Inference for Linear Systems with Unknown Coefficients*

Yuehao Bai
Department of Economics
University of Southern California
yuehao.bai@usc.edu

Kirill Ponomarev
Department of Economics
University of Chicago
kponomarev@uchicago.edu

Max Tabord-Meehan
Department of Economics
University of Toronto
m.tabordmeehan@utoronto.ca

Andres Santos
Department of Economics
University of California–Los Angeles
andres@econ.ucla.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Alexander Torgovitsky
Department of Economics
University of Chicago
torgovitsky@uchicago.edu

April 27, 2026

Abstract

This paper considers the problem of testing whether there exists a solution satisfying certain non-negativity constraints to a linear system of equations. Importantly and in contrast to some prior work, we allow all parameters in the system of equations, including the slope coefficients, to be unknown. For this reason, we describe the linear system as having unknown (as opposed to known) coefficients. This hypothesis testing problem arises naturally when constructing confidence sets for possibly partially identified parameters in the analysis of nonparametric instrumental variables models, treatment effect models, and random coefficient models, among other settings. To rule out certain instances in which the testing problem is impossible, in the sense that the power of any test will be bounded by its size, we begin our analysis by characterizing the closure of the null hypothesis with respect to the total variation distance. We then use this characterization to develop novel testing procedures based on sample-splitting. We establish the validity of our testing procedures under weak and interpretable conditions on the linear system. An important feature of these conditions is that they permit the dimensionality of the problem to grow rapidly with the sample size. A further attractive property of our tests is that they do not require simulation to compute suitable critical values. We illustrate the practical relevance of our theoretical results in a simulation study.

KEYWORDS: Linear programming, linear (in)equalities, partial identification, uniform inference, treatment effects, nonparametric instrumental variables

JEL classification codes: C31, C35, C36

*We thank Geonwoo Kim for outstanding research assistance. Shaikh acknowledges financial support from National Science Foundation Grant SES-2419008.

1 Introduction

Given an independent and identically distributed (i.i.d.) sample $\{Z_i\}_{i=1}^n$ with Z_i distributed according to $P \in \mathbf{P}$, this paper studies the hypothesis testing problem

$$H_0 : P \in \mathbf{P}_0 \quad \text{vs.} \quad H_1 : P \in \mathbf{P} \setminus \mathbf{P}_0 \ , \tag{1}$$

where $\mathbf{P}$ is a "large" set of distributions satisfying conditions described below and

$$\mathbf{P}_0 := \{P \in \mathbf{P} : A_0(P)x_0 + A_1(P)x_1 = \beta(P) \text{ for some } x_0 \in \mathbb{R}^{d_0}, x_1 \in \mathbb{R}^{d_1}, x_1 \geq 0\} \ . \tag{2}$$

Here, "$x_1 \geq 0$" signifies that all coordinates of $x_1 \in \mathbb{R}^{d_1}$ are non-negative, $A_0(P)$ is a $p \times d_0$ matrix with $d_0 \geq 0$, $A_1(P)$ is a $p \times d_1$ matrix, and $\beta(P)$ is a $p \times 1$ vector.

As discussed further in Section 2, the testing problem described above arises naturally in many settings of empirical interest, including (i) inference for linear functionals of structural functions in nonparametric instrumental variables (NPIV) models with shape restrictions, as in Freyberger and Horowitz (2015); (ii) inference for average marginal effects in nonlinear models with random coefficients, as in Fox et al. (2011); (iii) inference for treatment effect parameters that are partially identified through the marginal treatment response (MTR) framework of Mogstad et al. (2018); (iv) inference under "synthetic parallel trends" with convex weights as considerd in Liu (2025); and (v) inference for counterfactual choice probabilities in the distribution-free binary choice model in Gu and Russell (2023). We also note that the null hypothesis (1) subsumes as a special case the setting where $A_0(P)$ and $A_1(P)$ are *known*, for which there exist many other examples (see, for instance, the examples discussed in Fang et al., 2023).

We demonstrate in Section 3 that testing (1) in certain special cases is impossible (in the sense that the power of any test is bounded by its size), by observing that the null set $\mathbf{P}_0$ is dense in $\mathbf{P}$ with respect to the total variation metric. Accordingly, the first contribution of the paper is to obtain a characterization of the closure of $\mathbf{P}_0$ which guides the construction of our test. Exploiting Farkas' lemma, we show that (a subset of) the closure of $\mathbf{P}_0$ with respect to the total variation metric can be described in terms of a set of linear inequality restrictions involving the projection of $(A_1(P), \beta(P))$ onto the orthogonal complement of the column span of $A_0(P)$. Specifically, the characterization amounts to verifying if, for every unit vector, the minimum of a collection of linear inequalities is non-positive.

Building on this observation, in Section 4 we develop an inference procedure based on sample splitting. Using the first subsample, we construct a unit vector for which we expect that our characterization of the null is most strongly violated. We then use the second subsample to formally test whether or not our inequalities are violated. By virtue of this sample-splitting construction, the resulting test is straightforward to implement and involves no simulation: in particular, the construction of the "violating" unit vector requires solving two linear programs, and the test statistic is provided in closed form with rejection based on a comparison with the appropriate quantile of a standard normal distribution. Moreover, the uniform asymptotic validity of our test is established under weak and interpretable regularity conditions while allowing the dimensions of p, d_0, d_1 to grow with the sample size n.

We propose two variants of our test. The first, which we term the "direct" method, tests if the union of *all* the inequalities in our characterization hold. The second, which we term the "screening" method, tests whether or not a *single* inequality in our characterization is non-positive, under the restriction that the other inequalities are positive with high probability. In our simulation study, the screening method typically yields shorter confidence intervals through test inversion, at the cost of introducing one tuning parameter when selecting the unit vector in the first subsample.

Inference procedures for testing (1) and closely related null hypotheses have gained increasing attention in the literature. Bai et al. (2022), Andrews et al. (2023), Cox and Shi (2023), and Fang et al. (2023) all propose inference procedures which could be applied to (1) whenever $A_0(P)$ and $A_1(P)$ do not depend on P (i.e. they are known quantities). The method proposed in our paper immediately applies to this setting as a special case. There is also a literature for the closely related problem of inference for the *value* of a linear program: see in particular Freyberger and Horowitz (2015), Cho et al. (2024), Gafarov (2025), Voronin (2025), Goff and Mbakop (2025). As explained in Cox et al. (2025), these methods can be used to address the same empirical problems as those that we consider in Section 2, but there is a subtle technical difference between the two settings, and in general these methods are neither tuning parameter nor simulation-free.

Two recent papers that propose inference procedures which could be used to test (1) are Cox et al. (2025) and Goff and Mbakop (2025). The theoretical results in Cox et al. (2025) and Goff and Mbakop (2025) require rank conditions which, as argued in Liu (2025), could be difficult to verify or may even be violated in certain settings of empirical interest. In contrast, we are able to establish the asymptotic validity of our procedures under weak and interpretable regularity conditions on the ranks of $A_0(P)$ and $A_1(P)$. Moreover, neither paper establishes the validity of their tests in a high-dimensional regime where p, d_1 and d_0 are allowed to grow with sample size, as we do in this paper.[1] We illustrate the finite-sample performance of our procedures using simulation designs based on the ones considered previously in these papers, as well as some high-dimensional counterparts that these papers do not consider.

The remainder of the paper is organized as follows. In Section 2, we describe several examples of empirical problems of interest that can be accommodated in our framework. Our main results are contained in Sections 3 and 4: Section 3 presents our characterization of the closure of the null hypothesis, whereas Section 4 describes the test and establishes its uniform asymptotic validity. In Section 5, we study the finite-sample behavior of our proposed tests in a simulation study. Proofs of all results are collected in the Appendix.

2 Examples

In this section, we present a collection of motivating examples, two of which we revisit in the simulation study in Section 5. Goff and Mbakop (2025) and Cox et al. (2025) discuss several additional examples that provide further motivation.

[1] We note that Goff and Mbakop (2025) explain that they derive their results in a semi high-dimensional regime, where the number of unknown coefficients in the linear system must remain bounded.

Example 2.1 (Nonparametric instrumental variables with shape restrictions)**.** Consider the NPIV model studied in Freyberger and Horowitz (2015), in which

$$\begin{aligned} Y &= g(X) + U \ , \\ E_P[U|W] &= 0 \ , \end{aligned} \tag{3}$$

with X a possibly endogenous explanatory variable supported on $\mathcal{X} = \{x_1, \ldots, x_H\}$, and W an instrumental variable supported on $\mathcal{W} = \{w_1, \ldots, w_K\}$ with $K < H$. Define

$$\begin{aligned} \pi_{hk}(P) &= P\{X = x_h, W = w_k\} \\ m_k(P) &= E_P[Y 1\{W = w_k\}] \\ \Pi(P) &= (\pi_{hk}(P))_{1 \le h \le H, 1 \le k \le K} \\ m(P) &= (m_1(P), \ldots, m_K(P))' \ . \end{aligned}$$

Then, $g = (g(x_1), \ldots, g(x_H)) \in \mathbb{R}^H$ satisfies $\Pi(P)'g = m(P)$. Assume additionally that the vector g satisfies shape constraints encoded as $Sg \le 0$ for some matrix $S \in \mathbb{R}^{M \times H}$.

Suppose we wish to test the null hypothesis $H_0 : L(g) = L_0 \in \mathbb{R}$, where $L(g)$ represents a generic linear functional $L(g) = c'g$. Putting everything together, we obtain the following system:

$$A_0(P) = \begin{pmatrix} \Pi(P)' \\ S \\ c' \end{pmatrix} \qquad A_1(P) = \begin{pmatrix} \mathbf{0}_{K \times M} \\ \mathbf{I}_M \\ \mathbf{0}_{1 \times M} \end{pmatrix} \qquad \beta(P) = \begin{pmatrix} m(P) \\ \mathbf{0}_{M \times 1} \\ L_0 \end{pmatrix} \ .$$

In this example, $A_1(P)$ is known and $A_0(P)$ and $\beta(P)$ are unknown. ■

Example 2.2 (Average marginal effects in nonlinear models with random coefficients)**.** Fox et al. (2011) consider a class of nonlinear mixture models with discrete unobserved heterogeneity. A simple example is a static, binary choice logit model with random coefficients:

$$Y = 1\{C'W - U \ge 0\} \ ,$$

where $Y \in \{0, 1\}$ is an observed choice, W is a vector of observed explanatory variables, C is a vector of latent random coefficients, and U is a latent random variable that follows a standard logistic distribution, independently of (C, W). Under these assumptions, a consumer of type c with observables w chooses $Y = 1$ with probability

$$P\{Y = 1|W = w, C = c\} = \frac{1}{1 + \exp(-c'w)} := \ell(c'w) \ , \tag{4}$$

where $\ell(\cdot)$ is the standard logistic distribution function. Bajari et al. (2007) and Fox et al. (2011) assume C is independent of W and approximate the distribution of C using a discrete distribution with known support points $(c_1, \ldots, c_d)$ and unknown respective probabilities $\pi := (\pi_1, \ldots, \pi_d)$. Then (4) implies observed

probabilities

$$P\{Y = 1|W = w\} = \sum_{j=1}^{d} \pi_j \ell(c_j' w) \ . \tag{5}$$

A target parameter in this model is the average marginal effect (AME) of the kth explanatory variable (for example, Wooldridge, 2010, Section 2.2.5):

$$\alpha(P) := E_P\left[\frac{\partial}{\partial w_k}\ell(C'W)\right] = E_P[C_k \ell(C'W)(1 - \ell(C'W))] = \sum_{j=1}^{d} \pi_j \underbrace{c_{jk} E_P[\ell(c_j'W)(1 - \ell(c_j'W))]}_{\alpha_j(P)} \ , \tag{6}$$

where c_{jk} is the kth component of the jth support point, c_j.

Suppose we wish to test the null hypothesis $H_0 : \alpha(P) = \alpha_0$ for a vector of probabilities π that satisfies (5) at $p-2$ support points $w_1, \ldots, w_{p-2}$. This problem fits into form (2) with x_0 null, $x_1 = \pi$,

$$A_1(P) = \begin{pmatrix} \ell(c_1'w_1) & \cdots & \ell(c_d'w_1) \\ \vdots & \vdots & \vdots \\ \ell(c_1'w_{p-2}) & \cdots & \ell(c_d'w_{p-2}) \\ 1 & \cdots & 1 \\ \alpha_1(P) & \cdots & \alpha_d(P) \end{pmatrix} \qquad \beta(P) = \begin{pmatrix} P\{Y = 1|W = w_1\} \\ \vdots \\ P\{Y = 1|W = w_{p-2}\} \\ 1 \\ \alpha_0 \end{pmatrix} \ .$$

Notice that the dependence of $A_1(P)$ on P comes through the row corresponding to the AME. The same structure will generally appear in a linear program with known coefficients (Fang et al., 2023) when the target parameter is an object that averages over observed heterogeneity. ■

Example 2.3 (Instrumental variables with heterogeneous treatment effects)**.** Mogstad et al. (2018) develop an approach to marginal treatment effect analysis (Heckman and Vytlacil, 2005) that allows for shape constraints and partial identification. A binary treatment D produces two real-valued potential outcomes $Y(0)$, $Y(1)$, and an observed outcome $Y = (1 - D)Y(0) + DY(1)$. The researcher additionally observes an instrument Z. Treatment assignment satisfies the Imbens and Angrist (1994) monotonicity condition, which can be equivalently written with the threshold-crossing model $D = 1\{p(Z) \geq U\}$, where U is a uniformly distributed unobservable that is independent of Z and $p(z) := P\{D = 1|Z = z\}$ is the propensity score (Vytlacil, 2002). The marginal treatment response functions are assumed to take a linear-in-parameters form

$$E[Y(d)|U = u] = \theta(d)' b(d|u), \tag{7}$$

where $\theta(d)$ are unknown parameters and $b(d|u)$ are known basis functions.

The linear parameterization (7) implies that common target parameters can also be written as linear

functions of the parameters, $\theta := (\theta(0), \theta(1))$, taking the general form

$$\tau^\star = \theta(0)' E\left[\int_0^1 b(0|u)\omega^\star(0|u,Z)du\right] + \theta(1)' E\left[\int_0^1 b(1|u)\omega^\star(1|u,Z)du\right] = \sum_{d\in\{0,1\}} \theta(d)' E[t(d|Z)], \quad (8)$$

where $\omega^\star(d|u,z)$ are scalar weights that are known or identified and $t(d|z) = \int_0^1 b(d|u)\omega^\star(d|u,z)du$. Mogstad and Torgovitsky (2024) observe that if $Y(d)$ is mean independent of Z, conditional on U, then the linear-in-parameters form (7) also has implications for the observed outcome:

$$E_P[Y|D,Z] = \phi(D,Z)'\theta, \quad (9)$$

where $\theta := (\theta(0), \theta(1))$ and ϕ is a known function of (D,Z) that depends on the basis functions, b, and the propensity score, $p(Z)$. Bounds on $\tau^\star$ can be found by considering all values of (8) that can be produced by θ that satisfy (9) either for all (D,Z) or for some implied moments. Mogstad et al. (2018) propose using moments of the form $E_P[Ys(D,Z)]$, which can be shown from (9) to also be linear in θ. Shea and Torgovitsky (2023) alternatively propose using the normal equations implied by (9):

$$E_P[\phi(D,Z)\phi(D,Z)']\theta = E_P[\phi(D,Z)Y], \quad (10)$$

which has the advantage of always being the same dimension as θ and not requiring one to choose the s functions. In either case, shape constraints can be imposed on the marginal treatment response functions by constraining θ.

Consider testing whether $\tau^* = \tau_0$ for $\tau_0 \in \mathbb{R}$. Without shape constraints, the normal equation version of the problem can be phrased as (1) by taking $x_0 = \theta$, then setting

$$A_0(P) = \begin{pmatrix} E_P[\phi(D,Z)\phi(D,Z)'] \\ E_P[t(Z)]' \end{pmatrix} \quad \text{and} \quad \beta(P) = \begin{pmatrix} E_P[\phi(D,Z)Y] \\ \tau_0 \end{pmatrix}, \quad (11)$$

where $t(Z) = (t(0|Z)', t(1|Z)')'$. Shape constraints can be imposed by including appropriate slack variables. For example, if $Y \in \{0,1\}$ is binary and $b(d|u)$ are Bernstein polynomials, then the implied MTR function can be constrained to lie in $[0,1]$ by restricting all elements of θ to lie within $[0,1]$. To incorporate these shape constraints into (1), we would now let x_0 be null and set $x_1 = [\theta', s']'$, where s are slack variables. Instead of (11), we would take A_0 to be empty and set

$$A_1(P) = \begin{pmatrix} E_P[\phi(D,Z)\phi(D,Z)'] & 0_{d_\theta \times d_\theta} \\ E_P[t(Z)]' & 0_{1\times d_\theta} \\ \mathbf{I}_{d_\theta} & \mathbf{I}_{d_\theta} \end{pmatrix} \quad \text{and} \quad \beta(P) = \begin{pmatrix} E_P[\phi(D,Z)Y] \\ \tau_0 \\ 1_{d_\theta \times 1} \end{pmatrix},$$

where $0_{d_\theta \times d_\theta}$ is a d_θ-dimensional square matrix of zeros, $\theta_{1\times d_\theta}$ is a d_θ-dimensional row vector of zeros, $1_{d_\theta \times 1}$ is a d_θ-dimensional column vector of ones, and $\mathbf{I}_{d_\theta}$ is a d_θ-dimensional identity matrix. The new rows relative to (11) correspond to the constraint $\theta + s \leq 1$, which requires $\theta \leq 1$ because the slack variable s is non-negative. ■

Example 2.4 (Synthetic parallel trends with convex weights)**.** Consider the causal panel data setting presented in Liu (2025). There are K aggregate units indexed by $k \in \{1, \ldots, K\}$, observed over periods $t \in \{1, \ldots, T_0, T\}$, where $\{1, \ldots, T_0\}$ denote pre-treatment periods and T denotes a treatment period at which unit $k = 1$ is treated (units $k \geq 2$ are never treated). Let $\tilde{\mu}_t^k(1)$ and $\tilde{\mu}_t^k(0)$ denote potential aggregate outcomes for unit k at time t with and without treatment, respectively, and let the observed aggregate outcome be given by

$$\mu_t^k(P) = \tilde{\mu}_t^k(0) + \big(\tilde{\mu}_t^k(1) - \tilde{\mu}_t^k(0)\big) 1\{k = 1, t = T\} \ .$$

The target parameter is the effect on the treated unit at time T:

$$\tau = \tilde{\mu}_T^1(1) - \tilde{\mu}_T^1(0) \ .$$

Liu (2025) maintains the assumption of (convex) *synthetic parallel trends* (SPT); that is, there exists a set of weights $(\omega_k : 2 \leq k \leq K) \in \mathbb{R}^{K-1}$ with $\sum_{2 \leq k \leq K} \omega_k = 1$, $\omega_k \geq 0$ for all k such that for every $t \in \{2, \ldots, T\}$,

$$\sum_{k=2}^{K} \omega_k \, \Delta\tilde{\mu}_t^k(0) = \Delta\tilde{\mu}_t^1(0) \ ,$$

where $\Delta\tilde{\mu}_t^k(0) = \tilde{\mu}_t^k(0) - \tilde{\mu}_{t-1}^k(0)$. Let $\Delta\mu_t^k(P) = \Delta\tilde{\mu}_t^k(0)$ for all $k \geq 2$ and $t \geq 2$. Suppose we wish to test the null hypothesis $H_0 : \tau = \tau_0$. Then, under the convex SPT assumption, we obtain the following system:

$$A_1(P) = \begin{pmatrix} \Delta\mu_2^2(P) & \cdots & \Delta\mu_2^K(P) \\ \vdots & \ddots & \vdots \\ \Delta\mu_{T_0}^2(P) & \cdots & \Delta\mu_{T_0}^K(P) \\ \Delta\mu_T^2(P) & \cdots & \Delta\mu_T^K(P) \\ 1 & \cdots & 1 \end{pmatrix} \qquad \beta(P) = \begin{pmatrix} \Delta\mu_2^1(P) \\ \vdots \\ \Delta\mu_{T_0}^1(P) \\ \mu_T^1(P) - \tau_0 - \mu_{T_0}^1(P) \\ 1 \end{pmatrix} ,$$

and $A_0(P)$ does not exist. Liu (2025) also analyzes the setting where we drop the assumption of convexity, so that ω_k are not restricted to be non-negative. In this case, $A_0(P)$ is given by the above matrix of aggregate-outcome differences and $A_1(P)$ does not exist. ■

Example 2.5 (Distribution-free binary choice)**.** Gu and Russell (2023) consider binary choice models of the form

$$Y = 1\{\varphi(D, Z, U) \geq 0\} \ , \tag{12}$$

where Y is a binary outcome, φ is an unknown function, D is an endogenous regressor, Z is vector of exogenous regressors, and U is a vector of unobservables. The distribution of U is not restricted to lie in a parametric family, raising the possibility of partial identification. The authors observe that if D and Z are discrete, then the conditional distribution of Y implied by the model is determined by the mass placed on a

finite partition of the support of U into sets $\mathcal{U}_j$, $j = 1, \ldots, d_U$. In particular:

$$P\{Y = 1|D = d, Z = z\} = \sum_{j=1}^{d_U} 1\{j \in \mathcal{J}(d, z)\}\, \theta_j(d, z) \ , \tag{13}$$

where $\theta_j(d, z)$ is the mass that the distribution of U places on $\mathcal{U}_j$, conditional on $D = d, Z = z$, and the set $\mathcal{J}(d, z)$ collects the appropriate indices for sets that lead to $Y = 1$ when $D = d$ and $Z = z$. If the instrument Z is independent with U, then also

$$\sum_d \theta_j(d, z) P\{D = d|Z = z\} = \sum_d \theta_j(d, z') P\{D = d|Z = z'\} \quad \text{for all } z, z', \text{ and } j. \tag{14}$$

A natural target parameter in this model is the counterfactual choice probability $\pi(P) := P\{Y(d^\star) = 1\} = P\{\varphi(d^\star, Z, U) \geq 0\}$ at some fixed $d^\star$, which can also expressed as a linear function of the $\theta_j(d, z)$ if the sets $\mathcal{U}_j$ have been constructed to be sufficiently fine:

$$\pi(P) = \sum_{j=1}^{d_U} E_P\left[1\{j \in \mathcal{J}_\pi(d^\star, Z)\}\, \theta_j(D, Z)\right] = \sum_{j=1}^{d_U} \sum_{d,z} 1\{j \in \mathcal{J}_\pi(d^\star, z)\}\, \theta_j(d, z) P\{D = d, Z = z\} \ . \tag{15}$$

Similar observations have been used for multinomial choice models by Manski (2007), Tebaldi et al. (2023), and Gu et al. (2024).

Suppose we wish to test the null hypothesis $H_0 : \pi(P) = \pi_0$ for a vector of probabilities $\{\theta_j(d, z)\}_{j,d,z}$ that satisfies (13) and (14) when D and Z are both discrete. This problem fits into form (2) with x_0 null, x_1 taken to be the $\theta_j(d, z)$ arranged in a vector across (j, d, z), and $A_1(P)$ and $\beta(P)$ constructed from the linear functions (13)–(15), with (15) set equal to π_0 and additional sum-to-one constraints for each (d, z). The rows of $A_1(P)$ corresponding to (14)–(15) both depend on P. ■

3 A Useful Characterization of $\mathbf{P}_0$

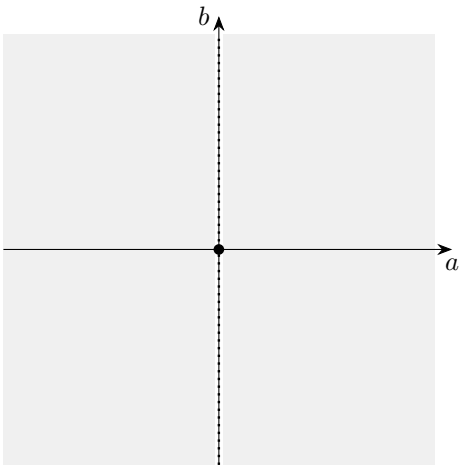
In this section, we provide a characterization of $\mathbf{P}_0$ that informs the construction of the test we present in Section 4. Before discussing the characterization formally, we motivate the need for such a characterization by demonstrating the impossibility of testing the null hypothesis that $P \in \mathbf{P}_0$ in a seemingly simple example. In particular, consider the case in which, for some scalars $a(P)$ and $b(P)$, the set $\mathbf{P}_0$ is given by

$$\mathbf{P}_0 = \{P \in \mathbf{P} : a(P)x = b(P) \text{ for some } x \in \mathbb{R}\} \ .$$

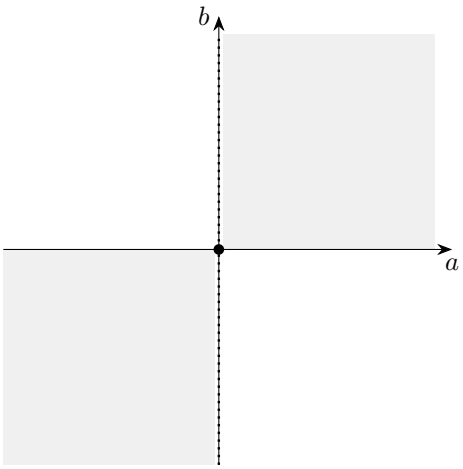
This is a special case of (2) where $d_0 = 1$ and $A_1(P)$ does not exist. In Figure 1(i) we plot the set

$$\mathbf{C}_0 = \{(a, b) \in \mathbb{R}^2 : ax = b \text{ for some } x \in \mathbb{R}\} \ ,$$

which represents the set of points $(a(P), b(P)) \in \mathbb{R}^2$ for which there exists a distribution $P \in \mathbf{P}_0$. From Figure 1(i) we notice immediately that the closure of $\mathbf{C}_0$ (as a subset of $\mathbb{R}^2$) is the entire space $\mathbb{R}^2$. This



(i) $H_0 : ax = b$ for some $x \in \mathbb{R}$



(ii) $H_0 : ax = b$ for some $x \geq 0$

Figure 1: Shaded region indicates the null set in the (a, b)-plane. The dotted line at $a = 0$ highlights points excluded from the null (except $(0, 0)$) but belonging to its closure.

observation suggests that the closure of $\mathbf{P}_0$ (with respect to the total variation metric) coincides with the entire set of distributions $\mathbf{P}$, and indeed we discuss conditions under which this is the case below. In contrast, Figure 1(ii) depicts the analogous set $\mathbf{C}_0$ for testing the null hypothesis given by

$$\mathbf{P}_0 = \{P \in \mathbf{P} : a(P)x = b(P) \text{ for some } x \in \mathbb{R}, x \geq 0\} \ ,$$

which is a special case of (2) where $d_1 = 1$ and $A_0(P)$ does not exist. Here, we see that the closure of $\mathbf{C}_0$ (as a subset of $\mathbb{R}^2$) is a strict subset of $\mathbb{R}^2$, which suggests that the closure of $\mathbf{P}_0$ in this example is a strict subset of $\mathbf{P}$.

