Counterfactual Mean-variance Optimization

Kwangho Kim*  Alan Mishler†  José R. Zubizarreta‡

August 2022

Abstract

We study a new class of estimands in causal inference, which are the solutions to a stochastic nonlinear optimization problem that in general cannot be obtained in closed form. The optimization problem describes the counterfactual state of a system after an intervention, and the solutions represent the optimal decisions in that counterfactual state. In particular, we develop a counterfactual mean-variance optimization approach, which can be used for optimal allocation of resources after an intervention. We propose a doubly-robust nonparametric estimator for the optimal solution of the counterfactual mean-variance program. We analyze rates of convergence and provide a closed-form expression for the asymptotic distribution of our estimator. Our analysis shows that the proposed estimator is robust against nuisance model misspecification, and can attain fast $\sqrt{n}$ rates with tractable inference even when using nonparametric methods. This result is applicable to general nonlinear optimization problems subject to linear constraints whose coefficients are unknown and must be estimated. In this way, our findings contribute to the literature in optimization as well as causal inference. We further discuss the problem of calibrating our counterfactual covariance estimator to improve the finite-sample properties of our proposed optimal solution estimators. Finally, we evaluate our methods via simulation, and apply them to problems in healthcare policy and portfolio construction.

Keywords: causal inference, counterfactual estimation, semiparametric theory, stochastic optimization

*Department of Health Care Policy, Harvard Medical School; kkim@hcp.med.harvard.edu
†J. P. Morgan AI Research, New York; alan@alanmishler.com
‡Department of Health Care Policy, Harvard Medical School; Departments of Biostatistics and Statistics, Harvard University; zubizarreta@hcp.med.harvard.edu
1 Introduction

Counterfactual or potential outcomes are used to describe how a unit would respond to a treatment or intervention, irrespective of whether the intervention actually takes place. They are commonly used in causal inference (Rubin 1974, Holland 1986, Höfler 2005). More recently, potential outcomes have also proven to be useful for counterfactual prediction and decision-making in medicine and epidemiology (e.g., Dickerman and Hernández 2020, Schulam and Saria 2017, Lin et al. 2021). In this work, we explore an optimization-based framework for estimating counterfactual quantities. In particular, we study a domain-general counterfactual mean-variance optimization problem that parameterizes a tradeoff between the mean and variance of an outcome in an intervened-upon world.

Traditional, non-counterfactual mean-variance optimization has been used to estimate optimal allocations of resources in various settings, including financial investment (Markowitz 1968), decision theory (Meyer 1987), product development (Cardozo and Smith 1983), healthcare policy (Fagefors and Lantz 2021, Qu et al. 2012), and electrical engineering (Delarue et al. 2011). In counterfactual mean-variance optimization, we aim to estimate optimal allocations under hypothetical interventions that change the distribution of outcomes, such as a healthcare policy intervention that modifies the compliance rate for patient scheduled appointments (Section 6.2).

A complication arises when our estimand is the optimal solution of a constrained optimization problem involving functions of counterfactual distributions (or simply counterfactuals), since accurate estimation of the counterfactuals does not necessarily guarantee accurate estimation of the estimand. This leads us to ask under what conditions these counterfactuals may be estimated with flexible nonparametric methods while preserving fast rates and achieving the property of double robustness (Bang and Robins 2005). To investigate this question, we leverage tools in nonlinear stochastic programming to analyze how the asymptotic properties of the estimated
counterfactual quantities in turn affect the asymptotic properties of estimators derived from the optimal solutions to the optimization problem.

1.1 Related Work

We analyze the counterfactual mean-variance optimization problem as a quadratic program (QP) where the objective function and, optionally, the constraints involve counterfactuals. Similar problems have been studied in causal inference settings, such as policy evaluation (e.g., Kita-gawa and Tetenov 2018, Athey and Wager 2021), optimal treatment regimes with limited resources (e.g., Luedtke and van der Laan 2016, Luedtke and Van Der Laan 2015), and nonparametric estimation using the projection approach, where nonparametric estimands are projected onto parametric models (e.g., Neugebauer and van der Laan 2007, Semenova and Chernozhukov 2021, Kennedy et al. 2021). They also appear in the context of algorithmic fairness (e.g., Mishler and Kennedy 2021, Mishler et al. 2021) and general counterfactual prediction problems (Coston et al. 2020), in the sense that these studies seek to minimize a loss function involving counterfactuals over a certain parameter space. However, despite its popularity in non-counterfactual settings, mean-variance optimization under counterfactual scenarios has not yet been studied. Furthermore, to our knowledge, no previous work has provided a formal analysis of the asymptotic properties of estimated optimal solutions in general nonlinear counterfactual optimization problems.

Our problem can be viewed as a stochastic program, as the coefficients of the program depend on an unknown probability distribution. The two standard approaches to solving stochastic programming problems are stochastic approximation (SA) and sample average approximation (SAA) (e.g., Nemirovski et al. 2009, Shapiro et al. 2014). In particular, there is a rich literature studying the statistical properties of SAA estimators; see, for example, Shapiro et al. (2014).
Chapter 5) for an excellent overview. The SA and SAA approaches are infeasible in our setting, however, as the relevant sample moments and stochastic (sub)gradients depend on unobserved counterfactuals. Moreover, these approaches cannot incorporate efficient estimators for counterfactual components, e.g., doubly-robust or semiparametric estimators with cross-fitting (Chernozhukov et al. 2017; Newey and Robins 2018). We therefore consider more general approaches beyond the SA and SAA settings (e.g., Shapiro 1991, 1993, 2000), which requires additional effort due to the inapplicability in our setting of certain standard stochastic assumptions about the behavior of the optimal solution and its estimator.

1.2 Contribution and Outline

We study counterfactual mean-variance optimization as a new tool to inform decisions under hypothetical scenarios, i.e. scenarios that do not obtain at present but may occur in the future. Our work lies at the intersection of the causal inference and optimization literatures, and is organized as follows. In Section 2, we describe the problem and present notation and assumptions. In Section 3, we analyze rates of convergence and derive a closed-form expression for the asymptotic distribution of estimated optimal solutions in general smooth nonlinear stochastic programming problems subject to linear constraints. Careful consideration is required since the particular form of the counterfactual parameters does not allow us to use standard results in the optimization literature. We provide conditions for achieving \( \sqrt{n} \)-consistency and tractable inference in terms of generic estimators for the program components. These results are not specific to mean-variance optimization; they are applicable to other nonlinear optimization problems whose components are unknown and must be estimated. For example, our method may be applied to a constrained counterfactual prediction problem, as discussed in Section 7. Crucially, it can be especially useful in causal inference when the estimand of interest cannot be expressed in closed form. Based
on these results, in Section 4 we propose a flexible doubly robust nonparametric estimator for the optimal solution of the counterfactual mean-variance program. In Section 5 we discuss the problem of calibrating the estimated counterfactual covariance matrix to improve the finite-sample properties of our proposed optimal solution estimators, which has applications to various other statistical and data science domains that rely on covariance matrix estimation. In Section 6 we analyze the finite sample behavior of our estimators in a simulation study and apply our method to problems in healthcare and investing. Section 7 concludes.

2 Problem Formulation

Suppose that we have access to an i.i.d. sample \((Z_1, \ldots, Z_n)\) of tuples \(Z = (Y, A, X) \sim \mathbb{P}\) for some distribution \(\mathbb{P}\), multiple outcomes \(Y = (Y_1, \ldots, Y_k) \in \mathbb{R}^k\) for fixed and finite \(k\), covariates \(X \in \mathcal{X} \subset \mathbb{R}^d\), and treatment \(A \in \mathcal{A} = \{0, 1\}\). For simplicity, here we assume \(A\) is binary, but in principle it can be multi-valued. For \(1 \leq i \leq k\), we let \(Y^a_i\) denote the counterfactual outcome that would be observed under treatment \(A = a, a \in \mathcal{A}\) (Rubin 1974). We focus on the following counterfactual mean-variance optimization problem where we wish to balance the “reward” (mean) against the “risk” (variance) in an intervened-upon world in which a treatment variable \(A\) is set to the value \(a\):

\[
\begin{align*}
\text{minimize} & \quad 1/2 w^\top \Sigma^a w - \lambda w^\top m^a \\
\text{subject to} & \quad w \in S^a,
\end{align*}
\]

\((P_{\text{MV}})\)

for \(m_a = (m^a_1, \ldots, m^a_k)^\top\), \(\Sigma^a = (\Sigma^a_{ij})\) where \(m^a_i = \mathbb{E}[Y^a_i], \Sigma^a_{ij} = \text{cov}(Y^a_i, Y^a_j), 1 \leq i, j \leq k\), and some compact subset \(\mathcal{W} \subset \mathbb{R}^k\). \(S^a\) is a feasible set possibly involving counterfactuals. As in the
standard QP, we restrict ourselves to linear constraints. Our default choice for $\mathcal{S}^a$ is the set

$$\mathcal{S}^a = \{ w \mid w^\top 1 = 1, w \geq 0, w^\top m^a \geq r_{\text{min}} \},$$

which requires the solution to be a vector of convex weights such that the weighted counterfactual mean is always greater than some known threshold $r_{\text{min}} \in \mathbb{R}$ set by the user. Although feasible sets with this form appear frequently in practice, they can be replaced with any combination of (counterfactual) linear constraints without affecting the subsequent results. $\lambda \geq 0$ is a user-determined risk tolerance coefficient that quantifies their tolerance towards the risk. The higher the value of $\lambda$, the larger the variance the user is willing to tolerate in order to maximize the reward.

The parameters $m^a$ and $\Sigma^a$ are unknown and must be estimated. To this end, we first need to express the counterfactual quantities in terms of an observed data distribution $\mathbb{P}$, which can be accomplished under the following standard causal assumptions (e.g., Imbens and Rubin 2015, Chapter 12).

- (C1) **Consistency**: $Y = Y^a$ if $A = a$
- (C2) **No unmeasured confounding**: $A \perp \perp Y^a \mid X$
- (C3) **Positivity**: $\mathbb{P}(A = a \mid X) > 0$ a.s.

Assumptions [C1] - [C3] will apply throughout this paper. Under these assumptions, $\forall i, j \in \{1, \ldots, k\}$, $\mathbb{E}[Y^a_i]$ and $\mathbb{E}[Y^a_i Y^a_j]$ are identified as $\mathbb{E}\{\mathbb{E}[Y_i \mid X, A = a]\}$ and $\mathbb{E}\{\mathbb{E}[Y_i Y_j \mid X, A = a]\}$, respectively, and are thus estimable from the observed sample. In Section 4 we discuss an efficient estimation strategy for $m^a$ and $\Sigma^a$. Note that even though we develop our estimator under the above set of causal assumptions, the results in Section 3 will show that our methods...
may be readily extended to other identification strategies as well (e.g., instrumental variables, mediation).

**Notation.** For clarity, we use the subscripts $i, j$ only to index the different outcome variables $Y_1, \ldots, Y_k$, and we reserve the subscript $t$ to index the samples $Z_1, \ldots Z_n$. For any fixed vector $v$ and matrix $M$, we let $\|v\|_2$ and $\|M\|_F$ denote the Euclidean norm (or $L_2$-norm) and Frobenius norm, respectively. $\| \cdot \|_2$ is understood as the spectral norm when it is used with a matrix. Let $\mathbb{P}_n$ denote the empirical measure over $(Z_1, \ldots, Z_n)$. Given a fixed operator $h$ (e.g., an estimated function), we let $\mathbb{P}$ denote the conditional expectation over a new independent observation $Z$, as in $\mathbb{P}(h) = \mathbb{P}\{h(Z)\} = \int h(z) d\mathbb{P}(z)$. Further, we use $\|h\|_{2, \mathbb{P}}$ to denote the $L_2(\mathbb{P})$ norm of $h$ defined by $\|h\|_{2, \mathbb{P}} = [\mathbb{P}(h^2)]^{1/2} = [\int h(z)^2 d\mathbb{P}(z)]^{1/2}$. Lastly, we let $s^*(\mathbb{P})$ denote the set of optimal solutions of an optimization program $\mathbb{P}$, and we let $\text{dist}(x, S) = \inf \{\|x - y\|_2 : y \in S\}$ denote the distance from a point $x$ to a set $S$.

### 3 Asymptotic Analysis of Optimal Solutions

This section is devoted to analyzing the rates of convergence and asymptotic distribution for estimated optimal solutions in the more general setting of nonlinear stochastic programming with a varying feasible set. For a set $\mathbb{T} \subset \mathbb{R}^{d_T}$, $d_T < \infty$, consider a finite-dimensional statistical functional $T : \mathbb{P} \to \mathbb{T}$ and a twice continuously differentiable real-valued function $f : \mathbb{R}^k \times \mathbb{T} \to \mathbb{R}$. Also consider some $C \in \mathbb{R}^{rC \times k}$, $d \in \mathbb{R}^k$, possibly depending on $T(\mathbb{P})$. We assume that $T(\mathbb{P}), C,$ and $d$ are unknown and must be estimated, while $f$ is deterministic and known. We wish to estimate the optimal solution(s) of the following general-form nonlinear program with linear constraints

$$\begin{align*}
\text{minimize} & \quad f(x, T) \\
\text{subject to} & \quad x \in S_{lin} := \{x \mid Cx \leq d\},
\end{align*}
$$

(P_{nl})
where $\mathbb{X}$ is a compact subset of $\mathbb{R}^k$. $T \equiv T(\mathbb{P})$ is an unknown component, not the decision variable, so we suppress the dependence on $T$ and write $f(x) \equiv f(x, T)$ in most cases. Let $x^* \in s^*(\mathbb{P}_{nl})$ be an optimal solution to $\mathbb{P}_{nl}$. For simplicity, we consider feasible sets with only inequality constraints, but one could also consider feasible sets defined by both equalities and inequalities (for example by letting $S_{lin} := \{ x \mid Cx \leq d, Fx = e \}$ for some $F \in \mathbb{R}^{r \times k}, e \in \mathbb{R}^k$) without affecting our results. 

The program $\mathbb{P}_{nl}$ is a nonlinear stochastic optimization problem with linear constraints. Since the “true” program ($\mathbb{P}_{nl}$) is not directly solvable, we estimate the optimal solution $x^*$ by computing $\hat{x}$, an optimal solution of the following approximating program

$$
\begin{align*}
\text{minimize} & \quad \hat{f}(x) \equiv f(x, \hat{T}) \\
\text{subject to} & \quad x \in \hat{S}_{lin} \equiv \{ x \mid \hat{C}x \leq \hat{d} \}.
\end{align*}$$

($\hat{\mathbb{P}}_{nl}$)

We study the case $S_{lin}, \hat{S}_{lin} \neq \emptyset$, i.e. where both programs are feasible. Note that in $\hat{\mathbb{P}}_{nl}$ all the relevant stochastic information in $\hat{f}$ is contained in the argument estimate $\hat{T}$.

$T$ can be viewed as a set of counterfactuals of interest. The program $\mathbb{P}_{MV}$ is a special case of $\mathbb{P}_{nl}$ with $T = (m_1, \ldots, m_k, \Sigma_{11}, \ldots, \Sigma_{kk})^\top \in \mathbb{R}^{k(k+3)/2}$ and $r_c = k + 3$. The optimal value and optimal solution of $\mathbb{P}_{nl}$ depend on the unknown components $T, C$, and $d$, but they may not be continuous in these unknown components. Consequently, requiring that $\hat{T}, \hat{C}$, and $\hat{d}$ are accurate estimators is not by itself sufficient to guarantee that $\hat{x}$ is an accurate estimator as well. Moreover, the asymptotic behavior of optimal solutions in stochastic programming with varying feasible sets (i.e., in which the feasible set is not the same for the true and the approximating programs) has not been studied as widely as the case of a fixed feasible set. Only a few such studies exist, including the seminal works of [King and Rockafellar (1993)] and [Shapiro (1993)] which discuss

---

1This is because equality constraints can be always expressed by pairs of inequality constraints. However, in this case, the regularity assumptions in Section 3 need to be slightly modified accordingly (Still 2018, Section 1.1)
sufficient conditions for obtaining a closed-form expression for the asymptotic distribution of optimal solutions in general stochastic programming.

A larger challenge arises from the particular form of our counterfactual parameters. Standard methods in the stochastic optimization literature typically assume that each unknown component has the functional form \( \mathbb{E}[g(x, \theta)] \), with some observable random element \( \theta \) and some fixed, known function \( g \). In such settings, the SAA approach is readily applicable, i.e., each unknown component can be approximated as \( \mathbb{P}_n[g(x, \theta)] = n^{-1} \sum_{i=1}^{n} g(x, \theta_i) \). However, this approach cannot be applied when the function \( g \) is unknown; for example, in the case of counterfactual components whose identifying expressions are given by expectations over unknown (nuisance) regression functions. This makes the application of classical methods for analysis (e.g., King and Rockafellar 1993; Shapiro 1993) challenging in our problem, because some of the required conditions are not readily satisfied. For example, it is commonly assumed that optimal solution estimators are consistent, i.e., \( \hat{x} \xrightarrow{P} x^* \), since this is considered a very weak assumption in the SAA setting. Yet this is not easily verifiable in our case.

