Many Proxy Controls

Ben Deaner
(Department of Economics UCL)
bendeaner@gmail.com

September 18, 2022

Abstract

NOTE: THIS IS A WORKING DRAFT AND IS STILL BEING REVISED

A recent literature considers causal inference using two vectors of noisy proxies for unobserved confounding factors. In this paper we consider linear models in which the vectors of proxies are potentially high-dimensional and there may be many unobserved confounders. A key insight is that if each group of proxies is strictly larger than the number of confounding factors then a matrix of nuisance parameters satisfies a rank restriction. We can exploit the rank restriction to reduce the number of free parameters to be estimated. The number of unobserved confounders is not known a priori but we show that it is identified, and we apply penalization methods to adapt to this quantity. We develop doubly-robust estimation and inference methods. We examine the asymptotic properties of these techniques and provide simulation evidence that they are effective.

Introduction

The key challenge for causal inference is the presence of confounding factors: variables that cause both treatments and outcomes. In observational studies some important confounders may be absent from the available covariates or crudely mismeasured. For example, suppose we wish to assess the effects of some intervention on a student's educational attainment. The pupil's academic ability is a potential confounding factor. Even the best measurements of academic aptitude (test scores, grades, etc.) are likely subject to error. We refer to noisy and possibly biased measurements of the underlying confounding factors as 'proxy controls'. If some confounders are unmeasured or mismeasured then standard methods that adjust for observed covariates do not recover a causal effect.

The problem of mismeasured confounding can persist even if the data contains a rich set of high-dimensional covariates. However, the many covariates constitute an ample source of potential proxies for the unobserved confounders. As such, high-dimensional settings are particularly apt for methods that can achieve causal inference using noisy proxies.

A class of empirical methods aims to consistently estimate causal effects using two vectors of noisy proxies for the underlying confounders. These two vectors of proxies must satisfy certain exclusion restrictions and must be sufficiently informative about the confounders. We refer to this class of methods as the 'double proxy' approach. Double proxy methods were introduced into the economic literature in the context of linear models by Zvi Griliches and coauthors in the 1970s, beginning with Griliches & Mason (1972). More recently, Miao *et al.*

(2018b) initiated a literature on identification and estimation of causal effects in nonlinear and nonparametric models using a double proxy approach. These methods are designed for use with a small, fixed number of proxies and unobserved confounders.

In this work we consider identification, estimation, and inference in linear models when there are many proxy controls. That is, when each of the two vectors of proxies is high-dimensional. In addition, we allow for the possibility that there are many unobserved confounding factors. We design estimation and inference methods tailored to this case. In high-dimensional settings standard asymptotic approximations based on a fixed number of proxies and confounders may be unreliable. Thus our asymptotic analysis allows the number of proxies and confounders to grow with the sample size.

A key insight in this work is that if there are strictly fewer unobserved confounders than there are proxies in each of the two vectors, then two matrices of nuisance parameters have a low-rank structure. We exploit this low-rank structure to reduce the number of free parameters to be estimated. This allows for more efficient estimation, particularly when the number of proxies is large. The number of confounders is generally unknown, and so we propose model selection methods that adapt to this quantity. The model selection methods are based on techniques from the literature on reduced-rank regression, particularly Bunea *et al.* (2011).

Selecting the correct number of unobserved confounders is important even in low-dimensional settings. Estimators that fail to adapt to the low-rank structure in the nuisance parameters or apply some kind of regularization may not be asymptotically normal even in fixed dimensions, leading to invalid inference.

Our proposed procedure is an example of a Double Machine Learning 2 (DML2) estimator of the kind analyzed in section 3.2 of Chernozhukov *et al.* (2018). Chernozhukov *et al.* (2018) shows that DML2 estimators are root-n consistent, asymptotically unbiased, and asymptotically Gaussian under relatively weak conditions on the nuisance parameter estimates. Our estimator is based on a doubly-robust score function. The estimator and corresponding confidence intervals have a closed-form, which ensures they are easy to compute. We develop asymptotic theory for the estimator and an associated inference method, and we provide extensive simulation evidence of the efficacy of our methods.

Related literature

Early papers that apply the double proxy approach in linear models include Griliches & Mason (1972), Chamberlain & Griliches (1975), and Griliches (1977). These works generally consider a single scalar unobserved confounder. An exception is Chamberlain & Griliches (1977) which allows for two scalar confounders.

Miao *et al.* (2018b) nonparametrically identifies the average structural function (and thus average causal effects) using double proxies and introduces a statistical test using proxy controls. Further works including Deaner (2021) and Tchetgen *et al.* (2020) build upon these identification results, they establish identification of conditional average causal effects, develop alternative characterizations of objects of interest, and adapt the approach to settings with panel data.

Nonparametric estimation with proxy controls was considered in Deaner (2019), and later by Tchetgen *et al.* (2020), Singh (2020), Cui *et al.* (2020), Kallus *et al.* (2021) and others. Miao *et al.* (2018a) consider estimation in parametric models when a 'confounding bridge' function is identified. Existing work applies to low-dimensional settings in which the number of proxies and confounding factors is small and treated as fixed.

In the context of panel data Imbens *et al.* (2021) provide a double proxy method in which the number of proxies can exceed the number of confounders. Imbens *et al.* (2021) uses l_2 penalization in order to ensure consistency and asymptotic normality when there are more proxies than confounders.

The analysis of Miao *et al.* (2018b), Miao *et al.* (2018a), and the classical works by Griliches and coauthors either implicitly or explicitly assume that the number of unobserved confounders is equal to the number of proxies in each group (see Subsection 1.1 for discussion).

Notation

The notation $M^\dagger$ denotes the Moore-Penrose pseudo-inverse of the matrix M. If M is positive semidefinite then $M^{1/2}$ denotes the unique positive semidefinite matrix square root of M and if M is strictly positive definite then $M^{-1/2}$ is the unique positive semidefinite matrix square root of M^{-1}. For any matrix M we let $\sigma_k(M)$ denote the k-th largest singular value of M.

For a matrix M we let $M_{[a:b,c:d]}$ be the sub-matrix of M consisting of the entries in rows a to b and columns c to d. $M_{[a:b,:]}$ is the sub-matrix of M consisting of rows a to b and $M_{[:,c:d]}$ is the sub-matrix of columns c to d. $M_{[a:b,c]}$ is shorthand for $M_{[a:b,c:c]}$ and similarly $M_{[a,c:d]} = M_{[a:a,c:d]}$.

If b is a vector then $\|b\|$ is the Euclidean norm of b. If M is a matrix then $\|M\| = \sup_{b\in\mathbb{R}^d:\|b\|=1} \|Mb\|$. For sequences a_n and b_n the notation $a_n \precsim b_n$ means that there exists a constant C so that $a_n \leq Cb_n$ for all n. $a_n \prec b_n$ means that $a_n \precsim b_n$ but not $b_n \precsim a_n$.

For any random column vector H we define $\Sigma_H = E[HH']$, however if H is the concatenation of two vectors, for example $H = (D', X')'$, we simply write Σ_{DX}. We define variables with D and X partialled out as follows. For $H = W$, V, Z, Y, X we define:

$$\begin{aligned}
\gamma_{H,0} &= \Sigma_D^\dagger E[DH'] \\
\omega_{H,0} &= \Sigma_{XD}^\dagger E[(X', D')'H'] \\
\tilde{H}(\gamma) &= H - \gamma' D \\
\bar{H}(\omega) &= H - \omega'(X', D')'
\end{aligned}$$

For notational convenience we sometimes write $\tilde{H} = \tilde{H}(\gamma_{H,0})$ and $\bar{H} = \bar{H}(\omega_{H,0})$.

1 Model and Identification

Let Y be an outcome of interest and X a vector of treatments. Let W be a vector of unobserved confounding factors. We assume the researcher has access to two vectors of proxies V and Z for the unobserved confounders W. In addition the researcher may have access to a vector of additional covariates D which could also confound X and Y.

The assumptions on the proxies V differ from those on the proxies Z. Due to this asymmetry we follow the existing literature and refer to V as 'outcome-aligned' proxies and Z as 'treatment-aligned' proxies, we explain these terms later in this section. Table 1 lists the relevant variables.

We aim to identify and estimate β_0, the coefficient on X in the first regression equation below. In order to avoid including intercepts we assume throughout that D contains a constant.

Table 1: List of Variables

Variable	Dimension	Description
Y	1	Outcome of interest.
X	d_X	Vector of treatments.
W	d_W	Unobserved confounding factors.
V	d_V	Outcome-aligned proxies for W.
Z	d_Z	Treatment-aligned proxies for W.
D	d_D	Additional conditioning variables.

$$Y = \beta_0' X + A_0 W + L_0 D + \varepsilon, \qquad E[\varepsilon(X', W', D')] = 0 \tag{1}$$
$$V = B_0 W + R_0 D + \upsilon, \qquad E[\upsilon(W', D')] = 0 \tag{2}$$

The regression models above are without loss of generality in the sense that they must hold for some coefficients β_0, A_0, B_0, etc. If the regressors are not linearly dependent then (1) and (2) uniquely define these coefficients.[1]

We cannot estimate β_0 from (1) directly because W is not observed. However, we leverage Assumption 1.1 below in order to derive moment conditions that do not involve W and thus allow us to identify β_0.

Assumption 1.1 (Model and Exclusion restrictions). i. (1) and (2) hold. ii. $E[\varepsilon Z'] = 0$ and $E[\upsilon(X', Z')] = 0$.

Assumption 1.1.ii imposes that Z is uncorrelated with the residual in regression model (1), and both Z and X are uncorrelated with the residual in (2). Note that the assumption is asymmetric in the two vectors of proxies Z and V. Unlike Z, V can be correlated with the residual from (1). X must be exogenous with respect the model (2) for V, but no such restriction need apply for the relationship between X and Z.

Assumption 1.1 is stated in terms of regression residuals, and the assumption defines β_0 not as a causal effect but as a vector of regression coefficients. In order to better assess when Assumption 1.1 is credible and β_0 has a causal interpretation, it is helpful to consider fully specified structural (i.e., causal) models that imply Assumption 1.1 holds with β_0 an average causal effect.

On the right in Figure 1.1.a is a linear structural equations model with uncorrelated residuals and coefficients that represent average causal effects. This model implies that Assumption 1.1 holds with $\beta_0' = a_{YX}$, where a_{YX} is the matrix of coefficients on X in the structural equation for Y. The directed graph on the left encodes the exclusion restrictions in the model: there is an arrow from one variable to another if and only if the variable appears in the other's structural equation. Thus arrows represent the possible presence of a causal effect. That the residuals are uncorrelated indicates that there are no omitted joint causes.[2]. In Figure 1.1.a we omit any additional observables D for the purpose of legibility but one could allow D to cause all the other variables.

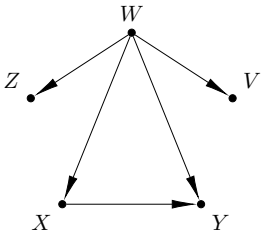
In Figure 1.1.a, W is a confounding factor: it appears in the equations for both Y and X. The proxies V and Z are excluded from all other equations which suggests they have no

[1] Our results continue to hold even if ε and υ are correlated with W, but we maintain this assumption because it provides a simple interpretation of β_0. We discuss this further in Appendix A.

[2] For further discussion of linear causal models see e.g., Pearl (2009)

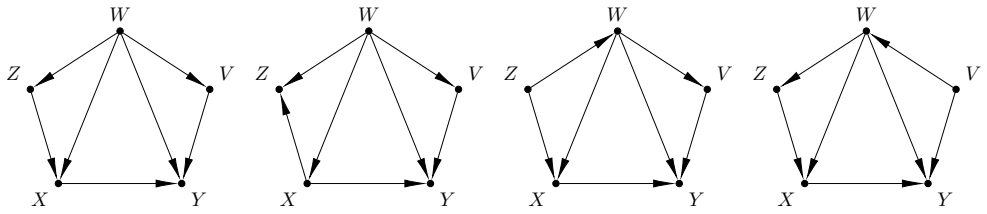
Figure 1.1: Causal Structure of Proxy Controls

(a) Simple Structural Model



- $Y{=}a_{YX}X + a_{YW}W + U_Y$
- $X{=}a_{XZ}Z + a_{XW}W + U_X$
- $V{=}a_{VW}W + U_V$
- $Z{=}a_{ZW}W + U_Z$
- $W{=}U_W$
- U_Y, U_X, U_V, U_Z, U_W all uncorrelated.

(b) Richer Structural Models



affect on treatments and outcomes. In addition, X and Y are excluded from the equations for V and Z.

The model in Figure 1.1.a is rather restrictive in that the proxies do cause treatments and outcomes. Suppose treatment is an educational intervention and the outcome is say, acceptance into university. In this case we can think of W as academic ability at the time of treatment and use test scores as proxies. The exclusion restrictions on Z and V in Figure 1.1.a are plausible in this setting if the tests are taken prior to treatment and the scores are privately observed by researchers with no impact on any student's life or education. However, if the scores in Z are from tests taken after treatment, then treatment may affect Z. In other settings the tests may determine eligibility for treatment. If the test scores in V determine whether a student receives extra tuition, then V could affect the outcome.

An advantage of the double proxy approach is that it can accommodate rich, direct causal interactions between the proxies, treatments and outcomes. The graphs for a number of additional linear causal models are given in Figure 1.1.b. Some of the causal diagrams in Figure 1.1.b are also featured in Miao *et al.* (2018b), Deaner (2021), and elsewhere. All of these models imply Assumption 1.1 holds with β_0 the causal effect of X on Y (that is, the matrix of coefficients on X in the structural equation for Y). These models in Figure 1.1 are not exhaustive: there are other linear structural models that imply Assumption 1.1. These models in Figure 1.1 are not exhaustive: there are other linear structural models that imply Assumption 1.1. In all of these models V does not directly cause nor is it caused by, X. Moreover, Z does not directly cause nor is it caused by Y. However, Z may cause or be caused by X (so Z is 'treatment aligned'), and V can cause Y (hence 'outcome-aligned').

Theorem 1.1 below refers to C_0 and G_0 which are the matrices of coefficients on Z and

X respectively from population regression of W on Z, X, and D.[3] Recall that variables with tildes have had D partialled out.

Theorem 1.1. *Under Assumption 1.1.i and 1.1.ii the following moment conditions hold:*

$$E\left[\left(\begin{pmatrix}\tilde{V}\\ \tilde{Y}-\beta_0'\tilde{X}\end{pmatrix}-\begin{pmatrix}B_0C_0 & B_0G_0\\ A_0C_0 & A_0G_0\end{pmatrix}\begin{pmatrix}\tilde{Z}\\ \tilde{X}\end{pmatrix}\right)(\tilde{Z}',\tilde{X}')\right]=0 \tag{3}$$

Theorem 1.1 states that under Assumption 1.1 a matrix of moment conditions hold. Assumptions 1.2-1.4 below ensure that the moment conditions in Theorem 1.1 identify β_0 as well as the number of confounding factors d_W.

Assumption 1.2 (V is sufficiently informative about W). B_0 has full column rank.

Assumption 1.3 (Z is sufficiently informative about W). C_0 has full row rank.

Assumption 1.4 (Full support). The matrix $E[(X',Z',D')'(X',Z',D')]$ is non-singular.

Assumptions 1.2 and 1.3 require that the proxies V and Z are sufficiently informative about the confounders W. In particular, each vector of proxies must be relevant instruments for W in the sense of linear Instrumental Variables (IV). More precisely, under Assumptions 1.1 and 1.4, Assumption 1.2 is equivalent to the 'rank condition for identification' in an IV model where W is a vector of endogenous regressors, D and X are exogenous regressors, and V is a vector of instruments. The same is true of Assumption 1.3 but with Z the vector of instruments. Note that Assumptions 1.2 and 1.3 imply an order condition: Z and V must each have weakly larger dimension than W.

Suppose Assumptions 1.1 and 1.4 hold, then 1.2 and 1.3 are equivalent to the following condition:

$$\sigma_{d_W}\left(E[\bar{V}\bar{W}']E[\bar{W}\bar{W}']^{\dagger}E[\bar{W}\bar{Z}']\right)>0 \tag{4}$$

We refer to the quantity on the left-hand side of the inequality as the 'measure of proxy informativeness'. We denote this quantity by $\underline{\sigma}_n$ and it plays an important role in our asymptotic analysis.

Assumption 1.4 is a very mild condition that none of the components of X, Z, and D are perfectly co-linear.

Theorem 1.2. *Under Assumptions 1.1-1.4, β_0 and d_W are identified. More precisely, suppose that for some r, $\beta\in\mathbb{R}^{d_X}$, $A\in\mathbb{R}^{1\times r}$, $B\in\mathbb{R}^{d_V\times r}$, $C\in\mathbb{R}^{r\times d_Z}$, and $G\in\mathbb{R}^{r\times d_X}$ satisfy the moment conditions below, and that B has full column rank:*

$$E\left[\left(\begin{pmatrix}\tilde{V}\\ \tilde{Y}-\beta'\tilde{X}\end{pmatrix}-\begin{pmatrix}BC & BG\\ AC & AG\end{pmatrix}\begin{pmatrix}\tilde{Z}\\ \tilde{X}\end{pmatrix}\right)(\tilde{Z}',\tilde{X}')\right]=0 \tag{5}$$

Then $\beta=\beta_0$, $d_W=rank(BC)=rank\big(B(C,G)\big)=rank\big((B',A')'C\big)$. Moreover, $BC=B_0C_0$, $BG=B_0G_0$, $AC=A_0C_0$, and $AG=A_0G_0$.

[3] Formally, C_0 and G_0 are of dimensions $d_W\times d_Z$ and $d_W\times d_X$ and are given by:

$$(C_0,G_0)=E[\tilde{W}(\tilde{Z}',\tilde{X}')]\Sigma^{\dagger}_{\tilde{Z}\tilde{X}}$$

Theorem 1.2 shows that under Assumptions 1.1-1.4 the object of interest β_0 and the number of confounders d_W are identified from the moment conditions in Theorem 1.1. In addition, the nuisance parameters B_0C_0, B_0G_0, A_0C_0, and A_0G_0 are also identified. Note that it is only these products that are identified: A_0, B_0, C_0, and G_0 are not themselves identified.[4]

The number of confounding factors d_W determines the rank of some matrices of nuisance parameters. If d_W is small then this constraint on the rank constitutes a substantial dimension reduction, which is useful for estimation. The number of unobserved confounders is generally unknown, but since it is identified this suggests we can adapt to this quantity using model selection methods.

Under Assumptions 1.1-1.4, β_0 could be estimated directly from the moment conditions in Theorem 1.1 using the Generalized Method of Moments (GMM) (Hansen (1982)). However, implementing such an estimator may be impractical in this setting. Suppose that d_W were known and we apply GMM enforcing one or more of the rank constraints, for example $rank(B_0C_0) = d_W$. If $d_W < \min\{d_V, d_Z\}$ then the GMM minimization problem does not have a closed-form solution and the problem is non-convex. Thus full GMM estimation would require a numerical optimization routine which may be computationally demanding and may not converge to a global minimum.

In light of the difficulties posed by full GMM estimation, we estimate β_0 by sequential method of moments. The sequential method also allows us to use existing penalized reduced-rank regression methods to estimate and adapt to the number of unobserved confounders. In a first-stage one estimates the relevant nuisance parameters, then in a second-stage one estimates β_0 by inverting a moment condition with the nuisance parameter estimates plugged in. The sequential method allows for estimates with a closed-form solution, even in the case with d_W unknown.

Corollary 1 states the moment conditions that we use for the sequential estimator. The corollary first provides an alternative set of moment conditions that identify β_0. We prove in Lemma 1 that (under Assumptions 1.1-1.4) this characterization of β_0 is in fact equivalent to that in Theorem 1.2. The alternative moment conditions depend on two nuisance parameters M_0 and ξ_0. The corollary then states that these nuisance parameters can be identified from moment conditions that do not involve β_0 and which are linear in parameters.

It is of note that, unlike the moment conditions in Theorem 1.1, which hold under Assumption 1.1 alone, the moment conditions in Corollary 1 need not hold (for any M_0 and ξ_0) without some of the remaining assumption 1.2-1.4.

Corollary 1. *Under Assumptions 1.1-1.4 β_0 is identified from (6) below:*

$$E\big[(\tilde{Y} - \beta_0'\tilde{X} - \xi_0'\tilde{V}')(\tilde{Z}', \tilde{X}')\big] = 0 \tag{6}$$

$$E\big[(\tilde{V} - M_0(\tilde{Z}', \tilde{X}')')(\tilde{Z}', \tilde{X}')\big] = 0 \tag{7}$$

$$E\big[((\bar{V}', \bar{Y}')' - Q_0\bar{Z})\bar{Z}'\big] = 0 \tag{8}$$

$M_0 = B_0(C_0, G_0)$ is the unique solution to (7). $Q_0 = (B_0', A_0')'C_0$ is the unique solution to (8). $rank(M_0) = rank(Q_0) = d_W$. (6) is satisfied by any ξ_0 that solves $\xi_0'B_0C_0 = A_0C_0$, and there exists a solution with $\|\xi_0\|_0 \leq d_W$.

[4] A_0, B_0, C_0, and G_0 are only identified up to non-singular transformations. More precisely, if A, B, C, and G satisfy (5) then so do matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{G}$ of the same dimensions where $(\tilde{B}, \tilde{A}')' = (B', A')'\Omega$ and $(\tilde{C}, \tilde{G}) = \Omega^{-1}(C, G)$ for any non-singular matrix Ω.

The first moment condition in Corollary 1, (6) suffices to identify β_0, but the other two conditions (7) (8) help to identify d_W. Moreover, ξ_0 can be written in terms of M_0 and Q_0 and thus ξ_0 can be obtained from (7) and (8), which motivates the sequential estimation strategy.

Corollary 1 suggest three different means of adapting to the number of confounding factors d_W. Firstly, d_W is the rank of M_0, secondly it is the rank of Q_0. Thirdly, there is a ξ_0 that satisfies (6) with at most d_W non-zero entries.

1.1 Comparison to Previous Results

Our analysis is related to that of Miao *et al.* (2018a) and a much older line of work surveyed in Griliches (1977). However, our results differ in important respects. Perhaps most important for our subsequent analysis, we link the ranks of nuisance parameter matrices to the number of unobserved confounders d_W, and we show this is identified. The rank restrictions can result in a substantial dimension reduction which may greatly reduce estimation error, particularly when there are many available proxies. Furthermore, the rank restrictions have important implications for inference even in when the number of available proxies is small.

Moreover, our identification results clarify the sense in which the proxies must be informative about the confounders in the linear model. Assumptions 1.2 and 1.3 appear absent from Griliches (1977) and related works, which implicitly assume the relevant moment conditions are invertible. This absence could be explained by the focus of those works on a single scalar unobservable in which case the rank conditions reduce to a simpler condition that the proxies are partially correlated with the confounder.

