Canonical correlation regression with noisy data

Isaac Meza
Department of Economics, Harvard University

Rahul Singh
Society of Fellows and Department of Economics, Harvard University

December 2025

Abstract

We study instrumental variable regression in data rich environments. The goal is to estimate a linear model from many noisy covariates and many noisy instruments. Our key assumption is that true covariates and true instruments are repetitive, though possibly different in nature; they each reflect a few underlying factors, however those underlying factors may be misaligned. We analyze a family of estimators based on two stage least squares with spectral regularization: canonical correlations between covariates and instruments are learned in the first stage, which are used as regressors in the second stage. As a theoretical contribution, we derive upper and lower bounds on estimation error, proving optimality of the method with noisy data. As a practical contribution, we provide guidance on which types of spectral regularization to use in different regimes.

We have an application in progress using the text and image embeddings from Bach et al. (2024) (R&R JASA).

Keywords: Factor model, instrumental variable, measurement error.

1 Introduction and related work

Linear instrumental variable regression is among the most widely used methods in economics, where it is increasingly implemented in data rich environments: settings where the economist has access to many noisy covariates and many noisy instruments. In macroeconomics and finance, covariates and instruments are often constructed from many noisy measurements of a few underlying factors. In applied microeconomics, covariates and instruments constructed from texts, surveys, and transactions are high dimensional and contaminated by measurement error.

In data rich environments, a common practice is to use some version of principal component analysis to compress the information in the covariates and instruments before running two stage least squares regression. However, there is little practical guidance on how to choose among the competing spectral procedures. Moreover, there is little theoretical guidance on whether the recovered low-dimensional signal is strong enough—and, more subtly, well-aligned enough—to recover the parameter of interest despite the endogeneity and noise. Our research question is how to optimally estimate the linear instrumental variable regression parameter in noisy, data rich environments where the signal of the covariates differs from the signal of the instruments.

Our key assumption is that the covariates and instruments are each repetitive in nature: they each are low rank, with a few factors providing the signal component to many noisy measurements. Crucially, we analyze not only regimes in which the covariate factors and instrument factors are well aligned, but also regimes in which they are poorly aligned. We interpret the extreme case in which the factors are perfectly aligned as a repeated measurement model. We interpret the other extreme case, in which the factors are perfectly misaligned, as an instrumental variable model with irrelevant instruments. We interpret the case of poorly aligned factors as an instrumental variable model with weak instruments. This paper provides a unified analysis across this continuum of regimes for a family of regularized estimators.

Our first contribution is to derive nonasymptotic upper bounds on the estimation error for a family of estimators that we refer to as canonical correlation regression. Different estimators in the family have different "first stages" and the same "second stage." The first stage may be principal component analysis applied separately to the covariates and instruments; canonical correlation analysis applied jointly to the covariates and instruments; or any interpolation thereof. The second stage is regression upon the learned signal from the first stage. By providing a unified analysis for this family of estimators in the presence of noise, we characterize which version an economist should use, depending upon the regime. Our upper bound analysis culminates in a phase diagram, summarizing when to use which estimator.

Our second contribution is to derive a nonasymptotic lower bound on the estimation error in the presence of endogeneity and noise. Like the upper bound, the lower bound is valid across a continuum of regimes. By combining our upper bound and lower bound, we conclude that canonical correlation regression is sometimes optimal in noisy, data rich environments. A particular variation of the estimator is optimal in particular regimes. Intuitively, those regimes are ones in which the instrument is strong enough.

We contribute to a mature literature on ordinary least squares and two stage least squares with noisy factor models, by considering the realistic scenario in which the covariate factors and instrument factors are not perfectly aligned. The regression model is an instrumental variable regression model in which the covariates and instruments coincide. In such case, the estimator simplifies to principal component analysis of the covariates, followed by least squares. This estimator has been extensively studied in the context of noisy data under various names: factor augmented regression, diffusion indices, or principal component regression (Stock and Watson, 2002; Bai and Ng, 2006; Agarwal et al., 2025; Agarwal and Singh, 2021). To the best of our knowledge, previous analyses of two stage least squares with noisy data

all place a strong assumption that we seek to relax: that the factors of the covariates and instruments are the same (Bai and Ng, 2010; Bai and Wang, 2016), and therefore that the instruments are essentially repeated measurements of the covariates. By contrast, in this work, we allow the factors of the covariates and the factors of the instruments to differ. We characterize the entire continuum from perfect alignment to perfect misalignment, which we view as the realistic continuum of settings faced by empirical economists.

