Many Proxy Controls

Ben Deaner
(Cowles Foundation, Yale University)
bendeaner@gmail.com

February 21, 2022

Abstract

A recent literature considers causal inference using noisy proxies for unobserved confounding factors. The proxies are divided into two groups that are independent conditional on the confounders. One set of proxies are 'negative control treatments' and the other are 'negative control outcomes'. Existing work applies to low-dimensional settings with a fixed number of proxies and confounders. In this paper we consider high-dimensional linear models with many proxies and possibly many confounders. A key insight is that if each group of proxies is strictly larger than the number of confounding factors this implies rank restrictions on matrices of nuisance parameters. 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 provide asymptotic analysis and provide simulation evidence that our methods are effective.

Introduction and Related Literature

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 subject to substantial measurement error. 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, and even the best measurements of ability (test scores, grades etc.) are likely subject to error. If some confounders are unmeasured or mismeasured then standard methods that adjust for observed covariates do not recover a causal effect.

Miao *et al.* (2018b) sparked a recent literature that considers identification and estimation of causal effects when a researcher observes noisy proxies for unobserved confounding factors. For example, one may observe test scores which are proxies for academic ability. 'Proxy' here refers simply to a covariate that is informative about, but mismeasures, some variable of interest. Two groups of proxies are required and these two sets must be uncorrelated conditional on confounders. One group of proxies is a set of negative control treatments: variables that have no direct causal effect on outcomes. The other group of proxies are negative control outcomes: variables that are not directly affected by the treatments. We refer to these proxies for the unobserved confounders as 'proxy controls'.

Compared to standard factor-analytic methods, the proxy control approach has the advantage that the factor structure itself need not be identified. More precisely, one need not identify the distribution of unobserved confounders nor their causal effects. Thus proxy control methods may be applied even when the assumptions required for identification of the factor structure do not hold.

The proxy control approach is particularly apt for causal inference with high-dimensional data, that is, data that contain many covariates. In these settings one may hope to use these covariates to adjust for confounding. But the availability of many covariates is no guarantee that standard methods recover a causal effect. Despite their numbers, the covariates may still provide only a noisy signal for the underlying confounders. However, the high-dimensional covariates are a rich source of potential proxies.

Miao *et al.* (2018b) provides conditions under which the average structural function is nonparametrically identified using proxy controls. Additional identification results are provided in Deaner (2021) and Tchetgen *et al.* (2020) among others.

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 this work we consider identification, estimation, and inference in linear models when the set of proxy controls and the set of confounding factors are high-dimensional. In high-dimensional settings standard asymptotic approximations based on a fixed number of proxies and confounders may be misleading. 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 group, 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 allow us to adapt to this quantity. The model selection methods are based on techniques from the literature on reduced-rank regression, particularly Bunea *et al.* (2011).

A pleasing feature of our analysis is that the sparsity (in the form of rank restrictions) follows from the structure of the model, and the degree of sparsity is tied directly to an interpretable feature of the model, namely the number of unobserved confounders. This contrasts with standard high-dimensional regression methods which typically assume some form of sparsity on the nuisance parameters a priori.

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.

The use of a linear model allows us to weaken the identifying assumptions in nonparametric proxy control analysis. We need only assume variables are uncorrelated rather than independent, we can replace statistical completeness with more intuitive full-rank assumptions, and we avoid the need for regularity conditions like Assumption A3 in Miao *et al.*

(2018a).

Linear proxy control models have a long history in economics, dating back to work by Zvi Griliches in the 1970s (Griliches & Mason (1972) and Griliches (1977), see also Gary Chamberlain's unpublished PhD dissertation). Griliches (1977) considers two scalar proxies for a single confounding factor in a model that can be understood as a special case of the one we present here. To the best of our knowledge no existing work exploits the dimension reduction when there are fewer confounders than proxies, nor does any existing work allow for a growing number of proxies or confounders.

In sum, our contributions are as follows. We provide a set of identifying assumptions in the linear proxy control model. We present novel estimation methods that allow us to exploit the low-rank structure in the nuisance parameters when the number of unobserved confounders is less than the number of proxies in each group. We develop asymptotic theory for the estimator and an associated inference method, and we provide simulation evidence of the efficacy of our methods.

1 Model and Identification

Let Y_i be an outcome of interest and X_i a vector of treatments. Let W_i be a vector of unobserved confounding factors. We assume the researcher has access to two sets of proxies V_i and Z_i for the unobserved confounders W_i.

The assumptions on the proxies V_i differ from those on the proxies Z_i. V_i is a vector of 'negative control outcomes', which means that treatment has no direct effect on V_i. Z_i is a vector of 'negative control treatments' which means that Z_i has no direct causal effect on the outcome Y_i.

In addition the researcher may have access to a vector of observable confounders which we denote by D_i. Table 1 lists the relevant variables.

Table 1: List of Variables

Variable	Dimension	Description
Y_i	1	Outcome of interest.
X_i	d_X	Vector of treatments.
W_i	d_W	Vector of unobserved confounding factors.
D_i	d_D	Vector of observed confounding factors.
V_i	d_V	Vector of proxies for W (negative control outcomes).
Z_i	d_Z	Vector of proxies for W (negative control treatments).

We assume the following linear models for the scalar outcome Y_i and the vector of negative control outcome proxies V_i. Our interest is in β_0, the vector of coefficients on the treatments X_i in (1), which we assume has some causal interpretation. In order to avoid the need to include intercepts we assume D_i contains a constant.

$$Y_i = \beta_0' X_i + A_0 W_i + L_0 D_i + \varepsilon_i \tag{1}$$

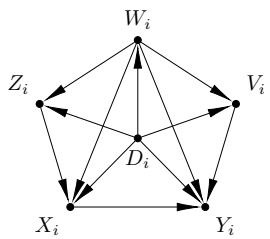
$$V_i = B_0 W_i + R_0 D_i + \upsilon_i \tag{2}$$

Assumption 1.1 (Model and Exclusion restrictions). i. Equations (1) and (2) hold. ii. $E[\varepsilon_i(X_i', Z_i', D_i')] = 0$ and $E[\upsilon_i(X_i', Z_i', D_i')] = 0$.

Assumption 1.1 places conditions on the residuals from equations (1) and (2). In order to justify Assumption 1.1 we suggest applied researchers specify a complete linear causal model for Y_i, V_i, Z_i, X_i, W_i, and D_i, and check whether the exclusion restrictions in that model imply Assumption 1.1. This task is made easier with the use a causal diagrams. Some examples of models in which Assumption 1.1 holds are given in Figure 1.1.

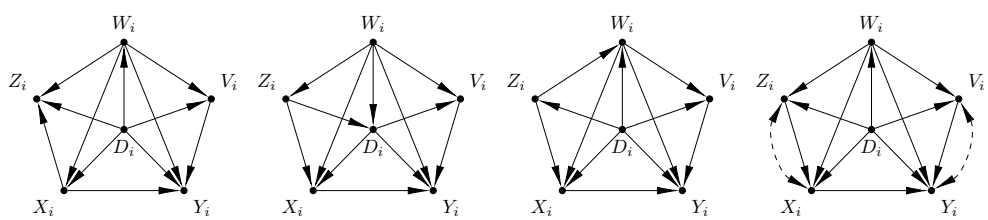
Figure 1.1: Causal Structure of Proxy Controls

(a) Linear Causal Model



- $Y_i = \beta_0' X_i + a_1 V_i + a_2 W_i + a_3 D_i + U_{Y,i}$
- $X_i = a_4 Z_i + a_5 W_i + a_6 D_i + U_{X,i}$
- $V_i = a_7 W_i + a_8 D_i + U_{V,i}$
- $Z_i = a_9 W_i + a_{10} D_i + U_{Z,i}$
- $W_i = a_{11} D_i + U_{W,i}$
- $D_i = U_{D,i}$
- $U_{Y,i}$, $U_{X,i}$, $U_{V,i}$, $U_{Z,i}$, $U_{W,i}$, $U_{D,i}$ all uncorrelated.

(b) Additional Models



The causal diagram in Subfigure 1.1.(a) is associated with the linear causal model to the right of this sub-figure. If there is an arrow pointing from a variable A to a variable B in the diagram, then A is said to be a 'parent' of B. In the corresponding linear causal model, each variable is a linear function of all of its parents. The coefficients on each variable are non-random and the error terms $U_{Y,i}$, $U_{X,i}$, etc. are all uncorrelated with each other.

The diagram is 'causal' in that there is an arrow from A to B if and only if A has a direct causal effect on B (by 'direct' we mean that the effect is not mediated by any included variables), and there are no omitted variables that jointly cause (or 'confound') any of those variables included in the diagram. Thus if a A is excluded from the equation for B, this should be taken to mean that A does not directly cause B. The coefficients β_0, a_1, a_2,...,a_{11} are understood to measure average causal effects. The absence of omitted confounders justifies the lack of correlation in the error terms. If $U_{Y,i}$ and $U_{X,i}$ were correlated this would suggest there exists an omitted joint cause of Y_i and X_i.

The linear causal model in Subfigure 1.1.(a) implies that Assumption 1.1 holds with β_0 the average causal effect of X_i on Y_i. This is not the only linear causal model that implies Assumption 1.1. All of the causal diagrams in Subfigure 1.1.(b) imply Assumption 1.1 with β_0 the causal effect of X_i on Y_i. These examples are not exhaustive. For a given causal diagram one can apply the tools in Pearl (2009) to determine whether Assumption 1.1 holds with β_0 a causal effect. In the rightmost diagram the dashed arrows indicate there are additional variables that jointly cause both Z_i and X_i, and also additional variables that jointly cause both Y_i and V_i.

Some of the causal diagrams in Figure 1.1 are also featured in Miao *et al.* (2018b), Deaner (2021), and elsewhere.

Before we state our first result let us introduce some additional notation. We will define variables with the observed confounders D_i and treatments X_i partialled out. Define the following objects for $H = W, V, Z, Y, X$:

$$\begin{aligned}
\gamma_{H,0} &= E[D_i D_i']^+ E[D_i H_i'] \\
\omega_{H,0} &= E[(X_i', D_i')'(X_i', D_i')]^+ E[(X_i', D_i')' H_i'] \\
\tilde{H}_i(\gamma) &= H_i - \gamma' D_i \\
\bar{H}_i(\omega) &= H_i - \omega'(X_i', D_i')'
\end{aligned}$$

The notation M^+ denotes the Moore-Penrose pseudo-inverse of the matrix M. For notational convenience we sometimes write $\tilde{H}_i = \tilde{H}_i(\gamma_{H,0})$ and $\bar{H}_i = \bar{H}_i(\omega_{H,0})$. So for example, $\tilde{X}_i$ is X_i with D_i partialled out and $\bar{Z}_i$ is Z_i with both D_i and X_i partialled out.

It will also be useful to define the matrices C_0, G_0, which are of dimensions $d_W \times d_Z$ and $d_W \times d_X$ respectively and are given below.

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

Note that (C_0, G_0) is a block matrix consisting of C_0 concatenated horizontally with G_0. We use this notation thoughout.