In order to make the analysis tractable, we focus on smooth stochastic programs of the form \( \mathbb{P}_{nl} \). In what follows, we show that the asymptotic properties of \( \hat{x} \) can be analyzed without assuming any particular form for the unknown components or their estimators.

For a feasible point \( \bar{x} \in S_{\text{lin}} \) we define the active index set for the inequality constraints.

**Definition 3.1 (Active set).** For \( \bar{x} \in S_{\text{lin}} \), we define the active index set \( J_0 \) by

\[
J_0(\bar{x}) = \{1 \leq j \leq r_C \mid C_j^T \bar{x} - d_j = 0\},
\]

where \( C_j^T \) is the \( j \)-th row of \( C \).

Next, we introduce conditions required to characterize the rates of convergence.
(A1) \[ \max \left\{ \| \hat{T} - T \|_2, \| \hat{C} - C \|_F, \| \hat{d} - d \|_2 \right\} = o_P(1) \]

(A2) For each \( x^* \in s^* \left( \mathcal{P}_{nl} \right) \),

\[ \varsigma^\top \nabla_x^2 f(x^*) \varsigma > 0, \quad \forall \varsigma \in \{ \varsigma \in \mathbb{R}^k \mid \nabla_x f(x^*)^\top \varsigma \leq 0, C_j^\top \varsigma \leq 0, j \in J_0(x^*) \} \setminus \{0\} \].

(A1) is a mild consistency assumption, with no requirement on rates of convergence. (A2) ensures that the quadratic growth condition holds at each \( x^* \), which implies that each \( x^* \) is locally isolated (Bonnans and Ioffe 1995). This type of second-order condition is widely used in nonlinear programming (e.g., Still 2018, Theorem 2.4). (A2) holds, for example, if \( \nabla_x^2 f(x^*) \) is positive definite. Under these two conditions, the consistency properties of \( \hat{x} \) can be established, as stated in the following theorem.

**Theorem 3.1.** Under (A1) [A2] we have

\[ \text{dist}(\hat{x}, s^* \left( \mathcal{P}_{nl} \right)) = O_P \left( \max \left\{ \| \hat{T} - T \|_2, \| \hat{C} - C \|_F, \| \hat{d} - d \|_2 \right\} \right). \]

A proof of Theorem 3.1 and all other results can be found in the appendix. Theorem 3.1 implies that the rates of convergence for estimating the optimal solutions of \( \mathcal{P}_{nl} \) are essentially given by the rates at which each component is estimated. In Appendix B.1 we show that the above result can be extended to more general smooth nonlinear constraints.

For the purposes of inference on \( \hat{x} \), it is also desirable to characterize the asymptotic distribution of \( \hat{x} \), which requires stronger assumptions than those required for consistency. Asymptotic properties of optimal solutions in stochastic programming are typically studied using a generalization of the delta method for directionally differentiable mappings (e.g., Shapiro 1993, 2000, Shapiro et al. 2014). Inference on optimal solution estimators is typically conducted using the
bootstrap. Thus, asymptotic normality is of particular interest, since without it the consistency of the bootstrap is no longer guaranteed for the solution estimators (Fang and Santos 2019).

We define two additional regularity conditions with respect to \((\mathcal{P}_{nl})\); these conditions are commonly invoked to study the asymptotic behavior of optimal solutions in nonlinear programming.

**Definition 3.2 (LICQ).** *The linear independence constraint qualification (LICQ) is satisfied at \(\bar{x} \in S_{\text{lin}}\) if the vectors \(C_j^T, j \in J_0(\bar{x})\) are linearly independent.*

**Definition 3.3 (SC).** *Strict Complementarity (SC) is satisfied at \(\bar{x} \in S_{\text{lin}}\) if, with multipliers (dual variables) \(\bar{\gamma} \geq 0\), the Karush-Kuhn-Tucker (KKT) conditions

\[
\nabla_x f(\bar{x}, T) + \sum_{j \in J_0(\bar{x})} \bar{\gamma}_j C_j^T \bar{x} = 0, \quad \text{diag}(\bar{\gamma})(C \bar{x} - d) = 0
\]

are satisfied such that \(\bar{\gamma}_j > 0, \forall j \in J_0(\bar{x})\).*

LICQ is arguably one of the most widely-used constraint qualifications that render the first-order KKT conditions necessary at optimal solutions points. SC means that if the \(j\)-th inequality constraint is active then the corresponding dual variable is strictly positive, so exactly one of \(\bar{\gamma}_j\) and \((C\bar{x})_j - d_j\) is zero for each \(1 \leq j \leq r_C\). SC is widely used in nonlinear programming, particularly in the context of parametric optimization (e.g., Still 2018; Shapiro et al. 2014). Now, we introduce additional assumptions required for analyzing the asymptotic distribution of \(\hat{x}\).

(A3) Program \((\mathcal{P}_{nl})\) has a unique optimal solution \(x^*\) (i.e., \(s^*(\mathcal{P}_{nl}) \equiv \{x^*\}\) is a singleton)

(A4) LICQ and SC hold at \(x^*\) with the corresponding multipliers \(\gamma^*\)

(A5) For some random vector \(\Upsilon \in \mathbb{R}^{k + |J_0(x^*)|}\) and \(|J_0(x^*)| \times k\) matrix \(C_{ac} = [C_j, j \in J_0(x^*)]\)

\[
\sqrt{n} \left[ \nabla_{x} f(\hat{x}) - \nabla_{x} f(x^*) + \sum_{j \in J_0(x^*)} \gamma_j^* \{\hat{C}_j - C_j\} \right] \overset{d}{\to} \Upsilon
\]
\begin{align*}
\text{(A6) } \max \left\{ \|\hat{T} - T\|_2, \|\hat{C} - C\|_F, \|\hat{d} - d\|_2 \right\} = O_p(n^{-1/2})
\end{align*}

Assumption (A5) roughly says that the KKT coefficients of the approximating program jointly converge in distribution to the KKT coefficients of the true program at \( \sqrt{n} \) rates (Shapiro 1993). Assumption (A6) is stronger than (A1), requiring \( \sqrt{n} \)-consistency rather than consistency at any rate. Note that under (A3), if LICQ holds at \( x^* \), then the corresponding multipliers \( \gamma^* \) are determined uniquely (Wachsmuth 2013). In the next theorem, we provide a closed-form expression for the asymptotic distribution of \( \hat{x} \).

**Theorem 3.2** (Asymptotic Distribution). Assume that (A2) - (A6) hold, and let \( \gamma^*_\text{ac} = [\gamma^*_j, j \in J_0(x^*)] \). Then

\[
n^{\frac{1}{2}} (\hat{x} - x^*) = \left[ \nabla^2_x f(x^*) \quad C^\top_{\text{ac}} \right]^{-1} \left[ \begin{array}{c} 1 \\ \operatorname{diag}(\gamma^*_{\text{ac}})1 \end{array} \right]^\top \Upsilon + o_p(1).
\]

Here, \( \Upsilon \) is an arbitrary random vector from Assumption (A5). If \( \Upsilon \) has a multivariate normal distribution, the above theorem gives explicit conditions under which \( \hat{x} \) is \( \sqrt{n} \)-consistent and asymptotically normal. In our proof, we follow Shapiro (1993) and use an expansion of \( \hat{x} \) in terms of an auxiliary parametric program. To derive the closed form, we compute the directional derivative of optimal solutions in the parametric program based on an appropriate form of the implicit function theorem (Dontchev and Rockafellar 2009). To our knowledge, the asymptotic results in theorems 3.1 and 3.2 are new in the causal inference and optimization literatures.

We note that the asymptotics of the optimal values for \( \hat{P}_{nl} \) can also be derived directly from Theorems 3.1 and 3.2 without requiring extra conditions. We omit these results here for conciseness.
4 Estimation and Inference

Based on the results in the previous section, here we propose a nonparametric estimator for the optimal solutions of $P_{MV}$ and show that the proposed estimator can achieve $\sqrt{n}$ rates and asymptotic normality under weak conditions. To this end, we first present efficient estimators for the counterfactual mean ($m^o$) and covariance ($\Sigma^o$) by leveraging tools in causal inference. To simplify notation, we first introduce the following nuisance functions

$\pi_a(X) = \mathbb{P}[A = a \mid X],$
$\mu_i(X, a) = \mathbb{E}[Y_i \mid X, A = a],$
$\sigma_{ij}(X, a) = \mathbb{E}[Y_i Y_j \mid X, A = a],$

and let $\hat{\pi}_a, \hat{\mu}_i,$ and $\hat{\sigma}_{ij}$ be some estimators of $\pi_a, \mu_i,$ and $\sigma_{ij},$ respectively. Further, we let

$\phi_i^o(Z; \eta_i) = \frac{\mathbb{I}(A = a)}{\pi_a(X)} \{Y_i - \mu_i(X, A)\} + \mu_i(X, a),$
$\phi_{ij}^o(Z; \eta_{ij}) = \frac{\mathbb{I}(A = a)}{\pi_a(X)} \{Y_i Y_j - \sigma_{ij}(X, A)\} + \sigma_{ij}(X, a),$

denote the uncentered efficient influence functions for the parameters $\psi_i^o \equiv \mathbb{E}[Y_i^o] = \mathbb{E}\{\mathbb{E}[Y_i \mid X, A = a]\}$ and $\psi_{ij}^o \equiv \mathbb{E}[Y_i^o Y_j^o] = \mathbb{E}\{\mathbb{E}[Y_i Y_j \mid X, A = a]\},$ with the relevant nuisance functions collectively denoted by $\eta_i = \{\pi_a(X), \mu_i(X, A)\}$ and $\eta_{ij} = \{\pi_a(X), \sigma_{ij}(X, A)\},$ respectively.

Estimation of mean counterfactual outcomes such as $\psi_i^o$ under the standard identification assumptions $\text{[C1]}$-$\text{[C3]}$ has been extensively studied in the causal inference literature, e.g., for the average treatment effect. The most widely-used estimators include the plug-in (PI) regression (or g-computation) and inverse probability weighting (IPW) estimators, which utilize information from the outcome regressions ($\mu_j$’s) and propensity score $\pi_a,$ respectively (e.g., Hernán and
Robins 2010).

Semiparametric (or doubly robust) estimators utilize both the outcome regression and the propensity score; they can be understood as an enhanced form of the PI or IPW estimators with an additional bias correction term (Robins and Rotnitzky 1995; Robins and Wang 2000). Semiparametric estimators have the appealing properties that 1) the fast parametric $\sqrt{n}$-rate is attainable even when all the nuisance functions are estimated flexibly at slower rates, and if this rate is attained then 2) they achieve asymptotic normality with semiparametric efficiency (Kennedy 2016).

For each $m_a^i$, $\Sigma_{ij}^a$, the corresponding semiparametric estimators are defined as

$$\hat{M}_i^a = \hat{M}_i^a(Z; \hat{\eta}_i) = \mathbb{P}_n \left\{ \hat{\phi}_i^a(Z) \right\} ,$$

$$\hat{\Sigma}_{ij}^a = \hat{\Sigma}_{ij}^a(Z; \hat{\eta}_i, \hat{\eta}_j, \hat{\eta}_{ij}) = \mathbb{P}_n \left\{ \hat{\phi}_{ij}^a(Z) \right\} - \mathbb{P}_n \left\{ \hat{\phi}_i^a(Z) \right\} \mathbb{P}_n \left\{ \hat{\phi}_j^a(Z) \right\} ,$$

where $\hat{\phi}_i^a(Z) = \phi_i^a(Z; \hat{\eta}_i)$ and $\hat{\phi}_{ij}^a(Z) = \phi_{ij}^a(Z; \hat{\eta}_{ij})$. First, we introduce the following assumptions for our nuisance function estimators.

(B1) $\hat{\pi}_i, \hat{\mu}_i, \hat{\mu}_j, \hat{\sigma}_{ij}$ are constructed using a single separate iid sample $D_0^n = \{Z_{a+1}, ..., Z_{2n}\}$

(B2) $\mathbb{P}(\hat{\pi}_a \in [\epsilon, 1-\epsilon]) = 1$ for some $\epsilon > 0$

(B3) $||\hat{\mu}_i - \mu_i||_{2,P} = o_\mathbb{P}(1)$, $||\hat{\pi}_i - \pi_i||_{2,P} = o_\mathbb{P}(1)$, $||\hat{\sigma}_{ij} - \sigma_{ij}||_{2,P} = o_\mathbb{P}(1)$

(B4) $||\hat{\pi}_a - \pi_a||_{2,P} \max_i ||\hat{\mu}_i - \mu_i||_{2,P} = o_\mathbb{P}(n^{-1/2})$, $||\hat{\pi}_a - \pi_a||_{2,P} \max_{i,j} ||\hat{\sigma}_{ij} - \sigma_{ij}||_{2,P} = o_\mathbb{P}(n^{-1/2})$

The above assumptions are commonly used in semiparametric estimation in causal inference (e.g., Kennedy 2016). Here we use sample splitting (B1) to allow for arbitrarily complex nuisance estimators, so that the nuisance estimators are constructed on a single separate independent
sample of the same size $n$ as the estimation sample on which $\text{P}_n$ operates. (See Remark 1). If one is willing to rely on appropriate empirical process conditions (e.g., Donsker-type conditions), then the nuisance estimators can be estimated on the same sample without (B1); however, this would limit the flexibility of the nuisance estimators. The requirement (B4) that the second-order nuisance errors converge to zero at faster than $\sqrt{n}$ rates is a sufficient condition commonly found in standard semiparametric settings with finite-dimensional parameters.

**Remark 1 (Sample splitting).** For nuisance estimation, we can always create separate independent samples by splitting the data in half (or in folds) at random; furthermore, full sample size efficiency can be attained by swapping the samples as in cross-fitting (e.g., Zheng and Van Der Laan 2010; Kennedy 2016; Chernozhukov et al. 2017; Newey and Robins 2018). Following previous studies (e.g., Kennedy 2020; Kennedy et al. 2021), for simplicity in the exposition we use a single split procedure in our analysis, as the extension to averages across independent splits is straightforward.

In what follows, we give conditions under which our proposed estimators in (1) and (2) are $\sqrt{n}$-consistent, asymptotically normal, and efficient.

**Lemma 4.1.** Under Assumptions (B1) - (B3) we have

\[
\|\hat{m}^a - m^a\|_2 = O_P \left( \|\hat{\pi}_a - \pi_a\|_{2,P} \max_i \|\hat{\mu}_i - \mu_i\|_{2,P} \right) + O_P \left( n^{-1/2} \right),
\]

\[
\|\hat{\Sigma}^a - \Sigma^a\|_2 = O_P \left( \|\hat{\pi}_a - \pi_a\|_{2,P} \max_i \|\hat{\mu}_i - \mu_i\|_{2,P} + \max_{i,j} \|\hat{\sigma}_{ij} - \sigma_{ij}\|_{2,P} \right) + O_P \left( n^{-1/2} \right).
\]

If we further assume the nonparametric conditions (B4) then

\[
\sqrt{n}(\hat{m}^a_i - m^a_i) \overset{d}{\to} N(0, \text{var}(\phi^a_i)),
\]

(3)

\[2\]In fact it suffices if this sample is of size $O(n)$, i.e. of the same order as the sample on which $\text{P}_n$ operates.
\[ \sqrt{n} \left( \widetilde{\Sigma}_{ij} - \Sigma_{ij} \right) \xrightarrow{d} \left[ -\psi_j^a, -\psi_i^a, 1 \right] \mathcal{N} (0, \text{cov} \left( [\phi_i, \phi_j, \phi_{ij}]^T \right)) , \quad (4) \]

and \( \widehat{m}^a, \widehat{\Sigma}^a \) are efficient, meaning that there exist no other regular asymptotically linear estimators that are asymptotically unbiased and have smaller variance.

The result in Lemma 4.1 is essentially due to the fact that our estimators are built from the efficient influence function, which yields second-order nuisance errors and double-robustness.

Now that we have discussed efficient estimation strategies for the counterfactual mean and covariance, we turn to estimation and inference for the optimal solutions of \( \mathcal{P}_{MV} \). Let \( \widehat{w} \) be an optimal solution of the approximating program of \( \mathcal{P}_{MV} \) in which we replace \( m^a, \Sigma^a \) with their estimates \( \widehat{m}^a, \widehat{\Sigma}^a \), respectively. Then \( \widehat{w} \) is our proposed estimator for the optimal solution to \( \mathcal{P}_{MV} \). Let \( r_n \) be any sequence such that

\[ \| \widehat{\pi}_a - \pi_a \|_{2,p} \left\{ \text{max}_j \| \hat{\mu}_j - \mu_j \|_{2,p} + \text{max}_{i,j} \| \hat{\sigma}_{ij} - \sigma_{ij} \|_{2,p} \right\} = O_P (r_n) . \quad (5) \]

Then the subsequent result follows from Theorem 3.1 and Lemma 4.1.

