Abstract

A recent literature considers causal inference using noisy proxies for unobserved confounding factors. The proxies are divided into two sets 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 work we consider linear models with many proxy controls and possibly many confounders. A key insight is that if each group of proxies is strictly larger than the number of confounding factors, then a matrix of nuisance parameters has a low-rank structure and a vector of nuisance parameters has a sparse structure. We can exploit the rank-restriction and sparsity 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 provide an estimator with a closed-form as well as a doubly-robust estimator that must be evaluated using numerical methods. We provide conditions under which our doubly-robust estimator is uniformly root-\(n\) consistent, asymptotically centered normal, and our suggested confidence intervals have asymptotically correct coverage. We provide simulation evidence that our methods achieve better performance than existing approaches in high dimensions, particularly when the number of proxies is substantially larger than the number of confounders.

1 Introduction and Related Literature

A recent, rapidly growing literature considers the problem of causal inference when a researcher observes only proxies for unobserved confounding factors. For example, one may wish to use test scores to proxy for ability, an unobserved confounder. The proxies are divided into two groups which are independent conditional on the unobserved confounders. One group 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.

Compared to standard factor-analytic methods, the proxy control approach has the advantage that the factor structure itself need not be identified. That is, neither the distribution of unobserved factors nor the causal effects of these factors must be identified. Thus proxy control methods may be applied even when the assumptions required for identification of the factor structure do not hold. However, this limits the application of proxy control methods to settings in which the factor structure itself is not of interest.
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 the rich covariates to adjust for confounding. Standard methods treat the covariates as controls, and are valid only if the confounders are non-random once we condition on the covariates. This condition is known as ‘unconfoundedness’, ‘ignorability’, or ‘selection on observables’. It holds, for example, if all of the confounders are included among the covariates.

By contrast, proxy control methods do not require that the confounders perfectly explain the confounders, only that they are informative about them. Proxy control methods can be valid when the covariates are only noisy proxies for the confounders. Thus proxy control methods allow us to exploit the richness of the covariates without having so assume unconfoundedness. However, to relax this assumption we must place other restrictions on the relationships between the relevant variables.

Nonparametric identification with proxy controls is achieved using the conditional independence assumptions listed in Sub-Figure 1.(a). \( Y_i \) is an outcome of interest for an individual \( i \) and \( X_i \) is a vector of treatments assigned to \( i \). \( Y_i(\cdot) \) is individual \( i \)'s treatment response function, or equivalently, \( Y_i(x) \) is individual \( i \)'s potential outcome from the counterfactual treatment level \( x \). \( W_i \) is a vector of unobserved confounding factors. \( D_i \) is a vector of observed characteristics like age and gender. \( Z_i \) and \( V_i \) are two groups of proxies for \( W_i \). \( Z_i \) is a vector of negative control treatments and \( V_i \) is a vector of negative control outcomes.

Sub-Figure 1.(b) contains two causal diagrams. The diagram on the left, taken from Tchetgen et al. (2020) differs from that on the right in that it assumes \( Z_i \) is a vector of pre-treatment variables. In the diagram on the right, \( Z_i \) is a vector of post-treatment variables, as in Deaner (2021). Each diagram is associated with a nonparametric structural equations model (Pearl 2009) that implies the conditions in Sub-Figure 1.(a). Note that this is not the only causal structure that implies the conditions in Sub-Figure 1.(a), for example we could allow simultaneous causation between \( V_i \) and \( W_i \) and between \( X_i \) and \( Z_i \) (see Deaner 2021).

1.(b) imposes that \( Z_i \) and \( V_i \) have no direct causal effect on each other. However, \( Z_i \) and \( V_i \) may both depend on \( D_i \) and \( X_i \). The dependence on \( W_i \) means that we can understand these variables as noisy proxies for \( W_i \). Correspondingly, condition 1.(a).3 implies that, after accounting for \( W_i \) and \( D_i \), the variables \( V_i \) and \( Z_i \) are independent.

The conditions in Figure 1 restrict the manner in which the treatments and outcomes are related to the proxies \( Z_i \) and \( V_i \). These restrictions are asymmetric. 1.(b) imposes that the negative control treatments \( Z_i \) can both cause and be caused by the treatment \( X_i \). However, the negative control outcomes \( V_i \) cannot cause or be caused by the treatments \( X_i \). On the other hand, the diagrams in 1.(b) allow \( V_i \) to directly cause the outcomes \( Y_i \) whereas \( Z_i \) cannot directly cause \( Y_i \).

Miao et al. (2018b) show that if the conditional independence restrictions in Sub-Figure 1.(a) hold along with two statistical completeness assumptions and some regularity conditions, then the average structural function is nonparametrically identified. The average structural function is equal to \( E[Y_i(x)] \) in potential outcomes notation or \( E[Y_i|do(x)] \) in the notation of Pearl 2009. Deaner 2021 shows that under the same conditional independence restrictions, related regularity conditions, and weaker completeness conditions, one can identify the conditional average structural function \( E[Y_i(x)|X_i = x] \). Note that \[ Y_i(x_1)|X_i = x_2 \]
Deaner (2021) allows for additional covariates $D_i$ by assuming they are included in both $W_i$, $Z_i$, and $V_i$. Alternatively, one can apply the results in Miao et al. (2018b) and Deaner (2021) within one particular stratum of $D_i$.

Deaner (2021) also provides an alternative characterization of the conditional average structural function and provides sufficient conditions for the conditional independence restrictions in the context of panel data.

Nonparametric estimation in this setting was considered by Deaner (2019), and later by Tchetgen et al. (2020), Singh (2020), Cui et al. (2020), and Kallus et al. (2021). Miao et al. (2018a) consider estimation in parametric models when a ‘confounding bridge’ function is identified.

In this work we consider identification, estimation, and inference in linear models when the set of proxy controls and the set of confounding factors may be high-dimensional. In particular, we allow for the possibility that the number of proxies and confounders grows with the sample size. Existing work applies to low-dimensional settings in which the number of proxies and confounding factors is treated as fixed.

A key insight in this work is that if there are strictly fewer confounders than there are proxies in $V_i$ and $Z_i$, then a matrix of nuisance parameters has a low rank structure. In addition, a vector of nuisance parameters has a sparse structure. We exploit this low-rank structure and sparsity 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, so we propose a model selection methods that allows us to adapt to this quantity. The model selection methods are based on techniques from the literature on reduced-rank regression and $\ell_1$-penalized regression.

We present three different estimation methods. The first employs a known bound on the number of confounders, the second selects the number of confounders using the data. These first two methods are based on techniques from the literature on reduced-rank regression and have closed-forms. The third estimator is doubly-robust and is an example of a 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.

Deaner (2018) earlier considered nonparametric estimation in an equivalent setting that arises in the context of panel data. Deaner (2019) was expanded to include a method for nonparametric inference in
and asymptotically Gaussian, under relatively weak conditions on the nuisance parameter estimates. A disadvantage of the doubly-robust estimator (compared to our alternative methods) is that it does not have a closed-form, however the numerical part of the algorithm can be carried out using any standard Lasso procedure.

Linearity also allows us to weaken the identifying assumptions used in the nonparametric case. In particular, we need only assume variables are uncorrelated rather than independent, we can replace statistical completeness with more intuitive full rank conditions, and we avoid the need for regularity conditions like Assumption A3 in [Miao et al.] (2018a).

The linear proxy control model has a long history, dating back to the work of Griliches (1977) who considers two scalar proxies for a single confounding factor. However, to the best of our knowledge no existing work exploits the dimension reduction when there are fewer confounders than proxies in $V_i$ or $Z_i$, 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 and sparsity in the nuisance parameters when the number of unobserved confounders is less than the number of proxies in $Z_i$ and $V_i$. We develop asymptotic theory for the estimator and an associated inference method, and we provide simulation evidence of the efficacy of our methods.

## 2 Model and Identification

Suppose that for each individual $i$, the potential outcome $Y_i(x)$ from treatment $x$ is linear in $x$ and differs between individuals only by an additive constant:

$$Y_i(x) = x'\beta_0 + U_i$$

$\beta_0$ in the above is a vector of parameters that is the same for all individuals. $U_i$ is a random scalar that represents heterogeneity in potential outcomes. Note that the model implies treatment effects are the same for all individuals. By the definition of potential outcomes, the realized outcome $Y_i$ satisfies $Y_i = Y_i(X_i)$ and so:

$$Y_i = X'_i\beta_0 + U_i$$  (1)

Where $X_i$ is the realized treatment and is a column vector of length $d_X$. Let $W_i$ be a length-$d_W$ column vector of latent variables and let $V_i$ be a length-$d_V$ vector of proxies. Let $D_i$ be a length-$d_D$ vector of observed control variables whose first entry is 1. We assume the following two relationships are linear:

$$U_i = A_0 W_i + L_0 D_i + \epsilon_i$$  (2)

$$V_i = B_0 W_i + R_0 D_i + \nu_i$$  (3)

Where $E[\epsilon_i(W_i', D_i')] = E[\nu_i(W_i', D_i')] = 0$. We can always find matrices $A_0$, $B_0$, $L_0$, and $R_0$ that satisfy these relationships, however we will place assumptions directly on $\epsilon_i$ and $\nu_i$ and so the linear specifications do incur a loss of generality.

$V_i$ does not directly enter the expression for $U_i$, however this does not rule out the possibility that $V_i$ has a direct causal effect on $Y_i$. We could make this more explicit by including $V_i$ in the expression for $Y_i$ as follows:

$$Y_i = X'_i\beta_0 + F_0 W_i + K_0 D_i + \chi_0 V_i + \epsilon_i$$  (4)
We wish to identify and estimate \( \beta_0 \) in the model for \( Y_i(\cdot) \). Identification of \( \beta_0 \) immediately implies identification of average treatment effects (ATEs):

\[
E[Y_i(x_1) - Y_i(x_2)] = (x_1 - x_2)'\beta_0
\]

And also the conditional average structural function (CASF):

\[
E[Y_i(x_1)|X_i = x_2] = (x_1 - x_2)'\beta_0 + E[Y_i|x_i = x_2]
\]

Before we state our identifying assumptions let us introduce some additional notation. In particular we define the following objects for each of the variables \( H_i = W_i, V_i, Z_i, Y_i, X_i \):

\[
\begin{align*}
\gamma_{H,0} &= E[D_iD_i']^+E[D_iH_i] \\
\omega_{H,0} &= E[(X_i'|D_i')(X_i'|D_i')] + E[(X_i'|D_i')'H_i'] \\
\tilde{H}_i(\gamma) &= H_i - \gamma'D_i \\
\tilde{H}_i(\omega) &= H_i - \omega'(X_i'|D_i')'
\end{align*}
\]

\( M^+ \) denotes the Moore-Penrose pseudo-inverse of a matrix \( M \). For notational convenience we sometimes write \( \tilde{H}_i = \tilde{H}_i(\gamma_{H,0}) \) and \( \tilde{H}_i = \tilde{H}_i(\omega_{H,0}) \). So for example, \( \bar{X}_i \) is the residual from population linear regression of \( X_i \) on \( D_i \), or in other words, \( X_i \) with \( D_i \) partialled out. Similarly, \( \bar{Z}_i \) is \( Z_i \) with both \( D_i \) and \( X_i \) partialled out.

**Assumption 1.1 (Model and Exclusion restrictions).** i. \([1]\), \([2]\), and \([3]\) hold. ii. \( E[\varepsilon_iX_i] = 0 \). iii. \( E[\varepsilon_iZ_i] = 0 \), \( E[v_iZ_i] = 0 \), and \( E[v_iX_i] = 0 \).

**Assumption 1.2 (\( V_i \) is sufficiently informative about \( W_i \)).** \( E[W_iV_i'] \) has full row rank.

**Assumption 1.3 (\( Z_i \) is sufficiently informative about \( W_i \)).** \( E[W_i\bar{Z}_i'] \) has full row rank.

**Assumption 1.4 (Full support).** i. \( E[\bar{X}_i,\bar{X}_i'] \) is non-singular. ii. \( E[\bar{Z}_i,\bar{Z}_i'] \) and \( E[D_iD_i'] \) are non-singular.

Assumptions 1.1.i, 1.1.ii, and 1.1.iii are analogous to conditions 1., 2., and 3., in Sub-Figure 1.(a). Because we impose linearity, we only require zero partial correlations rather than conditional independence.

We can formulate an equivalent set of conditions to Assumptions 1.1.ii and 1.1.iii in terms of the model \([4]\). Recall that \( \varepsilon_i = \varepsilon_i + \chi_0v_i \). If \( E[v_iX_i] = 0 \) and \( E[v_iZ_i] = 0 \) then the remaining conditions in Assumptions 1.1.ii and 1.1.iii hold if and only if \( E[e_iX_i] = 0 \) and \( E[e_iZ_i] = 0 \).

Assumption 1.2 requires that once we have accounted for \( X_i \) and \( D_i \), \( V_i \) 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. Note that the assumption is equivalent to 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. 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 can only hold if the vectors $Z_i$ and $V_i$ each have weakly larger dimension than $W_i$.

Assumption 1.4.i is a very mild condition that after partialling out the observed controls $D_i$, the treatments are not perfectly colinear. Assumption 1.4.ii is without loss of generality for the identification results because we can ignore any linearly dependent components of $Z_i$ and $D_i$.

**Theorem 1.** Under Assumptions 1.1-1.4 $\beta_0$ and $d_W$ are identified. In particular, let $d_W \leq r \leq d_V$, then there exists a vector $\beta$ and matrices $A$, $B$, $C$, and $G$ of respective dimensions $1 \times r$, $d_V \times r$, $r \times d_Z$, and $r \times d_X$, with $\text{rank}(B) \leq r$ so that:

$$E\left[\left(\begin{array}{c}
\tilde{V}_i - \beta'\tilde{X}_i \\
(BC\ AC \ AG) \begin{pmatrix}
\tilde{Z}_i \\
\tilde{X}_i
\end{pmatrix}
\end{array}\right)\begin{pmatrix}
\tilde{Z}_i' \\
\tilde{X}_i'
\end{pmatrix}\right] = 0 \quad (5)$$

For any such a solution, $\beta = \beta_0$, and $d_W$ is the smallest value of $r$ such that a solution exists.

Theorem 1 provides a characterization of $\beta_0$ and $d_W$ in terms of a set of moment conditions. The moment conditions involve nuisance parameters $A$, $B$, $C$, and $G$, which are not uniquely determined. One solution to the moment conditions is $\beta = \beta_0$, $A = A_0$, $B = B_0$, $C = C_0$, and $G = G_0$, where $\beta_0$, $A_0$, and $B_0$ are as defined [1], [2], and [3], and $C_0$ and $G_0$ are respectively the population coefficients from linear regression of $	ilde{W}_i$ on $Z_i$ and $X_i$.

$\beta_0$ could be estimated directly from the moment conditions in Theorem 1 using the Generalized Method of Moments (GMM) [Hansen (1982)]. However, such an estimation method is computationally challenging. It involves minimizing a GMM objective jointly over $\beta$ and matrices $A$, $B$, $C$, and $G$. Further complicating matters, the moment conditions are non-linear in parameters, the solutions $A$, $B$, $C$, and $G$ are only unique up to non-singular transformations, and (for $r \leq d_V$) the rank constraint on $B$ is non-linear.

As we discuss in Section 3, a more computationally expedient estimator can be attained using an alternative characterization of $\beta_0$ and $d_W$ that is equivalent to that in Theorem 1. The equivalent characterization is captured in Corollary 1 below.

**Corollary 1.** Under Assumptions 1.1-1.4 $\beta_0$ and $d_W$ are identified. In particular, if and only if $\beta = \beta_0$ then there exists $M \in \mathbb{R}^{d_V \times (d_Z + d_X)}$ and $\xi \in \mathbb{R}^{d_V}$ so that the moment conditions below hold:

$$E[(\tilde{V}_i - M(\tilde{Z}_i', \tilde{X}_i'))(\tilde{Z}_i', \tilde{X}_i')] = 0 \quad (6)$$

$$E[(\tilde{Y}_i - \beta'\tilde{X}_i - \xi'\tilde{M}(\tilde{Z}_i, \tilde{X}_i))(\tilde{Z}_i', \tilde{X}_i')) = 0 \quad (7)$$

The unique solution $M_0$ to (6) has $\text{rank}(M_0) = d_W$. Moreover, there exists a $\xi$ with $||\xi||_0 \leq d_W$ so that $\beta_0$, $M_0$, and $\xi$ satisfy (7).

$\beta_0$ is the unique value of $\beta$ that satisfies the moment conditions in Corollary 1 for some $M$ and $\xi$. There is a unique matrix $M$ that satisfies the moment conditions, and we refer to this solution as $M_0$, and note that $M_0 = B_0(C_0, G_0)$. If $d_W \leq d_V$ then there is not a unique choice of $\xi$ with $||\xi||_0 \leq d_W$ that satisfies the moment conditions. However, it is useful to refer to one particular solution $\xi_0$ which satisfies the moment conditions, we let $\xi_0$ be the (generically unique) solution with minimal $\ell_1$ norm. Note that if the minimal $\ell_1$ solution is unique then $||\xi_0||_0 \leq d_W$. $M_0$ and $\xi_0$ are nuisance parameters and are not of direct interest.²

²$||v||_0$ is the number of non-zero entries in the vector $v$. 
Corollary 1 suggest two different means of adapting to the number of confounding factors \(d_W\). Firstly, \(d_W\) is the rank of \(M_0\). Secondly, there is a solution \(\beta, M, \xi\) to the moment conditions where \(\xi\) has at most \(d_W\) non-zero entries. We exploit these results in Subsection 3.2.

2.1 Discussion

To the best of our knowledge, the characterization of \(\beta_0\) and the number of confounders \(d_W\) using the moment condition in (5) is original, and similarly for (6) and (7).

The characterization of \(\beta_0\) is distinct from the characterization using instrumental variables-type moment conditions in Miao et al. (2018a) and implicit in Griliches (1977). As we discuss below, when our assumptions hold and \(r < d_V\), our characterization provides additional restrictions that help identify \(\beta_0\).

First let us compare with Miao et al. (2018a). For simplicity let us assume there are no additional controls \(D_i\). Miao et al. (2018a) assume the existence of a function called a ‘confounding bridge’ which then plays a key role in their analysis. A confounding bridge 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 \(\epsilon_i\) and \(v_i\) are mean independent of \(W_i\) (rather than just uncorrelated with \(W_i\)), then our model admits a confounding bridge of the form \(b(v, x) = \beta_0' x + A_0(B_0' QB_0)^{-1} B_0' Q v\), where \(Q\) is any non-singular matrix and \(A_0\) and \(B_0\) are as defined in (2) and (3).

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 may depend on the matrix \(Q^4\). Even if the confounding bridge is unique, in order to identify the bridge, \(Z_i\) must be 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 \(d_V = d_W\).

Applying Miao et al. (2018a) in our model amounts to using GMM to estimate solutions \(\beta\) and \(\gamma\) to the following moment condition \(\S\):

\[
E[(Y_i - \beta' X_i - \xi(V_i))(Z_i', X_i')] = 0
\]

Under Assumptions 1.1-1.4, the condition above is satisfied when \(\beta = \beta_0\) and \(\xi = A_0(B_0' QB_0)^{-1} B_0' Q\) for any non-singular \(Q\). When \(d_W < d_V\), the solution \(\xi\) is generally not unique, however \(\beta_0\) may still be point identified by the moment condition. \(\S\) is a standard instrumental variables (IV) moment condition. Griliches (1977) suggests a method for estimation with scalar proxy controls that amounts to performing a standard IV procedure to empirically solve \(\S\).

\(^4\) Under Assumptions 1.1-1.4 \(B_0\) has full column rank and so \(B_0' QB_0\) is non-singular. See the proof of Theorem 1.

\(^5\) Under Assumptions 1.1-1.4 \(B_0' B_0\) is non-singular. If \(B_0' B_0\) is non-singular then \((B_0' QB_0)^{-1} B_0' Q = (B_0' B_0)^{-1} B_0' Q\) for all non-singular \(Q\) if and only if \(B_0\) is a square matrix, i.e., \(d_W = d_V\). Strictly speaking, even if \(d_W \neq d_V\) the confounding bridge may be unique for certain values of \(A_0\) (for example, if \(A_0\) is a matrix of zeros).