Theorem 1. *Under Assumption 1.1 the following moment conditions hold:*

$$E\left[\left(\begin{pmatrix} \tilde{V}_i \\ \tilde{Y}_i - \beta_0' \tilde{X}_i \end{pmatrix} - \begin{pmatrix} B_0 C_0 & B_0 G_0 \\ A_0 C_0 & A_0 G_0 \end{pmatrix} \begin{pmatrix} \tilde{Z}_i \\ \tilde{X}_i \end{pmatrix}\right)(\tilde{Z}_i', \tilde{X}_i')\right] = 0 \tag{3}$$

Theorem 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 identify β_0 as well as the number of confounding factors d_W.

Assumption 1.2 (V_i is sufficiently informative about W_i). B_0 has full column rank.

Assumption 1.3 (Z_i is sufficiently informative about W_i). C_0 has full row rank.

Assumption 1.4 (Full support). The matrix $E[(X_i', Z_i', D_i')'(X_i', Z_i', D_i')]$ is non-singular.

Assumption 1.2 requires that the vector V_i is a sufficiently informative proxy for the confounders W_i. The assumption replaces the statistical completeness condition on V_i required in the nonparametric setting. If $E[\upsilon_i W_i] = 0$ and Assumption 1.4 holds, this assumption is equivalent to full row rank of $E[\bar{W}_i \bar{V}_i']$. This is precisely the rank condition for identification in linear instrumental variables (IV) estimation: W_i takes the role of the endogenous regressors, D_i and X_i take the role of the exogenous regressors, and V_i acts as a vector of instruments.

Similarly, Assumption 1.3 requires that after accounting for the treatments and observed controls, Z_i is sufficiently informative about the confounders. Under Assumption 1.4 the condition is equivalent to full row rank of $E[\bar{W}_i \bar{Z}_i']$. Again, this is the same condition required for identification in a linear IV model in which Z_i is a vector of instruments for W_i, and the variables X_i and D_i are exogenous regressors.

Note that Assumptions 1.2 and 1.3 imply an order condition: Z_i and V_i must each have weakly larger dimension than W_i.

Assumption 1.4 is a very mild condition that none of the components of X_i, Z_i, and D_i are perfectly co-linear.

Theorem 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 B has full column rank:*

$$E\left[\left(\begin{pmatrix} \tilde{V}_i \\ \tilde{Y}_i - \beta'\tilde{X}_i \end{pmatrix} - \begin{pmatrix} BC & BG \\ AC & AG \end{pmatrix}\begin{pmatrix} \tilde{Z}_i \\ \tilde{X}_i \end{pmatrix}\right)(\tilde{Z}_i', \tilde{X}_i')\right] = 0 \tag{4}$$

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 $AG = A_0G_0$.

Theorem 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. 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.[1]

The theorem shows that the number of confounding factors d_W determines the rank of some matrices of nuisance parameters in the moment condition. If d_W is small then this constraint on the rank constitutes a substantial dimension reduction in the nuisance parameters, 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.

A subtle point in the theorem is the condition that B, like B_0, has full column rank. There could be a β, A, B, C, and G that satisfy (4) so that $\beta \neq \beta_0$, but then B must not have full column rank. It would also be sufficient to impose that C has full row rank.

Under Assumptions 1.1-1.4, β_0 could be estimated directly from the moment conditions in Theorem 1 using the Generalized Method of Moments (GMM) (Hansen (1982)). However, this presents some computational difficulty. Suppose that d_W were known and we apply GMM enforcing one of 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.

In order to avoid this computational problem we suggest a method to 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 and then in a second-stage 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 this characterization of β_0 is in fact equivalent to that in Theorem 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.

[1] 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 (4) 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.

Corollary 1. *Under Assumptions 1.1-1.4 β_0 is identified from the moment conditions below:*

$$E\big[(\tilde{V}_i - M_0(\tilde{Z}_i', \tilde{X}_i')')(\tilde{Z}_i', \tilde{X}_i')\big] = 0 \tag{5}$$

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

$M_0 = B_0(C_0, G_0)$ is the unique solution to (5) and $rank(M_0) = d_W$. (6) is satisfied by any ξ_0 that solves $\xi_0 B_0 C_0 = A_0 C_0$, and there exists a solution with $||\xi_0||_0 \leq d_W$. B_0C_0 and A_0C_0 (and thus the set of solutions ξ_0) are identified by the moment conditions below, which have unique solution $Q_0 = (B_0', A_0')'C_0$ and $rank(Q_0) = d_W$:

$$E\big[((\bar{V}_i', \bar{Y}_i')' - Q_0\bar{Z}_i)\bar{Z}_i\big] = 0 \tag{7}$$

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, d_W is the rank of Q_0. Thirdly, there is a ξ_0 that satisfies the moment conditions with d_W non-zero entries.

1.1 Comparison to Existing Results

Our results are related to those of Miao *et al.* (2018a) and Griliches (1977). However, our results differ in important respects. We link the rank of the nuisance parameter matrices to the number of unobserved confounders d_W which we show is identified, and we provide additional moment conditions that help identify β_0 when d_W is smaller than d_V or d_Z. The rank restrictions can result in a substantial dimension reduction which can greatly reduce estimation error, particularly when there are many available proxies.

First let us compare with Miao *et al.* (2018a). For simplicity let us assume there are no observed confounders D_i. 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_i, with probability 1:

$$E[Y_i|W_i, X_i = x] = E[b(V_i, x)|W_i, X_i = x]$$

Suppose our Assumptions 1.1-1.4 hold and ϵ_i and υ_i are mean independent of W_i (rather than just uncorrelated with W_i), then 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 $\xi_0 B_0 C_0 = A_0 C_0$.[2] Even if the confounding bridge is unique, in order to identify the bridge, Z_i must be a vector relevant instruments for V_i after controlling for X_i (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

[2] 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, this is generically false.

to d_W.[3] Griliches (1977) suggests estimation using instrumental variables that amounts to the use of the moment condition (6) but the analysis of Griliches (1977) is limited to scalar proxies and unobserved confounders. Corollary 1 imposes an additional moment condition (5) which is important when d_W is smaller than the number of proxies. For some intuition note that (5) and (6) together imply the following moment condition:

$$E\big[(\tilde{Y}_i - \beta_0 \tilde{X}_i - \xi_0 M_0(\tilde{Z}_i', \tilde{X}_i'))(\tilde{Z}_i', \tilde{X}_i')\big] = 0$$

In the moment condition above ξ_0' can be replaced by its projection onto the d_W-dimensional column space of M_0, and thus ξ_0 is effectively of dimension d_W rather than d_V. This is related to the result in Corollary 1 that there exists a sparse solution ξ_0 which has no more than d_W non-zero entries. The additional moment conditions in Corollary 1 suggest multiple means of estimating, and thus adapting to d_W.

In sum, our analysis, unlike existing results, applies to cases in which the number of proxies in V_i is strictly greater than the number of unobserved confounders. Moreover, our results show how one can adapt to the unknown number of confounders d_W and thus reduce the number of free nuisance parameters to be estimated when d_W is smaller than d_V or d_Z.

2 Estimation and Inference

We now present an estimator motivated by the results in Corollary 1. In a first stage one estimates nuisance parameters M_0, ξ_0, the parameters involved in partialling out D_i, and an additional matrix μ_0 which we introduce in Subsection 3.1. μ_0 is used to orthoganize the moment conditions (5) and (6) to the nuisance parameters so that the resulting moment condition is doubly robust. In a second stage we plug the nuisance parameter estimates into an empirical doubly robust moment condition and solve for an estimate of β_0.

The estimator is an example of a DML 2 (Double Machine Learning 2) estimator as developed in Chernozhukov *et al.* (2018). Following Chernozhukov *et al.* (2018) we use sample-splitting to reduce bias.

In order to adapt to the number of latent confounding factors d_W, we estimate M_0 by penalized reduced-rank regression and likewise for Q_0, which we then use to obtain an estimate of ξ_0. These procedures produce an estimate of d_W as a byproduct. In the appendix we also specify an alternative method for estimating ξ_0 which uses the fact (stated in Corollary 1) that there is a sparse solution ξ_0 with at most d_W non-zero entries.

The estimator has a closed-form. In Subsection 3.1 we specify the doubly-robust moment condition and our second stage estimator. In Subsection 3.2 we present estimates of the nuisance parameters. Subsection 3.3 specifies confidence intervals ans standard errors.

Let us introduce some additional notation. We assume we have access to a sample of size n, $\{Y_i, X_i, Z_i, V_i, D_i\}_{i=1}^n$. We let X be the matrix whose i^{th} row is X_i' and similarly for Z, V, Y, and D. 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 for $M_{[a,c:d]} = M_{[a:a,c:d]}$.

[3] Miao *et al.* (2018a) allow the instruments (Z_i', X_i') to be replaced with any vector of transformations $q(Z_i, X_i)$ 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 Doubly-Robust Score and Second-Stage Estimator

We now define an estimator of β_0 which uses nuisance parameter estimates obtained in a first stage. We develop methods for obtaining the nuisance parameter estimates in the subsequent subsection. The estimator is a DML 2 estimator of the kind in Chernozhukov *et al.* (2018). The estimator solves an empirical moment condition with a doubly robust score function. The score function is motivated by the moment conditions (5) and (6) in Corollary 1.

For notational convenience let us collect the parameters involved in partialling out D_i into a single parameter $\gamma_0 = (\gamma_{0,X}, \gamma_{0,Z}, \gamma_{0,X}, \gamma_{0,Z}, \gamma_{0,V}, \gamma_{0,Y})$. Note that we duplicate $\gamma_{0,X}$ and $\gamma_{0,Z}$. This is because $\gamma_{0,X}$ and $\gamma_{0,Z}$ appear in the moment conditions in two places and it is useful for analytical purposes to treat these instances as different parameters.

To define the doubly robust score, let us define the vector-valued random function g_i to be the score function from the moment condition (6):

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

In the above $\gamma = (\gamma_{X,1}, \gamma_{Z,1}, \gamma_{X,2}, \gamma_{Z,2}, \gamma_V, \gamma_Y)$. We can then rewrite (6) as follows:

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

The moment condition above is not doubly robust. If ξ_0 is replaced by some choices of $\xi \neq \xi_0$ the condition no longer holds.

However, the moment conditions are robust to each component of γ_0. For example suppose we replace γ_0 by $\gamma = (\gamma_{X,1}, \gamma_{0,Z}, \gamma_{0,X}, \gamma_{0,Z}, \gamma_{0,V}, \gamma_{0,Y})$ for some $\gamma_{X,1} \neq \gamma_{0,X}$, the moment conditions still hold, that is $E\big[g_i(\beta_0; M_0, \xi_0, \gamma)\big] = 0$.