**Corollary 4.1** (Rates of convergence for \( \widehat{w} \)). Assume that \( \Sigma^a \) is positive definite and let \( w^* \equiv s^* (\mathcal{P}_{MV}) \). Then under Assumptions (B1) (B2) (B3)

\[ \| \widehat{w} - w^* \|_2 = O_P (r_n \lor n^{-1/2}) . \]

If we additionally assume the second-order nonparametric conditions in (B4) then this becomes

\[ \| \widehat{w} - w^* \|_2 = O_P (n^{-1/2}) . \]

Corollary 4.1 gives our first main result in this section: it shows that the proposed estimator is
consistent at rates that are second-order in nuisance estimation errors under mild conditions. This enables us to attain fast $\sqrt{n}$ rates even when we estimate all the nuisance regression functions at much slower rates; for example, it suffices that all the nuisance functions converge to their true values at a faster-than-$n^{1/4}$ rate in $L_2(\mathbb{F})$ norm. This allows us to utilize a wide class of non- or semi-parametric regression tools, depending on structural constraints (Kennedy 2016, Section 4).

As another important consequence of Theorem 3.2 and Lemma 4.1, the next corollary gives sufficient conditions to yield $\sqrt{n}$-consistency and asymptotic normality for $\hat{w}$.

**Corollary 4.2** (Asymptotic distribution for $\hat{w}$). Assume that $\Sigma^a$ is positive definite and Assumptions (B1)-(B4) hold. Further assume that the LICQ and SC conditions hold at $w^*$. Then $\sqrt{n}(\hat{w} - w^*)$ converges in distribution to a zero-mean normal random variable.

An exact closed-form for the variance of the asymptotic normal distribution can be found based on Theorem 3.2 and Lemma 4.1. Corollary 4.2 gives our second main result, and means that asymptotically valid confidence intervals and hypothesis tests can be constructed via the bootstrap. Unlike in the case of $\hat{m}^a$, $\hat{\Sigma}^a$, it is not necessarily clear whether we can conclude that $\hat{w}$ is statistically efficient, due to the inherent complexity of the optimal solution mapping in the presence of constraints. We conjecture that one may show that $\hat{w}$ attains the semiparametric efficiency bound possibly under slightly stronger regularity assumptions, but we leave this for future work.

## 5 Calibration of $\hat{\Sigma}^a$

In the previous section, we developed a semiparametric estimator for the counterfactual covariance matrix, which is the quadratic component of our approximating program of $P_{MV}$. However, unless some special structure is imposed on the dependencies (e.g., diagonality or a factor model),
estimated covariance matrices can be ill-conditioned or non-positive semi-definite, which in our setting may dramatically degrade the quality of our optimal solution estimates. In this section, we discuss two calibration methods for our counterfactual covariance estimator, which help us to avoid these pitfalls, without affecting the rates of convergence for our optimal solution estimator.

5.1 Optimal Linear Shrinkage Estimation

If \( \hat{\Sigma}^a \) is ill-conditioned or (near) singular, then the solution to a system of linear equations involving \( \hat{\Sigma}^a \) is prone to large numerical errors. Consequently, there can be large errors in each iteration of commonly-used algorithms for solving our approximating program, which may cause those algorithms to fail to converge (by making the search direction deviate from a descent direction). This will seriously degrade estimation accuracy of the optimal solutions in finite samples, notwithstanding the favorable asymptotic properties of our proposed estimator.

Covariance shrinkage is a promising approach to handle this issue (e.g., Yang and Berger 1994, Daniels and Kass 1999, 2001, Ledoit and Wolf 2020). The basic idea is to achieve a bias–variance tradeoff by shrinking \( \hat{\Sigma}^a \) towards a certain target (a reference prior). Here, we develop a linear shrinkage estimator for \( \Sigma^a \) by adapting the method by Ledoit and Wolf (2004), one of the most widely-used techniques to regularize the sample covariance matrix. However, the adaptation of their method is not straightforward since the counterfactual outcomes are unobservable, which means that the use of the sample covariance matrix is not feasible in our setting.

Let \( \mathbb{I} \) denote the identity matrix. Our goal is to find the optimal linear combination of \( \mathbb{I} \) and \( \hat{\Sigma}^a \) with minimum expected quadratic loss, which is represented by the solution of the following
program

\[ \text{minimize} \quad \mathbb{P} \left\| \Sigma - \Sigma^a \right\|_F^2 \]
\[ \text{subject to} \quad \Sigma = \rho \nu \mathbb{I} + (1 - \rho) \hat{\Sigma}^a. \]

Let \( \Sigma^*_S := \rho^* \nu^* \mathbb{I} + (1 - \rho^*) \hat{\Sigma}^a \) where \((\rho^*, \nu^*)\) is the optimal solution of (6). \( \Sigma^*_S \) can be regarded as an oracle estimator that reduces the expected error of \( \hat{\Sigma}^a \) in the Frobenius norm (conditional on the nuisance parameter estimates) by shrinking it toward the matrix \( \nu^* \mathbb{I} \). It is an oracle in the sense that the optimal shrinkage parameters \( \rho^* \) and \( \nu^* \) are unknown. In parallel to Ledoit and Wolf (2004), we propose to estimate \( \rho^* \) and \( \nu^* \) by 
\( \hat{\rho} = \frac{\hat{\beta}^2}{\hat{\delta}^2} \) and 
\( \hat{\nu} = \frac{1}{k} \sum_{i=1}^{k} \hat{\Sigma}_{ii}^a \), respectively, where 
\( \hat{\beta}^2 = \left\| \hat{\Sigma}^a - \hat{\nu} \mathbb{I} \right\|_F^2 \) and 
\( \hat{\delta}^2 = \min \{ \hat{\beta}^2, \hat{\delta}^2 \} \) with 
\( \hat{\beta}^2 = \frac{1}{n^2} \sum_{t=1}^{n} \left\| \hat{\Sigma}_t - \hat{\Sigma}^a \right\|_F^2 \), and the \((i, j)\)-entry of the matrix \( \hat{\Sigma}_t \) is defined by 
\( \hat{\phi}_{ij}^a(Z_t) - \mathbb{P}_n(\hat{\phi}_{ij}^a(Z)) \mathbb{P}_n(\hat{\phi}_{ij}^a(Z)) \), \( 1 \leq i, j \leq k, 1 \leq t \leq n \).
Consequently, our proposed estimator for the optimal linear shrinkage is given by

\[ \hat{\Sigma}^*_S = \hat{\rho} \hat{\nu} \mathbb{I} + (1 - \hat{\rho}) \hat{\Sigma}^a. \]

The following theorem establishes the consistency of \( \hat{\Sigma}^*_S \).

**Theorem 5.1.** Assume that (B1) - (B3) hold and that we have an initial estimate \( \hat{\Sigma}^a \) via (2). Further assume that \( \forall i, j \leq k, Y_i^a \)’s have finite fourth moments. Recall that \( r_n \) is the rate given in (5) and let \( \hat{\Sigma}_S \) denote the corresponding estimate for the optimal solution derived by substituting \( \hat{\Sigma}_S \) for \( \hat{\Sigma}^a \). Then,

\[ \left\| \hat{\Sigma}^*_S - \Sigma^*_S \right\|_F = O_P \left( n^{-1/2} \vee r_n \right), \quad \left\| \hat{\Sigma}_S - \Sigma^a \right\|_F = O_P \left( n^{-1/2} \vee r_n \right), \]

Note that \( \forall i, j, \hat{\Sigma}_{ij}^a \) depends on the nuisance estimates \( \hat{\eta}_i, \hat{\eta}_j, \hat{\eta}_{ij} \), each of which is a function of a separate independent sample \( D_0^a \). (See Assumption (B1)) So in our notation, \( \mathbb{P} \left\| \Sigma - \Sigma^a \right\|_F^2 = \mathbb{E} \left[ \left\| \Sigma - \Sigma^a \right\|_F^2 | D_0^a \right] \).
\[ \| \hat{w}_S - w^* \|_2 = O_p \left( n^{-1/2} \lor r_n \right). \]

To the best of our knowledge, although we only consider the case of fixed and finite \( k \), our proposed estimator \( \hat{\Sigma}_S^* \) in (7) is the first attempt to apply the idea of shrinkage estimation to counterfactual covariance matrices. \( \hat{\Sigma}_S^* \) is guaranteed to be non-singular. Using \( \hat{\Sigma}_S^* \) addresses the limitations of \( \hat{\Sigma}^a \) described above, and Theorem 5.1 shows that it does not affect the rates of convergence for our optimal solution estimator \( \hat{w} \).

### 5.2 Positive Definite Correction

Although in Section 4 we assume that \( \Sigma^a \) is positive definite (PD), there is no guarantee that \( \hat{\Sigma}^a \) is PD too. Thus, we may have an approximating program that is not (strictly) convex. This can impose some practical difficulties, because common solution methods may get stuck in local optima, and also it precludes the use of fast QP solvers (e.g. [Stellato et al. 2020]) for large-scale data sets.

PD correction methods which replace \( \hat{\Sigma}^a \) with a nearby PD matrix \( \hat{\Sigma}_{cor}^* \) can be used to address this issue. The following theorem shows that any PD correction method can be used without affecting our previous rate results, as long as \( \hat{\Sigma}_{cor}^* \) and \( \hat{\Sigma}^a \) get arbitrarily close in probability in \( n \) at sufficiently fast rates. For example, \( \hat{\Sigma}_{cor}^* \) could be constructed using the model-free methods of [Huang et al. 2017], with a minimum eigenvalue threshold that decays toward 0 in \( n \).

**Theorem 5.2.** Assume that \( \Sigma^a \) is PD and let \( \hat{\Sigma}_{cor}^* \) denote a symmetric PD matrix indexed by \( \hat{\Sigma}^a \) such that \( \| \hat{\Sigma}_{cor}^* - \hat{\Sigma}^a \|_2 = o_p(1) \), and let \( \hat{w}_{cor} \) be the corresponding estimate for the optimal solution derived by substituting \( \hat{\Sigma}_{cor}^* \) for \( \hat{\Sigma}^a \). Then under Assumptions (B1) - (B3)

\[ \| \hat{w}_{cor} - w^* \|_2 = O_p \left( \| \hat{\Sigma}_{cor}^* - \hat{\Sigma}^a \|_2 \lor r_n \lor n^{-1/2} \right). \]
If $\hat{\Sigma}_{\text{cor}}^{*} = \hat{\Sigma}^{a}$ whenever $\hat{\Sigma}^{a}$ is PD, meaning that $\hat{\Sigma}^{a}$ is only replaced when it is not PD, then the righthand side simplifies to $O_p\left(r_n \vee n^{-1/2}\right)$.

Like the shrinkage method of the previous section, PD corrections that satisfy the conditions of Theorem 5.2 provide an approach to improve the stability of the covariance matrix estimator, especially when (strict or strong) convexity is desired, thereby improving the finite-sample performance of the optimal solution estimator.

6 Empirical Studies

In this section, we illustrate the performance of our proposed estimator in simulated and real-world datasets.

6.1 Kang and Schafer’s Study with Multivariate Outcomes

Without loss of generality, we consider one counterfactual scenario with $A = 1$. To estimate the counterfactual means and covariance matrices, we employ three methods: the PI, IPW, and our proposed SP estimator. Estimator performance is assessed via integrated bias and root-mean-squared error (RMSE), defined by

$$ \text{bias} = \frac{1}{k} \sum_{i=1}^{k} \left| \frac{1}{B} \sum_{j=1}^{B} \hat{w}_i^j - w_i^* \right|, \quad \text{RMSE} = \frac{1}{k} \sum_{i=1}^{k} \left[ \frac{1}{B} \sum_{s=1}^{B} \left\{ \hat{w}_i^s - w_i^* \right\}^2 \right]^{1/2}, $$

across $B = 250$ simulations, where $w^*$ are the optimal weights of the true QP and $\hat{w}^s$ are estimates of $w^*$ at the $s$-th simulation. Our data generation is based on the simulation study by Kang.
and Schafer (2007), modified to accommodate multivariate outcomes as follows:

\[ X = (X_1, X_2, X_3, X_4) \sim N(0, I), \]

\[ \pi_a(X) = \expit(-0.5X_1 + 0.25X_2 - 0.125X_3 - 0.05X_4), \]

\[ (Y_i \mid X, A) \sim N(\mu_i(X, A), V_i), \]

where for \( i = 1, \ldots, k \),

\[ \mu_i(X, A) = b_i + A(d_i + X_\alpha), \]

\[ b_i \sim \text{Unif}(-0.5, 0.5), \quad d_i \sim \text{Unif}(-0.25, 0.25), \]

\[ \alpha = (0.1 + u_{i1}, -0.1 + u_{i2}, 0.2 + u_{i3}, -0.2 + u_{i4})^\top, \]

\[ u_{i1}, u_{i2}, u_{i3}, u_{i4} \sim \text{Unif}(-0.5, 0.5), \]

\[ V_i \sim \text{Unif}(1.5, 3). \]

Here, \( \text{Unif}(l, u) \) denotes the uniform distribution over the interval \([l, u]\). We use sample sizes \( n = 500, 1000, 2500 \). Throughout this section, to estimate the nuisance regression functions, we use the cross-validation super learner ensemble estimator implemented in the SuperLearner R package to combine generalized additive models, multivariate adaptive regression splines, and random forests. We estimate all the nuisance functions on a separate independent set with equal sample size. Also, for simplicity the minimum level of the weighted mean outcome \( r_{\min} \) is set to \(-\infty\).

We consider two versions for each of the three estimators, depending on how each of the nuisance functions are estimated: using the baseline covariates \( X \) or using transformed covariates.
Figure 1: Finite sample performance of the three estimators based on non-transformed covariates $X$, across different sample sizes.

$\widetilde{X}$, based on the same transformations as in Kang and Schafer (2007), i.e.,

$$\widetilde{X} = \left( \exp(X_1/2), X_2/(1 + \exp(X_1)) + 10, (X_1X_3/25 + 0.6)^{2}, (X_2 + X_4 + 20)^2 \right).$$

When the transformed covariates $\widetilde{X}$ are used, estimation of the nuisance functions is more challenging, and at each round of simulation we estimate either the propensity score ($\pi_a$) or outcome regressions ($\mu_i, \sigma_{ij}$) using $\widetilde{X}$ with equal chance. In other words, $\widetilde{X}$ is used to estimate $\pi_a$ for roughly $B/2$ simulations and to estimate ($\mu_i, \sigma_{ij}$) for the remaining simulations.

Our results use the shrinkage estimator $\hat{\Sigma}_S^{*}$ developed in Section 5.1 as it shows a slight improvement in RMSE than the PD correction method. In general, we achieve between 10 and 30 percent relative improvement with the proposed calibration methods (see Appendix A for details). The results are presented in Figures 1, 2.

In Figure 1, the proposed estimator performs as well or slightly better than the PI or IPW estimators. However, in Figure 2, when one of the nuisance estimators is based on $\widetilde{X}$, the proposed estimator gives substantially smaller bias and RMSE in general, and performs better with $n$ than do the other methods. This interesting behavior follows from the results in Section 4 that the proposed estimator has second-order multiplicative bias and thus it is sufficient to require $n^{1/4}$
Figure 2: Finite sample performance of the three estimators with the transformed covariates. At each simulation, either the propensity score ($\pi$) or outcome regression models ($\mu, \sigma$) is estimated using the transformed covariates $\tilde{X}$, not $X$.

Rates on nuisance estimation in order for this estimator to attain $\sqrt{n}$ rates, while the PI and IPW directly inherit the slower-than-$\sqrt{n}$ rates at which the nuisance parameters are estimated and are expected to be converge particularly slowly when $\tilde{X}$ is used. This behavior appears to hold regardless of the value of the decision variable $k$, although we have slightly larger bias and RMSE for $k = 10$ than $k = 3$.

### 6.2 Optimal Medical Appointment Scheduling

Medical providers have finite time to provide care for large populations of patients. In order to accommodate patients’ scheduling needs and their own staffing needs, providers must choose how many appointment slots to reserve for fixed appointments, which are scheduled in advance, vs. open-access appointments, which are scheduled on short notice, often the same day that patients request them. Providers naturally wish to maximize the daily utilization rate, i.e. the proportion of slots each day in which patients are actually seen, while minimizing variance in this rate across days. The utilization rate depends in part on the patient no-show rate, which can be quite high for fixed appointments (Qu et al., 2012).
Qu et al. (2012) used mean-variance optimization to identify optimal proportions of fixed vs. open-access appointments across different types of providers. Their study assumed that all relevant distributional parameters were known and did not need to be estimated from data. Based on their approach, we illustrate how our counterfactual framework can be used for reliable decision support in healthcare when interventions exist that can change the distribution of outcomes.