Because it is impossible to test a null hypothesis for which $\mathbf{P}_0$ is dense in $\mathbf{P}$ with respect to the total variation metric, and more generally that it is impossible for any test to have non-trivial power against alternatives which lie on the *boundary* of $\mathbf{P}_0$ (see for instance Romano, 2004), our test is based on a characterization of the closure of $\mathbf{P}_0$ relative to the total variation metric, which we denote by $\text{cl}(\mathbf{P}_0)$. Towards that end, define

$$\mathbf{C}_0 := \left\{(A_0, A_1, b) \in \mathbb{R}^{p \times d_0} \times \mathbb{R}^{p \times d_1} \times \mathbb{R}^p : A_0 x_0 + A_1 x_1 = b \text{ for some } x_0 \in \mathbb{R}^{d_0}, x_1 \in \mathbb{R}^{d_1}, x_1 \geq 0\right\} \ , \quad (16)$$

so that $\mathbf{P}_0 = \{P \in \mathbf{P} : (A_0(P), A_1(P), \beta(P)) \in \mathbf{C}_0\}$. Accordingly, we begin by deriving a characterization of the closure of $\mathbf{C}_0$ with respect to the Euclidean topology, and then relate this characterization back to $\text{cl}(\mathbf{P}_0)$. First, consider the following alternative representation of $\mathbf{C}_0$ based on pre-multiplying the equation in (16) by the annihilator of A_0, which we denote by M_0.

Lemma 3.1. *Let M_0 denote the projection operator onto the orthogonal complement of the column space of A_0. Then*

$$\mathbf{C}_0 = \{(A_0, A_1, b) : M_0 A_1 x_1 = M_0 b \text{ for some } x_1 \in \mathbb{R}^{d_1}, x_1 \geq 0\} \ .$$

Let $a_1, \ldots, a_{d_1}$ denote the columns of A_1. The columns of $M_0 A_1$ are then given by $M_0 a_1, \ldots, M_0 a_{d_1}$. Given this alternative representation of $\mathbf{C}_0$, we can apply Farkas' lemma to conclude that $(A_0, A_1, b) \in \mathbf{C}_0$ if and only if for all $y \in \mathbb{R}^p$, either there exists some $1 \leq j \leq d_1$ such that $a_j' M_0 y < 0$ or $b' M_0 y \geq 0$. Consider the set of triples (A_0, A_1, b) obtained by weakening the strict inequalities $a_j' M_0 y < 0$ to weak inequalities:

$$\bar{\mathbf{C}}_0 := \left\{ (A_0, A_1, b) \in \mathbb{R}^{p \times d_0} \times \mathbb{R}^{p \times d_1} \times \mathbb{R}^p : \sup_{y \in \mathbb{R}^p} \min \left\{ \min_{1 \leq j \leq d_1} a_j' M_0 y, -b' M_0 y \right\} \leq 0 \right\} . \tag{17}$$

Theorem 3.1 formalizes the sense in which replacing these strict inequalities with weak inequalities relates to the closure of the set $\mathbf{C}_0$.

Theorem 3.1. *Let*

$$\mathbf{C}^{\mathrm{RD}} := \{ (A_0, A_1, b) \in \mathbb{R}^{p \times d_0} \times \mathbb{R}^{p \times d_1} \times \mathbb{R}^p : \mathrm{rank}(A_0) < d_0 \} . \tag{18}$$

Then,

$$\mathrm{cl}(\mathbf{C}_0) = \bar{\mathbf{C}}_0 \cup \mathbf{C}^{\mathrm{RD}} ,$$

where $\mathrm{cl}(\mathbf{C}_0)$ *denotes the closure of* $\mathbf{C}_0$ *in the Euclidean topology.*

Theorem 3.1 shows that the closure of $\mathbf{C}_0$ can be characterized by combining the set $\bar{\mathbf{C}}_0$, which describes the set of triples obtained by weakening the inequalities in the conclusion of Farkas' lemma, along with the set of triples for which A_0 is rank deficient. Note the theorem implies that, if $p < d_0$, then the closure of $\mathbf{C}_0$ becomes $\mathbb{R}^{p \times d_0} \times \mathbb{R}^{p \times d_1} \times \mathbb{R}^p$. As a result, we implicitly assume that $p \geq d_0$ for the rest of the paper. This characterization of the closure forms the basis for the test which we present in Section 4.

Finally, we relate $\mathrm{cl}(\mathbf{C}_0)$ to the closure $\mathrm{cl}(\mathbf{P}_0)$ in the total variation distance. Let $\widetilde{\mathbf{P}}_0 := \{P \in \mathbf{P} : (A_0(P), A_1(P), \beta(P)) \in \mathrm{cl}(\mathbf{C}_0)\}$ denote the pre-image of $\mathrm{cl}(\mathbf{C}_0)$ in $\mathbf{P}$. We claim that in general $\mathbf{P}_0 \subseteq \widetilde{\mathbf{P}}_0 \subseteq \mathrm{cl}(\mathbf{P}_0)$. Indeed, it follows by construction that $\mathbf{P}_0 \subseteq \widetilde{\mathbf{P}}_0$. As a result, any test that controls size on $\widetilde{\mathbf{P}}_0$ will necessarily control size on $\mathbf{P}_0$. Meanwhile, we generally expect $\widetilde{\mathbf{P}}_0 \subseteq \mathrm{cl}(\mathbf{P}_0)$, in which case we will not have power against any distribution in $\widetilde{\mathbf{P}}_0$. By definition, this will be the case whenever $\mathbf{P}$ is "rich" enough in the sense that for every $P \in \mathbf{P}$ such that $(A_0(P), A_1(P), \beta(P)) \in \mathrm{cl}(\mathbf{C}_0)$, there exists a sequence $P_n \in \mathbf{P}_0$ such that P_n converges to P in the total variation metric.

Low-level sufficient conditions for this "richness" property could be obtained, for example, if we view P as the distribution of a random vector in $\mathbb{R}^{p(d_0+d_1+1)}$ and define $\mathrm{vec}(A_0(P), A_1(P), \beta(P))$ to be the corresponding vector of means, where vec is the vec-operator (see, for instance, Chapter 2 of Magnus and Neudecker, 2019). In this case, $\mathbf{P}$ is rich enough if it contains a sufficiently large collection of normal location families. To illustrate, let $\mu = \mathrm{vec}(A_0(P), A_1(P), \beta(P))$ where $(A_0(P), A_1(P), \beta(P)) \in \mathrm{cl}(\mathbf{C}_0)$. Then by definition there exists a sequence of vectors μ_n corresponding to triplets in $\mathbf{C}_0$ such that $\mu_n \to \mu$. Let $H = \mathrm{span}\{\mu_n - \mu : n \geq 1\}$ and let Σ be a positive semi-definite matrix whose range is exactly H, so that $\mu_n - \mu \in \mathrm{range}(\Sigma)$ for all $n \geq 1$. Then, the sequence of distributions $N(\mu_n, \Sigma)$ converges to $N(\mu, \Sigma)$ in the total variation metric by Lemma C.7 in the appendix.

Remark 3.1. Following Fang et al. (2021), it may seem natural to first transform (2) into standard form

$$\mathbf{P}_0 = \{P \in \mathbf{P} : (A(P), \beta(P)) \in \mathbf{C}_0^{\text{alt}}\} \ , \tag{19}$$

where $\mathbf{C}_0^{\text{alt}} = \{(A, b) \in \mathbb{R}^{p\times(2d_0+d_1)} \times \mathbb{R}^p : Ax = b \text{ for some } x \in \mathbb{R}^{2d_0+d_1}, x \geq 0\}$. Indeed, given a triple $(A_0, A_1, b) \in \mathbf{C}_0$, we can obtain a pair $(A, b) \in \mathbf{C}_0^{\text{alt}}$ by defining $A = (A_0 \quad -A_0 \quad A_1)$. However, this transformation would not help provide a useful characterization of the closure as presented in this section. Let $\tilde{a}_j$ for $1 \leq j \leq 2d_0 + d_1$ denote the columns of A. By applying the reasoning we used to obtain $\bar{\mathbf{C}}_0$ in (17), we obtain the set

$$\bar{\mathbf{C}}_0^{\text{alt}} = \left\{(A, b) \in \mathbb{R}^{p\times(2d_0+d_1)} \times \mathbb{R}^p : \sup_{y\in\mathbb{R}^p} \min \left\{ \min_{1\leq j\leq 2d_0+d_1} \tilde{a}_j' y, -b'y \right\} \leq 0 \right\} \ .$$

Note that the $d_0 + 1$ to $2d_0$-th columns of A are simply the negatives of the first d_0 columns, so that there always exists a j for which $\tilde{a}_j' y \leq 0$. As a result, $\bar{\mathbf{C}}_0^{\text{alt}}$ recovers the *entire* space $\mathbb{R}^{p\times(2d_0+d_1)} \times \mathbb{R}^p$ and thus does not help to provide a useful characterization of the closure. ■

4 The Test

4.1 Description of the Test

Let $\{Z_i\}_{i=1}^n$ be i.i.d. with Z_i distributed according to $P \in \mathbf{P}$. Recall from Section 3 that the closure of the null space can be characterized as the set of distributions P for which either $A_0(P)$ is rank-deficient, or

$$\sup_{y\in\mathbb{R}^p} \min \left\{ \min_{1\leq j\leq d_1} a_j(P)' M_0(P) y, -\beta(P)' M_0(P) y \right\} \leq 0 \ . \tag{20}$$

Recognizing that whether condition (20) holds is not affected by norm constraints on y, and defining $b_j(P) := a_j(P)$ for $1 \leq j \leq d_1$, $b_{d_1+1}(P) := -\beta(P)$, and $J = \{1, 2, \ldots, d_1 + 1\}$, we may rewrite (20) as

$$\min_{j\in J} \ b_j(P)' M_0(P) y \leq 0 \qquad \text{for all } y \in \mathbb{R}^p \text{ such that } \|y\|_1 \leq 1 \ , \tag{21}$$

where $\|y\|_1 = \sum_{i=1}^p |y_i|$ for any $(y_1, \ldots, y_p) \in \mathbb{R}^p$. In what follows, we consider the following equivalent formulation of (21): For any $J^* \subseteq J$ and $(J^*)^c := J \setminus J^*$, condition (21) is equivalent to the statement

$$\min_{j\in J^*} b_j(P)' M_0(P) y \leq 0 \qquad \text{for all } y \in \mathbb{R}^p \text{ such that } \|y\|_1 \leq 1 \text{ and } \min_{j\in(J^*)^c} b_j'(P) M_0(P) y > 0 \ . \tag{22}$$

Our test uses a sample-splitting procedure in which one sample split is used to select a $y \in \mathbb{R}^p$ and a second split is used to test whether the inequalities in (22) hold at the selected y. We consider two proposals for $(J^*)^c$. Our first proposal takes $(J^*)^c$ to be those $j \in J$ for which $b_j(P)' M_0(P)$ is known deterministically – i.e. for which $b_j(P)' M_0(P)$ does not depend on P. In other words, we test the condition $\min_{j\in J^*} b_j(P)' M_0(P) y \leq 0$ using a unit vector y that is known to satisfy $\min_{j\in(J^*)^c} b_j(P)' M_0(P) y > 0$.

We call this method the "direct" method in what follows. Our second proposal sets $J^* = \{j^*\}$ for some non-random j^*. In this case, we test the condition $b_{j^*}(P)'M_0(P)y \leq 0$ using a unit vector y such that $\min_{j\in(J^*)^c} b_j(P)'M_0(P)y > 0$ holds with high probability. We call this method the "screening" method. In all of our examples, we set $j^* = d_1 + 1$, which is natural in settings where we perform test inversion to construct a confidence set for a scalar parameter whose null value only enters the vector $b_{d_1+1}(P)$; see, e.g., the examples in Section 2. We show via simulation in Section 5 that the screening method often generates shorter confidence intervals than the direct method, at the cost of introducing an additional tuning parameter which determines the amount of "screening" that is performed in the first sample split.

We next present a high-level description of the test and defer the details of the construction of its specific components to Sections 4.2 and 4.3. To construct the test, we first randomly split the data into two samples $\{Z_i\}_{i\in I_{1,n}}$, $\{Z_i\}_{i\in I_{2,n}}$ of sizes n_1 and n_2, where $I_{1,n} \cup I_{2,n} = \{1, \ldots, n\}$ and $I_{1,n} \cap I_{2,n} = \emptyset$. In what follows, we always assume that $n_2 \to \infty$ as $n \to \infty$ and allow n_1 to be fixed for the direct method but require $n_1 \to \infty$ for the screening method. Throughout, we use the superscript (k) to denote when a given quantity is a function of only the kth split. Using the first sample split $\{Z_i\}_{i\in I_{1,n}}$, we construct a vector $\hat{y}_n^{(1)}$ which represents a direction in which the weak inequality in (22) appears to be "most violated" — we discuss how to construct such a vector in Section 4.3.

Next, given suitable estimators $\hat{b}_{j,n}^{(2)}$ and $\hat{M}_{0,n}^{(2)}$ for $b_j(P)$ and $M_0(P)$ computed in the second sample split $\{Z_i\}_{i\in I_{2,n}}$, we define the test statistic

$$T_n := \min_{j\in J^*} \frac{\sqrt{n_2}(\hat{b}_{j,n}^{(2)})'\hat{M}_{0,n}^{(2)}\hat{y}_n^{(1)}}{\hat{\sigma}_{j,n}^{(2)}(\hat{y}_n^{(1)})} , \tag{23}$$

where $\hat{\sigma}_{j,n}^{(2)}(\hat{y}_n^{(1)})$ is an estimator for the asymptotic standard deviation of $\sqrt{n_2}(\hat{b}_{j,n}^{(2)})'\hat{M}_{0,n}^{(2)}\hat{y}_n^{(1)}$. Finally, we set

$$\phi_n := 1\{T_n > z_{1-\alpha}\} , \tag{24}$$

where, for $\alpha \in (0, 1)$, $z_{1-\alpha}$ denotes the $1-\alpha$ quantile of a standard normal distribution.

Remark 4.1. It may occur that for some $j \in J^*$ the asymptotic variance of $\sqrt{n_2}(\hat{b}_{j,n}^{(2)})'\hat{M}_{0,n}^{(2)}\hat{y}_n^{(1)}$ is zero. To avoid degeneracy of the corresponding standard error, we define $\hat{\sigma}_{j,n}^{(2)}(\hat{y}_n^{(1)})$ using a small truncation, as described in Section 4.2. This modification is technically motivated and typically has no effect on the test in practice. Indeed, if the numerator of (23) is negative for any $j \in J^*$, then we fail to reject for any choice of truncation. If the numerator of (23) is positive for all $j \in J^*$ then the choice of truncation can only induce a failure to reject if the estimated variance falls below the truncation threshold for at least one index that attains the minimum in the truncated version of T_n. ■

Remark 4.2. In practice, researchers may want to reduce the uncertainty introduced by sample splitting by aggregating the test results obtained from multiple different splits of the data. This can be accomplished by appropriately aggregating the (upper bounds on) p-values produced by the test described in (24). To that end, we have found the exchangeable improvement to the "twice the average" p-value, as described in Gasparin et al. (2025), works well in simulations. ■

4.2 Properties of the Test

In this section, we present results establishing the asymptotic validity of our test and a detailed construction of the standard deviation in (23). When stating our assumptions, we will suppress the superscript (k), with the understanding that all assumptions stated on the entire sample will also hold when applied on the sample splits $\{Z_i\}_{I_{1,n}}$ and $\{Z_i\}_{i\in I_{2,n}}$, under suitable scaling.

Our first assumption imposes conditions on the estimators $\hat{A}_{0,n}$ for $A_0(P)$ and $\hat{b}_{j,n}$ for $b_j(P)$ with $1 \leq j \leq d_1+1$. In its statement, $\|\cdot\|_2$ denotes the Euclidean norm, $\|\cdot\|_{2,2}$ denotes the operator norm of a matrix when the domain and range are endowed with $\|\cdot\|_2$, and $a \vee b = \max(a,b)$ for any $a, b \in \mathbb{R}$.

Assumption 4.1. Let $\{Z_i\}_{i=1}^n$ be i.i.d. with marginal distribution $P \in \mathbf{P}$. Then,

(a) There are $\Psi(Z_i, P) \in \mathbb{R}^{p\times d_0}$ with $E_P[\Psi(Z_i,P)] = 0$, $\varphi_j(Z_i,P) \in \mathbb{R}^p$ with $E_P[\varphi_j(Z_i,P)] = 0$ for $1 \leq j \leq d_1+1$, and a_n for which $a_n/\sqrt{n} \to 0$ such that uniformly in $P \in \mathbf{P}$,

$$\left\|\sqrt{n}(\hat{A}_{0,n} - A_0(P)) - \frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\Psi(Z_i,P)\right\|_{2,2} = O_P(a_n/\sqrt{n}) \tag{25}$$

$$\max_{1\leq j\leq d_1+1}\left\|\sqrt{n}(\hat{b}_{j,n} - b_j(P)) - \frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\varphi_j(Z_i,P)\right\|_2 = O_P(a_n/\sqrt{n})\ . \tag{26}$$

(b) For each $p \geq 1$, there are $1 \leq K_{0,p}, K_{1,p} < \infty$ such that $\|\Psi(Z_i,P)\|_{2,2} \leq K_{0,p}$ with probability one and

$$\sup_{P\in\mathbf{P}}\Big(\|E_P[\Psi(Z_i,P)\Psi(Z_i,P)']\|_{2,2} \vee \|E_P[\Psi(Z_i,P)'\Psi(Z_i,P)]\|_{2,2} \vee \max_{1\leq j\leq d_1+1} E_P[\varphi_j(Z_i,P)'\varphi_j(Z_i,P)]\Big) \leq K_{1,p}\ .$$

Assumption 4.1(a) requires our estimators for $A_0(P)$ and $b_j(P)$ for $1 \leq j \leq d_1+1$ to be asymptotically linear with influence functions whose moments are disciplined by Assumption 4.1(b). Assumption 4.1(a) is automatically satisfied with $a_n = 0$ whenever the entries of $A_0(P)$ and $b_j(P)$ are expectations and the entries of $\hat{A}_{0,n}$ and $\hat{b}_{j,n}$ are the corresponding sample means.

Our second assumption imposes boundedness and non-degeneracy conditions on $A_0(P)$ and $b_j(P)$. In its statement, $\bar{s}(A_0(P))$ and $\underline{s}(A_0(P))$ denote the maximum and minimum singular value of $A_0(P)$.

Assumption 4.2. $A_0(P)$ and $b_j(P)$ are such that

(a) $\sup_{P\in\mathbf{P}} \bar{s}(A_0(P)) \leq \bar{s}_p$ for some $\bar{s}_p$ satisfying $1 \leq \bar{s}_p < \infty$.

(b) $\inf_{P\in\mathbf{P}} \underline{s}(A_0(P)) \geq \underline{s} > 0$ for some $\underline{s}$ not depending on p.

(c) $\sup_{P\in\mathbf{P}} \max_{1\leq j\leq d_1+1} \|b_j(P)\|_2 \leq K_{2,p}$ for some $K_{2,p}$ satisfying $1 \leq K_{2,p} < \infty$.

Assumption 4.2(a) requires the maximum singular value of $A_0(P)$ is bounded above uniformly in $\mathbf{P}$, with the bound possibly depending on p. Assumption 4.2(b) ensures that $A_0(P)'A_0(P)$ is bounded away from degeneracy. This rules out the set of rank deficient matrices $\mathbf{C}^{\mathrm{RD}}$ in our characterization of the closure of the null hypothesis, as defined in Theorem 3.1. Assumption 4.2(c) requires that the Euclidean norm of $b_j(P)$ is bounded uniformly in $P \in \mathbf{P}$, with the bound possibly depending on p.

We note that Assumption 4.2(b) can be dropped if we let $n_1 \to \infty$ and apply the test ϕ_n in (24) only when $\underline{s}(\hat{A}^{(1)}_{0,n}) > \tau$ for some pre-specified small value $\tau > 0$ (and do not reject the null hypothesis when $s(\hat{A}^{(1)}_{0,n}) \leq \tau$). We emphasize, however, that even if we maintain Assumption 4.2(b), our assumptions impose no requirements on the rank of $A_1(P)$. In contrast, the assumptions underlying Cox et al. (2025) and Goff and Mbakop (2025) implicitly restrict the ranks of both $A_0(P)$ and $A_1(P)$.

Assumptions 4.1 and 4.2 are instrumental in obtaining an asymptotic expansion for our test statistic. In particular, letting $A_0^\dagger(P)$ denote the Moore-Penrose pseudoinverse of $A_0(P)$, we will show that for every $1 \leq j \leq d_1 + 1$ the vector $\sqrt{n}(\hat{b}'_{j,n}\hat{M}_{0,n} - b_j(P)'M_0(P))'$ is asymptotically linear with influence function

$$\xi_j(Z_i, P) := M_0(P)\varphi_j(Z_i, P) - M_0(P)\Psi(Z_i, P)A_0^\dagger(P)b_j(P) - A_0^\dagger(P)'\Psi(Z_i, P)'M_0(P)b_j(P) \ . \tag{27}$$

Our third assumption imposes moment restrictions on the influence function $\xi_j(Z, P)$.

Assumption 4.3. There is a constant $K_\xi < \infty$ not depending on p such that the following holds:

$$\sup_{P \in \mathbf{P}} \sup_{\|y\|_1 \leq 1} \max_{1 \leq j \leq d_1+1} E_P[|\xi_j(Z_i, P)'y|^3]^{1/3} \leq K_\xi.$$

Assumption 4.3 ensures that we are able to couple an influence function for our test statistic to a Gaussian random variable uniformly in $P \in \mathbf{P}$. The requirement of Assumption 4.3 is satisfied, for example, if the third moments of the entries of $\xi_j(Z, P)$ are uniformly bounded across $1 \leq j \leq d_1 + 1$ and $P \in \mathbf{P}$.

Our fourth assumption ensures that the inequalities not examined by our test statistic (i.e., those in $(J^*)^c$) are indeed positive when evaluated at a $\hat{y}_n^{(1)}$ not equal to zero, as required by the characterization of the null hypothesis in (22). We describe methods to construct such a $\hat{y}_n^{(1)}$ in Section 4.3 below.