We contribute to the mature literatures on high dimensional and nonparametric instrumental variable regression, with possibly weak instruments, by considering the realistic scenario in which the covariates and instruments are noisy. The literature on many weak instruments is vast and growing (Andrews, 2016, 2018; Mikusheva and Sun, 2022; Lim et al., 2024b,a), and some papers explicitly consider regularization (Carrasco and Tchuente, 2016; Ding et al., 2025), though all of these analyses appear to require clean data. For sparse and clean data, other works consider regularized approaches, under first stage restricted eigenvalue conditions (Belloni et al., 2014; Carrasco, 2012; Gautier and Rose, 2011). Finally, for nonparametric relationships in clean data, several works consider spectral regularization (Carrasco et al., 2007; Darolles et al., 2011; Chen and Christensen, 2018; Sun et al., 2024; Meunier et al., 2025; van der Laan et al., 2025). The linear versions of these estimators often belong to the family of estimators we call canonical correlation regression. Our analysis is complementary to these various works because we consider measurement error.

Finally, we contribute to the literature on canonical correlation analysis by emphasizing its central role in economic data analysis. Canonical correlation analysis is a classical tool (Hotelling, 1936); see e.g., Bykhovskaya and Gorin (2025a,b) for recent reviews. We argue that it is often implicit in the analysis of instrumental variables, and is, in fact, a crucially important lens through which to view many estimators and many regimes. The asymptotic

literature on canonical correlation analysis (Benaych-Georges and Nadakuditi, 2012; Onatski et al., 2013, 2014; Dobriban, 2017; Bao et al., 2019) clarifies the meaning of our finite-sample bounds. In particular, it helps us to interpret the weak-alignment settings in which sample canonical correlations are unreliable for instrumental variable estimation.

Section 2 introduces the continuum of models. Section 3 describes the family of estimators. Section 4 derives the upper bounds on estimation error, and summarizes them with phase diagrams. Section 5 derives the minimax lower bound. Section 6 illustrates our theory with simulations. Section 7 concludes.

2 Many noisy covariates and instruments

We formalize a data rich environment with many noisy covariates and instruments. We assume factor structure but allow a continuum of models in which the factors of covariates and instruments may be well aligned or poorly aligned.

As running convention, for a matrix A, we denote its transpose by $A^\top$ and its pseudo-inverse by $A^\dagger$. We denote its operator norm by $\|A\|_2$, and its rth singular value by $\sigma_r(A)$. Finally, we denote the projection onto the column space of A by proj_A. For a vector v, we denote its ℓ_2 norm by $\|v\|_2$. For scalars a and b, we write $a \lesssim b$ when $a \leq Cb$ for some universal constant C. Expectations are with respect to the randomness of the structural disturbance ε^* introduced below: $\mathbb{E}(\cdot \,|\, \cdot) = \int (\cdot) \mathrm{d}\mathbb{P}(\varepsilon^* \,|\, \cdot)$.

2.1 Instrumental variable model

Consider the linear instrumental variable regression model with a scalar outcome, high dimensional covariates, and high dimensional instruments. In matrix notation, we express these quantities as $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $W \in \mathbb{R}^{n \times q}$, respectively. We posit that the outcome

Y is generated by the linear model

$$Y = X\beta_0 + \varepsilon^*, \quad \mathbb{E}(\varepsilon^* \mid W) = 0,$$

where $\varepsilon^* \in \mathbb{R}^n$ is a structural disturbance. In this causal model, $\mathbb{E}(\varepsilon^* \mid X) \neq 0$ yet $\mathbb{E}(\varepsilon^* \mid W) = 0$; the covariates X are endogenous, so we leverage the instruments W. The causal parameter of interest is the coefficient vector $\beta^* \in \mathbb{R}^p$, which we aim to estimate well. We place a weak regularity condition on this coefficient.