Let us compare our results with Miao *et al.* (2018a). For simplicity let us assume there are no observed confounders D. Miao *et al.* (2018a) assume the existence of a function that they call a 'confounding bridge' which then plays a key role in their analysis. This is a function b with the property that for each x in the support of X, with probability 1:

$$E[Y|W, X = x] = E[b(V, x)|W, X = x]$$

Suppose our Assumptions 1.1-1.4 hold and ϵ and υ are mean independent of W (rather than just uncorrelated with W), then under Assumptions 1.1-1.4 our model admits a confounding bridge of the form $b(v, x) = \beta_0' x + \xi_0' v$, where ξ_0 is any solution to $\xi_0' B_0 C_0 = A_0 C_0$ (just as in Corollary 1).

Miao *et al.* (2018a) impose assumptions that imply the confounding bridge is unique and point identified. In our model it may be neither unique nor point identified. In fact, under Assumptions 1.1-1.4 the confounding bridge is generally not unique unless $d_V = d_W$, otherwise it is generically true that there are multiple solutions to $\xi_0' B_0 C_0 = A_0 C_0$.[5] Even if the confounding bridge is unique, in order to identify the bridge, Z must be a vector of relevant instruments for V after controlling for X (see Assumption 5 in Miao *et al.* (2018a)). Again, under our assumptions this is only possible when $d_V = d_W$. Thus the analysis of Miao *et al.* (2018a) can only apply in our model when there are the same number of negative outcome proxies as instruments.

Applying the methods of Miao *et al.* (2018a) in our model amounts to using GMM to estimate solutions β_0 and ξ_0 to the moment condition (6) without any of the constraints related

[5] Under Assumptions 1.1-1.4 C_0 has full row rank and so there is a unique solution if and only if A_0 is in the row space of B_0. Since the row space of B_0 is of dimension d_W and A_0 is a row vector of length $d_V > d_W$, this is generically false.

to d_W.[6] Griliches (1977) suggests estimation using instrumental variables that amounts to the use of the moment condition (6), again without any dimension reduction.

When $d_W < d_V$, estimation without any rank constraints can lead to invalid inference. The reason is that the nuisance parameter ξ_0 that satisfies (6) is not unique (and thus not identified). Method of moments estimates of ξ_0 that do not incorporate rank-selection generally do not converge in probability to any fixed limit. As we show in Appendix C, this is generally not a problem for consistency (at least in finite dimensions) because the resulting estimates of ξ_0 are stochastically bounded. However, this typically leads to estimates of β_0 that are not asymptotically normal, in which case standard inference methods can be misleading. Imbens *et al.* (2021) avoids problems of inference by using l_2 penalization rather than rank selection.

2 Estimation and Inference

We now present an estimator motivated by the results in Corollary 1. Our estimator is an example of a DML 2 (Double Machine Learning 2) estimator as developed in Chernozhukov *et al.* (2018). In order to apply DML 2 estimation in our setting, we must orthogonalize the moment condition (6). The score function in (6) is denoted by $g(\beta, \xi)$ and is given below:

$$g(\beta, \xi) = (\tilde{Y} - \beta' \tilde{X} - \xi' \tilde{V})(\tilde{Z}', \tilde{X}')'$$

For notational convenience we leave the dependence of the score on the nuisance parameters involved in partialling out D (that is, $\gamma_{Y,0}$, $\gamma_{X,0}$ etc.) implicit.

In order to orthogonalize the moment condition (6), we pre-multiply the score function by a matrix μ_0 with the following formula:

$$\mu_0 = G_\beta' \Omega^{-1/2} \big(I - \Omega^{-1/2} G_\xi (\Omega^{-1/2} G_\xi)^\dagger\big) \Omega^{-1/2} \tag{9}$$

In the above G_β and G_ξ are the matrices of derivatives of $E[g(\beta_0, \xi_0)]$ with respect to β_0 and ξ_0 respectively. They have formulas $G_\beta = -E[(\tilde{Z}_i', \tilde{X}_i')' \tilde{X}_i']$ and $G_\xi = -\Sigma_{\tilde{Z}\tilde{X}} M_0'$. The matrix Ω is the variance-covariance of $g(\beta_0, \xi_0)$.

pre-multiplication by μ_0 results in a doubly-robust moment condition:

$$\mu_0 E[\big(\tilde{Y} - \beta_0 \tilde{X} - \xi_0 \tilde{V}'\big)(\tilde{Z}', \tilde{X}')] = 0 \tag{10}$$

The moment condition above is doubly robust. If we replace ξ_0 in the above by $\xi \neq \xi_0$ then the condition still holds. Similarly, if we replace μ_0 by $\mu \neq \mu_0$ the condition continues to hold. The moment condition is also robust to the coefficients involved in partialling out D. We show this formally in the proof Lemma 2 in the Appendix.

In fact, in the formula for μ_0, G_β could be any matrix and Ω any strictly positive-definite matrix and the resulting moment condition would still be doubly robust. However, the choices of Ω and G_β above are efficient (see e.g., subsection 2.2.2 of Chernozhukov *et al.* (2018) for discussion).

We now specify our estimator. We assume we have access to an iid sample of the variables Y, X, Z, V, D of size n and we index the individual observations by 'i'.

[6] Miao *et al.* (2018a) allow the instruments (Z', X') to be replaced with any vector of transformations $q(Z, X)$ with finite variance. However, in our model if q is nonlinear then the resulting moment conditions are valid only when the zero correlation conditions in Assumption 1.1.ii are strengthened to mean independence.

2.1 Nuisance Parameter Estimates

The doubly-robust moment condition (10) depends on nuisance parameters ξ_0 and μ_0 which we estimate in a first stage. The key step is to attain estimates of M_0 and Q_0 from the conditions (7) and (8) using adaptive reduced-rank regression. ξ_0 and μ_0 can then be written in terms of M_0 and Q_0 and some sample means. In addition, D is partialled out from the variables in (10). We must obtain estimates of the regression coefficients used to partial out D.

First, let us consider estimators for the coefficients used to partial out D, namely $\gamma_{0,X}$, $\gamma_{0,Z}$, $\gamma_{0,V}$, and $\gamma_{0,Y}$. If the vector of additional covariates D is low-dimensional then one can simply ordinary least-squares, for example the estimate of $\gamma_{0,V}$ is:

$$\hat{\gamma}_V = (\frac{1}{n}\sum_{i=1}^{n} D_i D_i')^{\dagger} \frac{1}{n}\sum_{i=1}^{n} D_i V_i'$$

Similarly, one may estimate the the coefficients involved in partialling out both D and X, which are $\omega_{0,Z}$, $\omega_{0,V}$, and $\omega_{0,Y}$, by least-squares. For example, estimate of $\omega_{0,Z}$ by $\hat{\omega}_Z$ are:

$$\hat{\omega}_Z = (\frac{1}{n}\sum_{i=1}^{n} (X_i', D_i')'(X_i', D_i'))^{\dagger} \frac{1}{n}\sum_{i=1}^{n} (X_i', D_i')' Z_i'$$

We use hats to indicate a variable with D partialled out in the sample, so for example $\hat{V}_i = V_i - \hat{\gamma}_V' D_i$, or equivalently $\hat{V}_i = \tilde{V}_i(\hat{\gamma}_V)$. Check marks indicate a variables with both D and X partialled out in the sample, so for example $\check{Z}_i = Z_i - \hat{\omega}_Z'(X_i', D_i')'$ or equivalently $\check{Z}_i = \bar{Z}_i(\hat{\omega}_Z)$.

In some cases D_i may be high-dimensional and the true partialling out parameters may be sparse or approximate sparse. In this case we suggest that Lasso regression (Tibshirani (1996)) be used in place of ordinary least squares.

Now we turn to estimation of ξ_0 and μ_0 which requires estimates of M_0 and Q_0. Corollary 1 states that M_0 and Q_0 are the unique solutions to the moment conditions (7) and (8). These are standard least-squares moment conditions and so M_0 and Q_0 minimize sum-of-squares objectives. The corollary also states that M_0 and Q_0 are each of rank d_W. We can define least-squares estimates of M_0 and Q_0 that are subject to rank constraints as follows:

$$\hat{M}_r = \underset{rank(M)\leq r}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \|\hat{V}_i - M(\hat{Z}_i', \hat{X}_i')'\|^2 \tag{11}$$

$$\hat{Q}_r = \underset{rank(Q)\leq r}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \|(\check{V}_i', \check{Y}_i')' - Q\check{Z}_i\|^2 \tag{12}$$

$\hat{M}_r$ and $\hat{Q}_r$ are reduced-rank regression estimates and thus have closed-form solutions (Reinsel & Velu (1998), Izenman (1975)). The formulas for the solutions and other algorithmic details are provided in Appendix B.

Note that r, the bound on the rank, determines the number of free parameters in the minimization problem. If r is small compared to $\min\{d_V, d_Z\}$ then the constraint imparts a considerable dimension reduction. Ideally we would set $r = d_W$. However, d_W is generally not known a priori, but since it is identified one can adapt to this quantity.

In order to adapt to the unknown number of confounders d_W let us replace the constrained least-squares problems (11) and (12) with penalized least-squares problems. Let

$\lambda_{M,n}$ and $\lambda_{Q,n}$ be positive scalars that control the degree of regularization. We define penalized estimators as follows:

$$\hat{M} = \underset{rank(M)\leq d_Z}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \|\hat{V}_i - M(\hat{Z}_i', \hat{X}_i')'\|^2 + \lambda_{M,n} rank(M), \tag{13}$$

$$\hat{Q} = \underset{rank(Q)\leq d_V}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \|(\check{V}_i', \check{Y}_i')' - Q\check{Z}_i\|^2 + \lambda_{Q,n} rank(Q) \tag{14}$$

Bunea *et al.* (2011) provide the formula for the solution to penalized reduced-rank regression problems like (13) and (14). Their results show $\hat{M} = \hat{M}_{\hat{r}_M}$ and $\hat{Q} = \hat{Q}_{\hat{r}_Q}$, where $\hat{r}_M$ and $\hat{r}_Q$ are estimators of the number of unobserved confounders d_W. The formulas for $\hat{r}_M$ and $\hat{r}_Q$ are provided in the appendix.

The penalty parameters $\lambda_{M,n}$ and $\lambda_{Q,n}$ can be chosen by cross-validation or by a plug-in formula provided by Bunea *et al.* (2011). In our simulations we find cross-validation generally selects the correct rank with higher frequency and so we take this as our preferred method.

Alternative adaptive reduced-rank regression methods are available in the literature. One such a method replaces the term $rank(M)$ in (13) with the nuclear norm of M, commonly denoted $\|M\|_*$.[7] Penalizing the rank has the advantage that the solution has a closed-form. Another alternative, proposed in Bing & Wegkamp (2019) uses estimates which again take the form $\hat{M} = \hat{M}_{\hat{r}_M}$ and $\hat{Q} = \hat{Q}_{\hat{r}_Q}$, but where the selected ranks $\hat{r}_M$ and $\hat{r}_Q$ are chosen using an iterative method.

Having obtained estimates of M_0 and Q_0 we can estiamte μ_0 and ξ_0. Corollary 1 states that ξ_0 solves $\xi_0' B_0 C_0 = A_0 C_0$ and that $B_0 C_0$ and $A_0 C_0$ both sub-matrices of Q_0. We replace $B_0 C_0$ and $A_0 C_0$ in this equation with the corresponding sub-matrices of $\hat{Q}$ and take as our estimate $\hat{\xi}$ the solution with smallest Euclidean norm:

$$\hat{\xi}' = \hat{Q}_{[d_V+1,:]} \hat{Q}^{\dagger}_{[1:d_V,:]} \tag{15}$$

In order to estimate μ_0 we replace G_ξ, G_β, and Ω in the formula for μ_0 with estimates $\hat{G}_\xi$, $\hat{G}_\beta$, and $\hat{\Omega}$. The resulting estimate $\hat{\mu}$ is given below:

$$\hat{\mu} = \hat{G}_\beta' \hat{\Omega}^{-1/2} \big(I - \hat{\Omega}^{-1/2} \hat{G}_\xi (\hat{\Omega}^{-1/2} \hat{G}_\xi)^{\dagger}\big) \hat{\Omega}^{-1/2} \tag{16}$$

The estimates of G_ξ and $\hat{G}_\beta$ are given below. Note the dependence of $\hat{G}_\xi$ on $\hat{M}$, the estimate of M_0.

$$\hat{G}_\xi = -\frac{1}{n}\sum_{i=1}^{n} (\hat{Z}_i', \hat{X}_i')'(\hat{Z}_i', \hat{X}_i')\hat{M}'$$

$$\hat{G}_\beta = -\frac{1}{n}\sum_{i=1}^{n} (\hat{Z}_i', \hat{X}_i')'\hat{X}_i'$$

For $\hat{\Omega}$ we suggest either the identity matrix or the following estimate obtained by a two-step procedure. Recall that the efficient choice of Ω is the variance matrix of $g(\beta_0, \xi_0)$. Let $\tilde{\beta}$ be an initial estimate of $\hat{\beta}$ obtained using the identity for $\hat{\Omega}$. We can then estimate the

[7] See Chen *et al.* (2013) for some analysis of nuclear norm penalization in reduced-rank regression.

efficient Ω by letting $\hat{\Omega}$ be the sample variance-covariance matrix of $\hat{g}_i(\tilde{\beta}, \hat{\xi})$ which is defined below:

$$\hat{g}_i(\tilde{\beta}, \hat{\xi}) = (\hat{Y}_i - \beta' \hat{X}_i - \xi' \hat{V}_i)(\hat{Z}_i', \hat{X}_i')'$$

2.2 Second-Stage Estimator

We estimate β_0 by inverting an empirical analogue of (10) with the nuisance parameters replaced with the estimates specified in the previous sub-section. Following Chernozhukov *et al.* (2018) we employ sample-splitting to reduce bias.

For the purpose of exposition let us first specify a version of the estimator without sample-splitting. Define an estimate of the score function for the robust moment condition (10) evaluated at observation i as follows:

$$\hat{\psi}_i(\beta, \xi, \mu) = \mu(\hat{Y}_i - \beta' \hat{X}_i - \xi' \hat{V}_i)(\hat{Z}_i', \hat{X}_i')' \tag{17}$$

Note that the above is an estimate of the true score because we have partialled out D from the variables on the right-hand-side empirically rather than exactly. An estimate of β_0 that does not employ sample-splitting solves the empirical moment condition $\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_i(\beta; \hat{\xi}, \hat{\mu}) = 0$. The solution $\check{\beta}$ has the following formula:

$$\check{\beta} = \big(\sum_{i=1}^{n} \hat{X}_i(\hat{Z}_i', \hat{X}_i')\hat{\mu}'\big)^{\dagger} \sum_{i=1}^{n} \hat{\mu}(\hat{Z}_i', \hat{X}_i')'(\hat{Y}_i - \hat{\xi}' \hat{V}_i')$$

To apply sample splitting, one partitions the data into J sub-samples. Let $\{\mathcal{I}_j\}_{j=1}^{J}$ be a partition of $\{1, ..., n\}$ and let n_j be the number of entries in $\mathcal{I}_j$. Thus each index $i = 1, ..., n$ is a member of precisely one subset $\mathcal{I}_j$ and $\sum_{j=1}^{J} n_j = n$. We will use the shorthand $\mathcal{I}_{-j}$ to denote all the elements of $\{1, ..., n\}$ that are not in $\mathcal{I}_j$ (i.e., the complement of $\mathcal{I}_j$).

For each $j = 1, ..., J$ the researcher evaluates each of the nuisance parameter estimates using only the observations with indices in $\mathcal{I}_{-j}$, that is, the data outside of the j^{th} subsample. Thus, for each j, one obtains estimates $\hat{\xi}_j$, $\hat{\mu}_j$ of ξ_0 and μ_0. We also suggest sample-splitting when partialling out D and X. For each j one obtains an estimate $\hat{\gamma}_{j,X}$ of $\gamma_{X,0}$ using only the data in $\mathcal{I}_{-j}$ and likewise for $\gamma_{Y,0}$, $\gamma_{V,0}$, etc. Thus for each j one obtains a separate estimate of $\tilde{X}_i$ denoted by $\hat{X}_{j,i} = X_i - \hat{\gamma}_{j,X}' D_i$, and likewise for the other variables. This results in a separate estimate for the score function for each j:

$$\hat{\psi}_{j,i}(\beta, \xi, \mu) = \mu(\hat{Y}_{j,i} - \beta' \hat{X}_{j,i} - \xi' \hat{V}_{j,i})(\hat{Z}_{j,i}', \hat{X}_{j,i}')'$$

The estimate $\hat{\beta}$ with sample-splitting solves the following empirical analogue of the doubly robust moment condition (10):

$$\sum_{j=1}^{J} \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} \hat{\psi}_{j,i}(\hat{\beta}; \hat{\xi}_j, \hat{\mu}_j) = 0$$

The solution has the formula below:

$$\hat{\beta} = \big(\sum_{j=1}^{J} \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} \hat{X}_{j,i}(\hat{Z}_{j,i}', \hat{X}_{j,i}')\hat{\mu}_j'\big)^{\dagger} \sum_{j=1}^{J} \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} \hat{\mu}_j(\hat{Z}_{j,i}', \hat{X}_{j,i}')'(\hat{Y}_{j,i} - \hat{\xi} \hat{V}_{j,i}')$$

2.3 Inference

Chernozhukov *et al.* (2018) suggests a variance estimator for DML2 estimators. In the case of our estimator $\hat{\beta}$ the variance estimate is as follows:

$$\hat{s}^2 = \frac{1}{n}\sum_{j=1}^{J}\sum_{i\in\mathcal{I}_j}\hat{S}^{-1}\hat{\psi}_{j,i}\hat{\psi}_{j,i}'(\hat{S}^{-1})'$$

Where we let $\hat{\psi}_{j,i} = \hat{\psi}_{j,i}(\hat{\beta}, \hat{\xi}_j, \hat{\mu}_j)$ and define the matrix $\hat{S}$ by:

$$\hat{S} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i\in\mathcal{I}_j}\hat{\mu}_j(\hat{Z}_i', \hat{X}_i')'\hat{X}_i'$$

Note that the above estimates the symmetric matrix $S_0 = \mu_0 E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i']$.

If the variance estimator is consistent and $\hat{\beta}$ is asymptotically Gaussian centered at β_0, then a confidence interval for $l'\beta_0$ (where l is some vector) can be obtained as follows:

$$CI = \left[l'\hat{\beta} \pm \Phi^{-1}(1-\alpha/2)\sqrt{l'\hat{s}^2 l/n}\right]$$

The formula above is suggested in Chernozhukov *et al.* (2018). Φ is the cumulative distribution function of a standard Gaussian random variable.

3 Consistency and Asymptotic Normality

The methods in the previous section estimate a parameter of interest β_0 in the presence of possibly high-dimensional nuisance parameters. We take the standard approach to asymptotic analysis in such settings which is to find conditions under which the estimates are root-n consistent and admit an asymptotic Gaussian approximation.

For readers who wish to skip the more technical results in this section, some of the key takeaways are as follows. Firstly, to ensure that $\hat{\xi}$ and $\hat{\mu}$ are consistent, we require not only that the reduced-rank regression estimates $\hat{M}$ and $\hat{Q}$ are consistent, but that the rank d_W be selected correctly with probability approaching 1.

Standard IV estimation does not impose any rank-restrictions. Thus if we were to apply IV in our setting (as in Griliches (1977)), the resulting estimates of ξ_0 are generally inconsistent when the number of proxies in each group is strictly larger than the number of confounders. As we discuss in detail in Appendix C, despite the inconsistent estimation of ξ_0, standard IV estimators of β_0 are in fact consistent under fixed-dimensional asymptotics (under some weak regularity conditions), but they generally have a non-standard distribution.

Secondly, the quality of the nuisance parameter estimates depends upon the measure of proxy informativeness introduced in Section 1. If this quantity is small then $\hat{\xi}$ and $\hat{\mu}$ may be subject to substantial error compared with $\hat{M}$ and $\hat{Q}$.

Finally, suppose that $\|\hat{\xi} - \xi_0\| = O_p(\delta_\xi)$ and $\|\hat{\mu} - \mu_0\| = O_p(\delta_\mu)$. A key condition that we use to establish root-n asymptotic normality of the doubly-robust estimator $\hat{\beta}$ is that the product $\delta_\xi\delta_\mu$ goes to zero faster than $n^{-1/2}$.

It is helpful to introduce some additional notation. Firstly, we make the dependence of the robust core function on the parameters involved in partialling out D explicit:

$$\psi(\beta, \xi, \mu, \gamma) = \mu\big(\tilde{Y}(\gamma_Y) - \beta'\tilde{X}(\gamma_{X,1}) - \xi'\tilde{V}(\gamma_V)\big)\big(\tilde{Z}(\gamma_Z)', \tilde{X}(\gamma_{X,2})'\big)' \tag{18}$$

Note the distinction between $\gamma_{X,1}$ and $\gamma_{X,2}$. While both of these quantities take the place of the same parameter $\gamma_{X,0}$ it is useful to treat them as separately for analytical purposes. The argument γ in the score function collects $\gamma_{X,1}$, $\gamma_{X,2}$, γ_Y etc. into one parameter $\gamma = (\gamma_{X,1}, \gamma_{X,2}, \gamma_Z, \gamma_V, \gamma_Y)$ with true value $\gamma_0 = (\gamma_{X,0}, \gamma_{X,0}, \gamma_{Z,0}, \gamma_{V,0}, \gamma_{Y,0})$. For convenience we write $\psi = \psi(\beta_0, \xi_0, \mu_0, \gamma_0)$.

We define the residual $\tilde{\epsilon}$ as follows:

$$\tilde{\epsilon} = \tilde{Y} - \tilde{X}'\beta_0 - \tilde{V}'\xi_0'$$

Relative to some γ we define γ_ϵ by $\gamma_\epsilon = \gamma_Y - \gamma_V\xi_0' - \gamma_{X,2}\beta_0$ and $\gamma_{\epsilon,0}$ is defined in the same way but with γ_0 in place of γ.

3.1 Asymptotic Analaysis of the Doubly-Robust Estimator

We begin by establishing root-n consistency and asymptotic normality of the doubly-robust estimator with sample splitting as defined in (2.2). These results apply for any choice of nuisance parameter estimators $\hat{\xi}$ and $\hat{mu}$ not just those specified in Section 2. We analyze our suggested nuisance parameter estimates later in this section.

The doubly-robust estimator is a Double-Machine Learning 2 (DML2) estimator of the kind analyzed in section 2 in Chernozhukov *et al.* (2018). DML2 estimators (along with the DML1 estimators in Chernozhukov *et al.* (2018)) have the advantage that they are root-n consistent and centered asymptotically normal under relatively weak conditions on the nuisance parameter estimates.