Estimation of β_0 is based on a moment condition with a doubly robust score function ψ_i of the form below:

$$\psi_i(\beta_0; \xi_0, \gamma_0, \mu_0) = \mu_0 g_i(\beta_0; \xi_0, \gamma_0) \tag{8}$$

If the moment conditions in Corollary 1 hold then $E\big[\psi_i(\beta_0; \xi_0, \gamma_0, \mu_0)\big] = 0$. The doubly robust score depends on an additional matrix of nuisance parameters μ_0 of dimension $d_X \times (d_Z + d_X)$. To define μ_0 let us first define a matrix G_η.

$$G_\eta = -E[(\tilde{Z}_i', \tilde{X}_i')'(\tilde{Z}_i', \tilde{X}_i')]M_0'$$

Note that G_η is the matrix of derivatives of $E[g_i(\beta_0; \xi_0, \gamma_0)]$ with respect to ξ_0.

Let G_β be some $(d_Z + d_X) \times d_X$ matrix and Ω a non-singular $(d_Z + d_X) \times (d_Z + d_X)$ matrix. We then define μ_0 by:

$$\mu_0 = G_\beta' \Omega^{-1} - G_\beta' \Omega^{-1} G_\eta (G_\eta' \Omega^{-1} G_\eta)^+ G_\eta' \Omega^{-1} \tag{9}$$

With μ_0 defined in this way the score function (8) is doubly robust. We take Ω to be the variance-covariance matrix of $g_i(\beta_0; M_0, \xi_0, \gamma_0)$ and set G_β as follows:

$$G_\beta = -E\big[(\tilde{Z}_i', \tilde{X}_i')' \tilde{X}_i'\big]$$

With μ_0 defined as above, the score (8) is efficient in the sense that it yields a GMM estimate that has smallest asymptotic variance out of all GMM estimators based on a linear

combination of the components of g_i (see e.g., subsection 2.2.2 of Chernozhukov *et al.* (2018) for discussion).

An estimate of β_0 is obtained by solving the empirical analogue of the moment condition $E\big[\psi_i(\beta_0;\xi_0,\gamma_0,\mu_0)\big] = 0$. Recent work including Chernozhukov *et al.* (2016) and Chernozhukov *et al.* (2018) shows that there are advantages to sample-splitting in doubly robust and locally robust estimators. Before we specify the estimator that employs sample-splitting it may be helpful to first describe an estimate that does not use sample splitting.

Let $\hat{\xi}$, $\hat{\gamma}$, and $\hat{\mu}$ be estimates of ξ_0, γ_0, and μ_0. $\hat{\gamma}$ can be further decomposed into $\hat{\gamma}_X$, $\hat{\gamma}_Z$, $\hat{\gamma}_V$, and $\hat{\gamma}_Y$, which estimate $\gamma_{0,X}$, $\gamma_{0,Z}$, $\gamma_{0,V}$, and $\gamma_{0,Y}$. We let $\hat{X}_i = \tilde{X}_i(\hat{\gamma}_X)$ and similarly for $\hat{Z}_i$, $\hat{V}_i$, and $\hat{Y}_i$.

An estimate of β_0 that does not employ sample-splitting solves the empirical moment condition $\sum_{i=1}^n \psi_i(\beta;\hat{\xi},\hat{\gamma},\hat{\mu}) = 0$. This estimate $\hat{\beta}$ has the following formula:

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

The estimator with sample splitting is similar to the above. We partition 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 jth subsample. Thus, for each j, one obtains estimates $\hat{\xi}_j$, $\hat{\mu}_j$, and $\hat{\gamma}_j$ of ξ_0, μ_0, and γ_0. The estimate $\hat{\beta}$ with sample-splitting solves the equation below:

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

In the formula for $\hat{\beta}$, which is given below, $\hat{X}_{j,i} = \tilde{X}_i(\hat{\gamma}_{j,X})$ where $\hat{\gamma}_{j,X}$ is the component of $\hat{\gamma}_j$ that estimates $\gamma_{0,X}$, and similarly for $\hat{Z}_{j,i}$, $\hat{V}_{j,i}$, and $\hat{Y}_{j,i}$.

$$\hat{\beta} = \big(\sum_{j=1}^J \sum_{i\in\mathcal{I}_j} \hat{X}_{j,i}(\hat{Z}_{j,i}',\hat{X}_{j,i}')\hat{\mu}_j'\big)^+ \sum_{j=1}^J \sum_{i\in\mathcal{I}_j} \hat{\mu}_j(\hat{Z}_{j,i}',\hat{X}_{j,i}')'(\hat{Y}_{j,i} - \hat{\xi}\hat{V}_{j,i}') \tag{10}$$

2.2 Nuisance Parameter Estimates

We now present estimates of the nuisance parameters ξ_0, μ_0, and γ_0 which can then be plugged into the second stage estimator (10). The doubly-robust estimator with sample-splitting requires that for each $j = 1,...,J$ we estimate the nuisance parameters using only data outside the jth sub-sample i.e., with indices in $\mathcal{I}_{-j}$. Here we describe estimators that use the whole sample but these can easily adapted to the sample-splitting case by dropping the data with indices in $\mathcal{I}_j$.

First let us consider estimators for γ_0 which is composed of $\gamma_{0,X}$, $\gamma_{0,Z}$, $\gamma_{0,V}$, and $\gamma_{0,Y}$. If the vector of additional covariates D_i is relatively low-dimensional then we can use ordinary

least-squares. For $H = V$, Z, Y, X we estimate $\gamma_{0,H}$ as follows:

$$\begin{aligned} \hat{\gamma}_H &= (D'D)^+ D'H \\ \hat{H}_i &= \tilde{H}_i(\hat{\gamma}_H) \end{aligned} \tag{11}$$

In some cases D_i may be high-dimensional. In addition, we may believe that only a subset of these covariates are linearly predictive of V_i, Z_i, X_i, and Y_i. In this case $\gamma_{H,0}$ may be sparse or approximately sparse for some $H \in \{V, Z, X, Y\}$.

To exploit this sparsity or approximate sparsity we suggest Lasso regression (Tibshirani (1996)) instead of ordinary least squares. For $H = V$, Z, X, Y let $\lambda_{H,n,\gamma}$ be a scalar penalty parameter and define the Lasso estimate $\hat{\gamma}_H$ as follows:

$$\hat{\gamma}_H = \underset{\gamma \in \mathbb{R}^{d_D}}{\operatorname{argmin}} \sum_{i=1}^{n} (H_i - \gamma D_i)^2 + \lambda_{H,n,\gamma} ||\gamma||_1$$

Where $|| \cdot ||_1$ is the ℓ_1-norm. A number of methods exist for choosing the penalty parameters in Lasso regression, for example cross-validation.

We also define variables with both D_i and X_i partialled out in the sample. For $H = Z$, V, Y we define the following.

$$\begin{aligned} \hat{\omega}_H &= \big((X, D)'(X, D)\big)^+ (X, D)'H \\ \check{H}_i &= \bar{H}_i(\hat{\omega}) \end{aligned}$$

Again, if D_i is high-dimensional we could use Lasso to estimate $\hat{\omega}_H$:

$$\hat{\omega}_H = \underset{\omega \in \mathbb{R}^{d_D}}{\operatorname{argmin}} \sum_{i=1}^{n} (H_i - (X_i', D_i')\omega)^2 + \lambda_{H,n,\omega} ||\omega||_1$$

Corollary 1 states that M_0 and Q_0 are the unique solutions to the moment conditions (5) and (7). 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. Thus if $r \geq d_W$ then the matrices M_0 and Q_0 are the solutions to the constrained least-squares problems below.

$$\begin{aligned} M_0 &= \underset{rank(M) \leq r}{\operatorname{argmin}} \; E\big[||\tilde{V}_i - M(\tilde{Z}_i', \tilde{X}_i')'||^2\big] \\ Q_0 &= \underset{rank(Q) \leq r}{\operatorname{argmin}} \; E\big[||(\tilde{V}_i', \bar{Y}_i)' - Q\bar{Z}_i||^2\big] \end{aligned}$$

To estimate M_0 and Q_0 given a value of $r \geq d_W$ we can minimize empirical analogues of the above. Instead of an expected sum of squares we use the sample expectation and we partial out D_i and X_i using the data. An estimate $\hat{M}_r$ of M_0 and $\hat{Q}_r$ of Q_0 are given below.

$$\hat{M}_r = \underset{rank(M) \leq r}{\operatorname{argmin}} \; ||\hat{V} - (\hat{Z}, \hat{X})M'||_F^2 \tag{12}$$

$$\hat{Q}_r = \underset{rank(Q) \leq r}{\operatorname{argmin}} \; ||(\check{V}, \check{Y}) - \check{Z}Q'||_F^2 \tag{13}$$

Where $||\cdot||_F^2$ is the squared Frobenius norm (the sum of the squared entries of the matrix).

$\hat{M}_r$ and $\hat{Q}_r$ are reduced-rank regression estimates and thus have closed-form solutions (Reinsel & Velu (1998), Izenman (1975)). The formulas 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:

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

Then we have:

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

Note that r 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 we can adapt to this quantity.

In order to adapt to the unknown number of confounders d_W we replace the constrained least-squares problems (12) and (13) with unconstrained penalized least-squares problems as follows:

$$\hat{M} = \underset{rank(M)\leq d_Z}{\text{argmin}} ||\hat{V} - (\hat{Z},\hat{X})M'||_F^2 + \lambda_{M,n} rank(M), \tag{14}$$

$$\hat{Q} = \underset{rank(Q)\leq d_V}{\text{argmin}} ||(\check{V},\check{Y}) - \check{Z}Q'||_F^2 + \lambda_{Q,n} rank(Q) \tag{15}$$

$\lambda_{M,n}$ and $\lambda_{Q,n}$ are positive scalars that control the degree of regularization.

One could replace the rank penalties with some other penalty that induces a low-rank structure. For example, instead of $rank(M)$ in (14) we could instead use the nuclear norm of M, commonly denoted $||M||_*$.[4] Penalizing the rank has the advantage that the solution has a closed-form.

Bunea *et al.* (2011) provide the formula for the solution to a penalized reduced-rank regression problem like (14) and (15). 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. In particular, $\hat{r}_M$ is whichever is smaller: d_Z or the number of eigenvalues of the matrix $\hat{V}'(\hat{Z},\hat{X})\hat{\Sigma}^+_{\hat{Z}\hat{X}}(\hat{Z},\hat{X})'\hat{V}$ that exceed $\lambda_{M,n}$. Similarly $\hat{r}_Q$ is the minumum of d_V and the number of eigenvalues of $(\check{V},\check{Y})'\check{Z}\hat{\Sigma}^+_{\check{Z}}\check{Z}'(\check{V},\check{Y})$ that exceed $\lambda_{Q,n}$.

The penalty parameters $\lambda_{M,n}$ and $\lambda_{Q,n}$ can be chosen in a number of ways. Bunea *et al.* (2011) suggest plug-in formulas that are motivated by the assumption that the regression residuals are independent and normally distributed. In our simulations we choose the penalty parameters by cross-validation.