We generated a simulated dataset of 10,000 observations describing patient appointments, appointment types, an intervention to improve patient attendance, and utilization rates as follows:

\[
X \sim \text{Unif}(-1, 1)
\]

\[
\pi_1(X) = \expit(0.6 + 0.1 \times X^3)
\]

\[
Y_o \mid A, X \sim \text{Beta}(1 + A/5, X^2)
\]

\[
Y_f \mid A, X \sim \text{Beta}(0.1 + A/2, X^2)
\]

with \(Y_o \perp Y_f \mid A, X\). Here, the intervention \(A \in \{0, 1\}\) represents two types of appointment reminders. These patient prompts have been shown to reduce no-show rates for fixed appointments, with calls from staff \((A = 1)\) leading to greater improvements than automated reminders \((A = 0)\) (Parikh et al. 2010). Suppose medical providers are interested in optimal proportions of the two appointment types under policies that set \(A = 0\) or \(A = 1\) across all appointments, and have data in which \(A\) varies; for example, they may have been running a trial of a new automated calling system \((A = 1)\), or they may have previously made reminder phone calls only occasionally and are considering hiring staff to make phone calls in advance of every appointment. \(Y_o\) and \(Y_f\) represent the observed utilization rates, the proportions of daily open-access and fixed appointments, respectively, in which providers see patients. For simplicity, we assume that the utilization rates do not depend on the number of appointments of each type offered. \(X\) represents
Table 1: Means (variances) in utilization rates for open-access ($Y_o$) vs fixed ($Y_f$) appointments.

| $A$ | $Y_o$   | $Y_f$   |
|-----|---------|---------|
| 0   | 0.79 (0.08) | 0.68 (0.14) |
| 1   | 0.81 (0.07) | 0.79 (0.08) |

a synthesis of variables that affect $A$ and/or $Y_o$ and $Y_f$, such as weather conditions (which affect both staff availability and patient no-show rates) or the severity of patient comorbidities.

Since open-access appointments are made on very short notice, it is likely that the benefit of personal reminders relative to automated ones is greater for fixed appointments than for open-access appointments. Our data generating process reflects this view, with Table 1 showing that personal calls increase utilization rates substantially relative to automated ones for fixed appointments, but have only a small effect for open-access appointments.

Figure 3 shows the estimated optimal proportion of open appointments under the two counterfactual conditions and the observable condition, across a range of values of the risk tolerance parameter $\lambda$. Because utilization is higher on average for open-access appointments, the optimal proportion increases as risk tolerance increases. Because personal calls ($A = 1$) increase utilization for fixed appointments compared to automated ones ($A = 0$), the optimal proportion of open-access appointments is smaller if personal calls were to be implemented globally than if automated calls were to be implemented globally, or if the provider were to continue with the current practice of automated reminder calls for some appointments or personally made for others.

Figure 4 shows the estimated counterfactual Pareto-efficient frontiers for the three conditions, with each curve spanning $\lambda \in [0, 2]$. The leftmost point on each curve represents $\lambda = 0$, i.e., the composition of appointments that minimizes variance without regard to the mean. The $A = 1$ curve dominates the others, meaning that for any given mean utilization rate (variance), the lowest variance (highest mean utilization rate) is achieved by the personal reminder policy. This suggests
Figure 3: Estimated optimal proportion of open appointments under the two counterfactual poli-
cies and the observable policy ("Obs"), with 95% bootstrap confidence intervals, across different
risk tolerances.

Figure 4: Pareto-efficient frontiers under the two counterfactual policies and the observable pol-
icy. Each point on a each curve represents an estimated optimum allocation of fixed vs. open-
access appointments for a given value of the risk tolerance parameter $\lambda$.

that the provider may achieve the best mean-variance tradeoff by choosing the personal reminder
policy ($A \equiv 1$).

6.3 Counterfactual portfolio modeling

We next illustrate our method in the context of financial portfolio modeling. We consider monthly
returns from six Vanguard index funds representing different asset classes, the same funds used
in [Kim et al. (2021)]. We use daily adjusted closing prices collected from Yahoo Finance.

The intervention or treatment we consider is the federal funds effective rate, which is the
average rate at which banks lend each other money overnight. This rate is influenced by the
federal funds target rate, which is set by the Federal Reserve (Fed). For each month, we let $A = 1$ if the effective rate increased with respect to the previous month, and $A = 0$ otherwise. For example, if the effective rate for October was 3.0 and the effective rate for November was 3.25, then we would have $A = 1$ for November. As covariates, we include the Consumer Price Index (CPI) and the unemployment rate, which correspond to the Fed’s dual mandate to promote maximum employment and price stability (Federal Reserve System Publication[2021]). We also include the five factors from the Fama and French asset pricing model, which aim to explain long-term expected portfolio returns (Fama and French[2015]).

Our data span 2011-2020 (120 months), of which 58 months involve rate increases ($A = 1$). Figure[5] shows the estimated optimal weights for a range of values of the risk tolerance parameter $\lambda$, under the two counterfactual scenarios as well as the traditional observable setting. (We use the word “scenarios” to emphasize the fact that investors have no control over Fed policy.) As expected, for small values of $\lambda$, the portfolios all tilt heavily toward the U.S. total bond market (VBTLX), which has the lowest return and lowest volatility among the asset classes. As $\lambda$ increases, the portfolios tilt instead toward assets with higher return and higher volatility. In the $A = 1$ scenario, the portfolio tilts toward corporate bonds (VWEAX), whereas in the other two scenarios, the portfolio tilts toward large cap stocks (VFIAX), suggesting that the volatility-return tradeoffs differ in a (counterfactual) environment in which rates are rising versus an environment in which they are steady or falling.

The weights in the counterfactual scenarios represent the optimal portfolios were the mean and variance to be computed from the counterfactual return distributions. These are different from weight estimates using a subset of observed data in which $A$ is equal to 0 or 1 (denoted by “$|A = 0$” and “$|A = 1$” in Figure[5]); the latter may be affected by confounding, whereas in the former we adjust for potential confounders in order to isolate effects solely caused by the Fed’s
Figure 5: Estimated optimal asset weights in the observational and counterfactual scenarios.

decision. These counterfactual portfolios, which have never been studied in the literature, may be of intrinsic scientific interest. They can provide additional insight into portfolio robustness, by illuminating the sensitivity of the portfolio weights to surprise rate hikes or cuts. For example, consider an asset manager who believes that the current Fed is more aggressive toward inflation than previous Feds. Then the observational weights may be based on an underestimation of the Fed’s likelihood of raising rates, and the asset manager may wish to tilt their portfolio toward the weights in the $A = 1$ scenario. We present this as a heuristic argument for now, and leave a more thorough analysis of the uses of our framework in practice for future work.

7 Discussion

In this paper, we proposed counterfactual mean-variance optimization, a new framework for optimal decision making under hypothetical interventions. First, we characterized rates of convergence and provided a closed-form expression for the asymptotic distribution of optimal solutions in a general class of nonlinear stochastic optimization problems. Based on these results, we developed a doubly robust estimator that attains $\sqrt{n}$-consistency and asymptotic normality even
when using flexible nonparametric regression methods. We also discussed calibration methods for our counterfactual covariance matrix estimator, which help to avoid ill-conditioning or non-positive (semi-)definiteness. Finally, we studied our methods via simulation, and applied them to problems in healthcare operations research and financial portfolio construction.

Some of our theoretical developments can be harnessed outside the mean-variance framework. For example, our asymptotic results in Section 3 can be applied to doubly-robust counterfactual prediction via the following constrained and regularized estimation

$$\begin{align*}
\minimize_{\beta \in \mathbb{R}^k} & \ R \left( \beta^T b(X), Y^a \right) + \lambda_{\text{penalty}} J(\beta) \\
\text{subject to} & \ \beta \in S_{\text{lin}}(Y^a),
\end{align*}$$

for some risk $\mathcal{R}$, penalty $J$, finite basis set $b(\cdot)$, and tuning parameter $\lambda_{\text{penalty}} \geq 0$. Here, we are estimating a projection of $\mathbb{E}[Y^a | X]$ onto a finite-dimensional model with the constraint set $S_{\text{lin}}$ which allows us to use flexible penalization, or to incorporate prior information (Gaines et al. 2018) or (counterfactual) fairness constraints (Mishler and Kennedy 2021). Since we focused on stochastic programs of the form $(P_{\text{nl}})$, only specific types of risk can be used for $\mathcal{R}$ (e.g., the mean squared error). However, in upcoming companion papers, we consider a more general counterfactual prediction problem with a far broader class of risk functions. Furthermore, our proposed shrinkage method in Section 5.1 provides an opportunity to expand a wide range of applications of covariance matrix estimation (e.g., Ledoit and Wolf 2020, Section 2.4) to counterfactual settings.

There are a number of other potential opportunities for future work as well. First, our methods may be extended to other settings such as optimal policy estimation under constraints, such as resource or interpretability constraints, both of which frequently arise in medicine, the social sciences, and elsewhere. Furthermore, the objective function in $P_{\text{MV}}$ can be extended to include
other more complex reward and risk measures beyond the mean and variance. For example, to account for asymmetrical data distributions, one may use Value-at-Risk by replacing $w^\top \Sigma w$ with $Q_\alpha(w^\top m^a)$, where $Q_\alpha(\cdot)$ is the $\alpha$-quantile. Lastly, while we required linear constraints in this work, it is also of interest to extend our methods to general nonlinear constraints, and to see if they yield any new insights about counterfactual resource allocations.

**References**

Athey, S. and Wager, S. (2021), “Policy learning with observational data,” *Econometrica*, 89, 133–161.

Bang, H. and Robins, J. M. (2005), “Doubly robust estimation in missing data and causal inference models,” *Biometrics*, 61, 962–973.

Bonnans, J. F. and Ioffe, A. D. (1995), “Quadratic growth and stability in convex programming problems with multiple solutions,” *J. Convex Anal*, 2, 41–57.

Cardozo, R. N. and Smith, Jr., D. K. (1983), “Applying Financial Portfolio Theory to Product Portfolio Decisions: An Empirical Study,” *Journal of Marketing*, 47, 11.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017), “Double/Debiased/Neyman Machine Learning of Treatment Effects,” *American Economic Review*, 107, 261–65.

Coston, A., Kennedy, E., and Chouldechova, A. (2020), “Counterfactual Predictions under Runtime Confounding,” in *Advances in neural information processing systems*, vol. 33, pp. 4150–4162.
Daniels, M. J. and Kass, R. E. (1999), “Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models,” *Journal of the American Statistical Association*, 94, 1254–1263.

— (2001), “Shrinkage estimators for covariance matrices,” *Biometrics*, 57, 1173–1184.

Delarue, E., De Jonghe, C., Belmans, R., and D’haeseleer, W. (2011), “Applying portfolio theory to the electricity sector: Energy versus power,” *Energy Economics*, 33, 12–23.

Dickerman, B. A. and Hernán, M. A. (2020), “Counterfactual prediction is not only for causal inference,” *European Journal of Epidemiology*, 35, 615–617.

Dontchev, A. L. and Rockafellar, R. T. (2009), *Implicit functions and solution mappings*, vol. 543, Springer.

Fagefors, C. and Lantz, B. (2021), “Application of Portfolio Theory to Healthcare Capacity Management,” *International Journal of Environmental Research and Public Health*, 18, 659.

Faigle, U., Kern, W., and Still, G. (2013), *Algorithmic principles of mathematical programming*, vol. 24, Springer Science & Business Media.

Fama, E. F. and French, K. R. (2015), “A five-factor asset pricing model,” *Journal of Financial Economics*, 116, 1–22.

Fang, Z. and Santos, A. (2019), “Inference on directionally differentiable functions,” *The Review of Economic Studies*, 86, 377–412.

Federal Reserve System Publication (2021), “The Fed Explained: What the Central Bank Does,”

Gaines, B. R., Kim, J., and Zhou, H. (2018), “Algorithms for fitting the constrained lasso,” *Journal of Computational and Graphical Statistics*, 27, 861–871.
Hernán, M. A. and Robins, J. M. (2010), “Causal inference,”.

Höfler, M. (2005), “Causal inference based on counterfactuals,” *BMC medical research methodology*, 5, 1–12.

Holland, P. W. (1986), “Statistics and causal inference,” *Journal of the American statistical Association*, 81, 945–960.

Huang, C., Farewell, D., and Pan, J. (2017), “A calibration method for non-positive definite covariance matrix in multivariate data analysis,” *Journal of Multivariate Analysis*, 157, 45–52.

Imbens, G. W. and Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Kang, J. D. and Schafer, J. L. (2007), “Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data,” *Statistical science*, 523–539.

Kennedy, E. H. (2016), “Semiparametric theory and empirical processes in causal inference,” in *Statistical causal inferences and their applications in public health research*, Springer, pp. 141–167.

— (2020), “Optimal doubly robust estimation of heterogeneous causal effects,” *arXiv preprint arXiv:2004.14497*.

Kennedy, E. H., Balakrishnan, S., and Wasserman, L. (2021), “Semiparametric counterfactual density estimation,” *arXiv preprint arXiv:2102.12034*.

Kim, J. H., Lee, Y., Kim, W. C., and Fabozzi, F. J. (2021), “Mean–Variance Optimization for Asset Allocation,” *The Journal of Portfolio Management*, 48, 17.
Kim, K., Kim, J., and Kennedy, E. H. (2018), “Causal effects based on distributional distances,” *arXiv preprint arXiv:1806.02935*.

King, A. J. and Rockafellar, R. T. (1993), “Asymptotic theory for solutions in statistical estimation and stochastic programming,” *Mathematics of Operations Research*, 18, 148–162.

Kitagawa, T. and Tetenov, A. (2018), “Who should be treated? empirical welfare maximization methods for treatment choice,” *Econometrica*, 86, 591–616.

Ledoit, O. and Wolf, M. (2004), “A well-conditioned estimator for large-dimensional covariance matrices,” *Journal of multivariate analysis*, 88, 365–411.

— (2020), “The power of (non-) linear shrinking: A review and guide to covariance matrix estimation,” *Journal of Financial Econometrics*.

Lin, L., Sperrin, M., Jenkins, D. A., Martin, G. P., and Peek, N. (2021), “A scoping review of causal methods enabling predictions under hypothetical interventions,” *Diagnostic and prognostic research*, 5, 1–16.

Luedtke, A. R. and Van Der Laan, M. J. (2015), “Optimal Dynamic Treatments in Resource-Limited Settings,” .

Luedtke, A. R. and van der Laan, M. J. (2016), “Optimal individualized treatments in resource-limited settings,” *The international journal of biostatistics*, 12, 283–303.

Markowitz, H. M. (1968), *Portfolio selection*, Yale university press.

Meyer, J. (1987), “Two-moment decision models and expected utility maximization,” *The American Economic Review*, 421–430.

Mishler, A. and Kennedy, E. (2021), “FADE: FAir Double Ensemble Learning for Observable and Counterfactual Outcomes,” *arXiv preprint arXiv:2109.00173*.
Mishler, A., Kennedy, E. H., and Chouldechova, A. (2021), “Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 386–400.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009), “Robust stochastic approximation approach to stochastic programming,” *SIAM Journal on Optimization*, 19, 1574–1609.

Neugebauer, R. and van der Laan, M. (2007), “Nonparametric causal effects based on marginal structural models,” *Journal of Statistical Planning and Inference*, 137, 419–434.

Newey, W. K. and Robins, J. R. (2018), “Cross-fitting and fast remainder rates for semiparametric estimation,” *arXiv preprint arXiv:1801.09138*.

Parikh, A., Gupta, K., Wilson, A. C., Fields, K., Cosgrove, N. M., and Kostis, J. B. (2010), “The Effectiveness of Outpatient Appointment Reminder Systems in Reducing No-Show Rates,” *The American Journal of Medicine*, 123, 542–548.

Qu, X., Rardin, R. L., and Williams, J. A. S. (2012), “A mean–variance model to optimize the fixed versus open appointment percentages in open access scheduling systems,” *Decision Support Systems*, 53, 554–564.

Robins, J. M. and Rotnitzky, A. (1995), “Semiparametric efficiency in multivariate regression models with missing data,” *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M. and Wang, N. (2000), “Inference for imputation estimators,” *Biometrika*, 87, 113–124.

Rubin, D. B. (1974), “Estimating causal effects of treatments in randomized and nonrandomized studies.” *Journal of Educational Psychology*, 66, 688.
Schulam, P. and Saria, S. (2017), “Reliable decision support using counterfactual models,” *Advances in Neural Information Processing Systems*, 30, 1697–1708.