Assumption 1 (Parameter class)**.** *The target parameter β^* is the minimal ℓ_2 norm element of the equivalence class $\{\beta : X\beta = X\beta_0\}$. It satisfies $\|\beta^*\|_2 \leq B$ for a fixed radius $B > 0$.*

The crux of our problem is that the economist does not directly observe (Y, X, W); instead, the economist observes (Y, Z_X, Z_W), where $Z_X \in \mathbb{R}^{n \times p}$ are noisy measurements of the covariates, and $Z_W \in \mathbb{R}^{n \times q}$ are noisy measurements of the instruments. We define the perturbations as $H_X = Z_X - X$ and $H_W = Z_W - W$. Under these definitions,

$$Z_X = X + H_X, \quad Z_W = W + H_W,$$

and we have not yet placed any independence assumptions on the noise. These noise terms can capture measurement errors, discretizations, or privacy mechanisms applied to the true covariates and instruments.

What independence conditions are required on the structural disturbance ε^* and the noise terms H_X and H_W? For the structural disturbance, we have already assumed validity of the instrument via $\mathbb{E}(\varepsilon^* \mid W) = 0$. Throughout the paper, we also maintain that the structural disturbance is independent across observations and has bounded variance.

Assumption 2 (Structural disturbance)**.** *Conditional on (X, W, H_X, H_W), the structural disturbances $\varepsilon_1^*, \ldots, \varepsilon_n^*$ are independent with $\mathbb{E}(\varepsilon_i^* \mid W_i) = 0$ and $\mathbb{E}\{(\varepsilon_i^*)^2 \mid X, W, H_X, H_W\} \leq \bar{\sigma}^2$.*

For the noises, we remain largely agnostic: the noise terms on different covariates, on differ-

ent instruments, and on covariates and instruments, may be weakly correlated and heteroscedastic. Our general, non-asymptotic results only use the independence and randomness of the structural disturbance; in this sense, they are conditional upon the signal and noise (X,W,H_X,H_W).

An economist may place additional structure on the problem to make our general results concrete. For example, the operator norms of the noise H_X and H_W appear in our upper bounds. If the economist assumes the noise is sub-Gaussian and independent across observations, then these operator norms are small with high probability, appealing to the randomness of the noise. When deriving the lower bounds, we consider Gaussian noise for simplicity.

2.2 Main assumption: Concentrated signal, diffuse noise

Our strongest assumption is that the covariates and instruments each have factor structure. Algebraically, we assume they are low rank. Geometrically, we assume that the covariates and the instruments each have a low dimensional approximating subspace, though they are high dimensional. Moreover, the dimension of the covariate subspace is less than or equal to the dimension of the instrument subspace, generalizing the standard "first stage" rank condition in instrumental variable analysis.

Assumption 3 (Low effective dimension and enough effective instruments)**.** *We assume that* $\operatorname{rank}(X) = k \ll \min(n,p)$ *and* $\operatorname{rank}(W) = \ell \ll \min(n,q)$ *with* $\ell \geq k$.

Since the covariates and instruments are low rank, i.e., well approximated by low dimensional subspaces, we now introduce notation to describe those subspaces. Denote the singular value decompositions of the covariates and instruments by

$$X = U_* \Sigma_* V_*^\top, \quad W = \widetilde{U}_* \widetilde{\Sigma}_* \widetilde{V}_*^\top.$$

The diagonal matrices $\Sigma_* \in \mathbb{R}^{k \times k}$ and $\widetilde{\Sigma}_* \in \mathbb{R}^{\ell \times \ell}$ contain the singular values. They are pre and post multiplied by matrices that contain left and right singular vectors, respectively. The

spans of these singular vectors define the low dimensional approximating subspaces.

To focus on the subspaces, it is sometimes helpful to discard the scaling by the singular values. For any rank-r matrix $A = U_A \Sigma_A V_A^\top$, define its "whitened" factorization $\underline{A} = U_A V_A^\top$. In particular,

$$\underline{X} = U_* V_*^\top, \quad \underline{W} = \widetilde{U}_* \widetilde{V}_*^\top.$$

In our results, the continuum of models is indexed by the alignment between the covariate factors and the instrument factors. Geometrically, this is the alignment between the covariate subspace and the instrument subspace. Algebraically, it is captured by an object we call the subspace overlap matrix

$$\widetilde{U}_*^\top U_* \in \mathbb{R}^{\ell \times k}.$$

The singular values of $\widetilde{U}_*^\top U_*$ lie in [0,1] and equal the cosines of the principal angles between $\text{span}(U_*)$ and $\text{span}(\widetilde{U}_*)$. Equivalently, they are the canonical correlations between $\underline{X}$ and $\underline{W}$.