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, if \(q\) is nonlinear then the resulting moment conditions are valid only when \(\epsilon_i\) and \(v_i\) are mean independent of \(W_i\) rather than just uncorrelated with \(W_i\).
Corollary 1 imposes additional structure compared to (5). Firstly, Corollary 1 shows that there is a $\xi$ with $d_W$ non-zero entries that satisfies the moment condition. Secondly, if (6) in Corollary 1 holds, then (8) is equivalent to (7). Thus Corollary 1 supplements (6) with an additional set of moment conditions (9). (6) provides additional identifying power when $d_W < \min\{d_V, d_Z + d_X\}$ (there are fewer confounding factors than proxies). If $d_W$ is unknown, then (6) helps to identify $d_W$, because $d_W$ is rank of the lowest-rank solution $M$ to (6).

Conversely, if $d_W = d_V$ then the additional moment conditions (6) do not help identify $\beta_0$ because there is no reduced-rank restriction on the solution $M$ to (6). Further, if $d_W = d_V$, then the restriction that $\xi$ have only $d_W$ non-zero entries is trivial.

In sum, the results in Theorem 1 and Corollary 1 provide additional identifying power over existing results whenever there are fewer confounding factors than proxies, and also help identify the number of confounding factors. Our results show precisely how the number of latent factors implies sparsity and low-rank structure within the nuisance parameters.

3 Estimation

For a given $r$, we could estimate $\beta_0$ by GMM from the moment condition (5) in Theorem 1. However, the moment condition is non-linear in parameters, the objective is generally non-convex, the rank restriction on $B$ is non-linear when $r \leq d_V$, and the solutions $A, B, C, D$ are (at most) unique up to a normalization. Thus direct minimization of the GMM objective by standard numerical methods may be computationally infeasible, particularly in high dimensions. This problem also applies when $r$ is chosen using model selection (in which case we select an $r$ that estimates $d_W$).

In order to avoid the computational difficulty of joint GMM estimation using (5), we suggest a sequential method of moments estimator (see Newey & McFadden (1994)) based on the moment conditions in Corollary 1. We use the first set of moment conditions (6) to estimate the solution $M_0$. We plug the estimate into the moment conditions (7) and use these to estimate the remaining coefficients, including $\beta_0$. If $d_W \leq \min\{d_V, d_Z + d_X\}$ then the solution $M_0$ to (6) has a reduced-rank structure. Imposing this reduced-rank structure in estimation reduces the number of free nuisance parameters that must be estimated. For estimation of $M_0$, we apply methods from reduced-rank regression. These methods provide closed-form solutions, even when the rank is unknown and must be chosen by model selection. We can also impose a sparse structure on an estimate of $\xi_0$ (recall $\xi = \xi_0$ satisfies the moment conditions in Corollary 1 and $||\xi_0||_0 \leq d_W$).

To induce sparsity we use $\ell_1$-penalization. In the sub-section on adaptive estimation we propose an estimator with a closed-form that does not induce sparsity in the estimate of $\xi_0$, and also a doubly-robust (but more computationally burdensome) estimator that does impose that $\xi_0$ is sparse.

Let us introduce some notation. Let $n$ be the number of available observations, let $X$ be the matrix with $n$ rows whose $i^{th}$ row is equal to $X_i'$ and similarly for $Y, V, Z,$ and $D$. Recall that the partialled out variables $\bar{X_i}, \bar{Z_i}, \bar{V}_i,$ etc., depend on population linear regression parameters. In estimation we must replace these population coefficients with...
sample analogues. For each \( H = V, Z, Y, X \) we define the objects below:

\[
\hat{\gamma}_H = (D'D)^+D'H \\
\hat{\omega}_H = ((X, D)'(X, D))^+(X, D)'H \\
\hat{H} = H - D\hat{\gamma}_H \\
\hat{H} = H - (X, D)\hat{\omega}_H
\]

We also let \( \hat{H}_i \) be the transpose of the \( i \)th column vector of \( \hat{H} \) and \( \hat{H}_i \) the transpose of the \( i \)th column vector of \( \hat{H} \). Thus \( \hat{\gamma}_X \) is a sample estimate of \( \gamma_{X,0} \), \( \hat{X}_i \) is an estimate of \( X_i \), and similarly for the other variables.

### 3.1 Estimation With a Given Rank Restriction

Corollary 1 states that \( M_0 \) has rank equal to \( d_W \), the number of latent confounding factors. If \( r \) upper bounds \( d_W \) then \( \text{rank}(M_0) \leq r \). In this sub-section we consider estimation under this restriction on the rank of \( M \) for a fixed choice of \( r \). Of particular interest is the case of \( d_W \) known, then we can set \( r = d_W \).

To estimate \( \beta_0 \) we first find a rank \( \leq r \) matrix \( \hat{M}_r \) that approximately solves an empirical analogue of (6) in Corollary 1. In particular, we let \( \hat{M}_r \) solve the minimization problem below:

\[
\hat{M}_r = \underset{\text{rank}(M) \leq r}{\arg \min} \| \hat{V} - (\hat{Z}, \hat{X})M' \|_F^2
\]  

(9)

Where \( \| \cdot \|_F^2 \) is the squared Frobenius norm (the sum of the squared entries of the matrix). \( \hat{M}_r \) has a closed-form solution described in Reinsel & Velu \cite{reinsel1998} and originally due to Izenman \cite{izenman1975}. To describe the solution, let \( \hat{\Omega} = (\hat{Z}, \hat{X})'(\hat{Z}, \hat{X}) \) and define a matrix \( \hat{Q} \) given by:

\[
\hat{Q} = \hat{V}'(\hat{Z}, \hat{X})\hat{\Omega}^+(\hat{V}'(\hat{Z}, \hat{X}))'
\]  

(10)

Let \( \hat{E} = \text{eigen}(\hat{Q}) \) be the matrix whose columns are the right eigenvectors of \( \hat{Q} \) normalized so that \( \hat{E}\hat{E}' \) is the identity and ordered so that the \( k \)th column of \( \hat{E} \) corresponds to the \( k \)th largest eigenvalue. Let \( \hat{E}_{[1:r]} \) denote the sub-matrix of the first \( r \) columns of \( \hat{E} \). Then we have:

\[
\hat{M}_r = \hat{\Omega}^+(\hat{Z}, \hat{X})'(\hat{V}'(\hat{Z}, \hat{X}))'
\]  

(11)

Having evaluated \( \hat{M}_r \) we solve an empirical analogue of (7) with \( M \) in the moment condition replaced by \( \hat{M}_r \). In particular, our estimate of \( \beta_0 \) is the vector \( \hat{\beta}_r \) that minimizes the least-squares objective below:

\[
\hat{\beta}_r = \underset{\beta \in \mathbb{R}^d_X}{\arg \min} \| \hat{Y} - \hat{X}\beta - (\hat{Z}, \hat{X})\hat{M}_r'\xi \|_2^2
\]  

(12)

Where \( \| \cdot \|_2 \) is the Euclidean norm.

Corollary 1 states that there exists \( \xi \) that satisfies the moment conditions with \( \| \xi \|_0 \leq r \). We could impose this restriction by adding \( \| \xi \|_0 \leq r \) as a constraint in the minimization problem (12). That is, instead of taking the minimum over \( \xi \in \mathbb{R}^{d_V} \) we could take the minimum over \( \xi \in \mathbb{R}^{d_V} : \| \xi \|_0 \leq r \). However, if \( \hat{M}_r \) is of rank \( r \), then adding this restriction has no effect on the resulting estimator of \( \beta_0 \). This is because for any vector \( \xi \) and rank-\( r \) matrix \( M \), \( M'\xi \) always equals \( M'\xi^* \) for some vector \( \xi^* \) with only \( r \) non-zero components and vice-versa.
The minimizer $\hat{\beta}_r$ in \cite{12} has a closed-form given below.

$$\hat{\beta}_r = (I_{d_X}, 0_{d_X \times d_V}) \left( \tilde{J}_r(\tilde{Z}, \tilde{X})' \hat{\Omega}_r(\tilde{Z}, \tilde{X}) \right)^+ \tilde{J}_r(\tilde{Z}, \tilde{X})' \hat{Y}$$

$I_{d_X}$ is the $d_X \times d_X$ identity matrix, $0_{d_X \times d_V}$ is a $d_X \times d_V$ matrix of zeros, and $\tilde{J}_r = \left( \begin{pmatrix} 0_{d_X \times d_V} \\ I_{d_X} \end{pmatrix}, M_r \right)$. The complete procedure is detailed step-by-step in Algorithm 1 below.

Algorithm 1 Sequential estimate with a fixed $r$

Returns $\hat{\beta}_r$, the estimate of $\beta_0$.

```plaintext
1: for $H = W, V, Z, Y, X$ do
2: \[ \tilde{\gamma}_H \leftarrow (D'D)^+ D'H \]
3: \[ \tilde{H} \leftarrow H - D\tilde{\gamma}_H \]
4: end for
5: \[ \tilde{\Omega} \leftarrow (\tilde{Z}, \tilde{X})' (\tilde{Z}, \tilde{X}) \]
6: \[ \tilde{Q} \leftarrow \tilde{V}''(\tilde{Z}, \tilde{X}) \tilde{\Omega}^{+} (\tilde{V}'(\tilde{Z}, \tilde{X}))' \]
7: \[ \tilde{E} \leftarrow \text{eigen}(\tilde{Q}) \]
8: \[ M_r = \tilde{\Omega}^{+}(\tilde{Z}, \tilde{X})' \tilde{V} \tilde{E}_{[1, r]} \tilde{E}'_{[1, r]} \]
9: \[ \tilde{J}_r \leftarrow \left( \begin{pmatrix} 0_{d_X \times d_V} \\ I_{d_X} \end{pmatrix}, M_r \right) \]
10: \[ \hat{\beta}_r \leftarrow (I_{d_X}, 0_{d_X \times d_V}) \left( \tilde{J}_r(\tilde{Z}, \tilde{X})'(\tilde{Z}, \tilde{X}) \tilde{J}_r \right)^+ \tilde{J}_r(\tilde{Z}, \tilde{X})' \hat{Y} \]
```

3.2 Adaptive Estimation

Algorithm \cite{4} applies for a specific choice of rank restriction $r$ on the estimate of $M_0$. A smaller value of $r$ results in a greater dimension reduction in the nuisance parameters one must estimate. However, if $r < d_W$ (where $d_W$ is the number of unobserved confounders) then the moment conditions are misspecified, i.e., they have no solution. Ideally, $r$ would be chosen to equal $d_W$ which ensures the number of free parameters is minimized and the moment condition correctly specified. However, $d_W$ is generally unknown. We can choose the rank restriction $r$ by model selection and thus adapt to the unknown quantity $d_W$. These methods also result in an estimate of $d_W$, which may be of interest in itself.

We perform model selection by adding penalty terms in estimation so as to induce a low-rank structure on our estimate of $M_0$ and/or a sparse structure on our estimate of $\xi_0$ (recall $\xi_0$ is a choice of $\xi$ that satisfies the moment conditions and has $d_W$ non-zero entries). If we only apply a penalty in the estimation of $M_0$ and use a particular penalty described below, the resulting estimator has a closed-form. If we penalize in both the estimation of $M_0$ and $\xi_0$ this generally results in an estimator without a closed-form solution, but the required numerical minimization can be performed using a standard Lasso estimator \cite{Tibshirani1996}. The combination of both penalties results in a doubly-robust estimation method at the cost of additional computation.

In this sub-section we present both an estimator with a closed form that only applies penalization in the estimation of $M_0$, and a doubly-robust estimator without a closed-form that applies penalization in estimation of both $M_0$ and $\xi_0$.

Doubly-robust estimators are insensitive to estimation error in any one of a number of nuisance parameters. They are Neyman Orthogonal/Locally Robust, and generally allow
for valid inference using a standard root-n Gaussian approximation under weaker conditions than estimators that are not robust (see for example Chernozhukov et al. (2016) and Chernozhukov et al. (2018)).

Let us first describe the estimator of $M_0$ with an adaptive rank restriction. We replace the first-stage objective (9) with a penalized objective which we minimize over all $d_V \times (d_Z + d_X)$ matrices $M$ (rather than those with rank weakly less than $r$). The penalized objective is:

$$||\hat{V} - (\hat{Z}, \hat{X})M'||_F^2 + \lambda_n pen(M)$$

Where $pen(\cdot)$ is a penalty function and $\lambda_n$ is a scalar penalty parameter that may depend on the data. The penalty function should induce a low-rank structure in the minimizer, for example one could let $pen(M)$ return the nuclear norm of $M$, which is the sum of the singular values of $M$. We focus on the case in which $pen(M)$ simply returns the rank of the matrix $M$. The corresponding objective is given below:

$$||\hat{V} - (\hat{Z}, \hat{X})M'||_F^2 + \lambda_n rank(M)$$

Minimization of the objective above is a penalized reduced-rank regression problem as considered in Bunea et al. (2011), and admits a closed-form solution. Bunea et al. (2011) show that the solution $\hat{M}$ is a matrix of rank $\hat{r}$, where $\hat{r}$ is the number of eigenvalues of $\hat{Q}$ that exceed $\lambda_n$. Recall that $\hat{Q}$ is defined in (10). $\hat{M}$ is equal to $\hat{M}_r$ defined in (11) but with $r$ set to the data-dependent $\hat{r}$. Note that $\hat{r}$ can be understood as an estimator of the number of unobserved confounders $d_W$.

Having obtained the matrix $\hat{M}$ that minimizes the penalized objective, we may proceed as in the previous subsection to attain an estimate $\hat{\beta}$ of $\beta_0$. The full algorithm is given in detail Algorithm 2 below. The first five steps provide the estimates $\hat{r}$ and $\hat{M}$. The procedure does not require any numerical optimization nor simulation.

Algorithm 2 Sequential estimate with rank selection

Returns estimates $\hat{\beta}_r$ of $\beta_0$ and $\hat{r}$ of $d_W$.

1: for $H = W, V, Z, Y, X$ do
2: $\hat{H} \leftarrow (D^D)^+D'H$
3: $\hat{H} \leftarrow H - D\hat{H}$
4: end for
5: $\hat{\Omega} \leftarrow (\hat{Z}, \hat{X})'(\hat{Z}, \hat{X})$
6: $\hat{Q} \leftarrow V'(\hat{Z}, \hat{X})\hat{\Omega}^+(\hat{V}'(\hat{Z}, \hat{X}))'$
7: $\hat{E} \leftarrow eigen(\hat{Q})$
8: $\hat{r} \leftarrow \#\{\text{eigenvalues of } \hat{Q} \geq \lambda_n\}$
9: $M \leftarrow \hat{\Omega}^+(\hat{Z}, \hat{X})'\hat{V}\hat{E}_{1:1:1:1:1}{\hat{E}_1:1:1:1:1}^T$
10: $\hat{J} \leftarrow \left(\begin{array}{c} 0_{d_X \times d_V} \\ \hat{M} \end{array}\right)$
11: $\hat{\beta}_r \leftarrow (I_{d_X}, 0_{d_X \times d_Y})(\hat{J}'(\hat{Z}, \hat{X})'(\hat{Z}, \hat{X})\hat{J})^+\hat{J}'(\hat{Z}, \hat{X})\hat{Y}$

In our simulations we select the penalty parameter $\lambda_n$ by cross-validation. More precisely, let $\{\mathcal{I}_j\}_{j=1}^J$ be a partition of the indices $\{1, ..., n\}$. Let $\mathcal{I}_{-j}$ denote all the elements of $\{1, ..., n\}$ that are not in $\mathcal{I}_j$. For a given value of $\lambda$ of the penalty $\lambda_n$, let $M_{j, \lambda}$ be the result from carrying out steps 2 to 9 in Algorithm 2 using only the data with indices in $\mathcal{I}_{-j}$. We choose
Let \( \lambda_n \) to be the value of \( \lambda \) that minimizes the cross validation objective below:

\[
\sum_{j=1}^{J} \sum_{i \in I_j} \| \hat{V}_i - \hat{M}_{j,\lambda_n}^i (\hat{Z}_i', \hat{X}_i') \|^2
\]

### 3.2.1 Doubly-Robust Estimation

Corollary 1 states that there exists a vector \( \xi_0 \) with at most \( d_W \) non-zero entries so that the moment conditions are satisfied when \( \xi = \xi_0, \beta = \beta_0, \) and \( M = M_0. \)

The estimator in Algorithm \( \mathbb{2} \) does not impose sparsity on an estimate of \( \xi_0. \) We propose a doubly-robust estimator that exploits both the sparsity in \( \xi_0 \) and the low-rank structure of \( M_0. \)

Before we describe the procedure as a whole, let us motivate an estimator of \( \xi_0 \) that induces sparsity. Using (6) to substitute out \( M \) from (7) and then partialling out \( X_i \) from the resulting equation, we get the following:

\[
E[\hat{Z}_i' (\hat{Y}_i - \xi_0^i \hat{V}_i)] = 0
\]  

(14)

Our estimate of \( \xi_0 \) is indirectly based on the moment condition above. However, it is useful to adapt the moment condition so that \( \xi_0 \) may take the form of a penalized least-squares estimator. Let \( M_0_{0,[:1:d_Z]} \) be the sub-matrix of \( M_0 \) that contains its first \( d_Z \) columns. If we substitute the definition of \( M_0 \) into the condition above and multiply both sides by \( M_{0,[:1:d_Z]}' \) then we get the following moment condition:

\[
E[M_{0,[:1:d_Z]} \hat{Z}_i (\hat{Y}_i - \xi_0^i M_0_{0,[:1:d_Z]} \hat{Z}_i)] = 0
\]

\( \xi_0 \) satisfies the condition above if and only if it minimizes the following least squares criterion:

\[
E[(\hat{Y}_i - \xi_0^i M_{0,[:1:d_Z]} \hat{Z}_i)^2]
\]

To estimate \( \xi_0 \), we minimize a penalized empirical analogue of the criterion above. In particular, our estimate of \( \xi_0 \) is the vector \( \xi \) that minimizes the empirical objective below:

\[
|| \hat{Y} - \hat{Z} \hat{M}_{d_Y,[:1:d_Z]} ||_F^2 + \delta_n || \xi ||_1
\]  

(15)

Where \( || \cdot ||_1 \) is the \( \ell_1 \) norm and \( \delta_n \) is a penalty parameter. \( \hat{M}_{d_Y,[:1:d_Z]} \) is the sub-matrix consisting of the first \( d_Z \) columns of \( \hat{M}_{d_Y} \), which is the estimator of \( M_0 \) defined in Algorithm 1 for \( r = d_Y \). Note that the estimate \( M_{d_Y} \) does not have a restricted rank.

Minimization of (15) is an \( \ell_1 \)-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, \( \delta_n \) could be chosen using cross-validation. In our simulations we normalize each of the regressors and the outcomes so that they have mean zero and unit variance and then perform Lasso with the penalty parameter equal to \( d_Y / n. \)

We now turn to the definition of the doubly-robust moment condition. \( M_0 \) only enters the doubly-robust moment condition indirectly through a parameter \( \mu_0 \) defined below:

\[
\mu_0 = E[X_i (\hat{Z}_i', \hat{X}_i')|M_0_{0,[:1:d_Z]} E[Z_i \hat{Z}_i'] M_{0,[:1:d_Z]}'] + M_{0,[:1:d_Z]}
\]

Our estimate \( \mu_0 \) is a sample analogue of the above. We replace the expectations with sample averages, we replace the partialled out variables with their sample counterparts, and
we replace $M_0$ with the adaptive reduced-rank estimate $\hat{M}$ from Algorithm 2. The estimate $\hat{\mu}$ is given below:

$$\hat{\mu} = \bar{X}'(\bar{Z}, \bar{X})M'(\hat{M}_{i[1:d_2]}^\prime \bar{Z} \hat{M}_{i[1:d_2]}^\prime + \hat{M}_{0[1:d_2]}$$

In the above $\hat{M}_{i[1:d_2]}$ is the sub-matrix that contains the first $d_2$ columns of $\hat{M}$.

We can now define the doubly-robust moment condition. The condition involves a number of nuisance parameters. These nuisance parameters are $\mu_0$, $\xi_0$, and all the nuisance parameters involved in partialling out $D_i$ and $X_i$. The nuisance parameters involved in partialling out are $\gamma_{Y,0}$, $\gamma_{Y,0}$, $\gamma_{X,0}$, $\omega_{Z,0}$, $\omega_{Y,0}$, and $\omega_{V,0}$.