Our asymptotic analysis is based on Theorems 3.1 and 3.2 in Chernozhukov *et al.* (2018). The Assumptions 1.1-1.4 and 3.1-3.2 (stated below) act as primitive conditions for the assumptions in that paper.

In order to derive results that are uniform over a growing sequence of parameter spaces we suppose that for each sample size n, the data generating process, denoted by P, belongs to some set $\mathcal{P}_n$. Our results then rely on conditions that restrict $\mathcal{P}_n$.

Assumption 3.1 (Convergence rates of the nuisance parameter estimates). There is a sequence α_n with $\alpha_n \to 0$ so that if $P \in \mathcal{P}_n$ then with probability at least $1 - \alpha_n$ the following hold for $j = 1, ..., J$. i. $\|\hat{\mu}_j - \mu_0\| \leq \delta_\mu \prec d_X^{-1/2}$. ii. $\|\hat{\xi}_j - \xi_0\| \leq \delta_\xi \prec d_X^{-1/2}$. iii. For $H = X, V, Z, \epsilon$, $\|\hat{\gamma}_{H,j} - \gamma_{H,0}\| \leq \delta_{\gamma,H} \prec d_X^{-1/2}$.

Assumption 3.2 (Restrictions on the DGP). There are constants $c > 0$ and $q > 2$ so that if $P \in \mathcal{P}_n$ the following hold: i. $E[\tilde{\epsilon}^2] \leq c$, the eigenvalues of $\Sigma_{\tilde{V}}$, $\Sigma_{\tilde{Z}\tilde{X}}$, and Σ_D are bounded below by $1/c$ and above by c. ii. For each $H \in \{\tilde{V}, (\tilde{Z}', \tilde{X}')'\}$, $\|E[HH'|D]\| \leq c$ with probability 1, and for each $H \in \{\tilde{V}, \tilde{\epsilon}\}$, $\|E[HH'|\tilde{Z}, \tilde{X}]\| \leq c$ with probability 1. iii. For any matrix $A \in \mathbb{R}^{d_X, d_D}$ and vector b, $E\big[\|ADD'b\|^2\big] \leq d_X\|A\|^2\|b\|^2c^2$. iv. $\|M_0\|, \|\Omega^{-1}\|, \|\xi_0\| \leq c$. v. If $\|\xi - \xi_0\|, \|\mu - \mu_0\|, \|\gamma_H - \gamma_{H,0}\| \leq 1/c$ for each $H \in \{\epsilon, V, Z, X\}$, then $E\big[\|\psi(\beta_0; \xi, \gamma, \mu)\|^q\big]^{1/q} \leq c$ and $E\big[\|\mu(\tilde{Z}', \tilde{X}')'\tilde{X}'\|^q\big]^{1/q} \leq c$.

Assumption 3.1 imposes convergence rates for each of the nuisance parameter estimates. Note that the convergence rates are required to hold uniformly over sequences of DGPs in $\{\mathcal{P}_n\}_{n=1}^\infty$.

Assumption 3.2 imposes bounds on the rates at which the magnitudes of some population objects grow with the sample size. Note that Assumption 3.2.v requires that some higher-order moments of the score exist and be bounded uniformly over n when the nuisance

parameter arguments are close to the true values. Existence of higher-order moments is a standard assumption in problems with growing dimension as an assumption of this kind is generally required for an application of a multivariate central limit theorem.

Theorem 3.2. *Suppose that for each n, $P \in \mathcal{P}_n$ so that Assumptions 1.1-1.4, 3.1, and 3.2 hold, the singular values of S_0 are bounded uniformly below and away from zero, and the eigenvalues of $E[\psi\psi']$ are bounded uniformly above and below away from zero.*

In addition, suppose that $\delta_\mu \delta_\xi \prec n^{-1/2}$ and $(\sqrt{n}+\sqrt{d_X})(\delta_{\gamma,X}+\delta_{\gamma,Z})(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi) \prec 1$. Then uniformly over all $P \in \mathcal{P}_n$, $\hat{\beta}$ is root-n consistent and asymptotically normal:

$$\sqrt{n}s^{-1}(\beta_0 - \hat{\beta}) \rightsquigarrow N(0, I)$$

Where the asymptotic variance s^2 is given by: $s^2 = S_0^{-1}E[\psi\psi']S_0^{-1}$. Moreover, the estimator $\hat{s}^2$ is consistent for s^2 and the confidence intervals described earlier in this section have asymptotically correct coverage.

Theorem 3.2 establishes uniform root-n consistency of the estimator and asymptotic validity of the confidence intervals. In addition to Assumptions 1.1-1.4, 3.1, and 3.2, the theorem requires a number of conditions on the rates at which the nuisance parameters converge.

The condition that the singular values of S_0 are bounded uniformly below and that $E[\psi\psi']$ has eigenvalues bounded uniformly above ensures that the asymptotic variance of the estimator is uniformly bounded.

3.2 Consistency of the Nuisance Parameter Estimates

We now bound the error in the nuisance parameters ξ_0 and μ_0. In this subsection ξ_0 refers specifically to the solution to $\xi_0 B_0 C_0 = A_0 C_0$ with minimal Euclidean norm. A minimum-norm solution to moment conditions with proxy controls is also targeted by Imbens *et al.* (2021). Throughout this subsection $\hat{\xi}$ and $\hat{\mu}$ refer to the estimators of ξ_0 and μ_0 described in Section 2.

Proposition 3.1 relates the degree of error in the nuisance parameter estimates $\hat{\xi}$ and $\hat{\mu}$ to the quality of the corresponding estimates of M_0 and Q_0.

Proposition 3.1. *Suppose that Assumptions 1.1-1.4 hold and there exists $0 < c_1 < \infty$ so that $\|B_0(C_0, G_0)\|, \|A_0 C_0\| \leq c_1$, and the eigenvalues of Ω, $\Sigma_{\bar{Z}}$, and $\Sigma_{\tilde{Z}\tilde{X}}$ are bounded above and below by c_1 and $1/c_1$. Let $\{\alpha_n\}_{n=1}^\infty$ be a sequence in $(0,1]$ with $\alpha_n \to 0$.*

a. *Suppose that with probability at least $1-\alpha_n$, $rank(\hat{Q}_{[1:d_V,:]}) = d_W$ and $\|\hat{Q}_{[1:d_V,:]} - B_0 C_0\| \leq c_1 \underline{\sigma}_n \alpha_n$. Then there is a constant $c < \infty$ that depends only on $\{\alpha_n\}_{n=1}^\infty$ and c_1 so that with probability at least $1-\alpha_n$:*

$$\|\hat{\xi} - \xi_0\| \leq c(1 + \underline{\sigma}_n^{-2})\|\hat{Q} - Q_0\|$$

b. *Suppose that with probability at least $1-\alpha_n$, $rank(\hat{M}) = d_W$, $\hat{\Omega}$ and $\hat{\Sigma}_{\hat{Z}\hat{X}}$ have full rank, and $c_1 \underline{\sigma}_n \alpha_n$ exceeds $\|M_0 - \hat{M}\|$, $\|\hat{\Sigma}_{\hat{Z}\hat{X}} - \Sigma_{\tilde{Z}\tilde{X}}\|$, and $\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\|$. Then there is an $m \in \mathbb{N}$ and constant $c < \infty$ that depend only on $\{\alpha_n\}_{n=1}^\infty$ and c_1 so that for all $n \geq m$, with probability at least $1-\alpha_n$:*

$$\|\hat{\mu} - \mu_0\| \leq c(1 + \underline{\sigma}_n^{-2})\big(\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| + \|\hat{\Sigma}_{\hat{Z}\hat{X}} - \Sigma_{\tilde{Z}\tilde{X}}\| + \|\hat{M} - M_0\|\big)$$

Proposition 3.1 is non-asymptotic in that it bounds the finite-sample estimation error in $\hat{\xi}$ and $\hat{\mu}$. Formulas for the constants c and m can be found in the proof of the result.

Recall that the measure of proxy informativeness $\underline{\sigma}_n$, is defined as the d_W-th smallest singular value on the LHS of (4). When $\underline{\sigma}_n$ is small (suggesting uninformative proxies) Proposition 3.1 allows for the possibility that the error in ξ_0 and μ_0 is large compared to level the error in the intermediate estimates of M_0 and Q_0.

Proposition 3.1 only establishes consistency of the nuisance parameter estimates if the rank is estimated consistently. We discuss the need for consistent rank selection and how this may be relaxed in detail in Appendix C. Note that a linear IV estimation strategy (like that of Griliches (1977)) necessarily fails to select the correct rank when d_W is strictly smaller than d_V. As we discuss in the appendix, the corresponding estimates of β_0 may be consistent nonetheless. However, these estimates generally have a non-standard asymptotic distribution, even under standard finite-dimensional asymptotics. This means that standard inference methods based on IV estimation are invalid in this setting.

Bunea *et al.* (2011) provide finite-sample statistical results for their adaptive reduced-rank regression method. If $\underline{\sigma}_n$ is fixed and strictly positive, then consistent rank selection simply requires that the penalty parameters go to zero sufficiently slowly. The rate at which these penalties may go to zero depends on the noise levels $r_{M,n}$ and $r_{Q,n}$ defined below.

$$r_{M,n} = \|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1/2} \frac{1}{n} \sum_i^n (\hat{Z}_i', \hat{X}_i')' U_{i,M}'\|$$

$$r_{Q,n} = \|\hat{\Sigma}_{\check{Z}}^{-1/2} \frac{1}{n} \sum_i^n \check{Z}_i U_{Q,i}'\|$$

The regression residuals $U_{M,i}$ and $U_{Q,i}$ are defined as follows:

$$U_{M,i} = \hat{V}_i - M_0(\hat{Z}_i', \hat{X}_i')'$$
$$U_{Q,i} = (\check{V}_i', \check{Y}_i)' - Q_0 \check{Z}_i$$

Proposition 3.2 below applies some of the finite-sample results in Bunea *et al.* (2011) to our setting. '$\|\cdot\|_F^2$' denotes the squared Frobenius norm (the sum of the squared entries of the matrix).

Proposition 3.2 (Bunea et. al.). ***a.*** *Suppose Assumptions 1.1-1.4 hold and there are scalars $a > 0$ and $b > 0$ so that with probability at least $1 - \alpha_n$, $\|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1}\| \leq b$ and:*

$$(1+a) r_{M,n}^2 < \lambda_{M,n} < \left(1 + (1+a)^{-1/2}\right)^{-2} b^{-1} \underline{\sigma}_n^2 / \|\Sigma_{\bar{Z}}^{-1}\|^2$$

Then $P(rank(\hat{M}) = d_W) \geq 1 - 2\alpha_n$ and with probability at least $1 - \alpha_n$:

$$\|\hat{M} - M_0\|_F^2 \leq 2b(1 + 2/a)\lambda_{M,n} d_W$$

b. *Suppose Assumptions 1.1-1.4 hold and there are scalars $a > 0$ and $b > 0$ so that with probability at least $1 - \alpha_n$, $\|\hat{\Sigma}_{\check{Z}}^{-1}\| \leq b$ and:*

$$(1+a) r_{Q,n}^2 < \lambda_{Q,n} < \left(1 + (1+a)^{-1/2}\right)^{-2} b^{-1} \underline{\sigma}_n^2 / \|\Sigma_{\bar{Z}}^{-1}\|^2$$

Then $P(rank(\hat{Q}) = d_W) \geq 1 - 2\alpha_n$ and with probability at least $1 - \alpha_n$:

$$\|\hat{Q} - Q_0\|_F^2 \leq 2b(1 + 2/a)\lambda_{Q,n} d_W$$

Proposition 3.2 bounds the Frobenius norms of the errors in $\hat{M}$ and $\hat{Q}$. Note that the Euclidean matrix norm is always bounded above by the Frobenius norm, as is the Euclidean norm of any sub-vector of a matrix.

Bunea *et al.* (2011) provide results that (applied to our setting) show that if all the entries of $U_{M,i}$ and $U_{Q,i}$ are jointly independent, independent of $\hat{Z}$ and $\hat{X}$, and jointly standard Gaussian, then $r_{M,n}$ and $r_{Q,n}$ converge to zero in probability at respective rates $(\sqrt{d_V} + \sqrt{d_Z + d_X})/\sqrt{n}$ and $(\sqrt{1 + d_V} + \sqrt{d_Z})/\sqrt{n}$. These rates apply even in very high dimensions. Note that the same rates apply if $U_{M,i}$ and $U_{Q,i}$ are multivariate Gaussian with second moment matrices whose smallest eigenvalues are both bounded below away from zero by a constant.

4 Simulation Study

In order to assess the efficacy of the methods we present in Section 2 we carry out a Monte Carlo simulation. We implement our methods on a number of simulated datasets. For each simulation, we draw observations independently and identically from the following model:

$$\begin{aligned} V_i &= B_0 W_i + \upsilon_i \\ X_i &= T_0 W_i + \epsilon_i \\ Z_i &= P_0 W_i + Q_0 X_i + \eta_i \\ Y_i &= X_i' \beta_0 + F_0 W_i + \chi_0 V_i + e_i \end{aligned}$$

The residuals υ_i, ϵ_i, η_i, and e_i are drawn independently of each other from zero-mean Gaussian distributions: $W_i \sim N(0, I)$, $\upsilon_i \sim N(0, \Sigma_V)$, $\epsilon_i \sim N(0, \Sigma_X)$, $\eta_i \sim N(0, \Sigma_Z)$, and $e_i \sim N(0, \Sigma_Y)$. Note that we do not include additional controls D_i in our simulations however in estimation we include an intercept (i.e, we treat D_i as a constant).

In each simulation we must choose parameters β_0, B_0, P_0, Q_0, T_0, F_0, χ_0, Σ_Y, Σ_V, Σ_X, and Σ_Z. Rather than use a fixed value of each parameter in all of our simulations, we draw the parameters at random in each simulation. Thus our simulation results show the weighted average performance of our estimators over a parameter space.

We draw the parameters as follows. The elements of the coefficient matrices β_0, L_0, T_0, F_0, and χ_0 are all independently mean-zero normal with variance equal to the square root of the number of columns of the matrix. For example, the elements of F_0 are all independent with distribution $N(0, 1/\sqrt{d_W})$. This choice of the variances of the normal distributions ensures that the ratio of the variance in each variable to the residual variance remains roughly constant as the dimension changes.

The matrices B_0 and P_0 are generated in order so that we obtain a pre-specified value of $\underline{\sigma}_n$. Let N_1 be a $d_V \times d_V$ matrix of independent standard normals and N_2 be a $d_W \times d_W$ matrix of independent standard normals. We draw B_0 as follows:

$$B_0 \sim \sqrt{\underline{\sigma}_n} (N_1' N_1)^{-1/2} N_1 (I, 0)' N_2' (N_2' N_2)^{-1/2}$$

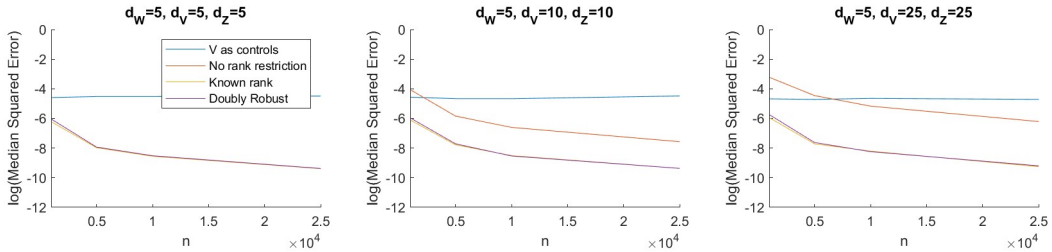
We draw P_0 independently of B_0 using a similar formula:

$$P_0 \sim \sqrt{\underline{\sigma}_n} (N_3' N_3)^{-1/2} N_3 (I, 0)' N_4' (N_4' N_4)^{-1/2} (I + T_0' T_0)$$

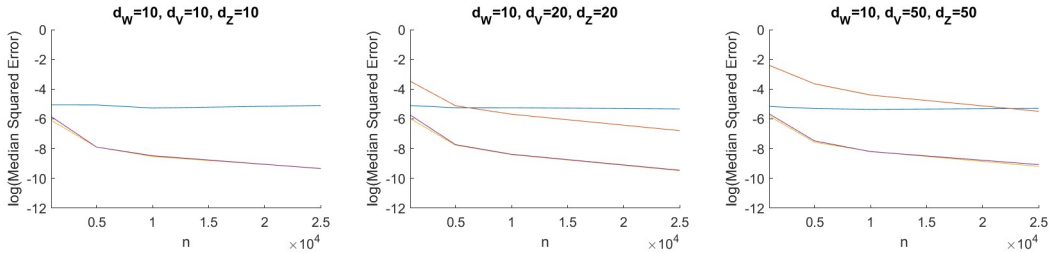
where N_3 is a $d_Z \times d_Z$ matrix of independent standard normals and N_4 is a $d_W \times d_W$ matrix of independent standard normals.

The covariance matrices have a re-scaled inverse Wishart distribution, for example $d_V p \Sigma_V^{-1} \sim \mathcal{W}_{d_V}(I, d_V p)$. The natural number p is a hyper-parameter that determines the degrees of freedom of the Wishart distribution.

We are left with hyperparameters $\underline{\sigma}_n$, p, d_W, d_X, d_V, d_Z, and the sample size n. In all of our simulations we let $d_X = 1$ so that there is a single treatment of interest. We set $p = 2$ which means the covariance matrices are concentrated around the identity. In all of our simulations $d_Z = d_V$ so there are the same number of proxies in Z_i as in V_i. We carry out simulations for a range of choices for the remaining hyperparameters $\underline{\sigma}_n$, d_W, d_V, and n.



(a) $d_W = 5$



(b) $d_W = 10$

Figure 4.1: Simulated Median Squared Errors, $\underline{\sigma}_n = 1$

Median Squared Errors on the y-axes are the medians of $\|\hat{\beta} - \beta_0\|^2$ over 1000 simulated datasets for various estimators $\hat{\beta}$. The different figures correspond to different choices for the number of confounding factors d_W, and the numbers of proxies d_V and d_Z.

Figure 4.1 shows the median-squared errors of alternative estimators for a variety of different hyperparameters. In all cases in Figure 4.1 we set $s = 1$. The estimators that are compared are: (in blue) a naive least-squares estimator that simply treats V_i as a set of controls, (in red) the proxy control estimator with no rank restriction, (in yellow) an infeasible estimator that imposes the rank restriction d_W, and (in purple) our doubly-robust estimator.[8]

For the implementation of the doubly-robust estimator we split the sample into five evenly-sized sub-samples. We select the penalty parameters from a dense grid of 100 values using cross-validation with five folds.

Keeping the number of confounders fixed but increasing the number of proxies leads to remarkably little loss in performance for the doubly robust estimator (in purple). The

[8] For the infeasible estimator we perform ordinary least-squares regression of $\tilde{Y}_i$ on $\tilde{X}_i$ and $\tilde{M}(\tilde{Z}_i', \tilde{X}_i')'$, where $\tilde{M}$ is a reduced-rank regression estimate of M_0 that imposes the rank d_W on the estimate.

doubly robust estimator achieves a performance that is near indistinguishable in all but the smallest samples from that of the infeasible estimator (in yellow). As we move from left to right in Figure 4.1 we see that the median squared error of this estimator stays roughly constant, with the only apparent exception occurring in the smallest sample size in the the rightmost sub-figures. Likewise, the performance varies little with the number of confounders d_W.

The proxy control estimator with no rank restriction (in red) is equivalent to the two-stage least squares strategy of Griliches (1977) in which V_i is a vector of endogenous regressors, X_i is a vector of exogenous regressors, and Z_i is a vector of instruments. As we discuss extensively in Appendix C along with formal results and further simulation evidence, this estimator is consistent (in a fixed-dimensional regime) under our identifying assumptions and some weak regularity conditions but has a non-standard asymptotic distribution when d_V is strictly larger than d_W. When the number of proxies in each group is equal to the number of confounders (the leftmost sub-figures) this estimator has nearly identical performance to our doubly robust procedure. This is to be expected as in these sub-figures there is no rank restriction for our estimator to exploit. However, unlike the doubly robust estimator (and the infeasible estimator), this procedure exhibits substantially worse performance as the number of proxies increases. This loss is apparent even in large samples.

The naive estimator (in blue) is inconsistent in this model, and this is clear from Figure 4.1 which shows that the median squared error of this estimator does not fall as the sample size grows. Nonetheless, in the setting with 10 confounders and 50 proxies in each group, the naive estimator outperforms the proxy estimator with no rank restrictions in all but the very largest samples. The doubly robust estimator has a lower median-squared error than the naive estimator in nearly all cases, the exceptions occurring in the leftmost sub-figures where the estimators have almost identical performance.

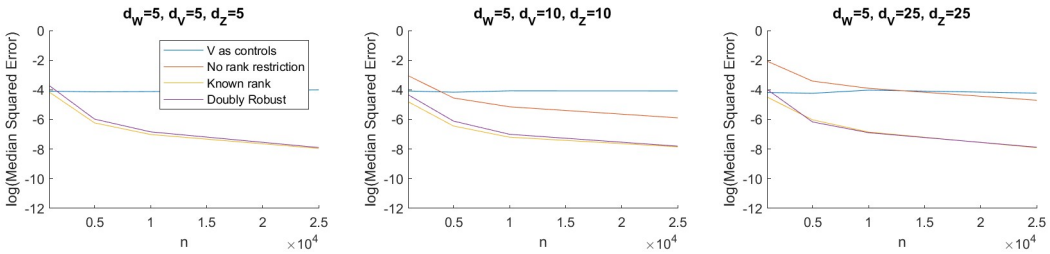
Figure 4.2 contains the same results for the case in which $\underline{\sigma}_n = 0.25$. Recall from Section 3 that $\underline{\sigma}_n$ controls the informativeness of the proxies relative to noise levels. A smaller value of $\underline{\sigma}_n$ is thus likely to be less favorable for our analysis. Indeed, all of the estimators perform worse in this setting (apart from the naive estimator which performs slightly better) and the proxy estimators perform markedly worse relative to the naive estimator. Nonetheless, our estimator still outperforms the naive estimator apart from in the smaller samples, and attains a level of performance that is close to that of the naive estimator, particularly with large sample sizes.

As in the case of $\underline{\sigma}_n = 1$, the estimator that does not impose a rank restriction performs substantially worse when there are many proxies compared to the number of unobserved confounders.

