$$\overline{M}(B, \widetilde{D}, w_1, \eta_1) = B^2 \max_{t \in [\underline{\gamma}_1, \overline{\gamma}_1]} \frac{w_1^2 - 2w_1 t + t^2}{t^2 + (1-t)^2} = B^2 \max_{t \in [\underline{\gamma}_1, \overline{\gamma}_1]} \frac{(w_1 - t)^2}{t^2 + (1-t)^2}. \tag{D.31}$$

The above problem has the same structure as (D.17) except that the support on t is different. It can be analyzed using a similar argument as in Example D.3. △