Semenova, V. and Chernozhukov, V. (2021), “Debiased machine learning of conditional average treatment effects and other causal functions,” *The Econometrics Journal*, 24, 264–289.

Shapiro, A. (1991), “Asymptotic analysis of stochastic programs,” *Annals of Operations Research*, 30, 169–186.

— (1993), “Asymptotic behavior of optimal solutions in stochastic programming,” *Mathematics of Operations Research*, 18, 829–845.

— (2000), “Statistical inference of stochastic optimization problems,” in *Probabilistic constrained optimization*, Springer, pp. 282–307.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014), *Lectures on stochastic programming: modeling and theory*, SIAM.

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. (2020), “OSQP: an operator splitting solver for quadratic programs,” *Mathematical Programming Computation*, 12, 637–672.

Still, G. (2018), “Lectures on parametric optimization: An introduction,” *Optimization Online*.

Wachsmuth, G. (2013), “On LICQ and the uniqueness of Lagrange multipliers,” *Operations Research Letters*, 41, 78–80.

Yang, R. and Berger, J. O. (1994), “Estimation of a covariance matrix using the reference prior,” *The Annals of Statistics*, 1195–1211.

Zheng, W. and Van Der Laan, M. J. (2010), “Asymptotic theory for cross-valiyeard targeted maximum likelihood estimation,” *Working Paper 273*. 

36
Disclaimer  This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.
A Relative Improvement in RMSE through Covariance Calibration

We give additional simulation results showing that our proposed calibration methods improve upon the optimal solution estimator based on $\hat{\Sigma}^a$ using the same simulation setup as Section 6.1. For each calibration method $\hat{\Sigma}^* \in \{\hat{\Sigma}_S^*, \hat{\Sigma}_cor^*\}$, we compute the percentage relative improvement in RMSE using the following formula:

$$\frac{\text{RMSE}(\hat{\Sigma}^a) - \text{RMSE}(\hat{\Sigma}^*)}{\text{RMSE}(\hat{\Sigma}^*)},$$

where the RMSE of the optimal solution estimator $\hat{w}$ is computed in the same way as Section 6.1 with $\hat{\Sigma}^a$ or its calibrated version $\hat{\Sigma}^*$. Again, we construct all the nuisance estimators on the independent, separate sample with the same size. For $\hat{\Sigma}_cor^*$ we compute the the nearest PD matrix by using R function nearPD, which implements the algorithm of Higham (2002), and then forces positive definiteness if needed.

The results are presented in Figures 6 and 7. Both calibration methods appear to significantly improve upon the original optimal solution estimator that is computed without covariance calibration, although the improvement becomes less substantial as sample size increases. Without calibration, we have the increased number of suboptimal solutions (i.e., a solution that has not converged). Amongst the three estimators, the semiparametric estimator achieves the largest improvement. When the transformed covariates are used, the relative improvement in RMSE has been mostly wiped out for the PI and IPW estimators, but still the substantial improvement is
Figure 6: Relative improvement in RMSE of the three estimators through the covariance shrinkage, based on the non-transformed covariates $X$ (left) and the transformed covariates (right).

Figure 7: Relative improvement in RMSE of the three estimators through the PD correction, based on the non-transformed covariates $X$ (left) and the transformed covariates (right).

observed for the semiparametric estimator. In general, larger improvements are observed for the shrinkage estimator than the PD correction with the given simulation setup.

B Proofs

Extra notation. First, we introduce some extra notation used throughout in the proofs. We let $\langle M_1, M_2 \rangle := \text{tr} \left( M_1^T M_2 \right) / k$ for $k \times k$ matrices $M_1, M_2$ (so $\|M_1\|_F = \sqrt{\langle M_1, M_1 \rangle}$). We let $\mathbb{B}_\delta(\bar{z})$ denote the open ball with radius $\delta > 0$ around the point $\bar{z}$ with $\| \cdot \|_2$ (unless otherwise mentioned), i.e., $\mathbb{B}_\delta(\bar{z}) = \{ z \mid \| z - \bar{z} \|_2 < \delta \}$. We use $C^r(S)$ to denote a set of functions that
are \( r \) times continuously differentiable on \( S \). Further, we let \( o_{x \to \alpha}(1) \) denote some function of \( x \) that converges to 0 when \( x \to \alpha \), i.e., a function \( h_{\alpha}(x) \) such that \( \lim_{x \to \alpha} \|h_{\alpha}(x)\| = 0 \). This is to distinguish from typical little-o \((o)\) asymptotic notation with respect to \( n \) in our proofs.

### B.1 Proof of Theorem 3.1

First, we discuss the notion of Lipschitz stability of local minimizers in general nonlinear parametric optimization. Let \( \Xi \subset \mathbb{R}^q \) be some finite-dimensional open parameter set. For \( \theta \in \Xi \), consider a parametric program

\[
\begin{align*}
\text{minimize} & \quad f(x, \theta) \\
\text{subject to} & \quad x \in S_{g}(\theta) = \{x \mid g_j(x, \theta) \leq 0, j \in J\}.
\end{align*}
\]

where \( f, g_j \in C^2(\mathbb{R}^k \times \Xi) \). Let \( x^*(\theta) \in s^*(P(\theta)) \), a local minimizer of \( P(\theta) \) with the corresponding parameter \( \theta \). Next, we define the Lipschitz stability property of local minimizers.

**Definition B.1** (Lipschitz stability of local minimizer). \( x_0 \equiv x^*(\theta_0) \) is called Lipschitz stable if there exist some constants \( L, \epsilon > 0 \) and a local minimizer \( x(\theta) \) of \( P(\theta) \) such that

\[
\|x(\theta) - x_0\| \leq L\|\theta_0 - \theta\| \quad \forall \theta \in B_{\delta}(\theta_0).
\]

See, for example, Still (2018), Chapter 6 for more details for Lipschitz and other types of stability result of local minimizers in smooth \((C^2)\) nonlinear parametric optimization.

Assume that the parameter \( \theta \) unknown and is to be estimated from data with an estimator \( \hat{\theta} \).

The next lemma show that if the Lipschitz stability result holds for each local minimizer of \( P(\theta) \) then the estimation error of the solutions is bounded by that of the parameters.

**Lemma B.1.** Suppose a local minimizer \( x^* \equiv x^*(\theta) \) of \( P(\theta) \) is Lipschitz stable. Provided that \( \hat{\theta} \)
converges in probability to $\theta$, we have

$$\text{dist}(x^*, s^*(P(\hat{\theta}))) = O_P \left( \|\hat{\theta} - \theta\| \right).$$

**Proof.** By definition of Lipschitz stability, there exist constants $L, \epsilon > 0$ and a local solution $x(\hat{\theta})$ of $[P(\theta)]$ such that

$$\|x(\hat{\theta}) - x^*\| \leq L\|\theta - \hat{\theta}\|, \forall \hat{\theta} \in B_\delta(\theta)$$

Given that $\hat{\theta} \overset{p}{\rightarrow} \theta$ we have

$$\|\hat{x} - x^*\| = \|\hat{x} - x^*\| 1 \left\{ \hat{\theta} \in B_\delta(\theta) \right\} + \|\hat{x} - x^*\| 1 \left\{ \hat{\theta} \notin B_\delta(\theta) \right\}$$

$$= \|\hat{x} - x^*\| + \left( 1 \left\{ \hat{\theta} \in B_\delta(\theta) \right\} - 1 \right) \|\hat{x} - x^*\| + \|\hat{x} - x^*\| 1 \left\{ \hat{\theta} \notin B_\delta(\theta) \right\}$$

$$= O \left( \|\theta - \hat{\theta}\| \right) + o_P \left( \|\hat{x} - x^*\| \right),$$

where the last equality follows by $1 \left\{ \hat{\theta} \notin B_\delta(\theta) \right\} = o_P(1)$ and $1 \left\{ \hat{\theta} \in B_\delta(\theta) \right\} - 1 = o_P(1)$. Hence we have

$$\frac{\|\hat{x} - x^*\|}{\|\hat{\theta} - \theta\|} = O_P(1),$$

and the result follows from $\text{dist}(x^*, s^*(P(\hat{\theta}))) \leq \|\hat{x} - x^*\|$. 

Then the proof of Theorem 3.1 immediately follows by the following argument.

**Proof.** Note that our problem can be viewed as a special case of the parametric program $[P(\theta)]$. By virtue of linearity of our constraints, the strong second order condition (Still 2018, Definition 6.1) is satisfied for each $x^* \in s^*(P_{nl})$ under Assumption (A2). Hence, by Theorem 6.4 of Still (2018), $x^*$ is Lipschitz stable. The rest of the proof is straightforward by Lemma B.1.
B.2 Proof of Theorem 3.2

Proof. Let \( \gamma \in \mathbb{R}^{rc} \) be the multipliers (the dual variables) associated with the following constrained program

\[
\begin{align*}
\text{minimize} \quad & f(x) \\
\text{subject to} \quad & Cx \leq d.
\end{align*}
\]

(\( P_c \))

Then consider the following unconstrained program

\[
\begin{align*}
\text{minimize} \quad & f(x) + \sum_{j=1}^{rc} \gamma_j C_j^T x \\
\end{align*}
\]

(\( P_u \))

We shall show that solving \( P_u \) are equivalent to solving \( P_c \). Since \( P_c \) is a convex program only with linear constraints, Slater’s condition is satisfied and thereby strong duality holds. Thus, any optimal solution to \( P_c \) minimizes \( f(x) + \sum_{j=1}^{rc} \gamma_j \{ C_j^T x - d_j \} \), and consequently being an optimal solution to \( P_u \). Further, since \( f \) is strictly convex on \( C \) both \( P_c \) and \( P_u \) have an unique solution. Therefore, the unique solution to \( P_c \) is also the unique solution to \( P_u \), and the perfect equivalence holds.

Let \( x^* \) denote the optimal solution to \( P_c \) (and also to \( P_u \)). Further, let

\[
\begin{align*}
\text{minimize} \quad & \hat{f}(x) \\
\text{subject to} \quad & \hat{C}x \leq \hat{d}
\end{align*}
\]

(\( \hat{P}_c \))

be an approximating program for \( P_c \). Since \( \hat{P}_c \) and \( \hat{P}_e \) have feasible sets that consist of only linear inequalities, the KKT optimality conditions hold for the both (Faigle et al. 2013, Theorem 5.4). Let \( \hat{\gamma} \) denote the multipliers associated with the program \( \hat{P}_c \). Then from the complementary
slackness, for each \( j \in \{1, \ldots, r_C \} \), we have \( \gamma_j(C_jx^* - d_j) = 0 \) and \( \tilde{\gamma}_j(C_j\tilde{x} - \tilde{d}_j) = 0 \), and thus

\[
\gamma_j(C_jx^* - d_j) - \tilde{\gamma}_j(C_j\tilde{x} - \tilde{d}_j) = (\gamma_j - \tilde{\gamma}_j)(C_jx^* - d_j) + \tilde{\gamma}_j \left\{ (C_jx^* - d_j) - (\tilde{C}_j\tilde{x} - \tilde{d}_j) \right\}
\]

\[
= (\gamma_j - \tilde{\gamma}_j)(C_jx^* - d_j) + \tilde{\gamma}_j C_j(x^* - \tilde{x}) + \tilde{\gamma}_j \left\{ (C_j - \tilde{C}_j)\tilde{x} + \tilde{d}_j - d_j \right\}
\]

\[
= 0.
\]

Let be \( J_0(x^*) \) the active index set for \( P_c \). Then, from the above

\[
j \notin J_0(x^*) \Rightarrow \tilde{\gamma}_j - \gamma_j = O \left( \|\tilde{x} - x^*\|_2 + \|\tilde{C}_j - C_j\|_2 + |\tilde{d}_j - d_j| \right).
\]

On the other hand, by the stationarity conditions (or the dual conditions) from \( P_c \) and \( \hat{P}_c \), again by adding and subtracting terms, it follows that From the stationary conditions, by adding and subtracting terms it follows

\[
0 = \nabla_x \hat{f}(\tilde{x}) - \nabla_x f(x^*) + \sum_{j \in J_0(x^*)} \left\{ \tilde{\gamma}_j \tilde{C}_j^T - \gamma_j C_j^T \right\}
\]

\[
= \nabla_x \hat{f}(\tilde{x}) - \nabla_x \hat{f}(x^*) + \nabla_x \hat{f}(x^*) - \nabla_x f(x^*) + \sum_{j \in J_0(x^*)} \left\{ (\tilde{\gamma}_j - \gamma_j)C_j^T + \tilde{\gamma}_j (\tilde{C}_j^T - C_j^T) \right\}
\]

\[
= (\tilde{x} - x^*)^T \nabla_x^2 \hat{f}(x^*) + (\tilde{x} - x^*) o_{\tilde{x} \rightarrow x^*}(1) + \nabla_x \hat{f}(x^*) - \nabla_x f(x^*)
\]

\[
+ \sum_{j \in J_0(x^*)} \left\{ (\tilde{\gamma}_j - \gamma_j)C_j^T + \tilde{\gamma}_j (\tilde{C}_j^T - C_j^T) \right\},
\]

where the last equality follows by Taylor’s theorem. Since we have assumed non-zero rows in \( C \), by the Cauchy–Schwarz and triangle inequalities we obtain

\[
j \in J_0(x^*) \Rightarrow |\tilde{\gamma}_j - \gamma_j| = O \left( \|\tilde{x} - x^*\|_2 + \|\nabla_x \hat{f}(x^*) - \nabla_x f(x^*)\|_2 + \|\tilde{C}_j - C_j\|_2 \right).
\]
Putting things together,

\[ \| \hat{\gamma} - \gamma \|_2 = O \left( \| \hat{x} - x^* \|_2 + \| \nabla_x \hat{f}(x^*) - \nabla_x f(x^*) \|_2 + \sum_{j=1}^{r_C} \| \hat{C}_j - C_j \|_2 + \sum_{j=1}^{r_C} | \hat{d}_j - d_j | \right) \]

\[ = O \left( \| \hat{x} - x^* \|_2 + \| \nabla_x \hat{f}(x^*) - \nabla_x f(x^*) \|_2 + \| \hat{C} - C \|_F + \| \hat{d} - d \|_2 \right). \] \hspace{1cm} (8)

Next, let

\[ \text{minimize}_{x \in \mathbb{C}} \quad \hat{f}(x) + \sum_{j=1}^{r_C} \hat{\gamma}_j \hat{C}_j^\top x \]

be an approximating program of the unconstrained program \( P_u \). Due to the positive definiteness of \( \nabla^2_x f \), by Still (2018, Theorem 2.4) the quadratic growth condition holds at \( x^* \), the optimal solution of the program \( P_u \). Then by Shapiro (1991, Lemma 4.1) and the mean value theorem, there exists a neighborhood \( \mathbb{B}_\delta(x^*) \) for some \( \delta > 0 \) such that

\[ \| \hat{x} - x^* \|_2 \lesssim \left( \sup_{x' \in \mathbb{C} \cup \mathbb{B}_\delta(x^*)} \| \nabla_x \hat{f}(x') - \nabla_x f(x') \|_2 + \sum_{j=1}^{r_C} \{ \hat{\gamma}_j \hat{C}_j - \gamma_j C_j \} \right) \]

\[ \lesssim \left( \sup_{x' \in \mathbb{C} \cup \mathbb{B}_\delta(x^*)} \| \nabla_x \hat{f}(x') - \nabla_x f(x') \|_2 + \sum_{j=1}^{r_C} \| \hat{C}_j - C_j \|_2 + \sum_{j=1}^{r_C} | \hat{\gamma}_j - \gamma_j | \right) \]

\[ \lesssim \left( \sup_{x' \in \mathbb{C} \cup \mathbb{B}_\delta(x^*)} \| \nabla_x \hat{f}(x') - \nabla_x f(x') \|_2 + \| \hat{C} - C \|_F + \| \hat{\gamma} - \gamma \|_2 \right). \] \hspace{1cm} (9)

Now, from (8) and (9) we finally obtain

\[ \| \hat{x} - x^* \|_2 = O \left( \sup_{x' \in \mathbb{C} \cup \mathbb{B}_\delta(x^*)} \| \nabla_x \hat{f}(x') - \nabla_x f(x') \|_2 + \| \hat{C} - C \|_F + \| \hat{d} - d \|_2 \right). \]
Proof. Let \( \gamma \) and \( \hat{\gamma} \) be the multipliers for \( P_{nl} \) and \( \hat{P}_{nl} \) respectively. First we aim to show that

\[
\|\hat{\gamma} - \gamma\|_2 = O_P\left(\|\hat{x} - x^*\|_2 + n^{-1/2}\right).
\]  

(10)