Assumption 4.4. For $\mathcal{Y}(P; J^*) := \{y \in \mathbb{R}^p : \|y\|_1 \leq 1 \text{ and } b_j'(P)M_0(P)y > 0 \text{ for all } j \in (J^*)^c\}$, we have

$$\lim_{n \to \infty} \inf_{P \in \mathbf{P}_0} P\big\{\{\hat{y}_n^{(1)} \in \mathcal{Y}(P; J^*)\} \cup \{\hat{y}_n^{(1)} = 0\}\big\} = 1 \ .$$

Assumptions 4.4 automatically holds, for instance, for the direct method which either sets J^* to equal J (so $(J^*)^c$ is empty) or $(J^*)^c$ to only contains coordinates j for which $b_j'M_0(P)$ is known. In this case, $\hat{y}_n^{(1)}$ can be chosen to belong to $\mathcal{Y}(P; J^*)$ with probability one for any n_1, and our asymptotics only require that $n_2 \to \infty$. In contrast, the screening method intuitively conducts a pre-test in the first fold to ensure that $\hat{y}_n^{(1)} \in \mathcal{Y}(P; J^*)$ with high probability. In this case, our asymptotics therefore require $n_1 \to \infty$ in order for Assumption 4.4 to be satisfied. We also note that Assumption 4.4 allows $\hat{y}_n^{(1)}$ to equal zero when it does not belong to $\mathcal{Y}(P; J^*)$. This flexibility is important because our test never rejects when $\hat{y}_n^{(1)}$ is zero (since then $T_n = 0$; see (23)). Therefore, setting $\hat{y}_n^{(1)} = 0$ allows us to decide not to reject after examining the first sample split — e.g., if we fail to reject in the screening method pre-test.

We now turn to the construction of the standard error $\hat{\sigma}_{j,n}(y)$ employed in the construction of our test statistic. To this end, we let vec denote the vec-operator and $\otimes$ denote the Kronecker product (see, for

instance, Chapter 2 of Magnus and Neudecker, 2019). We further define the asymptotic covariance matrix

$$V_j(P) := \mathrm{Var}_P\left[\begin{pmatrix}\mathrm{vec}\,\Psi(Z_i,P)\\ \varphi_j(Z_i,P)\end{pmatrix}\right]$$

for our estimator $\hat{A}_{0,n}$ and $\hat{b}_{j,n}$ and let $\hat{V}_{j,n}$ denote an estimator for $V_j(P)$. For each fixed $y \in \mathbb{R}^p$ and $1 \le j \le d_1+1$, we apply the Delta method to the function $(A_0(P), b_j(P)) \mapsto b_j(P)'M_0(P)y$ to obtain the asymptotic variance of the estimator $\hat{b}_{j,n}'\hat{M}_{0,n}y$. Accordingly, defining the gradient

$$D_j(P;y) := \begin{pmatrix}-\Big(A_0^\dagger(P)y \otimes M_0(P)b_j(P) + A_0^\dagger(P)b_j(P) \otimes M_0(P)y\Big)\\ M_0(P)y\end{pmatrix},$$

it is possible to show that the asymptotic variance of $\hat{b}_{j,n}'\hat{M}_{0,n}y$ equals $\sigma_j^2(P;y) := D_j(P;y)'V_j(P)D_j(P;y)$. By analogy, for $\hat{A}_{0,n}^\dagger$ the Moore-Penrose pseudoinverse of $\hat{A}_{0,n}$, we estimate $D_j(P;y)$ by setting

$$\hat{D}_{j,n}(y) = \begin{pmatrix}-\big(\hat{A}_{0,n}^\dagger y \otimes \hat{M}_{0,n}\hat{b}_{j,n} + \hat{A}_{0,n}^\dagger \hat{b}_{j,n} \otimes \hat{M}_{0,n}y\big)\\ \hat{M}_{0,n}y\end{pmatrix}.$$

As an estimator for the asymptotic variance we then set $\hat{\sigma}_{j,n}^2(y) = (\hat{D}_{j,n}(y)'\hat{V}_{j,n}\hat{D}_{j,n}(y)) \vee \underline{\sigma}^2$ for $1 \le j \le d_1+1$, where, as previously discussed in Remark 4.1, $\underline{\sigma} > 0$ is a small positive constant which we include to address potential (near) degeneracies in $\sigma_j^2(P;y)$.

Our fifth assumption imposes that our estimator $\hat{V}_{j,n}$ is suitably uniformly consistent.

Assumption 4.5. $\|\hat{V}_{j,n} - V_j(P)\|_{2,2} = o_P(K_{2,p}^{-2})$ uniformly in $P \in \mathbf{P}$ and $1 \le j \le d_1+1$.

Assumption 4.5 enables us to show that the standard errors $\hat{\sigma}_{j,n}(y)$ are suitably uniformly consistent in both $P \in \mathbf{P}$ and $1 \le j \le d_1+1$. It requires that $\hat{V}_{j,n}$ converge to $V_j(P)$ at a rate faster than $K_{2,p}^2$, where $K_{2,p}$ depends on p and is specified in Assumption 4.2(c). We state Assumption 4.5 as a high level condition because the structure of $V_j(P)$ is dictated by the influence functions of our estimators (as introduced in Assumption 4.1). In applications for which our estimators are sample means, and hence $\hat{V}_{j,n}$ are sample covariance matrices, sufficient conditions for Assumption 4.5 are readily available from the literature; see, e.g., Chapter 6 in Wainwright (2019).

Our final assumption imposes conditions on the rates of convergence of our moment and error bounds.

Assumption 4.6. The following rate restrictions hold as $n \to \infty$:

(a) $K_{2,p}(K_{0,p} \vee K_{1,p})\log(1+p) = o(\sqrt{n_2})$.
(b) $\bar{s}_p^2(K_{0,p} \vee K_{1,p})\log(1+p) = o(n_2)$.
(c) $K_{2,p}a_{n_2} = o(\sqrt{n_2})$.

In Remark 4.3 below, we discuss high-level sufficient conditions which guarantee Assumption 4.6 holds, as well as how these conditions differ when $A_0(P)$ needs to be estimated versus when it is known (or does not

exist). At this point we emphasize, however, that Assumptions 4.6 are automatically satisfied in asymptotic regimes in which p and d_1 are fixed. Moreover, we note that Assumption 4.6 only restricts the dimension d_1 through Assumptions 4.1 and 4.2(c), which impose that the linearization error, the second moment of the influence function $\varphi_j(Z,P)$, and the norm of $\|b_j(P)\|_2$ be bounded uniformly in $1 \leq j \leq d_1$. If these bounds do not grow with d_1, then Assumption 4.6 in fact leaves the dimension d_1 unrestricted. As we discuss in the next section, however, certain approaches for selecting $\hat{y}_n^{(1)}$ may impose restrictions on the dimension d_1 relative to the sample size in the first split $\{Z_i\}_{i \in I_{n,1}}$.

Our next, main, result establishes that our proposed test is uniformly consistent in level over $\mathbf{P}_0$.

Theorem 4.1. *Suppose Assumptions 4.1–4.6 hold. Then, for any $\alpha < 0.5$ it follows that*

$$\limsup_{n\to\infty} \sup_{P \in \mathbf{P}_0} E_P[\phi_n] \leq \alpha \ .$$

Remark 4.3. While the constants specified by Assumptions 4.1 and 4.2 are application specific, in many instances we can expect the bounds $K_{0,p} \vee K_{1,p} \lesssim p$, $K_{2,p} \vee \overline{s}_p \lesssim \sqrt{p}$, and $a_n \lesssim p$ (or $a_n = 0$) to hold. Under these conditions, Assumption 4.6 holds provided that $p^3 = o(n_2)$ (up to logs), while Assumption 4.5 can also be shown to hold when $p^3 = o(n_2)$ (up to logs) under suitable moment restrictions. These rates simplify when the matrix $A_0(P)$ does not exist or does not depend on P. In this case, an inspection of the proof of Theorem 4.1 reveals that Assumptions 4.6(a)(b) are no longer needed, Assumption 4.6(c) can be weakened to $a_n = o(\sqrt{n_2})$, and Assumption 4.5 can be weakened by replacing $K_{2,p}$ with one. Thus, when $A_0(P)$ does not exist or is known, Theorem 4.1 requires $p = o(n_2)$ and $a_n = o(\sqrt{n_2})$, with the latter requirement automatically holding when there is no linearization error or being implied by $p^2 = o(n_2)$ when $a_n \lesssim p$. ■

4.3 Procedure for selecting $\hat{y}_n^{(1)}$

In this section we describe a procedure for selecting $\hat{y}_n^{(1)}$ that satisfies Assumption 4.4. Given suitable estimators $\hat{b}_{j,n}^{(1)}$ and $\hat{M}_{0,n}^{(1)}$, we select $\hat{y}_n^{(1)}$ as the solution to the following optimization problem:

$$\begin{aligned} \hat{y}_n^{(1)} \in \operatorname*{argmax}_{\|y\|_1 \leq 1} \min_{j \in J^*} \ & \frac{\sqrt{n_1}(\hat{b}_{j,n}^{(1)})' \hat{M}_{0,n}^{(1)} y}{\hat{\omega}_{j,n}} \\ \text{subject to } & \sqrt{n_1}(\hat{b}_{j,n}^{(1)})' \hat{M}_{0,n}^{(1)} y \geq \hat{\omega}_{j,n} \text{ for all } j \in (J^*)^c \ , \end{aligned} \tag{28}$$

where $\hat{\omega}_{j,n} > 0$ are positive scaling parameters that we specify below. If the optimization problem in (28) is found to be infeasible, then we set $\hat{y}_n^{(1)} = 0$. In Appendix A, we explain how this optimization problem can be re-formulated as a linear program. Note that the specific choice of $\hat{\omega}_{j,n}$ for $j \in J^*$ will not affect the validity of the procedure, although these should be carefully selected to ensure the test has good power; see the discussion following Lemma 4.1 for details. In contrast, as we demonstrate in Lemma 4.1, the choice of $\hat{\omega}_{j,n}$ for $j \in (J^*)^c$ is relevant for the screening method. In the case of the direct method, that is, when $(J^*)^c$ contains only those j for which $b_j(P)' M_0(P)$ is known, the specific choice of $\hat{\omega}_{j,n}$ for $j \in (J^*)^c$ is immaterial in practice: we can simply set the constraints in (28) to ensure that $b_j(P)' M_0(P) y$ is strictly positive for every $j \in (J^*)^c$.

Lemma 4.1. *Let $\hat{y}_n^{(1)}$ be defined as in (28) and $\hat{\omega}_{j,n} > 0$ for all $j \in J$. Then, the following statements hold:*

(a) Suppose that $b_j(P)'M_0(P)$ is known for all $j \in (J^*)^c$. Then, it follows that Assumption 4.4 holds.

(b) Suppose Assumptions 4.1, 4.2, 4.3 and 4.6 hold, $K_{1,p}(K_{0,p} \vee K_{1,p}) \log(1+p) d_1^{1/3} = o(np^{2/3})$, and $\max_{j \in (J^*)^c}\{(pd_1)^{1/3}/\hat{\omega}_{j,n}\} = o_P(1)$ uniformly in $P \in \mathbf{P}$. Then, it follows that Assumption 4.4 holds.

(c) Suppose Assumptions 4.1, 4.2, 4.3 and 4.6 hold, and that the following moment restrictions hold

$$\sup_{P \in \mathbf{P}} E_P\left[\max_{1 \le j \le d_1+1} \|\xi_j(Z,P)\|_\infty^2\right] \le C_{\xi,p}^2 \qquad \sup_{P \in \mathbf{P}} E_P\left[\max_{1 \le j \le d_1+1} \|\varphi_j(Z,P)\|_2^2\right] \le C_{\varphi,p}^2 \ . \tag{29}$$

If $pC_{\varphi,p}^2 \log^2(p+d_1)(K_{0,p} \vee K_{1,p}) = o(n)$ and $\max_{j \in (J^*)^c}\{C_{\xi,p}\sqrt{\log(p+d_1)}/\hat{\omega}_{j,n}\} = o_P(1)$ uniformly in $P \in \mathbf{P}$, then it follows that Assumption 4.4 holds.

Part (a) of Lemma 4.1 formally states that for the direct method, the sole constraint on the weights $\hat{\omega}_{j,n}$ is that they be positive. Parts (b) and (c) specify the requirements on the weights $\hat{\omega}_{j,n}$ when the coordinates $j \in (J^*)^c$ may be such that $b_j(P)'M_0(P)$ depends on P, as in the screening method. In particular part (b) shows that Assumption 4.4 is satisfied under rate restrictions on the growth of d_1 and that $\hat{\omega}_{j,n}$ diverge to infinity faster than $(pd_1)^{1/3}$. As we show in part (c), however, these restrictions can be considerably weakened provided we strengthen our moment conditions by assuming that (29) holds. Under such a requirement, part (c) weakens the rate restrictions on d_1 and the rate at which $\hat{\omega}_{j,n}$ must diverge to be logarithmic.

Although the conditions of Lemma 4.1 are satisfied for a wide range of choices of $\hat{\omega}_{j,n}$, careful choices of these weights should be used to ensure that our test has good power properties. First, $\hat{\omega}_{j,n}$ should be selected so that, under the alternative, (28) is feasible with probability tending to one: for this we specify $\hat{\omega}_{j,n}$ to satisfy $\hat{\omega}_{j,n} = o_P(\sqrt{n_1})$. Second, the weights $\hat{\omega}_{j,n}$ should be selected in a way that "appropriately weights" the degree of violation of each component. To this end, we've found that in simulations it works well to set $\hat{\omega}_{j,n} = c_n \cdot \hat{\sigma}_{j,n}^{(1)}(\hat{y}_{0,n})$ for some sequence c_n diverging to infinity and $\hat{\sigma}_{j,n}^{(1)}(y)$ our (truncated) estimate of the asymptotic standard deviation of $\sqrt{n_1}(\hat{b}_{j,n}^{(1)})'\hat{M}_{0,n}y$ evaluated at $\hat{y}_{0,n}$. To maintain the linear structure of the optimization problem, we employ a preliminary direction $\hat{y}_{0,n}$ computed by solving the linear program

$$\begin{aligned} \hat{y}_{0,n} \in \operatorname*{argmax}_{\|y\|_1 \le 1} \min_{j \in J^*} \ & \sqrt{n_1}(\hat{b}_{j,n}^{(1)})'\hat{M}_{0,n}^{(1)}y \\ \text{subject to } & \sqrt{n_1}(\hat{b}_{j,n}^{(1)})'\hat{M}_{0,n}^{(1)}y \ge 0 \text{ for all } j \in (J^*)^c \ . \end{aligned} \tag{30}$$

With regards to the sequence c_n, we have found that the following choices work well in practice: in a low dimensional regime where p and d_1 could be considered fixed, we set $c_n = \sqrt{\log(\log(n_1))}$; in a high dimensional regime where p or d_1 diverges with sample size, we set $c_n = \sqrt{\log(\log(\log(n_1))) \times \log(p+d_1)}$.

Remark 4.4. In the case of the direct method, in which $b_j(P)'M_0(P)$ is known for all $j \in (J^*)^c$, it is straightforward to verify that Assumption 4.4 is in fact satisfied for any $\hat{y}_n^{(1)}$ such that

$$\sqrt{n_1}(\hat{b}_{j,n}^{(1)})'\hat{M}_{0,n}^{(1)}y \ge \eta \text{ for all } j \in (J^*)^c$$

for some $\eta > 0$, not just $\hat{y}_n^{(1)}$ defined in (28) and discussed in Lemma 4.1. ■

5 Simulations

In this section, we illustrate the finite-sample performance of our procedure via simulation. We consider three distinct settings, each inspired by earlier related papers. In all designs, we examine the performance of the direct and the screening method with $n_1 = n_2 = n/2$. In all cases we set $\hat{\omega}_{j,n}$ to equal $c_n \cdot \hat{\sigma}^{(1)}_{j,n}(\hat{y}_{0n})$ for $\hat{y}_{0,n}$ as defined in (30) and $c_n = \sqrt{\log(\log(\log(n_1))) \times \log(p + d_1)}$.

5.1 Cox et al. (2025)

We first consider the simple one-sided model described in Cox et al. (2025). Let $C := (C_1, \dots C_H)'$ and $X := (X_1, \dots, X_H)'$ be random vectors and let $H \geq 0$ index the number of inequalities to be considered. In our notation, the design can be described as $A_0(P) = \nu(P) := (E_P[C_1], \dots, E_P[C_H])'$, $A_1(P) = \mathbf{I}_H$, $\beta(P) = -\mu(P) - \mathbf{v}\theta$, where $\mu(P) := (E_P[X_1], \dots, E_P[X_H])'$ and $\mathbf{v} := (1, 1, 0, \dots, 0)' \in \mathbb{R}^H$. It can be shown that the identified set of θ is $(-\infty, 0]$. Under P, (C, X) are distributed according to

$$X \sim N(\mu(P), I_H),$$
$$C \sim N(\nu(P), 2I_H),$$

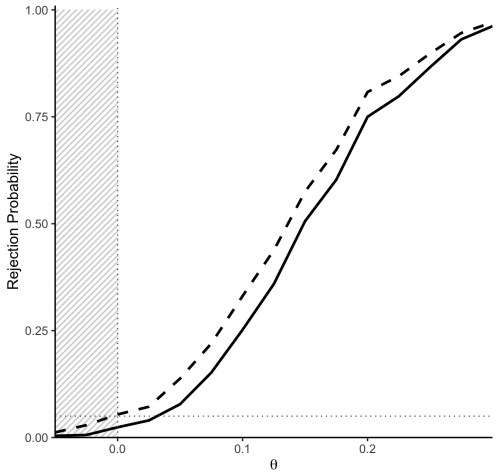
with $\mu(P) = (-1, 1, 1, \dots, 1)'$, $\nu(P) = (1, -1, -1, \dots, -1)'$. Accordingly, given a random sample from (X, C), $\hat{A}_{0,n}$ and $\hat{b}_{j,n}$ are computed by taking sample averages.

Figures 2 and 3 present the rejection probabilities from $1,000$ draws for both of our proposed tests for the null hypothesis that $P \in \mathbf{P}_0$ at a 5% significance level, over a grid of values of θ and across different choices of dimension H and sample size n (the direct method is represented by the solid line and the screening method by the dashed line). We see that both the direct and screening methods control size in all cases within the identified set for θ. Outside of the identified set, the rejection probabilities for the screening method are slightly larger than for the direct method, although the differences become negligible for large H and/or large n.

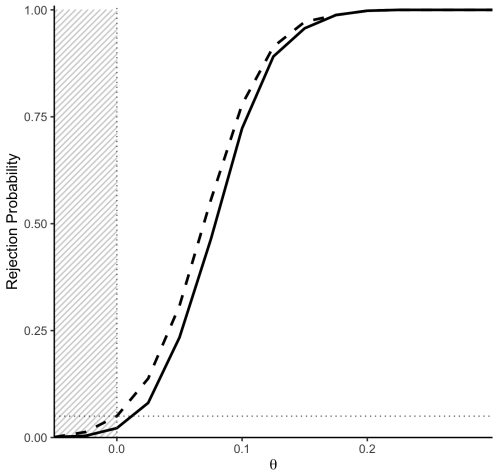
5.2 Goff and Mbakop (2025)

We revisit a simple instance of Example 2.3 proposed by Goff and Mbakop (2025), which is loosely based on the application in Mogstad et al. (2017) to the bed net data used by Dupas (2014). The outcome Y is a binary indicator of using a new type of anti-malarial bed net, the treatment D is an indicator for purchasing the bed net, and the instrument Z is a randomly-assigned price for the bed net. The context implies that $Y(0) = 0$. Suppose that the instrument is binary and that there are no covariates. The researcher assumes that $E[Y(1)|U = u] = \theta_0 + \theta_1 u + \theta_2 u^2$ is a weakly decreasing, quadratic function of u, which must be contained within $[0, 1]$ because $Y(1)$ is binary. These shape restrictions are imposed through four linear constraints:

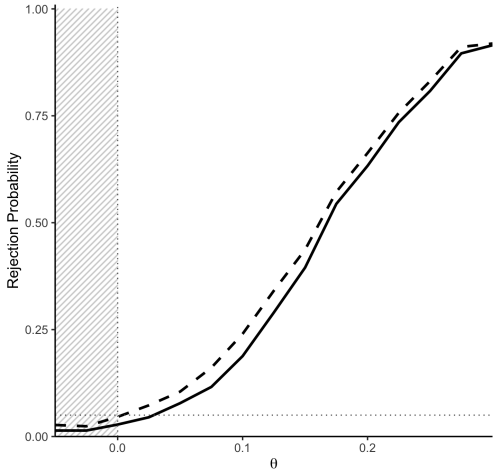
$$0 \leq \theta_0 \leq 1, \quad 0 \leq \theta_0 + \theta_1 + \theta_2, \quad \theta_1 \leq 0, \quad \text{and} \quad \theta_1 + 2\theta_2 \leq 0, \tag{31}$$



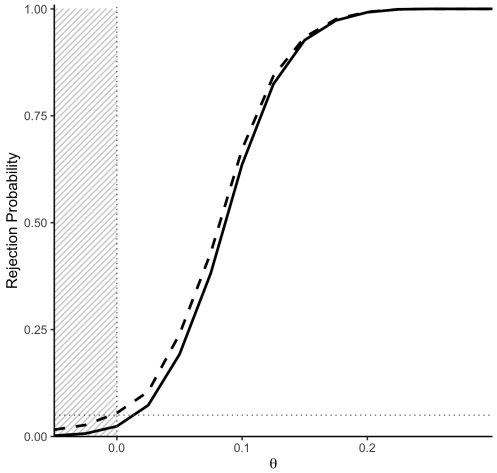
(a) $H = 3,\ n = 500$



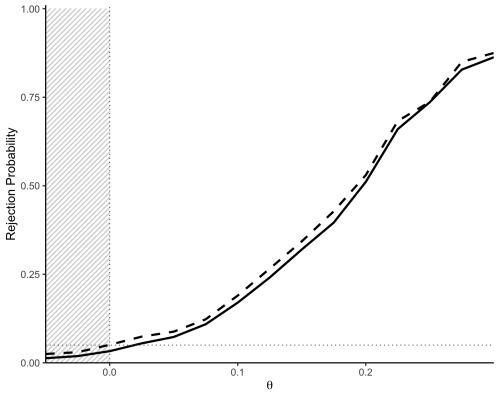
(b) $H = 3,\ n = 2000$



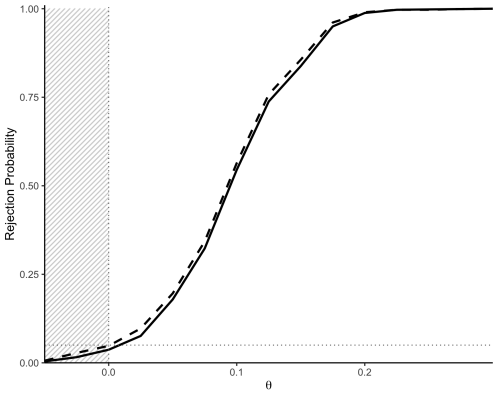
(c) $H = 10,\ n = 500$



(d) $H = 10,\ n = 2000$

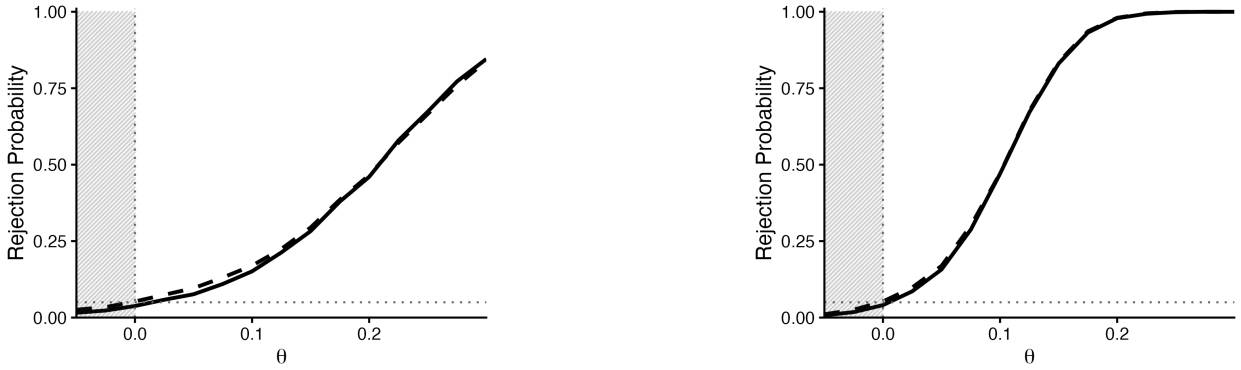


(e) $H = 50,\ n = 500$



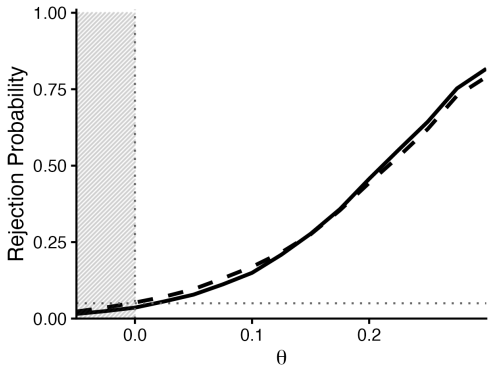
(f) $H = 50,\ n = 2000$

Figure 2: Cox et al. (2025) simulation rejection curves, arranged by H (rows) and n (columns). In each plot, the hypothesized value of θ is on the horizontal axis. The shaded region is the identified set for θ. The dashed line represents the screening method and the solid line represents the direct method.