Assumption 4 (Relevant but possibly weak instruments)**.** *We assume $\widetilde{U}_*^\top U_*$ is full rank.*

The rank of $\widetilde{U}_*^\top U_*$ is the number of covariate factors that can be learned from the instrument factors. By assuming $\widetilde{U}_*^\top U_*$ is full rank, we ensure that all of the directions of the covariate factors can be detected by the directions of the instrument factors. If $\widetilde{U}_*^\top U_*$ were rank deficient, then some covariate factors would be lost; the instruments would be irrelevant to those covariate factors. Importantly, we allow the smallest singular value of $\widetilde{U}_*^\top U_*$ to be close to zero, meaning that some covariate factors may only have weak instruments. In this way, we consider a continuum of instrumental variable models.

So far, we have placed assumptions on the signal: the covariates and instruments have factor structure, and the covariate factors can be at least weakly detected by the instrument factors. Now, we place an assumption on the noise: it is dominated by the signal in a particular sense.

Assumption 5 (Low noise to signal ratio)**.** *Assume that* $\|H_X\|_2 \lesssim \sigma_k(X)$ *and* $\|H_W\|_2 \lesssim \sigma_\ell(W)$.

Assumption 5 essentially means that the signal is concentrated while the noise is diffuse. By a concentrated signal, we mean that the smallest nonzero singular values of X and W are large enough, similar to strong factor, pervasiveness, and incoherence assumptions (Bai and Wang, 2016). By diffuse noise, we mean that the largest singular values of H_X and H_W are small enough, which is the case when their tails are well behaved. For example, Assumption 5 is satisfied when X and W have balanced nonzero singular values while H_X and H_W are sub-Gaussian, weakly dependent across columns, and independent across rows.

3 Canonical correlation regression

We describe the family of two stage least squares estimators with spectral regularization. Then we articulate our research question in more detail.

3.1 Family of estimators

A consequence of Assumption 5 is that spectral regularization will be able to separate signal from noise. In this paper, we study a family of spectrally regularized two stage least squares estimators, which differ in how they regularize in the first stage. We will prove that different forms of first stage regularization lead to different performance, depending on the alignment between covariate factors and instrument factors as well as the strength of those factors.

To state the family of estimators, we extend the singular value decomposition notation from the true covariates and instruments to their noisy measurements. For the noisy analogues, we drop the subscript $*$. For example, we write

$$Z_X = U\Sigma V^\top, \quad Z_W = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top.$$

Let $U_k\Sigma_k V_k^\top$ denote the rank-k truncated singular value decomposition of Z_X, and $\widetilde{U}_\ell\widetilde{\Sigma}_\ell\widetilde{V}_\ell^\top$ the rank-ℓ truncated decomposition of Z_W. When convenient, we abbreviate

$$\widehat{X}=U_k\Sigma_k V_k^\top,\quad \widehat{W}=\widetilde{U}_\ell\widetilde{\Sigma}_\ell\widetilde{V}_\ell^\top.$$

Finally, recall the whitening notation $\underline{\widehat{X}}=U_kV_k^\top$ and $\underline{\widehat{W}}=\widetilde{U}_\ell\widetilde{V}_\ell^\top$.

The spirit of two stage least squares is to project the covariates onto the instruments in the first stage, then to project the outcomes onto these projections in the second stage. Let $\breve{P}\in\mathbb{R}^{n\times p}$ be a high level representation of the projected covariates from the first stage. Then the abstract formulation of the second stage is

$$\widehat{\beta}=\operatorname*{argmin}_\beta\|Y-\breve{P}\beta\|_2^2.$$

Compactly, the solution can be written as $\widehat{\beta}=\breve{P}^\dagger Y$.

As a warm up, we write the oracle 2SLS estimator that has access to clean data (Y,X,W). The oracle first stage would be the projection of X onto W.