Corollary 1 states that ξ_0 solves $\xi_0 B_0 C_0 = A_0 C_0$. B_0C_0 and A_0C_0 sub-matrices of Q_0. If we replace B_0C_0 and A_0C_0 with the corresponding sub-matrices of $\hat{Q}$ the resulting equation is $\xi_0\hat{Q}_{[1:d_V,:]} = \hat{Q}_{[d_V+1,:]}$. Our estimate $\hat{\xi}$ is the solution with smallest Euclidean norm and has the following formula:

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

$$\hat{\xi} = \hat{Q}_{[d_V+1,:]}\hat{Q}'_{[1:d_V,:]}(\hat{Q}_{[1:d_V,:]}\hat{Q}'_{[1:d_V,:]})^+$$

For the estimate of μ_0 we replace the objects in (9) with their sample analogues. An estimate G_η is given below:

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

The estimate of μ_0 is then given by:

$$\hat{\mu} = \hat{G}'_\beta\hat{\Omega}^{-1} - \hat{G}'_\beta\hat{\Omega}^{-1}\hat{G}_\eta(\hat{G}'_\eta\hat{\Omega}^{-1}\hat{G}_\eta)^+\hat{G}'_\eta\hat{\Omega}^{-1}$$

$\hat{G}_\beta$ and $\hat{\Omega}$ are estimates of G_β and Ω. G_β can be estimated by its sample analogue below:

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

The efficient choice of Ω is the variance matrix of $g_i(\beta_0; \xi_0, \gamma_0)$. Let $\tilde{\beta}$ be an initial estimate of $\hat{\beta}$ that uses the identity in place of $\hat{\Omega}$. We can then estimate the efficient Ω by letting $\hat{\Omega}$ be the sample variance-covariance matrix of $g_i(\tilde{\beta}; \hat{\xi}, \hat{\gamma})$.

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:

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

The matrix $\hat{S}$ is defined below:

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

Note that the above is estimate of $S_0 = \mu_0 E[(\tilde{Z}'_i, \tilde{X}'_i)'\tilde{X}'_i]$. For each $j \in 1, ..., J$ and $i \in \mathcal{I}_j$ we define $\hat{\psi}_i = \psi_i(\hat{\beta}; \hat{\xi}_j, \hat{\gamma}_j, \hat{\mu}_j)$.

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:

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

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.

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 asymptotically normal centered at the true parameter under relatively weak conditions on the rates at which the nuisance parameters converge.

We now present high-level assumptions for root-n consistency and asymptotic normality of the doubly-robust estimator with sample splitting as defined in (**??**). Note that our results apply for any choice of estimators for the nuisance parameters not just those specified in Section 2.

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 some parameter space, we suppose that for each sample size n, the data generating process, denoted by P, belongs to some set $\mathcal{P}_n$. The Assumptions below then restrict $\mathcal{P}_n$.

It is helpful to introduce some additional notation. For any random column vector H_i we let $\Sigma_H = E[H_i H_i']$, however in the case of $H_i = (\tilde{Z}_i', \tilde{X}_i')'$ we write $\Sigma_{\tilde{Z}\tilde{X}}$. If b is a vector then $||b||$ is the Euclidean norm of b. If A is a matrix then $||A|| = \sup_{b \in \mathbb{R}^d : ||b||=1} ||Ab||$. For sequences a_n and b_n the notation $a_n \precsim b_n$ means that there exists a constant C so that $a_n \leq C b_n$ for all n. $a_n \prec b_n$ means that $a_n \precsim b_n$ but not $b_n \precsim a_n$. Finally, we define $\tilde{\epsilon}_i$ as follows:

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

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. $||(\hat{\mu}_j - \mu_0)\Sigma_{\tilde{Z},\tilde{X}}^{1/2}|| \precsim \delta_\mu$. ii. $||\Sigma_{\tilde{V}}^{1/2}(\hat{\xi}_j - \xi_0)|| \leq \delta_\xi$. iii. For $H = X, V$, $||\Sigma_{\tilde{H}}^{-1/2}(\hat{\gamma}_{H,j} - \gamma_{H,0})\Sigma_D^{1/2}|| \precsim \delta_{\gamma,H}$ and in addition:

$$||\Sigma_{\tilde{Z},\tilde{X}}^{-1/2}(\hat{\gamma}_{Z,j}' - \gamma_{Z,0}', \hat{\gamma}_{X,j}' - \gamma_{X,0}')'\Sigma_D^{1/2}|| \precsim \delta_{\gamma,Z} + \delta_{\gamma,X}$$
$$||\Sigma_D^{1/2}\big((\gamma_Y - \gamma_{0,Y}) - (\gamma_V - \gamma_{0,V})\xi_0' - (\gamma_X - \gamma_{0,X})\beta_0\big)|| \precsim \delta_{\gamma,\epsilon}$$

Assumption 3.2 (Restrictions on the DGP). If $P \in \mathcal{P}_n$ the following hold: i. $E[\tilde{\epsilon}_i^2] \precsim 1$, $||\Sigma_{\tilde{X}}|| \precsim 1$. ii. There is a constant $q > 2$ so that for each $H_i \in \{\tilde{X}_i, \tilde{V}_i, (\tilde{Z}_i', \tilde{X}_i')', D_i, \tilde{\epsilon}_i\}$:

$$E\big[||\Sigma_H^{-1/2} H_i D_i' \Sigma_D^{-1/2}||^q\big]^{1/q} \precsim \sqrt{d_H d_D}$$

Where $\Sigma_H = E[H_i H_i']$ and d_H is the length of H_i. Similarly:

$$E\big[||\Sigma_{\tilde{V}}^{-1/2}\tilde{V}_i(\tilde{Z}_i', \tilde{X}_i')\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}||^q\big]^{1/q} \precsim \sqrt{d_V(d_X + d_Z)}$$

iii. There is a constant c so that for each $H_i \in \{\tilde{V}_i, (\tilde{Z}_i', \tilde{X}_i')'\}$, $||E[\Sigma_H^{-1/2} H_i H_i' \Sigma_H^{-1/2}|D_i]|| \leq c$, and for each $H_i \in \{\tilde{V}_i, \tilde{\epsilon}_i\}$, $||E[\Sigma_H^{-1/2} H_i H_i' \Sigma_H^{-1/2}|\tilde{Z}_i, \tilde{X}_i]|| \leq c$. In both cases $\Sigma_H = E[H_i H_i']$.

iv. $||\Sigma_{\tilde{X}}^{1/2}\beta_0||$, $||\Sigma_{\tilde{X}}^{1/2}\xi_0'||$, $||\mu_0\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||$, and $||\Sigma_{\tilde{Z}\tilde{X}}^{1/2}M_0'\Sigma_{\tilde{V}}^{-1/2}||$ are all uniformly bounded above and below away from zero. v. $E\big[||\psi_i(\beta_0;\xi_0,\gamma_0,\mu_0)||^q\big]^{1/q} \precsim 1$, $E\big[||\mu_0(\tilde{Z}_i',\tilde{X}_i')'\tilde{X}_i'||^q\big]^{1/q} \precsim 1$.

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.ii requires that some higher-order moments exist. 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. Assumption 3.2.iii states that some conditional moments are almost surely bounded by a fixed constant.

To motivate the rates in Assumption 3.2.ii note that:

$$E\big[||\Sigma_H^{-1/2}H_iD_i'\Sigma_D^{-1/2}||^q\big]^{1/q} \leq \sqrt{E\big[\big(||\Sigma_H^{-1/2}H_i||^2\cdot||\Sigma_D^{-1/2}D_i||^2\big)^q\big]^{1/q}}$$

The term $||\Sigma_H^{-1/2}H_i||^2\cdot||\Sigma_D^{-1/2}D_i||^2$ can be written as a sum of d_Dd_H scalar random variables. Thus the RHS above is the square-root of the L_q norm of the sum of d_Dd_H random scalars. This is bounded by the square root of the sum of the L_q norms of the d_Dd_H random variables. Thus, if the norms of each of these scalars is uniformly bounded we get a rate $\sqrt{d_Dd_H}$.

Theorem 3. *Suppose that for each n, $P \in \mathcal{P}_n$ so that Assumptions 1.1-1.4, 3.1, and 3.2 hold. In addition, suppose that the singular values of S_0 are bounded uniformly below and away from zero, and the eigenvalues of $E[\psi_i\psi_i']$ are bounded uniformly above and below away from zero.*

Moreover, suppose that the following conditions hold, $\delta_\mu \precsim d_X^{-1/2}(d_X+d_Z)^{-1/2}$, $\delta_\xi \precsim (d_X+d_Z)^{-1/2}d_V^{-1/2}$, $\delta_{\gamma,Z} \precsim d_X^{-1/2}d_D^{-1/2}$, and:

$$\delta_{\gamma,X},\delta_{\gamma,\epsilon},\delta_{\gamma,V}\delta_\xi \precsim (d_X+d_Z)^{-1/2}d_D^{-1/2}$$

Moreover, suppose that $\delta_\mu\delta_\xi \prec n^{-1/2}$,$(\delta_{\gamma,X}+\delta_{\gamma,Z})(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi) \prec \min\{n^{-1/2},d_D^{-1}\}$ and:

$$\delta_\mu,\delta_\xi,\delta_{\gamma,\epsilon},\delta_{\gamma,V}\delta_\xi,\delta_{\gamma,Z},\delta_{\gamma,X} \prec d_X^{-1/2}$$

Then uniformly over all $P \in \mathcal{P}_n$, $\hat{\beta}$ is root-n consistent and asymptotically normal:

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

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

Theorem 3 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. In effect, these additional conditions restrict the rate at which the dimensions of the variables can grow with the sample size.

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

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:

$$
\begin{aligned}
V_i &= B_0 W_i + \upsilon_i \\
X_i &= T_0 W_i + \epsilon_i \\
Z_i &= C_0 W_i + G_0 X_i + \eta_i \\
Y_i &= X_i' \beta_0 + F_0 W_i + \chi_0 V_i + e_i
\end{aligned}
$$

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, C_0, G_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, G_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 C_0 are generated so that their non-zero singular values are equal to s. Let N_1 be a $d_V \times d_V$ matrix of independent random normals and N_2 be a $d_W \times d_W$ matrix of random normals. We set $B_0 \sim s(N_1' N_1)^{-1/2} N_1 (I, 0)' N_2' (N_2' N_2)^{-1/2}$ and similarly for C_0.

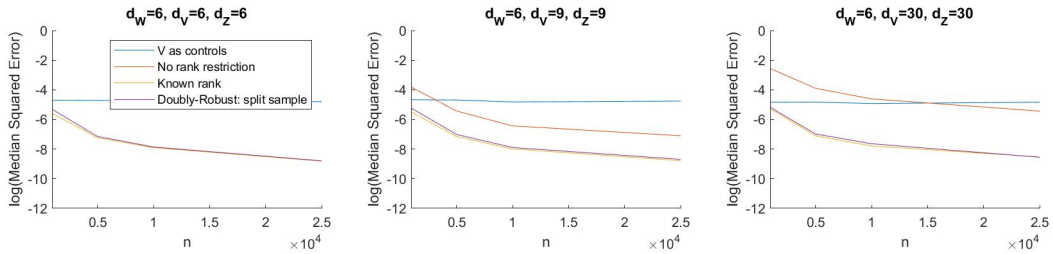
The covariance matrices have a re-scaled inverse Wishart distribution, for example $d_V p \Sigma_V^{-1} \sim W_{d_V}(I, d_V p)$. The natural number p is a hyper-parameter that determines the degrees of freedom of the Wishart distribution.