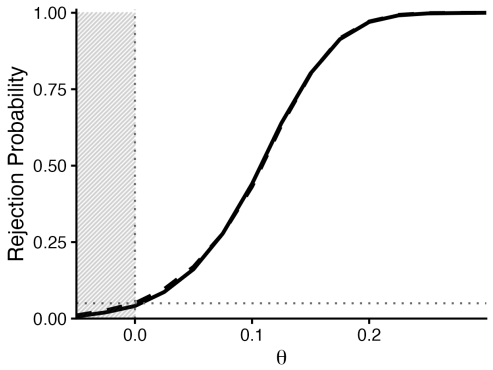


(a) $H = 200,\ n = 500$

(b) $H = 200,\ n = 2000$



(c) $H = 500,\ n = 500$



(d) $H = 500,\ n = 2000$

Figure 3: Cox et al. (2025) simulation rejection curves, arranged by H (rows) and n (columns). In each plot, the hypothesized value of θ is on the horizontal axis. The shaded region is the identified set for θ. The dashed line represents the screening method and the solid line represents the direct method.

which imply $E[Y(1)|U=u] \leq 1$ for all u, because $E[Y(1)|U=0] = \theta_0 \leq 1$ and the derivative of $E[Y(1)|U=u]$ is $\theta_1 + 2\theta_2 \leq 0$.

It will be convenient to use an alternative parameterization. Set $\tilde{\theta}_1 = -\theta_1$, $\delta = \theta_0 + \theta_1 + \theta_2$, so that $\theta_0, \tilde{\theta}_1, \delta \geq 0$. Then the constraints (31) can be simplified to

$$\theta_0 + s_1 = 1 \quad \text{and} \quad -2\theta_0 + \tilde{\theta}_1 + 2\delta + s_2 = 0, \tag{32}$$

where $s_1, s_2 \geq 0$ are slack variables. The researcher matches the moments $E_P[YD|Z=z]$ for $z = 0, 1$, creating the two restrictions

$$\begin{aligned}
\left(p(0) - \frac{p(0)^3}{3}\right)\theta_0 + \left(\frac{p(0)^3}{3} - \frac{p(0)^2}{2}\right)\tilde{\theta}_1 + \frac{p(0)^3}{3}\delta &= E_P[YD|Z=0] \\
\left(p(1) - \frac{p(1)^3}{3}\right)\theta_0 + \left(\frac{p(1)^3}{3} - \frac{p(1)^2}{2}\right)\tilde{\theta}_1 + \frac{p(1)^3}{3}\delta &= E_P[YD|Z=1]\ .
\end{aligned} \tag{33}$$

The null hypothesis of interest is $H_0 : E_P[Y(1)] = \tau_0$, where

$$E_P[Y(1)] = \int_0^1 \theta_0 - \tilde{\theta}_1 u + (\delta - \theta_0 + \tilde{\theta}_1)u^2\, du = \frac{2}{3}\theta_0 - \frac{1}{6}\tilde{\theta}_0 + \frac{1}{3}\delta. \tag{34}$$

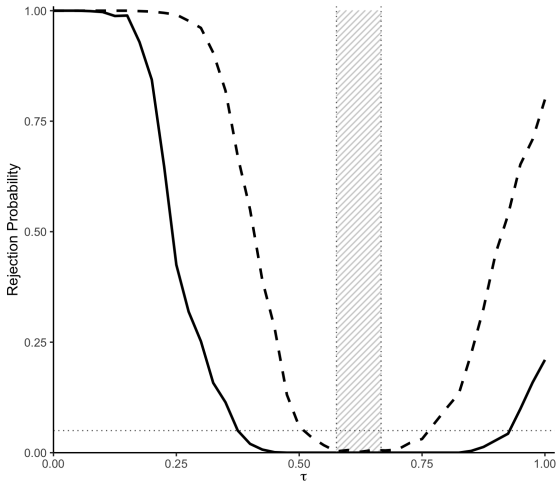
We set $x_1 = (\theta_0, \tilde{\theta}_1, \delta, s_1, s_2)'$, so that the problem defined by (32)–(34) fits in the form (1) with

$$A_1(P) = \begin{pmatrix}
p(0) - \frac{p(0)^3}{3} & \frac{p(0)^3}{3} - \frac{p(0)^2}{2} & \frac{p(0)^3}{3} & 0 & 0 \\
p(1) - \frac{p(1)^3}{3} & \frac{p(1)^3}{3} - \frac{p(1)^2}{2} & \frac{p(1)^3}{3} & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
-2 & 1 & 2 & 0 & 1 \\
\frac{2}{3} & -\frac{1}{6} & \frac{1}{3} & 0 & 0
\end{pmatrix} \qquad \beta(P) = \begin{pmatrix}
E_P[YD|Z=0] \\
E_P[YD|Z=1] \\
1 \\
0 \\
\tau_0
\end{pmatrix},$$

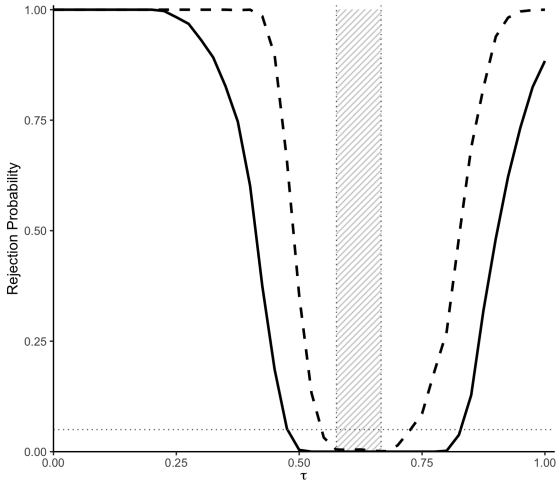
and with both x_0 and $A_0(P)$ null.

The simulation design is specified as $P\{Z=1\} = 0.5 = P\{Z=0\}$, $p(0) = 1/3$, $p(1) = 2/3$, $Y(0) = 0$, $Y(1) = \mathbf{1}\{V \leq \theta_0 + \theta_1 U + \theta_2 U^2\}$ where $V|Z, U \sim U[0,1]$ and $(\theta_0, \theta_1, \theta_2) = (1, -1, 0.5)$. Given this design, the identified set for the ATE is $[0.58, 0.67]$. Using a random sample from (Y, D, Z), $\hat{b}_{j,n}$ is once again computed using sample analogs (recall that $A_0(P)$ does not exist given this parametrization).

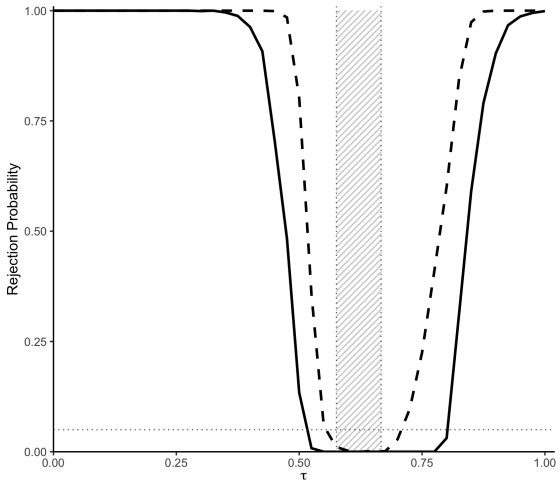
Figure 4 presents the rejection probabilities from $1,000$ draws for tests of the null hypothesis $H_0 : E[Y(1)] = \tau_0$ at a 5% significance level, for a grid of values of τ_0. Both tests control size within the identified set for all sample sizes, however, the rejection probability remains below 5% well outside of the identified set in small samples. At all sample sizes, we find that the direct method is more conservative outside of the identified set than the screening method in this example.



(a) $n = 500$



(b) $n = 2000$



(c) $n = 5000$

Figure 4: Goff and Mbakop (2025) simulation rejection curves, arranged by n. In each plot, the hypothesized value of τ_0 is on the horizontal axis. The shaded region is the identified set. The dashed line represents the screening method and the solid line represents the direct method.

5.3 Freyberger and Horowitz (2015)

Finally, we consider the model presented in Freyberger and Horowitz (2015), as described in Example 2.1. The simulation design is specified as follows: $\mathcal{X} = \{2, 3, 4, 5, 6, 7\}$ and $\mathcal{W} = \{0, 1\}$, with $\pi_{h,k} = \Pr(X = x_h, W = w_k)$ given by

$$\Pi = \begin{pmatrix} 0.20 & 0.15 \\ 0.10 & 0.12 \\ 0.06 & 0.07 \\ 0.05 & 0.08 \\ 0.03 & 0.06 \\ 0.03 & 0.05 \end{pmatrix} .$$

The vector $g = (g(2), \ldots, g(7))'$ is given by $(23, 17, 13, 11, 9, 8)'$. The instrument is distributed as $Z_i \sim N(0, 1)$, independently of (X_i, W_i). Finally, the outcome equation is

$$Y_i = g(X_i) + U_i \ ,$$

where the error is given by

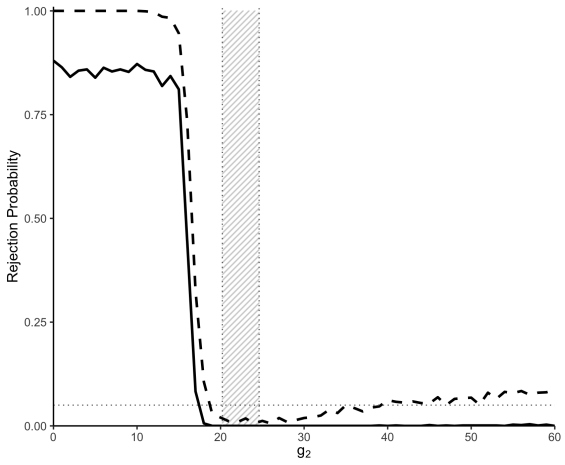
$$U_i = X_i Z_i^2 - E_P[X_i | W_i] \ .$$

The shape constraint we maintain is that the structural function g is decreasing, i.e., $Sg \leq 0$ for

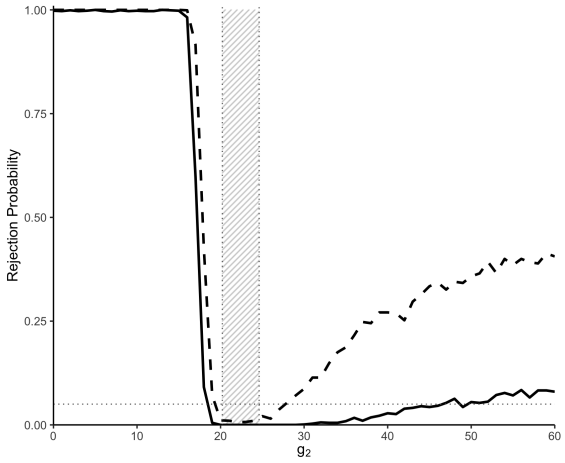
$$S = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} ,$$

and our functional of interest is $L(g) = g(2)$. Given this design the identified set for $L(g)$ is $[20.21, 24.61]$. Using a random sample from (Y, X, W, Z), $\hat{A}_{0,n}$ and $\hat{b}_{j,n}$ are computed using sample analogs, using the known value of S and c.

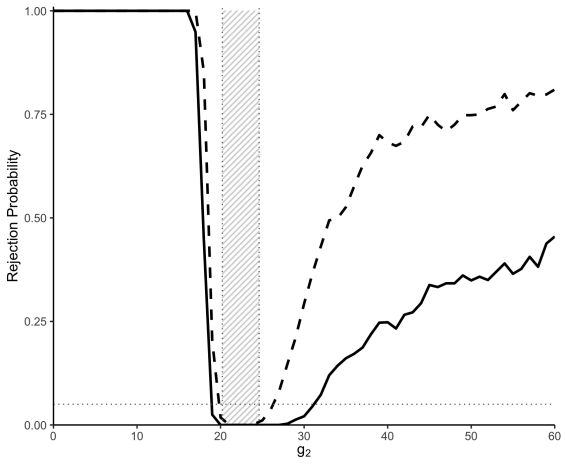
Figure 5 presents the rejection probabilities from $1,000$ draws for tests of the null hypothesis $H_0 : L(g) = L_0$ at a 5% significance level, for a grid of values of L_0. This design seems particularly challenging, with the rejection probabilities outside of the identified set being highly asymmetric. One possible source for this asymmetry is that the shape constraints are violated to the left of the identified set, but are not violated to the right. Beyond this observation, our findings are similar to the previous two examples; both tests control size within the identified set for all sample sizes, and we find that the direct method is more conservative outside of the identified set than the screening method.



(a) $n = 500$



(b) $n = 2000$



(c) $n = 5000$

Figure 5: Freyberger and Horowitz (2015) simulation rejection curves, arranged by n. In each plot, the hypothesized value of L_0 is on the horizontal axis. The shaded region is the identified set for $L(g)$. The dashed line represents the screening method and the solid line represents the direct method.

References

Andrews, I., Roth, J. and Pakes, A. (2023). Inference for Linear Conditional Moment Inequalities. *The Review of Economic Studies*, **90** 2763–2791. URL https://doi.org/10.1093/restud/rdad004.

Bai, Y., Santos, A. and Shaikh, A. M. (2022). On testing systems of linear inequalities with known coefficients. Tech. rep., Working Paper.

Bajari, P., Fox, J. T. and Ryan, S. P. (2007). Linear Regression Estimation of Discrete Choice Models with Nonparametric Distributions of Random Coefficients. *American Economic Review*, **97** 459–463.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Fernández-Val, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, **213** 4–29.

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to Linear Optimization*, vol. 6. Athena Scientific Belmont, MA.

Bhatia, R. (2013). *Matrix analysis*, vol. 169. Springer Science & Business Media.

Bogachev, V. I. (1998). *Gaussian measures*. 62, American Mathematical Soc.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Chen, Y. M., Chen, X. S. and Li, W. (2016). On perturbation bounds for orthogonal projections. *Numerical Algorithms*, **73** 433–444.

Cho, J., and Russell, T. M. (2024). Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments. *Journal of Business & Economic Statistics*, **42** 563–578. Publisher: ASA Website _eprint: https://doi.org/10.1080/07350015.2023.2203768, URL https://doi.org/10.1080/07350015.2023.2203768.

Cox, G. and Shi, X. (2023). Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models. *The Review of Economic Studies*, **90** 201–228. URL https://doi.org/10.1093/restud/rdac015.

Cox, G. F., Shi, X. and Shimizu, Y. (2025). Testing Inequalities Linear in Nuisance Parameters. ArXiv:2510.27633 [stat], URL http://arxiv.org/abs/2510.27633.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, **82** 197–228. URL http://dx.doi.org/10.3982/ECTA9508.

Fang, Z., Santos, A., Shaikh, A. M. and Torgovitsky, A. (2021). Inference for Large-Scale Linear Systems with Known Coefficients. ArXiv:2009.08568 [econ], URL http://arxiv.org/abs/2009.08568.

Fang, Z., Santos, A., Shaikh, A. M. and Torgovitsky, A. (2023). Inference for large-scale linear systems with known coefficients. *Econometrica*, **91** 299–327.

Fox, J. T., Kim, K. I., Ryan, S. P. and Bajari, P. (2011). A simple estimator for the distribution of random coefficients. *Quantitative Economics*, **2** 381–418.

Freyberger, J. and Horowitz, J. L. (2015). Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics*, **189** 41–53. URL https://www.sciencedirect.com/science/article/pii/S0304407615001918.

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

Gasparin, M., Wang, R. and Ramdas, A. (2025). Combining exchangeable p-values. *Proceedings of the National Academy of Sciences*, **122** e2410849122.

Goff, L. and Mbakop, E. (2025). Inference on the value of a linear program. ArXiv:2506.06776 [econ], URL http://arxiv.org/abs/2506.06776.

Gu, J., Russell, T. and Stringham, T. (2024). Counterfactual Identification and Latent Space Enumeration in Discrete Outcome Models. 4188109.

Gu, J. and Russell, T. M. (2023). Partial identification in nonseparable binary response models with endogenous regressors. *Journal of Econometrics*, **235** 528–562.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, **73** 669–738.

Hiriart-Urruty, J.-B. and Le, H. Y. (2013). A variational approach of the rank function. *TOP*, **21** 207–240. URL https://doi.org/10.1007/s11750-013-0283-y.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** 467–475.

Liu, Y. (2025). Synthetic parallel trends. *arXiv preprint arXiv:2511.05870*.

Luenberger, D. G. (1969). *Optimization by vector space methods*. John Wiley & Sons, Inc., New York-London-Sydney.

Magnus, J. and Neudecker, H. (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics, Wiley.

Manski, C. F. (2007). Partial Indentification of Counterfactual Choice Probabilities. *International Economic Review*, **48** 1393–1410. 4542017.

Mogstad, M., Santos, A. and Torgovitsky, A. (2017). Using instrumental variables for inference about policy relevant treatment parameters. *NBER Working Paper*.

Mogstad, M., Santos, A. and Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, **86** 1589–1619.

Mogstad, M. and Torgovitsky, A. (2024). Instrumental variables with unobserved heterogeneity in treatment effects. In *Handbook of Labor Economics*, vol. 5. Elsevier, 1–114. URL https://linkinghub.elsevier.com/retrieve/pii/S1573446324000038.

Pollard, D. (2002). *A user's guide to measure theoretic probability*. 8, Cambridge University Press.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Romano, J. P. (2004). On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics*, **31** 567–584.

Shea, J. and Torgovitsky, A. (2023). ivmte: An r package for extrapolating instrumental variable estimates away from compliers. *Observational Studies*, **9** 1–42. URL https://muse.jhu.edu/pub/56/article/883476.

Tebaldi, P., Torgovitsky, A. and Yang, H. (2023). Nonparametric Estimates of Demand in the California Health Insurance Exchange. *Econometrica*, **91** 107–146.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, **12** 389–434.

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics, Springer-Verlag, New York. URL https://www.springer.com/gp/book/9780387946405.

Voronin, A. (2025). Linear programming approach to partially identified econometric models. ArXiv:2503.14940 [econ], URL http://arxiv.org/abs/2503.14940.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, **70** 331–341.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge university press.

Wedin, P.-Å. (1973). Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, **13** 217–232.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press.

A Details for Section 4.3

Here, we present the linear programming reformulation of (28):

$$\begin{aligned}
\max_{t,y^+,y^-} \quad & t \\
\text{s.t.} \quad & \sqrt{n_1}(\hat{b}^{(1)}_{j,n})'\hat{M}^{(1)}_{0,n}(y^+ - y^-) \geq t\hat{\omega}_{j,n}, && \forall j \in J^* \\
& \sqrt{n_1}(\hat{b}^{(1)}_{j,n})'\hat{M}^{(1)}_{0,n}(y^+ - y^-) \geq \hat{\omega}_{j,n}, && \forall j \in (J^*)^c \\
& \mathbf{1}_p' y^+ + \mathbf{1}_p' y^- \leq 1 \\
& y^+ \geq 0, y^- \geq 0 \ ,
\end{aligned}$$

where $\mathbf{1}_p$ is the $p \times 1$ vector of ones.

B Proofs of Main Results

In the rest of the appendix, for $1 \leq p, q \leq \infty$, let $\|\cdot\|_{p,q}$ denote the (p,q)-operator norm of a matrix. We write $a_n \lesssim b_n$ to denote that there exists a universal constant c such that $a_n \leq cb_n$ for all $n \geq 1$.

B.1 Proof of Lemma 3.1

First, note that if $(A_0, A_1, b) \in \mathbf{C}_0$, then there exists $x_0 \in \mathbb{R}^{d_0}$, $x_1 \in \mathbb{R}^{d_1}$ with $x_1 \geq 0$ such that $A_0x_0 + A_1x_1 = b$. By the definition of M_0, $M_0A_0x_0 = 0$, so that $M_0A_1x_1 = M_0b$ with $x_1 \geq 0$, which implies

$$\mathbf{C}_0 \subseteq \{(A_0, A_1, b) : M_0A_1x_1 = M_0b \text{ for some } x_1 \in \mathbb{R}^{d_1}, x_1 \geq 0\}.$$

For the reverse inclusion, suppose there exists a solution $\tilde{x}_1 \geq 0$ to the equation $M_0A_1\tilde{x}_1 = M_0b$. By definition of M_0, we may write $M_0A_1\tilde{x}_1 = A_1\tilde{x}_1 - \Pi_0A_1\tilde{x}_1$, where Π_0 is the projection onto the column space of A_0. Next use that $\Pi_0A_1\tilde{x}_1 = A_0\tilde{x}_0$ for some $\tilde{x}_0 \in \mathbb{R}^{d_0}$ by definition of Π_0 and note that, similarly, we can write $M_0b = b - \Pi_0b = b - A_0\tilde{b}$ for some $\tilde{b} \in \mathbb{R}^{d_0}$. Therefore we obtain that

$$A_0(\tilde{b} - \tilde{x}_0) + A_1\tilde{x}_1 = b \ ,$$

which implies $\{(A_0, A_1, b) : M_0A_1x_1 = M_0b \text{ for some } x_1 \in \mathbb{R}^{d_1}, x_1 \geq 0\} \subseteq \mathbf{C}_0$, as desired. ■

B.2 Proof of Theorem 3.1

First note that Lemma C.1 establishes that $\text{cl}(\mathbf{C}_0) \cap (\mathbf{C}^{\text{RD}})^c \subseteq \bar{\mathbf{C}}_0$, which allows us to conclude that

$$\text{cl}(\mathbf{C}_0) \subseteq \bar{\mathbf{C}}_0 \cup \mathbf{C}^{\text{RD}}. \tag{35}$$

To establish the reverse inclusion, i.e. $\bar{\mathbf{C}}_0 \cup \mathbf{C}^{\mathrm{RD}} \subseteq \mathrm{cl}(\mathbf{C}_0)$, we begin by defining the set $\mathbf{C}^{\mathrm{NP}}$ to be given by

$$\mathbf{C}^{\mathrm{NP}} := \{(A_0, A_1, b) \in \mathbb{R}^{p \times d_0} \times \mathbb{R}^{p \times d_1} \times \mathbb{R}^p : \{M_0 A_1 x_1 : x_1 \geq 0\} \text{ is not a pointed cone}\} \,. \tag{36}$$

Since $\bar{\mathbf{C}}_0 \cap (\mathbf{C}^{\mathrm{NP}})^c \subseteq \mathbf{C}_0$ by Lemma C.3, and $\mathbf{C}^{\mathrm{NP}} \cup \mathbf{C}^{\mathrm{RD}} \subseteq \mathrm{cl}(\mathbf{C}_0)$ by Lemmas C.2 and C.4, we obtain

$$\bar{\mathbf{C}}_0 \cup \mathbf{C}^{\mathrm{RD}} = (\bar{\mathbf{C}}_0 \cap (\mathbf{C}^{\mathrm{NP}})^c) \cup (\bar{\mathbf{C}}_0 \cap \mathbf{C}^{\mathrm{NP}}) \cup \mathbf{C}^{\mathrm{RD}} \subseteq \mathrm{cl}(\mathbf{C}_0),$$

which together with (35) establishes the claim of the theorem. ■

B.3 Proof of Theorem 4.1

In what follows, it will be helpful to decompose the test statistic T_n into the following two components:

$$T_n = \min_{j \in J^*} \Bigg(\underbrace{\frac{\sqrt{n_2}\big((\hat{b}^{(2)}_{j,n})' \hat{M}^{(2)}_{0,n} \hat{y}^{(1)}_n - b_j(P)' M_0(P) \hat{y}^{(1)}_n\big)}{\hat{\sigma}^{(2)}_{j,n}(\hat{y}^{(1)}_n)}}_{:= \hat{\mathbb{G}}^{(2)}_{j,n}(\hat{y}^{(1)}_n)} + \frac{\sqrt{n_2} b_j(P)' M_0(P) \hat{y}^{(1)}_n}{\hat{\sigma}^{(2)}_{j,n}(\hat{y}^{(1)}_n)} \Bigg) \,. \tag{37}$$

Next, note Assumption 4.4 and $\hat{y}^{(1)}_n = 0$ implying that $T_n = 0$ and hence $1\{T_n > z_{1-\alpha}\} = 0$ yield that

$$\begin{aligned} \sup_{P \in \mathbf{P}_0} P\{T_n > z_{1-\alpha}\} &= \sup_{P \in \mathbf{P}_0} P\Big\{T_n > z_{1-\alpha} \text{ and } \hat{y}^{(1)}_n \in \mathcal{Y}(P; J^*) \cup \{0\}\Big\} + o(1) \\ &= \sup_{P \in \mathbf{P}_0} P\Big\{T_n > z_{1-\alpha} \text{ and } \hat{y}^{(1)}_n \in \mathcal{Y}(P; J^*)\Big\} + o(1) \,. \end{aligned} \tag{38}$$

Moreover, note that for any $P \in \mathbf{P}_0$, the event $\hat{y}^{(1)}_n \in \mathcal{Y}(P; J^*)$ implies that $b_{\hat{k}_n}(P)' M_0(P) \hat{y}^{(1)}_n \leq 0$ for some $\hat{k}_n \in J^*$ depending on $\hat{y}^{(1)}_n$ and P (though we leave the dependence implicit to avoid notational clutter). In particular, we obtain from the decomposition in (37) that we must have

$$1\Big\{T_n > z_{1-\alpha} \text{ and } \hat{y}^{(1)}_n \in \mathcal{Y}(P; J^*)\Big\} \leq 1\Big\{\hat{\mathbb{G}}^{(2)}_{\hat{k}_n,n}(\hat{y}^{(1)}_n) > z_{1-\alpha} \text{ and } \hat{y}^{(1)}_n \in \mathcal{Y}(P; J^*)\Big\} \,. \tag{39}$$

Next, note that result (39) together with $\hat{y}^{(1)}_n$ and $\hat{k}_n$ being functions of only the first split $\{Z_i\}_{i \in I_{n,1}}$ and therefore being independent of $\hat{\mathbb{G}}^{(2)}_{k,n}(y)$ for any $k \in J^*$ and $y \in \mathcal{Y}(P; J^*)$ we can conclude that

$$\begin{aligned} &\sup_{P \in \mathbf{P}_0} P\Big\{T_n > z_{1-\alpha} \text{ and } \hat{y}^{(1)}_n \in \mathcal{Y}(P; J^*)\Big\} \\ &\quad \leq \sup_{P \in \mathbf{P}_0} E_P\Big[P\Big\{\hat{\mathbb{G}}^{(2)}_{\hat{k}_n,n}(\hat{y}^{(1)}_n) > z_{1-\alpha} \Big| \{Z_i\}_{i \in I_{n,1}}\Big\} \cdot 1\{\hat{y}^{(1)}_n \in \mathcal{Y}(P; J^*)\}\Big] \\ &\quad \leq \sup_{P \in \mathbf{P}_0} E_P\Bigg[\max_{k \in J^*} \sup_{y \in \mathcal{Y}(P; J^*)} P\Big\{\hat{\mathbb{G}}^{(2)}_{k,n}(y) > z_{1-\alpha} \Big| \{Z_i\}_{i \in I_{n,1}}\Big\} \cdot 1\{\hat{y}^{(1)}_n \in \mathcal{Y}(P; J^*)\}\Bigg] \\ &\quad \leq \sup_{P \in \mathbf{P}_0} \max_{k \in J^*} \sup_{y \in \mathcal{Y}(P; J^*)} P\Big\{\hat{\mathbb{G}}^{(2)}_{k,n}(y) > z_{1-\alpha}\Big\} \,, \end{aligned} \tag{40}$$

where in the final inequality we used the bound $1\{\hat{y}_n^{(1)} \in \mathcal{Y}(P;J^*)\} \leq 1$ and again used that $\hat{\mathbb{G}}_{k,n}^{(2)}(y)$ is independent of the first split $\{Z_i\}_{i\in I_{n,1}}$.