Example 1 (Oracle 2SLS)**.** The 2SLS estimator using the clean covariates is $\beta^*=\breve{P}^\dagger Y$, where

$$\breve{P}=\operatorname{proj}_W X=(\widetilde{U}_*\widetilde{U}_*^\top)(U_*\Sigma_*V_*^\top)=\widetilde{U}_*(I)(\widetilde{U}_*^\top U_*)(\Sigma_*)V_*^\top.$$

Intuitively, the first factor $\widetilde{U}_*$ is the signal of W. The last factor $V_*^\top$ is the signal of X. In the middle, we have the subspace overlap matrix $\widetilde{U}_*^\top U_*$, i.e. the true canonical correlations. It is pre-multiplied by the left weight I and post multiplied by the right weight Σ_*.

In practice, the economist actually observes (Y,Z_X,Z_W). Some sensible estimators spectrally regularize Z_X and Z_W in the first stage, either separately or jointly, before conducting least squares in the second stage.

Example 2 (PCA-2SLS)**.** Consider a 2SLS estimator that, in the first stage, conducts PCA on Z_X and Z_W separately, and then projects the former onto the latter. This estimator is

$\breve{P}^\dagger Y$, where

$$\breve{P} = \text{proj}_{\widehat{W}} \widehat{X} = (\widetilde{U}_\ell \widetilde{U}_\ell^\top)(U_k \Sigma_k V_k^\top) = (\widetilde{U}_\ell)(I)(\widetilde{U}_\ell^\top U_k)(\Sigma_k) V_k^\top.$$

This example is clearly an empirical analogue of the previous example. Instead of the true canonical correlations $\widetilde{U}_*^\top U_*$, it contains the estimated canonical correlations $\widetilde{U}_\ell^\top U_k$ as the middle factor. These are multiplied by the same left weight I and the analogous right weight Σ_k.

Alternative first stage procedures amount to different left and right weights on the estimated canonical correlations.

Example 3 (Whiten-2SLS)**.** Consider a 2SLS estimator that, in the first stage, conducts PCA on Z_X and Z_W separately, whitens them, and projects the former onto the latter. This estimator is $\breve{P}^\dagger Y$ where

$$\breve{P} = \text{proj}_{\underline{\widehat{W}}} \underline{\widehat{X}} = (\widetilde{U}_\ell \widetilde{U}_\ell^\top)(U_k V_k^\top) = (\widetilde{U}_\ell)(I)(\widetilde{U}_\ell^\top U_k)(I) V_k^\top.$$

The left weight remains I but now the right weight becomes I as well.

Example 4 (CCA-2SLS)**.** Consider a 2SLS estimator that, in the first stage, conducts CCA on Z_X and Z_W jointly, by forming a whitened cross-moment and truncating the spectrum. Concretely, define the joint CCA as $\widehat{\underline{\widehat{W}}^\top \underline{\widehat{X}}}$, where the outer $\widehat{\cdot}$ denotes the rank-k truncation. This estimator is $\breve{P}^\dagger Y$ where

$$\begin{aligned}\breve{P} &= \widehat{W}(\widehat{W}^\top \widehat{W})^\dagger \widehat{\underline{\widehat{W}}^\top \underline{\widehat{X}}} = (\widetilde{U}_\ell \widetilde{\Sigma}_\ell \widetilde{V}_\ell^\top)(\widetilde{V}_\ell \widetilde{\Sigma}_\ell^2 \widetilde{V}_\ell^\top)^\dagger \left(\widehat{\widetilde{V}_\ell \widetilde{U}_\ell^\top U_k V_k^\top} \right) \\ &= \widetilde{U}_\ell \widetilde{\Sigma}_\ell^{-1} \widetilde{V}_\ell^\top \left(\widetilde{V}_\ell \widetilde{U}_\ell^\top U_k V_k^\top \right) = (\widetilde{U}_\ell)(\widetilde{\Sigma}_\ell^{-1})(\widetilde{U}_\ell^\top U_k)(I) V_k^\top.\end{aligned}$$

In this final example, the left weight becomes $\widetilde{\Sigma}_\ell^{-1}$ while the right weight becomes I.

Definition 1 (Canonical correlation regression)**.** A canonical correlation regression is an estimator of the form $\breve{P}^\dagger Y$ where

$$\breve{P} = \widetilde{U}_\ell A_L (\widetilde{U}_\ell^\top U_k) A_R V_k^\top.$$

Here, $A_L \in \mathbb{R}^{\ell \times \ell}$ and $A_R \in \mathbb{R}^{k \times k}$ are diagonal weighting matrices. They apply different weights to the empirical canonical correlations $\widetilde{U}_\ell^\top U_k$.