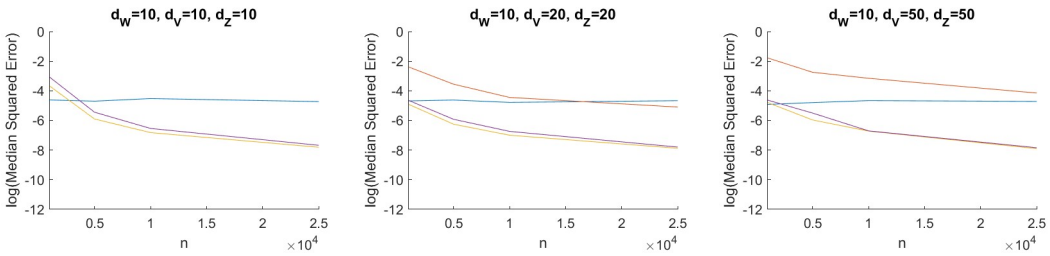
Figure 4.3 shows the percentage of simulations in which 99%, 95%, and 90% confidence intervals cover the true parameter β_0 (recall β_0 is drawn at random in each simulation). The confidence intervals are those based on a Gaussian approximation for the doubly-robust estimator as described in Section 3. In all cases the coverage is close to nominal level in large samples. In small samples the confidence intervals undercover, particularly when there are many proxies and many confounders.

Figure 4.4 shows the coverage is less favorable setting with $\underline{\sigma}_n = 0.25$ at least in small and moderate samples. There is very substantial undercoverage in small and medium sized samples, particularly when the number of proxies is large. Even in the largest samples the 90% interval undercovers by as close to 5% in some cases.

In Table 1 we give the proportion of simulations in which both the rank of $\hat{M}$ and $\hat{Q}$, are equal to the number of confounders d_W (which is the rank of the matrices M_0 and Q_0).



(a) $d_W = 5$



(b) $d_W = 10$

Figure 4.2: Simulated Median Squared Errors, $\underline{\sigma}_n = 0.25$

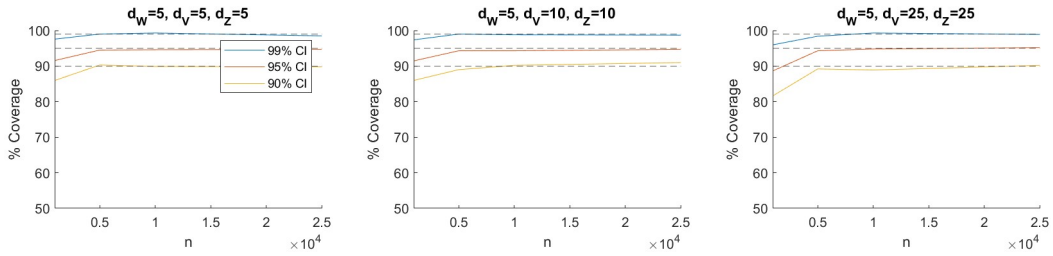
Median Squared Errors on the y-axes are the medians of $\|\hat{\beta} - \beta_0\|^2$ over 1000 simulated datasets for various estimators $\hat{\beta}$. The different figures correspond to different choices for the number of confounding factors d_W, and the numbers of proxies d_V and d_Z.

These estimates were evaluated using the whole sample rather than a sub-sample.

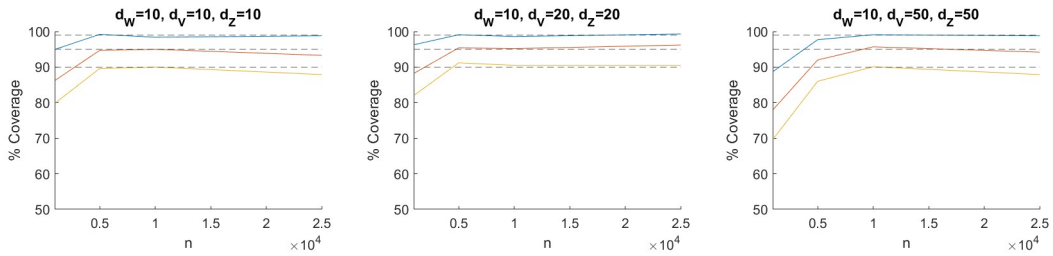
As we would expect, the probability of correct rank selection increases with the sample size. The frequency is lower when the measure of proxy informativeness $\underline{\sigma}_n$ is smaller. A small value of $\underline{\sigma}_n$ means that the smallest non-zero singular values of Q_0 and M_0 are close to zero and are therefore more difficult to distinguish from noise. This makes it harder to estimate the number of non-zero singular values, and thus the ranks of these matrices. In the case of a small $\underline{\sigma}_n$ and small sample size, a smaller number of proxies is associated with a greater probability of correct rank selection. In larger samples and with more informative proxies there is no clear trend.

5 Conclusion

We present novel identification results for the linear model with proxy controls. Our identification results suggest method of moments estimators that can take advantage of the dimension reduction when the number of unobserved confounding factors is smaller than the number of proxies. We present model selection methods that adapt to the unknown number of confounding factors. We provide conditions for uniform root-n consistency of our estimates and asymptotic validity of an inference procedure. Our simulation results suggest that our estimators are more effective than proxy control methods that do not exploit the dimension reduction, particularly when the the number of proxies substantially exceeds the number of unobserved confounders. In the latter case inference based on our doubly-robust adaptive proxy control method performs well.



(a) $d_W = 5$



(b) $d_W = 10$

Figure 4.3: Simulated Confidence Interval Coverage, $\underline{\sigma}_n = 1$

Confidence interval coverage of the treatment parameter. on the y-axes are percentages of 1000 simulated datasets in which confidence intervals contain β_0. The different figures correspond to different choices for the number of confounding factors d_W, and the numbers of proxies d_V and d_Z.

References

Bing, Xin, & Wegkamp, Marten H. 2019. Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. *The Annals of Statistics*, **47**.

Bunea, Florentina, She, Yiyuan, & Wegkamp, Marten H. 2011. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, **39**.

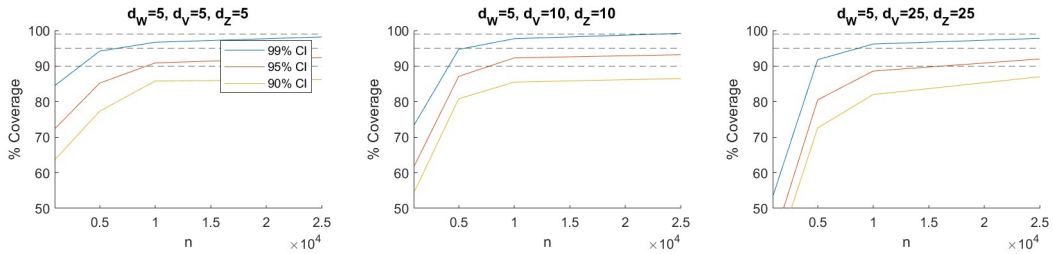
Chamberlain, Gary, & Griliches, Zvi. 1975. Unobservables with a Variance-Components Structure: Ability, Schooling, and the Economic Success of Brothers. *International Economic Review*, **16**, 422.

Chamberlain, Gary, & Griliches, Zvi. 1977. *Kinometrics: Determinants of Socioeconomic Success within and between Families.* Vol. 7. North-Holland. Chap. 4, pages 97–124.

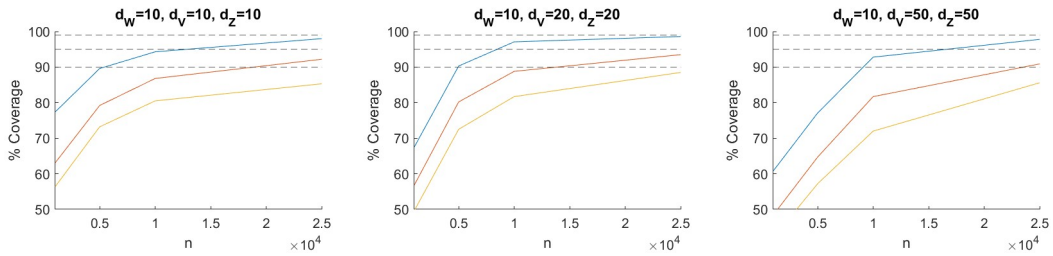
Chen, K., Dong, H., & Chan, K.-S. 2013. Reduced rank regression via adaptive nuclear norm penalization. *Biometrik*, **100**, 901–920.

Chernozhukov, Victor, Chetverikov, Denis, Demirer, Mert, Duflo, Esther, Hansen, Christian, Newey, Whitney, & Robins, James. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21**, C1–C68.

Cui, Yifan, Pu, Hongming, Shi, Xu, Miao, Wang, & Tchetgen, Eric Tchetgen. 2020 (Nov.). *Semiparametric proximal causal inference.*



(a) $d_W = 5$



(b) $d_W = 10$

Figure 4.4: Simulated Confidence Interval Coverage, $\underline{\sigma}_n = 0.25$

Confidence interval coverage of the treatment parameter. on the y-axes are percentages of 1000 simulated datasets in which confidence intervals contain β_0. The different figures correspond to different choices for the number of confounding factors d_W, and the numbers of proxies d_V and d_Z.

Deaner, Ben. 2019. Proxy Controls and Panel Data. Dec.

Deaner, Ben. 2021. Proxy Controls and Panel Data. Jan.

Griliches, Zvi. 1977. Estimating the Returns to Schooling: Some Econometric Problems. *Econometrica*, **45**, 1.

Griliches, Zvi, & Mason, William M. 1972. Education, Income, and Ability. *Journal of Political Economy*, **80**, S74–S103.

Hansen, Lars Peter. 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, **50**, 1029.

Horn, Roger A., & Johnson, Charles R. 1991. *Topics in Matrix Analysis*.

Imbens, Guido, Kallus, Nathan, & Mao, Xiaojie. 2021. Controlling for Unmeasured Confounding in Panel Data Using Minimal Bridge Functions: From Two-Way Fixed Effects to Factor Models. Aug. Ar.

Izenman, Alan Julian. 1975. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, **5**, 248–264.

Kallus, Nathan, Mao, Xiaojie, & Uehara, Masatoshi. 2021. Causal Inference Under Unmeasured Confounding With Negative Controls: A Minimax Learning Approach. Mar.

Miao, Wang, Shi, Xu, & Tchetgen, Eric Tchetgen. 2018a. A Confounding Bridge Approach for Double Negative Control Inference on Causal Effects. Aug.

Table 2: Frequency of Correct Rank Selection

			Sample Size			
$\underline{\sigma}_n$	d_W	$d_V = d_Z$	1000	5000	10000	25000
0.25	5	5	0.576	0.952	0.983	0.996
0.25	5	10	0.045	0.65	0.813	0.893
0.25	5	25	0	0.127	0.696	0.982
0.25	10	10	0.432	0.937	0.992	1
0.25	10	20	0	0.356	0.734	0.943
0.25	10	50	0	0	0.084	0.905
1	5	5	0.987	0.999	1	1
1	5	10	0.8	0.909	0.916	0.933
1	5	25	0.703	0.989	0.993	1
1	10	10	0.994	1	1	1
1	10	20	0.773	0.956	0.983	0.983
1	10	50	0.075	0.999	1	1

Figures are the proportion of the 1000 simulated datasets in which both the estimated rank of M_0 and the estimated rank of Q_0 are equal to the number of unobserved confounders d_W. Rows corresponds to different choices of d_W, d_V, and d_Z, columns correspond to different choices of the sample size n.

Miao, Wang, Geng, Zhi, & Tchetgen, Eric J. Tchetgen. 2018b. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, **105**, 987–993.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference (Second Edition)*. Cambridge University Press.

Reinsel, Gregory C., & Velu, Raja P. 1998. *Multivariate Reduced-Rank Regression*.

Singh, Rahul. 2020. Kernel Methods for Unobserved Confounding: Negative Controls, Proxies, and Instruments. Dec.

Stewart, G. W. 1977. On the Perturbation of Pseudo-Inverses, Projections and Linear Least Squares Problems. *SIAM Review*, **19**, 634–662.

Tchetgen, Eric J. Tchetgen, Ying, Andrew, Cui, Yifan, Shi, Xu, & Miao, Wang. 2020 (Sept.). *An Introduction to Proximal Causal Learning*. Appeared on Arxiv 23 Sep 2020.

Tibshirani, Robert. 1996. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society*, **58**, 267–288.

A Relaxing Assumption 1.1

In Theorems 1.1 and 1.2, the following weaker condition would suffice in lieu of Assumption 1.1:

$$Y = \beta_0' X + A_0 W + L_0 D + \varepsilon, \qquad E[\varepsilon(X', D')] = 0 \tag{19}$$

$$V = B_0 W + R_0 D + \upsilon, \qquad E[\upsilon D'] = 0 \tag{20}$$

Assumption 1.1*. i. (19) and (20) hold. ii. $E[\varepsilon Z'] = 0$ and $E[\upsilon(X', Z')] = 0$.

Assumption 1.1* weakens Assumption 1.1 by dropping the requirement (in 1.1.i) that $E[\varepsilon W'] = 0$ and $E[\upsilon W'] = 0$. Thus Assumption 1.1* is weaker than Assumption 1.1. However, Assumption 1.1.i can be understood as a definition rather than an 'assumption' in the usual sense of the word. In particular, Assumption 1.1.i defines the coefficients β_0, A_0, B_0, etc. and the residuals ε and υ. This is because (generically) there exist unique values of these coefficients and residuals that satisfy Assumption 1.1.i. By contrast, Assumption 1.1*.i does not uniquely determine these coefficients and residuals, leaving their interpretation ambiguous. Ambiguity in the definition of β_0 is problematic because identification of this quantity is the main goal of Section 1.

B Additional Algorithmic Details

The formulas for the solutions to the rank-restricted regression problems (11) and (12) are as follows. Let $\hat{\Sigma}_{\hat{Z}\hat{X}} = (\hat{Z}, \hat{X})'(\hat{Z}, \hat{X})/n$ and $\hat{\Sigma}_{\check{Z}} = \check{Z}'\check{Z}/n$ and define $\hat{E}_M$ and $\hat{E}_Q$ by:

$$\hat{E}_M = eigen\big(\hat{V}'(\hat{Z}, \hat{X})\hat{\Sigma}^{\dagger}_{\hat{Z}\hat{X}}(\hat{Z}, \hat{X})'\hat{V}\big)$$
$$\hat{E}_Q = eigen\big((\check{V}, \check{Y})'\check{Z}\hat{\Sigma}^{\dagger}_{\check{Z}}\check{Z}'(\check{V}, \check{Y})\big)$$

Then we have:

$$\hat{M}_r = \hat{\Sigma}^{\dagger}_{\hat{Z}\hat{X}}(\hat{Z}, \hat{X})'\hat{V}\hat{E}_{M,[:,1:r]}\hat{E}'_{M,[:,1:r]}$$
$$\hat{Q}_r = \hat{\Sigma}^{\dagger}_{\check{Z}}\check{Z}'(\check{V}, \check{Y})\hat{E}_{Q,[1:r,:]}\hat{E}'_{Q,[1:r,:]}$$

As stated in Section 2, the solutions to the rank-penalized least squares problems (13) and (14) are identical to the solutions to (11) and (12) but with the rank restriction r set in each equation to $\hat{r}_M$ and $\hat{r}_Q$ respectively. The formulas for $\hat{r}_M$ and $\hat{r}_Q$ are as follows.

$\hat{r}_M$ is whichever is smaller: d_Z or the number of eigenvalues of the matrix $\hat{V}'(\hat{Z}, \hat{X})\hat{\Sigma}^{\dagger}_{\hat{Z}\hat{X}}(\hat{Z}, \hat{X})'\hat{V}$ that exceed $\lambda_{M,n}$. Similarly $\hat{r}_Q$ is the minimum of d_V and the number of eigenvalues of $(\check{V}, \check{Y})'\check{Z}\hat{\Sigma}^{\dagger}_{\check{Z}}\check{Z}'(\check{V}, \check{Y})$ that exceed $\lambda_{Q,n}$.

In our setting the plug-in penalty parameters suggested by Bunea *et al.* (2011) are:

$$\lambda_{M,n} = C({d_V}^{1/2} + rank\big((\hat{Z}, \hat{X})\big)^{1/2})^2 s_M^2/n$$
$$\lambda_{Q,n} = C\big((d_V + 1)^{1/2} + rank(\check{Z})^{1/2}\big)^2 s_Q^2/n$$

Where C is a constant with $C > 1$. In our simulations we set $C = 1.1$. s_M^2 is an estimate of the residual variance from regression of $\hat{V}_i$ on $\hat{Z}_i$ and $\hat{X}_i$, s_Q^2 is an estimate of the residual variance from regression of $(\check{V}_i', \check{Y}_i)'$ on $\check{Z}_i$. The estimates are:

$$s_M^2 = \|\hat{Y} - (\hat{Z}, \hat{X})\big((\hat{Z}, \hat{X})'(\hat{Z}, \hat{X})\big)^{\dagger}(\hat{Z}, \hat{X})'\hat{Y}\|_F^2/\big(nd_V - rank\big((\hat{Z}, \hat{X})\big)d_V\big)$$
$$s_Q^2 = \|(\check{V}, \check{Y}) - \check{Z}(\check{Z}'\check{Z})^{\dagger}\check{Z}'(\check{V}, \check{Y})\|_F^2/\big(n(d_V + 1) - rank(\check{Z})(d_V + 1)\big)$$

C Consistency and non-Gaussianity of Two-Stage Least-Squares

Recall the moment condition (6) in Corollary 1.1 which identifies β_0:

$$E\big[(\tilde{Y}_i - \beta_0'\tilde{X}_i - \xi_0'\tilde{V}_i)(\tilde{Z}_i', \tilde{X}_i')'\big] = 0$$

We can rewrite this as:

$$E\big[(\tilde{Y}_i - \beta_0'\tilde{X}_i)(\tilde{Z}_i', \tilde{X}_i')'\big] - \chi_0 = 0$$

Where χ_0 is defined as follows:

$$\begin{aligned}
\chi_0 &= E[\xi_0\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')'] \\
&= A_0C_0(C_0B_0)^\dagger B_0G_0 \\
&= Q_{0,[d_V+1,:]}Q^\dagger_{0,[1:d_V,:]}E\big[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')'\big]
\end{aligned}$$

Where $Q_{0,[d_V+1,:]} = A_0C_0$ and $Q_{0,[1:d_V,:]} = C_0B_0$ are submatrices of $Q_0 = (B_0', A_0')'C_0$. To obtain an estimate of β_0 we effectively replace χ_0 with an estimate. An estimate $\hat{\chi}$ without sample splitting is given below:

$$\begin{aligned}
\hat{\chi} &= \frac{1}{n}\sum_{i=1}^n \hat{\xi}'\hat{V}_i(\hat{Z}_i', \hat{X}_i')' \\
&= \hat{Q}_{[d_V+1,:]}\hat{Q}^\dagger_{[1:d_V,:]}\frac{1}{n}\sum_{i=1}^n \hat{V}_i(\hat{Z}_i', \hat{X}_i')'
\end{aligned}$$

In order to establish consistency of $\hat{\chi}$ (and thus consistency of a corresponding estimate $\hat{\beta}$ of β_0), we would typically decompose:

$$\begin{aligned}
\hat{\chi} - \chi_0 = \xi_0'\big(&\frac{1}{n}\sum_{i=1}^n \hat{V}_i(\hat{Z}_i', \hat{X}_i')' - E[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')']\big) \\
&+ (\hat{\xi} - \xi_0)'E[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')'] \\
&+ (\hat{\xi} - \xi_0)'\big(\frac{1}{n}\sum_{i=1}^n \hat{V}_i(\hat{Z}_i', \hat{X}_i')' - E[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')']\big)
\end{aligned}$$

Under standard conditions, the terms on the first two rows of the RHS above are root-n asymptotically zero-mean normal and the term on the third row above is of second order. Thus $\hat{\chi} - \chi_0$ is root-n asymptotically zero-mean Gaussian.

However, a problem arises when the rank is inconsistently estimated. Let $\hat{d}_W = rank(\hat{Q}^\dagger_{[1:d_V,:]})$ and recall that $d_W = rank(Q_{0,[1:d_V,:]})$. Asymptoptic analysis of $\hat{\chi}$ is complicated by the following fact. If $\|\hat{Q} - Q_0\| \prec_p 1$ but $P\big(\hat{d}_W = d_W\big) \not\to^p 1$, then it is necessarily the case that $\|Q^\dagger_{0,[1:d_V,:]} - \hat{Q}^\dagger_{[1:d_V,:]}\|$ does not converge in probability to zero. This follows from a well-known result (see Stewart (1977)) that for any matrices A and B, if $rank(A) \neq rank(B)$ then:

$$\|A^\dagger - B^\dagger\| \geq \frac{1}{\|A - B\|}$$

Moreover, if $rank(A) > rank(B)$ then $\|A^\dagger\| \geq \frac{1}{\|A-B\|}$. This contrasts with the case of $rank(A) = rank(B)$ which implies that, for $\|A^\dagger\|\|A - B\| < 1$:

$$\|A^\dagger - B^\dagger\| \leq \frac{\|A - B\|}{1 - \|A^\dagger\|\|A - B\|}$$

If $\hat{Q}$ is a consistent least squares estimator of Q_0 and the variables are continuously distributed, then $\hat{Q}_{[1:d_V,:]}$ has rank d_V with probability 1, but $Q_{0[1:d_V,:]}$ has rank d_W. As such, if $d_W < d_V$ then both $\|\hat{Q}^\dagger_{[1:d_V,:]}\|$ and $\|Q^\dagger_{0,[1:d_V,:]} - \hat{Q}^\dagger_{[1:d_V,:]}\|$ must diverge.

An interesting consequence of the arguments above is that if $\hat{Q}$ is consistent for Q_0, and $\|(B_0C_0)^\dagger\| \precsim_p 1$, then the probability that $\hat{d}_W$ strictly underestimates d_W must vanish. Thus the problematic case is restricted exclusively to the overestimation of d_W. To see this recall that $\|(B_0C_0)^\dagger\| = \|Q^\dagger_{0[1:d_V,:]}\|$, so if $\hat{d}_W < d_W$ then $\|(B_0C_0)^\dagger\| \geq \frac{1}{\|\hat{Q}_{[1:d_V,:]} - Q_{0,[1:d_V,:]}\|}$, but by consistency of $\hat{Q}$ the RHS diverges, so unless the probability of $\hat{d}_W < d_W$ goes to zero we have $1 \prec_p \|(B_0C_0)^\dagger\|$, which yields a contradiction.