We next aim to apply Lemma C.8 to show $\hat{\mathbb{G}}_{k,n}^{(2)}(y) - \mathbb{G}_{k,n}^{(2)}(P;y)$ converges in probability to zero uniformly, where

$$\mathbb{G}_{k,n}^{(2)}(P;y) := \frac{1}{\sqrt{n_2}} \sum_{i\in I_{n,2}} \frac{y'\xi_k(Z_i,P)}{(\sigma_j(P;y) \vee \underline{\sigma})} .$$

To verify the conditions of Lemma C.8 note Assumption 4.3 and Jensen's inequality imply $\mathrm{Var}_P[y'\xi_j(Z,P)]$ is bounded uniformly in $P \in \mathbf{P}$ and $1 \leq j \leq d_1+1$. Also note Assumptions 4.6(a)(b) imply $(K_{1,p}K_{2,p}^2 \vee \bar{s}_p^2)(K_{0,p} \vee K_{1,p})\log(1+p) = o(n)$ (for $p \geq 2$) and Assumptions 4.6(a)(b)(c) imply $(K_{1,p}^{1/2}K_{2,p} \vee \bar{s}_p)a_n = o(n)$. By Assumptions 4.1, 4.2, and 4.5 we may therefore apply Lemma C.8, which together with Assumption 4.6 implies

$$\limsup_{n_2\to\infty} \sup_{P\in\mathbf{P}} \max_{1\leq j\leq d_1+1} \sup_{\|y\|_1\leq 1} P\left\{\left|\hat{\mathbb{G}}_{k,n}^{(2)}(y) - \mathbb{G}_{k,n}^{(2)}(P;y)\right| > \epsilon\right\} = 0$$

for any $\epsilon > 0$. Hence, since $\mathcal{Y}(P;J^*) \subseteq \{y \in \mathbb{R}^p : \|y\|_1 \leq 1\}$ by definition, we obtain for any $\epsilon > 0$ that

$$\limsup_{n_2\to\infty} \sup_{P\in\mathbf{P}_0} \max_{k\in J^*} \sup_{y\in\mathcal{Y}(P;J^*)} P\left\{\hat{\mathbb{G}}_{k,n}^{(2)}(y) > z_{1-\alpha}\right\} \leq \limsup_{n_2\to\infty} \sup_{P\in\mathbf{P}_0} \max_{k\in J^*} \sup_{y\in\mathcal{Y}(P;J^*)} P\left\{\mathbb{G}_{k,n}^{(2)}(y) > z_{1-\alpha} - \epsilon\right\}. \tag{41}$$

Next set $\eta_{n_2}(\delta) := (K_\xi/\delta \vee 1)^3 \log(n_2)/\sqrt{n_2}$ and note that for any $\epsilon, \delta > 0$ such that $z_{1-\alpha} - \epsilon - 3\delta/\underline{\sigma} > 0$, Lemma C.14 implies that there is a universal $C < \infty$ and standard normal $\mathbb{Z}$ such that

$$\begin{aligned}
&\limsup_{n_2\to\infty} \sup_{P\in\mathbf{P}_0} \max_{k\in J^*} \sup_{y\in\mathcal{Y}(P;J^*)} P\left\{\mathbb{G}_{k,n}^{(2)}(y) > z_{1-\alpha} - \epsilon\right\} \\
&\quad\leq \limsup_{n_2\to\infty} \sup_{P\in\mathbf{P}_0} \max_{1\leq j\leq d_1+1} \sup_{\|y\|_1\leq 1} \{(1 - P\{\sigma_j(P;y)\mathbb{Z} \leq (z_{1-\alpha} - \epsilon)(\sigma_j(P;y) \vee \underline{\sigma}) - 3\delta\}) + C\eta_{n_2}(\delta)\} \\
&\quad\leq P\Big\{\mathbb{Z} > z_{1-\alpha} - \epsilon - \frac{3\delta}{\underline{\sigma}}\Big\},
\end{aligned} \tag{42}$$

where in the final inequality we used that $z_{1-\alpha} - \epsilon - 3\delta/\underline{\sigma} > 0$ and $\sigma_j(P;y) \leq \sigma_j(P;y) \vee \underline{\sigma}$. Since $\epsilon, \delta > 0$ are arbitrary and $P\{\mathbb{Z} > z_{1-\alpha}\} = \alpha$, the claim of the theorem follows from results (38), (40), (41), and (42). ■

B.4 Proof of Lemma 4.1

Proof. Part (a) of the lemma is immediate from the definition of $\mathcal{Y}(P;J^*)$.

To establish parts (b) and (c) we first establish a number of preliminary steps that are common to both arguments. First define the event $\mathcal{E}_n$ on which the optimization problem in (28) is feasible by setting

$$\mathcal{E}_n := \{\text{There is } y \in \mathbb{R}^p \text{ s.t. } \|y\|_1 \leq 1 \text{ and } \sqrt{n_1}(\hat{b}_{j,n}^{(1)})'\hat{M}_{0,n}^{(1)}y \geq \hat{\omega}_{j,n} \text{ for all } j \in (J^*)^c\} . \tag{43}$$

Further define the positive scalar $t_n := \min_{j\in(J^*)^c} \hat{\omega}_{j,n}$ and set $\Delta_n(P)$ to equal the maximal deviation

$$\Delta_n(P) := \max_{1\leq j\leq d_1+1} \sup_{\|y\|_1\leq 1} |\sqrt{n_1}\big((\hat{b}_{j,n}^{(1)})'\hat{M}_{0,n}^{(1)}y - b_j(P)'M_0(P)y\big)| . \tag{44}$$

Next, note that since $\hat{y}_n^{(1)} = 0$ whenever the event $\mathcal{E}_n^c$ occurs we obtain from the definition of t_n that

$$\inf_{P\in\mathbf{P}_0:\mathcal{Y}(P;J^*)=\emptyset} P\left\{\hat{y}_n^{(1)}=0\right\} \geq \inf_{P\in\mathbf{P}_0:\mathcal{Y}(P;J^*)=\emptyset} P\left\{\{Z_i\}_{i\in I_{n,1}}\in\mathcal{E}_n^c\right\}$$
$$\geq \inf_{P\in\mathbf{P}_0:\mathcal{Y}(P;J^*)=\emptyset} P\left\{\min_{j\in(J^*)^c}\sqrt{n_1}(\hat{b}_{j,n}^{(1)})'\hat{M}_{0,n}^{(1)}y<t_n \text{ for all } \|y\|_1\leq 1\right\} \geq \inf_{P\in\mathbf{P}} P\left\{\Delta_n(P)<t_n\right\} \quad (45)$$

where the final inequality follows from the definition of $\Delta_n(P)$ and the fact that $\mathcal{Y}(P;J^*)$ being empty implies that for every y with $\|y\|_1\leq 1$ there is a $j\in(J^*)^c$ for which $b_j(P)'M_0(P)y\leq 0$. Moreover, since $0\notin\mathcal{Y}(P;J^*)$ and the event $\mathcal{E}_n^c$ implies that $\hat{y}_n^{(1)}=0$ we obtain by definition of $\mathcal{Y}(P;J^*)$ that

$$\inf_{P\in\mathbf{P}_0:\mathcal{Y}(P;J^*)\neq\emptyset} P\left\{\hat{y}_n^{(1)}\in\mathcal{Y}(P;J^*)\right\} = \inf_{P\in\mathbf{P}_0} P\left\{\min_{j\in(J^*)^c}\sqrt{n_1}b_j(P)'M_0(P)\hat{y}_n^{(1)}>0 \text{ and } \{Z_i\}_{i\in I_{n,1}}\in\mathcal{E}_n\right\}$$
$$\geq \inf_{P\in\mathbf{P}} P\left\{t_n>\Delta_n(P) \text{ and } \{Z_i\}_{i\in I_{n,1}}\in\mathcal{E}_n\right\}, \quad (46)$$

where the inequality follows from $\{Z_i\}_{i\in I_{n,1}}\in\mathcal{E}_n$ implying that $\sqrt{n_1}(\hat{b}_{j,n}^{(1)})'\hat{M}_{0,n}^{(1)}\hat{y}_n\geq\hat{\omega}_{j,n}$ for all $j\in(J^*)^c$ and the definitions of t_n and $\Delta_n(P)$. Furthermore, also note that since the event $\mathcal{E}_n^c$ implies $\hat{y}_n^{(1)}=0$ we have

$$\inf_{P\in\mathbf{P}_0:\mathcal{Y}(P;J^*)\neq\emptyset} P\left\{\hat{y}_n^{(1)}=0\right\} \geq \inf_{P\in\mathbf{P}_0:\mathcal{Y}(P;J^*)\neq\emptyset} P\left\{t_n>\Delta_n(P) \text{ and } \{Z_i\}_{i\in I_{n,1}}\in\mathcal{E}_n^c\right\}. \quad (47)$$

Results (45), (46), and (47) and the events $\hat{y}_n^{(1)}\in\mathcal{Y}(P;J^*)$ and $\hat{y}_n^{(1)}=0$ being mutually exclusive yield

$$\inf_{P\in\mathbf{P}_0} P\left\{\{\hat{y}_n^{(1)}\in\mathcal{Y}(P;J^*)\}\cup\{\hat{y}_n^{(1)}=0\}\right\} \geq \inf_{P\in\mathbf{P}} P\left\{t_n>\Delta_n(P)\right\}. \quad (48)$$

Parts (b) and (c) of the lemma therefore follow from (48) provided that we can show that

$$\lim_{n\to\infty}\inf_{P\in\mathbf{P}} P\left\{t_n>\Delta_n(P)\right\}=1. \quad (49)$$

In particular, part (b) follows by noting that Lemma C.15(a) implies $\Delta_n(P)=O_P((pd_1)^{1/3})$ uniformly in $P\in\mathbf{P}$ and therefore the assumption $\max_{j\in(J^*)^c}\{(pd_1)^{1/3}/\hat{\omega}_{j,n}\}=o_P(1)$ uniformly in $P\in\mathbf{P}$ implies (49). Similarly, part (c) follows from Lemma C.15(b) implying $\Delta_n(P)=O_P(C_{\xi,p}\sqrt{\log(p+d_1)})$ uniformly in $P\in\mathbf{P}$, which together with $\max_{j\in(J^*)^c}\{C_{\xi,p}\sqrt{\log(p+d_1)}/\hat{\omega}_{j,n}\}=o_P(1)$ uniformly in $P\in\mathbf{P}$ yields (49). ■

C Auxiliary Lemmas

C.1 Lemmas for Section 3

The following lemmas are employed in the proof of Theorem 3.1. In what follows we let $\|(A_0,A_1,b)\|_{\sup}$ denote the largest absolute value of the entries of the triple (A_0,A_1,b), and $\mathbf{C}_0$, $\mathbf{C}_0^{\mathrm{RD}}$, $\bar{\mathbf{C}}_0$, and $\mathbf{C}^{\mathrm{NP}}$ be as defined in (16), (18), (17), and (36) respectively.

Lemma C.1. $\text{cl}(\mathbf{C}_0) \cap (\mathbf{C}^{\text{RD}})^c \subseteq \bar{\mathbf{C}}_0$.

PROOF. Fix some triple $(A_0, A_1, b) \in \text{cl}(\mathbf{C}_0) \cap (\mathbf{C}^{\text{RD}})^c$. Since this triple lies in the closure of $\mathbf{C}_0$, there exists a sequence of triples $\{(A_{0n}, A_{1n}, b_n)\}_{n\geq 1} \in \mathbf{C}_0$ which converges to it. Let M_{0n} denote the annihilator matrix for A_{0n}. Because $(A_0, A_1, b) \in (\mathbf{C}^{\text{RD}})^c$, we have $\text{rank}(A_0) = d_0$. Since the rank function is lower-semicontinuous (see equation (4) in Hiriart-Urruty and Le, 2013), it follows that $\text{rank}(A_{0n}) = d_0$ for n large enough and thus $M_{0n} = I - A_{0n}(A'_{0n}A_{0n})^{-1}A'_{0n}$, which implies $M_{0n} \to M_0$.

Let $c_1, \ldots, c_{d_1}$ denote the columns of M_0A_1 and similarly set $c_{n,1}, \ldots, c_{d_1,n}$ to be the columns of $M_{0n}A_{1n}$. Since $(A_{0n}, A_{1n}, b_n) \in \mathbf{C}_0$, Lemma 3.1 and Farkas' lemma imply that for each fixed $y \in \mathbb{R}^p$, either there exists a $1 \leq j(n) \leq d_1$ such that $c'_{n,j(n)}y < 0$ or $b'_nM_{0n}y \geq 0$. If for all n large enough $b'_nM_{0n}y \geq 0$, then $(M_{0n}, b_n) \to (M_0, b)$ implies that $b'M_0y \geq 0$ by continuity. Otherwise, since d_1 is finite, there exists $1 \leq j^* \leq d_1$ and a subsequence (indexed by n_k) along which $j(n_k) \equiv j^*$; that is, $c'_{n_k,j^*}y < 0$ for all k. It then follows by continuity again that $c_{n_k,j^*} \to c_{j^*}$, and thus $c'_{n_k,j^*}y \to c'_{j^*}y \leq 0$. Since y was arbitrary, we obtain by the definition of $\bar{\mathbf{C}}_0$ in (17) that $(A_0, A_1, b) \in \bar{\mathbf{C}}_0$, as desired. ■

Lemma C.2. $\mathbf{C}^{\text{RD}} \subseteq \text{cl}(\mathbf{C}_0)$.

PROOF. Fix some triple $(A_0, A_1, b) \in \mathbf{C}^{\text{RD}}$. To establish the result, we construct a new triple $(A_0^\epsilon, A_1, b) \in \mathbf{C}_0$ such that $\|(A_0^\epsilon, A_1, b) - (A_0, A_1, b)\|_{\text{sup}} < \|b\|_\infty\epsilon$ for any $\epsilon > 0$. Let v_j for $1 \leq j \leq d_0$ denote the columns of A_0. Since $(A_0, A_1, b) \in \mathbf{C}^{\text{RD}}$, there exists scalars λ_j for $1 \leq j \leq d_0$, not all zero, such that $\sum_{1\leq j\leq d_0} \lambda_j v_j = 0$. Assume $\lambda_1 \neq 0$ without loss of generality and define A_0^ϵ as the matrix with columns given by $v_1 + b\epsilon, v_2, \ldots, v_{d_0}$. Letting $x_0 = (1/\epsilon, \lambda_2/(\lambda_1\epsilon), \ldots, \lambda_d/(\lambda_1\epsilon))'$ and $x_1 = 0$, it is immediate that

$$A_0^\epsilon x_0 + A_1x_1 = b \ ,$$

which implies $(A_0^\epsilon, A_1, b) \in \mathbf{C}_0$. Moreover, by construction, $\|(A_0, A_1, b) - (A_0^\epsilon, A_1, b)\|_{\text{sup}} < \|b\|_\infty\epsilon$. Since $\epsilon > 0$ is arbitrary, we conclude that $\mathbf{C}^{\text{RD}} \subseteq \text{cl}(\mathbf{C}_0)$. ■

Lemma C.3. $\bar{\mathbf{C}}_0 \cap (\mathbf{C}^{\text{NP}})^c \subseteq \mathbf{C}_0$.

PROOF. Suppose $(A_0, A_1, b) \in \bar{\mathbf{C}}_0 \cap (\mathbf{C}^{\text{NP}})^c$ and let $c_1, \ldots, c_{d_1}$ denote the columns of M_0A_1. Since the triple $(A_0, A_1, b) \in \bar{\mathbf{C}}_0$, it follows that for all $y \in \mathbb{R}^p$, if $c'_jy > 0$ for $1 \leq j \leq d_1$, then $b'M_0y \geq 0$. Moreover, since $(A_0, A_1, b) \in (\mathbf{C}^{\text{NP}})^c$ implies the cone $K := \{M_0A_1x_1 : x_1 \geq 0\}$ is pointed, we can apply Lemma C.5 to conclude that for all $y \in \mathbb{R}^p$, if $c'_jy \geq 0$ for all $1 \leq j \leq d_1$, then $b'M_0y \geq 0$. It then follows from Farkas' lemma and Lemma 3.1 that $(A_0, A_1, b) \in \mathbf{C}_0$ as desired. ■

Lemma C.4. $\mathbf{C}^{\text{NP}} \subseteq \text{cl}(\mathbf{C}_0)$.

PROOF. Fix some triple $(A_0, A_1, b) \in \mathbf{C}^{\text{NP}}$. We will construct a new triple $(A_0, A_1^\epsilon, b) \in \mathbf{C}_0$ such that $\|(A_0, A_1^\epsilon, b) - (A_0, A_1, b)\|_{\text{sup}} < \|M_0b\|_\infty\epsilon$ for any $\epsilon > 0$. To this end, note that since $(A_0, A_1, b) \in \mathbf{C}^{\text{NP}}$ implies $\{M_0A_1x_1 : x_1 \geq 0\}$ is not pointed, it follows that there is a $z \neq 0$ such that $z = M_0A_1x_1 = -M_0A_1\tilde{x}_1$ for some $x_1, \tilde{x}_1 \geq 0$. In particular, letting $c_1, \ldots, c_{d_1}$ denote the columns of M_0A_1, we conclude that there

are scalars $\lambda_j \geq 0$ for $1 \leq j \leq d_1$, not all zero, such that $\sum_{1 \leq j \leq d_1} \lambda_j c_j = 0$. Assume $\lambda_1 > 0$ without loss of generality and define $A_1^\epsilon = \Pi_0 A_1 + (c_1 + \epsilon M_0 b, c_2, \ldots, c_{d_1})$, where recall Π_0 denotes the projection matrix onto the column space of A_0. Letting $\hat{x}_1 = (1/\epsilon, \lambda_2/(\lambda_1 \epsilon), \ldots, \lambda_d/(\lambda_1 \epsilon))'$ we then obtain by direct calculation

$$-\Pi_0 A_1 \hat{x}_1 + \Pi_0 b + A_1^\epsilon \hat{x}_1 = b$$

and $\|(A_0, A_1^\epsilon, b) - (A_0, A_1, b)\|_{\sup} \leq \|M_0 b\|_\infty \epsilon$. Further note $-\Pi_0 A_1 \hat{x}_1 + \Pi_0 b = A_0 \hat{x}_0$ for some $\hat{x}_0$. Therefore, since $\hat{x}_1 \geq 0$ by construction, it follows that $(A_0, A_1^\epsilon, b) \in \mathbf{C}_0$. Because ϵ is arbitrary, the result follows. ■

Lemma C.5. *Let A be a $p \times d$ matrix, b be a $p \times 1$ vector, and $a_1, \ldots, a_d$ denote the columns of A. Suppose the cone $K = \{Ax : x \geq 0\}$ is pointed (meaning $K \cap (-K) = \{0\}$). Then, the following are equivalent:*

(a) For all $y \in \mathbb{R}^p$, if $a_j' y \geq 0$ for $1 \leq j \leq d$, then $b'y \geq 0$.

(b) For all $y \in \mathbb{R}^p$, if $a_j' y > 0$ for $1 \leq j \leq d$, then $b'y \geq 0$.

Proof. It is clear that if statement (a) is true, then (b) must be true as well. To show (b) implies (a), note that K is a nonempty closed convex cone, where closedness follows from Corollary 2.5 and Theorem 4.9 in Bertsimas and Tsitsiklis (1997). Because K is finitely-generated, its dual cone K^* satisfies $K^* = \{y \in \mathbb{R}^p : a_j' y \geq 0 \text{ for } 1 \leq j \leq d\}$. Similarly, Lemma C.6(a) implies the interior of K^* equals $K_o^* = \{y \in \mathbb{R}^p : a_j' y > 0 \text{ for } 1 \leq j \leq d\}$. Moreover, since K is pointed, Lemma C.6(c) further implies that $K_o^* \neq \emptyset$, which together with Lemma 6.3 in Rockafellar (1970) allows us to conclude that $\text{cl}(K_o^*) = K^*$. Hence, since part (a) states $b'y \geq 0$ for all $y \in K^*$ and part (b) states $b'y \geq 0$ for all $y \in K_o^*$, the fact that part (b) implies part (a) follows from continuity of $y \mapsto b'y$ and $\text{cl}(K_o^*) = K^*$. ■

Lemma C.6. *Let $K \neq \{0\}$ be a nonempty closed convex cone in $\mathbb{R}^k$ and recall that its dual cone K^* is defined to be $K^* := \{y \in \mathbb{R}^k : y'x \geq 0 \text{ for all } x \in K\}$. Then, it follows that:*

(a) The interior of K^* relative to $\mathbb{R}^k$ equals $K_o^* = \{y \in \mathbb{R}^k : y'x > 0 \text{ for all } x \in K \setminus \{0\}\}$.

(b) $K^{**} = K$ for K^{**} the dual cone of K^*.

(c) K is pointed if and only if $K_o^* \neq \emptyset$.

Proof. The claims of the lemma are contained in Exercise 2.31 in Boyd and Vandenberghe (2004) but we include a proof for completeness and correct a mistake in the statement.

To show (a), fix a $y \in K_o^*$, so that $y'x > 0$ for all $x \in K \setminus \{0\}$. Then, $y'x > 0$ for all $x \in K \cap \{x \in \mathbb{R}^k : \|x\| = 1\}$. The function $x \mapsto y'x$ is continuous and attains its minimum in $K \cap \{x \in \mathbb{R}^k : \|x\| = 1\}$ because the set is compact. Denote the minimum by m_y. Then, for all $\|u\| < m_y/2$, $(y+u)'x > 0$ for all $x \in K \cap \{x \in \mathbb{R}^k : \|x\| = 1\}$, and therefore $(y+u)'x \geq 0$ for all $x \in K$. Therefore, the ball about y of radius $m_y/2$ is contained in K^*, which implies K_o^* is contained in the interior of K^*. For the converse direction, suppose $y \in K^*$ and $y'x = 0$ for some $x \in K \setminus \{0\}$. Pick u such that $u'x < 0$, which is possible because $x \neq 0$. For such a u, $(y + \epsilon u)'x < 0$, which implies $y + \epsilon u \notin K^*$ for any $\epsilon > 0$. It follows that y is not in the interior of K^* and therefore that the interior of K^* is contained in K_o^*, which establishes part (a).