The doubly-robust moment condition is as follows:

$$E[g_i(\beta_0; \xi_0, \mu_0, \gamma_{Y,0}, \gamma_{Y,0}, \gamma_{X,0}, \omega_{Z,0}, \omega_{Y,0}, \omega_{V,0})] = 0$$

The score function $g_i(\cdot)$ in the moment condition is defined below:

$$g_i(\beta; \xi, \mu, \gamma_Y, \gamma_V, \gamma_X, 1, \gamma_X, 2, \omega_Z, \omega_Y, \omega_V)$$

$$= \bar{X}_i(\gamma_X, 1)(\bar{Y}_i(\gamma_Y) - \xi' \bar{Y}_i(\gamma_V) - \beta' \bar{X}_i(\gamma_X, 2)) - \mu \bar{Z}_i(\omega_Z) (\bar{Y}_i(\omega_Y) - \xi' \bar{V}_i(\omega_V))$$

The moment condition is doubly-robust because it holds when we replace a single one of its nuisance parameter arguments with any alternative value. For example, for any vector $\xi$ (not just the correct value $\xi_0$), the moment condition holds:

$$E[g_i(\beta_0; \xi, \mu_0, \gamma_{Y,0}, \gamma_{V,0}, \gamma_{X,0}, \omega_{Z,0}, \omega_{Y,0}, \omega_{V,0})] = 0$$

We prove the validity of the moment condition (under Assumptions 1.1-1.4) and establish that it is doubly-robust in Appendix B.

The doubly robust estimate $\hat{\beta}_{DR}$ of $\beta$, sets an empirical expectation of the score function to zero when the nuisance parameters are replaced with first-stage estimates. As we discuss below, it may be advantageous to employ sample-splitting in doubly-robust estimation, but for ease of exposition we begin by defining a doubly robust estimator without sample-splitting.

Recall the definitions of the estimates $\hat{\xi}$, $\hat{\mu}$, $\hat{\gamma}_Y$, $\hat{\gamma}_V$, $\hat{\gamma}_X$, $\hat{\omega}_Z$, $\hat{\omega}_Y$, and $\hat{\omega}_V$. $\hat{\beta}_{DR}$ solves:

$$\frac{1}{n} \sum_{i=1}^{n} g_i(\hat{\beta}_{DR}; \hat{\xi}, \hat{\mu}, \hat{\gamma}_Y, \hat{\gamma}_V, \hat{\gamma}_X, \hat{\omega}_Z, \hat{\omega}_Y, \hat{\omega}_V) = 0$$

We can write the solution succinctly as follows:

$$\hat{\beta}_{DR} = \hat{\Sigma}^{-1} \left( \bar{X}' (\bar{Y} - \hat{V} \hat{\xi}) - \hat{\mu} \bar{Z}' (\bar{Y} - \hat{V} \hat{\xi}) \right) / n$$

In the above $\hat{\Sigma} = \bar{X}' \bar{X} / n$. Recent work including [Chernozhukov et al. (2018)] and [Chernozhukov et al. (2016)] shows that there may be advantages to sample-splitting in doubly robust and locally robust estimators. Below we describe a version of the estimator above that employs sample splitting.

We partition the data into $J$ sub-samples. In particular, let $\{I_j\}^J_{j=1}$ be a partition of $\{1, \ldots, n\}$ and let $n_j$ be the number of entries in $I_j$. Thus each index $i = 1, \ldots, n$ is a member

Note that $\gamma_{X,0}$ appears twice as an argument in the moment condition, this is not an error. $\gamma_{X,0}$ enters the score function in two different places, and for analytical purposes it is useful to think of these two occurrences as separate parameters.
of precisely one subset $I_j$ and $\sum_{j=1}^J n_j = n$. We will use the shorthand $I_{-j}$ to denote all the elements of $\{1, \ldots, n\}$ that are not in $I_j$ (i.e., the complement of $I_j$).

For each $j = 1, \ldots, J$ the researcher evaluates each of the nuisance parameter estimates using only the observations with indices in $I_{-j}$; that is, the data outside of the $j^{th}$ subsample. Thus, for each $j$, one evaluates estimates $\hat{\xi}_j$, $\hat{\mu}_j$, $\hat{\gamma}_{Y,j}$, $\hat{\gamma}_{V,j}$, $\hat{\gamma}_{X,j}$, $\hat{\omega}_{Z,j}$, $\hat{\omega}_{Y,j}$, and $\hat{\omega}_{V,j}$. These estimates are calculated using only data outside the $j^{th}$ subsample, but are otherwise identical to $\xi$, $\hat{\mu}$, $\hat{\gamma}_Y$, $\hat{\gamma}_V$, $\hat{\gamma}_X$, $\hat{\omega}_Z$, $\hat{\omega}_Y$, and $\hat{\omega}_V$ respectively.

The estimate $\hat{\beta}_{DR}$ with sample-splitting satisfies the formula:

$$
\frac{1}{n} \sum_{j=1}^J \sum_{i \in I_j} g_i(\hat{\beta}_{DR}; \hat{\xi}_j, \hat{\mu}_j, \hat{\gamma}_{Y,j}, \hat{\gamma}_{V,j}, \hat{\gamma}_{X,j}, \hat{\omega}_{Z,j}, \hat{\omega}_{Y,j}, \hat{\omega}_{V,j}) = 0
$$

The solution to the above is as follows:

$$
\hat{\beta}_{DR} = \hat{\Sigma}_X^{-1} \frac{1}{n} \sum_{j=1}^J \sum_{i \in I_j} (\hat{X}_{i,j} (\hat{Y}_{i,j} - \hat{V}_{i,j} \hat{\xi}_j) + \hat{\mu}_j \hat{Z}_{i,j} (\hat{Y}_{i,j} - \hat{V}_{i,j} \hat{\xi}_j))
$$

In the above $\hat{\Sigma}_X = \sum_{j=1}^J \sum_{i \in I_j} \hat{X}_{i,j} \hat{X}_{i,j}' / n$. The full doubly-robust procedure, with sample-splitting, is detailed in Algorithm 3 in Appendix A.

### 3.3 Many Additional Covariates

In some cases the vector of additional covariates $D_i$ may be high-dimensional, that is, there may be many available additional covariates. 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 $\gamma_{H,0}$ and/or $\omega_{H,0}$ may be sparse or approximately sparse for some $H \in \{V, Z, X, Y\}$.

To exploit this sparsity or approximate sparsity, we can replace the linear regression estimators $\hat{\gamma}_H$ and $\hat{\omega}_H$ with Lasso estimators. Let $\eta_{H,n,\gamma}$ and $\eta_{H,n,\omega}$ be a scalar penalty parameter. For $H = V, Z, X, Y$ we can define the following alternative estimators $\hat{\gamma}_H$ and $\hat{\omega}_H$:

$$
\hat{\gamma}_H = \text{argmin}_{\gamma \in \mathbb{R}^d} \sum_{i=1}^n (H_i - \gamma D_i)^2 + \eta_{H,n,\gamma} ||\gamma||_1
$$

$$
\hat{\omega}_H = \text{argmin}_{\omega \in \mathbb{R}^{d+g+x}} \sum_{i=1}^n (H_i - \omega (X_i', D_i'))^2 + \eta_{H,n,\omega} ||\omega||_1
$$

We can then use these estimates in place of the linear regression estimates in the algorithms detailed in this section.

A number of methods exist for choosing the penalty parameters in Lasso regression, for example cross-validation.

### 4 Inference and Consistency

The methods in the previous section estimate a low-dimensional parameter of interest $\beta_0$ (and perhaps $d_W$) 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
and strictly positive definite. We focus on the doubly-robust estimator with sample-splitting \( \hat{\beta}_{DR} \) detailed in the previous section.

The doubly-robust estimator is a Double-Machine Learning 2 (DML2) estimator of the kind analyzed in section 3.2 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. Chernozhukov et al. (2018) suggest a variance estimator for DML2 estimators. In the case of \( \hat{\beta}_{DR} \) this simplifies to:

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{J} \sum_{i \in I_j} \hat{\Sigma}_X^+ g_i \hat{g}_i \hat{\Sigma}_X^+
\]

For each \( j \in 1, ..., J \) and \( i \in I_j \), \( \hat{g}_i \) is defined as follows:

\[
\hat{g}_i = g_i(\hat{\beta}_{DR}; \hat{\theta}_j, \hat{\mu}_j, \hat{\gamma}_V, \hat{\gamma}_Y, \hat{\gamma}_X, \hat{\omega}_Z, \hat{\omega}_Y, \hat{\omega}_V)
\]

\( \hat{g}_i \) is a sample analogue of \( g_i \), which is defined by:

\[
g_i = g_i(\beta_0; \xi_0, \mu_0, \gamma_0, \gamma_V, \gamma_Y, \gamma_X, \omega_Z, \omega_Y, \omega_V)
\]

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

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

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

We now present high-level assumptions for root-\( n \) consistency and asymptotic normality of the doubly-robust estimator with sample splitting as defined in (18). Note that our results apply for any choice of estimators for the nuisance parameters not just those specified in Section 3.

Our asymptotic analysis is based on Theorems 3.1 and 3.2 in Chernozhukov et al. (2018). The Assumptions 1.1-1.4 and 4.1-4.3 (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 \( \mathcal{P}_n \).

For notational convenience, for any random column vector \( H \) we let \( \Sigma_H = E[H_i H_i'] \), however in the case of \( H_i = (X_i', D_i')' \) we write \( \Sigma_{XD} \). As in previous sections, if \( b \) is a vector then \( \|b\| \) is the Euclidean norm of \( b \). If \( A \) is a matrix then \( \|A\| \) is the Euclidean matrix norm of \( A \), so if \( A \) has \( d \) rows then \( \|A\| = \sup_{b \in \mathbb{R}^d : \|b\|=1} \|Ab\| \). For a positive semi-definite matrix \( A \), \( A^{1/2} \) is the unique positive semi-definite matrix \( B \) so that \( BB = A \) and if \( A \) is strictly positive definite then \( A^{-1/2} \) is the inverse of \( A^{1/2} \).

**Assumption 4.1 (Restrictions on the DGP).** If \( P \in \mathcal{P}_n \) the following hold with all scalars on the right-hand sides finite. i. \( \Sigma_X, \Sigma_{\bar{X}}, \Sigma_{\bar{Y}}, \Sigma_{\bar{Z}}, \Sigma_D, \) and \( \Sigma_{XD} \), are finite, and strictly positive definite. \( E[Y_i^2] > 0 \) and \( E[\bar{Y}_i^2] > 0 \). ii. \( E[Y_i^2] \leq \sigma^2_Y, \|\Sigma_X^{-1/2} \beta_0\| \leq \bar{\beta}, \|\Sigma_{\bar{X}}\| \leq \bar{\sigma}_X, \|\Sigma_{\bar{Y}}^{-1/2} \xi_0\| \leq \bar{\xi}, \|\mu_0 \Sigma_{\bar{Z}}^{-1/2}\| \leq \bar{\mu} \). iii. With probability 1 for \( H = X, V, Y \)
the following hold for 
\[ \| \Sigma_H^{-1/2} E[\tilde{H}, \tilde{H}'|D_i] \|^{1/2} \| \leq \sigma_{H|D}, \text{ for } H = Z, V, Y \quad \| \Sigma_H^{-1/2} E[\tilde{H}, \tilde{H}'|X_i, D_i] \|^{1/2} \| \leq \sigma_{H|X,D}. \]

\[ \| \Sigma_V^{-1/2} E[\tilde{V}, \tilde{V}'|\tilde{X}_i] \|^{1/2} \| \leq \sigma_{V|\tilde{X}}, \quad \| \Sigma_V^{-1/2} E[\tilde{V}, \tilde{V}'|\tilde{Z}_i] \|^{1/2} \| \leq \sigma_{V|\tilde{Z}}, \quad \text{and } E[\tilde{Y}_i^2 | \tilde{Z}_i] / E[\tilde{Y}_i^2] \leq \sigma_{\tilde{Y}|\tilde{Z}}^2. \]

iv. \[ E[\| \Sigma_H^{-1/2} D_i \|] \] and \[ E[\| \Sigma_H^{-1/2} (X_i, D_i') \|] \leq (d_D + d_X) s_X^2. \]

**Assumption 4.2 (Convergence rates of the nuisance parameter estimates).** There is a sequence \( \alpha_n \) 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. \[ \| (\tilde{\mu}_j - \mu_0) \Sigma_H^{-1/2} \| \leq \delta_{\mu,n}. \]

ii. \[ \| \Sigma_H^{-1/2} (\tilde{\xi}_j - \xi_0) \| \leq \delta_{\xi,n}. \]

iii. For \( H = X, V, Y, \)

\[ \| \Sigma_H^{-1/2} (\tilde{\gamma}_{H,j} - \gamma_{H,0}) \Sigma_H^{-1/2} \| \leq \delta_{\gamma,H,n}. \]

iv. For \( H = Z, V, Y, \)

\[ \| \Sigma_H^{-1/2} (\omega_{H,j} - \omega_{H,0}) \Sigma_H^{-1/2} \| \leq \delta_{\omega,H,n}. \]

**Assumption 4.3 (Higher order moments).** For some \( q > 2 \), if \( P \in \mathcal{P}_n \) then the following inequalities hold with the scalars on the right-hand sides finite.

i. \[ E[||X_i, V_i'|\xi_0||^q]^{1/q} \leq \tilde{S}_q, \]

\[ E[||\mu_0 Z_i V_i'|\xi_0||^q]^{1/q} \leq \tilde{S}_q, \]

\[ E[||\Sigma_X^{-1/2} X_i H_i \Sigma_H^{-1/2} ||^q] \leq d_D S_{q,H,D}, \text{ for } H = X, V, Y, \]

\[ E[||\Sigma_X^{-1/2} X_i H_i \Sigma_H^{-1/2} ||^q] \leq d_X d_H S_{q,H,D}. \]

ii. For \( H = X, V, Y, \)

\[ E[||\Sigma_X^{-1/2} (X_i, D_i') H_i \Sigma_H^{-1/2} ||^q] \leq (d_X + d_D) d_H S_{q,X,D}. \]

Assumption 4.1 imposes bounds on the magnitudes of some population objects. The bounds must apply for all sample sizes \( n \). The bounds given in the assumption do not change with the sample size apart from in the case of 4.1.iv. The upper bound \( d_D s_X^2 \) in 4.1.iv may grow with the sample size at because \( d_D \), the number of additional controls, may grow with \( n \), likewise for \((d_D + d_X) s_X^2 \). This flexibility is important because \( E[||\Sigma_H^{-1/2} D_i||^q] \) is equal to the sum of fourth moments of each of the \( d_D \) components of \( \Sigma_H^{-1/2} D_i \). Suppose the bound on \( E[||\Sigma_H^{-1/2} D_i||^q] \) were fixed, this would imply that as \( d_D \) grows, the average fourth moment of each component of \( \Sigma_H^{-1/2} D_i \) would have to go to zero. Note that the bounds in 4.1.iii are trivially satisfied under conditional homoskedasticity. In that case each of the upper bounds in 4.1.iii simply equals 1.

Assumption 4.2 simply imposes convergence rates for each of the nuisance parameter estimates. Note that the convergence rates are required to hold uniformly over sequences of DGPUs in \( \{ \mathcal{P}_n \}_{n=1}^\infty \).

Assumption 4.3 imposes some bounds on higher-order moments. This condition is required to hold for consistency of the variance estimate \( \hat{\sigma} \) and thus for asymptotic validity of the confidence intervals.

**Theorem 2.** Suppose that for each \( n, P \in \mathcal{P}_n \) so that Assumptions 1.1-1.4, Assumptions 4.1, 4.2, and 4.3 all hold, \( d_X \) is fixed, and all of the eigenvalues of \( E[g_i g_i'] \) are bounded away from zero. Suppose that the nuisance parameter estimates are all consistent. That is, \( \delta_{\mu,n} \to 0, \delta_{\xi,n} \to 0, \) for \( H = X, Y \) we have \( \delta_{\gamma,H,n} < 1 \), and for \( H = Z, Y, V \) \( \delta_{\omega,H,n} < 1 \). Moreover, suppose that:

\[ (\sqrt{d_D} + \sqrt{n}) \delta_{\gamma,X,n} (\delta_{\gamma,Y,n} + \delta_{\gamma,V,n} + \delta_{\gamma,X,n}) < 1 \] (17)

\[ (\sqrt{d_D} + \sqrt{n}) \delta_{\omega,Z,n} (\delta_{\omega,Y,n} + \delta_{\omega,V,n}) < 1 \] (18)

Finally suppose that there exits a constant \( c \) so that:

\[ c \geq d_D^{1/8} \delta_{\xi,n} + d_D^{2/8} (\delta_{\gamma,Y,n} + \delta_{\gamma,V,n} + \delta_{\gamma,X,n}) + (\delta_{\mu,n} + \delta_{\xi,n}) d_Z^{1/8} d_V^{1/8} \]

\[ + d_X^{1/8} (d_Z^{1/4} + d_D^{1/4} \delta_{\omega,Z,n}) (\delta_{\omega,Y,n} + \delta_{\omega,V,n}) \] (19)
Where \( q > 2 \) is the constant in Assumption 4.3. Then uniformly over all \( \mathcal{P}_n \), \( \hat{\beta} \) is root-\( n \) consistent and asymptotically normal:

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

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

Theorem 2 establishes uniform root-\( n \) consistency of the estimator and asymptotic validity of the confidence intervals. The theorem requires conditions (17), (18), and (19). One set of primitive conditions for (17) and (18) is that \( d_D \) grows more slowly than \( n \) and all of the nuisance parameter estimates go to zero at a strictly faster rate than \( n^{-1/4} \).

The condition (19) directly restricts the rate at which \( d_D, d_V, \) and \( d_Z \) may grow with the sample size. The condition is weaker when \( q \) is large. Indeed, if the nuisance parameter estimates are consistent then (19) must hold for a sufficiently large value of \( q \). However when \( q \) is larger then Assumption 4.3 requires the existence of even higher order moments.

5 Simulation Study

In order to assess the efficacy of the methods we present in Section 3 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:

\[
\begin{align*}
V_i &= B_0 W_i + \nu_i \\
X_i &= T_0 W_i + \epsilon_i \\
Z_i &= C_0 W_i + G_0 X_i + \eta_i \\
Y_i &= X_i \beta_0 + F_0 W_i + \chi_0 V_i + e_i \\
\end{align*}
\]

The residuals \( \nu_i, \epsilon_i, \eta_i, \) and \( e_i \) are drawn independently of each other from zero mean Gaussian distributions: \( W_i \sim N(0, I), \nu_i \sim N(0, \Sigma_V), \epsilon_i \sim N(0, \Sigma_X), \eta_i \sim N(0, \Sigma_Z), \) and \( e_i \sim N(0, \Sigma_Y) \). Note that we do not include additional controls \( D_i \) in our simulations.

In each simulation we must choose parameters \( \beta_0, B_0, C_0, G_0, T_0, F_0, \chi_0, \Sigma_Y, \Sigma_X, \Sigma_Z \). 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 \( \beta_0, B_0, C_0, G_0, T_0, F_0, \) and \( \chi_0 \) 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 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 \( p, d_W, d_X, d_V, d_Z, \) 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 \(Z\) as in \(V\). We carry out simulations for a range of choices for the remaining hyperparameters \(d_W\), \(d_V\), and \(n\).

Figure 5.1 shows the mean-squared errors of alternative estimators for a variety of different hyperparameters. The estimators that are compared are: a naive least-squares estimator that simply treats \(V\) as a set of controls, and three different proxy control estimators. These are the proxy control estimator in Algorithm 1 with \(r = d_V\) (i.e., no rank restriction), the computationally expedient adaptive estimator in Algorithm 2, and the doubly-robust estimator with sample-splitting in Algorithm 3.

The proxy control estimator with no rank restriction is equivalent to the two-stage least squares strategy of Griliches (1977) in which \(V\) is a vector of endogenous regressors, \(X\) is a vector of exogenous regressors, and \(Z\) is a vector of instruments. In all cases shows in Figure 5.1 this estimator performs worse (in terms of median squared error) than the other two proxy control estimators. The difference is particularly stark when the ratio of the number of proxies to the number of confounders is large.

In the left-most sub-figures in Figure 5.1, the number of proxies in each group is the same as the number of unobserved confounders. Therefore there is no reduced-rank structure or sparsity in the true parameters. Nonetheless, the model selection estimators out-perform the unrestricted estimator. This may be due to the regularizing effect of the reduced-rank regression and \(\ell_1\) penalization.