Fortunately, $\hat{\chi}$ may converge in probability to χ even when the rank is inconsistently estimated. Demonstrating this requires a rather more involved argument which we provide in the proof of Theorem A.3.1 below. In effect, the theorem demonstrates that, although $\hat{\xi}$ is generally inconsistent for ξ_0, the term $\hat{\xi} - \xi_0$ is stoachstically bounded. Stochastic boundedness of $\hat{\xi} - \xi_0$ ensures consistency of $\hat{\chi}$ (assuming ξ_0 has bounded norm and $\frac{1}{n}\sum_{i=1}^n \hat{V}_i(\hat{Z}_i', \hat{X}_i')'$ is consistent). However, if $\hat{\xi} - \xi_0$ does not converge to a constant, then $\hat{\chi}$ (and thus $\hat{\beta}$) is generally not root-n asymptotically normal.

In sum. If the rank is estimated inconsistently (as is necessarily the case for 2SLS estimation when $d_W < d_V$) then the resulting proxy control estimator may be consistent but it may not have an asymptotic normal distribution.

We demonstrate the non-normality of the asymptotic distribution of 2SLS using simulation evidence. The following figure contains a QQ plot of the simulated distribution of the 2SLS estimator centered and rescaled to have mean zero and unit variance. The data were drawn from the same model as in Section 4 but with all residuals standard normal, and all coefficients equal to unity except with a zero coefficients on X_i in the equation for Z_i and V_i in the equation for Y_i. The simulation has $d_X = 1$, $d_Z = d_V = 2$, and $d_W = 1$ and $n = 10000$.

The QQ-plot was generated using 1000 simulation draws and the top and bottom 1% outliers are trimmed. The figure shows that the distribution is highly non-Gaussian, with very heavy tails.

Despite the non-standard asymptotic distribution, our simulations in Section 4 clearly demonstrate consistency of the 2SLS estimator. Theorem A.3.1 provides a formal consistency result for estimates which, like 2SLS, do not correctly select the rank. The argument depends strongly on the Generalized $\sin\theta$ Theorem from Wedin (1972), which we adapt for our purposes in a Lemma stated at the end of this subsection.

Theorem A.3.1. *Suppose that* $\|C_0^\dagger\|, \|B_0^\dagger\| \precsim 1$ *and* $\|A_0\|, \|B_0\|, \|C_0\|, \|G_0\| \precsim 1$. *In addi-*

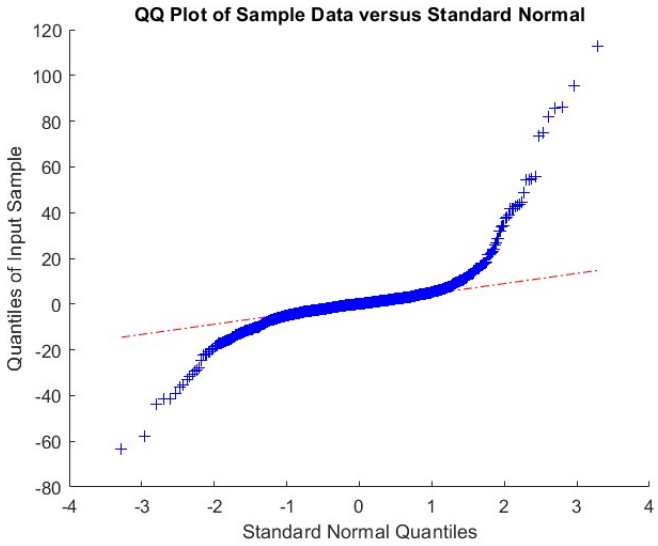


Figure C.1: QQ-plot of 2SLS estimates

tion for sequences $\delta_a, \delta_b, \delta_c \prec 1$ *suppose that:*

$$\|\hat{Q}_{[d_V+1,:]} - Q_{0,[d_V+1,:]}\| \precsim_p \delta_a$$
$$\|\hat{Q}_{[1:d_V,:]} - Q_{0,[1:d_V,:]}\| \precsim_p \delta_b$$
$$\|\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i(\hat{Z}_i', \hat{X}_i')' - E\big[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')'\big]\| \precsim_p \delta_c$$

If $\|\hat{Q}^{\dagger}_{[1:d_V,:]}\| \precsim_p \delta_b^{-1}$ *then:*

$$\|\hat{\chi} - \chi_0\| \precsim_p \delta_a + \delta_b + \delta_c + \frac{\delta_a \delta_c}{\delta_b}$$

Discussion

As we discuss above, if the rank is not consistently estimated then $\|\hat{Q}^{\dagger}_{[1:d_V,:]}\|$ diverges, and in fact it must diverge at least as quickly as $\|\hat{Q}_{[1:d_V,:]} - Q_{0,[1:d_V,:]}\|$. Thus the condition $\|\hat{Q}^{\dagger}_{[1:d_V,:]}\| \precsim_p \delta_b^{-1}$ states that it diverges no more quickly than this lower bound. In the 2SLS case in fixed dimensions $\delta_a, \delta_b, \delta_c \asymp_p n^{-1/2}$ under weak conditions. In this setting we would expect the singular values of the linear regression estimate $\hat{Q}_{[1:d_V,:]}$ to concentrate around the those of $Q_{0,[1:d_V,:]}$ at exactly rate $\sqrt{n}$ in thich case $\|\hat{Q}^{\dagger}_{[1:d_V,:]}\| \precsim_p \sqrt{n}$ the condition of the theorem holds.

Proof. Define ω_0, ω_1, ω_2, and ω_3 as follows:

$$\begin{aligned}
\omega_0 &= Q^{\dagger}_{0,[1:d_V,:]} - \hat{Q}^{\dagger}_{[1:d_V,:]} \\
\omega_1 &= A_0 C_0 \big(Q^{\dagger}_{0,[1:d_V,:]} - \hat{Q}^{\dagger}_{[1:d_V,:]}\big) \\
\omega_2 &= \big(Q^{\dagger}_{0,[1:d_V,:]} - \hat{Q}^{\dagger}_{[1:d_V,:]}\big) B_0 G_0 \\
\omega_3 &= A_0 C_0 \big(Q^{\dagger}_{0,[1:d_V,:]} - \hat{Q}^{\dagger}_{[1:d_V,:]}\big) B_0 G_0
\end{aligned}$$

We can decompose $\hat{\chi} - \chi_0$ as follows:

$$\begin{aligned}
&\hat{\chi} - \chi_0 \\
=&(\hat{Q}_{[d_V+1,:]} - Q_{0,[d_V+1,:]})(C_0 B_0)^{\dagger} E\big[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')'\big] \\
&+\omega_3 + (\hat{Q}_{[d_V+1,:]} - Q_{0,[d_V+1,:]})\omega_2 \\
&+Q_{0,[d_V+1,:]}(C_0 B_0)^{\dagger}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i(\hat{Z}_i', \hat{X}_i')' - E\big[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')'\big]\right) \\
&+\omega_1\left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i(\hat{Z}_i', \hat{X}_i')' - E\big[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')'\big]\right) \\
&+(\hat{Q}_{[d_V+1,:]} - Q_{0,[d_V+1,:]})(C_0 B_0)^{\dagger}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i(\hat{Z}_i', \hat{X}_i')' - E\big[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')'\big]\right) \\
&+(\hat{Q}_{[d_V+1,:]} - Q_{0,[d_V+1,:]})\omega_0\left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i(\hat{Z}_i', \hat{X}_i')' - E\big[\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')'\big]\right)
\end{aligned}$$

By a well-known property of the pseudo-inverse $\|(C_0 B_0)^{\dagger}\| \leq \|C_0^{\dagger}\|\|B_0^{\dagger}\| \precsim 1$. Applying the triangle inequality, definition of the matrix norm, and the rate conditions in the statement of the theorem, we get:

$$\begin{aligned}
\|\hat{\chi} - \chi_0\| \precsim_p &\|\omega_3\| + \delta_a(1 + \|\omega_2\|) + \delta_c(1 + \|\omega_1\|) \\
&+\delta_a\delta_c(1 + \|\omega_0\|)
\end{aligned}$$

It remains to bound $\|\omega_0\|$, $\|\omega_1\|$, $\|\omega_2\|$, and $\|\omega_3\|$. The first of these is straight forward, recall $\|\hat{Q}^{\dagger}_{[1:d_V,:]}\| \asymp_p \delta_b^{-1}$ and note that $\|Q^{\dagger}_{0,[1:d_V,:]}\| = \underline{\sigma}^{-1} \precsim 1$. So by the triangle inequality:

$$\|\omega_0\| \leq \|Q^{\dagger}_{0,[1:d_V,:]}\| + \|\hat{Q}^{\dagger}_{[1:d_V,:]}\| \precsim_p \frac{1}{\|Q_{0,[1:d_V,:]} - \hat{Q}_{[1:d_V,:]}\|}$$

We can rewrite ω_1 as:

$$\begin{aligned}
\omega_1 &= A_0 C_0 \big(Q^{\dagger}_{0,[1:d_V,:]} - \hat{Q}^{\dagger}_{[1:d_V,:]}\big) \\
&= A_0 B_0^{\dagger} Q_{0,[1:d_V,:]} \big(Q^{\dagger}_{0,[1:d_V,:]} - \hat{Q}^{\dagger}_{[1:d_V,:]}\big)
\end{aligned}$$

And so:

$$\|\omega_1\| \leq \|A_0\|\|B_0^{\dagger}\|\|Q_{0,[1:d_V,:]}\big(Q^{\dagger}_{0,[1:d_V,:]} - \hat{Q}^{\dagger}_{[1:d_V,:]}\big)\|$$

By similar reasoning

$$\|\omega_2\| \leq \|C_0^\dagger\|\|G_0\|\|Q_{0,[1:d_V,:]}(Q^\dagger_{0,[1:d_V,:]} - \hat{Q}^\dagger_{[1:d_V,:]})\|$$

And again by similar deductions:

$$\|\omega_3\| \leq \|A_0\|\|B_0^\dagger\|\|C_0^\dagger\|\|G_0\|\|Q_{0,[1:d_V,:]}(Q^\dagger_{0,[1:d_V,:]} - \hat{Q}^\dagger_{[1:d_V,:]})Q'_{0,[1:d_V,:]}\|$$

In order to bound the term $\|Q_{0,[1:d_V,:]}(Q^\dagger_{0,[1:d_V,:]} - \hat{Q}^\dagger_{[1:d_V,:]})\|$ we apply Lemma A.3.1 which is a slight adaptation of a theorem from Wedin (1972). We get that if $\hat{d}_W \geq d_W$ then:

$$\begin{aligned}\|Q_{0,[1:d_V,:]}(Q^\dagger_{0,[1:d_V,:]} - \hat{Q}^\dagger_{[1:d_V,:]})\| \leq \|Q_{0,[1:d_V,:]}\| &\frac{2\|Q_{0,[1:d_V,:]} - \hat{Q}_{[1:d_V,:]}\|}{\|(B_0C_0)^\dagger\|^{-1} - 2\|Q_{0,[1:d_V,:]} - \hat{Q}_{[1:d_V,:]}\|} \\ &+ \frac{1}{2}\|(B_0C_0)^\dagger\|\|\hat{Q}^\dagger_{[1:d_V,:]}\|\|Q_{0,[1:d_V,:]}\|\|Q_{0,[1:d_V,:]} - \hat{Q}_{[1:d_V,:]}\| \\ &\asymp_p 1\end{aligned}$$

Also using Lemma A.3.1 we get that for $\hat{d}_W \geq d_W$:

$$\begin{aligned}&\|Q_{0,[1:d_V,:]}(Q^\dagger_{0,[1:d_V,:]} - \hat{Q}^\dagger_{[1:d_V,:]})Q'_{0,[1:d_V,:]}\| \\ \leq &\|Q_{0,[1:d_V,:]}\|^2\frac{2\|Q_{0,[1:d_V,:]} - \hat{Q}_{[1:d_V,:]}\|}{\|(B_0C_0)^\dagger\| - 2\|Q_{0,[1:d_V,:]} - \hat{Q}_{[1:d_V,:]}\|} \\ +&\frac{1}{4}\|(B_0C_0)^\dagger\|^2\|\hat{Q}^\dagger_{[1:d_V,:]}\|\|Q_{0,[1:d_V,:]}\|^2\|Q_{0,[1:d_V,:]} - \hat{Q}_{[1:d_V,:]}\|^2 \\ \asymp &\|Q_{0,[1:d_V,:]} - \hat{Q}_{[1:d_V,:]}\| \\ \lesssim_p &\delta_b\end{aligned}$$

Recall that $\|(B_0C_0)^\dagger\| \lesssim_p 1$ and consistency of $\hat{Q}_{[1:d_V,:]}$ imply that $P(\hat{d}_W \geq d_W) \to 1$ and so the bounds above hold with probability approaching 1. In all we get:

$$\|\hat{\chi} - \chi_0\| \lesssim_p \delta_a + \delta_b + \delta_c + \frac{\delta_a\delta_c}{\delta_b}$$

□

Lemma A.3.1 (Wedin). *Consider $k \times l$ matrices A and B with $r = rank(A) < rank(B) = p$. Let σ_r denote the r^{th} smallest non-zero singular value of A. Suppose $\sigma_r > 0$ and $2\|B - A\| < \sigma_r$, then:*

$$\|A(B^\dagger - A^\dagger)\| \leq \|A\|\frac{2\|B - A\|}{\|A^\dagger\|^{-1} - 2\|B - A\|} + \frac{1}{2}\|A^\dagger\|\|B^\dagger\|\|A\|\|A - B\|$$

Where $\tilde{\sigma}_p$ is the smallest non-zero singular value of B. Moreover:

$$\|A(B^\dagger - A^\dagger)A'\| \leq \|A\|^2\frac{2\|B - A\|}{\|A^\dagger\|^{-1} - 2\|B - A\|} + \frac{1}{4}\|A^\dagger\|^2\|B^\dagger\|\|A\|^2\|A - B\|^2$$

Proof. Let the singular values of A ordered in terms of size be $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{\min\{k,l\}-1} \geq \sigma_{\min\{k,l\}}$ and the singular values of B ordered in terms of size be $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq ... \geq \tilde{\sigma}_{\min\{k,l\}-1} \geq \tilde{\sigma}_{\min\{k,l\}}$. Since B has rank r, $\sigma_s = 0$ for all $r < s \leq \min\{l,k\}$. We decompose B into two matrices $B_1 = \tilde{U}_1\tilde{\Sigma}_1\tilde{V}_1'$ and $B_0 = \tilde{U}_0\tilde{\Sigma}_0\tilde{V}_0'$ using the compact singular value decomposition of B:

$$\begin{aligned} B &= (\tilde{U}_1, \tilde{U}_0) \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & \tilde{\Sigma}_0 \end{pmatrix} (\tilde{V}_1, \tilde{V}_0)' \\ &= \tilde{U}_1\tilde{\Sigma}_1\tilde{V}_1' + \tilde{U}_0\tilde{\Sigma}_0\tilde{V}_0' \\ &= B_1 + B_0 \end{aligned}$$

In the above, $\tilde{\Sigma}_1$ is the $r \times r$ diagonal matrix whose diagonal entires are the r largest signular values of B and $\tilde{\Sigma}_2$ is the diagonal marix whose entires are the reaming singular values of B. The block matrix $(\tilde{U}_1, \tilde{U}_0)$ is the orthonormal matrix of left singular vectors of B with $\tilde{U}_1$ of dimension $k \times r$ and $\tilde{U}_0$ of dimension $k \times (p-r)$. The block matrix $(\tilde{V}_1, \tilde{V}_0)$ is the orthonormal matrix of right singular vectors of B with $\tilde{V}_1$ of dimension $l \times r$ and $\tilde{V}_0$ of dimension $l \times (p-r)$. Applying a similar decomposition for A would simply yield $A_1 = A$ and $A_0 = 0$.

Note that $rank(B_1) = rank(A) = r$. The pseudoinverse of B is then:

$$\begin{aligned} B^\dagger &= (\tilde{V}_1, \tilde{V}_0) \begin{pmatrix} \tilde{\Sigma}_1^\dagger & 0 \\ 0 & \tilde{\Sigma}_0^\dagger \end{pmatrix} (\tilde{U}_1, \tilde{U}_0)' \\ &= \tilde{V}_1\tilde{\Sigma}_1^\dagger\tilde{U}_1' + \tilde{V}_0\tilde{\Sigma}_0^\dagger\tilde{U}_0' \\ &= B_1^\dagger + B_0^\dagger \end{aligned}$$

Note that $\|A^\dagger\| = \sigma_r^{-1}$ and $\|B_0^\dagger\| = \tilde{\sigma}_p^{-1}$. It is helpful to note the well-known result that for each $1 \leq s \leq \min\{l,k\}$, $|\sigma_s - \tilde{\sigma}_s| \leq \|B - A\|$.

Bounding $\|A(B^\dagger - A^\dagger)\|$

Applying the triangle inequality and the definition of the matrix norm we get:

$$\begin{aligned} &\|A(B^\dagger - A^\dagger)\| \\ \leq &\|A\|\|B_1^\dagger - A^\dagger\| + \|AB_0^\dagger\| \end{aligned} \tag{21}$$

Using the triangle inequality and the fact (already established) that $\tilde{\sigma}_s \leq \|B - A\|$ for $s > r$, we get:

$$\begin{aligned} \|A - B_1\| &\leq \|B - A\| + \|B_0\| \\ &= \|B - A\| + \tilde{\sigma}_{r+1} \\ &\leq 2\|B - A\| \end{aligned}$$

So using $2\|A^\dagger\|\|B-A\| = 2\sigma_r^{-1}\|B-A\| < 1$ we get:

$$\begin{aligned}&\|B_1^\dagger - A^\dagger\| \\ \leq &\frac{\|A^\dagger\|\|A-B_1\|}{1-\|A^\dagger\|\|B-B_1\|} \\ \leq &\frac{2\sigma_r^{-1}\|B-A\|}{1-2\sigma_r^{-1}\|B-A\|}\end{aligned}$$

Now consider the term $\|AB_0^\dagger\|$. Let $P_{\mathcal{R}(B_1')}$ be the orthogonal projection onto the range of B_1' and similarly for $P_{\mathcal{R}(A')}$. By the construction of B_0 and B_1, $\mathcal{R}(B_1')$ is a subspace the kernel of B_0. By elementary properties of the pseudoinverse, the kernel of $(B_0')^\dagger$ is identical to the kernel of B_0, thus $\mathcal{R}(B_1')$ is a subspace of the kernel of $(B_0')^\dagger$ and so $(B_0')^\dagger = (B_0')^\dagger(I - P_{\mathcal{R}(B_1')})$. Moreover by definition of the projection $P_{\mathcal{R}(A')}A' = A'$. Thus we have:

$$(B_0')^\dagger A' = (B_0')^\dagger(I - P_{\mathcal{R}(B_1')})P_{\mathcal{R}(A')}A'$$

And so applying the triangle inequality and definition of the matrix norm:

$$\begin{aligned}\|AB_0^\dagger\| &\leq \|A\|\|(I - P_{\mathcal{R}(B_1')})P_{\mathcal{R}(A_1')}\|\|B_0^\dagger\| \\ &= \tilde{\sigma}_p^{-1}\|A\|\|(I - P_{\mathcal{R}(B_1')})P_{\mathcal{R}(A')}\|\end{aligned}$$

In order to bound the quantity $\|(I - P_{\mathcal{R}(B_1')})P_{\mathcal{R}(A')}\|$ we apply the 'Generalized $\sin\theta$ Theorem' from Wedin (1972). In our setting this theorem states that if $\tilde{\sigma}_r \geq \delta + \alpha$ and $\sigma_{\min\{k,l\}} \leq \alpha$ for some $\alpha \geq 0$ and $\delta > 0$ then:[9]

$$\|(I - P_{\mathcal{R}(B_1')})P_{\mathcal{R}(A')}\| \leq \frac{\|A-B\|}{\delta}$$

In our setting $\sigma_{\min\{k,l\}} = 0$ by $|\tilde{\sigma}_r - \sigma_r| \leq \|B-A\|$ and by assumption $2\|B-A\| < \sigma_r$, and so $\tilde{\sigma}_r > \frac{1}{2}\sigma_r$ so letting $\alpha = 0$ and $\delta = \frac{1}{2}\sigma_r$ and combining we get:

$$\|AB_0^\dagger\| \leq \frac{1}{2}\sigma_r^{-1}\tilde{\sigma}_p^{-1}\|A\|\|A-B\|$$

In all:

$$\|A(B^\dagger - A^\dagger)\| \leq \|A\|\frac{2\|B-A\|}{\sigma_r - 2\|B-A\|} + \frac{1}{2}\sigma_r^{-1}\tilde{\sigma}_p^{-1}\|A\|\|A-B\|$$

Bounding $\|A(B^\dagger - A^\dagger)A'\|$

Applying a similar decomposition to the above we get:

$$\begin{aligned}&\|A(B^\dagger - A^\dagger)A'\| \\ \leq &\|A\|^2\|B_1^\dagger - A^\dagger\| + \|AB_0^\dagger A'\| \qquad (22)\end{aligned}$$

[9] Strictly speaking the bound in Wedin (1972) replaces $\|A-B\|$ with a weakly smaller quantity. In Wedin the matrix A_1 appears in place of A, but recall that in our case $A_1 = A$.

We already attained a bound on $\|B_1^\dagger - A^\dagger\|$. it remains to bound $\|AB_0^\dagger A'\|$. Using our previous reasoning:

$$A(B_0')^\dagger A' = A\big((I - P_{\mathcal{R}(B_1')})P_{\mathcal{R}(A')}\big)'(B_0')^\dagger(I - P_{\mathcal{R}(B_1')})P_{\mathcal{R}(A')}A'$$

And so applying the triangle inequality and definitiopn of the matrix norm:

$$\|AB_0^\dagger A'\| \leq \|A\|^2\|(I - P_{\mathcal{R}(B_1')})P_{\mathcal{R}(A_1')}\|^2\|B_0^\dagger\|$$

We then use the bound on $\|(I - P_{\mathcal{R}(B_1')})P_{\mathcal{R}(A_1')}\|$ which we attained earlier. □

D An alternative estimate of ξ_0

In Section 2 we develop an estimator of ξ_0 that is a function of the reduced-rank estimator of Q_0. However, one can also estimate ξ_0 directly. In this case, instead of using the rank restrictions on M_0 and Q_0 for dimension reduction we instead take advantage of the sparsity result in Corollary 1.

To motivate the estimator, note that the moment conditions in Corollary 1 imply the following condition.

$$E\big[\bar{Z}_i(\bar{Y}_i - \bar{V}_i\xi_0')\big] = 0$$

Substituting 7 and multiplying by $M_{0,[:,1:d_Z]}$ we get:

$$E\big[M_{0,[:,1:d_Z]}\bar{Z}_i(\bar{Y}_i - \xi_0' M_{0,[:,1:d_Z]}\bar{Z}_i)\big] = 0$$

Recall that Corollary 1 states there is a solution ξ_0 to the moment conditions which has at most d_W non-zero entries. To estimate ξ_0, we perform ℓ_1 penalized two-stage least-squares. The ℓ_1 penalization induces sparsity in the estimate of ξ_0. In particular, the alternative estimate of ξ_0 is the vector ξ that minimizes the empirical objective below:

$$\|\check{Y} - \check{Z}\tilde{M}'\xi\|_F^2 + \lambda_{\xi,n}\|\xi\|_1 \tag{23}$$

Where $\|\cdot\|_1$ is the ℓ_1 norm and $\lambda_{\xi,n}$ is a penalty parameter. $\tilde{M}$ is the matrix of regression estimates from multiple linear regression of $\bar{V}_i$ on $\bar{Z}_i$ without any rank restrictions or penalization.