Since \( P_{nl} \) and \( \hat{P}_{nl} \) have feasible sets that consist of only linear inequalities, the KKT optimality conditions hold for the both \cite[Theorem 5.4]{Faigle2013}. Then from the complementary slackness, for each \( j \in \{1, \ldots, r_C\} \), we have

\[
\gamma_j(C_jx^* - d_j) = 0 \quad \text{and} \quad \hat{\gamma}_j(C_j\hat{x} - \hat{d}_j) = 0,
\]

and thus

\[
\gamma_j(C_jx^* - d_j) - \hat{\gamma}_j(C_j\hat{x} - \hat{d}_j) = (\gamma_j - \hat{\gamma}_j)(C_jx^* - d_j) + \hat{\gamma}_j\left\{ (C_jx^* - d_j) - (\hat{C}_j\hat{x} - \hat{d}_j) \right\}
\]

\[
= (\gamma_j - \hat{\gamma}_j)(C_jx^* - d_j) + \hat{\gamma}_jC_j(x^* - \hat{x}) + \hat{\gamma}_j\left\{ (C_j - \hat{C}_j)\hat{x} + \hat{d}_j - d_j \right\}
\]

\[
= 0.
\]

Let be \( J_0(x^*) \) the active index set for \( \hat{P}_{nl} \). Then, from the above

\[
j \notin J_0(x^*) \implies \hat{\gamma}_j - \gamma_j = O\left(\|\hat{x} - x^*\|_2 + \|\hat{C}_j - C_j\|_2 + |\hat{d}_j - d_j|\right).
\]

On the other hand, by the stationarity conditions (or the dual conditions) from \( P_{nl} \) and \( \hat{P}_{nl} \), again by adding and subtracting terms, it follows that From the stationary conditions, by adding and subtracting terms it follows

\[
0 = \nabla_x \hat{f}(\hat{x}) - \nabla_x f(x^*) + \sum_{j \in J_0(x^*)} \left\{ \hat{\gamma}_j \hat{C}_j^T - \gamma_j C_j^T \right\}
\]

\[
= \nabla_x \hat{f}(\hat{x}) - \nabla_x \hat{f}(x^*) + \nabla_x \hat{f}(x^*) - \nabla_x f(x^*) + \sum_{j \in J_0(x^*)} \left\{ (\hat{\gamma}_j - \gamma_j)C_j^T + \hat{\gamma}_j(\hat{C}_j^T - C_j^T) \right\}
\]

\[
= (\hat{x} - x^*)^T \nabla^2_x \hat{f}(x^*) + (\hat{x} - x^*)o_{\hat{x} - x^*}(1) + \nabla_x \hat{f}(x^*) - \nabla_x f(x^*)
\]

\[
+ \sum_{j \in J_0(x^*)} \left\{ (\hat{\gamma}_j - \gamma_j)C_j^T + \hat{\gamma}_j(\hat{C}_j^T - C_j^T) \right\},
\]

8
where the last equality follows by Taylor’s theorem. Since we have assumed non-zero rows in $C$, by the Cauchy–Schwarz and triangle inequalities we obtain

$$j \in J_0(x^*) \Rightarrow |\hat{\gamma}_j - \gamma_j| = O \left( \|\hat{x} - x^*\|_2 + \|\nabla_x \hat{f}(x^*) - \nabla_x f(x^*)\|_2 + \|\hat{C}_j - C_j\|_2 \right).$$

Putting things together,

$$\|\hat{\gamma} - \gamma\|_2 = O \left( \|\hat{x} - x^*\|_2 + \|\nabla_x \hat{f}(x^*) - \nabla_x f(x^*)\|_2 + \sum_{j=1}^{r_C} \|\hat{C}_j - C_j\|_2 + \sum_{j=1}^{r_C} |\hat{d}_j - d_j| \right)$$

$$= O \left( \|\hat{x} - x^*\|_2 + \|\nabla_x \hat{f}(x^*) - \nabla_x f(x^*)\|_2 + \|\hat{C} - C\|_F + \|\hat{d} - d\|_2 \right),$$

and thus under the nonparametric condition (A6), we obtain (10).

Next, consider the following perturbed parametrized program $P_{\xi}$:

$$\begin{align*}
\text{minimize} & \quad f(x) + x^T \xi_1 \\
\text{subject to} & \quad Cx - d - \xi_2 \leq 0,
\end{align*}$$

(P_{\xi})

for a parameter $\xi = (\xi_1, \xi_2) \in \mathbb{R}^k \times \mathbb{R}^{r_C}$. $P_{\xi}$ can be viewed as a perturbed program of $\hat{P}_{nl}$; for $\xi = 0$, $P_{\xi}$ coincides with the program $\hat{P}_{nl}$. Here, the parameter $\xi$ will play a role of medium that contain all relevant stochastic information in $P_{\xi}$ (Shapiro 1993). Let $\bar{x}(\xi)$ denote the solution of the program $P_{\xi}$. Clearly, we get $\bar{x}(0) = x^*$.

By Theorem 3.1 and the nonparametric condition (A6), $\hat{x} \xrightarrow{P} x^*$ at the rate of $n^{1/2}$. Also, by positive definiteness of $\nabla^2_x f(x^*)$ and the fact that we have only linear constraints, the uniform version of the quadratic growth condition also holds at $\bar{x}(\xi)$ (see Shapiro (1993) Assumption...
A3)). Hence, given that the SC condition holds, by Shapiro (1993, Theorem 3.1) we get

\[ \hat{x} = \bar{x}(\xi) + o_P(n^{-1/2}) \]  \hspace{1cm} (11)

where

\[ \zeta = \begin{bmatrix} \nabla_{\hat{x}} f(x^*) - \nabla_{x} f(x^*) - \sum_{j \in J_0(x^*)} \gamma_j \left\{ \tilde{C}_j - C_j \right\} \\ -(\tilde{C}_{ac} - C_{ac})x^* \end{bmatrix} \equiv \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}. \]

If \( \bar{x}(\xi) \) is differentiable at \( \xi = 0 \) in the sense of Frechet, we have

\[ \bar{x}(\xi) - x^* = D_0 \bar{x}(\xi) + o(||\xi||), \]

where the mapping \( D_0 \bar{x}(\cdot) : \mathbb{R}^k \rightarrow \mathbb{R}^k \) is the directional derivative of \( \bar{x}(\cdot) \) at \( \xi = 0 \). Thus in this case, letting \( \xi = \zeta \) yields

\[ n^{1/2} (\hat{x} - x^*) = D_0 \bar{x}(n^{1/2} \zeta) + o_P(1), \]

which follows by (11) and that \( \bar{x}(0) = x^* \).

Now we shall show that such mapping \( D_0 \bar{x}(\cdot) \) exists and is indeed linear. To this end, we will show that \( \bar{x}(\xi) \) is locally totally differentiable at \( \xi = 0 \), followed by applying an appropriate form of the implicit function theorem. Define a vector-valued function \( H \in \mathbb{R}^{(k+|J_0(x^*)|)} \) by

\[ H(x, \xi, \gamma) = \begin{bmatrix} \nabla_{x} f(x) + C^T \gamma + \xi_1 \\ \text{diag}(\gamma)(Cx - d - \xi_2) \end{bmatrix}. \]

Let the solution of \( H(x, \xi, \gamma) = 0 \) be \( \bar{x}(\xi), \tilde{\gamma}(\xi) \). Due to the SC condition, \( \bar{x}(\xi), \tilde{\gamma}(\xi) \) satisfies...
the KKT conditions for \((\mathcal{P}_\xi)\). Then by the classical implicit function theorem (e.g., Dontchev and Rockafellar 2009, Theorem 1B.1), there always exists a neighborhood \(B_{\bar{r}}(0)\), for some \(\bar{r} > 0\), of \(\xi = 0\) such that \(\bar{x}(\xi)\) and its total derivative exist for \(\forall \xi \in B_{\bar{r}}(0)\). In particular, the derivative at \(\xi = 0\) is computed by

\[
\nabla_\xi \bar{x}(0) = -J_{x,\gamma}H(\bar{x}(0), 0, \bar{\gamma}(0))^{-1} \left[ J_\xi H(\bar{x}(0), 0, \bar{\gamma}(0)) \right],
\]

where in our case \(\bar{x}(0) = x^*, \bar{\gamma}(0) = \gamma^*,\) and thus

\[
J_{x,\gamma}H(\bar{x}(0), 0, \bar{\gamma}(0)) = \begin{bmatrix} \nabla^2 f(x^*) & C_{ac} \\ C_{ac}^\top & 0 \end{bmatrix},
\]

and

\[
J_\xi H(\bar{x}(0), 0, \bar{\gamma}(0)) = \begin{bmatrix} 1 \\ -\text{diag}(\gamma)1 \end{bmatrix}.
\]

Here the inverse of \(J_{x,\gamma}H(\bar{x}(0), 0, \bar{\gamma}(0))\) always exists under the LICQ condition (Assumption (A4)). Therefore we obtain that

\[
D_0 \bar{x}(n^{1/2}\zeta) = \begin{bmatrix} \nabla^2 f(x^*) & C_{ac} \\ C_{ac}^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ -\text{diag}(\gamma^*)1 \end{bmatrix} n^{1/2}\zeta.
\]

Since we assume that

\[
n^{1/2}\zeta = n^{1/2} \left[ \nabla_\bar{x} \tilde{f}(x^*) - \nabla_\bar{x} f(x^*) - \sum_{j \in J_0(x^*)} \gamma^*_j \left\{ \tilde{C}_{j}^\top - C_j^\top \right\} \right] \xrightarrow{d} \Upsilon,
\]

11
for any random variable $\Upsilon$, by Slutsky’s theorem we finally get the desired result

$$n^{1/2} (\hat{x} - x^*) \xrightarrow{d} \begin{bmatrix} \nabla^2_x f(x^*) & C_{ac} \\ C_{ac}^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ -\text{diag}(\gamma_j^*) 1 \end{bmatrix} \Upsilon.$$

\[\square\]

B.3 Proof of Lemma 4.1

Proof. Recall that we have

$$\phi_i(Z; \eta_i) = \frac{1}{\pi(X)} (A = a) \{Y_i - \mu_i(X,A)\} + \mu_i(X,a),$$

$$\phi_{ij}(Z; \eta_{ij}) = \frac{1}{\pi(X)} (A = a) \{Y_i Y_j - \sigma_{ij}(X,A)\} + \sigma_{ij}(X,a),$$

as the uncentered efficient influence functions for the parameter $\psi_i = \mathbb{E}[Y_i^a] = \mathbb{E}\{\mathbb{E}[Y_i | X, A = a]\}$ and $\psi_{ij} = \mathbb{E}[Y_i^a Y_j^a] = \mathbb{E}\{\mathbb{E}[Y_i Y_j | X, A = a]\}$ with the relevant nuisance functions $\eta_i = \{\pi(X), \mu_i(X,A)\}$, $\eta_j = \{\pi(X), \sigma_{ij}(X,A)\}$, respectively.

Recall that our proposed estimators in (1), (2) are given by

$$\hat{m}_i^a = \hat{\psi}_i,$$

$$\hat{\Sigma}_{ij}^a = \hat{\psi}_{ij} - \hat{\psi}_i \hat{\psi}_j,$$
where

$$\hat{\psi}_i = \mathbb{P}_n \{ \phi_i(Z; \hat{\eta}_i) \},$$

$$\hat{\psi}_j = \mathbb{P}_n \{ \phi_j(Z; \hat{\eta}_j) \},$$

$$\hat{\psi}_{ij} = \mathbb{P}_n \{ \phi_{ij}(Z; \hat{\eta}_{ij}) \}.$$

$$\hat{\psi}_{ij}, \hat{\psi}_i, \hat{\psi}_j$$ are semiparametric estimators for the mean outcomes $$\psi_{ij} = \mathbb{E}[Y_i Y_j], \psi_i = \mathbb{E}[Y_i],$$ $$\psi_j = \mathbb{E}[Y_j].$$ Hence with (B2), Together with either the Donsker condition or sample splitting, it follows that by Kennedy (2016),

$$\hat{\psi}_i - \psi_i = (\mathbb{P}_n - \mathbb{P}) \phi_i(Z) + O \left( \| \hat{\pi}_a - \pi_a \|_{2,\mathbb{P}} \| \hat{\mu}_i - \mu_i \|_{2,\mathbb{P}} \right) + O_P \left( \frac{\| \hat{\psi}_i - \psi_i \|}{\sqrt{n}} \right),$$

$$\hat{\psi}_j - \psi_j = (\mathbb{P}_n - \mathbb{P}) \phi_j(Z) + O \left( \| \hat{\pi}_a - \pi_a \|_{2,\mathbb{P}} \| \hat{\mu}_j - \mu_j \|_{2,\mathbb{P}} \right) + O_P \left( \frac{\| \hat{\psi}_j - \psi_j \|}{\sqrt{n}} \right),$$

$$\hat{\psi}_{ij} - \psi_{ij} = (\mathbb{P}_n - \mathbb{P}) \phi_{ij}(Z) + O \left( \| \hat{\pi}_a - \pi_a \|_{2,\mathbb{P}} \| \hat{\sigma}_{ij} - \sigma_{ij} \|_{2,\mathbb{P}} \right) + O_P \left( \frac{\| \hat{\psi}_{ij} - \psi_{ij} \|}{\sqrt{n}} \right),$$

and thus by the central limit theorem and the given consistency conditions,

$$\hat{m}_{i}^a - m_{i}^a = O_P \left( \| \hat{\pi}_a - \pi_a \|_{2,\mathbb{P}} \| \hat{\mu}_i - \mu_i \|_{2,\mathbb{P}} \right) + O_P \left( n^{-1/2} \right),$$

$$\hat{\Sigma}_{ij} - \Sigma_{ij} = O_P \left( \| \hat{\pi}_a - \pi_a \|_{2,\mathbb{P}} \left\{ \| \hat{\mu}_i - \mu_i \|_{2,\mathbb{P}} + \| \hat{\mu}_j - \mu_j \|_{2,\mathbb{P}} + \| \hat{\sigma}_{ij} - \sigma_{ij} \|_{2,\mathbb{P}} \right\} \right) + O_P \left( n^{-1/2} \right).$$
Since $k$ is finite, we have

$$\|\hat{\Sigma}^a - \Sigma^a\|_2 \leq \|\hat{\Sigma}^a - \Sigma^a\|_F$$

$$= \left( \sum_{i,j} |\hat{\Sigma}_{ij}^a - \Sigma_{ij}^a|^2 \right)^{1/2}$$

$$\leq \sum_{i,j} |\hat{\Sigma}_{ij}^a - \Sigma_{ij}^a|$$

$$= O_P \left( \|\hat{\pi}_a - \pi_a\|_{2,P} \left\{ \max_i \|\hat{\mu}_i - \mu_i\|_{2,P} + \max_{i,j} \|\hat{\sigma}_{ij} - \sigma_{ij}\|_{2,P} \right\} \right).$$

The case for $\hat{m}^a$ is straightforward and omitted here. Now we have the approximation-by-averages representation

$$\sqrt{n} \left\{ \begin{pmatrix} \hat{\psi}_i \\ \hat{\psi}_j \\ \hat{\psi}_{ij} \end{pmatrix} - \begin{pmatrix} \psi_i \\ \psi_j \\ \psi_{ij} \end{pmatrix} \right\} = \sqrt{n} (P_n - P) \begin{pmatrix} \phi_i(Z;\eta_i) \\ \phi_j(Z;\eta_j) \\ \phi_{ij}(Z;\eta_{ij}) \end{pmatrix} + o_P(1)$$

$$\xrightarrow{d} N \left( \begin{pmatrix} \phi_i(Z;\eta_i) \\ \phi_j(Z;\eta_j) \\ \phi_{ij}(Z;\eta_{ij}) \end{pmatrix} \right).$$

Now for a vector $(y_1, y_2, y_3)^\top \in \mathbb{R}^3$, define a function $g : \mathbb{R}^3 \to \mathbb{R}$ such that $g(y_1, y_2, y_3) = y_3 - y_1 y_2$. Also let $\psi \equiv (\psi_i, \psi_j, \psi_{ij})^\top$, $\hat{\psi} \equiv (\hat{\psi}_i, \hat{\psi}_j, \hat{\psi}_{ij})^\top$, and $\phi \equiv (\phi_{ij}, \phi_i, \phi_j)^\top$. Then by the
delta method, it follows that

\[
\sqrt{n} \left( g(\hat{\psi}) - g(\psi) \right) = \sqrt{n} \left( \hat{\Sigma}_{ij} - \Sigma_{ij} \right)
\]

\[
\xrightarrow{d} \left[ \nabla g(\psi) \right]^\top \mathcal{N} \left( 0, \text{cov} \begin{pmatrix} \phi_i(Z; \eta_i) \\ \phi_j(Z; \eta_j) \\ \phi_{ij}(Z; \eta_{ij}) \end{pmatrix} \right).
\]

\[\Box\]

### B.4 Proof of Theorem 5.1

Recall that \( \Sigma^*_S = \rho^*_1 \mathbb{I} + \rho^*_2 \hat{\Sigma}^a \) where \((\rho^*_1, \rho^*_2)\) is the solution of the program (6). Let us define

\[
\hat{\Sigma}^*_S = \nu^* \mathbb{I} + (1 - \nu^*) \hat{\Sigma}^a \quad \text{with} \quad \nu^* := \frac{\beta^2}{\delta^2},
\]

where \( \nu = \langle \Sigma^a, \mathbb{I} \rangle \), \( \beta^2 = \mathbb{P} \| \hat{\Sigma}^a - \Sigma^a \|_F^2 \) and \( \delta^2 = \mathbb{P} \| \hat{\Sigma}^a - \nu \mathbb{I} \|_F^2 \). The next lemma shows that \( \hat{\Sigma}^*_S \) converges in probability to \( \Sigma^*_S \) under very weak conditions.