Figure 5.1: Simulated Median Squared Errors

Median Squared Errors on the y-axes are the medians of \(||\hat{\beta} - \beta_0||^2\) over 500 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 \(d_Z\).
indeed the mean squared error of this estimator is stable for all but the smallest sample sizes shown in the figures. Nonetheless, the estimator regularly out-performs the unrestricted proxy control estimator in small samples, and it outperforms this estimator even in the largest samples in the most high dimensional case where $d_V = d_Z = 100$. The naive method involves estimation of a $d_V$-dimensional nuisance parameter, whereas in the unrestricted proxy control method the nuisance parameters have dimension $d_V (1 + d_Z + d_X)$. The larger dimension of the nuisance parameters may lead to high variance of the unrestricted proxy control method, which may dominate the bias in the naive estimator in finite samples.

The naive estimator out-performs all the other methods when $d_W = d_V = d_Z = 20$, even in large samples. When $d_W = d_V = d_Z = 10$ the naive estimator out-performs the other methods for sample-sizes in the bottom half of the range we examine, but performs worse than all three proxy methods in the largest samples. When the number of proxies in each group exceeds the number of confounders, the adaptive proxy control methods substantially outperformed the naive estimator in all but the very smallest sample sizes we examine.

Figure 5.2: Simulated Confidence Interval Coverage
Confidence interval coverage of the treatment parameter. on the y-axes are percentages of 500 simulated datasets in which confidence intervals contain $\beta_0$. The different figures correspond to different choices for the number of confounding factors $d_W$, and the numbers of proxies $d_V$ and $d_Z$.

Figure 5.2 shows the percentage of simulations in which 99%, 95%, and 90% confidence intervals cover the true parameter $\beta_0$ (recall $\beta_0$ 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 4.

We see from the left-most sub-figures that the confidence intervals greatly under-cover when the number of proxies in $V_i$ and $Z_i$ equal the number of confounding factors. When there are more proxies than confounders the coverage is close to the desired levels, particularly in large samples.

In Table 1 we give the proportion of simulations in which the rank $r$ selected according to Algorithm 2, is equal to the number of confounders $d_W$ (which is the rank of the matrix $M_0$). In each row we see that the probability of correctly selecting the rank increases with
the sample size. For all choices of $d_W$, $d_V$, and $d_Z$ other than those in the first two rows of the table, the algorithm selects the correct rank in at least 87% of simulations when the sample size is 10,000 or higher, and at least 69% of the time when the sample size is 5000 or higher.

However, when $d_V = d_Z = d_W = 10$ the algorithm selects the incorrect rank in over 80% of simulations, even with 100,000 observations. When $d_V = d_Z = d_W = 20$ the incorrect rank is chosen over 90% of the time. This may explain the under-coverage when $d_V = d_Z = d_W$ as seen in Figure 5.2. In these cases the algorithm regularly selects a rank that is smaller than the true rank (the algorithm cannot select a rank smaller than the minimum of $d_V$ and $d_Z$), and this leads to bias in the resulting coefficient estimates. This bias then leads to under-coverage of the confidence intervals, and because the performance of the rank selection only gradually increases with the sample size, this bias decreases more slowly than the variance, leading to increasingly poor coverage as the sample size grows.

Table 1: Frequency of Correct Rank Selection

| $d_W$ | $d_V = d_Z$ | 1000 | 5000 | 10000 | 50000 | 100000 |
|-------|-------------|------|------|-------|-------|--------|
| 10    | 10          | 0.048| 0.062| 0.078 | 0.108 | 0.138  |
| 20    | 20          | 0.068| 0.066| 0.072 | 0.076 | 0.088  |
| 10    | 20          | 0.478| 0.922| 0.984 | 0.992 | 0.986  |
| 20    | 40          | 0.204| 0.696| 0.876 | 1.000 | 1.000  |
| 10    | 50          | 0.700| 0.998| 1.000 | 1.000 | 1.000  |
| 20    | 100         | 0.062| 0.940| 1.000 | 1.000 | 1.000  |

Figures are the proportion of the 500 simulated datasets in which the $r$ selected by Algorithm 2 is equal to $d_W$. Rows correspond to different choices of $d_W$, $d_V$, and $d_Z$, columns correspond to different choices of the sample size $n$.

6 Conclusion

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

References

Bunea, Florentina, She, Yiyuan, & Wegkamp, Marten H. 2011. Optimal selection of reduced rank estimators of high-dimensional matrices. The Annals of Statistics, 39.
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. 2018. Nonparametric Estimation and Identification in Non-Separable Models Using Panel Data. Sept.

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

Deaner, Ben. 2020. Proxy Controls and Panel Data. Oct.

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

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

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.

Newey, Whitney K., & McFadden, Daniel. 1994. Chapter 36 Large sample estimation and hypothesis testing, Handbook of Econometrics.

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 Algorithm 3

Below is a full step-by-step description of the doubly robust estimation procedure with sample splitting.

Algorithm 3 Doubly-robust estimation with sample-splitting

Returns estimates \( \hat{\beta}_{DR} \) of \( \beta_0 \).

1: for \( j = 1, 2, \ldots, J \) do
2:   for \( H = W, V, Z, Y, X \) do
3:     \( \hat{\gamma}_{H,j} \leftarrow (\sum_{i \in I_{-j}} D_i D_i') + \sum_{i \in I_{-j}} D_i H_i \)
4:     \( \hat{\omega}_{H,j} \leftarrow (\sum_{i \in I_{-j}} (X_i', D_i')(X_i', D_i')) + \sum_{i \in I_{-j}} (X_i', D_i')' H_i \)
5:   for \( i \in I_j \) do
6:     \( \hat{H}_i \leftarrow \hat{H}_i - \hat{\gamma}_{H,j} D_i \)
7:     \( \hat{H}_i \leftarrow \hat{H}_i - \hat{\omega}_{H,j} (X_i', D_i)' \)
8:   end for
9: end for
10: \( \hat{\Omega}_j \leftarrow \sum_{i \in I_{-j}} (\hat{Z}_{j,i}', \hat{X}_{j,i}') (\hat{Z}_{j,i}', \hat{X}_{j,i}) \)
11: \( \hat{Q}_j \leftarrow \sum_{i \in I_{-j}} \hat{V}_{j,i} (\hat{Z}_{j,i}', \hat{X}_{j,i}) \hat{\Omega}_j^{-1} \sum_{i \in I_{-j}} \hat{V}_{j,i} (\hat{Z}_{j,i}', \hat{X}_{j,i})' \)
12: \( \hat{E}_j \leftarrow \text{eigen}(\hat{Q}_j) \)
13: \( \hat{r}_j \leftarrow \#\{\text{eigenvalues of } \hat{Q} \geq \lambda_n\} \)
14: \( \hat{M}_j \leftarrow \hat{\Omega}_j^{-1} \sum_{k \in I_{-j}} (\hat{Z}_{j,i}', \hat{X}_{j,i})' \hat{V}_{j,i} \hat{E}_{j,k}^{-1} \hat{E}_{j,k}' \)
15: \( \hat{\gamma}_j \leftarrow \sum_{i \in I_{-j}} \hat{X}_{j,i} (\hat{Z}_{j,i}', \hat{X}_{j,i}) \hat{M}_j' (\hat{M}_j, [i,j],1:d_v (\hat{Z}_{j,i}, \hat{M}_{j}, [i,j],1:d_v))' (\hat{M}_j', [i,j],1:d_v) \)
16: \( \hat{M}_{j,dv} \leftarrow \hat{\Omega}_j^{-1} \sum_{i \in I_{-j}} (\hat{Z}_{j,i}', \hat{X}_{j,i})' \hat{V}_{j,i} \)
17: \( \hat{\xi}_j \leftarrow \text{argmin}_{\xi \in \mathbb{R}^{dv}} \sum_{i \in I_{-j}} (\hat{Y}_i - \hat{Z}_{j,i} \hat{M}_{j,dv}, [i,j],1,d_v) \xi)^2 + \delta_n ||\xi||_1 \)
18: \( \hat{\Sigma}_X \leftarrow \sum_{j=1}^J \sum_{i \in I_j} \hat{X}_{i,j} \hat{X}_{i,j}' / n \)
19: end for
20: \( \hat{\beta}_{DR} \rightarrow \hat{\beta} + \frac{1}{n} \sum_{j=1}^J \sum_{i \in I_j} (\hat{X}_{i,j} (\hat{Y}_{i,j} - \hat{V}_{i,j} \hat{\xi}_j)) - \hat{\mu}_j \hat{Z}_{i,j} (\hat{Y}_{i,j} - \hat{V}_{i,j} \hat{\xi}_j)) \)

B Proofs

Proof Theorem 1. Let \( C_0 = E[\hat{Z}_i \hat{Z}_i']^{-1} E[\hat{Z}_i \hat{W}_i'] \) and \( G_0' = E[\hat{X}_i \hat{X}_i']^{-1} E[\hat{X}_i \hat{W}_i] \) (the inverses are well defined under Assumption 1.4), then we can write:

\[
\hat{W}_i = C_0 \hat{Z}_i + G_0 \hat{X}_i + a_i
\]

Where \( E[a_i] = 0 \) and \( E[\hat{Z}_i a_i'] = E[\hat{X}_i a_i'] = 0 \). Note that by the properties of partialling out \( E[\hat{Z} W'] = E[\hat{X} W'] \), and so under Assumption 1.3 \( C_0 \) has full row rank.

Partialling out \( D_i \) from both sides of (3) and using \( E[D_i v_i'] = 0 \) we get:

\[
\hat{V}_i = B_0 \hat{W}_i + v_i
\]

Under Assumptions 1.2.iii and 1.3 \( B_0 \) has full column rank. To see this note that \( E[X_i v_i'] = 0 \) and \( E[D_i v_i'] = 0 \), which implies \( E[X_i v_i'] = 0 \) and so \( \hat{V}_i = \hat{V}_i \). Moreover,
by properties of partialling out, \( E[\tilde{V}W'] = E[\tilde{V}\tilde{W}'] \). So \( E[\tilde{V}W'] = E[\tilde{V}\tilde{W}'] \), and thus \( B_0 = E[\tilde{V}W']E[WW']^{-1} \). By Assumption 1.2 \( E[WW'] \) has full column rank, and thus \( B_0 \) has full column rank.

Using (2) to substitute for \( U_i \) in equation (1), we get:

\[
Y_i = X'_i\beta_0 + A_0W_i + L_0D_i + \epsilon_i
\]

Partialling out \( D_i \) from both sides of the above and using that \( E[D_i\epsilon_i] = 0 \) we get:

\[
\tilde{Y}_i = \tilde{X}'_i\beta_0 + A_0\tilde{W}_i + \epsilon_i
\]  

(22)

We will use equations (20), (21), and (22), and Assumption 1.2 to derive a set of moment conditions.

Substituting (20) for \( \tilde{W}_i \) in (21) we get:

\[
\tilde{V}_i = (B_0C_0)\tilde{Z}_i + (B_0G_0)\tilde{X}_i + (B_0a_i + \nu_i)
\]

And so, since \( E[\tilde{X}_i\tilde{Z}_i'] = 0 \):

\[
E[\tilde{X}_i\tilde{V}_i'] = E[\tilde{X}_i\tilde{X}_i']G'_0B'_0 + E[\tilde{X}_i\nu_i']
\]  

(23)

Where the second line follows from Assumption 1.1.iii and \( E[D_i\nu_i'] = 0 \). Furthermore we see:

\[
E[\tilde{Z}_i\tilde{V}_i'] = E[\tilde{Z}_i\tilde{Z}_i']C'_0B'_0 + E[\tilde{Z}_i\nu_i']
\]  

(24)

Where the second line again follows from Assumption 1.1.iii and \( E[D_i\nu_i'] = 0 \).

Substituting (20) for \( \tilde{W}_i \) in (22) we get:

\[
\tilde{Y}_i = \tilde{X}'_i\beta_0 + A_0C_0\tilde{Z}_i + A_0G_0\tilde{X}_i + A_0a_i + \epsilon_i
\]

And so:

\[
E[\tilde{X}_i\tilde{Y}_i] = E[\tilde{X}_i\tilde{X}_i'](\beta_0 + G'_0A'_0) + E[\tilde{X}_i\alpha_i']A'_0 + E[\tilde{X}_i\epsilon_i]
\]  

(25)

Where the second line follows from Assumption 1.1.ii, Assumption 1.1.i which states \( E[D_i\epsilon_i] = 0 \), and \( E[\tilde{X}_i\alpha_i'] = 0 \). Furthermore:

\[
E[\tilde{Z}_i\tilde{Y}_i] = E[\tilde{Z}_i\tilde{Z}_i']C'_0A'_0 + E[\tilde{Z}_i\alpha_i']A'_0 + E[\tilde{Z}_i\epsilon_i]
\]  

(26)

Where the second line follows from Assumption 1.1.iii, \( E[D_i\epsilon_i] = 0 \) and \( E[\tilde{Z}_i\alpha_i'] = 0 \). Because \( E[\tilde{X}\tilde{Z}'] = 0 \) we can combine (23), (24), (25), and (26) into one moment condition as follows:
\[
E \left[ \left( \begin{array}{c} \tilde{V}_i \\ \tilde{Y}_i \\
\end{array} \right) - \left( \begin{array}{cc} B_0C_0 & B_0G_0 \\ A_0C_0 & A_0G_0 + \beta'_0 \\
\end{array} \right) \left( \begin{array}{c} \tilde{Z}_i \\ \tilde{X}_i \\
\end{array} \right) \right] \left( \begin{array}{c} \tilde{Z}_i' \\ \tilde{X}_i' \\
\end{array} \right) = 0
\]

Note that \( E[\tilde{Z}_i\tilde{Z}_i'] = E[\tilde{Z}_i\tilde{Z}_i'] \), and \( E[\tilde{Z}_i\tilde{X}_i'] = 0 \), so the above is equivalent to:

\[
E \left[ \left( \begin{array}{c} \tilde{V}_i \\ \tilde{Y}_i \\
\end{array} \right) - \left( \begin{array}{cc} B_0C_0 & B_0G_0 \\ A_0C_0 & A_0G_0 + \beta'_0 \\
\end{array} \right) \left( \begin{array}{c} \tilde{Z}_i \\ \tilde{X}_i \\
\end{array} \right) \right] \left( \begin{array}{c} \tilde{Z}_i' \\ \tilde{X}_i' \\
\end{array} \right) = 0 \tag{27}
\]

Moreover, \((\tilde{Z}_i', \tilde{X}_i') = (\tilde{Z}_i', \tilde{X}_i')O\) where \(O\) is the matrix defined by:

\[
O = \left( \begin{array}{cc} I_{dg} & 0 \\ -E[\tilde{X}_i\tilde{X}_i']E[\tilde{X}_i\tilde{Z}_i'] & I_{d_X} \end{array} \right)
\]

Under Assumption 1.4 this matrix is non-singular and so (27) is equivalent to

\[
E \left[ \left( \begin{array}{c} \tilde{V}_i \\ \tilde{Y}_i \\
\end{array} \right) - \left( \begin{array}{cc} B_0C_0 & B_0G_0 \\ A_0C_0 & A_0G_0 + \beta'_0 \\
\end{array} \right) \left( \begin{array}{c} \tilde{Z}_i \\ \tilde{X}_i \\
\end{array} \right) \right] \left( \begin{array}{c} \tilde{Z}_i' \\ \tilde{X}_i' \\
\end{array} \right) = 0 \tag{28}
\]

So the moment condition \([5]\) is satisfied by \(r = d_W\), \(A = A_0\), \(B = B_0\), \(C = C_0\), and \(G = G_0\), and \(\beta = \beta_0\).

We also need to show there exists a solution to \([5]\) whenever \(d_W < r \leq d_V\). We can construct a solution as follows. Firstly, let \(\beta = \beta_0\). Let the first \(d_W\) rows of \(C\) and \(G\) equal \(C_0\) and \(G_0\) respectively, let the final \(r - d_W\) rows of \(C\) and \(G\) have all entries equal to 0. Let the first \(d_W\) entries of \(A\) equal \(A_0\) and the final \(r - d_W\) entries be zeros. Finally, let the first \(d_W\) columns of \(B\) equal \(B_0\), and the final \(r - d_W\) columns be any linearly independent vectors in the orthogonal complement of the null space of \(B_0\) (such a collection must exist because \(B\) has \(d_V\) rows and \(r \leq d_V\)). Then \(B\) has full column rank, the matrices are of the required dimensions and we have \(BC = B_0C_0\), \(BG = B_0G_0\), \(AC = A_0C_0\), and \(AG = A_0G_0\) so by (28) the solution satisfies the moment condition.

Now, consider a vector \(\beta\), natural number \(r \leq \min\{d_V, d_Z\}\), and matrices \(A, B, C,\) and \(G\) of respective dimensions \(1 \times r, d_V \times r, r \times d_Z\) and \(r \times d_X\), with \(B\) and \(C\) of rank \(r\) so that:

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

We will prove the first statement of the theorem by showing that \(\beta = \beta_0\). We will then prove the second statement of the theorem by showing that \(r \geq d_W\), since we have already shown there exists a solution with \(r = d_W\), we then have that the smallest \(r\) such that a solution exists is \(d_W\).

Under Assumption 1.4 \(E[\tilde{Z}_i'\tilde{X}_i'](\tilde{Z}_i', \tilde{X}_i')\) is non-singular, so we can combine the above with (28) to get the following four equalities:

\[
\begin{align*}
B_0C_0 &= BC \tag{29} \\
B_0G_0 &= BG \tag{30} \\
A_0C_0 &= AC \tag{31} \\
A_0G_0 - \beta'_0 &= AG - \beta' \tag{32}
\end{align*}
\]
Recall that $C_0$ has full row rank and thus $C_0C_0'$ is non-singular. Define $Q = C_0'(C_0C_0')^{-1}C_0$. Post-multiplying both sides of (31) by $Q$ we get $A_0G_0 = ACQ$ and substituting this into (32) gives:

$$ACQ - \beta_0' = AG - \beta'$$

(33)

Now, post-multiplying both sides of (29) by $Q$ we get $B_0G_0 = BCQ$. Substituting into (30) $BG = BCQ$. Premultiplying both sides by $A(B'B)^{-1}B'$ (recall $B$ has full column rank and so $B'B$ is non-singular) we get $AG = ACQ$. Substituting into (33) we get $\beta = \beta_0$, as required.

Next we show $r \geq d_W$, which establishes identification of $d_W$. Post-multiplying both sides of (29) by $C_0'(C_0C_0')^{-1}$ (recall $C_0C_0'$ is non-singular), we get $B_0 = BCC_0'(C_0C_0')^{-1}$. Recall $B_0$ is of full column rank, that is $\text{rank}(B_0) = d_W$, and so $BCC_0'(C_0C_0')^{-1}$ is of rank $d_W$. For any two matrices $M_1$ and $M_2$, $\text{rank}(M_1M_2) \leq \min\{\text{rank}(M_1), \text{rank}(M_2)\}$, and so:

$$d_W = \text{rank}(BCC_0'(C_0C_0')^{-1}) \leq \text{rank}(B)$$

Since $B$ has $r$ columns, $\text{rank}(B) \leq r$, and so $r \geq d_W$. □

**Lemma 1.** Consider a vector $\beta$ and natural number $r \leq d_V$. There exist matrices $A$, $B$, $C$, and $G$ of dimensions $1 \times r$, $d_V \times r$, $r \times d_Z$, and $r \times d_X$, with $\text{rank}(B) = r$ that satisfy (3) if and only if there exists a $d_V \times (d_Z + d_X)$ matrix $M$ with $\text{rank}(M) = r$ and a vector $\xi \in \mathbb{R}^{d_V}$ with at most $r$ non-zero entries so that:

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

**Proof of Lemma 1.** First let us show the ‘only if’. Fix $r \leq d_V$ and $\beta$. Suppose $A$, $B$, $C$, and $G$ have dimensions $1 \times r$, $d_V \times r$, $r \times d_Z$ and $r \times d_X$, with $B$ and $C$ of rank $r$ and satisfy (3).

$B$ is $d_V \times r$ and has rank $r$, thus $B$ is of full row rank and so $B'B$ is non-singular for any non-singular matrix $Q$ so letting $\xi = QB(B'QB)^{-1}A'$ we have:

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

Let $M = B(C,G)$, then $M$ has rank $r$. Now, since $B'$ is $r \times d_V$ with rank $r$, it has an $r$-dimensional range and $d_V - r$ dimensional null space. So we can choose $Q$ so that it has $r$ linearly independent rows in the range of $B$ and $d_V - r$ linearly independent rows in the null-space of $B$, in which case $Q$ is non-singular and $\xi = QB(B'QB)^{-1}A'$ has at most $r$ non-zero entries (the entries corresponding to the rows of $Q$ in the null space are all zero). So the ‘if’ holds.