We are left with hyperparameters s, 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 s, d_W, d_V, and n.

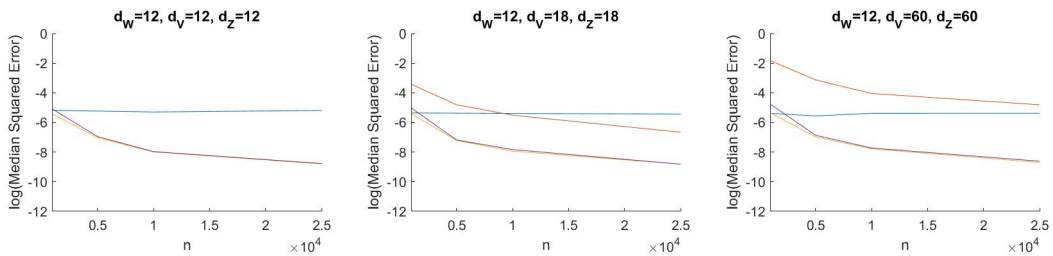
Figure 4.1 shows the median-squared errors of alternative estimators for a variety of different hyperparameters. In all cases in Figure 4.1 we set $s = 1$. The estimators that are compared are: (in blue) a naive least-squares estimator that simply treats V_i as a set of controls, (in red) the proxy control estimator with no rank restriction, (in yellow) an infeasible estimator that imposes the rank restriction d_W, and (in purple) our doubly-robust estimator.[5]

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

[5] 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.



(a) $d_W = 6$



(b) $d_W = 12$

Figure 4.1: Simulated Median Squared Errors, $s = 1$

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.

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.

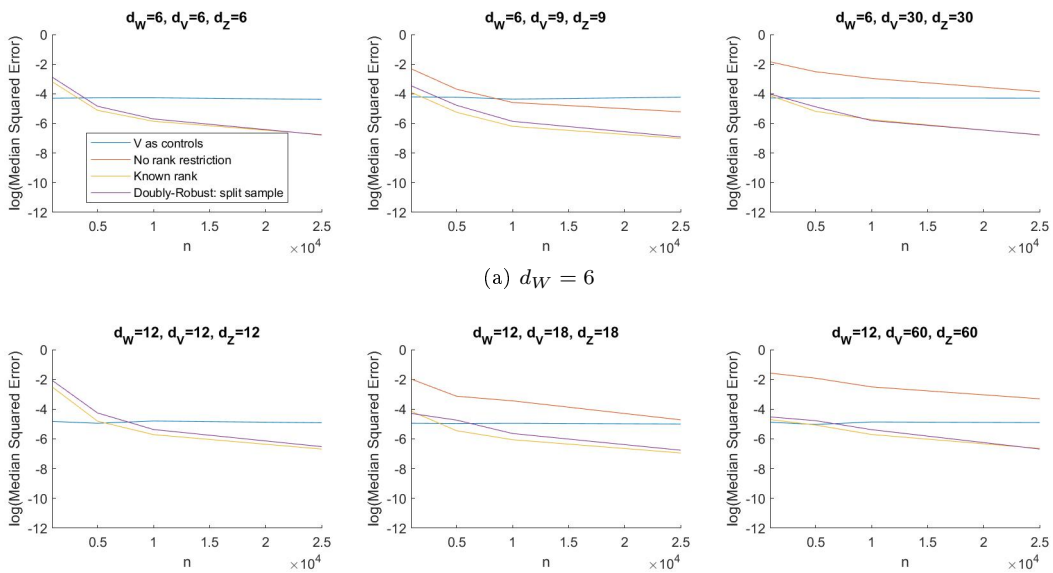
The proxy control estimator with no rank restriction (in red) is equivalent to the two-stage least squares strategy of Griliches (1977) in which V_i is a vector of endogenous regressors, X_i is a vector of exogenous regressors, and Z_i is a vector of instruments. When the number of proxies in each group 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, 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 12 confounders and 60 proxies in each group, the naive estimator outperforms the proxy estimator with no rank restrictions. 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 $s = 0.5$. Recall that s is the level of the singular values of B_0 and C_0, and thus controls the informativeness of the

proxies relative to noise levels. A smaller value of s is thus likely to be less favorable for our analysis. Indeed, all of the estimators perform worse in this setting (including the naive estimator) and the proxy estimators perform 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 $s = 1$, the estimator that does not impose a rank restriction performs substantially worse when there are many proxies compared to the number of unobserved confounders.



(a) $d_W = 6$

(b) $d_W = 12$

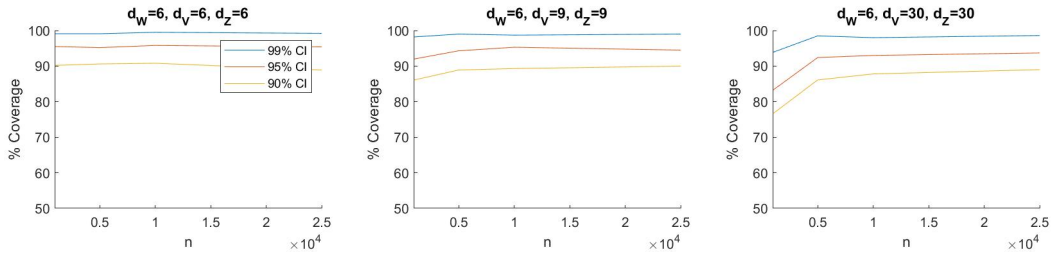
Figure 4.2: Simulated Median Squared Errors, $s = 0.5$

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.

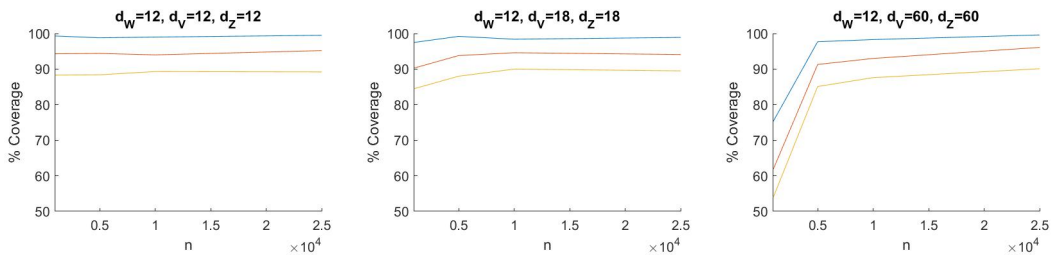
Figure 4.3 shows the percentage of simulations in which 99%, 95%, and 90% confidence intervals cover the true parameter β_0 (recall β_0 is drawn at random in each simulation). The confidence intervals are those based on a Gaussian approximation for the doubly-robust estimator as described in Section 3. In all cases the coverage is close to nominal level in large samples. In small samples the coverage is close to nominal apart from in the case with a very large number of proxies ($d_V = d_Z = 60$ in the bottom right sub-figure).

Figure 4.4 shows the coverage in the less favorable setting with $s = 0.5$. There is substantial undercoverage in the rightmost sub-figures with many proxies, although this appears to improve with the sample size. The middle subfigures show that with a moderate number of proxies compared to unobserved confounders the confidence intervals severely undercover in small samples and moderately undercover in larger samples.

In Table 1 we give the proportion of simulations in which the rank of the estimated nuisance parameter $\hat{M}$, is equal to the number of confounders d_W (which is the rank of the matrix M_0). The estimate $\hat{M}$ is attained using the full sample. The figures in Table 1



(a) $d_W = 6$



(b) $d_W = 12$

Figure 4.3: Simulated Confidence Interval Coverage, $s = 1$

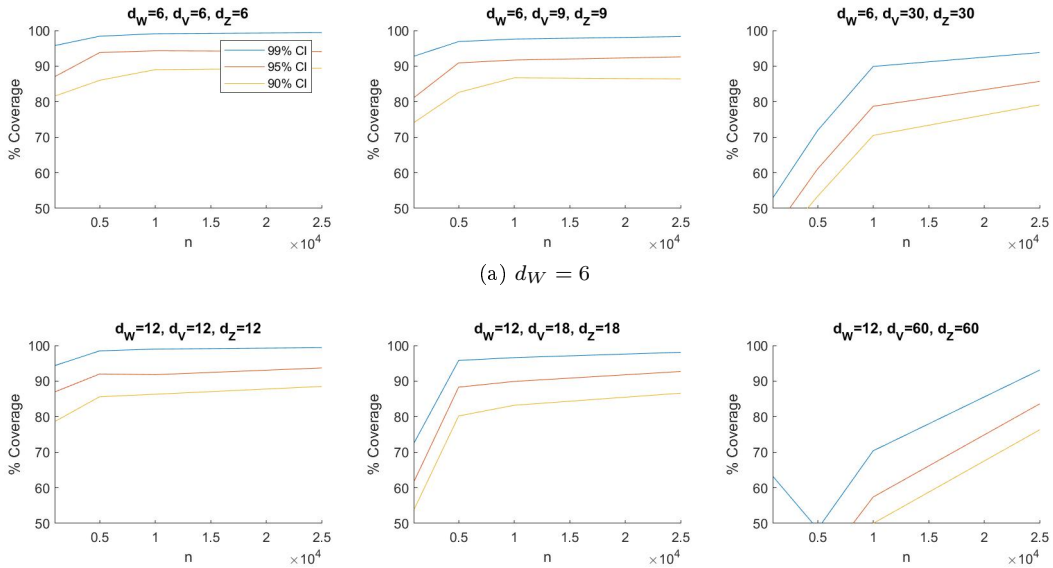
Confidence interval coverage of the treatment parameter. on the y-axes are percentages of 1000 simulated datasets in which confidence intervals contain β_0. The different figures correspond to different choices for the number of confounding factors d_W, and the numbers of proxies d_V and d_Z.

correspond to the favorable $s = 1$ case, whereas those in Table 2 are for the less favorable $s = 0.5$ setting.

In each case the proportion is generally increasing with the sample size. In Table 1 we see that in small samples ($n = 1000$) the rank selection is less accurate when the ratio of the number of proxies to the number of unobserved confounders is larger. Curiously, for $n > 1000$ the worst performance occurs for $d_V/d_W = 1.5$. When $s = 0.5$ there is a similar trend in the case of $d_W = 6$ although worse performance when $d_V/d_W = 1.5$ only occurs for the largest samples. One possible explanation is that this nonlinearity reflects a change in the balance between two opposing effects. On the one hand many proxies provide many signals regarding the latent factors, which increases accuracy, but on the other hand the presence of many proxies increases the risk that sample correlation between the components of $\tilde{Z}_i$ and $\tilde{V}_i$ creates the illusion of additional factors.

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



(a) $d_W = 6$

(b) $d_W = 12$

Figure 4.4: Simulated Confidence Interval Coverage, $s = 1$

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.

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.