Part (b) follows from Theorem 14.1 in Rockafellar (1970).

To show (c), we prove its contrapositive: K is not pointed if and only if $K_o^* = \emptyset$. First suppose K is not pointed. Then, there exists $0 \neq x \in K \cap (-K)$, which implies $x \in K$ and $-x \in K$. Since there cannot exist a y such that $y'x > 0$ and $y'(-x) > 0$, it follows that $K_o^* = \emptyset$. For the converse direction, suppose $K_o^* = \emptyset$. By part (a), it then follows that K^* is a convex cone with empty interior relative to $\mathbb{R}^k$. Therefore, K^* must be contained in a proper subspace of $\mathbb{R}^k$, which implies there is a $0 \neq z \in \mathbb{R}^k$ such that $z'y = 0$ for all $y \in K^*$. In particular, we must then have $(-z)'y = 0$ for all $y \in K^*$, so both z and $-z$ lie in K^{**}, which implies K^{**} is not pointed and, by part (b), that K is not pointed either. ■

Lemma C.7. *Let Σ be a $k \times k$ symmetric positive semi-definite matrix and $\mu_n \to \mu \in \mathbb{R}^k$ as $n \to \infty$. Further suppose $\mu_n - \mu \in \text{range}(\Sigma)$ for all n. Then, $\text{TV}(N(\mu_n, \Sigma), N(\mu, \Sigma)) \to 0$ as $n \to \infty$.*

Proof. Let H denote the Cameron-Martin space of $\gamma := N(\mu, \Sigma)$, and note that by Lemma 2.4.1 in Bogachev (1998), $H = \text{range}(\Sigma)$ and the Cameron-Martin norm $\|\cdot\|_H$ satisfies $\|h\|_H^2 = h'\Sigma^\dagger h$ for any $h \in H$ and $\Sigma^\dagger$ the Moore-Penrose pseudoinverse of Σ. By Lemma 2.4.4 in Bogachev (1998) and $\mu_n - \mu \in H$ we then obtain

$$2 - 2\exp\{-\frac{1}{8}(\mu_n - \mu)'\Sigma^\dagger(\mu_n - \mu)\} \leq \text{TV}(N(\mu_n, \Sigma), N(\mu, \Sigma)) \leq 2(1 - \exp\{-\frac{1}{4}(\mu_n - \mu)'\Sigma^\dagger(\mu_n - \mu)\})^{1/2},$$

and therefore the claim of the lemma follows from $\mu_n \to \mu$. ■

C.2 Lemmas for Section 4

For notational simplicity, in this subsection we suppress the superscript (k) indicating sample split for all sample objects. Further recall that for any $u = (u_1, \ldots, u_r)' \in \mathbb{R}^r$, we set $\|u\|_q = (\sum_{i=1}^r |u_i|^q)^{1/q}$ for any $1 \leq q < \infty$ and $\|u\|_\infty = \max_{1 \leq i \leq r} |u_i|$. In addition, for any $1 \leq r_1, r_2 \leq \infty$ and matrix V, set $\|V\|_{r_1, r_2} = \sup_{\|u\|_{r_1} \leq 1} \|Vu\|_{r_2}$ and $\underline{s}(V)$ and $\bar{s}(V)$ to denote the smallest and largest singular values of V.

Lemma C.8. *Suppose Assumptions 4.1 and 4.2 hold, $\sup_{\|y\|_1 \leq 1} \text{Var}_P[y'\xi_j(Z, P)] < B < \infty$ for all $P \in \mathbf{P}$, $1 \leq j \leq d_1 + 1$ and $p \geq 1$, and that $\hat{V}_{j,n}$ is such that uniformly in $P \in \mathbf{P}$ and $1 \leq j \leq 1 + d_1$ we have*

$$\|\hat{V}_{j,n} - V_j(P)\|_{2,2} = O_P(\delta_n) \ .$$

Further suppose that $(K_{1,p}K_{2,p}^2 \vee \bar{s}_p^2)(K_{0,p} \vee K_{1,p}) \log(1+p) = o(n)$ and $(K_{1,p}^{1/2}K_{2,p} \vee \bar{s}_p)a_n = o(n)$, and define

$$r_n := \frac{K_{2,p}(\log(1+p)(K_{0,p} \vee K_{1,p}) + a_n)}{\sqrt{n}}$$
$$q_n := \sqrt{K_{1,p}}K_{2,p}\left(\sqrt{\frac{(K_{0,p} \vee K_{1,p})\log(1+p)}{n}} + \frac{a_n}{n}\right) + K_{2,p}^2\delta_n \ .$$

It then follows that uniformly in $P \in \mathbf{P}$, $1 \leq j \leq d_1 + 1$, and $y \in \mathbb{R}^p$ with $\|y\|_1 \leq 1$ we have that

$$\left|\frac{\sqrt{n}(\hat{b}_{j,n}'\hat{M}_{0,n}y - b_j(P)'M_0(P)y)}{\hat{\sigma}_{j,n}(y)} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{y'\xi_j(Z_i, P)}{(\sigma_j(P; y) \vee \underline{\sigma})}\right| = O_P\left(\frac{r_n}{\underline{\sigma}} + \frac{q_n}{\underline{\sigma}^3}\right) \ .$$

PROOF. To begin, note that Lemma C.9 implies that uniformly in $P \in \mathbf{P}$ and $1 \leq j \leq d_1 + 1$ we have

$$\sup_{\|y\|_2 \leq 1} |\sqrt{n}(\hat{b}'_{j,n}\hat{M}_{0,n}y - b_j(P)'M_0(P)y) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_j(Z_i,P)'y| = O_P(r_n) \ . \tag{50}$$

Next note that $\hat{\sigma}_{j,n}(y) \geq \underline{\sigma} > 0$ by construction together with Lemma C.18 allow us to conclude that

$$\begin{aligned}\sup_{\|y\|_1 \leq 1}\Big|\frac{1}{\hat{\sigma}_{j,n}(y)} - \frac{1}{\sigma_j(P;y) \vee \underline{\sigma}}\Big| &\leq \sup_{\|y\|_1 \leq 1}\frac{1}{2\underline{\sigma}^3}|\hat{\sigma}^2_{j,n}(y) - (\sigma^2_j(P;y) \vee \underline{\sigma}^2)| \\ &\leq \sup_{\|y\|_1 \leq 1}\frac{1}{2\underline{\sigma}^3}|\hat{D}_{j,n}(y)'\hat{V}_{j,n}\hat{D}_{j,n}(y) - \sigma^2_j(P;y)| = O_P\Big(\frac{q_n}{\underline{\sigma}^3}\Big), \end{aligned} \tag{51}$$

where the second inequality follows from $|(x_1 \vee \underline{\sigma}^2) - (x_2 \vee \underline{\sigma}^2)| \leq |x_1 - x_2|$ for any $x_1, x_2 \in \mathbb{R}$, and the final result holds uniformly in $P \in \mathbf{P}$ and $1 \leq j \leq d_1 + 1$ by Lemma C.11. Moreover, note that we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} y'\xi_j(Z_i,P) = O_P(1) \tag{52}$$

uniformly in $P \in \mathbf{P}$, $1 \leq j \leq d_1 + 1$, and $y \in \mathbb{R}^p$ with $\|y\|_1 \leq 1$ due to Markov's inequality and the condition $\sup_{\|y\|_1 \leq 1} \mathrm{Var}_P[y'\xi_j(Z,P)] < B < \infty$. Since $\hat{\sigma}_{j,n}(y) \geq \underline{\sigma}$, results (50), (51), and (52) then allow us to conclude uniformly in $P \in \mathbf{P}$, $1 \leq j \leq d_1 + 1$, and $y \in \mathbb{R}^p$ with $\|y\|_1 \leq 1$ that

$$\begin{aligned}&\frac{\sqrt{n}(\hat{b}'_{j,n}\hat{M}_{0,n}y - b_j(P)'M_0(P)y)}{\hat{\sigma}_{j,n}(y)} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{y'\xi_j(Z_i,P)}{(\sigma_j(P;y) \vee \underline{\sigma})} \\ &\qquad = \Big(\frac{1}{\hat{\sigma}_{j,n}(y)} - \frac{1}{\sigma_j(P;y) \vee \underline{\sigma}}\Big)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_j(Z_i,P)'y + O_P\Big(\frac{r_n}{\underline{\sigma}}\Big) = O_P\Big(\frac{r_n}{\underline{\sigma}} + \frac{q_n}{\underline{\sigma}^3}\Big), \end{aligned} \tag{53}$$

which establishes the claim of the lemma. ■

Lemma C.9. *Let Assumptions 4.1 and 4.2 hold, and suppose* $\bar{s}_p^2(K_{0,p} \vee K_{1,p})\log(1+p)/n = o(1)$ *and* $\bar{s}_p a_n/n = o(1)$. *Then, it follows that uniformly in* $P \in \mathbf{P}$ *and* $1 \leq j \leq d_1 + 1$ *we have*

$$\sup_{\|y\|_2 \leq 1}\Big|\sqrt{n}(\hat{b}'_{j,n}\hat{M}_{0,n}y - b_j(P)'M_0(P)y) - \frac{1}{\sqrt{n}}\sum_{1 \leq i \leq n}\xi_j(Z_i,P)'y\Big| = O_P\Big(\frac{K_{2,p}(\log(1+p)(K_{0,p} \vee K_{1,p}) + a_n)}{\sqrt{n}}\Big) \ .$$

PROOF. Let $D(A_0(P))[H] := -(M_0(P)HA_0^{\dagger}(P) + (A_0^{\dagger}(P))'H'M_0(P))$ for any $p \times d_0$ matrix H and define

$$\begin{aligned}\hat{S}_{j,n}(P) &:= M_0(P)\sqrt{n}(\hat{b}_{j,n} - b_j(P)) + D(A_0(P))[\sqrt{n}(\hat{A}_{0,n} - A_0(P))]b_j(P) \\ S^*_{j,n}(P) &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_j(Z_i,P) \ . \end{aligned}$$

Next note that Lemma C.12, Assumption 4.2(a), $\bar{s}_p \geq 1$, and $(K_{0,p} \vee K_{1,p})\log(1+p)/n = o(1)$ by hypothesis allow us to conclude that uniformly in $P \in \mathbf{P}$ we have

$$\|A_0(P)\|_{2,2}\|\hat{A}_{0,n} - A_0(P)\|_{2,2} + \|\hat{A}_{0,n} - A_0(P)\|^2_{2,2} = O_P\Big(\bar{s}_p\Big(\sqrt{\frac{(K_{0,p} \vee K_{1,p})\log(1+p)}{n}} + \frac{a_n}{n}\Big)\Big) \ . \tag{54}$$

Together with Assumption 4.2(b), $\bar{s}_p^2(K_{0,p} \vee K_{1,p}) \log(1+p)/n = o(1)$ and $\bar{s}_p a_n/n = o(1)$, result (54) implies that $\|A_0(P)\|_{2,2}\|\hat{A}_{0,n} - A_0(P)\|_{2,2} + \|\hat{A}_{0,n} - A_0(P)\|_{2,2}^2 < \underline{s}(A_0(P))^2/2$ with probability tending to one uniformly in $P \in \mathbf{P}$. We may therefore apply Lemma C.13(a) together with Lemma C.12 and Assumptions 4.2(b)(c) to conclude

$$\sup_{y \in \mathbb{R}^p : \|y\|_2 \leq 1} |\sqrt{n}(\hat{b}_{j,n}' \hat{M}_{0,n} y - b_j(P)' M_0(P) y) - \hat{S}_{j,n}(P)' y| = O_P\Big(\frac{K_{2,p}}{\sqrt{n}}((K_{0,p} \vee K_{1,p}) \log(1+p) + \frac{a_n^2}{n^{3/2}})\Big) \quad (55)$$

uniformly in $P \in \mathbf{P}$ and $1 \leq j \leq d_1 + 1$. Furthermore, by definition of $\xi_j(Z_i, P)$ we also have that

$$\begin{aligned}
\|\hat{S}_{j,n}(P) - S_{j,n}^*(P)\|_2 \leq & \Big\| M_0(P)(\sqrt{n}(\hat{b}_{j,n} - b_j(P)) - \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \varphi_j(Z_i, P)) \Big\|_2 \\
& + \Big\| M_0(P)\Big(\sqrt{n}(\hat{A}_{0,n} - A_0(P)) - \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \Psi(Z_i, P)\Big) A_0^\dagger(P) b_j(P) \Big\|_2 \\
& + \Big\| A_0^\dagger(P)'\Big(\sqrt{n}(\hat{A}_{0,n} - A_0(P)) - \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \Psi(Z_i, P)\Big)' M_0(P) b_j(P) \Big\|_2 . \quad (56)
\end{aligned}$$

Next, note that using Lemma C.20, $\|M_0(P)\|_{2,2} \leq 1$ because $M_0(P)$ is a projection matrix, result (56), and Assumptions 4.1(a) and 4.2(b)(c) we can conclude that uniformly in $P \in \mathbf{P}$ and $1 \leq j \leq d_1 + 1$

$$\|\hat{S}_{j,n}(P) - S_{j,n}^*(P)\|_2 = O_P\Big(\frac{a_n K_{2,p}}{\sqrt{n}}\Big) . \quad (57)$$

The claim of the lemma therefore follows from results (55), (57), the Cauchy-Schwarz inequality and $a_n/\sqrt{n} = o(1)$ by Assumption 4.1. ■

Lemma C.10. *Suppose* $\operatorname{rank}(A_0(P)) = d_0$ *and* $V_j(P) < \infty$. *Then,* $\operatorname{Var}_P[\xi_j(Z_i, P)' y] = \sigma_j^2(P; y)$.

Proof. Using $\operatorname{vec}(ABC) = (C' \otimes A)\operatorname{vec}(B)$, $(A \otimes B)' = A' \otimes B'$, and the definition of $D_j(P; y)$ we obtain

$$\begin{aligned}
&\xi_j(Z_i, P)' y \\
&\quad = y' M_0(P) \varphi_j(Z_i, P) - y' M_0(P) \Psi(Z_i, P) A_0^\dagger(P) b_j(P) - y' A_0^\dagger(P)' \Psi(Z_i, P)' M_0(P) b_j(P) \\
&\quad = y' M_0(P) \varphi_j(Z_i, P) - \operatorname{vec}\left(y' M_0(P) \Psi(Z_i, P) A_0^\dagger(P) b_j(P)\right) - \operatorname{vec}\left(b_j(P)' M_0(P) \Psi(Z_i, P) A_0^\dagger(P) y\right) \\
&\quad = y' M_0(P) \varphi_j(Z_i, P) - \{(A_0^\dagger(P) b_j(P))' \otimes (y' M_0(P)) + (A_0^\dagger(P) y)' \otimes (b_j(P)' M_0(P))\} \operatorname{vec}(\Psi(Z_i, P)) \\
&\quad = D_j(P; y)' \begin{pmatrix} \operatorname{vec} \Psi(Z_i, P) \\ \varphi_j(Z_i, P) \end{pmatrix} ,
\end{aligned}$$

which establishes the claim of the lemma. ■

Lemma C.11. *Suppose Assumptions 4.1 and 4.2 hold, and suppose* $\sup_{\|y\|_1 \leq 1} \operatorname{Var}_P[y' \xi_j(Z, P)] < B < \infty$ *for all* $P \in \mathbf{P}$, $1 \leq j \leq d_1 + 1$ *and* $p \geq 1$. *If* $\hat{V}_{j,n}$ *is such that uniformly in* $P \in \mathbf{P}$ *and* $1 \leq j \leq 1 + d_1$ *we have*

$$\|\hat{V}_{j,n} - V_j(P)\|_{2,2} = O_P(\delta_n),$$

and in addition $(K_{1,p}K_{2,p}^2 \vee \bar{s}_p^2)(K_{0,p} \vee K_{1,p})\log(1+p) = o(n)$ *and* $(K_{1,p}^{1/2}K_{2,p} \vee \bar{s}_p)a_n = o(n)$*, then it follows*

$$\sup_{y\in\mathbb{R}^p:\|y\|_1\leq 1} |\hat{D}_{j,n}(y)'\hat{V}_{j,n}\hat{D}_{j,n}(y) - \sigma_j^2(P;y)| = O_P\Big(\sqrt{K_{1,p}}K_{2,p}(\sqrt{\frac{(K_{0,p}\vee K_{1,p})\log(1+p)}{n}} + \frac{a_n}{n}) + K_{2,p}^2\delta_n\Big)$$

uniformly in $P \in \mathbf{P}$ *and* $1 \leq j \leq d_1 + 1$.

Proof. To begin the proof, first note that direct calculation yields the following decomposition

$$\begin{aligned}
&\hat{D}_{j,n}(y)'\hat{V}_{j,n}\hat{D}_{j,n}(y) - D_j(P;y)'V_j(P)D_j(P;y) \\
&= 2D_j(P;y)'V_j(P)(\hat{D}_{j,n}(y) - D_j(P;y)) + D_j(P;y)'(\hat{V}_{j,n} - V_j(P))D_j(P;y) \\
&\quad + (\hat{D}_{j,n}(y) - D_j(P;y))'V_j(P)(\hat{D}_{j,n}(y) - D_j(P;y)) + 2D_j(P;y)'(\hat{V}_{j,n} - V_j(P))(\hat{D}_{j,n}(y) - D_j(P;y)) \\
&\quad + (\hat{D}_{j,n}(y) - D_j(P;y))'(\hat{V}_{j,n} - V_j(P))(\hat{D}_{j,n}(y) - D_j(P;y)) \ . \qquad (58)
\end{aligned}$$

Next, let $V_j^{1/2}(P)$ denote the unique positive semi-definite square root of $V_j(P)$. Then note that (58) and $V_j(P) = V_j^{1/2}(P)V_j^{1/2}(P)$ imply that uniformly in P and $1 \leq j \leq d_1 + 1$ we have

$$\begin{aligned}
&|\hat{D}_{j,n}(y)'\hat{V}_{j,n}\hat{D}_{j,n}(y) - D_j(P;y)'V_j(P)D_j(P;y)| \\
&\quad \lesssim \|V_j^{1/2}(P)(\hat{D}_{j,n}(y) - D_j(P;y))\|_2 \cdot (\|V_j^{1/2}(P)D_j(P;y)\|_2 + \|V_j^{1/2}(P)(\hat{D}_{j,n}(y) - D_j(P;y))\|_2) \\
&\qquad + \|\hat{V}_{j,n} - V_j(P)\|_{2,2} \cdot (\|D_j(P;y)\|_2^2 + \|\hat{D}_{j,n}(y) - D_j(P;y)\|_2^2) \ . \qquad (59)
\end{aligned}$$

To control the terms on the right hand side (59), first use that $\|a \otimes b\|_2 = \|a\|_2 \cdot \|b\|_2$ for any vectors a and b, the triangle inequality, and the definition of $D_j(P;y)$ to obtain for any y with $\|y\|_1 \leq 1$ that

$$\begin{aligned}
\|D_j(P;y)\|_2 &\leq \|M_0(P)y\|_2 + \|A_0^\dagger(P)y\|_2 \cdot \|M_0(P)b_j(P)\|_2 + \|A_0^\dagger(P)b_j(P)\|_2 \cdot \|M_0(P)y\|_2 \\
&\leq 1 + 2\frac{\|b_j(P)\|_2}{\underline{s}(A_0(P))} \ , \qquad (60)
\end{aligned}$$

where the final inequality follows from Lemma C.20, $\|M_0(P)\|_{2,2} \leq 1$ because $M_0(P)$ is a projection matrix, and $\|y\|_2 \leq \|y\|_1 \leq 1$. Furthermore, by similar arguments we also obtain the following inequality

$$\begin{aligned}
\|\hat{D}_{j,n}(y) - D_j(P;y)\|_2 \leq \|(\hat{M}_{0,n} - M_0(P))y\|_2 + \|(\hat{A}_{0,n}^\dagger y) \otimes (\hat{M}_{0,n}\hat{b}_{j,n}) - (A_0(P)^\dagger y) \otimes (M_0(P)b_j(P))\|_2 \\
+ \|(\hat{A}_{0,n}^\dagger \hat{b}_{j,n}) \otimes (\hat{M}_{0,n}y) - (A_0^\dagger(P)b_j(P)) \otimes (M_0(P)y)\|_2 \ . \qquad (61)
\end{aligned}$$

Next note that by Weyl's perturbation inequality (see, e.g., Corollary III.2.6 in Bhatia (2013)) we have

$$\begin{aligned}
\underline{s}^2(\hat{A}_{0,n}) &\geq \underline{s}^2(A_0(P)) - \|\hat{A}_{0,n}'\hat{A}_{0,n} - A_0(P)'A_0(P)\|_{2,2} \\
&\geq \underline{s}^2(A_0(P)) - 2\|A_0(P)\|_{2,2}\|\hat{A}_{0,n} - A_0(P)\|_{2,2} - \|\hat{A}_{0,n} - A_0(P)\|_{2,2}^2 \ . \qquad (62)
\end{aligned}$$

In particular, since $\|\hat{A}_{0,n} - A_0(P)\|_{2,2}(1 \vee \|A_0(P)\|_{2,2}) = o_P(1)$ uniformly in $P \in \mathbf{P}$ by Lemma C.12,

$\bar{s}_p^2(K_{0,p} \vee K_{1,p}) \log(1+p) = o(n)$, and $\bar{s}_p a_n = o(n)$, it follows from (62) and Assumption 4.2(a)(b) that

$$\frac{1}{\underline{s}(\hat{A}_{0,n})} = O_P(1) \tag{63}$$

uniformly in $P \in \mathbf{P}$. Therefore, applying Theorem 2.5 in Chen et al. (2016) and Lemma C.12 yields that

$$\|\hat{M}_{0,n} - M_0(P)\|_{2,2} = O_P\Big(\sqrt{\frac{(K_{0,p} \vee K_{1,p}) \log(1+p)}{n}} + \frac{a_n}{n}\Big) \tag{64}$$

uniformly in $P \in \mathbf{P}$. Similarly, result (63), Assumption 4.2(b), Lemma C.12, and Theorem 4.1 in Wedin (1973) allow us to conclude that uniformly in $P \in \mathbf{P}$ we have

$$\|\hat{A}_{0,n}^\dagger - A_0^\dagger(P)\|_{2,2} = O_P\Big(\sqrt{\frac{(K_{0,p} \vee K_{1,p}) \log(1+p)}{n}} + \frac{a_n}{n}\Big)\,. \tag{65}$$

Next, again use that $\|a \otimes b\|_2 = \|a\|_2 \cdot \|b\|_2$ for any vectors a and b, Lemma C.20, Assumption 4.2(b), and that $\hat{M}_{0,n}$ is a projection matrix to obtain that for any y with $\|y\|_1 \le 1$ we have

$$\begin{aligned}
&\|\hat{A}_{0,n}^\dagger \hat{b}_{j,n} \otimes \hat{M}_{0,n} y - A_0^\dagger(P) b_j(P) \otimes M_0(P) y\|_2 \\
&\quad \le \|\hat{A}_{0,n}^\dagger \hat{b}_{j,n} - A_0^\dagger(P) b_j(P)\|_2 \cdot \|\hat{M}_{0,n} y\|_2 + \|A_0^\dagger(P) b_j(P)\|_2 \cdot \|(\hat{M}_{0,n} - M_0(P)) y\|_2 \\
&\quad \lesssim \|\hat{A}_{0,n}^\dagger - A_0^\dagger(P)\|_{2,2} \cdot \|b_j(P)\|_2 + \frac{1}{\underline{s}(\hat{A}_{0,n})} \|\hat{b}_{j,n} - b_j(P)\|_2 + \|b_j(P)\|_2 \cdot \|\hat{M}_{0,n} - M_0(P)\|_{2,2}\,.
\end{aligned} \tag{66}$$

Furthermore, by similar arguments it also follows that for any y with $\|y\|_1 \le 1$ we have the upper bound

$$\begin{aligned}
&\|(\hat{A}_{0,n}^\dagger y) \otimes (\hat{M}_{0,n} \hat{b}_{j,n}) - (A_0^\dagger(P) y) \otimes (M_0(P) b_j(P))\|_2 \\
&\quad \le \|(\hat{A}_{0,n}^\dagger - A_0^\dagger(P)) y\|_2 \cdot \|M_0(P) b_j(P)\|_2 + \|\hat{A}_{0,n}^\dagger y\|_2 \cdot \|\hat{M}_{0,n} \hat{b}_{j,n} - M_0(P) b_j(P)\|_2 \\
&\quad \le \|\hat{A}_{0,n}^\dagger - A_0^\dagger(P)\|_{2,2} \cdot \|b_j(P)\|_2 + \frac{1}{\underline{s}(\hat{A}_{0,n})} (\|\hat{M}_{0,n} - M_0(P)\|_{2,2} \cdot \|b_j(P)\|_2 + \|\hat{b}_{j,n} - b_j(P)\|_2).
\end{aligned} \tag{67}$$