**Lemma B.2.** Suppose that \( \hat{\pi}, \hat{\mu}_j, \hat{\sigma}_{ij} \) are consistent. Then

\[
\| \hat{\Sigma}^*_S - \Sigma^*_S \|_F = O_P \left( \| \hat{\pi} - \pi \|_2, \Sigma^k_{i,j=1} \left\{ \| \hat{\mu}_j - \mu_j \|_2, \| \hat{\sigma}_{ij} - \sigma_{ij} \|_2 \right\} \right),
\]

and thus \( \hat{\Sigma}^*_S \xrightarrow{P} \Sigma^*_S \) in the Frobenius norm.

**Proof.** Recall that we are interested in the following optimization program

\[
\text{minimize} \quad \mathbb{P} \| \rho \nu \mathbb{I} + (1 - \rho) \hat{\Sigma}^a - \Sigma^a \|_F^2
\]

over \( \rho, \nu \in \mathbb{R} \).
Now consider an unconstrained parametric program

$$\text{minimize } f(\rho, \nu; \Omega) \quad (P(\Omega))$$

over $\rho, \nu \in \mathbb{R}$

with $\Omega$ as the parameter. Since $f \in C^2$ with respect to $(\rho, \nu)$ and its Hessian is positive definite (note that we tacitly assumed $\hat{\Sigma}^a \neq \Sigma^a$, otherwise the result is trivial), a local minimizer of the program $P(\Omega)$ is Lipschitz stable. Let $\hat{y} = (\hat{\rho}, \hat{\nu}), y_0 = (\rho_0, \nu_0)$ be the solutions of $P(\hat{\Omega}), P(0_{k \times k})$. 

Now, for a $k \times k$ real-valued matrix $\Omega$, define a function $f$ indexed by $\Omega$ as

$$f(\rho, \nu; \Omega) = \rho^2 \|\nu \mathbb{I} - \Sigma^a\|_F^2 + (1 - \rho)^2 \|\hat{\Sigma}^a - \Sigma^a\|_F^2 + 2\rho(1 - \rho) \langle (\nu \mathbb{I} - \Sigma^a), \Omega \rangle.$$

So, if we let $\hat{\Omega}$ denote a matrix whose $(i, j)$-component is given by $P(\hat{\Sigma}^a_{ij} - \Sigma^a_{ij})$ ($1 \leq i, j \leq k$) and $0_{k \times k}$ denote a matrix of $k^2$ zeros, then we may write

$$f \left( \rho, \nu; \hat{\Omega} \right) = \rho^2 \|\nu \mathbb{I} - \Sigma^a\|_F^2 + (1 - \rho)^2 \|\hat{\Sigma}^a - \Sigma^a\|_F^2 + 2\rho(1 - \rho) \langle (\nu \mathbb{I} - \Sigma^a), \hat{\Omega} \rangle,$$

$$f \left( \rho, \nu; 0_{k \times k} \right) = \rho^2 \|\nu \mathbb{I} - \Sigma^a\|_F^2 + (1 - \rho)^2 \|\hat{\Sigma}^a - \Sigma^a\|_F^2.$$
respectively. Then by Lemma B.1, it follows that
\[
\|\hat{y} - y_0\|_2 = O_P\left(\|\hat{\Omega}\|_2\right)
\]
\[
= O_P\left(\|\mathbb{P}\left\{\hat{\Sigma}^a - \Sigma^a\right\}\|_F\right)
\]
\[
= O_P\left(\sum_{i,j} \|\mathbb{P}\left\{\hat{\Sigma}_{ij}^a - \Sigma_{ij}^a\right\}\|_F\right)
\]
\[
= O_P\left(\sum_i \|\hat{\pi} - \pi\|_{2,\mathbb{P}} \|\hat{\mu}_i - \mu_i\|_{2,\mathbb{P}} + \sum_j \|\hat{\pi} - \pi\|_{2,\mathbb{P}} \|\hat{\mu}_j - \mu_j\|_{2,\mathbb{P}} + \sum_{ij} \|\hat{\pi} - \pi\|_{2,\mathbb{P}} \|\hat{\sigma}_{ij} - \sigma_{ij}\|_{2,\mathbb{P}}\right)
\]
\[
= O_P\left(\|\hat{\pi} - \pi\|_{2,\mathbb{P}} \Sigma_{ij} \left\{\|\hat{\mu}_j - \mu_j\|_{2,\mathbb{P}} + \|\hat{\sigma}_{ij} - \sigma_{ij}\|_{2,\mathbb{P}}\right\}\right),
\]
where the fourth line follows by rearranging the second-order remainder terms of the estimators \(\psi_{ij}, \psi_i, \psi_j, 1 \leq i, j \leq k\) defined in the appendix B.3.

The program \(P(\hat{\Omega})\) is equivalent to (13). Moreover, using the same logic used in Theorem 2.1 of Ledoit and Wolf (2004), it can be deduced that the solution of the program \(P(0_{k \times k})\) is given by \((\varpi^*, \nu)\) defined in (12): i.e., \(y_0 = (\varpi^*, \nu)\). Hence, by (14) and the given consistency conditions, we finally obtain the desired result:
\[
\|\hat{\Sigma}^*_{S} - \Sigma^*_{S}\|_F = O_P\left(\|\hat{\pi} - \pi\|_{2,\mathbb{P}} \Sigma_{ij} \left\{\|\hat{\mu}_j - \mu_j\|_{2,\mathbb{P}} + \|\hat{\sigma}_{ij} - \sigma_{ij}\|_{2,\mathbb{P}}\right\}\right)
\]
\[
= o_P(1).
\]

Next, we show that our proposed estimator (7) converges in probability to \(\hat{\Sigma}^*_{S}\), which concludes the first part of our proof of Theorem 5.1.

Lemma B.3. Let \(\|\hat{\pi} - \pi\|_{2,\mathbb{P}} \Sigma_{ij} \left\{\|\hat{\mu}_j - \mu_j\|_{2,\mathbb{P}} + \|\hat{\sigma}_{ij} - \sigma_{ij}\|_{2,\mathbb{P}}\right\} = O_P(r(n))\). Then,
\[
\|\hat{\Sigma}^*_{S} - \Sigma^*_{S}\|_F = O_P\left(n^{-1/2} \vee r(n)\right).
\]
Proof. It suffices to show that \( \hat{\nu}, \hat{\delta}, \hat{\beta} \) are consistent at the specified rate. Let \( \hat{S}^a \) denote the (virtual) sample covariance matrix that can be computed from \( Y_1^a, \ldots, Y_n^a \) (note that \( \hat{S}^a \) can never be computed in reality).

i) \( \hat{\nu} - \nu = O_p \left( n^{-1/2} \lor r(n) \right) \). \( \forall 1 \leq i, j \leq k \), we have \( \hat{\Sigma}_{ij}^a - \Sigma_{ij}^a = O_p(r(n)) \) by Lemma 4.1. It also follows \( \hat{S}_{ij}^a - \Sigma_{ij}^a = O_p(n^{-1/2}) \) by the central limit theorem. Hence by the continuous mapping theorem,

\[
\| \hat{\Sigma}^a - \hat{S}^a \|_F = O_p \left( n^{-1/2} \lor r(n) \right).
\]

Now we have

\[
|\hat{\nu} - \nu| = \left| \langle \hat{\Sigma}^a, I \rangle - \langle \Sigma^a, I \rangle \right| \\
= \left| \langle \hat{\Sigma}^a - \hat{S}^a + \hat{S}^a, I \rangle - \langle \Sigma^a, I \rangle \right| \\
\leq \left| \langle \hat{\Sigma}^a - \hat{S}^a, I \rangle \right| + \left| \langle \hat{S}^a, I \rangle - \langle \Sigma^a, I \rangle \right| \\
= O_p \left( n^{-1/2} \lor r(n) \right) + O_p(n^{-1/2}),
\]

which yields the desired conclusion.

ii) \( \hat{\delta} - \delta = O_p \left( n^{-1/2} \lor r(n) \right) \). First note that

\[
\hat{\delta} - \delta \\
= \| \hat{\Sigma}^a - \hat{S}^a \|_F - \left( \| \hat{\Sigma}^a - \nu I \|_F^2 \right)^{1/2} \\
= \| \hat{\Sigma}^a - \hat{S}^a + \hat{S}^a - \langle \hat{\Sigma}^a - \hat{S}^a + \hat{S}^a, I \rangle I \|_F - \left( \| \hat{\Sigma}^a - \hat{S}^a + \hat{S}^a - \nu I \|_F^2 \right)^{1/2} \\
\leq \| \hat{\Sigma}^a - \hat{S}^a \|_F + \| \hat{S}^a - \langle \hat{\Sigma}^a, I \rangle I \|_F + \| \langle \hat{\Sigma}^a - \hat{S}^a, I \rangle I \|_F - \left( \| \hat{\Sigma}^a - \nu I \|_F^2 \right)^{1/2} + \left( \| \hat{\Sigma}^a - \hat{S}^a \|_F^2 \right)^{1/2},
\]

where the second last inequality follows by Jensen’s Inequality and the last by the triangle inequality and the fact that \( \sqrt{\|A + B\|_F^2} \geq \sqrt{\|A\|_F^2} - \sqrt{\|B\|_F^2}, \forall A, B \in \mathbb{R}^{k \times k} \).
From part i), we know $\|\hat{\Sigma}^a - \hat{S}^a\|_F = O_p\left(\sqrt{n} \lor r(n)\right)$. Since $\|\langle \hat{\Sigma}^a - \hat{S}^a, I\rangle\|_F \leq \|\hat{\Sigma}^a - \hat{S}^a\|_F$, it follows $\|\langle \hat{\Sigma}^a - \hat{S}^a, I\rangle\|_F = O_p\left(\sqrt{n} \lor r(n)\right)$. Thus the first and third terms in the last display converge at the desired rate.

For the fifth term, by the triangle inequality

$$\left(\|\hat{\Sigma}^a - \hat{S}^a\|_F^2\right)^{1/2} \equiv \|\hat{\Sigma}^a - \hat{S}^a\|_{F,P} \leq \|\hat{\Sigma}^a - \Sigma^a\|_{F,P} + \|\hat{S}^a - \Sigma^a\|_{F,P},$$

where we view $\|\cdot\|_{F,P}$ as an element-wise $L_2(\mathbb{P})$-norm for matrix. By Proposition B.1, we have

$$\|\hat{\Sigma}^a - \Sigma^a\|_{F,P} \leq \sum_{i,j} \|\hat{\Sigma}^a_{ij}\|_2^2, \sum_{i,j} \|\hat{\Sigma}^a_{ij} - \Sigma^a_{ij}\|_2^2 = O_p\left(\frac{1}{\sqrt{n}}\right) + o_p\left(r(n)\right).$$

Further, by Theorem 3.1 in Ledoit and Wolf (2004) it follows that

$$\|\hat{S}^a - \Sigma^a\|_{F,P} = O\left(\frac{1}{\sqrt{n}}\right).$$

Therefore, we get $\|\hat{\Sigma}^a - \hat{S}^a\|_{F,P} = O_p\left(\sqrt{n} \lor r(n)\right)$. Bringing these results together, we have

$$\|\hat{\Sigma}^a - \hat{S}^a\|_F + \|\hat{S}^a - \langle \hat{S}^a, I\rangle\|_F + \left(\mathbb{P}\|\hat{\Sigma}^a - \hat{S}^a\|_F^2\right)^{1/2} + \|\hat{S}^a - \langle \hat{S}^a, I\rangle\|_F - \left(\mathbb{P}\|\hat{\Sigma}^a - \nu\|_F^2\right)^{1/2} = O_p\left(\sqrt{n} \lor r(n)\right) + \|\hat{S}^a - \langle \hat{S}^a, I\rangle\|_F - \left(\mathbb{P}\|\hat{S}^a - \nu\|_F^2\right)^{1/2}. $$

However, since we consider the case of fixed $p$, Lemma 3.3 of Ledoit and Wolf (2004) implies

$$\|\hat{S}^a - \langle \hat{S}^a, I\rangle\|_F - \left(\mathbb{P}\|\hat{S}^a - \nu\|_F^2\right)^{1/2} = O\left(\frac{1}{\sqrt{n}}\right),$$

19
which finally leads to

\[ \hat{\delta} - \delta = O_P \left( n^{-1/2} \lor r(n) \right). \]

Similarly, one can also show that

\[ \delta - \hat{\delta} \leq \left( \mathbb{P} \| \hat{S}^a - \nu \|_F^2 \right)^{1/2} - \left\| \hat{S}^a - \langle \hat{S}^a, I \rangle \right\|_F + O_P \left( n^{-1/2} \lor r(n) \right) \]

\[ = O_P \left( n^{-1/2} \lor r(n) \right). \]

Hence, we obtain \( \hat{\delta} - \delta = O_P \left( n^{-1/2} \lor r(n) \right). \)

\textbf{iii) } \hat{\beta} - \beta = O_P \left( n^{-1/2} \lor r(n) \right). \] This in fact follows because with fixed \( k, \) each of \( \hat{\beta}, \beta \) vanishes quickly to zero. To show this, first let \( \hat{S}^a = \sum_{t=1}^n \hat{S}_t^a \) and consider the following quantity

\[ \sqrt{\mathbb{P} \| \hat{\Sigma}^a - \hat{S}^a + \hat{S}^a - \Sigma^a \|_F^2}. \]

By the Cauchy-Schwarz inequality we have

\[ \mathbb{P} \| \hat{\Sigma}^a - \hat{S}^a + \hat{S}^a - \Sigma^a \|_F^2 = \mathbb{P} \| \hat{\Sigma}^a - \hat{S}^a \|_F^2 + 2 \mathbb{P} \left\{ \sum_{ij} (\hat{\Sigma}_{ij}^a - \hat{S}_{ij}^a) (\hat{S}_{ij}^a - \Sigma_{ij}^a) \right\} + \mathbb{P} \| \hat{S}^a - \Sigma^a \|_F^2 \]

\[ \leq \mathbb{P} \| \hat{\Sigma}^a - \hat{S}^a \|_F^2 + 2 \mathbb{P} \left\{ \| (\hat{\Sigma}^a - \hat{S}^a) \|_F \| (\hat{S}^a - \Sigma^a) \|_F \right\} + \mathbb{P} \| \hat{S}^a - \Sigma^a \|_F^2 \]

\[ \leq \mathbb{P} \| \hat{\Sigma}^a - \hat{S}^a \|_F^2 + 2 \sqrt{\mathbb{P} \| (\hat{\Sigma}^a - \hat{S}^a) \|_F^2} \sqrt{\mathbb{P} \| (\hat{S}^a - \Sigma^a) \|_F^2} + \mathbb{P} \| \hat{S}^a - \Sigma^a \|_F^2. \]

In part ii), we showed \( \sqrt{\mathbb{P} \| (\hat{\Sigma}^a - \hat{S}^a) \|_F^2} = O_P \left( n^{-1/2} \lor r(n) \right) \) and \( \sqrt{\mathbb{P} \| (\hat{S}^a - \Sigma^a) \|_F^2} = O \left( n^{-1/2} \right). \) Hence it follows

\[ \sqrt{\mathbb{P} \| (\hat{\Sigma}^a - \hat{S}^a) \|_F^2} \sqrt{\mathbb{P} \| (\hat{S}^a - \Sigma^a) \|_F^2} = O_P \left( n^{-1/2} \lor r(n) \right) O_P \left( n^{-1/2} \right), \]
and consequently, we have

\[ \sqrt{\mathbb{P}||\hat{\Sigma} - \hat{S}_t + \hat{S}_a - \Sigma^a||_F^2} = O_p(n^{-1/2} \vee r(n)). \]

Next, we consider

\[ \sqrt{\frac{1}{n^2} \sum_{t=1}^{n} ||\hat{\Sigma}_t - \hat{S}_t + \hat{S}_a - \Sigma^a||_F^2}. \]

We shall first show that

\[ \frac{1}{n} \sum_{t=1}^{n} ||\hat{\Sigma}_t - \hat{S}_t||_F^2 = O_p(1). \]