References

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

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

Chernozhukov, Victor, Escanciano, Juan Carlos, Ichimura, Hidehiko, Newey, Whitney K., & Robins, James M. 2016. Locally Robust Semiparametric Estimation. July.

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*.

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

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

Table 2: Frequency of Correct Rank Selection, $s = 1$

		Sample Size			
d_W	$d_V = d_Z$	1000	5000	10000	25000
6	6	0.994	1	1	1
6	9	0.830	0.884	0.896	0.903
6	30	0.755	0.999	0.998	0.998
12	12	0.994	1	1	1
12	18	0.795	0.939	0.956	0.960
12	60	0.033	1	1	1

Figures are the proportion of the 1000 simulated datasets in which the estimated rank of M_0 is 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.

Table 3: Frequency of Correct Rank Selection, $s = 0.5$

		Sample Size			
d_W	$d_V = d_Z$	1000	5000	10000	25000
6	6	0.803	0.956	0.991	0.998
6	9	0.236	0.734	0.826	0.876
6	30	0	0.129	0.738	0.983
12	12	0.526	0.965	0.995	1
12	18	0.018	0.538	0.783	0.908
12	60	0	0	0.041	0.907

Figures are the proportion of the 1000 simulated datasets in which the estimated rank of M_0 is 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.

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.

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.

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.

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 An alternative estimate of ξ_0

In Section 2 we develop an estimator of ξ_0 that is a simple function of the reduced-rank estimators of M_0 and 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.

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

ξ_0 satisfies the condition above if and only if it minimizes the following least squares criterion:

$$E\big[(\bar{Y}_i - \xi_0' M_{0,[:,1:d_Z]}\bar{Z}_i)^2\big]$$

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 minimize a penalized empirical analogue of the criterion above. The penalization induces sparsity in the estimate of ξ_0. In particular, our estimate of ξ_0 is the vector ξ that minimizes the empirical objective below:

$$||\check{Y} - \check{Z}\tilde{M}'_{[:,1:d_Z]}\xi||_F^2 + \delta_n||\xi||_1 \tag{16}$$

Where $||\cdot||_1$ is the ℓ_1 norm and δ_n is a penalty parameter. $\tilde{M}$ is the matrix of regression estimates from multiple linear regression of $\tilde{V}_i$ on $(\tilde{Z}_i', \tilde{X}_i')'$.

Minimization of (16) 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, δ_n could be chosen using cross-validation.

B Proofs

Proof Theorem 1. By Assumption 1.1.ii $E[\varepsilon_i(Z_i', X_i')] = 0$ and $E[\varepsilon_i D_i'] = 0$ and so $E[\varepsilon_i(\tilde{Z}_i', \tilde{X}_i')] = 0$ and by the same reasoning $E[\upsilon_i(\tilde{Z}_i', \tilde{X}_i')] = 0$

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

$$\tilde{V}_i = B_0\tilde{W}_i + u_i$$
$$\tilde{Y}_i = \beta_0'\tilde{X}_i + A_0\tilde{W}_i + \epsilon_i$$

And so:

$$E\big[(\tilde{V}_i - B_0\tilde{W}_i)(\tilde{Z}_i', \tilde{X}_i')\big] = 0 \tag{17}$$

$$E\big[(\tilde{Y}_i - \beta_0'\tilde{X}_i - A_0\tilde{W}_i)(\tilde{Z}_i', \tilde{X}_i')\big] = 0 \tag{18}$$

Recall the definition of C_0 and G_0:

$$(C_0, G_0) = E\big[\tilde{W}_i(Z_i', X_i')\big]E\big[(Z_i', X_i')'(Z_i', X_i')\big]^+$$

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

$$(C_0, G_0)E\big[(Z_i', X_i')'(Z_i', X_i')\big] = E\big[\tilde{W}_i(Z_i', X_i')\big]$$

Using the above to substitute out $E\big[\tilde{W}_i(Z_i', X_i')\big]$ from (17) and (18) we get:

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

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

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

□

Proof Theorem 2. **Step 1: Prove the rank conditions on the nuisance parameters**

Under Assumption 1.4 $C_0' = E[\bar{Z}_i\bar{Z}_i']^{-1}E[\bar{Z}_i\tilde{W}_i']$ and so by Assumption 1.3 C_0 has full row rank. By Assumption 1.1 $E[D_i\upsilon_i'] = 0$ and $E[W_i\upsilon_i'] = 0$ and so $E[\tilde{W}_i\upsilon_i] = 0$. Since $\tilde{V}_i = B_0\tilde{W}_i + \upsilon_i$ we then have $B_0 = E[\tilde{V}_i\tilde{W}_i']E[\tilde{W}_i\tilde{W}_i']^{-1}$. Also by Assumption 1.1, $E[X_i\upsilon_i'] = 0$ and $E[D_i\upsilon_i'] = 0$, which implies $E[\tilde{X}_i\upsilon_i'] = 0$ and thus $\tilde{V}_i = \bar{V}_i$ and thus $B_0 = E[\tilde{V}_i\tilde{W}_i']E[\tilde{W}_i\tilde{W}_i']^{-1}$. So by Assumption 1.2 B_0 has full column rank.

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

Step 2:

From Theorem 1 we have:

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

Suppose the following holds:

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

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

$$B_0C_0 = BC \tag{19}$$

$$B_0G_0 = BG \tag{20}$$

$$A_0C_0 = AC \tag{21}$$

$$A_0G_0 - \beta_0' = AG - \beta' \tag{22}$$

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 (21) by M we get $A_0G_0 = ACM$ and substituting this into (22) gives:

$$ACM - \beta_0' = AG - \beta' \tag{23}$$

Now, post-multiplying both sides of (19) by M we get $B_0G_0 = BCM$. Substituting into (20) $BG = BCM$. 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 (23) we get $\beta = \beta_0$, as required.

□

Lemma 1. *There exist matrices A, B, C, and G with so that B has full column rank and β, A, B, C, and G satisfy (4) if and only if there exists a matrix M and a vector ξ so that:*

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

Proof of Lemma 1. First let us prove the '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 (**??**) we get:

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

Let $M = B(C, G)$, then we get the result. Now the 'only if'. 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 $A = \xi' B$ and we are done. □

Proof of Corollary 1. Theorem 2 and Lemma 1 together show that (5) and (6) identify β_0. By Theorem 1 (5) is satisfied by $M_0 = B_0(C_0, G_0)$. By Assumption 1.4 $E[(\tilde{Z}_i', \tilde{X}_i')(\tilde{Z}_i', \tilde{X}_i')']$ is non-singular and so this M_0 is the unique solution. By Theorem 2 we then have $rank(M_0) = d_W$. Next we show that ξ_0 satisfies (6) if and only if $\xi_0 B_0 C_0 = A_0 C_0$. Substituting $M_0 = B_0(C_0, G_0)$ and combining (5) and (6) we get:

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

Again, using that $E[(\tilde{Z}_i', \tilde{X}_i')(\tilde{Z}_i', \tilde{X}_i')']$ 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 'if'. For the 'only if', Theorem 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 $\xi_0 M_0 = A_0(C_0, G_0)$, substituting into (3) gives the result.

Next we will show that there exists a ξ_0 with $||\xi_0|| \leq d_W$. By Theorem 2, $rank(B_0C_0) = d_W$. Since B_0C_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 2) is identified from (7). First note that $\bar{Z}_i$ is a linear combination of $\tilde{Z}_i$ and $\tilde{X}_i$ and so (3) implies:

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

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

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

Multiplying out $\begin{pmatrix}B_0C_0 & B_0G_0\\ A_0C_0 & A_0G_0\end{pmatrix}\begin{pmatrix}\bar{Z}_i\\ 0\end{pmatrix}$ and substituting $Q_0 = (A_0', B_0')'C_0$ we get (7). Q_0 is the unique solution to (7) because $E[\bar{Z}_i\bar{Z}_i']$ is non-singular by Assumption 1.4.

□

Lemma 2. *Under Assumptions 1.1-1.4 $\psi_i(\beta_0; M_0, \xi_0, \gamma_0, \mu_0)$ 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:

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

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:

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

Substituting the definition of μ_0 the RHS becomes:

$$\begin{aligned}\mu_0 G_\eta &= (G_\beta'\Omega^{-1} - G_\beta'\Omega^{-1}G_\eta(G_\eta'\Omega^{-1}G_\eta)^+G_\eta'\Omega^{-1})G_\eta\\ &= 0\end{aligned}$$

The final equality follows by the elementary property of the Moore-Penrose pseudo-inverse that for any matrix A, $A(A'A)^+A'A = AA^+A = A$, 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 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 Gateux differentiable, and 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 2 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 Gauteax 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 jth subsample:

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

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 derive the following three rates:

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

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

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

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

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

Which is $o(1)$ under the conditions in the Theorem.

We will 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 $\tilde{H}_i = \tilde{H}_i(\gamma_{H,0})$. Define $\bar{\beta} = ||\Sigma_{\tilde{X}}^{1/2}\beta_0||$, $\bar{\xi} = ||\Sigma_{\tilde{X}}^{1/2}\xi_0'||$, $\bar{\mu} = ||\mu_0\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||$, and $\bar{M} = ||\Sigma_{\tilde{Z}\tilde{X}}^{1/2}M_0'\Sigma_{\tilde{V}}^{-1/2}||$.

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 $||(\mu-\mu_0)\Sigma^{1/2}_{\tilde{Z},\tilde{X}}|| \leq \delta_\mu$, $||\Sigma^{1/2}_{\tilde{V}}(\xi-\xi_0)'|| \leq \delta_\xi$, and:

$$
\begin{aligned}
||\Sigma^{-1/2}_{\tilde{V}}(\gamma_V-\gamma_{V,0})\Sigma^{1/2}_D|| &\leq \delta_{\gamma,V} \\
||\Sigma^{-1/2}_{\tilde{X}}(\gamma_X-\gamma_{X,0})\Sigma^{1/2}_D|| &\leq \delta_{\gamma,X} \\
||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}(\gamma'_Z-\gamma'_{Z,0},\gamma'_X-\gamma'_{X,0})\Sigma^{1/2}_D|| &\leq \delta_{\gamma,Z}+\delta_{\gamma,X} \\
||\big((\gamma_Y-\gamma_{Y,0})-(\gamma_X-\gamma_{X,0})\beta_0-(\gamma_V-\gamma_{V,0})\xi'_0\big)\Sigma^{1/2}_D|| &\leq \delta_{\gamma,\epsilon}
\end{aligned}
$$

In our discussion of the remaining conditions we take $\mathcal{T}_n$ to be this set.