Now the ‘only if’. Let $M$ and $\xi$ satisfy (3) and (7) with $\text{rank}(M) = r$, $M$ of dimension $d_V \times (d_Z + d_X)$. By elementary linear algebra, any $d_V \times (d_Z + d_X)$ matrix $M$ of rank $r$ can be written as the product $M = M_1M_2$ where $M_1$ is a $d_V \times r$ matrix and $M_2$ is a $r \times (d_Z + d_X)$ matrix and $M_1$ and $M_2$ have rank $r$. So let $B = M_1$, $(C,G) = M_2$, and $A = \xi' B$ and we are done. □

**Proof of Corollary 1.** Immediately follows from Theorem 1 and Lemma 1. □
Proof that the moment condition holds under Assumptions 1.1-1.4. Recall that the moment condition is defined as follows:

\[ E[g_i(\beta_0; \xi_0, \mu_0, \gamma_0, \gamma_0, \gamma_0, \omega_0, \omega_0, \omega_0, \omega_0, \omega_0)] \]

Using the first moment equality in Corollary 1 to substitute for \( M \) in the second moment equality we get:

\[ E[(\tilde{Z}_i', \tilde{X}_i')(\tilde{Y}_i - \tilde{X}_i'\beta_0) - \mu_0E[\tilde{Z}_i(\tilde{Y}_i - \xi_0\tilde{V}_i)]] = 0 \]

Where we have set \( \beta = \beta_0 \) and \( \xi = \xi_0 \) which satisfy the moment conditions in Corollary 1. The expression above implies:

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

Next, premultiplying both sides of \( (34) \) by \( (I_d, -E[\tilde{Z}_i\tilde{X}_i'\mu][\tilde{Z}_i\tilde{Z}_i']^+) \) we get:

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

By elementary properties of partialling out we get \( E[\tilde{Z}_i\tilde{Y}_i'] = E[\tilde{Z}_i\tilde{V}_i], E[\tilde{Z}_i\tilde{V}_i] = E[\tilde{Z}_i\tilde{Z}_i], \) and \( E[\tilde{Z}_i\tilde{X}_i] = E[\tilde{Z}_i\tilde{X}_i] = 0. \) And so the moment condition above implies:

\[ E[\tilde{Z}_i(\tilde{Y}_i - \xi_0\tilde{V}_i)] = 0 \]

Combining gives the result.

\[ \square \]

Proof that the doubly-robust moment condition is doubly-robust under Assumptions 1.1-1.4. We will go through each argument of \( g_i \) in turn and show that the moment condition is robust to that argument.

**Step 1**: Prove \( E[g_i(\beta_0; \xi, \mu_0, \gamma_0, \gamma_0, \gamma_0, \omega_0, \omega_0, \omega_0, \omega_0, \omega_0)] = 0 \) for all \( \xi \)

Note that for any \( \xi \):

\[ E[g_i(\beta_0; \xi, \mu_0, \Sigma_0, \gamma_0, \gamma_0, \gamma_0, \gamma_0, \omega_0, \omega_0, \omega_0, \omega_0)] \]

\[ = E[\tilde{X}_i(\tilde{Y}_i - \tilde{V}_i'\xi - \tilde{X}_i'\beta_0)] - \mu_0E[\tilde{Z}_i(\tilde{Y}_i - \xi\tilde{V}_i)] \]

\[ = E[\tilde{X}_i(\tilde{Y}_i - \tilde{X}_i'\beta_0)] - \mu_0E[\tilde{Z}_i\tilde{Y}_i] + (\mu_0E[\tilde{Z}_i\tilde{V}_i'] - E[\tilde{X}_i\tilde{V}_i'] )\xi \]

We will show that \( E[\tilde{X}_i\tilde{V}_i'] = \mu_0E[\tilde{Z}_i\tilde{V}_i'] \) and so the expression above is independent of \( \xi \). Since the moment condition holds when \( \xi = \xi_0 \), this proves the result. Substituting the definition of \( \mu_0 \) we have:

\[ \mu_0E[\tilde{Z}_i\tilde{V}_i'] \]

\[ = E[\tilde{X}_i(\tilde{Z}_i', \tilde{X}_i')M_0(M_0, \cdot, \cdot; d)\tilde{Z}_i^\prime M_0^\prime(M_0, \cdot, \cdot; d)\tilde{Z}_i^\prime + M_0\tilde{M}d)\tilde{Z}_i^\prime] \]

By the definition of \( M_0 \):

\[ E[(\tilde{V}_i - \tilde{M}_0(\tilde{Z}_i', \tilde{X}_i'))(\tilde{Z}_i', \tilde{X}_i') = 0 \]

Post-multiplying by \( (I_d, -E[\tilde{Z}_i\tilde{X}_i']\tilde{Z}_i\tilde{Z}_i']^+) \) we get:

\[ E[(\tilde{V}_i - \tilde{M}_0(\tilde{Z}_i', \tilde{X}_i'))\tilde{Z}_i'] = 0 \]
And then by the properties of partialling out:

\[ 0 = E \left[ (\tilde{V}_i - M_0(\tilde{Z}_i', \tilde{X}_i')) \tilde{Z}_i' \right] = E \left[ (\tilde{V}_i - M_0(\tilde{Z}_i', \tilde{X}_i')) \tilde{Z}_i' \right] \\
= E \left[ (\tilde{V}_i - M_0(\tilde{Z}_i', 0_{1 \times dx})) \tilde{Z}_i' \right] \\
= E \left[ (\tilde{V}_i - M_{0,\tilde{Z}_i, \tilde{X}_i} \tilde{Z}_i') \tilde{Z}_i' \right] \\
\]

And so we have \( M_{0,\tilde{Z}_i, \tilde{X}_i} E[\tilde{Z}_i \tilde{Z}_i'] = E[\tilde{V}_i \tilde{Z}_i'] \). Using this to substitute for \( E[\tilde{V}_i \tilde{Z}_i'] \) in (35) and simplifying we get:

\[ \mu_0 E[\tilde{Z}_i \tilde{V}_i'] = E[\tilde{X}_i(\tilde{Z}_i', \tilde{X}_i') M_0] \]

Recall that \( E \left[ (\tilde{V}_i - M_0(\tilde{Z}_i', \tilde{X}_i')) (\tilde{Z}_i', \tilde{X}_i') \right] = 0 \), which implies that \( E[\tilde{V}_i \tilde{X}_i] = E[M_0(\tilde{Z}_i', \tilde{X}_i') \tilde{X}_i] \). Substituting this into the above we get:

\[ \mu_0 E[\tilde{Z}_i \tilde{V}_i'] = E[\tilde{X}_i \tilde{V}_i] \]

As desired.

Step 2: Prove \( E[g_i(\beta_0; \xi, \mu, \Sigma_0, \gamma_Y, \gamma_Y, \gamma_X, \gamma_X, \omega_Z, \omega_Z, \omega_Y, \omega_Y)] = 0 \) for all \( \mu \)

In the proof that the moment condition holds we showed that:

\[ E[\tilde{Z}_i(\tilde{Y}_i - \xi' \tilde{V}_i)] = 0 \]

And so for any \( \mu \):

\[ E[g_i(\beta_0; \xi, \mu, \gamma_Y, \gamma_Y, \gamma_X, \gamma_X, \omega_Z, \omega_Z, \omega_Y, \omega_Y)] \\
= E[\tilde{X}_i(\tilde{Y}_i - \tilde{V}_i' \xi_0 - \tilde{X}_i' \beta_0)] - \mu E[\tilde{Z}_i(\tilde{Y}_i - \xi' \tilde{V}_i)] \\
= E[\tilde{X}_i(\tilde{Y}_i - \tilde{V}_i' \xi_0 - \tilde{X}_i' \beta_0)] \\
\]

The final expression above does not depend on \( \mu \). The moment condition holds for \( \mu = \mu_0 \) and so it must hold for all \( \mu \).

Step 3: Prove robustness for the remaining arguments

By the properties of partialling out, for any \( \gamma_Y \):

\[ E[\tilde{X}_i \tilde{Y}_i(\gamma_Y)] = E[\tilde{X}_i \tilde{Y}_i] - E[\tilde{X}_i \tilde{D}_i(\gamma_Y)] \\
= E[\tilde{X}_i \tilde{Y}_i] \\
\]

\( \gamma_Y \) only enters \( E[g_i(\beta_0; \xi_0, \mu_0, \gamma_Y, \gamma_Y, \gamma_X, \gamma_X, \omega_Z, \omega_Z, \omega_Y, \omega_Y)] \), through the expression above, so we are robust to \( \gamma_Y \). Similarly for any \( \gamma_Y \), \( E[\tilde{X}_i \tilde{V}_i(\gamma_Y)'(\gamma_Y)] = E[\tilde{X}_i \tilde{V}_i]' \), which shows we are robust to \( \gamma_Y \). For any \( \gamma_X, 2 \), \( E[\tilde{X}_i \tilde{X}_i(\gamma_X,2)] = E[\tilde{X}_i \tilde{X}_i]' \) so we are robust to \( \gamma_X,2 \). For any \( \gamma_X,1 \):

\[ E[\tilde{X}_i(\gamma_X,1)(\tilde{Y}_i - \tilde{V}_i' \xi_0 - \tilde{X}_i' \beta_0)] = E[\tilde{X}_i(\tilde{Y}_i - \tilde{V}_i' \xi_0 - \tilde{X}_i' \beta_0)] \\
\]

Which shows we are robust to \( \gamma_X,1 \). For any \( \omega_Y \), \( E[\tilde{Z}_i \tilde{Y}_i(\omega_Y)] = E[\tilde{Z}_i \tilde{Y}_i] \), for any \( \omega_Y \) we have \( E[\tilde{Z}_i \tilde{V}_i(\omega_Y)'(\omega_Y)] = E[\tilde{Z}_i \tilde{V}_i]' \), and for any \( \omega_Z \):

\[ E[\tilde{Z}_i(\omega_Z)(\tilde{Y}_i - \tilde{V}_i' \xi_0)] = E[\tilde{Z}_i(\tilde{Y}_i - \tilde{V}_i' \xi_0)] \\
\]

So the moment condition is also robust to \( \omega_Y, \omega_Y, \) and \( \omega_Z \).
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 \((\beta_0\) 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’). By Lemma 2 the moment condition is valid so a) hold. By Lemma 3 the score is doubly-robust and therefore Neyman-orthogonal so d) holds. The score is linear in \(\beta_0\) and it is linear in each of its arguments and is thus continuously twice Gateaux differentiable, so b) and c) hold. 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 \(\eta_0\) contain the true values of all the nuisance parameters so that:

\[
\eta_0 = (\mu_0, \xi_0, \gamma_{X,0}, \gamma_{Y,0}, \gamma_{V,0}, \omega_{Z,0}, \omega_{Y,0}, \omega_{V,0})
\]

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

\[
\hat{\eta}_j = (\hat{\mu}_j, \hat{\xi}_j, \hat{\gamma}_{X,j}, \hat{\gamma}_{Y,j}, \hat{\gamma}_{V,j}, \hat{\omega}_{Z,j}, \hat{\omega}_{Y,j}, \hat{\omega}_{V,j})
\]

Moreover, for some \(\eta = (\mu, \xi, \gamma_X, \gamma_Y, \gamma_V, \omega_Z, \omega_Y, \omega_V)\) we define \(g_i(\eta)\) as follows:

\[
g_i(\beta, \eta) = g_i(\beta, \mu, \xi, \gamma_X, \gamma_Y, \gamma_V, \omega_Z, \omega_Y, \omega_V)
\]

Assumption 3.2 of Chernozhukov et al. (2018) states \(E[g_i g_i']\) has eigenvalues all bounded below away from zero and that for each \(n\) there is a set \(T_n\), a sequence \(\alpha_n \to 0\), a constant \(c_1\), and a sequence \(\delta_n \to 0\), so that if \(P \in \mathcal{P}_n\) the conditions below all hold.

1. With probability at least \(1 - \alpha_n\), \(\hat{\eta}_j \in T_n\) for all \(j = 1, ..., J\).
2. \(\sup_{\eta \in T_n} E[\|g_i(\beta_0, \eta)\|^q]^{1/q} \leq c_1\)
3. \(\sup_{\eta \in T_n} E[\|\hat{X}_i(\gamma_X) - \hat{X}_i(\gamma_X')\|^q]^{1/q} \leq c_1\)
4. \(\sup_{\eta \in T_n} \|\Sigma_X - E[\hat{X}_i(\gamma_X) \hat{X}_i(\gamma_X')]\| \leq \delta_n\)
5. \(\sup_{\eta \in T_n} E[\|g_i(\beta_0, \eta_0) - g_i(\beta_0, \eta)\|^2]^{1/2} \leq \delta_n\)
6. \(\sup_{\eta \in T_n} E[\|g_i(\beta_0, \eta_0 + \eta - \eta_0)\|^2]^{1/2} \leq \delta_n/\sqrt{n}\)

We will show that the conditions above are satisfied with:

\[
\delta_n \leq c_{3,n} + c_{4,n} + \sqrt{n}c_{5,n}
\]

Where the sequences \(c_{3,n}\), \(c_{4,n}\), and \(c_{5,n}\) have rates \(c_{3,n} \lesssim \delta_{\gamma,\gamma,\gamma,\gamma,n}^2\).
\[ c_{4,n} \preceq \delta_{\gamma,Y,n} + \delta_{\gamma,X,n} + \delta_{\gamma,Y,n} + \delta_{\omega,Y,n} + \delta_{\omega,Z,n} + \delta_{\omega,V,n} + \delta_{\xi,n} + \delta_{\mu,n} + \sqrt{d_D} \delta_{\gamma,X,n} (\delta_{\gamma,Y,n} + \delta_{\gamma,V,n} + \delta_{\gamma,X,n}) + \sqrt{d_D} \delta_{\omega,Z,n} (\delta_{\omega,Y,n} + \delta_{\omega,V,n}) \]

and:

\[ c_{5,n} \preceq \delta_{\gamma,X,n} (\delta_{\gamma,Y,n} + \delta_{\gamma,V,n} + \delta_{\gamma,X,n}) + \delta_{\omega,Z,n} (\delta_{\omega,Y,n} + \delta_{\omega,V,n}) \]

Then under the premises of the Theorem \( \delta_n \prec 1 \).

The following notation will prove useful. For any \( \eta = (\mu, \xi, \gamma_X, \gamma_Y, \gamma_V, \omega_Z, \omega_Y, \omega_V) \), let \( \delta_{\xi,n} = \xi - \xi_0, \delta_{\mu,n} = \mu - \mu_0 \), for \( H = X, Y, V \) let \( \Delta_{\gamma,H} = \gamma_{H} - \gamma_{H,0} \), and for \( H = Z, Y, V \) let \( \Delta_{\omega,H} = \omega_{H} - \omega_{H,0} \).

Some facts we use repeatedly in our arguments are as follows. For any random vector \( Q \) and a non-random matrix \( A \) with number of columns equal to the length of \( Q \), then letting \( \Sigma_Q = E[QQ^\top] \) and \( r_A \) the number of rows of \( A \):

\[
||A\Sigma_Q^{1/2}||^2 \leq E[||AQ||^2] \leq r_A ||A\Sigma_Q^{1/2}||^2
\]

As an immediate corollary, if \( b \) is a column vector then \( E[||b^\top Q||^2] = ||\Sigma_Q^{1/2}b||^2 \). In addition, if \( R_i \) is a vector of residuals from population least squares regression of a variable \( Q_i \) on some regressors, then \( ||\Sigma_{R_i}^{1/2}b||^2 \leq ||\Sigma_Q^{1/2}b||^2 \). Thus we have \( ||\Sigma_{\nu}\xi_0|| \leq ||\Sigma_Q\xi_0|| \leq \xi \) and similarly, for any \( \xi \) we have \( ||\Sigma_{\nu}\Delta_{\xi}|| \leq ||\Sigma_Q\Delta_{\xi}|| \leq \delta_{\xi,n} \). Finally, \( E[Y_i^2] \leq E[\tilde{Y}_i^2] \leq \sigma_X^2 \), and \( ||\Sigma_{\nu}\xi|| \leq ||\Sigma_Q\xi|| \leq \sigma_X^2 \).

**Condition 1**

Under Assumption 4.2, the set \( \mathcal{T}_n \) defined as follows satisfies Condition 1 for some \( \alpha_n \rightarrow 0 \). \( \eta \in \mathcal{T}_n \) if and only if \( ||\Delta_{\nu} \Sigma_{\nu}^{1/2}|| \leq \delta_{\mu,n}, ||\Sigma_{\nu}^{1/2} \Delta_{\xi}|| \leq \delta_{\xi,n}, \) for \( H = Y, V, X \)

\[
||\Sigma_{R_i}^{-1/2} \Delta_{\gamma,H} \Sigma_{R_i}^{1/2}|| \leq \delta_{\gamma,H},\text{ and for } H = Z, Y, V \text{ we have } ||\Sigma_{R_i}^{-1/2} \Delta_{\gamma,H} \Sigma_{R_i}^{1/2}|| \leq \delta_{\omega,H}.
\]

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

We now proceed by bounding in turn \( m_n, m'_n, r_n, r'_n, \) and \( \lambda_n', \) the scalars on the left-hand sides of the equations in Conditions 2-6.

**Condition 2**

We will show that there is a sequence so that if \( P \in \mathcal{P}_n \) and \( \eta \in \mathcal{T}_n \) then \( E[||g_i(\beta_0, \eta)||^q]^{1/q} \leq c_{1,n} \) where:

\[
c_{1,n} \preceq 1 + d_{V}^{1/4} \delta_{\xi,n} + d_{D}^{3/4} (\delta_{\gamma,Y,n} + \delta_{\gamma,V,n} + \delta_{\gamma,X,n}) + (\delta_{\mu,n} + \delta_{\xi,n}) d_{V}^{1/4} d_{V}^{1/4} + d_{D}^{1/4} (d_{V}^{1/4} + d_{V}^{3/4} \delta_{\omega,Z,n}) (\delta_{\omega,Y,n} + \delta_{\omega,V,n})
\]

Under the premises of the Theorem the expression on the right-hand side above is bounded by a constant \( c \) given in the theorem and so \( c_{1,n} \preceq 1 \).