Minimization of (23) is an ℓ_1-penalized least squares problem and can be solved using any standard Lasso algorithm. A number of methods are available for selecting the penalty parameter in Lasso regression. For example, $\lambda_{\xi,n}$ could be chosen using cross-validation.

Note that in contrast to the methods in Section 2, which effectively estimate the minimum Euclidean norm solution to $\xi_0 B_0 C_0 = A_0 C_0$, the method above estimates the solution with smallest ℓ_1 norm.

E Proofs

Proof Theorem 1.1. By Assumption 1.1.ii $E[\varepsilon(Z', X')] = 0$ and $E[\varepsilon D'] = 0$ and so $E[\varepsilon(\tilde{Z}', \tilde{X}')] = 0$ and by the same reasoning $E[\upsilon(\tilde{Z}', \tilde{X}')] = 0$

Partialling out D from both sides of (1) and (2) and using that $E[\varepsilon D'] = 0$ and $E[\upsilon D'] = 0$ we get:

$$\tilde{V} = B_0\tilde{W} + u$$
$$\tilde{Y} = \beta_0'\tilde{X} + A_0\tilde{W} + \epsilon$$

And so:

$$E\big[(\tilde{V} - B_0\tilde{W})(\tilde{Z}', \tilde{X}')\big] = 0 \tag{24}$$
$$E\big[(\tilde{Y} - \beta_0'\tilde{X} - A_0\tilde{W})(\tilde{Z}', \tilde{X}')\big] = 0 \tag{25}$$

Recall the definition of C_0 and G_0:

$$(C_0, G_0) = E\big[\tilde{W}(Z', X')\big]E\big[(Z', X')'(Z', X')\big]^\dagger$$

The rows of $E\big[\tilde{W}(Z', X')\big]$ must all be in the row space of $E\big[(Z', X')'(Z', X')\big]$, and so by elementary properties of the pseudo-inverse:

$$(C_0, G_0)E\big[(Z', X')'(Z', X')\big] = E\big[\tilde{W}(Z', X')\big]$$

Using the above to substitute out $E\big[\tilde{W}(Z', X')\big]$ from (24) and (25) we get:

$$E\big[(\tilde{V} - B_0(C_0, G_0)(\tilde{Z}', \tilde{X}')')(\tilde{Z}', \tilde{X}')\big] = 0$$
$$E\big[\big(\tilde{Y} - \beta_0'\tilde{X} - A_0(C_0, G_0)(\tilde{Z}', \tilde{X}')'\big)(\tilde{Z}', \tilde{X}')\big] = 0$$

Stacking the condition above into a block matrix we get the result.

□

Proof Theorem 1.2. **Step 1: Prove the rank conditions on the nuisance parameters**

By Assumptions 1.2 and 1.3, B_0 has full column rank and C_0 has rank d_W. Thus the product B_0C_0 has rank d_W. Moreover, since C_0 rank d_W, (C_0, G_0) must have row rank of at least d_W and so (using that B_0 has full column rank of d_W) $B_0(C_0, G_0)$ has rank d_W. Since $(B_0', A_0')'$ has column rank of at least d_W, and C_0 has full row rank of d_W, $(B_0', A_0')'C_0$ has rank d_W.

Step 2:

From Theorem 1.1 we have:

$$E\left[\left(\begin{pmatrix}\tilde{V}\\ \tilde{Y} - \beta_0\tilde{X}\end{pmatrix} - \begin{pmatrix}B_0C_0 & B_0G_0\\ A_0C_0 & A_0G_0\end{pmatrix}\begin{pmatrix}\tilde{Z}\\ \tilde{X}\end{pmatrix}\right)(\tilde{Z}', \tilde{X}')\right] = 0$$

Suppose the following holds for B with full column rank:

$$E\left[\left(\begin{pmatrix}\tilde{V}\\ \tilde{Y} - \beta\tilde{X}\end{pmatrix} - \begin{pmatrix}BC & BG\\ AC & AG\end{pmatrix}\begin{pmatrix}\tilde{Z}\\ \tilde{X}\end{pmatrix}\right)(\tilde{Z}', \tilde{X}')\right] = 0$$

Under Assumption 1.4, $E\big[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')\big]$ is non-singular, and so we get the following four equalities:

$$B_0C_0 = BC \tag{26}$$
$$B_0G_0 = BG \tag{27}$$
$$A_0C_0 = AC \tag{28}$$
$$A_0G_0 - \beta_0' = AG - \beta' \tag{29}$$

It follows immediately from the above and the rank restrictions on B_0C_0, $B_0(C_0, G_0)$, and $(B_0', A_0')'C_0$ that BC, $B(C, G)$, and $(B', A')'C$ each have rank d_W.

Recall that C_0 has full row rank and thus C_0C_0' is non-singular. Define $M = C_0'(C_0C_0')^{-1}G_0$. Post-multiplying both sides of (28) by M we get $A_0G_0 = ACM$ and substituting this into (29) gives:

$$ACM - \beta_0' = AG - \beta' \tag{30}$$

Now, post-multiplying both sides of (26) by M we get $B_0G_0 = BCM$. Substituting into (27) $BG = BCM$. Premultiplying both sides by $A(B'B)^{-1}B'$ (recall B has full column rank and so $B'B$ is non-singular) we get $AG = ACM$. Substituting into (30) we get $\beta = \beta_0$, as required.

□

Lemma 1. *Under Assumption 1.4, there exist matrices A, B, C, and G with so that B has full column rank and β, A, B, C, and G satisfy (5) if and only if there exists a vector ξ so that:*

$$E\big[\big(\tilde{Y} - \beta'\tilde{X} - \xi'\tilde{V}\big)(\tilde{Z}', \tilde{X}')\big] = 0 \tag{31}$$

Proof of Lemma 1. First let us prove the 'only if'. B has full column rank and so $B'B$ is non-singular, so letting $\xi' = A(B'B)^{-1}B'$ we have $AC = \xi BC$ and $AG = \xi'BG$. Substituting into (5) we get:

$$E\left[\left(\begin{pmatrix}\tilde{V}\\ \tilde{Y} - \beta'\tilde{X}\end{pmatrix} - \begin{pmatrix}B(C,G)\\ \xi'B(C,G)\end{pmatrix}\begin{pmatrix}\tilde{Z}\\ \tilde{X}\end{pmatrix}\right)(\tilde{Z}', \tilde{X}')\right] = 0$$

Using the first d_V rows of the matrix equation above to substitute out $E[B(C, G)(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')]$ from the remaining rows we get (31).

Now the 'if'. By Assumption 1.4 $E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')]$ is non-singular and so there must exist a matrix M so that:

$$E\big[\big(\tilde{V} - M(\tilde{Z}', \tilde{X}')'\big)(\tilde{Z}', \tilde{X}')\big] = 0 \tag{32}$$

Using this to substitute out $E\big[\tilde{V}(\tilde{Z}', \tilde{X}')\big]$ from (31) we get:

$$E\big[\big(\tilde{Y} - \beta'\tilde{X} - \xi'M(\tilde{Z}', \tilde{X}')'\big)(\tilde{Z}', \tilde{X}')\big] = 0 \tag{33}$$

Any matrix M of can be written as the product $M = M_1M_2$ where M_1 has full column rank. So let $B = M_1$, $(C, G) = M_2$, and $A = \xi'B$ substituting this into (32) and (33) and stacking the resulting moment conditions we get (5). □

Proof of Corollary 1. Theorem 1.2 and Lemma 1 together show that (6) identifies β_0.

By Theorem 1.1 (7) is satisfied by $M_0 = B_0(C_0, G_0)$. By Assumption 1.4 $E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')]$ is non-singular and so this M_0 is the unique solution. By Theorem 1.2 we then have $rank(M_0) = d_W$.

Next we show that ξ_0 satisfies (6) if and only if $\xi_0' B_0 C_0 = A_0 C_0$. From Theorem 1 we get:

$$E[(\tilde{Y} - \beta_0' \tilde{X})(\tilde{Z}', \tilde{X}')] = A_0(C_0, G_0) E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')]$$

From (6) and (7) we have:

$$E[(\tilde{Y} - \beta_0' \tilde{X})(\tilde{Z}', \tilde{X}')] = \xi_0' M_0 E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')]$$

Substituting $M_0 = B_0(C_0, G_0)$ and combining we get:

$$\xi_0' B_0(C_0, G_0) E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')] = A_0(C_0, G_0) E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')]$$

Again, using that $E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')]$ is non-singular we get $\xi_0' B_0(C_0, G_0) = A_0(C_0, G_0)$ and thus $\xi_0 B_0 C_0 = A_0 C_0$ which proves the 'only if'. For the 'if', Theorem 1.2 states that C_0 has full row rank and so $\xi_0' B_0 C_0 = A_0 C_0$ implies $\xi_0' B_0 = A_0$ and thus $\xi_0' B_0(C_0, G_0) = A_0(C_0, G_0)$. Using $M_0 = B_0(C_0, G_0)$ we get:

$$\xi_0 M_0 E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')] = A_0(C_0, G_0) E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')]$$

Using (7) to substitute out $M_0 E[(\tilde{Z}', \tilde{X}')(\tilde{Z}', \tilde{X}')']$ we get:

$$\xi_0 E[\tilde{V}(\tilde{Z}', \tilde{X}')] = A_0(C_0, G_0) E[(\tilde{Z}', \tilde{X}')'(\tilde{Z}', \tilde{X}')]$$

Substituting into the moment condition Theorem 1.1 gives (6).

Next we will show that there exists a solution ξ_0 with $\|\xi_0\| \leq d_W$. By Theorem 1.2, $rank(B_0 C_0) = d_W$. Since $B_0 C_0$ has rank d_W, for any vector ξ, there is a ξ_0 with at most d_W non-zero entries so that $\xi_0' B_0 C_0 = \xi' B_0 C_0$. Since there exists at least one ξ so that $\xi' B_0 C_0 = A_0 C_0$ it follows that there is at least one ξ_0 with at most d_W non-zero entries and $\xi_0' B_0 C_0 = A_0 C_0$.

Finally we show that $Q_0 = (A_0', B_0')' C_0$ (which is of rank d_W by Theorem 1.2) is identified from (8). First note that $\bar{Z}$ is a linear combination of $\tilde{Z}$ and $\tilde{X}$ and so (3) implies:

$$E\left[\left(\begin{pmatrix} \tilde{V} \\ \tilde{Y} - \beta_0 \tilde{X} \end{pmatrix} - \begin{pmatrix} B_0 C_0 & B_0 G_0 \\ A_0 C_0 & A_0 G_0 \end{pmatrix} \begin{pmatrix} \tilde{Z} \\ \tilde{X} \end{pmatrix}\right) \bar{Z}'\right] = 0$$

By the properties of partialling out, $E[\tilde{Z}\bar{Z}'] = E[\bar{Z}\bar{Z}']$, $E[\tilde{V}\bar{Z}'] = E[\bar{V}\bar{Z}']$, etc. and $\bar{X} = 0$, and so the above is equivalent to the following:

$$E\left[\left(\begin{pmatrix} \bar{V} \\ \bar{Y} \end{pmatrix} - \begin{pmatrix} B_0 C_0 & B_0 G_0 \\ A_0 C_0 & A_0 G_0 \end{pmatrix} \begin{pmatrix} \bar{Z} \\ 0 \end{pmatrix}\right) \bar{Z}'\right] = 0$$

Multiplying out $\begin{pmatrix} B_0 C_0 & B_0 G_0 \\ A_0 C_0 & A_0 G_0 \end{pmatrix} \begin{pmatrix} \bar{Z} \\ 0 \end{pmatrix}$ and substituting $Q_0 = (A_0', B_0')' C_0$ we get (8). Q_0 is the unique solution to (8) because $E[\bar{Z}\bar{Z}']$ is non-singular by Assumption 1.4.

□

Lemma 2. *Under Assumptions 1.1-1.4 $\psi(\beta_0; M_0, \xi_0, \gamma_0, \mu_0)$ is doubly robust.*

Proof of Lemma 2. Recall that $\psi_i(\beta, \xi, \gamma, \mu) = \mu g_i(\beta, \xi, \gamma)$ where g_i is given by:

$$g_i(\beta, \xi, \gamma) = \begin{pmatrix} \tilde{Z}_i(\gamma_{Z,1}) \\ \tilde{X}_i(\gamma_{X,1}) \end{pmatrix} \left(\tilde{Y}_i(\gamma_Y) - \beta' \tilde{X}_i(\gamma_{X,2}) - \xi \tilde{V}_i(\gamma_{X,2})'\right)$$

Step 1: Show the score is robust to μ_0.

Corollary 1 immediately implies that $E\big[g_i(\beta_0;\xi_0,\gamma_0)\big]=0$ and so for any μ:

$$E\big[\psi_i(\beta_0;\xi_0,\gamma_0,\mu)\big]=\mu E\big[g_i(\beta_0;\xi_0,\gamma_0)\big]=0$$

And so the score function is doubly robust with respect to μ_0.

Step 2: Show the score is robust to ξ_0.

Consider the derivatives of $E\big[\psi_i(\beta;\xi,\gamma,\mu)\big]$ with respect to ξ with the other arguments set to their true values. With a little work one can show the derivatives are as follows:

$$\frac{\partial}{\partial\xi}E\big[\psi_i(\beta_0;M,\xi,\gamma_0,\mu_0)\big]=-\mu_0E[(\tilde{Z}_i',\tilde{X}_i')'\tilde{V}_i']$$

The derivatives does not depend on ξ. Therefore, if the derivative with respect to ξ is zero at ξ_0 then it is zero for all ξ. As in the main text, define G_η by:

$$G_\eta=\frac{\partial}{\partial\xi}E\big[g_i(\beta_0;\xi,\gamma_0)\big]\big|_{\xi=\xi_0}$$

Substituting the above we get:

$$\frac{\partial}{\partial\xi}E\big[\psi_i(\beta_0;\xi,\gamma_0,\mu_0)\big]\big|_{\xi=\xi_0}=\mu_0G_\eta$$

Substituting the definition of μ_0 the RHS becomes:

$$\begin{aligned}\mu_0G_\eta&=(G_\beta'\Omega^{-1}-G_\beta'\Omega^{-1}G_\eta(G_\eta'\Omega^{-1}G_\eta)^\dagger G_\eta'\Omega^{-1})G_\eta\\&=0\end{aligned}$$

The final equality follows by the elementary property of the Moore-Penrose pseudo-inverse that for any matrix A, $A(A'A)^\dagger A'A=AA^\dagger A=A$, even if A is nonsingular. So $\frac{\partial}{\partial\xi}E\big[\psi_i(\beta_0;\xi,\gamma_0,\mu_0)\big]\big|_{\xi=\xi_0}=0$ and thus $\frac{\partial}{\partial\xi}E\big[\psi_i(\beta_0;\xi,\gamma_0,\mu_0)\big]=0$ for all ξ. Since $E\big[\psi_i(\beta_0;\xi_0,\gamma_0,\mu_0)\big]=0$ it follows that $E\big[\psi_i(\beta_0;\xi_0,\gamma_0,\mu_0)\big]=0$ for all ξ.

Step 3: Show the score is robust to the components of γ_0.

Suppose γ differs from γ_0 only in that $\gamma_Y\neq\gamma_{0,Y}$. By the properties of partialling out, for any γ_Y:

$$\begin{aligned}E\big[(\tilde{Z}_i',\tilde{X}_i')\tilde{Y}_i(\gamma_Y)\big]&=E[(\tilde{Z}_i',\tilde{X}_i')Y_i]-E[(\tilde{Z}_i',\tilde{X}_i')D_i']\gamma_Y\\&=E[(\tilde{Z}_i',\tilde{X}_i')\tilde{Y}_i]\end{aligned}$$

γ_Y only enters $E\big[\psi_i(\beta_0;M,\xi_0,\gamma,\mu_0)\big]=0$, through the expression above, so we are robust to the γ_Y component of γ_0. By the same reasoning:

$$E\big[(\tilde{Z}_i',\tilde{X}_i')\tilde{V}_i(\gamma_V)'\big]=E[(\tilde{Z}_i',\tilde{X}_i')\tilde{V}_i']$$

And so we are robust to γ_V. We can follow similar steps to show we are robust to $\gamma_{1,X}$, $\gamma_{2,X}$, $\gamma_{1,Z}$, and $\gamma_{2,Z}$.

Note this is why we treat $\gamma_{0,X}$ as two different parameters in the two places it enters the score function and likewise for $\gamma_{0,Z}$. If $\gamma_X\neq\gamma_{0,X}$ when in general $E[\tilde{X}_i(\gamma_X)\tilde{X}_i(\gamma_X)']\neq E[\tilde{X}_i\tilde{X}_i']$ but $E[\tilde{X}_i(\gamma_{X,1})\tilde{X}_i(\gamma_{0,X})']=E[\tilde{X}_i\tilde{X}_i']$ regardless of $\gamma_{X,1}$.

□

Proof of Theorem 3.2. To prove the result we confirm that the conditions for Theorems 3.1 and 3.2 in Chernozhukov *et al.* (2018) hold. The result follows immediately from those theorems.

Theorems 3.1 and 3.2 in Chernozhukov *et al.* (2018) require Assumptions 3.1 and 3.2 in that paper. Let us begin with Assumption 3.1. This states that a) the true parameter (β_0 in our case) satisfies the moment condition. b) That the moment condition is linear in this parameter. c) That the map from the parameters to the moment is twice continuously Gateaux differentiable. d) That the score is Neyman orthogonal (or 'near Neyman orthogonal'). (e) S_0 has eigenvalues bounded above and below away from zero. By Lemma 2 the moment condition is valid so a) hold. By Lemma 3 the score is doubly-robust and therefore Neyman-orthogonal so d) holds. The score is linear in β_0 and it is linear in each of its arguments and is thus continuously twice Gateaux differentiable, so b) and c) hold. Condition (e) holds by supposition. Thus Assumption 3.1 of Chernozhukov *et al.* (2018) is satisfied.

We now show that Assumption 3.2 of Chernozhukov *et al.* (2018) holds. this constitutes the bulk of the proof. Below we restate this assumption as it applies in our setting. It will be convenient to collect all the nuisance parameters into one single parameter. In particular, let η_0 contain the true values of all the nuisance parameters so that:

$$\eta_0 = (\mu_0, \xi_0, \gamma_0)$$

In the above, the parentheses indicate an ordered set rather than horizontal concatenation of matrices. Similarly, let $\hat{\eta}_j$ be the collection of all the nuisance parameter estimates for the j^{th} sub-sample:

$$\hat{\eta}_j = (\hat{\mu}_j, \hat{\xi}_j, \hat{\gamma}_j)$$

Moreover, for some $\eta = (\mu, \xi, \gamma)$ we define $\psi_i(\beta, \eta) = \psi_i(\beta; \xi, \gamma, \mu)$ and use ψ_i as shorthand for $\psi_i(\beta_0, \eta_0)$.

Assumption 3.2 of Chernozhukov *et al.* (2018) states that there are sequences $\alpha_n \to 0$ and $\delta_n \to 0$, constants c_0 and c_1, and a sequence of sets $\mathcal{T}_n$ so that for each n if $P \in \mathcal{P}_n$ the conditions below all hold.

1. With probability at least $1 - \alpha_n$, $\hat{\eta}_j \in \mathcal{T}_n$ for all $j = 1, ..., J$.
2. $\sup_{\eta \in \mathcal{T}_n} E\big[\|\psi_i(\beta_0, \eta)\|^q\big]^{1/q} \leq c_1$
3. $\sup_{\eta \in \mathcal{T}_n} E\big[\|\mu E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\|^q\big]^{1/q} \leq c_1$
4. $\sup_{\eta \in \mathcal{T}_n} \|\mu_0 E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - \mu E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\| \leq \delta_n$
5. $\sup_{\eta \in \mathcal{T}_n} E\big[\|\psi_i(\beta_0, \eta_0) - \psi_i(\beta_0, \eta)\|^2\big]^{1/2} \leq \delta_n$
6. $\sup_{r \in (0,1), \eta \in \mathcal{T}_n} \|\frac{\partial^2}{\partial r^2} E\Big[\psi_i\big(\beta_0, \eta_0 + r(\eta - \eta_0)\big)\Big]\| \leq \delta_n / \sqrt{n}$
7. The eigenvalues of $E[\psi_i(\beta_0, \eta_0)\psi_i(\beta_0, \eta_0)']$ are bounded below by a constant c_0.

Note that condition 7 holds by supposition. For conditions 4, 5, and 6 we will show that the following rates apply uniformly over $P \in \mathcal{P}_n$:

$$\sup_{\eta\in\mathcal{T}_n} \|\mu_0 E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - \mu E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\|$$
$$\precsim \delta_\mu + (\delta_{\gamma,Z} + \delta_{\gamma,X})\delta_{\gamma,X}$$

$$\sup_{r\in(0,1),\eta\in\mathcal{T}_n} E\big[\|\psi_i(\beta_0,\eta_0) - \psi_i(\beta_0,\eta)\|^2\big]^{1/2}$$
$$\precsim \sqrt{d_X}(\delta_\mu + \delta_\xi + \delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi + \delta_{\gamma,Z} + \delta_{\gamma,X})$$
$$+ \sqrt{d_X}(\delta_{\gamma,X} + \delta_{\gamma,Z})(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi)$$

$$\sup_{r\in(0,1),\eta\in\mathcal{T}_n} \|\frac{\partial^2}{\partial r^2} E\big[\psi_i\big(\beta_0, \eta_0 + r(\eta - \eta_0)\big)\big]\|$$
$$\precsim \delta_\mu\delta_\xi + (\delta_{\gamma,X} + \delta_{\gamma,Z})(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi)$$

This implies that conditions 4, 5, and 6 hold with:

$$\begin{aligned}\delta_n \precsim\ & \sqrt{n}\delta_\mu\delta_\xi \\ & + \sqrt{d_X}(\delta_\mu + \delta_\xi + \delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi + \delta_{\gamma,Z} + \delta_{\gamma,X}) \\ & + (\sqrt{n} + \sqrt{d_X})(\delta_{\gamma,X} + \delta_{\gamma,Z})(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi)\end{aligned}$$

The premises of the theorem imply that each of the three terms in the sum above $o(1)$, and so we get that $\delta_n \prec 1$ as required.