Condition 2

We show that:

$$\sup_{\eta\in\mathcal{T}_n} E\big[||\psi_i(\beta_0,\eta)||^q\big]^{1/q} \precsim 1$$

Using notation introduced above we have:

$$
\begin{aligned}
&E\big[||\psi_i(\beta_0;\xi,\gamma,\mu)||^q\big]^{1/q} \\
=&E\Big[||\mu\tilde{R}_i(\gamma_R)\big(\tilde{Y}_i(\gamma_Y)-\tilde{V}_i(\gamma_V)'\xi-\tilde{X}_i(\gamma_X)'\beta_0\big)||^q\Big]^{1/q}
\end{aligned}
$$

Using the triangle inequality and the definition of the operator norm:

$$
\begin{aligned}
&E\Big[||\mu\tilde{R}_i(\gamma_R)\big(\tilde{Y}_i(\gamma_Y)-\tilde{V}_i(\gamma_V)'\xi-\tilde{X}_i(\gamma_X)'\beta_0\big)||^q\Big]^{1/q} \\
\leq&E\big[||\psi_i(\beta_0,\eta_0)||^q\big]^{1/q}+||(\mu-\mu_0)\Sigma^{1/2}_{\tilde{Z}\tilde{X}}||E\big[||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}\tilde{R}_i\tilde{\epsilon}_i||^q\big]^{1/q} \\
+&||\mu\Sigma^{1/2}_{\tilde{Z}\tilde{X}}||\cdot||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}(\gamma_R-\gamma_{R,0})\Sigma^{1/2}_D||E\big[||\Sigma^{-1/2}_D D_i\tilde{\epsilon}_i||^q\big]^{1/q} \\
+&||\mu\Sigma^{1/2}_{\tilde{Z}\tilde{X}}||E\big[||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}\tilde{R}_i\tilde{V}'_i\Sigma^{-1/2}_{\tilde{V}}||^q\big]^{1/q}||\Sigma^{1/2}_{\tilde{V}}(\xi-\xi_0)'|| \\
+&||\mu\Sigma^{1/2}_{\tilde{Z}\tilde{X}}||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}(\gamma_R-\gamma_{R,0})\Sigma^{1/2}_D||E\big[||\Sigma^{-1/2}_D D_i\tilde{V}'_i\Sigma^{-1/2}_{\tilde{V}}||^q\big]^{1/q}||\Sigma^{1/2}_{\tilde{V}}(\xi-\xi_0)'|| \\
+&||\mu\Sigma^{1/2}_{\tilde{Z}\tilde{X}}||\Bigg(E\big[||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}\tilde{R}_iD'_i\Sigma^{-1/2}_D||^q\big]^{1/q}+||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}(\gamma_R-\gamma_{R,0})\Sigma^{1/2}_D||E\big[||\Sigma^{-1/2}_D D_iD'_i\Sigma^{-1/2}_D||^q\big]^{1/q}\Bigg) \\
&\times\Bigg(||\Sigma^{1/2}_D\big((\gamma_Y-\gamma_{0,Y})-(\gamma_V-\gamma_{0,V})\xi'_0-(\gamma_X-\gamma_{0,X})\beta_0\big)|| \\
&+||\Sigma^{1/2}_D(\gamma_V-\gamma_{0,V})\Sigma^{-1/2}_{\tilde{V}}||\cdot||\Sigma^{1/2}_{\tilde{V}}(\xi-\xi_0)'||\Bigg)
\end{aligned}
$$

For $\eta\in\mathcal{T}_n$ we have:

$$
\begin{aligned}
&E\Big[||\mu\tilde{R}_i(\gamma_R)\big(\tilde{Y}_i(\gamma_Y)-\tilde{V}_i(\gamma_V)'\xi-\tilde{X}_i(\gamma_X)'\beta_0\big)||^q\Big]^{1/q}\\
\lesssim&E\big[||\psi_i(\beta_0,\eta_0)||^q\big]^{1/q}+||(\mu-\mu_0)\Sigma^{1/2}_{\tilde{Z}\tilde{X}}||E\big[||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}\tilde{R}_i\tilde{\epsilon}_i||^q\big]^{1/q}\\
&+(\delta_\mu+\bar{\mu})(\delta_{\gamma,X}+\delta_{\gamma,Z})E\big[||\Sigma^{-1/2}_D D_i\tilde{\epsilon}_i||^q\big]^{1/q}\\
&+(\delta_\mu+\bar{\mu})\delta_\xi E\big[||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}\tilde{R}_i\tilde{V}_i'\Sigma^{-1/2}_{\tilde{V}}||^q\big]^{1/q}\\
&+(\delta_\mu+\bar{\mu})(\delta_{\gamma,X}+\delta_{\gamma,Z})E\big[||\Sigma^{-1/2}_D D_i\tilde{V}_i'\Sigma^{-1/2}_{\tilde{V}}||^q\big]^{1/q}\delta_\xi\\
&+(\delta_\mu+\bar{\mu})\Big(E\big[||\Sigma^{-1/2}_{\tilde{Z}\tilde{X}}\tilde{R}_iD_i'\Sigma^{-1/2}_D||^q\big]^{1/q}+(\delta_{\gamma,X}+\delta_{\gamma,Z})E\big[||\Sigma^{-1/2}_D D_iD_i'\Sigma^{-1/2}_D||^q\big]^{1/q}\Big)\\
&\quad\times(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi)
\end{aligned}
$$

Using Assumption 3.2:

$$
\begin{aligned}
&E\Big[||\mu\tilde{R}_i(\gamma_R)\big(\tilde{Y}_i(\gamma_Y)-\tilde{V}_i(\gamma_V)'\xi-\tilde{X}_i(\gamma_X)'\beta_0\big)||^q\Big]^{1/q}\\
\lesssim&1+\delta_\mu(d_X+d_Z)^{1/2}+(\delta_\mu+\bar{\mu})(\delta_{\gamma,X}+\delta_{\gamma,Z})d_D^{1/2}\\
&+(\delta_\mu+\bar{\mu})\delta_\xi(d_X+d_Z)^{1/2}d_V^{1/2}+(\delta_\mu+\bar{\mu})(\delta_{\gamma,X}+\delta_{\gamma,Z})d_D^{1/2}d_V^{1/2}\delta_\xi\\
&+(\delta_\mu+\bar{\mu})\Big((d_X+d_Z)^{1/2}d_D^{1/2}+(\delta_{\gamma,X}+\delta_{\gamma,Z})d_D\Big)(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi)
\end{aligned}
$$

Under the conditions of the theorem the right-hand side is $O(1)$.

Condition 3

We now show that:

$$
\begin{aligned}
&\sup_{\eta\in\mathcal{T}_n}E\big[||\mu E[(\tilde{Z}_i(\gamma_Z)',\tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']||^q\big]^{1/q}\\
\lesssim&1
\end{aligned}
$$

Given our notation we have:

$$
\begin{aligned}
&E\big[||\mu E[(\tilde{Z}_i(\gamma_Z)',\tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']||^q\big]^{1/q}\\
=&E\Big[||\mu\tilde{R}_i(\gamma_R)\tilde{X}_i(\gamma_X)'||^q\Big]^{1/q}
\end{aligned}
$$

Using the triangle inequality and definition of the operator norm:

$$
\begin{aligned}
&E\left[||\mu\tilde{R}_i(\gamma_R)\tilde{X}_i(\gamma_X)'||^q\right]^{1/q} \\
\leq &E\left[||\mu_0\tilde{R}_i\tilde{X}_i'||^q\right]^{1/q} \\
+&||(\mu-\mu_0)\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||\big(E\big[||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}\tilde{R}_i\tilde{X}_i'||^q\big]^{1/q} \\
+&||\mu\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\gamma_R-\gamma_{R,0})\Sigma_D^{1/2}||E\big[||\Sigma_D^{-1/2}D_i\tilde{X}_i'||^q\big]^{1/q} \\
+&||\mu\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||\left(E\big[||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}\tilde{R}_iD_i'\Sigma_D^{-1/2}||^q\big]^{1/q}+||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\gamma_R-\gamma_{R,0})\Sigma_D^{1/2}||E\big[||\Sigma_D^{-1/2}D_iD_i'\Sigma_D^{-1/2}||^q\big]^{1/q}\right) \\
&\times||\Sigma_D^{1/2}(\gamma_X-\gamma_{0,X})\Sigma_{\tilde{X}}^{-1/2}||\cdot||\Sigma_{\tilde{X}}^{1/2}||
\end{aligned}
$$

For $\eta\in\mathcal{T}_n$ we have:

$$
\begin{aligned}
&E\left[||\mu\tilde{R}_i(\gamma_R)\tilde{X}_i(\gamma_X)'||^q\right]^{1/q} \\
\precsim &E\left[||\mu_0\tilde{R}_i\tilde{X}_i'||^q\right]^{1/q}+\delta_\mu E\big[||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}\tilde{R}_i\tilde{X}_i'||^q\big]^{1/q} \\
+&(\bar{\mu}+\delta_\mu)(\delta_{\gamma,Z}+\delta_{\gamma,X})E\big[||\Sigma_D^{-1/2}D_i\tilde{X}_i'||^q\big]^{1/q} \\
+&(\bar{\mu}+\delta_\mu)\delta_{\gamma,X}\left(E\big[||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}\tilde{R}_iD_i'\Sigma_D^{-1/2}||^q\big]^{1/q}+(\delta_{\gamma,Z}+\delta_{\gamma,X})E\big[||\Sigma_D^{-1/2}D_iD_i'\Sigma_D^{-1/2}||^q\big]^{1/q}\right)
\end{aligned}
$$

Using Assumption 3.2 we then get:

$$
\begin{aligned}
&E\left[||\mu\tilde{R}_i(\gamma_R)\tilde{X}_i(\gamma_X)'||^q\right]^{1/q} \\
\precsim &1+\delta_\mu(d_Z+d_X)^{1/2}d_X^{1/2} \\
&+(\bar{\mu}+\delta_\mu)(\delta_{\gamma,Z}+\delta_{\gamma,X})d_D^{1/2}d_X^{1/2} \\
&+(\bar{\mu}+\delta_\mu)\delta_{\gamma,X}\left((d_Z+d_X)^{1/2}d_D^{1/2}+(\delta_{\gamma,Z}+\delta_{\gamma,X})d_D\right)
\end{aligned}
$$

Under the conditions of the theorem the right-hand side is $O(1)$.