The initial factor is $\widetilde{U}_\ell$, i.e. the left singular vectors of the noisy instruments, while the final factor is V_k, i.e. the right singular vectors of the noisy covariates. The middle factor $\widetilde{U}_\ell^\top U_k$ is the empirical canonical correlations. The matrices $A_L \in \mathbb{R}^{\ell \times \ell}$ and $A_R \in \mathbb{R}^{k \times k}$ are diagonal weighting matrices that vary across different estimators within the family. By differently weighting the canonical correlations, they trade off signal usage against conditioning. As a consequence, different choices can dominate in different signal and noise regimes.

3.2 Research question in detail

A growing empirical literature uses some variation of principal component analysis to denoise and compress high dimensional covariates and instruments. These procedures are attractive because they are computationally simple and robust to a wide range of measurement errors, but they leave a fundamental question unanswered: how to optimally estimate the linear instrumental variable regression parameter in noisy, data rich environments. This question has several components.

First, many different estimators that combine spectral regularization and two stage least squares are possible. It is not clear how to compare them on a common footing, and how to select one as a user. Practical guidance is missing.

Second, the statistical limits of such procedures are incompletely understood. How does the magnitude of the measurement error, together with the strength and alignment of instruments, jointly determine the best achievable accuracy? Theoretical guidance is missing.

Third, results from high-dimensional canonical correlation analysis suggest that sufficiently weak correlations can be spectrally indistinguishable from noise. When is it the case that

some components of β^* may be impossible to learn?

In what follows, we develop theory for instrumental variable estimation with noisy, high dimensional data. Our analysis unifies across alignment regimes and estimators. We establish upper and lower bounds that clarify which directions of β^* are estimable, and at what rates.

4 Upper bound on estimation error

As our first contribution, we analyze the estimation error for the family of estimators we call canonical correlation regression. We consider noisy, data rich environments. Our nonasymptotic upper bound decomposes the mean square error into interpretable terms. The main bias–variance trade-off is governed by the choice of weights (A_L, A_R).

Our bound yields a "phase diagram" in terms of singular values and condition numbers of (X, W). In some regimes, PCA-style denoising is preferred, in others whitening is preferred, and in a wide intermediate region CCA and closely related choices of (A_L, A_R) dominate.

4.1 Key quantities

A few interpretable quantities will recur in our upper and lower bounds: noise to signal ratios, and condition numbers. The noise to signal ratios are

$$\mathtt{NSR}_X = \frac{\|H_X\|_2^2}{\sigma_k(X)^2}, \quad \mathtt{NSR}_W = \frac{\|H_W\|_2^2}{\sigma_\ell(W)^2}.$$

Intuitively, they measure the diffusion of the noise relative to the concentration of the signal, as measured by the largest singular value of the former and the smallest nonzero singular value of the latter. More concentrated signal and more diffuse noise, reflected by smaller noise to signal ratios, will lead to better rates.

The condition numbers are $\kappa(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$ and, when needed, the generalized

condition number $\kappa(A,A') = \sigma_{\max}(A)/\sigma_{\min}(A')$. We will see condition numbers for the weights (A_L,A_R) as well as for the canonical correlations $(\widetilde{U}_\ell^\top U_k)$. Intuitively, the condition numbers quantify how imbalanced the spectra of these objects are. Better conditioning, reflected by smaller condition numbers, will lead to better rates.

Finally, to lighten notation, we abbreviate a key object in the definition of the canonical correlation regression: the weighted canonical correlations are $\Delta = A_L(\widetilde{U}_\ell^\top U_k)A_R \in \mathbb{R}^{\ell\times k}$, so the second stage regressors can be expressed as $\check{P} = \widetilde{U}_\ell \Delta V_k^\top$. Denote $r = \mathrm{rank}(\Delta) \le \min\{\ell,k\}$. Different Δ are given by different choices of (A_L, A_R). Clearly, Δ is an estimator of the corresponding oracle quantity in Example 1: $\Delta_* = I(\widetilde{U}_*^\top U_*)\Sigma_*$. Considering alternative weights (A_L,A_R) introduces bias, but possibly reduces variance.