Before we begin confirming each of the conditions stated above, we first show that $\|\mu_0\| \leq c^2$. Note that by the definition of μ_0 and the definition of the matrix norm:

$$\|\mu_0\| \leq \|E[\tilde{X}_i(\tilde{Z}_i', \tilde{X}_i')]\|\|I - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\|\|\Omega^{-1}\|$$

The matrix $I - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger$ is idempotent and thus has matrix norm less that unity. $E[\tilde{X}_i(\tilde{Z}_i', \tilde{X}_i')]$ is a sub-matrix of $\Sigma_{\tilde{Z}\tilde{X}}$ and so $\|E[\tilde{X}_i(\tilde{Z}_i', \tilde{X}_i')]\| \leq \|\Sigma_{\tilde{Z}\tilde{X}}\|$. Combining we get:

$$\|\mu_0\| \leq \|\Sigma_{\tilde{Z}\tilde{X}}\|\|\Omega^{-1}\|$$

By Assumption 3.2.i, $\|\Sigma_{\tilde{Z}\tilde{X}}\| \leq c$, and by the conditions of the theorem $\|\Omega^{-1}\| \leq c$, and so $\|\mu_0\| \leq c^2$.

We will now consider conditions 1-6 in turn. In order to reduce the complexity of some of the expressions in the arguments below, we use the following notation: $\gamma_R = (\gamma_Z', \gamma_X')'$, $\gamma_H = (\gamma_Y', \gamma_V', \gamma_X')'$, $\tilde{R}_i(\gamma_R) = (\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'$, $\tilde{H}_i(\gamma_H) = (\tilde{Y}_i(\gamma_Y), \tilde{V}_i(\gamma_V)', \tilde{X}_i(\gamma_X)')'$ and $\zeta = (1, -\xi, -\beta_0')'$. In addition let $\gamma_{R,0} = (\gamma_{Z,0}', \gamma_{X,0}')'$, $\gamma_{H,0} = (\gamma_{Y,0}', \gamma_{V,0}', \gamma_{X,0}')'$, let $\tilde{R}_i = \tilde{R}_i(\gamma_{R,0})$ and $\tilde{H}_i = \tilde{H}_i(\gamma_{H,0})$.

Condition 1

Under Assumption 3.1, the set $\mathcal{T}_n$ defined as follows satisfies Condition 1 for some $\alpha_n \to 0$. $\eta \in \mathcal{T}_n$ if and only if $\|\mu - \mu_0\| \leq \delta_\mu$, $\|\xi - \xi_0\| \leq \delta_\xi$, and $\|\gamma_Q - \gamma_{Q,0}\| \leq \delta_{\gamma,Q}$ for $Q \in \{V, X, Z, \epsilon\}$. In our discussion of the remaining conditions we take $\mathcal{T}_n$ to be this set.

Condition 2

By Assumption 3.2.v it is enough to show that for n sufficiently large, for all $\eta \in \mathcal{T}_n$, $\|\mu - \mu_0\| \leq 1/c$, $\|\xi - \xi_0\| \leq 1/c$, and $\|\gamma_Q - \gamma_{Q,0}\| \leq 1/c$ for $Q \in \{V, X, Z, \epsilon\}$. By the definition of $\mathcal{T}_n$ this holds so long as:

$$\delta_\mu, \delta_\xi, \delta_{\gamma,X}, \delta_{\gamma,V}, \delta_{\gamma,Z}, \delta_{\gamma,\epsilon} \prec 1$$

And the above is true by the conditions of the Theorem.

Condition 3

This follows from the same argument as we used for Condition 2.

Condition 4

Next we show that, uniformly over all $P \in \mathcal{P}_n$:

$$\begin{aligned}
&\sup_{\eta \in \mathcal{T}_n} \|\mu_0 E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - \mu E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\| \\
\precsim &\delta_\mu + (\delta_{\gamma,Z} + \delta_{\gamma,X})\delta_{\gamma,X}
\end{aligned}$$

By the triangle inequality and definition of the matrix norm:

$$\begin{aligned}
&\|\mu_0 E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - \mu E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\| \\
\leq &\|\mu_0 - \mu\| \|E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i']\| \\
+ &(\|\mu_0 - \mu\| + \|\mu_0\|)\|E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\|
\end{aligned}$$

Using the properties of partialling out:

$$\begin{aligned}
&E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)'] \\
= &E[\big((\tilde{Z}_i', \tilde{X}_i')' - \tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')\big)(\tilde{X}_i(\gamma_X)' - \tilde{X}_i')] \\
= &(\gamma_{Z,0}' - \gamma_Z', \gamma_{X,0}' - \gamma_X')'\Sigma_D(\gamma_{X,0} - \gamma_X)
\end{aligned}$$

And so:

$$\begin{aligned}
&\|E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\| \\
\leq &\|(\gamma_{Z,0}' - \gamma_Z', \gamma_{X,0}' - \gamma_X')\|\|(\gamma_{X,0} - \gamma_X)\|\|\Sigma_D\|
\end{aligned}$$

Combining we get:

$$\begin{aligned}
&\|\mu_0 E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - \mu E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\| \\
\leq &\|\mu_0 - \mu\| \|E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i']\| \\
+ &(\|\mu_0 - \mu\| + \|\mu_0\|) \\
&\times \|(\gamma_{Z,0}' - \gamma_Z', \gamma_{X,0}' - \gamma_X')'\|\|\gamma_{X,0} - \gamma_X\|\|\Sigma_D\|
\end{aligned}$$

Note that:

$$\|E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i']\| \leq \|\Sigma_{\tilde{Z}\tilde{X}}\|$$

And so, if $\eta \in \mathcal{T}_n$ and Assumption 3.2 holds we get:

$$\begin{aligned}
&\|\mu_0 E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - \mu E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\| \\
\leq &c\delta_\mu + c(\delta_\mu + c^2)(\delta_{\gamma,Z} + \delta_{\gamma,X})\delta_{\gamma,X}
\end{aligned}$$

Under the conditions of the Theorem we then have uniformly over $P \in \mathcal{P}_n$:

$$\begin{aligned}&\|\mu_0 E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - \mu E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']\| \\ \precsim &\delta_\mu + (\delta_{\gamma,Z} + \delta_{\gamma,X})\delta_{\gamma,X}\end{aligned}$$

Condition 5
We will show that uniformly over $P \in \mathcal{P}_n$:

$$\begin{aligned}&\sup_{\eta \in \mathcal{T}_n} E\big[\|\psi_i(\beta_0, \eta_0) - \psi_i(\beta_0, \eta)\|^2\big]^{1/2} \\ \precsim &\sqrt{d_X}(\delta_\mu + \delta_\xi + \delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi + \delta_{\gamma,Z} + \delta_{\gamma,X}) \\ &+ d_D(\delta_{\gamma,X} + \delta_{\gamma,Z})(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi)\end{aligned}$$

In the notation introduced earlier in the proof:

$$\begin{aligned}&E\big[\|\psi_i(\beta_0, \eta_0) - \psi_i(\beta_0, \eta)\|^2\big]^{1/2} \\ =&E\big[\|\mu_0 \tilde{R}_i \tilde{H}_i'\zeta_0 - \mu \tilde{R}_i(\gamma_R)\tilde{H}_i(\gamma_H)'\zeta\|^2\big]^{1/2}\end{aligned}$$

Using the triangle inequality and definition of the operator norm:

$$\begin{aligned}&E\big[\|\psi_i(\beta_0, \eta_0) - \psi_i(\beta_0, \eta)\|^2\big]^{1/2} \\ \leq &E\left[\|(\mu - \mu_0)(\tilde{Z}_i', \tilde{X}_i')'\|^2 \tilde{\epsilon}_i^2\right]^{1/2} \\ &+ E\left[\|\mu(\tilde{Z}_i', \tilde{X}_i')'\|^2 \|\tilde{V}_i'(\xi - \xi_0)\|^2\right]^{1/2} \\ &+ E\left[\|\mu\big(\tilde{R}_i(\gamma_R)\tilde{H}_i(\gamma_H)' - \tilde{R}_i\tilde{H}_i'\big)\zeta\|^2\right]^{1/2} \qquad (34)\end{aligned}$$

Under Assumption 3.2.iii, the first term on the RHS is bounded by:

$$\begin{aligned}&E\left[\|(\mu - \mu_0)(\tilde{Z}_i', \tilde{X}_i')'\|^2 \tilde{\epsilon}_i^2\right]^{1/2} \\ =&E\left[\|(\mu - \mu_0)(\tilde{Z}_i', \tilde{X}_i')'\|^2 E[\tilde{\epsilon}_i^2 | \tilde{Z}_i, \tilde{X}_i]\right]^{1/2} \\ \leq &cE\left[\|(\mu - \mu_0)(\tilde{Z}_i', \tilde{X}_i')'\|^2\right]^{1/2} \\ \leq &\sqrt{d_X}c\|\mu - \mu_0\|\|\Sigma_{\tilde{Z}\tilde{X}}^{1/2}\| \\ \precsim &\sqrt{d_X}\delta_\mu\end{aligned}$$

Where the final inequality above assumes $\eta \in \mathcal{T}_n$. For the second term on the RHS of 34, if $\eta \in \mathcal{T}_n$ then:

$$
\begin{aligned}
&E\left[\|\mu(\tilde{Z}_i', \tilde{X}_i')'\|^2 \|\tilde{V}_i'(\xi - \xi_0)\|^2\right]^{1/2} \\
=&E\left[\|\mu(\tilde{Z}_i', \tilde{X}_i')'\|^2 \|E[\tilde{V}_i'\tilde{V}_i|\tilde{Z}_i, \tilde{X}_i]^{1/2}(\xi - \xi_0)\|^2\right]^{1/2} \\
\leq&E\left[\|\mu(\tilde{Z}_i', \tilde{X}_i')'\|^2 \|E[\tilde{V}_i'\tilde{V}_i|\tilde{Z}_i, \tilde{X}_i]\|\right]^{1/2} \|\xi - \xi_0\| \\
\precsim&\delta_\xi E\left[\|\mu(\tilde{Z}_i', \tilde{X}_i')'\|^2\right]^{1/2} \\
\precsim&\delta_\xi \sqrt{d_X}\|\mu\|\|\Sigma_{\tilde{Z}\tilde{X}}^{1/2}\| \\
\precsim&\delta_\xi \sqrt{d_X}(\delta_\mu + \|\mu_0\|)
\end{aligned}
$$

Next we will show that:

$$
\begin{aligned}
&E\left[\|\mu\big(\tilde{R}_i(\gamma_R)\tilde{H}_i(\gamma_H)' - \tilde{R}_i\tilde{H}_i'\big)\zeta\|^2\right]^{1/2} \\
\precsim&\sqrt{d_X}(\|\mu_0\| + \delta_\mu)(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi) \\
+&2\sqrt{d_X}(\|\mu_0\| + \delta_\mu)(\delta_{\gamma,Z} + \delta_{\gamma,X})(1 + \delta_\xi) \\
+&d_D(\delta_\mu + \|\mu_0\|)(\delta_{\gamma,X} + \delta_{\gamma,Z})(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi)
\end{aligned} \tag{35}
$$

To see this, we first apply the triangle inequality and Young's inequality to get:

$$
\begin{aligned}
&E\left[\|\mu\big(\tilde{R}_i(\gamma_R)\tilde{H}_i(\gamma_H)' - \tilde{R}_i\tilde{H}_i'\big)\zeta\|^2\right]^{1/2} \\
\leq&E\big[\|\mu\tilde{R}_i\|^2\|D_i'(\gamma_{H,0} - \gamma_H)'\zeta\|^2\big]^{1/2} \\
+&2E\big[\|\mu(\gamma_{R,0} - \gamma_R)D_i\|^2\tilde{\epsilon}_i^2\big]^{1/2} \\
+&2E\big[\|\mu(\gamma_{R,0} - \gamma_R)D_i\|^2\|\tilde{V}_i'(\xi - \xi_0)|^2\big]^{1/2} \\
+&E\big[\|\mu(\gamma_{R,0} - \gamma_R)D_iD_i'(\gamma_{H,0} - \gamma_H)'\zeta\|^2\big]^{1/2}
\end{aligned} \tag{36}
$$

To bound the above first note that under Assumption 3.2:

$$
\begin{aligned}
E[\|\mu\tilde{R}_i\|^2|D_i] &\leq d_X\|\mu E[\tilde{R}_i\tilde{R}_i'|D_i]^{1/2}\|^2 \\
&\leq d_X\|\mu\|^2 \cdot \|E[\tilde{R}_i\tilde{R}_i'|D_i]\| \\
&\precsim d_X(\|\mu_0\| + \delta_\mu)^2
\end{aligned}
$$

Where the final inequality above assumes $\eta \in \mathcal{T}_n$. Similarly:

$$
\begin{aligned}
E[\|\mu(\gamma_{R,0} - \gamma_R)D_i\|^2] &\leq d_X\|\mu(\gamma_{R,0} - \gamma_R)\Sigma_D^{1/2}\|^2 \\
&\leq d_X\|\mu\|^2 \cdot \|\gamma_{R,0} - \gamma_R\|^2\|\Sigma_D\| \\
&\precsim d_X(c^2 + \delta_\mu)^2(\delta_{\gamma,Z} + \delta_{\gamma,X})^2
\end{aligned}
$$

And moreover, if Assumption 3.2 holds and $\eta \in \mathcal{T}_n$:

$$\begin{aligned} E[|\tilde{V}_i'(\xi - \xi_0)|^2 | D_i] &\leq \|\xi - \xi_0\|^2 \cdot \|E[\tilde{V}_i \tilde{V}_i' | D_i]\| \\ &\precsim \delta_\xi^2 \end{aligned}$$

If $\eta \in \mathcal{T}_n$, then the using the definition of ζ, γ_H, and $\gamma_{H,0}$:

$$\begin{aligned} & E\big[\|D_i'(\gamma_{H,0} - \gamma_H)'\zeta\|^2\big]^{1/2} \\ =& \|\Sigma_D^{1/2}(\gamma_{H,0} - \gamma_H)'\zeta\| \\ \leq& \|(\gamma_Y - \gamma_{Y,0}) - (\gamma_X - \gamma_{X,0})\beta_0 - (\gamma_V - \gamma_{V,0})\xi_0'\| \|\Sigma_D^{1/2}\| \\ +& \|\gamma_{V,0} - \gamma_V\| \|\xi - \xi_0\| \|\Sigma_D^{1/2}\| \\ \precsim& \delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi \end{aligned}$$

If Assumption 3.2 holds, then using the law of iterated expectations and the above we get from 36:

$$\begin{aligned} & E\left[\|\mu\big(\tilde{R}_i(\gamma_R)\tilde{H}_i(\gamma_H)' - \tilde{R}_i\tilde{H}_i'\big)\zeta\|^2\right]^{1/2} \\ \precsim& \sqrt{d_X}(\|\mu_0\| + \delta_\mu)(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi) \\ +& \sqrt{d_X}(\|\mu_0\| + \delta_\mu)(\delta_{\gamma,Z} + \delta_{\gamma,X})(1 + \delta_\xi) \\ +& E\big[\|\mu(\gamma_{R,0} - \gamma_R)D_iD_i'(\gamma_{H,0} - \gamma_H)'\zeta\|^2\big]^{1/2} \end{aligned} \tag{37}$$

Applying Assumption 3.2:

$$\begin{aligned} & E\big[\|\mu(\gamma_{R,0} - \gamma_R)D_iD_i'(\gamma_{H,0} - \gamma_H)'\zeta\|^2\big] \\ \leq& d_X\|\mu(\gamma_{R,0} - \gamma_R)\|^2\|(\gamma_{H,0} - \gamma_H)'\zeta\|^2 c^2 \\ \leq& d_X(\delta_\mu + \|\mu_0\|)^2(\delta_{\gamma,X} + \delta_{\gamma,Z})^2(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi)^2 \end{aligned}$$

Where the last line assumes $\eta \in \mathcal{T}_n$. Combining we get 35, and in all:

$$\begin{aligned} & E\big[\|\psi_i(\beta_0, \eta_0) - \psi_i(\beta_0, \eta)\|^2\big]^{1/2} \\ \precsim& \sqrt{d_X}\delta_\mu + \delta_\xi\sqrt{d_X}(\delta_\mu + \|\mu_0\|) \\ +& \sqrt{d_X}(\|\mu_0\| + \delta_\mu)(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi) \\ +& \sqrt{d_X}(\|\mu_0\| + \delta_\mu)(\delta_{\gamma,Z} + \delta_{\gamma,X})(1 + \delta_\xi) \\ +& \sqrt{d_X}(\|\mu_0\| + \delta_\mu)(\delta_{\gamma,X} + \delta_{\gamma,Z})(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi) \end{aligned}$$

Under the conditions of the theorem we get:

$$\begin{aligned} & E\big[\|\psi_i(\beta_0, \eta_0) - \psi_i(\beta_0, \eta)\|^2\big]^{1/2} \\ \precsim& \sqrt{d_X}(\delta_\mu + \delta_\xi + \delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi + \delta_{\gamma,Z} + \delta_{\gamma,X}) \\ +& \sqrt{d_X}(\delta_{\gamma,X} + \delta_{\gamma,Z})(\delta_{\gamma,\epsilon} + \delta_{\gamma,V}\delta_\xi) \end{aligned}$$

Condition 6
Next we show that uniformly over $P \in \mathcal{P}_n$:

$$\begin{aligned}&\sup_{r\in(0,1),\eta\in\mathcal{T}_n} \|\frac{\partial^2}{\partial r^2} E\big[\psi_i\big(\beta_0, \eta_0 + r(\eta-\eta_0)\big)\big]\| \\ \lesssim &\delta_\mu\delta_\xi + (\delta_{\gamma,X}+\delta_{\gamma,Z})(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi)\end{aligned}$$

Twice differentiating we get:

$$\begin{aligned}&\frac{\partial^2}{\partial r^2} E\big[\psi_i\big(\beta_0, \eta_0 + r(\eta-\eta_0)\big)\big] \\ =&2(\mu-\mu_0)\Sigma_{\tilde{Z}\tilde{X}}M_0'(\xi_0-\xi)' \\ &+2\mu_0(\gamma_Z'-\gamma_{Z,0}', \gamma_X'-\gamma_{X,0}')'\Sigma_D \\ &\quad\times\big((\gamma_Y-\gamma_{0,Y})-(\gamma_V-\gamma_{0,V})'\xi_0'-(\gamma_X-\gamma_{0,X})'\beta_0\big) \\ &+6r\mu_0(\gamma_Z'-\gamma_{Z,0}', \gamma_X'-\gamma_{X,0}')'\Sigma_D(\gamma_V-\gamma_{0,V})'(\xi_0-\xi)' \\ &+6r(\mu-\mu_0)(\gamma_Z'-\gamma_{Z,0}', \gamma_X'-\gamma_{X,0}')'\Sigma_D \\ &\quad\times\big((\gamma_Y-\gamma_{0,Y})-(\gamma_V-\gamma_{0,V})'\xi_0'-(\gamma_X-\gamma_{0,X})'\beta_0\big) \\ &+12r^2(\mu-\mu_0)(\gamma_Z'-\gamma_{Z,0}', \gamma_X'-\gamma_{X,0}')'\Sigma_D(\gamma_V-\gamma_{0,V})'(\xi_0-\xi)'\end{aligned}$$

Where we have used that $E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{V}_i'] = \Sigma_{\tilde{Z}\tilde{X}}M_0'$. Applying the triangle inequality and the definition of the operator norm:

$$\begin{aligned}&\|\frac{\partial^2}{\partial r^2} E\big[\psi_i\big(\beta_0, \eta_0 + r(\eta-\eta_0)\big)\big]\| \\ \le&2\|\mu-\mu_0\|\|M_0\|\|\xi_0-\xi\|\|\Sigma_{\tilde{Z}\tilde{X}}\| \\ &+2\|\mu_0\|(\|\gamma_Z-\gamma_{Z,0}\|+\|\gamma_X-\gamma_{X,0}\|)\|\Sigma_D\|\|\gamma_\epsilon-\gamma_{0,\epsilon}\| \\ &+6r\|\mu_0\|(\|\gamma_Z-\gamma_{Z,0}\|+\|\gamma_X-\gamma_{X,0}\|)\|\gamma_V-\gamma_{0,V}\|\|\xi_0-\xi\|\|\Sigma_D\| \\ &+6r\|\mu-\mu_0\|(\|\gamma_Z-\gamma_{Z,0}\|+\|\gamma_X-\gamma_{X,0}\|)\|\Sigma_D\|\|\gamma_\epsilon-\gamma_{0,\epsilon}\| \\ &+12r^2\|\mu-\mu_0\|(\|\gamma_Z-\gamma_{Z,0}\|+\|\gamma_X-\gamma_{X,0}\|)\|\gamma_V-\gamma_{0,V}\|\|\xi_0-\xi\|\|\Sigma_D\|\end{aligned}$$

The expression above is maximized over $r \in [0,1]$ by $r = 1$. If Assumption 3.2 holds and $\eta \in \mathcal{T}_n$ then we get from the above:

$$\begin{aligned}&\|\frac{\partial^2}{\partial r^2} E\big[\psi_i\big(\beta_0, \eta_0 + r(\eta-\eta_0)\big)\big]\| \\ \le&2c^2\delta_\mu\delta_\xi \\ &+2c(c^2+3\delta_\mu)(\delta_{\gamma,Z}+\delta_{\gamma,X})\delta_{\gamma,\epsilon} \\ &+6c(c^2+4\delta_\mu)(\delta_{\gamma,Z}+\delta_{\gamma,X})\delta_{\gamma,V}\delta_\xi\end{aligned}$$

So using the conditions of the theorem, uniformly over $P \in \mathcal{P}_n$:

$$\|\frac{\partial^2}{\partial r^2}E\big[\psi_i\big(\beta_0,\eta_0+r(\eta-\eta_0)\big)\big]\|$$
$$\precsim\delta_\mu\delta_\xi+(\delta_{\gamma,X}+\delta_{\gamma,Z})(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi)$$

□

Proof of Proposition 3.1. Throughout the proof, for any matrix A with side lengths greater than k, $\sigma_k(A)$ denotes the k^{th} largest singular value of A and $\sigma_{\min}(A)$ the smallest singular value. We use the following fact about the Moore-Penrose pseudoinverse (which can be found in Stewart (1977)) for any two matrices A and B with $rank(A) = rank(B)$ and $\|A^\dagger\|\|A-B\| < 1$:

$$\|A^\dagger - B^\dagger\| \leq \frac{\|A^\dagger\|^2\|A-B\|}{1-\|A^\dagger\|\|A-B\|} \tag{38}$$

Another important inequality is that for any two real $K\times K$ matrices A and B, $\sigma_k(AB') \leq \sigma_k(A)\sigma_1(B)$ for $k = 1, ..., K$ (see for example Theorem 3.3.16 in Horn & Johnson (1991)) If B is non-singular then this inequality implies:

$$\sigma_k(A) = \sigma_k\big(AB'(B')^{-1}\big) \leq \sigma_k(AB')\sigma_1\big((B')^{-1}\big) = \sigma_k(AB')\|(B')^{-1}\| \tag{39}$$

Where we have used that $\sigma_1\big((B')^{-1}\big) = \|(B')^{-1}\|$. Note that this inequality easily extends to the case in which A is $L\times K$ with $L \neq K$ because $\sigma_k(AB') = \sigma_k\big((A'A)^{1/2}B'\big)$ and $\sigma_k(A) = \sigma_k\big((A'A)^{1/2}\big)$.