Let \( \eta \in \mathcal{T}_n \). In our setting \( E[||g_i(\beta_0, \eta)||^q]^{1/q} \) is equal to:

\[
E\left[||\tilde{X}_i(\gamma_X)(\tilde{Y}_i(\gamma_Y) - \tilde{V}_i(\gamma_V)^t \xi - \tilde{X}_i(\gamma_X)^t \beta_0) - \mu Z_i(\omega_Z)(Y_i(\omega_Y) - \tilde{V}_i(\omega_V)^t \xi)||^q\right]^{1/q}
\]

29
By the triangle inequality:

\[
E\left[ \left\| \tilde{X}_i(\gamma) (\tilde{Y}_i(\gamma) - \tilde{V}_i(\gamma)') - \tilde{X}_i(\gamma) (\beta_0) - \mu \tilde{Z}_i(\omega) (\tilde{Y}_i(\omega) - \tilde{V}_i(\omega)') \right\|_q \right]^{1/q}
\leq E\left[ \left\| \tilde{X}_i(\gamma) (\tilde{Y}_i(\gamma) - \tilde{V}_i(\gamma)') - \tilde{X}_i(\gamma) (\beta_0) \right\|_q \right]^{1/q}
+ E\left[ \left\| \mu \tilde{Z}_i(\omega) (\tilde{Y}_i(\omega) - \tilde{V}_i(\omega)') \right\|_q \right]^{1/q}
\]

(37)

Let us upper-bound the first term on the right-hand side above. Using the definitions of \( \tilde{X}_i(\cdot) \), \( \tilde{Y}_i(\cdot) \), and \( \tilde{V}_i(\cdot) \) and the triangle inequality we get:

\[
E\left[ \left\| \tilde{X}_i(\gamma) (\tilde{Y}_i(\gamma) - \tilde{V}_i(\gamma)') - \tilde{X}_i(\gamma) (\beta_0) \right\|_q \right]^{1/q}
\leq E\left[ \left\| \tilde{X}_i \tilde{Y}_i + \tilde{X}_i \tilde{\beta}_0 + \tilde{V}_i' \Delta \xi \right\|_q \right]^{1/q}
+ E\left[ \left\| \tilde{X}_i \tilde{\nu}_0 \right\|_q \right]^{1/q}
+ E\left[ \left\| \tilde{X}_i + \Delta_{\gamma,X} D_i (\Delta_{\gamma,Y} + \Delta_{\gamma,V} \gamma + \Delta_{\gamma,V} \beta_0) \right\|_q \right]^{1/q}
\]

We will bound each term on the right-hand side of the inequality above in turn. Again by the triangle inequality:

\[
E\left[ \left\| \tilde{X}_i \tilde{Y}_i \right\|_q \right]^{1/q} \leq \left\| \Sigma_{\tilde{X}}^{1/2} \right\|_q E\left[ \left\| \Sigma_{\tilde{X}}^{-1/2} \tilde{X}_i \tilde{Y}_i / E[\tilde{Y}_i^2]^{1/2} \right\|_q \right]^{1/q} E[\tilde{Y}_i^2]^{1/2}
\leq \sigma_X d_{\tilde{X}}^{1/2} S_{q,\tilde{X},Y}^{1/4} \sigma_Y
\]

Again, using the properties of the matrix norm and Assumptions 4.1 and 4.3:

\[
E\left[ \left\| \tilde{X}_i \tilde{\nu}_0 \right\|_q \right]^{1/q} \leq \left\| \Sigma_{\tilde{X}}^{1/2} \right\|_q E\left[ \left\| \Sigma_{\tilde{X}}^{-1/2} \tilde{X}_i \tilde{\nu}_0 / E[\tilde{\nu}_0^2]^{1/2} \right\|_q \right]^{1/q} E[\tilde{\nu}_0^2]^{1/2}
\leq \sigma_X d_{\tilde{X}}^{1/2} S_{q,\tilde{X},\tilde{\nu}}^{1/4} \sigma_{\tilde{\nu}}
\]

And following similar steps but also using the properties of \( T_n \) to get \( \left\| \Sigma_{\tilde{\nu}}^{-1/2} \Delta \xi \right\| \leq \delta_\xi \):

\[
E\left[ \left\| \tilde{X}_i \tilde{\nu}_i' \Delta \xi \right\|_q \right]^{1/q} \leq \left\| \Sigma_{\tilde{X}}^{1/2} \right\|_q E\left[ \left\| \Sigma_{\tilde{X}}^{-1/2} \tilde{X}_i \tilde{\nu}_i' / E[\tilde{\nu}_i'^2]^{1/2} \right\|_q \right]^{1/q} E[\tilde{\nu}_i'^2]^{1/2}
\leq \sigma_X d_{\tilde{X}}^{1/2} d_{\tilde{\nu}}^{1/2} S_{q,\tilde{X},\tilde{\nu}}^{1/4} \delta_\xi, n
\]

30
And so:

$$E[||\tilde{X}_i(\tilde{Y}_i + \tilde{X}_i^\prime 0 + \tilde{V}_i\Delta\xi)||^q]^{1/q}$$

$$\leq \sigma_X d_X^q (S_{1/q}^{1/q} \sigma_Y + d_X^q S_{1/q}^{1/q} \beta + d_Y^q S_{1/q}^{1/q} \delta\xi)$$

By Assumption 4.3, $E[||\tilde{X}_i\tilde{V}_i\xi_0||^q]^{1/q} \leq S_\xi$. Next note that, by the properties of the matrix norm:

$$E[||\tilde{X}_i + \Delta_{\gamma,X} D_{\gamma,Y} + \Delta_{\gamma,X}^\prime \xi + \Delta_{\gamma,X}^\prime 0||^q]^{1/q}$$

$$\leq E[||\tilde{X}_i + \Delta_{\gamma,X} D_{\gamma,Y} + \Delta_{\gamma,X}^\prime \xi + \Delta_{\gamma,X}^\prime 0||^q]^{1/q}$$

Consider the first term in the product on the right-hand side above. Applying the triangle inequality and the properties of the matrix norm:

$$E[||\tilde{X}_i + \Delta_{\gamma,X} D_{\gamma,Y} + \Delta_{\gamma,X}^\prime \xi + \Delta_{\gamma,X}^\prime 0||^q]^{1/q}$$

$$\leq \sigma_X (d_X^{1/q} S_{1/q}^{1/q} + d_Y^{1/q} S_{1/q}^{1/q} \delta\xi)$$

Where the last line follows by Assumptions 4.1 and 4.3. Next note that again applying the triangle inequality and properties of the matrix norm:

$$||\Delta_{\gamma,Y} + \Delta_{\gamma,X}^\prime \xi + \Delta_{\gamma,X}^\prime 0||^q$$

$$\leq \sigma_X (d_X^{1/q} S_{1/q}^{1/q} + d_Y^{1/q} S_{1/q}^{1/q} \delta\xi)$$

In all we get:

$$E[||\tilde{X}_i(\gamma_X)(\tilde{Y}_i(\gamma_Y) - \tilde{V}_i(\gamma_Y)\xi - \tilde{X}_i(\gamma_X)\beta_0)||^q]^{1/q}$$

$$\leq \sigma_X (d_X^{1/q} S_{1/q}^{1/q} + d_Y^{1/q} S_{1/q}^{1/q} \delta\xi) + \sigma_X (d_X^{1/q} S_{1/q}^{1/q} + d_Y^{1/q} S_{1/q}^{1/q} \delta\xi)$$

$$\leq 1 + d_Y^{2/q} \delta\xi$$
It remains to bound the second term on the right-hand side of (37). Using the definitions
of $\bar{Z}_i(\cdot)$, $\bar{Y}_i(\cdot)$, and $\bar{V}_i(\cdot)$ and the triangle inequality

$$E \left[ \left| \mu \bar{Z}_i(\omega_Z) (\bar{Y}_i(\omega_Y) - \bar{V}_i(\omega_Y)'\xi) \right|^q \right]^{1/q}$$

$$\leq E \left[ \left| \Delta_\mu \bar{Z}_i(\bar{Y}_i + \bar{V}_i'\xi_0) \right|^q \right]^{1/q}$$

$$+ E \left[ \left| \mu_0 \bar{Z}_i \bar{Y}_i \right|^q \right]^{1/q} + E \left[ \left| \mu_0 \bar{Z}_i \bar{V}_i' \xi_0 \right|^q \right]^{1/q}$$

$$+ E \left[ \left| \mu \bar{Z}_i \bar{V}_i' \Delta_\xi \right|^q \right]^{1/q}$$

$$+ E \left[ \left| \mu (\bar{Z}_i + \Delta_{\omega,Z}(X_i', D_i')) (X_i', D_i')(\Delta_{\omega,Y} - \Delta_{\omega,V}\xi) \right|^q \right]^{1/q}$$

(38)

Note that by the triangle inequality and properties of the matrix norm:

$$E \left[ \left| \Delta_\mu \bar{Z}_i(\bar{Y}_i + \bar{V}_i'\xi_0) \right|^q \right]^{1/q}$$

$$\leq \left| \Delta_\mu \Sigma_{\bar{Z}_i}^{1/2} \right| \left( E \left[ \left| \Sigma_{\bar{V}_i}^{-1/2} \bar{V}_i, E \left[ \left| \bar{Y}_i^2 \right|^q \right]^{1/2} \right] \right) \left| \left. \right| \Sigma_{\bar{V}_i}^{-1/2} \bar{V}_i, \Sigma_{\bar{V}_i}^{-1/2} \right| \left| \left. \right| \Sigma_{\bar{V}_i}^{-1/2} \xi_0 \right| \right)$$

$$\leq \delta_{\mu,n} \left( d_{\bar{Z}_i}^{1/q} S_{\bar{Z},\bar{V}}^{1/q} \sigma_Y + d_{\bar{V}_i}^{1/q} d_{\bar{V}_i}^{1/q} S_{\bar{Z},\bar{V}}^{1/q} \bar{\xi} \right)$$

Where the final line follows from Assumptions 4.1, 4.2, and the definition of $\mathcal{T}_n$. Next,
by Assumption 4.3:

$$E \left[ \left| \mu_0 \bar{Z}_i \bar{Y}_i \right|^q \right]^{1/q} + E \left[ \left| \mu_0 \bar{Z}_i \bar{V}_i' \xi_0 \right|^q \right]^{1/q} \leq 2 \bar{S}_q$$

Next applying the properties of the matrix norm and the triangle inequality and then Assumptions 4.2 and 4.3:

$$E \left[ \left| \mu \bar{Z}_i \bar{V}_i' \Delta_\xi \right|^q \right]^{1/q}$$

$$\leq \left( \left| \Delta_\mu \Sigma_{\bar{Z}_i}^{1/2} \right| + \left| \Delta_\mu \Sigma_{\bar{V}_i}^{1/2} \right| \right) E \left[ \left| \Sigma_{\bar{Z}_i}^{-1/2} \bar{Z}_i, \Sigma_{\bar{V}_i}^{-1/2} \right| \left| \left. \right| \Sigma_{\bar{V}_i}^{-1/2} \Delta_\xi \right| \right]$$

$$\leq (\bar{\mu} + \delta_{\mu,n}) \left( d_{\bar{Z}_i}^{1/q} d_{\bar{V}_i}^{1/q} S_{\bar{Z},\bar{V}}^{1/q} \delta_{\xi,n} \right)$$

Now we bound the final term in (38). Using basic properties of the matrix norm:

$$E \left[ \left| \mu (\bar{Z}_i + \Delta_{\omega,Z}(X_i', D_i')) (X_i', D_i')(\Delta_{\omega,Y} - \Delta_{\omega,V}\xi) \right|^q \right]^{1/q}$$

$$\leq \left| \mu \Sigma_{\bar{Z}_i}^{1/2} \right| E \left[ \left| \Sigma_{\bar{Z}_i}^{-1/2} \bar{Z}_i + \Delta_{\omega,Z}(X_i', D_i') \right| (X_i', D_i') \Sigma_{\bar{V}_i}^{-1/2} \right| \left| \left. \right| \Sigma_{\bar{V}_i}^{-1/2} \Delta_{\omega,V}\xi \right| \right]$$

By the triangle inequality and Assumptions 4.1 and the definition of $\mathcal{T}_n$ we have $\left| \mu \Sigma_{\bar{Z}_i}^{1/2} \right| \leq \bar{\mu} + \delta_{\mu,n}$. Next applying properties of the matrix norm and the triangle inequality and we
get:

\[
E \left[ \left\| \Sigma^{-1/2}_Z (\tilde{Z}_i + \Delta_{\omega,Z}(X'_i, D'_i)) (X'_i, D'_i) \Sigma^{-1/2}_X \right\|^q \right]^{1/q} \\
\leq E \left[ \left\| \Sigma^{-1/2}_Z \tilde{Z}_i (X'_i, D'_i) \Sigma^{-1/2}_X \right\|^q \right]^{1/q} \\
+ \left\| \Sigma^{-1/2}_Z \Delta_{\omega,Z} \Sigma^{-1/2}_X \right\| E \left[ \left\| \Sigma^{-1/2}_X (X'_i, D'_i) \Sigma^{-1/2}_X \right\|^q \right]^{1/q} \\
\leq d^{1/q}_Z (d_X + d_D)^{1/q} \bar{G}^{1/q}_{q,X,D,X,D} + \delta_{\omega,Z,n} (d_X + d_D)^{2/q} \bar{G}^{1/q}_{q,X,D,X,D}
\]

Where the final line follows from the properties of \( T_n \) and Assumption 4.3. Finally, using the triangle inequality, the properties of the matrix norm along with Assumptions 4.1 and 4.2:

\[
||\Sigma^{-1/2}_X (\Delta_{\omega,Y} - \Delta'_{\omega,V} \xi)|| \\
\leq ||\Sigma^{-1/2}_X \Delta_{\omega,Y} E[\bar{Y}_i^2]^{-1/2} ||E[\bar{Y}_i^2]^{1/2} + ||\Sigma^{-1/2}_X \Delta'_{\omega,V} \Sigma^{-1/2}_V ||(||\Sigma^{-1/2}_V \xi|| + ||\Sigma^{-1/2}_V \Delta_v||) \\
\leq \delta_{\omega,Y,n} \sigma_Y + \delta_{\omega,V,n} (\xi + \delta_{\xi,n})
\]

Combining we get:

\[
E \left[ ||\mu (\tilde{Z}_i + \Delta_{\omega,Z}(X'_i, D'_i)) (X'_i, D'_i) (\Delta_{\omega,Y} - \Delta'_{\omega,V} \xi)||^q \right]^{1/q} \\
\leq (\bar{\mu} + \delta_{\mu,n})(d^{1/q}_Z d^{1/q}_V \sigma_Y + d^{1/q}_Z d^{1/q}_V \sigma_{q,Z,V} \xi) \\
+ 2 \bar{G}^{1/q}_{q,Z,V} \\
+ (\bar{\mu} + \delta_{\mu,n})(d^{1/q}_Z d^{1/q}_V \sigma_{q,Z,V} \delta_{\xi,n}) \\
+ (\bar{\mu} + \delta_{\mu,n})(d^{1/q}_Z d^{1/q}_V \sigma_{q,Z,V} \delta_{\xi,n}) \\
+ (\bar{\mu} + \delta_{\mu,n})(d^{1/q}_Z d^{1/q}_V \sigma_{q,Z,V} \delta_{\xi,n}) \\
+ 1 + (\delta_{\mu,n} + \delta_{\xi,n}) d^{1/q}_Z d^{1/q}_V \\
+ d^{1/q}_D (d^{1/q}_Z + d^{1/q}_V \delta_{\omega,Z,n}) (\delta_{\omega,Y,n} + \delta_{\omega,V,n})
\]

And so:

\[
E \left[ ||\mu \tilde{Z}_i (\omega_Z)(\bar{Y}_i(\omega_Y) - \bar{V}_i(\omega_V) \xi)||^q \right]^{1/q} \\
\leq \delta_{\mu,n} (d^{1/q}_Z d^{1/q}_V \sigma_Y + d^{1/q}_Z d^{1/q}_V \sigma_{q,Z,V} \xi) \\
+ 2 \bar{G}^{1/q}_{q,Z,V} \\
+ (\bar{\mu} + \delta_{\mu,n})(d^{1/q}_Z d^{1/q}_V \sigma_{q,Z,V} \delta_{\xi,n}) \\
+ (\bar{\mu} + \delta_{\mu,n})(d^{1/q}_Z d^{1/q}_V \sigma_{q,Z,V} \delta_{\xi,n}) \\
+ (\bar{\mu} + \delta_{\mu,n})(d^{1/q}_Z d^{1/q}_V \sigma_{q,Z,V} \delta_{\xi,n}) \\
+ 1 + (\delta_{\mu,n} + \delta_{\xi,n}) d^{1/q}_Z d^{1/q}_V \\
+ d^{1/q}_D (d^{1/q}_Z + d^{1/q}_V \delta_{\omega,Z,n}) (\delta_{\omega,Y,n} + \delta_{\omega,V,n})
\]
Combining everything we get:
\[
c_{1,n} = \sigma_X d_X^{1/q} \left( \sigma_Y S^{1/q}_{q,X,Y} + d_X^{1/q} \beta S^{1/q}_{q,X,X} + d_Y^{1/q} S^{1/q}_{q,Y,Y} \delta_{\xi,n} \right) + \bar{S}_q
\]
\[
+ \sigma_X (d_X^{1/q} S^{1/q}_{q,X,D} + d_D^{1/q} S^{1/q}_{q,D,D}) d_Y^{1/q} \left( \delta_{Y,n} \sigma_Y + \delta_{Y,n} (\xi + \delta_{\xi,n}) + \delta_{Y,n} \tilde{\beta} \right)
\]
\[
+ \delta_{\mu,n} \left( d_{Z}^{1/q} S^{1/q}_{q,Z,Y} \sigma_Y + d_{V}^{1/q} S^{1/q}_{q,Z,V} \tilde{\xi} \right)
\]
\[
+ 2 \bar{S}_q
\]
\[
+ (\bar{\mu} + \delta_{\mu,n}) (d_{Z}^{1/q} d_{V}^{1/q} S^{1/q}_{q,Z,V} \delta_{\xi,n})
\]
\[
+ \left( (\bar{\mu} + \delta_{\mu,n}) (d_{Z}^{1/q} (d_X + d_D)^{1/q} S^{1/q}_{q,Z,X,D} + \delta_{\omega,n} (d_X + d_D)^{2/q} S^{1/q}_{q,X,D,D})
\right)
\]
\[
\times \left( \delta_{\omega,n} \sigma_Y + \delta_{\omega,n} (\xi + \delta_{\xi,n}) \right)
\]
\[
\lesssim 1 + d_{V}^{1/q} \delta_{\xi,n} + d_{D}^{2/q} (\delta_{Y,n} + \delta_{Y,n} + \delta_{Y,n} + \delta_{Y,n})
\]
\[
+ (\delta_{\mu,n} + \delta_{\xi,n}) d_{Z}^{1/q} d_{V}^{1/q}
\]
\[
+ d_{D}^{2/q} (d_{Z}^{1/q} + d_{V}^{1/q} \delta_{\omega,n}) \left( \delta_{\omega,n} + \delta_{\omega,n} \right)
\]

**Condition 3**

We now consider \( \sup_{\eta \in \mathcal{T}_n} E\left[ ||\tilde{X}_i(\gamma_X) \tilde{X}_i(\gamma_Y)||^q \right]^{1/q} \). We will show that if \( P \in \mathcal{P}_n \) and \( \eta \in \mathcal{T}_n \) then \( E\left[ ||\tilde{X}_i(\gamma_X) \tilde{X}_i(\gamma_Y)||^q \right]^{1/q} \leq c_{2,n} \) where:

\[
c_{2,n} = \sigma_X d_X^{2/q} S^{1/q}_{q,X,X}
\]
\[
+ \delta_{Y,n} \sigma_X d_{D}^{1/q} (d_X^{1/q} S^{1/q}_{q,X,D} + \delta_{Y,n} d_{D}^{1/q} S^{1/q}_{q,D,D})
\]
\[
\lesssim 1 + \delta_{Y,n} \sigma_X d_{D}^{1/q} (1 + \delta_{Y,n} d_{D}^{1/q})
\]

Under the premises of the Theorem the expression on the right-hand side above is bounded by a constant and so \( c_{2,n} \lesssim 1 \). Since \( c_{1,n} \lesssim 1 \) and \( c_{2,n} \lesssim 1 \) there must exist a sufficiently large constant \( c_1 \) so that Conditions 1 and 2 are satisfied.

Let \( \eta \in \mathcal{T}_n \). By the triangle inequality and the definition of \( \tilde{X}_i(\cdot) \) and \( \tilde{X}_i \):

\[
E\left[ ||\tilde{X}_i(\gamma_X) \tilde{X}_i(\gamma_Y)||^q \right]^{1/q}
\]
\[
\leq E\left[ ||\tilde{X}_i||^q \right]^{1/q}
\]
\[
+ E\left[ ||(\tilde{X}_i + \Delta_{Y,n} D_i) D'_i \Delta_{Y,X}||^q \right]^{1/q}
\]

By the properties of the matrix norm and Assumptions 4.1 and 4.3:

\[
E\left[ ||\tilde{X}_i \tilde{X}_i'||^q \right]^{1/q} \leq ||\Sigma X|| E\left[ ||\Sigma^{-1/2}_X \tilde{X}_i \Sigma^{-1/2}_X||^q \right]^{1/q}
\]
\[
\leq \sigma_X d^{2/q}_X S^{1/q}_{q,X,X}
\]

Again by the properties of the matrix norm and Assumption 4.1 and 4.2:
\[ E[\|\tilde{X}_i + \Delta_{\gamma,X} D_i D_i \Delta'_{\gamma,X} \| q]^{1/q} \]
\[ \leq E[\|\tilde{X}_i + \Delta_{\gamma,X} D_i D_i \Sigma_{D}^{-1/2} \| q]^{1/q} \| \Sigma_{D}^{-1/2} \Delta'_{\gamma,X} \Sigma_{\tilde{X}}^{-1/2} \| : \| \Sigma_{\tilde{X}} \|^{1/2} \]
\[ \leq E[\|\tilde{X}_i + \Delta_{\gamma,X} D_i D_i \Sigma_{D}^{-1/2} \| q]^{1/q} \delta_{\gamma,X} \sigma_{X} \]

Finally, using the triangle inequality and properties of the matrix norm and then Assumptions 4.1, 4.2, and 4.3:

\[ E[\|\tilde{X}_i + \Delta_{\gamma,X} D_i D_i \Sigma_{D}^{-1/2} \| q]^{1/q} \leq \| \Sigma_{\tilde{X}} \|^{1/2} E[\|\tilde{X}_i + \Delta_{\gamma,X} D_i D_i \Sigma_{D}^{-1/2} \| q]^{1/q} \]
\[ + \| \Sigma_{\tilde{X}} \|^{1/2} \| \Sigma_{\tilde{X}}^{-1/2} \Delta_{\gamma,X} \Sigma_{D}^{-1/2} \| E[\|\tilde{X}_i + \Delta_{\gamma,X} D_i D_i \Sigma_{D}^{-1/2} \| q]^{1/q} \]
\[ \leq \sigma_{X} d_{\delta}^{1/q} d_{\gamma}^{1/q} S_{q,\delta,D}^{1/q} + \sigma_{X} \delta_{\gamma,X} \sigma_{D}^{1/q} q_{D}^{1/q} \]

Combining gives the result.