Therefore, combining result (61) with the bounds in (66) and (67), and using results (63), (64), (65), Assumption 4.2(c), and Lemma C.12 we obtain uniformly in $P \in \mathbf{P}$ and $1 \le j \le d_1 + 1$ that

$$\sup_{\|y\|_2 \le 1} \|\hat{D}_{j,n}(y) - D_j(P; y)\|_2 = O_P\Big(K_{2,p}\big(\sqrt{\frac{(K_{0,p} \vee K_{1,p}) \log(1+p)}{n}} + \frac{a_n}{n}\big) + \sqrt{\frac{K_{1,p}}{n}}\Big). \tag{68}$$

Next, note that $\operatorname{vec}(ABC) = (C' \otimes A) \operatorname{vec}(B)$, $(A \otimes B)' = A' \otimes B'$ for any comformable matrices A, B, and C, the definitions of $\hat{D}_{j,n}(y)$ and $D_j(P; y)$, and direct calculation allow us to conclude that

$$\begin{aligned}
&(D_j(P; y) - \hat{D}_{j,n}(y))' \begin{pmatrix} \operatorname{vec} \Psi(Z_i, P) \\ \varphi_j(Z_i, P) \end{pmatrix} \\
&\quad = y'(\hat{A}_{0,n}^\dagger - A_0^\dagger(P))' \Psi(Z_i, P)' M_0(P) b_j(P) + y'(\hat{A}_{0,n}^\dagger)' \Psi(Z_i, P)' (\hat{M}_{0,n} \hat{b}_{j,n} - M_0(P) b_j(P)) \\
&\qquad + y'(\hat{M}_{0,n} - M_0(P)) \Psi(Z_i, P) A_0^\dagger(P) b_j(P) + y' \hat{M}_{0,n} \Psi(Z_i, P) (\hat{A}_{0,n}^\dagger \hat{b}_{j,n} - A_0^\dagger(P) b_j(P))
\end{aligned}$$

$$+ y'(M_0(P) - \hat{M}_{0,n})\varphi_j(Z_i, P) \ . \tag{69}$$

Also note that for any $v \in \mathbb{R}^p$ and $u \in \mathbb{R}^{d_0}$, it follows from $\|uu'\|_{2,2} \leq \|u\|_2^2$ and Assumption 4.1(b) that

$$E_P[(v'\Psi(Z,P)u)^2] = E_P[v'\Psi(Z,P)uu'\Psi(Z,P)'v] \leq \|u\|_2^2(v'E_P[\Psi(Z,P)\Psi(Z,P)']v) \leq K_{1,p}\|u\|_2^2\|v\|_2^2. \tag{70}$$

Therefore, using (69), (70), the definition of $V_j(P)$, Lemma C.19, Assumption 4.1(b), and that $\|\hat{M}_{0,n}\|_{2,2} \leq 1$ due to $\hat{M}_{0,n}$ being a projection matrix implies for any $y \in \mathbb{R}^p$ satisfying $\|y\|_1 \leq 1$ that

$$\begin{aligned}
&\|V_j^{1/2}(P)(\hat{D}_{j,n}(y) - D_j(P;y))\|_2^2 \\
&\quad \lesssim K_{1,p}\|(\hat{A}_{0,n}^\dagger - A_0^\dagger(P))'\|_{2,2}^2\|b_j(P)\|_2^2 + K_{1,p}\|(\hat{A}_{0,n}^\dagger)'\|_{2,2}^2\|\hat{M}_{0,n}\hat{b}_{j,n} - M_0(P)b_j(P)\|_2^2 \\
&\qquad + K_{1,p}\|\hat{M}_{0,n} - M_0(P)\|_{2,2}^2(1 + \|A_0^\dagger(P)b_j(P)\|_2^2) + K_{1,p}\|\hat{A}_{0,n}^\dagger\hat{b}_{j,n} - A_0^\dagger(P)b_j(P)\|_2^2 \ .
\end{aligned} \tag{71}$$

Further note that the arguments employed in (66) and (67) imply that uniformly in $P \in \mathbf{P}$ and $1 \leq j \leq d_1+1$

$$\begin{aligned}
\|\hat{M}_{0,n}\hat{b}_{j,n} - M_0(P)b_j(P)\|_2 &= O_P\Big(\sqrt{\frac{K_{1,p}}{n}} + K_{2,p}(\sqrt{\frac{(K_{0,p} \vee K_{1,p})\log(1+p)}{n}} + \frac{a_n}{n})\Big) \\
\|\hat{A}_{0,n}^\dagger\hat{b}_{j,n} - A_0^\dagger(P)b_j(P)\|_2 &= O_P\Big(\sqrt{\frac{K_{1,p}}{n}} + K_{2,p}(\sqrt{\frac{(K_{0,p} \vee K_{1,p})\log(1+p)}{n}} + \frac{a_n}{n})\Big) \ .
\end{aligned} \tag{72}$$

Therefore, combining results (71) and (72) and using Lemma C.20 and Assumptions 4.2(b)(c) imply

$$\|V_j^{1/2}(P)(\hat{D}_{j,n}(y) - D_j(P;y))\|_2 = O_P\Big(\sqrt{K_{1,p}}K_{2,p}(\sqrt{\frac{(K_{0,p} \vee K_{1,p})\log(1+p)}{n}} + \frac{a_n}{n})\Big) \tag{73}$$

uniformly in $P \in \mathbf{P}$ and $1 \leq j \leq d_1 + 1$. Finally, note that Lemma C.10 implies $\|V_j^{1/2}(P)D_j(P;y)\|_2^2 = \mathrm{Var}_P[y'\xi_j(Z,P)]$ for any y. Since $\mathrm{Var}_P[y'\xi_j(Z,P)]$ is uniformly bounded in $P \in \mathbf{P}$, $1 \leq j \leq d_1 + 1$, and y satisfying $\|y\|_1 \leq 1$, the lemma then follows from (59), (60), Assumptions 4.2(b)(c), and results (68), (73), and our rate conditions implying $\|\hat{D}_{j,n}(y) - D_j(P;y)\|_2 \vee \|V_j^{1/2}(P)(\hat{D}_{j,n}(y) - D_j(P;y))\|_2 = o_P(1)$ uniformly in $P \in \mathbf{P}$, $1 \leq j \leq d_1 + 1$ and y with $\|y\|_1 \leq 1$. ■

Lemma C.12. *Suppose Assumptions 4.1 and 4.2(b) hold. Then,*

(a) If $(K_{0,p} \vee K_{1,p})\log(p + d_0)/n = o(1)$, then it follows that uniformly in $P \in \mathbf{P}$ we have

$$\|\hat{A}_{0,n} - A_0(P)\|_{2,2} = O_P\left(\sqrt{\frac{(K_{0,p} \vee K_{1,p})\log(1+p)}{n}} + \frac{a_n}{n}\right) \ .$$

(b) Uniformly in $P \in \mathbf{P}$,

$$\max_{1 \leq j \leq d_1+1} \|\hat{b}_{j,n} - b_j(P)\|_2 = O_P\left(\sqrt{\frac{K_{1,p}(d_1+1)}{n}} + \frac{a_n}{n}\right) \ ,$$

and uniformly in $P \in \mathbf{P}$ *and* $1 \leq j \leq d_1 + 1$,

$$\|\hat{b}_{j,n} - b_j(P)\|_2 = O_P\left(\sqrt{\frac{K_{1,p}}{n}} + \frac{a_n}{n}\right) .$$

PROOF. To establish part (a), note that the triangle inequality and Assumption 4.1(a) imply that

$$\|\hat{A}_{0,n} - A_0(P)\|_{2,2} \leq \Big\|\frac{1}{n}\sum_{1\leq i\leq n}\Psi(Z_i, P)\Big\|_{2,2} + O_P\Big(\frac{a_n}{n}\Big)$$

uniformly in $P \in \mathbf{P}$. Moreover, Assumption 4.1(b) and Theorem 1.6 in Tropp (2012) yield for any $t \geq 0$,

$$\sup_{P\in\mathbf{P}} P\Big\{\Big\|\frac{1}{n}\sum_{1\leq i\leq n}\Psi(Z_i, P)\Big\|_{2,2} \geq t\Big\} \leq (p + d_0)\exp\left(\frac{-t^2/2}{K_{1,p}/n + K_{0,p}t/(3n)}\right) . \tag{74}$$

Next set $t = C((K_{0,p} \vee K_{1,p})\log(p + d_0)/n)^{1/2}$ for any $0 < C < \infty$ and note that $t = o(1)$ by hypothesis. Therefore, for n large enough we have that $K_{1,p}/n + K_{0,p}t/(3n) \leq 2(K_{1,p} \vee K_{0,p})/n$, which yields

$$\begin{aligned}\sup_{P\in\mathbf{P}} P\Big\{\Big\|\frac{1}{n}\sum_{1\leq i\leq n}\Psi(Z_i, P)\Big\|_{2,2} \geq C\frac{((K_{0,p} \vee K_{1,p})\log(p + d_0))^{1/2}}{\sqrt{n}}\Big\}& \\ \leq (p + d_0)\exp\left(-\frac{C^2}{4}\log(p + d_0)\right) = \exp\left(-\Big(\frac{C^2}{4} - 1\Big)\log(p + d_0)\right)& \end{aligned} \tag{75}$$

for n sufficiently large. Finally, note that $d_0 \leq p$ because of Assumption 4.2(b), which implies $\log(p + d_0) \leq 2\log(1 + p)$. Part (a) then follows from (74) and letting $C \to \infty$ in (75).

To establish part (b), we once again apply the triangle inequality and Assumption 4.1 to obtain

$$\max_{1\leq j\leq d_1+1}\|\hat{b}_{j,n} - b_j(P)\|_2 \leq \max_{1\leq j\leq d_1+1}\Big\|\frac{1}{n}\sum_{1\leq i\leq n}\varphi_j(Z_i, P)\Big\|_2 + O_P\Big(\frac{a_n}{n}\Big) .$$

Applying Lemma 2.2.2 in van der Vaart and Wellner (1996) with $\psi(x) = x^2$ we can then conclude

$$\begin{aligned}E_P\Big[\max_{1\leq j\leq d_1+1}\Big\|\frac{1}{n}\sum_{1\leq i\leq n}\varphi_j(Z_i, P)\Big\|_2^2\Big]^{1/2} &\leq \sqrt{d_1 + 1}\max_{1\leq j\leq d_1+1}E_P\Big[\Big\|\frac{1}{n}\sum_{1\leq i\leq n}\varphi_j(Z_i, P)\Big\|_2^2\Big]^{1/2} \\ &= \frac{\sqrt{d_1 + 1}}{\sqrt{n}}\max_{1\leq j\leq d_1+1}E_P[\varphi_j(Z_i, P)'\varphi_j(Z_i, P)]^{1/2} .\end{aligned} \tag{76}$$

Therefore, Markov's inequality, result (76), and Assumption 4.1(b), imply the bound

$$P\Big\{\max_{1\leq j\leq d_1+1}\Big\|\frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\varphi_j(Z_i, P)\Big\|_2 \geq t\Big\} \leq \frac{(d_1 + 1)K_{1,p}}{t^2} .$$

The first claim of part (b) then follows by setting $t = C\sqrt{(d_1 + 1)K_{1,p}}$ for some large C large. The second claim of part (b) follows by similar arguments. ■

Lemma C.13. *Define* $D(A_0(P))[H] = -\big(M_0(P)HA_0^\dagger(P) + (A_0^\dagger(P))'H'M_0(P)\big)$ *and for* $1 \le j \le d_1+1$ *set*

$$\begin{aligned}\hat{\varepsilon}_{j,n}(P;y) := \sqrt{n}(\hat{b}_{j,n}'\hat{M}_{0,n}y - b_j(P)'M_0(P)y) - b_j(P)'D(A_0(P))[\sqrt{n}(\hat{A}_{0,n} - A_0(P))]y \\ - \sqrt{n}(\hat{b}_{j,n} - b_j(P))'M_0(P)y\ .\end{aligned}$$

Suppose that $\inf_{P\in\mathbf{P}} \underline{s}(A_0(P))^2 > 0$ *and* $2\|A_0(P)\|_{2,2}\|\hat{A}_{0,n} - A_0(P)\|_{2,2} + \|\hat{A}_{0,n} - A_0(P)\|_{2,2}^2 < \underline{s}(A_0(P))^2/2$ *with probability approaching one uniformly in* $P \in \mathbf{P}$*. Then, it follows that*

(a) With probability approaching one uniformly in $P \in \mathbf{P}$ *and* $1 \le j \le d_1+1$*, it follows*

$$\sup_{y\in\mathbb{R}^p:\|y\|_2\le 1} |\hat{\varepsilon}_{j,n}(P;y)| \lesssim \frac{\sqrt{n}\|\hat{A}_{0,n} - A_0(P)\|_{2,2}^2\|b_j(P)\|_2}{\underline{s}(A_0(P))^2} + \frac{\sqrt{n}\|\hat{A}_{0,n} - A_0(P)\|_{2,2}\|\hat{b}_{j,n} - b_j(P)\|_2}{\underline{s}(A_0(P))}\ .$$

(b) With probability approaching one uniformly in $P \in \mathbf{P}$*,*

$$\begin{aligned}\sup_{y\in\mathbb{R}^p:\|y\|_2\le 1} \max_{1\le j\le d_1+1} |\hat{\varepsilon}_{j,n}(P;y)| \lesssim \frac{\sqrt{n}\|\hat{A}_{0,n} - A_0(P)\|_{2,2}^2 \max_{1\le j\le d_1+1}\|b_j(P)\|_2}{\underline{s}(A_0(P))^2} \\ + \frac{\sqrt{n}\|\hat{A}_{0,n} - A_0(P)\|_{2,2} \max_{1\le j\le d_1+1}\|\hat{b}_{j,n} - b_j(P)\|_2}{\underline{s}(A_0(P))}\ .\end{aligned}$$

Proof. For notational convenience, in the proof we suppress the dependence of $A_0(P)$, $M_0(P)$, and $b_j(P)$ on P and the subscript n in $\hat{A}_{0,n}$ and $\hat{M}_{0,n}$. To show part (a), we apply the decomposition

$$\hat{b}_{j,n}'\hat{M}_0y - b_j'M_0y = b_j'(\hat{M}_0 - M_0)y + (\hat{b}_{j,n} - b_j)'M_0y + (\hat{b}_{j,n} - b_j)'(\hat{M}_0 - M_0)y\ . \tag{77}$$

Consider the first summand in (77). According to Lemma C.16, the first differential or directional derivative of a function $f : \mathbb{R}^{p\times d_0} \to \mathbb{R}$ defined as $f(A_0) = b_j'M_0y$ with respect to A_0 in the direction H is $b_j'D(A_0)[H]y$. By the Mean-Value Theorem (see, e.g., Theorem 5.10 in Magnus and Neudecker (2019)), $f(\hat{A}_0) - f(A_0) = b_j'D(\tilde{A}_0)[\hat{A}_0 - A_0]y$, where $\tilde{A}_0 = t\hat{A}_0 + (1-t)A_0$ for some $t \in [0,1]$, so we can write

$$b_j'(\hat{M}_0 - M_0)y = b_j'D(A_0)[\hat{A}_0 - A_0]y + \underbrace{b_j'(D(\tilde{A}_0)[\hat{A}_0 - A_0] - D(A_0)[\hat{A}_0 - A_0])y}_{(R)}. \tag{78}$$

To bound the term (R) in result (78) uniformly in $P \in \mathbf{P}$ we decompose it into the terms

$$\begin{aligned}(R) = -b_j'\{\tilde{M}_0(\hat{A}_0 - A_0)\tilde{A}_0^\dagger - M_0(\hat{A}_0 - A_0)A_0^\dagger\}y \qquad (R.1)\\ - b_j'\{(\tilde{A}_0^\dagger)'(\hat{A}_0 - A_0)'\tilde{M}_0 - (A_0^\dagger)'(\hat{A}_0 - A_0)'M_0\}y\ . \qquad (R.2)\end{aligned}$$

First examining the terms $(R.1)$, we expand it further to obtain the decomposition:

$$\begin{aligned}(R.1) = -b_j'(\tilde{M}_0 - M_0)(\hat{A}_0 - A_0)A_0^\dagger y \qquad (R.1.A)\\ - b_j'M_0(\hat{A}_0 - A_0)(\tilde{A}_0^\dagger - A_0^\dagger)y \qquad (R.1.B)\\ + b_j'(\tilde{M}_0 - M_0)(\hat{A}_0 - A_0)(\tilde{A}_0^\dagger - A_0^\dagger)y \qquad (R.1.C)\ .\end{aligned}$$

Next note that the Cauchy-Schwarz inequality implies that $|a'b| \leq \|a\|_2\|b\|_2$ for any a and b. We therefore obtain that

$$\begin{aligned} |(R.1.A)| &\leq \|(A_0^\dagger)'(\hat{A}_0 - A_0)'(\tilde{M}_0 - M_0)b_j\|_2 \cdot \|y\|_2 \\ &\leq \|(A_0^\dagger)'\|_{2,2} \cdot \|(\hat{A}_0 - A_0)'\|_{2,2} \cdot \|\tilde{M}_0 - M_0\|_{2,2} \cdot \|b_j\|_2 \cdot \|y\|_2 \ . \end{aligned} \tag{79}$$

Similarly, we have

$$\begin{aligned} |(R.1.B)| &\leq \|(\tilde{A}_0^\dagger - A_0^\dagger)'(\hat{A}_0 - A_0)'M_0 b_j\|_2 \cdot \|y\|_2 \\ &\leq \|(\tilde{A}_0^\dagger - A_0^\dagger)'\|_{2,2} \cdot \|(\hat{A}_0 - A_0)'\|_{2,2} \cdot \|M_0\|_{2,2} \cdot \|b_j\|_2 \cdot \|y\|_2 \ . \end{aligned} \tag{80}$$

and

$$\begin{aligned} |(R.1.C)| &\leq \|(\tilde{A}_0^\dagger - A_0^\dagger)'(\hat{A}_0 - A_0)'(\tilde{M}_0 - M_0)b_j\|_\infty \cdot \|y\|_1 \\ &\leq \|(\tilde{A}_0^\dagger - A_0^\dagger)'\|_{2,2} \cdot \|(\hat{A}_0 - A_0)'\|_{2,2} \cdot \|\tilde{M}_0 - M_0\|_{2,2} \cdot \|b_j\|_2 \cdot \|y\|_1 \ . \end{aligned} \tag{81}$$

Proceeding to bound the term $(R.2)$ in our expansion for (R), we first decompose $(R.2)$ into the three terms

$$\begin{aligned} (R.2) = &-b_j'(\tilde{A}_0^\dagger - A_0^\dagger)'(\hat{A}_0 - A_0)'M_0 y && (R.2.A) \\ &- b_j'(\tilde{A}_0^\dagger)'(\hat{A}_0 - A_0)'(\tilde{M}_0 - M_0)y && (R.2.B) \\ &+ b_j'(\tilde{A}_0^\dagger - A_0^\dagger)'(\hat{A}_0 - A_0)'(\tilde{M}_0 - M_0)y && (R.2.C) \ . \end{aligned}$$

Using similar arguments to those employed in deriving (79), (80), and (81), we then obtain the bounds

$$\begin{aligned} |(R.2.A)| &\leq \|M_0(\hat{A}_0 - A_0)(\tilde{A}_0^\dagger - A_0^\dagger)b_j\|_2 \cdot \|y\|_2 \\ &\leq \|M_0\|_{2,2} \cdot \|\hat{A}_0 - A_0\|_{2,2} \cdot \|\tilde{A}_0^\dagger - A_0^\dagger\|_{2,2} \cdot \|b_j\|_2 \cdot \|y\|_2 \end{aligned} \tag{82}$$

and

$$\begin{aligned} |(R.2.B)| &\leq \|(\tilde{M}_0 - M_0)(\hat{A}_0 - A_0)A_0^\dagger b_j\|_2 \|y\|_2 \\ &\leq \|\tilde{M}_0 - M_0\|_{2,2} \cdot \|\hat{A}_0 - A_0\|_{2,2} \cdot \|A_0^\dagger\|_{2,2} \cdot \|b_j\|_2 \cdot \|y\|_2 \ . \end{aligned} \tag{83}$$

and

$$\begin{aligned} |(R.2.C)| &\leq \|(\tilde{M}_0 - M_0)(\hat{A}_0 - A_0)A_0^\dagger b_j\|_2 \|y\|_2 \\ &\leq \|\tilde{A}_0^\dagger - A_0^\dagger\|_{2,2} \cdot \|\hat{A}_0 - A_0\|_{2,2} \cdot \|\tilde{M}_0 - M_0\|_{2,2} \cdot \|b_j\|_2 \cdot \|y\|_2 \ . \end{aligned} \tag{84}$$

Next note that by Weyl's perturbation inequality (see, e.g., Corollary III.2.6 in Bhatia (2013)) we have

$$\underline{s}(\tilde{A}_0)^2 \geq \underline{s}(A_0)^2 - \|\tilde{A}_0'\tilde{A}_0 - A_0'A_0\|_{2,2} \geq \underline{s}(A_0)^2 - 2\|A_0\|_{2,2}\|\hat{A}_0 - A_0\|_{2,2} - \|\hat{A}_0 - A_0\|_{2,2}^2 \ , \tag{85}$$

where in the final inequality we used that $\|\tilde{A}_0 - A_0\|_{2,2} \leq \|\hat{A}_0 - A_0\|_{2,2}$ due to $\tilde{A}_0$ being a convex combination of $\hat{A}_0$ and A_0. Since, by hypothesis, $2\|A_0(P)\|_{2,2}\|\hat{A}_0 - A_0\|_{2,2} + \|\hat{A}_0 - A_0\|_{2,2}^2 < \underline{s}(A_0)^2/2$ with probability tending to one (uniformly in $P \in \mathbf{P}$) we obtain from (85) that with probability tending to one

$$\underline{s}(\tilde{A}_0)^2 \geq \frac{\underline{s}(A_0)^2}{2} > 0 \tag{86}$$

(uniformly in $P \in \mathbf{P}$). In particular, result (86) implies that $\text{rank}\{\tilde{A}_0\} = \text{rank}\{A_0\} = d_0$ with probability

tending to one, which together with Theorem 4.1 in Wedin (1973) gives us

$$\|\tilde{A}_0^\dagger - A_0^\dagger\|_{2,2} \lesssim \|\tilde{A}_0^\dagger\|_{2,2} \cdot \|A_0^\dagger\|_{2,2} \cdot \|\tilde{A}_0 - A_0\|_{2,2} = \frac{1}{\underline{s}(\tilde{A}_0)} \frac{1}{\underline{s}(A_0)} \|\hat{A}_0 - A_0\|_{2,2} \lesssim \frac{1}{\underline{s}(A_0)^2} \|\hat{A}_0 - A_0\|_{2,2} \ , \quad (87)$$

where the final inequality holds with probability tending to one uniformly in $P \in \mathbf{P}$ by (86). Furthermore, Theorem 2.5 in Chen et al. (2016) and $\|\tilde{A}_0 - A_0\|_{2,2} \leq \|\hat{A}_0 - A_0\|_{2,2}$ imply that

$$\|\tilde{M}_0 - M_0\|_{2,2} \leq \min\{\|\tilde{A}_0^\dagger\|_{2,2}, \|A_0^\dagger\|_{2,2}\} \cdot \|\hat{A}_0 - A_0\|_{2,2} \lesssim \frac{1}{\underline{s}(A_0)} \|\hat{A}_0 - A_0\|_{2,2}, \quad (88)$$

where the final inequality holds with probability tending to one uniformly in $P \in \mathbf{P}_0$ by Lemma C.20 and result (86). Further note that $\|S\|_{2,2} = \|S'\|_{2,2}$ for any matrix S by Theorem 6.5.1 in Luenberger (1969). Since $\|M_0\|_{2,2} \leq 1$ due to M_0 being a projection matrix, and $\|\hat{A}_0 - A_0\|_{2,2} \leq \underline{s}(A_0)/\sqrt{2}$ with probability tending to one uniformly in $P \in \mathbf{P}$ by hypothesis, we can then combine the bounds in results (79), (80), (81), (82), (83), and (84) with results (87) and (88) to conclude that

$$|(R)| \lesssim \frac{\|\hat{A}_0 - A_0\|_{2,2}^2 \|b_j\|_2}{\underline{s}(A_0)^2} \quad (89)$$

with probability approaching one uniformly in $P \in \mathbf{P}$ and $1 \leq j \leq d_1 + 1$. Moreover, by similar arguments

$$\begin{aligned} |(\hat{b}_{j,n} - b_j)'(\hat{M}_0 - M_0)y| &\leq \|(\hat{M}_0 - M_0)(\hat{b}_{j,n} - b_j)\|_\infty \|y\|_1 \\ &\leq \|\hat{M}_0 - M_0\|_{2,2} \|\hat{b}_{j,n} - b_j\|_2 \lesssim \frac{\|\hat{A}_0 - A_0\|_{2,2} \|\hat{b}_{j,n} - b_j\|_2}{\underline{s}(A_0)}, \end{aligned} \quad (90)$$

where the final inequality holds uniformly in $P \in \mathbf{P}$ by the same arguments used in (88). Part (a) of the lemma therefore follows from the decomposition in equations (77) and (78), and the bounds in (89) and (90). Part (b) can be proved similarly because the with probability approaching one statement above only depends on $\hat{A}_{0,n}$ and $A_0(P)$ and not $1 \leq j \leq d_1 + 1$. ■