Recall that $\underline{\sigma}_n$ is the d_Wth largest singular value of $E[\bar{V}\bar{W}']E[\bar{W}\bar{W}']^\dagger E[\bar{W}\bar{Z}']$. Under Assumption 1.1 this matrix is equal to $B_0C_0E[\bar{Z}\bar{Z}']$. From (39) and the fact that $E[\bar{Z}\bar{Z}']$ is nonsingular (by Assumption 1.4) it follows that:

$$\sigma_{d_W}\big(E[\bar{V}\bar{W}']E[\bar{W}\bar{W}']^\dagger E[\bar{W}\bar{Z}']\big) \leq \sigma_{d_W}(B_0C_0)\|E[\bar{Z}\bar{Z}']^{-1}\|$$

Because B_0C_0 has rank d_W we have $\|(B_0C_0)^\dagger\| = \sigma_{d_W}(B_0C_0)^{-1}$, using this and the definition of $\underline{\sigma}_n$ we get:

$$\|(B_0C_0)^\dagger\| \leq \underline{\sigma}_n^{-1}\|\Sigma_{\bar{Z}}^{-1}\| \tag{40}$$

Proof of part a.

Using the definitions of $\hat{\xi}$ and ξ_0 and applying the triangle inequality and definition of the operator norm:

$$\begin{aligned}\|\hat{\xi}-\xi_0\| \leq\ & \|(B_0C_0)^\dagger\|\|\hat{Q}_{[d_V+1,:]} - A_0C_0\| \\ &+ \|(B_0C_0)^\dagger - \hat{Q}^\dagger_{[1:d_V,:]}\|\|A_0C_0\| \\ &+ \|(B_0C_0)^\dagger - \hat{Q}^\dagger_{[1:d_V,:]}\|\|\hat{Q}_{[d_V+1,:]} - A_0C_0\|\end{aligned}$$

Recall that by Corollary 1 $rank(B_0C_0) = d_W$, and so, if $rank(\hat{Q}_{[d_V+1,:]}) = d_W$ and $\|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\| < \underline{\sigma}_n\|\Sigma_{\bar{Z}}^{-1}\|^{-1}$ then by (38):

$$\|(B_0C_0)^\dagger - \hat{Q}^\dagger_{[1:d_V,:]}\| \leq \frac{\underline{\sigma}_n^{-2}\|\Sigma_{\bar{Z}}^{-1}\|^2\|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\|}{1-\underline{\sigma}_n^{-1}\|\Sigma_{\bar{Z}}^{-1}\|\|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\|}$$

Where we have used (40). So if $rank(\hat{Q}_{[d_V+1,:]}) = d_W$, and $\|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\| \leq \alpha_n \|\Sigma_{\bar{Z}}^{-1}\| \underline{\sigma}_n$:

$$\|(B_0C_0)^\dagger - \hat{Q}^\dagger_{[1:d_V,:]}\| \leq \frac{1}{1-\alpha_n} \underline{\sigma}_n^{-2} \|\Sigma_{\bar{Z}}^{-1}\|^2 \|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\|$$

Combining we get:

$$\begin{aligned}
\|\hat{\xi} - \xi_0\| \leq\ & \underline{\sigma}_n^{-1} \|\Sigma_{\bar{Z}}^{-1}\| \|\hat{Q}_{[d_V+1,:]} - A_0C_0\| \\
& + \frac{1}{1-\alpha_n} \underline{\sigma}_n^{-2} \|\Sigma_{\bar{Z}}^{-1}\|^2 \|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\| \|A_0C_0\| \\
& + \frac{1}{1-\alpha_n} \underline{\sigma}_n^{-2} \|\Sigma_{\bar{Z}}^{-1}\|^2 \|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\| \|\hat{Q}_{[d_V+1,:]} - A_0C_0\|
\end{aligned}$$

Using $\underline{\sigma}_n^{-1} \|\Sigma_{\bar{Z}}^{-1}\| \|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\| \leq \alpha_n$ the above yields:

$$\begin{aligned}
\|\hat{\xi} - \xi_0\| \leq & \frac{\underline{\sigma}_n^{-1} \|\Sigma_{\bar{Z}}^{-1}\|}{1-\alpha_n} \|\hat{Q}_{[d_V+1,:]} - A_0C_0\| \\
& + \frac{\underline{\sigma}_n^{-2} \|\Sigma_{\bar{Z}}^{-1}\|^2}{1-\alpha_n} \|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\| \|A_0C_0\|
\end{aligned}$$

By supposition a number of quantities on the RHS are bounded above by c_1 so we get:

$$\begin{aligned}
& \|\hat{\xi} - \xi_0\| \\
\leq & \frac{\underline{\sigma}_n^{-1} c_1}{1-\alpha_n} \|\hat{Q}_{[d_V+1,:]} - A_0C_0\| + \frac{\underline{\sigma}_n^{-2} c_1^3}{1-\alpha_n} \|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\| \\
\leq & 2(c_1 + c_1^3) \frac{1 + \underline{\sigma}_n^{-2}}{1-\alpha_n} \left(\|\hat{Q}_{[d_V+1,:]} - A_0C_0\| + \|(B_0C_0) - \hat{Q}_{[1:d_V,:]}\| \right)
\end{aligned}$$

Where the final line uses that for any $a > 0$, $a + a^2 \leq 2(1 + a^2)$. Finally, note that for any A and B with equal number of columns, $\|A\| + \|B\| \leq 2\|(A', B')'\|$ so we get:

$$\|\hat{\xi} - \xi_0\| \leq 4(c_1 + c_1^3) \frac{1 + \underline{\sigma}_n^{-2}}{1-\alpha_n} \|Q_0 - \hat{Q}\|$$

Setting $c = 4(c_1 + c_1^3)/(1 - \max_n \alpha_n)$ gives the result.

Proof of part b.

Let us suppose that $rank(\hat{M}) = rank(M_0)$, $\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| \leq \underline{\sigma}_n \alpha_n / \|\Sigma_{\bar{Z}}^{-1}\|$, $\|\hat{M} - M_0\| \leq \underline{\sigma}_n \alpha_n / \|\Sigma_{\bar{Z}}^{-1}\|$, and $\|\hat{\Sigma}_{\tilde{Z}\tilde{X}} - \Sigma_{\tilde{Z}\tilde{X}}\| \leq \underline{\sigma}_n \alpha_n / \|\Sigma_{\bar{Z}}^{-1}\|$. Since this holds with probability at least $1 - \alpha_n$ our conclusions will likewise hold with at least this probability. We take m to be the smallest natural number so that $3c_1^2 \alpha_n (c_1 + 1)^2 < 1$ and $3\underline{\sigma}_n \alpha_n \leq 1$ and assume $n \geq m$.

Applying the triangle inequality and definition of the matrix norm:

$$\begin{aligned}
&\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}' - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'\| \\
\leq&\big(\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| + \|\Omega^{-1/2}\|\big) \\
&\times \big(\|\hat{\Sigma}_{\tilde{Z}\tilde{X}} - \Sigma_{\tilde{Z}\tilde{X}}\| + \|\Sigma_{\tilde{Z}\tilde{X}}\|\big) \\
&\times \big(\|\hat{M} - M_0\| + \|M_0\|\big) \\
&- \|\Omega^{-1/2}\|\|\Sigma_{\tilde{Z}\tilde{X}}\|\|M_0'\| \\
\leq&(c_1+1)^2 \\
&\times \big(\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| + \|\hat{\Sigma}_{\tilde{Z}\tilde{X}} - \Sigma_{\tilde{Z}\tilde{X}}\| + \|\hat{M} - M_0\|\big) \\
\leq&3(c_1+1)^2\underline{\sigma}_n\alpha_n/\|\Sigma_{\tilde{Z}}^{-1}\| \qquad (41)
\end{aligned}$$

Where for the penultimate inequality we have used that $\|\Omega^{-1/2}\|, \|\Sigma_{\tilde{Z}\tilde{X}}\|, \|M_0\| \leq c_1$ and $\alpha_n \leq 1$.

From (39) we get that the smallest non-zero singular value of $\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'$ is greater than the smallest non-zero singular value of M_0 times $\|\Sigma_{\tilde{Z}\tilde{X}}^{-1}\Omega^{1/2}\|$, and therefore:

$$\begin{aligned}
\|(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| &\leq \|M_0^\dagger\|\|\Sigma_{\tilde{Z}\tilde{X}}^{-1}\Omega^{1/2}\| \\
&\leq \|M_0^\dagger\|\|\Omega^{1/2}\|\|\Sigma_{\tilde{Z}\tilde{X}}^{-1}\| \\
&\leq c_1^2\|M_0^\dagger\|
\end{aligned}$$

Under Assumptions 1.1-1.4 $M_0 = B_0(C_0, G_0)$ and $rank(B_0C_0) = rank(M_0)$, thus M_0 has the same number of non-zero singular values as B_0C_0. The singular values of a sub-matrix are all weakly smaller than the singular values of the original matrix, thus $\underline{\sigma}_n$, the smallest non-zero singular value of B_0C_0 (which satisfies (40), is weakly less than $\|M_0^\dagger\|^{-1}$, the smallest singular value of M_0. So we get:

$$\begin{aligned}
\|(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| &\leq c_1^2\|M_0^\dagger\| \\
&\leq c_1^2\underline{\sigma}_n^{-1}\|\Sigma_{\tilde{Z}}^{-1}\| \qquad (42)
\end{aligned}$$

By (38), if $\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}'$ and $\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'$ have the same rank, and the denominator on the RHS below is strictly positive:

$$\begin{aligned}
&\|(\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}')^\dagger - (\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| \\
\leq&\frac{\|(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\|^2 \cdot \|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}' - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'\|}{1 - \|(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\|\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}' - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'\|} \\
\leq&\tilde{c}\underline{\sigma}_n^{-2}\|\Sigma_{\tilde{Z}}^{-1}\|^2\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}' - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'\|
\end{aligned}$$

Where $\tilde{c} = \frac{c_1^4}{1-3c_1^2\alpha_n(c_1+1)^2}$ which is finite for $n \geq m$, and the final line uses 41.

$\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}$ and $\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}$ have full rank, and so $rank(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0') = rank(M_0)$ and $rank(\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}') = rank(\hat{M})$. By supposition $rank(\hat{M}) = rank(M_0)$ and so the above holds.

Next note that applying the triangle inequality and the definition of the matrix norm:

$$\begin{aligned}
&\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}'(\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}')^\dagger - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| \\
\leq&\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}' - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'\|\|(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| \\
+&\|\Omega^{-1/2}\|\|\Sigma_{\tilde{Z}\tilde{X}}\|\|M_0\|\|(\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}')^\dagger - (\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| \\
+&\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}' - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'\|\|(\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}')^\dagger - (\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| \\
\leq&(c_1^2\underline{\sigma}_n^{-1}\|\Sigma_{\bar{Z}}^{-1}\| + c_1^3\tilde{c}\underline{\sigma}_n^{-2}\|\Sigma_{\bar{Z}}^{-1}\|^2)\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}' - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'\| \\
+&\tilde{c}\underline{\sigma}_n^{-2}\|\Sigma_{\bar{Z}}^{-1}\|^2\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}' - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'\|^2
\end{aligned}$$

Combining the above with 42, setting $c^* = (c_1^2 + c_1^3\tilde{c} + \tilde{c})(c_1+1)^2$, we get:

$$\begin{aligned}
&\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}'(\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}')^\dagger - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| \\
\leq&c^*(\underline{\sigma}_n^{-1}\|\Sigma_{\bar{Z}}^{-1}\| + \underline{\sigma}_n^{-2}\|\Sigma_{\bar{Z}}^{-1}\|^2)\big(\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| + \|\hat{\Sigma}_{\tilde{Z}\tilde{X}} - \Sigma_{\tilde{Z}\tilde{X}}\| + \|\hat{M} - M_0\|\big) \\
+&c^*\underline{\sigma}_n^{-2}\|\Sigma_{\bar{Z}}^{-1}\|^2\big(\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| + \|\hat{\Sigma}_{\tilde{Z}\tilde{X}} - \Sigma_{\tilde{Z}\tilde{X}}\| + \|\hat{M} - M_0\|\big)^2 \qquad (43)
\end{aligned}$$

For the final step, first note that the matrix norm of a sub-matrix is weakly smaller than the norm of the full matrix and so $\|E[\tilde{X}(\tilde{Z}', \tilde{X}')]\| \leq \|\Sigma_{\tilde{Z}\tilde{X}}\|$ and:

$$\|E[\tilde{X}(\tilde{Z}', \tilde{X}')] - \frac{1}{n}\sum_{i=1}^n \hat{X}_i(\hat{Z}_i', \hat{X}_i')\| \leq \|\Sigma_{\tilde{Z}\tilde{X}} - \hat{\Sigma}_{\tilde{Z}\hat{X}}\|$$

In addition, note that the norm of an idempotent matrix is one or zero and so:

$$\|I - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| \leq 1$$

Applying the triangle inequality and definition of the matrix norm along with the inequality above, we get:

$$\begin{aligned}
&\|\mu_0 - \hat{\mu}\| \\
\leq&\Bigg(\big(\|\hat{\Sigma}_{\tilde{Z}\tilde{X}} - \Sigma_{\tilde{Z}\tilde{X}}\| + \|\Sigma_{\tilde{Z}\tilde{X}}\|\big)\big(\|\Omega^{-1/2} - \hat{\Omega}^{-1/2}\| + \|\Omega^{-1/2}\|\big)^2 \\
&\times\big(\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}'(\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}')^\dagger - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\| + 1\big)\Bigg) \\
-&\|\Sigma_{\tilde{Z}\tilde{X}}\|\|\Omega^{-1/2}\|^2
\end{aligned}$$

And so:

$$\begin{aligned}
&\|\mu_0 - \hat{\mu}\| \\
\leq&2(c_1+1)^2\|\Omega^{-1/2} - \hat{\Omega}^{-1/2}\| + 2(c_1+1)^2\|\hat{\Sigma}_{\tilde{Z}\tilde{X}} - \Sigma_{\tilde{Z}\tilde{X}}\| \\
+&(c_1+1)^3\|\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}'(\hat{\Omega}^{-1/2}\hat{\Sigma}_{\tilde{Z}\tilde{X}}\hat{M}')^\dagger - \Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0'(\Omega^{-1/2}\Sigma_{\tilde{Z}\tilde{X}}M_0')^\dagger\|
\end{aligned}$$

Using (43) and taking $c^\circ = (c_1+1)^3 c^* + 2(c_1+1)^2$ we get that:

$$\begin{aligned}
&\|\mu_0 - \hat{\mu}\| \\
\leq& c^\circ(1 + \underline{\sigma}_n^{-1}\|\Sigma_{\bar{Z}}^{-1}\| + \underline{\sigma}_n^{-2}\|\Sigma_{\bar{Z}}^{-1}\|^2)\big(\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| + \|\hat{\Sigma}_{\bar{Z}\tilde{X}} - \Sigma_{\bar{Z}\tilde{X}}\| + \|\hat{M} - M_0\|\big) \\
&+ c^\circ \underline{\sigma}_n^{-2}\|\Sigma_{\bar{Z}}^{-1}\|^2\big(\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| + \|\hat{\Sigma}_{\bar{Z}\tilde{X}} - \Sigma_{\bar{Z}\tilde{X}}\| + \|\hat{M} - M_0\|\big)^2
\end{aligned}$$

Since $n \geq m$ with $3\underline{\sigma}_n\alpha_n/\|\Sigma_{\bar{Z}}^{-1}\| \leq 1$, by (41) we see that $\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| + \|\hat{\Sigma}_{\bar{Z}\tilde{X}} - \Sigma_{\bar{Z}\tilde{X}}\| + \|\hat{M} - M_0\| \leq 1$ and so:

$$\begin{aligned}
&\|\mu_0 - \hat{\mu}\| \\
\leq& c^\circ(1 + \underline{\sigma}_n^{-1}\|\Sigma_{\bar{Z}}^{-1}\| + \underline{\sigma}_n^{-2}\|\Sigma_{\bar{Z}}^{-1}\|^2)\big(\|\hat{\Omega}^{-1/2} - \Omega^{-1/2}\| + \|\hat{\Sigma}_{\bar{Z}\tilde{X}} - \Sigma_{\bar{Z}\tilde{X}}\| + \|\hat{M} - M_0\|\big)
\end{aligned}$$

Finally, for any $a \geq 0$, $1 + a + a^2 \leq 2(1+a^2)$, setting $c = 2c^\circ(1 + c_1^2)$ gives the result. □

Proof of Proposition 3.2. We use results in Bunea *et al.* (2011). Note that the objective and penalty in Bunea *et al.* (2011) are scaled up by a factor of n so the penalty, denoted by μ in Bunea *et al.* (2011), is $n\lambda_{M,n}$ or $n\lambda_{Q,n}$ in our setting.

The argument for part a. is as follows. For a proof of part b. one need only replace M by Q and $\hat{\Sigma}_{\hat{Z}\hat{X}}$ by $\hat{\Sigma}_{\hat{Z}}$ in the steps below.

The following is an immediate application of Theorem 7 in Bunea *et al.* (2011). For any $a > 0$, if $(1+a)r_{M,n}^2 \leq \lambda_{M,n}$ and then:

$$\|(\hat{Z}, \hat{X})(\hat{M} - M_0)'\|_F^2 \leq 2(1 + 2/a)n\lambda_{M,n}d_W$$

Note that:

$$\frac{1}{n}\|(\hat{Z}, \hat{X})(\hat{M} - M_0)'\|_F^2 = \|\hat{\Sigma}_{\hat{Z}\hat{X}}^{1/2}(\hat{M} - M_0)'\|_F^2$$

If $\|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1}\| \leq b$ then using elementary properties of the Frobenius norm the above implies:

$$\|\hat{M} - M_0\|_F^2 \leq \frac{b}{n}\|(\hat{Z}, \hat{X})(\hat{M} - M_0)'\|_F^2$$

By supposition with probability at least $1 - \alpha$, $(1+a)r_{M,n}^2 < \lambda_{M,n}$ and $\|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1}\| \leq b$, so combining the inequalities above we get that with probability at least $1 - \alpha$:

$$\|\hat{M} - M_0\|_F^2 \leq 2b(1 + 2/a)\lambda_{M,n}d_W$$

The following is an immediate application of Theorem 2 in Bunea *et al.* (2011): Suppose that for some $\delta \in (0, 1]$ and $s \leq d_W$ that $\sigma_s\big((\hat{Z}, \hat{X})M_0'\big) > (1+\delta)\sqrt{n\lambda_{M,n}}$ and $\sigma_{s+1}\big((\hat{Z}, \hat{X})M_0'\big) < (1-\delta)\sqrt{n\lambda_{M,n}}$ then (conditional on these events):

$$P\big(rank(\hat{M}) = s\big) \geq 1 - P(r_{M,n} \geq \delta\sqrt{\lambda_{M,n}}) \tag{44}$$

We apply the above with $s = d_W$ and $\delta = (1+a)^{-1/2}$. Under Assumptions 1.1-1.4 M_0 has rank d_W and so $\sigma_{d_W+1}\big((\hat{Z}, \hat{X})M_0'\big) = 0$ and so (assuming $\lambda_{M,n} > 0$) the condition $\sigma_{d_W+1}\big((\hat{Z}, \hat{X})M_0'\big) < (1-\delta)\sqrt{n\lambda_{M,n}}$ holds trivially. As for the other condition, suppose $\|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1}\| \leq b$ and $\lambda_{M,n} < \big(1 + (1+a)^{-1/2}\big)^{-2} b^{-1}\underline{\sigma}_n^2/\|\Sigma_{\bar{Z}}^{-1}\|^2$ then:

$$\sqrt{n}\|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1}\|^{-1/2}\underline{\sigma}_n/\|\Sigma_{\bar{Z}}^{-1}\| > (1+\delta)\sqrt{n\lambda_{M,n}} \tag{45}$$

Recall from the proof of Lemma 3.1 that for two real matrices A and B where B is a square matrix with side length greater than k, $\sigma_k(A)\|(B')^{-1}\|^{-1} \leq \sigma_k(AB')$. In addition, note that $\|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1/2}\|^{-1} = \|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1}\|^{-1/2}$ and $\sigma_{d_W}(M_0) \geq \underline{\sigma}_n/\|\Sigma_{\bar{Z}}^{-1}\|$ (again, see the proof of Lemma 3.1). It follows that:

$$\begin{aligned}\sigma_{d_W}\big((\hat{Z},\hat{X})M_0'\big) &= \sqrt{n}\sigma_{d_W}\big(\hat{\Sigma}_{\hat{Z}\hat{X}}^{1/2}M_0'\big)\\ &\geq \sqrt{n}\|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1}\|^{-1/2}\underline{\sigma}_n/\|\Sigma_{\bar{Z}}^{-1}\|\end{aligned}$$

Using (45) and $\delta = (1+a)^{-1/2}$ we then get:

$$\sigma_{d_W}\big((\hat{Z},\hat{X})M_0'\big) > (1+\delta)\sqrt{n\lambda_{M,n}}$$

As required. So conditional on the events $\|\hat{\Sigma}_{\hat{Z}\hat{X}}^{-1}\| \leq b$ and $\lambda_{M,n} < \big(1+(1+a)^{-1/2}\big)^{-2}b^{-1}\underline{\sigma}_n^2/\|\Sigma_{\bar{Z}}^{-1}\|^2$, 44 holds for our choice of δ and s:

$$\begin{aligned}P\big(rank(\hat{M}) = d_W\big) &\geq 1 - P(r_{M,n} \geq (1+a)^{-1/2}\sqrt{\lambda_{M,n}})\\ &= P\big((1+a)r_{M,n}^2 < \lambda_{M,n}\big)\\ &\geq 1-\alpha_n\end{aligned}$$

Where the final inequality holds by supposition. Since the above holds conditional on events which occur with probability at least $1-\alpha_n$, by the Bonferroni bound we get that, unconditionally:

$$P\big(rank(\hat{M}) = d_W\big) \geq 1 - 2\alpha_n$$

□