**Condition 4**

We now consider \( \sup_{\eta \in T_{n}} \| \Sigma_{\tilde{X}} - E[\tilde{X}_i (\gamma,X) \tilde{X}_i (\gamma,X)] \| \). We will show that if \( \eta \in T_{n} \) and \( P \in P_{n} \) then letting the sequence \( c_{3,n} = d_{X} \delta_{\gamma,X} \sigma_{X}^{2} \) we have

\[ \| \Sigma_{\tilde{X}} - E[\tilde{X}_i (\gamma,X) \tilde{X}_i (\gamma,X)] \| \leq c_{3,n} \]
\[ \leq \delta_{\gamma,X}^{2} \]

Using the properties of partialling out:

\[ \Sigma_{\tilde{X}} - E[\tilde{X}_i (\gamma,X) \tilde{X}_i (\gamma,X)] = E[(\tilde{X}_i - \tilde{X}_i (\gamma,X))(\tilde{X}_i - \tilde{X}_i (\gamma,X))'] \]
\[ = (\gamma_{X,0} - \gamma_{X}) \Sigma_{D} (\gamma_{X,0} - \gamma_{X})' \]

And so:

\[ \| \Sigma_{\tilde{X}} - E[\tilde{X}_i (\gamma,X) \tilde{X}_i (\gamma,X)] \| = \| (\gamma_{X,0} - \gamma_{X}) \Sigma_{D} (\gamma_{X,0} - \gamma_{X})' \| \]
\[ \leq E[\| D_i (\gamma_{X,0} - \gamma_{X})' \|^{2}] \]
\[ \leq d_{X} \| \Sigma_{D}^{-1/2} (\gamma_{X,0} - \gamma_{X})' \|^{2} \]
\[ \leq d_{X} \| \Sigma_{D}^{-1/2} (\gamma_{X,0} - \gamma_{X})' \Sigma^{-1/2} (\tilde{X}_{\tilde{X}})^{2} \| \]

Where the second line follows by \( [30] \) and the third line follows from the properties of the matrix norm. By the properties of \( T_{n} \), \( \| \Sigma_{D}^{-1/2} (\gamma_{X,0} - \gamma_{X})' \Sigma^{-1/2} \| \leq \delta_{\gamma,X,n} \), and Assumption 4.1 states that \( \| \Sigma_{\tilde{X}} \| \leq \sigma_{X}^{2} \), combining with the above gives the result.

**Condition 5**

We will show that if \( \eta \in T_{n} \) and \( P \in P_{n} \) then \( E[\| g_{\eta}(\beta_{0}, \eta_{0}) - g_{\eta}(\beta_{0}, \eta) \|^{2}]^{1/2} \leq c_{4,n} \)

where \( c_{4,n} \) has rate below:

\[ c_{4,n} \leq \delta_{\gamma,Y,n} + \delta_{\gamma,X,n} + \delta_{\gamma,V,n} + \delta_{\omega,Y,n} + \delta_{\omega,Z,n} + \delta_{\omega,V,n} + \delta_{\xi,n} + \delta_{\mu,n} \]
\[ + \sqrt{d_D} \delta_{\gamma,X,n} (\delta_{\gamma,Y,n} + \delta_{\gamma,V,n} + \delta_{\gamma,X,n}) \]
\[ + \sqrt{d_D} \delta_{\omega,Z,n} (\delta_{\omega,Y,n} + \delta_{\omega,V,n}) \]
Let $\eta \in \mathcal{T}_n$. In our setting $E[||g_i(\beta_0, \eta_0) - g_i(\beta_0, \eta)||^2]^{1/2}$ is equal to the following:

\[
E \left[ \left| \bar{X}_i(\gamma_X)(\bar{Y}_i(\gamma_Y) - \bar{V}_i(\gamma_Y)'\bar{\beta}_0) - \mu \bar{Z}_i(\omega_Z)(\bar{Y}_i(\omega_Y) - \bar{V}_i(\omega_Y)'\bar{\beta}_0) - \bar{X}_i(\bar{Y}_i - \bar{V}_i')\bar{\beta}_0 + \mu_0 \bar{Z}_i(\bar{Y}_i - \bar{V}_i')\bar{\beta}_0 \right|^2 \right]^{1/2}
\]

By the triangle inequality the above is bounded by:

\[
E \left[ \left| \bar{X}_i(\gamma_X)(\bar{Y}_i(\gamma_Y) - \bar{V}_i(\gamma_Y)'\bar{\beta}_0) \right|^2 \right]^{1/2}
+ E \left[ \left| \bar{X}_i(\gamma_X)(\bar{Y}_i(\gamma_Y)' - \bar{X}_i(\gamma_X)'\bar{\beta}_0) \right|^2 \right]^{1/2}
+ E \left[ \left| \mu \bar{Z}_i(\omega_Z)(\bar{Y}_i(\omega_Y) - \bar{V}_i(\omega_Y)'\bar{\beta}_0) \right|^2 \right]^{1/2}
+ E \left[ \left| \mu \bar{Z}_i(\omega_Z)(\bar{Y}_i(\omega_Y)' - \bar{X}_i(\gamma_X)'\bar{\beta}_0) \right|^2 \right]^{1/2}
+ E \left[ \left| \mu \bar{Z}_i(\omega_Z)(\bar{V}_i(\omega_Y)' - \bar{X}_i(\gamma_X)'\bar{\beta}_0) \right|^2 \right]^{1/2}
\]

(39)

Before we begin bound the terms in the expression above, we prove an inequality given below. Let $R_i$, $H_i$, and $Q_i$ be random vectors and let $R_i(\chi) = R_i - \chi Q_i$ for $\chi = \chi_H$ and $\chi_{R,0}$, and let $H_i(\chi) = H_i - \chi Q_i$ for $\chi = \chi_H$ and $\chi_{R,0}$. Define $\Sigma_R = E[R_i(\chi_H,0)R_i(\chi_{R,0})']$, $\Sigma_H = E[H_i(\chi_{H,0})H_i(\chi_{H,0})']$, and $\Sigma_Q = E[Q_iQ_i']$. Let $A$ be a matrix with $r_A$ rows and the number of columns equal to the length of $R_i$. Let $\zeta$ be a vector of the same length as $H_i$. Define $||A\Sigma_h^{1/2}|| = \tilde{A}$, $||\Sigma_h^{1/2}\zeta|| = \tilde{\zeta}$, $||\Sigma_h|| = \sigma_h^2$, and $E[||\Sigma_q^{1/2}Q_i||^2] = \sigma_q^2$. Let $\sigma_h^{1/2}$ and $\sigma_q^{1/2}$ be the smallest almost sure bounds on $||\Sigma_R^{1/2}E[R_i(\chi_H,0)R_i(\chi_{R,0})']\Sigma_q^{1/2}||$ and $||\Sigma_h^{1/2}E[H_i(\chi_{H,0})H_i(\chi_{H,0})']Q_i\Sigma_q^{1/2}||$ respectively. We will show that:

\[
E \left[ \left| A(R_i(\chi_R)H_i(\chi_H) - R_i(\chi_{R,0})H_i(\chi_{H,0})')\zeta \right|^2 \right]^{1/2}
\leq \sqrt{\tilde{A} \tilde{A} \tilde{\zeta} \tilde{\zeta} \tilde{\zeta} \tilde{\zeta}} ||\Sigma_h^{1/2}(\chi_H,0 - \chi_H)\Sigma_q^{1/2}||
+ \sqrt{\tilde{A} \tilde{A} \tilde{\zeta} \tilde{\zeta} \tilde{\zeta} \tilde{\zeta}} ||\Sigma_h^{1/2}(\chi_{R,0} - \chi_{R,0})\Sigma_q^{1/2}||
+ s \sqrt{\tilde{A} \tilde{A} \tilde{\zeta} \tilde{\zeta} \tilde{\zeta} \tilde{\zeta}} ||\Sigma_h^{1/2}(\chi_H,0 - \chi_H)\Sigma_q^{1/2}|| ||\Sigma_h^{1/2}(\chi_{R,0} - \chi_{R,0})\Sigma_q^{1/2}||
\]

(40)

To see this, first apply the triangle inequality and the properties of the matrix norm:

\[
E \left[ \left| A(R_i(\chi_R)H_i(\chi_H) - R_i(\chi_{R,0})H_i(\chi_{H,0})')\zeta \right|^2 \right]^{1/2}
\leq E\left[ ||AR_i(\chi_{R,0})Q_i(\chi_{R,0} - \chi_{R,0})'\zeta||^2 \right]^{1/2}
+ E\left[ ||AR_i(\chi_{R,0} - \chi_{R,0})Q_iH_i(\chi_{H,0})'\zeta||^2 \right]^{1/2}
+ E\left[ ||AR_i(\chi_{R,0} - \chi_{R,0})Q_i'H_i(\chi_{H,0} - \chi_{H,0})'\zeta||^2 \right]^{1/2}
\]

(41)

To help bound the above we will show that $E[||AR_i(\chi_{R,0})||^2|Q_i] \leq r_A \sigma_{RQ}^2 \tilde{A}$ and $E[||\zeta'H_i(\chi_{H,0})||^2|Q_i] \leq \sigma_{HQQ}^2 \tilde{\zeta}^2$ each with probability 1. To see this, we first apply[36] to get:

\[
E[||AR_i(\chi_{R,0})||^2|Q_i] \leq r_A ||AE[R_i(\chi_{R,0})R_i(\chi_{R,0})']|Q_i||^2
\]
Using the properties of the matrix norm we then get:

\[
E[||AR_i(\chi_R,0)||^2 Q_i] \leq r_A ||A\Sigma_R^{1/2}||^2 ||\Sigma_R^{-1/2}E[R_i(\chi_R,0)R_i(\chi_R,0)'|Q_i]^{1/2}||^2 \\
\leq r_A \tilde{\sigma}_{RQ}^2 A^2
\]

The inequality \(E[||\zeta' H_i(\chi_R,0)||^2 Q_i] \leq \tilde{\sigma}_H^2 Q \tilde{c}^2\) follows from the same steps (note that \(\zeta\) is a column vector and so \(\zeta'\) has one row, hence there is no term analogous to \(r_A\) in the upper bound).

Now we bound the first term on the right-hand side of \ref{eq:41}.

\[
E[||AR_i(\chi_R,0)Q_i'(\chi_{H,0} - \chi_H)'\zeta||^2] \\
= E\left[ E[||AR_i(\chi_R,0)||^2 Q_i] ||\zeta'(\chi_{H,0} - \chi_H)Q_i||^2 \right] \\
\leq r_A \tilde{\sigma}_{RQ}^2 A^2 E[||\zeta'(\chi_{H,0} - \chi_H)Q_i||^2] \\
= r_A \tilde{\sigma}_{RQ}^2 A^2 ||\zeta'(\chi_{H,0} - \chi_H)\Sigma_Q^{1/2}||^2 \\
\leq r_A \tilde{\sigma}_{RQ}^2 A^2 \tilde{c}^2 ||\Sigma_Q^{-1/2}(\chi_{H,0} - \chi_H)\Sigma_Q^{1/2}||^2
\]

Following similar steps we get:

\[
E[||A(\chi_{R,0} - \chi_R)Q_i H_i(\chi_{H,0})'\zeta||^2] \\
= E\left[ ||A(\chi_{R,0} - \chi_R)Q_i||^2 E[||H_i(\chi_{H,0})'\zeta||^2 Q_i] \right] \\
\leq \tilde{\sigma}_{H_Q}^2 \tilde{c}^2 E[||A(\chi_{R,0} - \chi_R)Q_i||^2] \\
\leq r_A \tilde{\sigma}_{H_Q}^2 \tilde{c}^2 ||A(\chi_{R,0} - \chi_R)\Sigma_Q^{1/2}||^2 \\
\leq r_A \tilde{\sigma}_{H_Q}^2 \tilde{c}^2 ||A(\chi_{R,0} - \chi_R)\Sigma_Q^{1/2}||^2
\]

Where the penultimate line uses \ref{eq:36} Finally, with repeated application of the properties of the matrix norm:

\[
E[||A(\chi_{R,0} - \chi_R)Q_i Q_i'(\chi_{H,0} - \chi_H)'\zeta||^2] \\
= E[||A(\chi_{R,0} - \chi_R)\Sigma_Q^{1/2}Q_i'(\chi_{H,0} - \chi_H)'\Sigma_Q^{-1/2}Q_iQ_i'(\chi_{H,0} - \chi_H)'\zeta||^2] \\
\leq E[||A(\chi_{R,0} - \chi_R)\Sigma_Q^{1/2}||^2 ||\Sigma_Q^{-1/2}Q_iQ_i'(\chi_{H,0} - \chi_H)'\Sigma_Q^{-1/2}||^2 ||\Sigma_Q^{-1/2}(\chi_{H,0} - \chi_H)'\zeta||^2] \\
= ||A(\chi_{R,0} - \chi_R)\Sigma_Q^{1/2}||^2 E[||\Sigma_Q^{-1/2}Q_iQ_i'(\chi_{H,0} - \chi_H)'\Sigma_Q^{-1/2}||^2 ||\Sigma_Q^{-1/2}(\chi_{H,0} - \chi_H)'\zeta||^2] \\
= E[||\Sigma_Q^{-1/2}Q_i||^2 ||A(\chi_{R,0} - \chi_R)\Sigma_Q^{1/2}||^2 ||\Sigma_Q^{-1/2}(\chi_{H,0} - \chi_H)'\zeta||^2] \\
\leq s_Q^2 d_Q ||\Sigma_Q^{1/2}(\chi_{H,0} - \chi_H)'\zeta||^2 ||A(\chi_{R,0} - \chi_R)\Sigma_Q^{1/2}|| \\
\leq s_Q^2 d_Q \tilde{\sigma}_Q^2 \tilde{c}^2 \\
\times ||\Sigma_Q^{-1/2}(\chi_{H,0} - \chi_H)\Sigma_Q^{1/2}||^2 \\
\times ||\Sigma_Q^{-1/2}(\chi_{R,0} - \chi_R)\Sigma_Q^{1/2}||^2
\]

37
Combining we get $\mathbb{R}$. We apply $A$ with $d_X \times d_X$ identity matrix, $\zeta = 1$, $R_s(\cdot) = \hat{X}_s(\cdot)$, $H_s(\cdot) = \hat{Y}_s(\cdot)$, $\chi = \gamma_X$, $\chi_{R,0} = \gamma_{X,0}$, $\chi_H = \gamma_Y$, and $\chi_{H,0} = \gamma_{Y,0}$. Note that in this case $r_A = d_X$, $\hat{A} \leq \sigma_X$, $\zeta \leq \sigma_Y$, $\hat{\sigma}_{H|Q} \leq \hat{\sigma}_{Y|D}$, $\hat{\sigma}_{R|Q} \leq \hat{\sigma}_{X|D}$, and $s_Q \leq s_D$. We get:

\[
E[||\hat{X}_i(\gamma_X)\hat{Y}_i(\gamma_Y) - \hat{X}_i\hat{Y}_i||^2]^{1/2} \\
\leq \sigma_X \sigma_Y \left( \sqrt{d_X \sigma_{X|D}} \delta_{\gamma,Y,n} + \sqrt{d_Y \sigma_{Y|D}} \delta_{X,Y,n} + \sqrt{d_D \sigma_{X,D}} \right)
\]

Where we have used the properties of $T_n$ for the bounds $\delta_{\gamma,Y,n}$ and $\delta_{X,Y,n}$. Next we bound the second term in $[39]$. By the triangle inequality:

\[
E[||\hat{X}_i(\gamma_X)\hat{V}_i(\gamma_Y)' - \hat{X}_i\hat{V}_i||^2]^{1/2} \\
\leq E[||\hat{X}_i(\gamma_X)\hat{V}_i(\gamma_Y)' - \hat{X}_i\hat{V}'_i||^2]^{1/2} \\
+ E[||\hat{X}_i(\gamma_X)\hat{V}_i(\gamma_Y)' - \hat{X}_i\hat{V}'_i - \hat{V}_i(\gamma_Y)'|\Delta|\delta_{\gamma,Y,n}||^2]^{1/2}
\]

Let us apply $A$ the identity, $\zeta = \xi_0$, $R_s(\cdot) = \hat{X}_s(\cdot)$, $H_s(\cdot) = \hat{V}_s(\cdot)$, $\chi = \gamma_X$, $\chi_{R,0} = \gamma_{X,0}$, $\chi_H = \gamma_Y$, and $\chi_{H,0} = \gamma_{Y,0}$. We get:

\[
E[||\hat{X}_i(\gamma_X)\hat{V}_i(\gamma_Y)' - \hat{X}_i\hat{V}_i||^2]^{1/2} \\
\leq \sigma_X \xi_0 \left( \sqrt{d_X \sigma_{X|D}} \delta_{\gamma,V,n} + \sqrt{d_Y \sigma_{Y|V}} \delta_{\gamma,X,n} + \sqrt{d_D \sigma_{X,D}} \right)
\]

Where we have again used the properties of $T_n$ for the bounds $\delta_{\gamma,V,n}$, $\delta_{\gamma,X,n}$, and $\delta_{\gamma,XV}$. Doing just as above but with $\zeta = \xi_0 - \xi$ we get:

\[
E[||\hat{X}_i(\gamma_X)\hat{V}_i(\gamma_Y)' - \hat{X}_i\hat{V}_i||^2]^{1/2} \\
\leq \sigma_X \xi_0 \left( \sqrt{d_X \sigma_{X|D}} \delta_{\gamma,Y,n} + \sqrt{d_Y \sigma_{Y|D}} \delta_{\gamma,X,n} + \sqrt{d_D \sigma_{X,D}} \right)
\]

Where we have used that if $\eta \in T_n ||\Sigma_{\hat{V}}^{1/2}(\xi - \xi)||^2 \leq \delta_{\xi,n}^2$. Finally we note that:

\[
E[||\hat{X}_i\hat{V}_i||^2]^{1/2} = E[||\hat{X}_i||^2]^{1/2} E[||\Sigma_{\hat{V}}^{-1/2}\hat{V}_i\Sigma_{\hat{V}}^{-1/2}|\hat{X}_i||^{1/2}||\Sigma_{\hat{V}}^{1/2}||^2]^{1/2} \\
\leq E[||\hat{X}_i||^2]^{1/2} E[||\Sigma_{\hat{V}}^{-1/2}\hat{V}_i\Sigma_{\hat{V}}^{-1/2}|\hat{X}_i||]^{1/2}||\Sigma_{\hat{V}}^{1/2}||^2 \Delta_{\xi} \\
\leq E[||\hat{X}_i||^2]^{1/2} \Delta_{\xi,n} \sigma_{\hat{V}|\hat{X}} \\
\leq \sqrt{d_X \sigma_X \delta_{\xi,n} \sigma_{\hat{V}|\hat{X}}}
\]

Where we have used that $E[||\hat{X}_i||^2] = trace(E[\hat{X}_i\hat{X}_i^T]) = d_X \sigma_X^2$. 