Lemma C.14. *Suppose Z_i, $1 \leq i \leq n$ is an i.i.d. sequence, $\mathbb{Z} \sim N(0,1)$, Assumption 4.3 holds, and define $S_{j,n}^*(P) := n^{-1/2} \sum_{1 \leq i \leq n} \xi_j(Z_i, P)$. Then, it follows that for some universal $C < \infty$ and any $\delta > 0$, $x \in \mathbb{R}$*

$$\sup_{P \in \mathbf{P}} \sup_{y \in \mathbb{R}^p : \|y\|_1 \leq 1} \max_{1 \leq j \leq d_1 + 1} \left\{ P\left\{S_{j,n}^*(P)'y \leq x\right\} - P\left(\sigma_j(P;y)\mathbb{Z} \leq x + 3\delta\right) \right\} \leq \left(\frac{K_\xi}{\delta} \vee 1\right)^3 \frac{C \log(n)}{\sqrt{n}}$$

$$\sup_{P \in \mathbf{P}} \sup_{y \in \mathbb{R}^p : \|y\|_1 \leq 1} \max_{1 \leq j \leq d_1 + 1} \left\{ P\left(\sigma_j(P;y)\mathbb{Z} \leq x - 3\delta\right) - P\left\{S_{j,n}^*(P)'y \leq x\right\} \right\} \leq \left(\frac{K_\xi}{\delta} \vee 1\right)^3 \frac{C \log(n)}{\sqrt{n}}.$$

Proof. First note that for any $y \in \mathbb{R}^p$, $\eta_{j,i}(P;y) := \xi_j(Z_i, P)'y$ is an i.i.d. sequence satisfying $E_P[\eta_{j,i}(P;y)] = 0$ and $\mathrm{Var}_P[\eta_{j,i}(P;y)] = \sigma_j^2(P;y)$ by Lemma C.10. Further note that Assumption 4.3 implies that

$$\sup_{P \in \mathbf{P}} \sup_{y \in \mathbb{R}^p : \|y\|_1 \leq 1} \max_{1 \leq j \leq d_1 + 1} \sum_{i=1}^n E_P\left[\left|\frac{\eta_{j,i}(P;y)}{\sqrt{n}}\right|^3\right] \leq \frac{K_\xi^3}{\sqrt{n}}. \quad (91)$$

Letting $g_{j,i}(P;y) \sim N(0, \sigma_j^2(P;y)/n)$ and $\mathbb{Z} \sim N(0,1)$, next use that $E[|\mathbb{Z}|^3] \leq 2$ to conclude that

$$\sup_{P\in\mathbf{P}} \sup_{y\in\mathbb{R}^p:\|y\|_1\leq 1} \max_{1\leq j\leq d_1+1} \sum_{i=1}^n E[|g_{j,i}(P;y)|^3] \leq \sup_{P\in\mathbf{P}} \sup_{y\in\mathbb{R}^p:\|y\|_1\leq 1} \max_{1\leq j\leq d_1+1} \frac{2\sigma_j^3(P;y)}{\sqrt{n}} \leq \frac{2K_\xi^3}{\sqrt{n}}, \tag{92}$$

where in the final inequality we used that $\sigma_j(P;y) = (E_P[\eta_{j,i}^2(P;y)])^{1/2} \leq (E_P[|\eta_{j,i}(P;y)|^3])^{1/3}$ and Assumption 4.3. Combining results (91) and (92) and applying Lemma 39 in Belloni et al. (2019), we then obtain that for each $\delta > 0$ there exists a random variable $T_j(P;y) \sim N(0, \sigma_j^2(P;y))$ satisfying

$$\begin{aligned}\sup_{P\in\mathbf{P}} \sup_{y\in\mathbb{R}^p:\|y\|_1\leq 1} \max_{1\leq j\leq d_1+1} P\Big\{\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_{j,i}(P;y) - T_j(P;y)\Big| > 3\delta\Big\} \\ \leq \min_{t\geq 0}\Big(2P(|\mathbb{Z}|>t) + \frac{3K_\xi^3}{\delta^3\sqrt{n}}t^2\Big) \leq \Big(\frac{K_\xi}{\delta}\vee 1\Big)^3 \frac{C\log(n)}{\sqrt{n}},\end{aligned} \tag{93}$$

where the final inequality holds for some $C < \infty$ by setting $t = \sqrt{\log(n)}$ and using the bound $P\{|\mathbb{Z}| > t\} \leq 2\exp\{-t^2/2\}$. Since $S_{j,n}^*(P)'y = \sum_{1\leq i\leq n} \eta_{j,i}(P;y)/\sqrt{n}$, we obtain from result (93) and Strassen's theorem (see, e.g., Theorem 10.3.8 in Pollard (2002)) that for any $\delta > 0$ and any $x \in \mathbb{R}$ we have that

$$\begin{aligned}\sup_{P\in\mathbf{P}} \sup_{y\in\mathbb{R}^p:\|y\|_1\leq 1} \max_{1\leq j\leq d_1+1} \Big\{P\left\{S_{j,n}^*(P)'y \leq x\right\} - P\left\{T_j(P;y) \leq x+3\delta\right\}\Big\} \leq \Big(\frac{K_\xi}{\delta}\vee 1\Big)^3 \frac{C\log(n)}{\sqrt{n}} \\ \sup_{P\in\mathbf{P}} \sup_{y\in\mathbb{R}^p:\|y\|_1\leq 1} \max_{1\leq j\leq d_1+1} \Big\{P\{T_j(P;y) \leq x-3\delta\} - P\left\{S_{j,n}^*(P)'y \leq x\right\}\Big\} \leq \Big(\frac{K_\xi}{\delta}\vee 1\Big)^3 \frac{C\log(n)}{\sqrt{n}}.\end{aligned}$$

The claim of the lemma then follows from $T_j(P;y) \sim N(0, \sigma_j^2(P;y))$. ■

Lemma C.15. *Let Assumptions 4.1, 4.2, 4.3, and 4.6 hold.*

(a) If $K_{1,p}(K_{0,p} \vee K_{1,p})\log(1+p)d_1^{1/3} = O(np^{2/3})$, then it follows that uniformly in $P \in \mathbf{P}$:

$$\max_{1\leq j\leq d_1+1} \sup_{\|y\|_1\leq 1} |\sqrt{n}(\hat{b}_{j,n}'\hat{M}_{0,n}y - b_j(P)'M_0(P)y)| = O_P((pd_1)^{1/3}) .$$

(b) Suppose in addition that for each $p \geq 1$ there are finite constants $C_{\xi,p}$ and $C_{\varphi,p}$ satisfying the inequalities

$$\begin{aligned}\sup_{P\in\mathbf{P}} E_P\Big[\max_{1\leq j\leq d_1+1} \|\xi_j(Z,P)\|_\infty^2\Big] \leq C_{\xi,p}^2 \\ \sup_{P\in\mathbf{P}} E_P\Big[\max_{1\leq j\leq d_1+1} \|\varphi_j(Z,P)\|_2^2\Big] \leq C_{\varphi,p}^2.\end{aligned}$$

If $pC_{\varphi,p}^2\log^2(p+d_1)(K_{0,p}\vee K_{1,p}) = o(n)$, then it follows that uniformly in $P \in \mathbf{P}$ we have:

$$\max_{1\leq j\leq d_1+1} \sup_{\|y\|_1\leq 1} |\sqrt{n}(\hat{b}_{j,n}'\hat{M}_{0,n}y - b_j(P)'M_0(P)y)| = O_P(C_{\xi,p}\sqrt{\log(p+d_1)}) .$$

Proof. We begin with some preliminary steps that apply to both claims of the lemma. For any matrices $A, H \in \mathbb{R}^{p\times d_0}$ with $\text{rank}(A) = d_0$ define $M(A) := \mathbf{I}_p - A(A'A)^{-1}A'$ where $\mathbf{I}_p$ denotes the $p \times p$ identity

matrix, and set $D(A)[H] := -M(A)H(A'A)^{-1}A' - A(A'A)^{-1}H'M(A)$ and the two terms

$$R_{1,j,n}(y) := \sqrt{n}(\hat{b}_{j,n} - b_j(P) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_j(Z_i,P))'M_0(P)y$$
$$R_{2,j,n}(y) := b_j(P)'D(A_0(P))[\sqrt{n}(\hat{A}_{0,n} - A_0(P)) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Z_i,P)]y\ . \tag{94}$$

Next note that Lemma C.12(a), Assumptions 4.2(a)(b), and Assumption 4.6 imply that $\|A_0(P)\|_{2,2}\|\hat{A}_{0,n} - A_0(P)\|_{2,2} + \|\hat{A}_{0,n} - A_0(P)\|_{2,2}^2 < \underline{s}(A_0(P))^2/2$ with probability tending to one uniformly in $P \in \mathbf{P}$ (see also the the arguments in (54) and (55) for additional details). We can therefore apply Lemma C.13 and use Assumptions 4.2(b)(c) and $\|y\|_2 \leq \|y\|_1$ to obtain that uniformly in $P \in \mathbf{P}$ that

$$\begin{aligned}
&\max_{1\leq j\leq d_1+1}\sup_{\|y\|_1\leq 1}\left|\sqrt{n}(\hat{b}_{j,n}'\hat{M}_{0,n}y - b_j(P)'M_0(P)y)\right| \\
&\quad\lesssim \max_{1\leq j\leq d_1+1}\sup_{\|y\|_1\leq 1}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}y'\xi_j(Z_i,P)\right| + \sup_{\|y\|_2\leq 1}\max_{1\leq j\leq d_1+1}(|R_{1,j,n}(y)| + |R_{2,j,n}(y)|) \\
&\qquad + K_{2,p}\sqrt{n}\|\hat{A}_{0,n} - A_0(P)\|_{2,2}^2 + \sqrt{n}\|\hat{A}_{0,n} - A_0(P)\|_{2,2}\max_{1\leq j\leq d_1+1}\|\hat{b}_{j,n} - b_j(P)\|_2\ .
\end{aligned} \tag{95}$$

Since $\|M_0(P)y\|_2 \leq 1$ for any $\|y\|_2 \leq 1$ due to $M_0(P)$ being a projection matrix, Assumption 4.1(a) implies

$$\max_{1\leq j\leq d_1+1}\sup_{\|y\|_2\leq 1}|R_{1,j,n}(y)| = O_P(a_n/\sqrt{n})\ . \tag{96}$$

Similarly, again using that $M_0(P)$ is a projection matrix, and that $A^\dagger(P) = (A_0(P)'A_0(P))^{-1}A_0(P)'$ (see, e.g., Proposition 6.12.1 in Luenberger (1969)) we obtain uniformly in $P \in \mathbf{P}$ that

$$\begin{aligned}
&\sup_{\|y\|_2\leq 1}\max_{1\leq j\leq d_1+1}|R_{2,j,n}(y)| \\
&\leq 2\max_{1\leq j\leq d_1+1}\|b_j(P)\|_2\|\sqrt{n}(\hat{A}_{0,n} - A_0(P)) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Z_i,P)\|_{2,2}\|A_0^\dagger(P)\|_{2,2} = O_P(K_{2,p}a_n/\sqrt{n})\ ,
\end{aligned} \tag{97}$$

where in the final result we employed Assumptions 4.1(a), 4.2(b)(c), and Lemma C.20. Therefore, combining results (95), (96), and (97) together with Lemma C.12, and Assumptions 4.1(a) and 4.6 yields

$$\begin{aligned}
&\max_{1\leq j\leq d_1+1}\sup_{\|y\|_1\leq 1}|\sqrt{n}(\hat{b}_{j,n}'\hat{M}_{0,n}y - b_j(P)'M_0(P)y)| \\
&\lesssim \max_{1\leq j\leq d_1+1}\sup_{\|y\|_1\leq 1}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}y'\xi_j(Z_i,P)\right| + \max_{1\leq j\leq d_1+1}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_j(Z_i,P)\right\|_2\|\hat{A}_{0,n} - A_0(P)\|_{2,2} + o_P(1)
\end{aligned} \tag{98}$$

uniformly in $P \in \mathbf{P}$.

The two parts of the lemma follow from the bound in (98) and employing the different assumptions to control the terms in the right hand side of (98). To establish part (a) let $\xi_{j,k}(Z,P)$ denote the k^{th} coordinate

of $\xi_j(Z,P) \in \mathbb{R}^p$ and use that $\sup_{\|y\|_1 \leq 1} |y'b| = \|b\|_\infty$ for any $b \in \mathbb{R}^p$ to obtain the upper bound

$$\begin{aligned} E_P\left[\max_{1\leq j\leq d_1+1}\sup_{\|y\|_1\leq 1}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n y'\xi_j(Z_i,P)\right|\right] &= E_P\left[\max_{1\leq j\leq d_1+1}\max_{1\leq k\leq p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{j,k}(Z_i,P)\right|\right] \\ &\lesssim (pd_1)^{1/3}\max_{1\leq j\leq d_1+1}\max_{1\leq k\leq p} E_P\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{j,k}(Z_i,P)\right|^3\right]^{1/3} \\ &\lesssim (pd_1)^{1/3}\max_{1\leq j\leq d_1+1}\sup_{\|y\|_1\leq 1} E_P[|\xi_j(Z,P)'y|^3]^{1/3}\ , \end{aligned}$$

where the first inequality follows from applying Lemma 2.2.2 in van der Vaart and Wellner (1996) with $\psi(x) = x^3$ and the second from Lemma C.17. Hence, Assumption 4.3 and Markov's inequality yield

$$\max_{1\leq j\leq d_1+1}\sup_{\|y\|_1\leq 1}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n y'\xi_j(Z_i,P)\right| = O_P((pd_1)^{1/3})\ . \tag{99}$$

Moreover, Lemma C.12 together with Assumptions 4.1(a) and Assumption 4.6 yield uniformly in $P \in \mathbf{P}$

$$\max_{1\leq j\leq d_1+1}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_j(Z_i,P)\right\|_2 \|\hat{A}_{0,n} - A_0(P)\|_{2,2} = O_P\left(1 \vee \frac{\sqrt{K_{1,p}(d_1+1)(K_{0,p}\vee K_{1,p})\log(1+p)}}{\sqrt{n}}\right). \tag{100}$$

Part (a) of the lemma therefore follows from results (98), (99), (100), and the rate condition $K_{1,p}(K_{0,p} \vee K_{1,p})\log(1+p)d_1^{1/3} = O(np^{2/3})$.

To establish part (b) define $\mathcal{F}_P := \{f : f(Z) = \xi_{j,k}(Z,P)$ for some $1 \leq j \leq d_1+1$ and $1 \leq k \leq p\}$ and note that $\mathcal{F}_P$ has envelope $F_P(Z) := \max_{1\leq j\leq d_1+1}\|\xi_j(Z,P)\|_\infty$. Moreover, since $|\mathcal{F}_P| \leq p(1+d_1)$, we can apply Theorem 2.14.1 in van der Vaart and Wellner (1996) to conclude that uniformly in $P \in \mathbf{P}$ we have

$$\begin{aligned} E_P\left[\max_{1\leq j\leq d_1+1}\sup_{\|y\|_1\leq 1}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n y'\xi_j(Z_i,P)\right|\right] &= E_P\left[\max_{1\leq j\leq d_1+1}\max_{1\leq k\leq p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{j,k}(Z_i,P)\right|\right] \\ &\lesssim \sqrt{\log(pd_1)}\cdot(E_P[F_P^2(Z)])^{1/2} \leq \sqrt{\log(pd_1)}C_{\xi,p}\ , \end{aligned} \tag{101}$$

where the final inequality follows by definition of F_P and $C_{\xi,p}$. Similarly, denote the $\|\cdot\|_2$-unit ball in $\mathbb{R}^p$ by $\mathcal{B}_p := \{u \in \mathbb{R}^p : \|u\|_2 \leq 1\}$ and define the class $\mathcal{G}_P := \bigcup_{j=1}^{d_1+1}\mathcal{G}_{j,P}$ where for each $1 \leq j \leq d_1+1$ we set $\mathcal{G}_{j,P} := \{f : f(Z) = u'\varphi_j(Z,P)$ for some $u \in \mathcal{B}_p\}$. Further note that, by the Cauchy-Schwarz inequality, $\mathcal{G}_P$ has envelope $G_P(Z) := \max_{1\leq j\leq d_1+1}\|\varphi_j(Z,P)\|_2$. Setting $\|g\|_{P,2} := (E_P[g^2(Z)])^{1/2}$ for any $g \in \mathcal{G}_P$ and letting $N_{[]}(\varepsilon, \mathcal{G}_P, \|\cdot\|_{P,2})$ denote the bracketing numbers of $\mathcal{G}_P$ under the norm $\|\cdot\|_{P,2}$, we then obtain from $\mathcal{G}_P := \bigcup_{j=1}^{d_1+1}\mathcal{G}_{j,P}$ and Theorem 2.7.11 in van der Vaart and Wellner (1996) that for any $\varepsilon \leq 1$ we have

$$\begin{aligned} N_{[]}(\varepsilon\|G_P\|_{P,2}, \mathcal{G}_P, \|\cdot\|_{P,2}) &\leq (d_1+1)\max_{1\leq j\leq d_1+1} N_{[]}(\varepsilon\|G\|_{P,2}, \mathcal{G}_{j,P}, \|\cdot\|_{P,2}) \\ &\leq (d_1+1)N(\varepsilon/2, \mathcal{B}_p, \|\cdot\|_2) \leq (d_1+1)\left(\frac{4}{\varepsilon}+1\right)^p \end{aligned} \tag{102}$$

where the final inequality follows from Lemma 14.27 in Bühlmann and Van De Geer (2011). Therefore, using

that $\sup_{\|u\|_2\leq 1}|u'b| = \|b\|_2$ for any $b\in\mathbb{R}^p$ and applying Theorem 2.14.2 in van der Vaart and Wellner (1996) we can conclude from result (102) and the definition of $C_{\varphi,p}$ and G_P that

$$E_P\left[\max_{1\leq j\leq d_1+1}\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varphi_j(Z_i,P)\Big\|_2\right] = E_P\left[\max_{1\leq j\leq d_1+1}\sup_{\|u\|_2\leq 1}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n u'\varphi_j(Z_i,P)\Big|\right]$$
$$\lesssim \int_0^1\sqrt{1+\log N_{[]}(\varepsilon\|G_P\|_{P,2},\mathcal{G}_P,\|\cdot\|_{P,2})}d\varepsilon\cdot\|G_P\|_{P,2}\lesssim\sqrt{p\log(1+d_1)}C_{\varphi,p}. \quad (103)$$

The second part of the lemma then follows from results (98), (101), (103), Markov's inequality, Lemma C.12 and the rate condition $C_{\varphi,p}^2 p\log^2(p+d_1)(K_{0,p}\vee K_{1,p}) = o(n)$. ■

Lemma C.16. *For any $A\in\mathbb{R}^{p\times d_0}$ with $\operatorname{rank}(A)=d_0$ define the function $A\mapsto M(A) := \mathbf{I}_p - A(A'A)^{-1}A'$. Then, $M:\mathbb{R}^{p\times d_0}\to\mathbb{R}^{p\times d_0}$ is differentiable at A and its derivative in the direction $H\in\mathbb{R}^{p\times d_0}$ equals*

$$D(A)[H] = -M(A)H(A'A)^{-1}A' - A(A'A)^{-1}H'M(A)\ .$$

Therefore, $A\mapsto b'M(A)y$ is differentiable at any A with $\operatorname{rank}(A)=d_0$ and its derivative is $H\mapsto b'D(A)[H]y$.

Proof. For any invertible $d_0\times d_0$ matrix S, Theorem 8.3 in Magnus and Neudecker (2019) implies that the derivative of the inverse map $S\mapsto S^{-1}$ is given by $H\mapsto -S^{-1}HS^{-1}$. Moreover, by the chain rule, the derivative of $A\mapsto A'A$ is given by $H\mapsto A'H+H'A$. Therefore, once again applying the chain rule we obtain that derivative of $A\mapsto M(A)$ at any A with $\operatorname{rank}(A)=d_0$ is given by

$$\begin{aligned}D(A)[H] &= -H(A'A)^{-1}A' - A(A'A)^{-1}H' + A(A'A)^{-1}(H'A+A'H)(A'A)^{-1}A'\\ &= -M(A)H(A'A)^{-1}A' - A(A'A)^{-1}H'M(A),\end{aligned}$$

as desired. The fact that the derivative of $b'M(A)y$ equals $H\mapsto b'D(A)[H]y$ follows from a second application of the chain rule. ■

Lemma C.17. *Let V_i, $1\leq i\leq n$ be i.i.d. random variables with $E[V_i]=0$ and $E[|V_i|^3]<\infty$. Then,*

$$E\Big[\Big|\frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}V_i\Big|^3\Big]\leq 3456E[|V_i|^3]\ .$$

Proof. For $(a)_+ = \max\{a,0\}$, it follows from Rosenthal's inequality (see, e.g., Theorem 15.11 in Boucheron et al. (2013)), the random variables V_i, $1\leq i\leq n$ being i.i.d., and the inequality $E[V_i^2]^{1/2}\leq E[|V_i|^3]^{1/3}$ that

$$E\Big[\Big(\frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}V_i\Big)_+^3\Big]^{1/3}\leq 6E[V_i^2]^{1/2}+6\Big(\frac{n}{n^{3/2}}E[|V_i|^3]\Big)^{1/3}\leq 6(1+n^{-1/6})E[|V_i|^3]^{1/3}\leq 12E[|V_i|^3]^{1/3}\ . \quad (104)$$

Next, note that we can apply the same arguments to $-V_i$ in place of V_i to obtain the upper bound

$$E\Big[\Big(\frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}(-V_i)\Big)_+^3\Big]^{1/3}\leq 12E[|V_i|^3]^{1/3}\ . \quad (105)$$

For $(a)_- = \min\{a, 0\}$, then note that for any random variable X we have $E[|X|^3] = E[X_+^3 - X_-^3] = E[X_+^3 + (-X)_+^3]$. The claim of the lemma therefore follows form combining (104) and (105). ■

Lemma C.18. *If $c_1 \geq \underline{\sigma}^2 > 0$ and $c_2 \geq \underline{\sigma}^2$, then*

$$|c_1^{-1/2} - c_2^{-1/2}| \leq \frac{1}{2\underline{\sigma}^3}|c_1 - c_2|.$$

PROOF. Consider the function $f(x) = x^{-1/2}$. For $x \geq \underline{\sigma}^2$, we have $|f'(x)| = \frac{1}{2}x^{-3/2} \leq \frac{1}{2}(\underline{\sigma}^2)^{-3/2} = 1/(2\underline{\sigma}^3)$. The conclusion now follows from the mean-value theorem. ■

Lemma C.19. *Let $X \in \mathbb{R}^p$ be random vector and suppose $E[XX'] < \infty$. Then, $\|E[XX']\|_{2,2} \leq E[X'X]$.*

PROOF. Because $E[XX']$ is symmetric and positive semi-definite, all its eigenvalues are non-negative and $\|E[XX']\|_{2,2}$ equals the largest eigenvalue of $E[XX']$. Using that $\text{trace}\{E[XX']\}$ equals the sum of the eigenvalues of $E[XX']$, we can then conclude that $\|E[XX']\|_{2,2} \leq \text{trace}\{E[XX']\} = E[X'X]$. ■

Lemma C.20. *Let A_0 be an arbitrary $p \times d_0$ matrix of rank d_0. Then, $\|A_0^\dagger\|_{2,2} = \|(A_0^\dagger)'\|_{2,2} = 1/\underline{s}(A_0)$.*

PROOF. The claim that $\|A_0^\dagger\|_{2,2} = \|(A_0^\dagger)'\|_{2,2}$ follows from Theorem 6.5.1 in Luenberger (1969). Since A_0 has rank d_0 it follows from Proposition 6.12.1 in Luenberger (1969) that $A_0^\dagger = (A_0'A_0)^{-1}A_0'$. Hence, we have

$$\|(A_0^\dagger)'\|_{2,2} = \sup_{\|x\|_2 \leq 1} (x'(A_0'A_0)^{-1}A_0'A_0(A_0'A_0)^{-1}x)^{1/2} = \sup_{\|x\|_2 \leq 1} (x'(A_0'A_0)^{-1}x)^{1/2} = \frac{1}{\underline{s}(A_0)} , \tag{106}$$

where the first equality follows by definition of $\|\cdot\|_{2,2}$ and the final one from $\|(A_0'A_0)^{-1}\|_{2,2} = 1/\underline{s}(A_0)$. ■