To this end, we note that

\[ \frac{1}{n} \sum_{t=1}^{n} ||\hat{\Sigma}_t - \Sigma^a||_F^2 = \sum_{ij} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left( \hat{\phi}_{ij}^a(Z_t) - \mathbb{P}_n \hat{\phi}_{ij}^a \mathbb{P}_n \hat{\phi}_{ij}^a - \Sigma_{ij}^a \right)^2 \right\} \]

\[ = \sum_{ij} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left( \hat{\phi}_{ij}^a(Z_t) - \mathbb{P}_n \hat{\phi}_{ij}^a + \mathbb{P}_n \hat{\phi}_{ij}^a + \mathbb{P}_n \hat{\phi}_{ij}^a \mathbb{P}_n \hat{\phi}_{ij}^a - \Sigma_{ij}^a \right)^2 \right\} \]

\[ \leq \sum_{ij} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left( \hat{\phi}_{ij}^a(Z_t) - \mathbb{P}_n \hat{\phi}_{ij}^a \right)^2 + \left( \mathbb{P}_n \hat{\phi}_{ij}^a - \mathbb{P}_n \hat{\phi}_{ij}^a \right)^2 \right\} \]

\[ + \left| \mathbb{P}_n \hat{\phi}_{ij}^a - \mathbb{P}_n \hat{\phi}_{ij}^a \mathbb{P}_n \hat{\phi}_{ij}^a - \Sigma_{ij}^a \right| \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left( \hat{\phi}_{ij}^a(Z_t) - \mathbb{P}_n \hat{\phi}_{ij}^a \right)^2}, \]

and that \( \forall 1 \leq i, j \leq k, \)

\[ \frac{1}{n} \sum_{t=1}^{n} \left( \hat{\phi}_{ij}^a(Z_t) - \mathbb{P}_n \hat{\phi}_{ij}^a \right)^2 \xrightarrow{p} \text{var} \left( \hat{\phi}_{ij}^a(Z) \mid Z_{n+1}, ..., Z_{2n} \right) \]

\[ \mathbb{P}_n \hat{\phi}_{ij}^a - \mathbb{P}_n \hat{\phi}_{ij}^a \mathbb{P}_n \hat{\phi}_{ij}^a - \Sigma_{ij}^a = O_p(1). \]
Then from the given boundedness conditions, \( \text{var} \left( \hat{\phi}_{ij}^a(Z) \mid Z_{n+1}, \ldots, Z_{2n} \right) \) is bounded in probability, and thus so is the RHS of the last inequality. Hence \( \frac{1}{n} \sum_{t=1}^{n} \| \tilde{\Sigma}_t - \hat{S}_t^a \|_F^2 = O_p(1) \).

Next, by the unbiasedness of the sample covariance estimator we get

\[
\frac{1}{n} \sum_{t=1}^{n} \| \tilde{\Sigma}_t - \hat{S}_t^a \|_F^2 = \sum_{ij} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left( \hat{S}_{t,ij}^a - \Sigma_{ij}^a \right)^2 \right\} \rightarrow \sum_{ij} \text{var} \left( \hat{S}_{ij}^a \right) < \infty
\]

as \( Y_i \)'s have finite fourth moments. Thus \( \frac{1}{n} \sum_{t=1}^{n} \| \tilde{\Sigma}_t - \hat{S}_t^a \|_F^2 = O_p(1) \).

Using these facts, and again by the Cauchy Schwarz inequality, we obtain

\[
\frac{1}{n} \sum_{t=1}^{n} \| \tilde{\Sigma}_t - \hat{S}_t^a \|_F^2 = O_p(1)
\]

as desired.

Moreover, the terms involving \( \| \hat{\Sigma}^a - \hat{S}^a \|_F \) converge at fast rates since \( \| \hat{\Sigma}^a - \hat{S}^a \|_F = O_F \left( n^{-1/2} \lor r(n) \right) \) as shown in part i). Therefore, we have

\[
\frac{1}{n^2} \sum_{i=1}^{n} \| \hat{\Sigma}_i - \tilde{\Sigma}_i^a + \hat{S}_i^a - \hat{S}_i^a + \hat{S}_i^a - \hat{\Sigma}^a \|_F^2 = O \left( n^{-1} \right) + \frac{1}{n^2} \sum_{i=1}^{n} \| \hat{S}_i^a - \hat{S}_i^a \|_F^2
\]

which follows by simple rearrangement and the Cauchy Schwarz inequality.

On the other hand, Lemma 3.4 of Ledoit and Wolf (2004) indicates that

\[
\frac{1}{n^2} \sum_{i=1}^{n} \| \hat{S}_i^a - \hat{\Sigma}^a \|_F^2 - \| \hat{S}_i^a - \Sigma^a \|_F^2 = O_F \left( n^{-1} \right).
\]

Now we can bring all the results together, to get to the conclusion

\[
\sqrt{\frac{1}{n^2} \sum_{t=1}^{n} \| \tilde{\Sigma}_t - \hat{S}_t^a + \hat{S}_t^a - \hat{S}_i^a + \hat{S}_i^a - \hat{\Sigma}^a \|_F^2} = O_F \left( n^{-1/2} \lor r(n) \right),
\]

22
which completes the proof.

The followings are the auxiliary technical results used for the proof of Lemma [B.3]

**Lemma B.4.** Let $P_n$ denote the empirical measure over an iid sample $(Z_1, \ldots, Z_n)$. Also we let $f$ and $\hat{f}$ be any function and its estimator constructed in a separate, independent sample $(Z_{n+1}, \ldots, Z_{2n})$, respectively. Then we have

$$
\|P_n \hat{f} - P f\|_{2,P} \leq \frac{\|f\|_{2,P}}{\sqrt{n}} + P \left( \hat{f} - f \right).
$$

**Proof.**

$$
\|P_n \hat{f} - P f\|_{2,P} = \| (P_n - P) \hat{f} - P (\hat{f} - f) \|_{2,P}
\leq \| (P_n - P) \hat{f} \|_{2,P} + \| P (\hat{f} - f) \|_{2,P}
\leq \sqrt{\frac{\text{var} \left[ \hat{f} \mid Z_{n+1}, \ldots, Z_{2n} \right]}{n}} + \| P (\hat{f} - f) \|_{2,P}
\leq \frac{\|f\|_{2,P}}{\sqrt{n}} + P \left( \hat{f} - f \right),
$$

where the third line follows by Lemma C.3 in [Kim et al. (2018)].

**Proposition B.1.** Suppose that $\forall 1 \leq i, j \leq k$, $\| \hat{\sigma}_{ij} \|_{2,P} = O_P(1)$ and $\| \hat{\mu}_i \|_{2,P} = O_P(1)$. Then, we have

$$
\| \hat{\Sigma}_{ij} - \Sigma_{ij} \|_{2,P} = O_P \left( \frac{1}{\sqrt{n}} \right) + o_P \left( r(n) \right),
$$

where $r(n)$ is defined in Lemma [B.3]

**Proof.** Recall that for our counterfactual covariance estimator $\hat{\Sigma}_{ij}^a$ defined in (2), we can write

$$
\hat{\Sigma}_{ij} - \Sigma_{ij} = P_n \hat{\phi}_{ij}^a - P_n \hat{\phi}'_i P_n \hat{\phi}_j^a - P \hat{\phi}_i^a P \hat{\phi}_j^a.
$$
Hence, by the result of Lemma [B.4] we obtain

\[
\|\hat{\Sigma}_{ij} - \Sigma_{ij}\|_{2,P} = \left\|P_n \hat{\phi}^a_{ij} - P_n \phi^a_{ij} P_n \hat{\phi}^a_{ij} - P \phi^a_{ij} P \phi^a_{ij}\right\|_{2,P}
\]

\[
\leq \left\|P_n \hat{\phi}^a_{ij} - P \phi^a_{ij}\right\|_{2,P} + \left\|P_n \phi^a_{ij} \left\{P_n \hat{\phi}^a_{ij} - P \phi^a_{ij}\right\}\right\|_{2,P} + \left\|P_n \hat{\phi}^a_{ij} \left\{P_n \phi^a_{ij} - P \phi^a_{ij}\right\}\right\|_{2,P}
\]

\[
\leq \left\|P_n \hat{\phi}^a_{ij} - P \phi^a_{ij}\right\|_{2,P} + \left\|P_n \phi^a_{ij}\right\|_{2,P} \left\|P_n \hat{\phi}^a_{ij} - P \phi^a_{ij}\right\|_{2,P} + \left\|P_n \hat{\phi}^a_{ij}\right\|_{2,P} \left\|P_n \phi^a_{ij} - P \phi^a_{ij}\right\|_{2,P}
\]

\[
\lesssim \frac{\|\hat{\phi}^a_{ij}\|_{2,P} + \|\phi^a_{ij}\|_{2,P} \|\hat{\phi}^a_{ij}\|_{2,P}}{\sqrt{n}} + \left(1 + \|\hat{\phi}^a_{ij}\|_{2,P} + \|\phi^a_{ij}\|_{2,P}\right) o_P (r(n))
\]

\[
= O_P \left(\frac{1}{\sqrt{n}}\right) + O_P \left(1\right) o_P (r(n))
\]

, which gives the result.

The second part of the proof of Theorem 5.1 immediately follows by Corollary 4.1 and Theorem 3.1.

B.5 Proof of Theorem 5.2

As in [Pnl] for arbitrary $k \times k$ matrix $\Sigma$ and $k \times 1$ vector $m$, define a parametric program $P(\Sigma, m)$

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} w^T \Sigma w - \tau w^T m \\
\text{subject to} & \quad w \in S(\Sigma, m),
\end{align*}
\]

(P(\Sigma, m))

by viewing $\Sigma$ and $m$ together as parameters. Then $P(\Sigma^a, m^a), P(\hat{\Sigma}^a_{cor}, \hat{m}^a)$ are our true and approximating programs, respectively.
Proof. By virtue of the quadratic growth condition, \( s^* (P(\Sigma^a, m^a)) \) is a singleton (Still 2018, Theorem 2.5), and \( w^* = s^* (P(\Sigma^a, m^a)) \). By the Lipschitz stability result for smooth \((C^2)\) parametric programming (Still 2018, Theorem 6.2), for each \( w^* \) there exist \( \varepsilon, L > 0 \) such that for all \( \bar{\Sigma} \in B_\varepsilon(\Sigma^a) \), \( \bar{m} \in B_\varepsilon(m^a) \) there exists at least one local minimizer \( \bar{x} \) of \( P(\bar{\Sigma}, \bar{m}) \) that satisfies

\[
\|\bar{x} - w^*\|_2 \leq L \left\{ \|\bar{\Sigma} - \Sigma^a\|_2 + \|\bar{m} - m^a\|_2 \right\}.
\]

Now, it is straightforward to see that

\[
\|\hat{w}_{cor} - w^*\|_2 \leq \|\hat{w}_{cor} - w^*\|_2 \mathbb{1} \left\{ \hat{\Sigma}^*_{cor} \in B_\varepsilon(\Sigma^a), \hat{m}^a \in B_\varepsilon(m^a) \right\}
+ \|\hat{w}_{cor} - w^*\|_2 \mathbb{1} \left\{ \hat{\Sigma}^*_{cor} \notin B_\varepsilon(\Sigma^a) \right\} + \mathbb{1} \left\{ \hat{m}^a \notin B_\varepsilon(m^a) \right\}.
\]

For the first term, it follows that

\[
\|\hat{w}_{cor} - w^*\|_2 \mathbb{1} \left\{ \hat{\Sigma}^*_{cor} \in B_\varepsilon(\Sigma^a), \hat{m}^a \in B_\varepsilon(m^a) \right\} \leq L \left( \|\hat{\Sigma}^*_{cor} - \Sigma^a\|_2 + \|\hat{m} - m^a\|_2 \right).
\]

Also \( \forall \varepsilon', \varepsilon > 0, \) we have that

\[
\mathbb{P} \left( \mathbb{1} \left\{ \hat{\Sigma}^*_{cor} \notin B_\varepsilon(\Sigma^a) \right\} > \varepsilon' \right) = \mathbb{P} \left( \hat{\Sigma}^*_{cor} \notin B_\varepsilon(\Sigma^a) \right)
= \mathbb{P} \left( \|\hat{\Sigma}^*_{cor} - \Sigma^a\|_2 \geq \varepsilon \right)
\leq \mathbb{P} \left( \|\hat{\Sigma}^*_{cor} - \hat{\Sigma}^a\|_2 + \|\hat{\Sigma}^a - \Sigma^a\|_2 \geq \varepsilon \right)
\rightarrow 0,
\]

which follows by the fact that \( \|\hat{\Sigma}^*_{cor} - \hat{\Sigma}^a\|_2 + \|\hat{\Sigma}^a - \Sigma^a\|_2 = o_P(1) \) under the given conditions.

Similarly, we obtain \( \mathbb{1} \left\{ \hat{m}^a \notin B_\varepsilon(m^a) \right\} = o_P(1) \) which immediately follows by that \( \|\hat{m} - m^a\|_2 = \)
Putting the pieces together, we obtain
\[
\| \hat{w}_{\text{cor}} - w^* \|_2 = O \left( \| \hat{\Sigma}_{\text{cor}}^{*} - \Sigma^a \|_2 + \| \hat{m} - m^a \|_2 \right) + o_P \left( \| \hat{w}_{\text{cor}} - w^* \|_2 \right)
\]
which yields
\[
\frac{\| \hat{w}_{\text{cor}} - w^* \|_2}{\| \hat{\Sigma}_{\text{cor}}^{*} - \Sigma^a \|_2 + \| \hat{m} - m^a \|_2} = O_P(1) .
\]
Hence, the result follows.

\[ \square \]

### B.6 Alternate proof of Theorem 5.2

**Proof.** Since \( \Sigma^a \) is assumed to be PD, the quadratic growth condition (A2) holds at \( w^* \), so we have
\[
\| \hat{w}_{\text{cor}} - w^* \|_2 = O_P \left( \| \hat{\Sigma}_{\text{cor}}^{*} - \Sigma^a \|_2 \lor r_n \lor n^{-1/2} \right)
\]
\[
\implies \| \hat{w}_{\text{cor}} - w^* \|_2 = O_P \left( \| \hat{\Sigma}_{\text{cor}}^{*} - \Sigma^a \|_2 \lor \| \hat{\Sigma}^a - \Sigma^a \|_2 \lor r_n \lor n^{-1/2} \right)
\]
\[
= O_P \left( \| \hat{\Sigma}_{\text{cor}}^{*} - \Sigma^a \|_2 \lor r_n \lor n^{-1/2} \right)
\]

where the first line follows from Theorem 3.1, and the last line follows from Lemma 4.1 which says that \( \| \hat{\Sigma}^a - \Sigma^a \|_2 = O_P(r_n) + O_P(n^{-1/2}) \).

For the second statement of the theorem, suppose that \( \hat{\Sigma}_{\text{cor}}^{*} = \hat{\Sigma}^a \) whenever \( \hat{\Sigma}^a \) is PD. Let \( S^k_+ \) denote the set of all PD \( k \times k \) matrices, and for any \( \epsilon > 0 \) let \( B_\epsilon(\Sigma^a) = \{ M \in S^{k \times k} : \| M - \Sigma^a \|_F \leq \epsilon \} \) denote the \( \epsilon \)-ball in Frobenius norm of all symmetric matrices around \( \Sigma^a \).
Since $S_k^+$ is an open set, we can fix a $\delta > 0$ such that $B_\delta(S^a) \subset S_k^+$. We have

$$P(\hat{\Sigma}^a_{\text{cor}} \neq \hat{\Sigma}^a) \leq P(\hat{\Sigma}^a \notin B_\delta(S^a))$$

$$= P(\|\hat{\Sigma}^a - \Sigma^a\|_F > \delta)$$

$$\to 0 \text{ as } n \to \infty$$

where the last line again follows from Lemma 4.1. Next, note that for any sequence of positive numbers $a_n$,

$$P(a_n^{-1}\|\hat{\Sigma}^a_{\text{cor}} - \hat{\Sigma}^a\|_F > \epsilon) = P(a_n^{-1}\|\hat{\Sigma}^a_{\text{cor}} - \hat{\Sigma}^a\| > \epsilon | \hat{\Sigma}^a_{\text{cor}} \neq \hat{\Sigma}^a) P(\hat{\Sigma}^a_{\text{cor}} \neq \hat{\Sigma}^a)$$

$$\leq P(\hat{\Sigma}^a_{\text{cor}} \neq \hat{\Sigma}^a)$$

$$\to 0 \text{ as } n \to \infty$$

so that $\|\hat{\Sigma}^a_{\text{cor}} - \hat{\Sigma}^a\|_F = o_P(a_n)$. Letting $a_n = r_n \vee n^{-1/2}$, we have that

$$\|\hat{w}_{\text{cor}} - w^*\|_2 = O_P(r_n \vee n^{-1/2})$$

as claimed.