Condition 4

Next we show that:

$$
\begin{aligned}
&\sup_{\eta\in\mathcal{T}_n}||\mu_0E[(\tilde{Z}_i',\tilde{X}_i')'\tilde{X}_i']-\mu E[(\tilde{Z}_i(\gamma_Z)',\tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']|| \\
\precsim &\delta_\mu+(\delta_{\gamma,Z}+\delta_{\gamma,X})\delta_{\gamma,X}
\end{aligned}
$$

By the triangle inequality and definition of the matrix norm:

$$
\begin{aligned}
&||\mu_0E[(\tilde{Z}_i',\tilde{X}_i')'\tilde{X}_i']-\mu E[(\tilde{Z}_i(\gamma_Z)',\tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']|| \\
\leq &||(\mu_0-\mu)\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||\cdot||E[\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\tilde{Z}_i',\tilde{X}_i')'\tilde{X}_i'\Sigma_{\tilde{X}}^{-1/2}]||\cdot||\Sigma_{\tilde{X}}^{1/2}|| \\
+&(||(\mu_0-\mu)\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||+||\mu_0\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||)||E[\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\tilde{Z}_i',\tilde{X}_i')'\tilde{X}_i']-E[\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\tilde{Z}_i(\gamma_Z)',\tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']||
\end{aligned}
$$

Using the properties of partialling out:

$$
\begin{aligned}
&E[\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - E[\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)'] \\
=&E[\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}\big((\tilde{Z}_i', \tilde{X}_i')' - \tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')\big)(\tilde{X}_i(\gamma_X)' - \tilde{X}_i')] \\
=&\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\gamma_{Z,0}' - \gamma_Z', \gamma_{X,0}' - \gamma_X')'\Sigma_D(\gamma_{X,0} - \gamma_X)
\end{aligned}
$$

And so:

$$
\begin{aligned}
&||E[\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - E[\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']|| \\
\leq&||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\gamma_{Z,0}' - \gamma_Z', \gamma_{X,0}' - \gamma_X')'\Sigma_D^{1/2}|| \cdot ||\Sigma_D^{1/2}(\gamma_{X,0} - \gamma_X)\Sigma_{\tilde{X}}^{-1/2}|| \cdot ||\Sigma_{\tilde{X}}^{1/2}||
\end{aligned}
$$

Combining we get:

$$
\begin{aligned}
&||\mu_0 E[(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'] - \mu E[(\tilde{Z}_i(\gamma_Z)', \tilde{X}_i(\gamma_X)')'\tilde{X}_i(\gamma_X)']|| \\
\leq&||(\mu_0 - \mu)\Sigma_{\tilde{Z}\tilde{X}}^{1/2}|| \cdot ||E[\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\tilde{Z}_i', \tilde{X}_i')'\tilde{X}_i'\Sigma_{\tilde{X}}^{-1/2}]|| \cdot ||\Sigma_{\tilde{X}}^{1/2}|| \\
+&(||(\mu_0 - \mu)\Sigma_{\tilde{Z}\tilde{X}}^{1/2}|| + ||\mu_0\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||) \\
&\times ||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\gamma_{Z,0}' - \gamma_Z', \gamma_{X,0}' - \gamma_X')'\Sigma_D^{1/2}|| \cdot ||\Sigma_D^{1/2}(\gamma_{X,0} - \gamma_X)\Sigma_{\tilde{X}}^{-1/2}|| \cdot ||\Sigma_{\tilde{X}}^{1/2}||
\end{aligned}
$$

Note that:

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

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

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

Under the conditions of the Theorem we then have:

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

Condition 5
We will show that:

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

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:

$$
\begin{aligned}
&E\big[||\psi_i(\beta_0,\eta_0)-\psi_i(\beta_0,\eta)||^2\big]^{1/2}\\
\leq &E\left[||(\mu-\mu_0)(\tilde{Z}_i',\tilde{X}_i')'||^2\tilde{\epsilon}_i^2\right]^{1/2}\\
+&E\left[||\mu(\tilde{Z}_i',\tilde{X}_i')'||^2||\tilde{V}_i'(\xi-\xi_0)||^2\right]^{1/2}\\
+&E\left[||\mu\big(\tilde{R}_i(\gamma_R)\tilde{H}_i(\gamma_H)'-\tilde{R}_i\tilde{H}_i'\big)\zeta||^2\right]^{1/2}
\end{aligned}
\tag{24}
$$

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

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

Where the final inequality above assumes $\eta\in\mathcal{T}_n$. For the second term on the RHS of (24), if $\eta\in\mathcal{T}_n$ then:

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

Next we will show that:

$$
\begin{aligned}
&E\Big[||\mu\big(\tilde{R}_i(\gamma_R)\tilde{H}_i(\gamma_H)' - \tilde{R}_i\tilde{H}_i'\big)\zeta||^2\Big]^{1/2} \\
&\precsim \sqrt{d_X}(\bar{\mu}+\delta_\mu)(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi)\sigma_{\tilde{Z}\tilde{X}|D} \\
&+2\sqrt{d_X}(\bar{\mu}+\delta_\mu)(\delta_{\gamma,Z}+\delta_{\gamma,X})(\sigma_{\tilde{\epsilon}|D}+\delta_\xi\sigma_{\tilde{V}|D}) \\
&+(\delta_\mu+\bar{\mu})(\delta_{\gamma,X}+\delta_{\gamma,Z})(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi)\sigma_{4,D}
\end{aligned} \tag{25}
$$

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

$$
\begin{aligned}
&E\Big[||\mu\big(\tilde{R}_i(\gamma_R)\tilde{H}_i(\gamma_H)' - \tilde{R}_i\tilde{H}_i'\big)\zeta||^2\Big]^{1/2} \\
&\leq E\big[||\mu\tilde{R}_i||^2||D_i'(\gamma_{H,0}-\gamma_H)'\zeta||^2\big]^{1/2} \\
&+2E\big[||\mu(\gamma_{R,0}-\gamma_R)D_i||^2\tilde{\epsilon}_i^2\big]^{1/2} \\
&+2E\big[||\mu(\gamma_{R,0}-\gamma_R)D_i||^2|\tilde{V}_i'(\xi-\xi_0)|^2\big]^{1/2} \\
&+E\big[||\mu(\gamma_{R,0}-\gamma_R)D_iD_i'(\gamma_{H,0}-\gamma_H)'\zeta||^2\big]^{1/2}
\end{aligned} \tag{26}
$$

To bound the above first note that under Assumption 3.2:

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

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

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

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

$$
\begin{aligned}
E[|\tilde{V}_i'(\xi-\xi_0)|^2|D_i] &\leq ||\Sigma_{\tilde{V}}^{1/2}(\xi-\xi_0)'||^2\cdot||E[\Sigma_{\tilde{V}}^{-1/2}\tilde{V}_i\tilde{V}_i'\Sigma_{\tilde{V}}^{-1/2}|D_i]|| \\
&\precsim \delta_\xi^2
\end{aligned}
$$

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

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

If Assumption 3.2 holds, then using the law of iterated expectations and the above we get from (26):

$$
\begin{aligned}
&E\Big[||\mu\big(\tilde{R}_i(\gamma_R)\tilde{H}_i(\gamma_H)' - \tilde{R}_i\tilde{H}_i')\zeta||^2\Big]^{1/2} \\
&\lesssim \sqrt{d_X}(\bar{\mu}+\delta_\mu)(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi) \\
&+\sqrt{d_X}(\bar{\mu}+\delta_\mu)(\delta_{\gamma,Z}+\delta_{\gamma,X})(1+\delta_\xi) \\
&+E\big[||\mu(\gamma_{R,0}-\gamma_R)D_iD_i'(\gamma_{H,0}-\gamma_H)'\zeta||^2\big]^{1/2}
\end{aligned} \tag{27}
$$

Finally, with repeated application of the properties of the matrix norm:

$$
\begin{aligned}
&E\big[||\mu(\gamma_{R,0}-\gamma_R)D_iD_i'(\gamma_{H,0}-\gamma_H)'\zeta||^2\big] \\
\leq& E\big[||\mu(\gamma_{R,0}-\gamma_R)\Sigma_D^{1/2}||^2\cdot||\Sigma_D^{-1/2}D_iD_i'\Sigma_D^{-1/2}||^2\cdot||\Sigma_D^{1/2}(\gamma_{H,0}-\gamma_H)'\zeta||^2\big] \\
\leq& ||\mu\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||^2\cdot||\Sigma_D^{1/2}(\gamma_{R,0}-\gamma_R)'\Sigma_{\tilde{Z}\tilde{X}}^{1/2}||^2\cdot||\Sigma_D^{1/2}(\gamma_{H,0}-\gamma_H)'\zeta||^2E[||\Sigma_D^{-1/2}D_iD_i'\Sigma_D^{-1/2}||^2] \\
\lesssim& d_D(\delta_\mu+\bar{\mu})^2(\delta_{\gamma,X}+\delta_{\gamma,Z})^2(\delta_{\gamma,\epsilon}+\delta_{\gamma,V}\delta_\xi)^2
\end{aligned}
$$

Where the last line assumes $\eta \in \mathcal{T}_n$ and uses $E[||\Sigma_D^{-1/2}D_iD_i'\Sigma_D^{-1/2}||^2] \lesssim d_D$ from Assumption 3.2.

Combining we get (25) and in all:

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

Under the conditions of the theorem we get:

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

Condition 6

Next we show that

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

Twice differentiating we get:

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

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:

$$
\begin{aligned}
&||\frac{\partial^2}{\partial r^2} E\big[\psi_i\big(\beta_0, \eta_0 + r(\eta - \eta_0)\big)\big]|| \\
\leq&2||(\mu - \mu_0)\Sigma_{\tilde{Z}\tilde{X}}^{1/2}|| \cdot ||\Sigma_{\tilde{Z}\tilde{X}}^{1/2} M_0' \Sigma_{\tilde{V}}^{-1/2}|| \cdot ||\Sigma_{\tilde{V}}^{1/2}(\xi_0 - \xi)'|| \\
+&2||\mu_0 \Sigma_{\tilde{Z}\tilde{X}}^{1/2}|| \cdot ||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\gamma_Z' - \gamma_{Z,0}', \gamma_X' - \gamma_{X,0}')'\Sigma_D^{1/2}|| \\
&\times ||\Sigma_D^{1/2}\big((\gamma_Y - \gamma_{0,Y}) - (\gamma_V - \gamma_{0,V})\xi_0' - (\gamma_X - \gamma_{0,X})\beta_0\big)|| \\
+&6r||\mu_0 \Sigma_{\tilde{Z}\tilde{X}}^{1/2}|| \cdot ||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\gamma_Z' - \gamma_{Z,0}', \gamma_X' - \gamma_{X,0}')'\Sigma_D^{1/2}|| \cdot ||\Sigma_D^{1/2}(\gamma_V - \gamma_{0,V})\Sigma_{\tilde{V}}^{-1/2}|| \cdot ||\Sigma_{\tilde{V}}^{1/2}(\xi_0 - \xi)|| \\
+&6r||(\mu - \mu_0)\Sigma_{\tilde{Z}\tilde{X}}^{1/2}|| \cdot ||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\gamma_Z' - \gamma_{Z,0}', \gamma_X' - \gamma_{X,0}')'\Sigma_D^{1/2}|| \\
&\times ||\Sigma_D^{1/2}\big((\gamma_Y - \gamma_{0,Y}) - (\gamma_V - \gamma_{0,V})\xi_0' - (\gamma_X - \gamma_{0,X})\beta_0\big)|| \\
+&12r^2||(\mu - \mu_0)\Sigma_{\tilde{Z}\tilde{X}}^{1/2}|| \cdot ||\Sigma_{\tilde{Z}\tilde{X}}^{-1/2}(\gamma_Z' - \gamma_{Z,0}', \gamma_X' - \gamma_{X,0}')'\Sigma_D^{1/2}|| \cdot ||\Sigma_D^{1/2}(\gamma_V - \gamma_{0,V})\Sigma_{\tilde{V}}^{1/2}|| \cdot ||\Sigma_{\tilde{V}}^{-1/2}(\xi_0 - \xi)||
\end{aligned}
$$

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

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

The conditions in the theorem then give the result.

□