38
So in all, the second term in \[39\] is bounded by:
\[
E[||\bar{X}_i(\gamma_X)Y_i(\gamma_Y)'\xi - \bar{X}_iY_i'\xi||_2]^{1/2} \\
\leq \sigma_X(\xi + \delta_{\xi,n})(\sqrt{d_X\delta_{X|D}\delta_{\gamma,Y,n}} + \sqrt{d_X\delta_{Y|D}\delta_{\gamma,X,n}} + \sqrt{d_DsD}\delta_{\gamma,Y,n}) \\
+ \sqrt{d_X\sigma_X\delta_{\xi,n}\delta_{\gamma,Y,X}}
\]
Next we bound the third term in \[39\] Once again let us apply \[40\] with \(A\) the identity, \(\zeta = \beta_0, R_i(\cdot) = H(\cdot) = X_i(\cdot), \chi_R = \chi_H = \gamma_X,\) and \(\chi_{R,0} = \chi_{H,0} = \gamma_{X,0}.\) We get, using the properties of \(T_n:\)
\[
E[||\bar{X}_i(\gamma_X)Y_i(\gamma_Y)'\beta_0 - \bar{X}_i\bar{X}_i'\beta_0||_2]^{1/2} \leq \sigma_X\tilde{\beta}(2\sqrt{d_X\delta_{X|D}\delta_{\gamma,X,n}} + \sqrt{d_DsD}\delta_{\gamma,X,n})
\]
Next we bound the fourth term in \[39\] By the triangle inequality:
\[
E[||\mu\bar{Z}_i(\omega_Z)\bar{Y}_i(\omega_Y) - \mu_0\bar{Z}_i\bar{Y}_i||_2]^{1/2} \\
\leq E[||\mu_0\bar{Z}_i(\omega_Z)\bar{Y}_i(\omega_Y) - \mu_0\bar{Z}_i\bar{Y}_i||_2]^{1/2} \\
+ E[||\Delta_\mu(\bar{Z}_i(\omega_Z)\bar{Y}_i(\omega_Y) - \bar{Z}_i\bar{Y}_i)||_2]^{1/2}
\]
Let us apply \[40\] with \(A = \mu_0\) the identity, \(\zeta = 1, R_i(\cdot) = \bar{Z}_i(\cdot), H_i(\cdot) = \bar{Y}_i(\cdot), \chi_R = \omega_Z,\)
\(\chi_{R,0} = \omega_{Z,0}, \chi_H = \omega_Y,\) and \(\chi_{H,0} = \omega_{Y,0}.\) Note that \(\mu_0\) has \(d_X\) rows so \(r_A = d_X.\) We get:
\[
E[||\mu_0\bar{Z}_i(\omega_Z)\bar{Y}_i(\omega_Y) - \mu_0\bar{Z}_i\bar{Y}_i||_2]^{1/2} \\
\leq \sigma_Y(\sqrt{d_X\delta_{\bar{Z}|X|D}\delta_{\omega,Y,n}} + \sqrt{d_X\delta_{\bar{Y}|X|D}\delta_{\omega,Z,n}} + \sqrt{d_X + d_DsXD}\delta_{\omega,Z,n}\delta_{\omega,Y,n})
\]
Repeating the above but with \(A = \mu_0 - \mu\) we get:
\[
E[||\mu_0\bar{Z}_i(\omega_Z)\bar{Y}_i(\omega_Y) - \mu_0\bar{Z}_i\bar{Y}_i||_2]^{1/2} \\
\leq \delta_{\mu,n}\sigma_Y(\sqrt{d_X\delta_{\bar{Z}|X|D}\delta_{\omega,Y,n}} + \sqrt{d_X\delta_{\bar{Y}|X|D}\delta_{\omega,Z,n}} + \sqrt{d_X + d_DsXD}\delta_{\omega,Z,n}\delta_{\omega,Y,n})
\]
And now note that:
\[
E[||\Delta_\mu\bar{Z}_i\bar{Y}_i||_2]^{1/2} = E[||\Delta_\mu\bar{Z}_i||_2^2E[\bar{Y}_i^2|\bar{Z}_i]]^{1/2} \\
\leq \sigma_{\bar{Y}|\tilde{Z}_i}\sigma_Y E[||\Delta_\mu\bar{Z}_i||_2]^{1/2} \\
\leq \sigma_{\bar{Y}|\tilde{Z}_i}\sigma_Y\sqrt{d_X\delta_{\mu,n}}^{1/2} \\
\leq \sqrt{d_X}\sigma_{\bar{Y}|\tilde{Z}_i}\sigma_Y\delta_{\mu,n}
\]
Where the final inequality uses \[36\] And so in all:
\[
E[||\mu\bar{Z}_i(\omega_Z)\bar{Y}_i(\omega_Y) - \mu_0\bar{Z}_i\bar{Y}_i||_2]^{1/2} \\
\leq \sigma_Y(\mu + \delta_{\mu,n})(\sqrt{d_X\delta_{\bar{Z}|X|D}\delta_{\omega,Y,n}} + \sqrt{d_X\delta_{\bar{Y}|X|D}\delta_{\omega,Z,n}} + \sqrt{d_X + d_DsXD}\delta_{\omega,Z,n}\delta_{\omega,Y,n}) \\
+ \sqrt{d_X}\sigma_{\bar{Y}|\tilde{Z}_i}\sigma_Y\delta_{\mu,n}
\]
Next we bound the final term in 39. By the triangle inequality:

\[
E[||\mu\bar{Z}_i(\omega_Z)\bar{V}_i(\omega_V)'\xi - \mu_0\bar{Z}_i\bar{V}_i'\xi_0||^2]^{1/2}
\leq E[||\mu_0\bar{Z}_i(\omega_Z)\bar{V}_i(\omega_V)'\xi_0 - \mu_0\bar{Z}_i\bar{V}_i'\xi_0||^2]^{1/2}
\]

\[
+ E\left[||\Delta_\mu(\bar{Z}_i(\omega_Z)\bar{V}_i(\omega_V)'\xi_0 - \mu_0\bar{Z}_i\bar{V}_i'\xi_0)||^2\right]^{1/2}
\]

\[
+ E\left[||\Delta_\mu(\bar{Z}_i(\omega_Z)\bar{V}_i(\omega_V)' - \bar{Z}_i\bar{V}_i')\Delta_\xi||^2\right]^{1/2}
\]

\[
+ E\left[||\mu_0(\bar{Z}_i(\omega_Z)\bar{V}_i(\omega_V)' - \bar{Z}_i\bar{V}_i')\Delta_\xi||^2\right]
\]

\[
+ E\left[||\mu_0\bar{Z}_i\bar{V}_i'\xi_0||^2\right]^{1/2}
\]

\[
+ E\left[||\Delta_\mu\bar{Z}_i\bar{V}_i'\Delta_\xi||^2\right]^{1/2}
\]

\[
+ E\left[||\Delta_\mu\bar{Z}_i\bar{V}_i'\Delta_\xi||^2\right]^{1/2}
\]

Let us apply 40 with \(A = \mu_0\) the identity, \(\zeta = \xi_0\), \(R_i() = \bar{Z}_i()\), \(H_i() = \bar{V}_i()\), \(\chi_R = \omega_Z\), \(\chi_{R,0} = \omega_{Z,0}\), \(\chi_H = \omega_V\), and \(\chi_{H,0} = \omega_{V,0}\). We get:

\[
E[||\mu_0\bar{Z}_i(\omega_Z)\bar{V}_i(\omega_V)'\xi_0 - \mu_0\bar{Z}_i\bar{V}_i'\xi_0||^2]^{1/2}
\]

\[
\leq \tilde{\mu}\xi(\sqrt{d_X\sigma_Z|X_D\delta_{\omega,V,n}} + \sqrt{d_X\sigma_{\bar{V}}|X_D\delta_{\omega,z,n}} + \sqrt{d_X + d_Ds}\delta_{\omega,z,n}\delta_{\omega,v,n})
\]

Repeating the above but with \(A = \mu_0 - \mu\) instead of \(A = \mu_0\) we get:

\[
E\left[||\Delta_\mu(\bar{Z}_i(\omega_Z)\bar{V}_i(\omega_V)'\xi_0 - \mu_0\bar{Z}_i\bar{V}_i'\xi_0)||^2\right]^{1/2}
\]

\[
\leq \delta_\mu,n\tilde{\xi}(\sqrt{d_X\sigma_Z|X_D\delta_{\omega,v,n}} + \sqrt{d_X\sigma_{\bar{V}}|X_D\delta_{\omega,z,n}} + \sqrt{d_X + d_Ds}\delta_{\omega,z,n}\delta_{\omega,v,n})
\]

Using the steps above but with \(\zeta = \xi_0 - \xi\) instead of \(\zeta = \xi_0\):

\[
E\left[||\Delta_\mu(\bar{Z}_i(\omega_Z)\bar{V}_i(\omega_V)' - \bar{Z}_i\bar{V}_i')\Delta_\xi||^2\right]^{1/2}
\]

\[
\leq \delta_\mu,n\delta_\xi,n(\sqrt{d_X\sigma_Z|X_D\delta_{\omega,v,n}} + \sqrt{d_X\sigma_{\bar{V}}|X_D\delta_{\omega,z,n}} + \sqrt{d_X + d_Ds}\delta_{\omega,z,n}\delta_{\omega,v,n})
\]

And applying the steps above but with \(A = \mu_0\) instead of \(A = \mu_0 - \mu\):

\[
E\left[||\Delta_\mu(\bar{Z}_i(\omega_Z)\bar{V}_i(\omega_V)'\xi_0 - \mu_0\bar{Z}_i\bar{V}_i'\xi_0)||^2\right]^{1/2}
\]

\[
\leq \tilde{\mu}\delta_\xi,n(\sqrt{d_X\sigma_Z|X_D\delta_{\omega,v,n}} + \sqrt{d_X\sigma_{\bar{V}}|X_D\delta_{\omega,z,n}} + \sqrt{d_X + d_Ds}\delta_{\omega,z,n}\delta_{\omega,v,n})
\]

Next note that:

\[
E[||\bar{V}_i'\xi_0||^2|\bar{Z}_i] = ||E[\bar{V}_i'\bar{V}_i|\bar{Z}_i]||^{1/2}\Sigma^{-1/2}_{\bar{V}_i}\Sigma_{\bar{V}_i'}\xi_0||^{1/2}
\]

\[
\leq ||E[\bar{V}_i'\bar{V}_i|\bar{Z}_i]||^{1/2}\Sigma^{-1/2}_{\bar{V}_i}\Sigma_{\bar{V}_i'}\xi_0||^{1/2}
\]

\[
\leq \tilde{\sigma}_{\bar{V}_i}\xi_0^2
\]
Applying the above we get:

\[
E[|\Delta_\mu \tilde{Z}_i \tilde{V}'_i \xi_0|^2] = E[|\Delta_\mu \tilde{Z}_i|^2 \cdot E[|\tilde{V}'_i \xi_0|^2 | \tilde{Z}_i]] \\
\leq \tilde{\sigma}_{\tilde{V}|Z}^2 \xi^2 E[|\Delta_\mu \tilde{Z}_i|^2] \\
\leq \tilde{\sigma}_{\tilde{V}|Z}^2 \xi^2 d_X |\Delta_\mu \Sigma_{\omega}^{1/2}|^2 \\
\leq d_X \delta_{\mu,n}^2 \tilde{\sigma}_{\tilde{V}|Z}^2
\]

Where the penultimate line follows from 36. Applying the same steps with \(\xi_0 - \xi\) in place of \(\xi_0\):

\[
E[|\Delta_\mu \tilde{Z}_i \tilde{V}'_i \Delta_\xi|^2]^{1/2} \leq \sqrt{d_X} \delta_{\mu,n} \tilde{\sigma}_{\tilde{V}|Z}
\]

Applying the same steps with \(\mu_0\) in place of \(\mu_0 - \mu\):

\[
E[|\mu_0 \tilde{Z}_i \tilde{V}'_i \Delta_\xi|^2]^{1/2} \leq \sqrt{d_X} \tilde{\mu} \delta_{\xi,n} \tilde{\sigma}_{\tilde{V}|Z}
\]

So the final term in 39 is bounded by:

\[
E[|\mu_0 \tilde{Z}_i (\omega V) V_i (\omega V)' \xi - \mu_0 \tilde{Z}_i \tilde{V}'_i \xi_0|^2]^{1/2} \\
\leq (\tilde{\mu} + \delta_{\mu,n})(\tilde{\xi} + \delta_{\xi,n})(\sqrt{d_X} \tilde{\sigma}_{Z}|D\delta_{\gamma,Y,n} + \sqrt{d_X} \tilde{\sigma}_{V}|D\delta_{\gamma,X,n} + \sqrt{d_X} + d_D s_X D\delta_{\omega,Z,n} \delta_{\omega,V,n}) \\
+ (\delta_{\mu,n} \tilde{\xi} + \tilde{\mu} \delta_{\xi,n} + \delta_{\mu,n} \delta_{\xi,n}) \sqrt{d_X} \tilde{\sigma}_{\tilde{V}|\tilde{Z}}
\]

Combining everything, we see that 39 and therefore \(E[||g_i(\beta_0, \eta_0) - g_i(\beta_0, \eta)||^2]^{1/2}\) is bounded by \(c_{4,n}\) given by:

\[
c_{4,n} = \sigma_X \sigma_Y (\sqrt{d_X} \tilde{\sigma}_{\tilde{X}|D}|D\delta_{\gamma,Y,n} + \sqrt{d_X} \tilde{\sigma}_{V}|D|D\delta_{\gamma,X,n} + \sqrt{d_X} s_D D|D\delta_{\gamma,Y,n}) \\
+ (\tilde{\mu} + \delta_{\mu,n})(\sqrt{d_X} \tilde{\sigma}_{Z}|X D|D\delta_{\omega,Y,n} + \sqrt{d_X} \tilde{\sigma}_{V}|X D|D\delta_{\omega,X,n} + \sqrt{d_X} + d_D s_X D|D\delta_{\omega,Z,n} \delta_{\omega,V,n}) \\
+ (\delta_{\mu,n} \tilde{\xi} + \tilde{\mu} \delta_{\xi,n} + \delta_{\mu,n} \delta_{\xi,n}) \sqrt{d_X} \tilde{\sigma}_{\tilde{V}|\tilde{Z}}
\]

The terms on the right hand side without \(n\) subscripts are independent of \(n\) with the sole exception of \(d_D\). As such we get:

\[
c_{4,n} \lesssim \delta_{\gamma,Y,n} + \delta_{\gamma,X,n} + \delta_{\gamma,Y,n} + \delta_{\gamma,X,n} + \delta_{\gamma,Y,n} + \delta_{\gamma,X,n} + \delta_{\gamma,Y,n} + \delta_{\gamma,X,n} + \sqrt{d_D} \delta_{\gamma,X,n} (\delta_{\gamma,Y,n} + \delta_{\gamma,X,n} + \delta_{\gamma,X,n}) \\
+ \sqrt{d_D} \delta_{\omega,Z,n} (\delta_{\omega,Y,n} + \delta_{\omega,X,n} + \delta_{\omega,X,n}) + \sqrt{d_D} s_X D|D\delta_{\omega,Z,n} \delta_{\omega,V,n}) \\
+ (\delta_{\mu,n} \tilde{\xi} + \tilde{\mu} \delta_{\xi,n} + \delta_{\mu,n} \delta_{\xi,n}) \sqrt{d_X} \tilde{\sigma}_{\tilde{V}|\tilde{Z}}
\]

Condition 6
We now consider \( \sup_{r \in (0, 1), \eta \in \mathcal{T}_n} \| \frac{\partial^2}{\partial r^2} E \left[ g_i(\beta_0, \eta_0 + r(\eta - \eta_0)) \right] \| \). We will show that if \( r \in (0, 1) \), \( \eta \in \mathcal{T}_n \) and \( P \in \mathcal{P}_n \) then \( \| \frac{\partial^2}{\partial r^2} E \left[ g_i(\beta_0, \eta_0 + r(\eta - \eta_0)) \right] \| \) is bounded by \( c_{5,n} \) given below:

\[
c_{5,n} = 2\sigma_X \delta_{\gamma,X,n}(\delta_{\gamma,Y,n}\sigma_Y + \delta_{\gamma,V,n}\xi + 2\delta_{\gamma,V,n}\xi + \delta_{\gamma,X,n}\beta) + 6(\bar{\mu} + \delta_{\mu,n})\delta_{\omega,Z,n}(\delta_{\omega,Y,n}\sigma_Y + \delta_{\omega,V,n}\xi + 3\delta_{\omega,V,n}\xi) \lesssim \delta_{\gamma,X,n}(\delta_{\gamma,Y,n} + \delta_{\gamma,V,n} + \delta_{\gamma,X,n}) + \delta_{\omega,Z,n}(\delta_{\omega,Y,n} + \delta_{\omega,V,n})
\]

From the definition of \( g_i \) we have the following:

\[
E \left[ g_i(\beta_0, \eta_0 + r(\eta - \eta_0)) \right] = E \left[ \tilde{X}_i(\gamma X_0 + r\Delta_{\gamma,X}) \tilde{Y}_i(\gamma Y_0 + r\Delta_{\gamma,Y}) - \tilde{V}_i(\gamma V_0 + r\Delta_{\gamma,V}) \right]
\]

Substituting definitions into the right-hand side and then using out assumptions to simplify the expression this yields the following:

\[
E \left[ g_i(\beta_0, \eta_0 + r(\eta - \eta_0)) \right] = r^2 \Delta_{\gamma,X} \Sigma_D (\Delta'_{\gamma,Y} - \Delta'_{\gamma,V}(\xi_0 + r\Delta_\xi) - \Delta'_{\gamma,X}\beta_0) - r^2 \mu_0 \Delta'_{\omega,Z} \Sigma_X D (\Delta'_{\omega,Y} - \Delta'_{\omega,V}(\xi_0 + r\Delta_\xi)) - r^3 \Delta_\mu \Delta'_{\omega,Z} \Sigma_X D (\Delta'_{\omega,Y} - \Delta'_{\omega,V}(\xi_0 + r\Delta_\xi))
\]

Taking second derivatives we get the equality below:

\[
\frac{\partial^2}{\partial r^2} E \left[ g_i(\beta_0, \eta_0 + r(\eta - \eta_0)) \right] = 2\Delta_{\gamma,X} \Sigma_D (\Delta'_{\gamma,Y} - \Delta'_{\gamma,V}(\xi_0 + 3r\Delta_\xi) - \Delta'_{\gamma,X}\beta_0) - 2\mu_0 \Delta'_{\omega,Z} \Sigma_X D (\Delta'_{\omega,Y} - \Delta'_{\omega,V}(\xi_0 + 3r\Delta_\xi)) - 6\Delta_\mu \Delta'_{\omega,Z} \Sigma_X D (\Delta'_{\omega,Y} - \Delta'_{\omega,V}(\xi_0 + 2r\Delta_\xi))
\]

With repeated application of the triangle inequality, Cauchy-Schwartz, and the properties of the matrix norm, and using the properties of \( \mathcal{T}_n \) we get:

\[
\| \frac{\partial^2}{\partial r^2} E \left[ g_i(\beta_0, \eta_0 + r(\eta - \eta_0)) \right] \|
\]

\[
= 2\| \Sigma^{1/2}_X \| \delta_{\gamma,X,n}(\delta_{\gamma,Y,n} E[Y_{\hat{\gamma}}^2]^{1/2} + \delta_{\omega,V,n}(\| \Sigma^{1/2}_Y \xi_0 \| + 2r\| \Sigma^{1/2}_Y \Delta_\xi \|) + \delta_{\gamma,X}(\| \Sigma^{1/2}_Y \beta_0 \|))
\]

\[
+ 2\| \mu_0 \Sigma^{1/2}_Z \| \delta_{\omega,Z,n}(\delta_{\gamma,Y,n} E[Y_{\hat{\gamma}}^2]^{1/2} + \delta_{\omega,V,n}(\| \Sigma^{1/2}_Y \xi_0 \| + 3r\| \Sigma^{1/2}_Y \Delta_\xi \|))
\]

\[
+ 6r\| \delta_\mu \Sigma^{1/2}_Z \| \delta_{\omega,Z,n}(\delta_{\gamma,Y,n} E[Y_{\hat{\gamma}}^2]^{1/2} + \delta_{\omega,V,n}(\| \Sigma^{1/2}_Y \xi_0 \| + 2r\| \Sigma^{1/2}_Y \Delta_\xi \|))
\]

42
The expression above is maximized over $r \in [0, 1]$ by setting $r = 1$. Simplifying and employing properties of $T_n$ and Assumption 4.1 we get the bound below:

$$c_{5,n} = 2\sigma_X \delta_{\gamma,X,n} (\delta_{\gamma,Y,n} \sigma_Y + \delta_{\gamma,V,n} \bar{\xi} + 2\delta_{\gamma,V,n} \delta_{\xi,n} + \delta_{\gamma,X,n} \bar{\beta})$$

$$+ 6(\bar{\mu} + \delta_{\mu,n})\delta_{\omega,Z,n} (\delta_{\omega,Y,n} \sigma_Y + \delta_{\omega,V,n} \bar{\xi} + 3\delta_{\omega,V,n} \delta_{\xi,n})$$

$$\leq \delta_{\gamma,X,n} (\delta_{\gamma,Y,n} + \delta_{\gamma,V,n} + \delta_{\gamma,X,n}) + \delta_{\omega,Z,n} (\delta_{\omega,Y,n} + \delta_{\omega,V,n})$$

$\square$