4.2 Decomposing mean square error

As an initial step, we decompose the mean square error $\|\widehat{\beta} - \beta^*\|_2^2$ into three terms. The decomposition reflects the weighted canonical correlations Δ. Specifically, we decompose the space of covariates $\mathbb{R}^p$ space via three operators:

$$\Pi_{\text{row}} = V_k \text{proj}_\Delta V_k^\top, \quad \Pi_{\text{null}} = V_k(I_k - \text{proj}_\Delta)V_k^\top, \quad \Pi_\perp = I - V_k V_k^\top.$$

These operators decompose the covariates based on their estimated signal. The projection onto the estimated signal is $V_k V_k^\top$. Clearly the initial two operators add up to the projection $V_k V_k^\top$, while the third operator is the rejection $I - V_k V_k^\top$. We call the third operator $\Pi_\perp$

The initial two operators further decompose the estimated covariate signal $V_k V_k^\top$ in a way that reflects the estimator, i.e. in a way that reflects the choice of weighted canonical correlations Δ. The first operator Π_{row} subsets to the directions of the estimated covariate signal that are preserved by the first stage of canonical correlation regression. Equivalently, it subsets to the row space of $\check{P}$. The second operator Π_{null} is the remainder.

Our decomposition of mean square error, formally derived in Lemma A.1, is

$$\|\hat{\beta}-\beta^*\|_2^2=\|\Pi_{\text{row}}(\hat{\beta}-\beta^*)\|_2^2+\|(\Pi^*_{\text{null}}-\Pi_{\text{null}})\beta^*\|_2^2+\|(\Pi^*_{\perp}-\Pi_{\perp})\beta^*\|_2^2,$$

where $\Pi^*_{\text{null}}=V_*(I-\text{proj}_{\Delta_*})V_*^\top$ and $\Pi^*_{\perp}=I-V_*V_*^\top$.

We bound the first term in Theorem 1 below. We bound the second and third terms in Theorem 2 below. The first term generally dominates the second and third.

These three terms have an intuitive interpretation. The first term is the error of estimating the coefficient, restricted to the directions that canonical correlation regression can actually learn, namely the row space of $\breve{P}$. The second term is error from extracting the causally relevant signal from the covariates using the estimated signal from the instruments. The third term is the error from estimating the signal of the noisy covariates.

4.3 First main result

To begin, we bound the first term in the mean square error.

Theorem 1 (Upper bound on projected mean square error)**.** *Under Assumptions 1–5, and A.1,*

$$\begin{aligned}\mathbb{E}\|\Pi_{row}\left(\widehat{\beta}-\beta^*\right)\|_2^2\lesssim&\frac{\|(I-A_L)\|_2^2\sigma_{\max}(X)^2+\|A_L\|_2^2(\mathit{NSR}_X+\mathit{NSR}_W)\sigma_{\max}(X)^2+\|A_L\|_2^2\|\Sigma_*-A_R\|_2^2}{c_\ell^2c_k^2\sigma_{\min}(\widetilde{U}_\ell^\top U_k)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}\\&+\mathit{NSR}_X\left(\frac{\kappa(\widetilde{U}_\ell^\top U_k)^2\kappa(A_L)^2\kappa(A_R)^2}{c_\ell^2c_k^2}\right)\\&+\frac{\bar{\sigma}^2\text{rank}(\Delta)}{c_\ell^2c_k^2\sigma_{\min}(\widetilde{U}_\ell^\top U_k)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}\end{aligned}$$

for some constants c_ℓ,c_k bounded away from zero.

We interpret the upper bound as having three components: a weighting bias, a conditioning bias, and a variance.

The weighting bias is the first term. Some objects that feature prominently are $\|(I-A_L)\|_2^2$, $\|A_L\|_2^2$, and $\|\Sigma_*-A_R\|_2^2$, which are biases from weights that deviate from $A_L=I$ and $A_R=\Sigma_*$, which appear in the oracle estimator's Δ_*. This term is larger when the noise to signal

ratios are worse, and when the canonical correlations are closer to zero. We may view the denominator as the effective strength of the first stage of canonical correlation regression.

The conditioning bias is the second term. Some objects that feature prominently are the condition numbers of the weights, $\kappa(A_L)$ and $\kappa(A_R)$, and of the canonical correlations $\kappa(\widetilde{U}_\ell^\top U_k)$. This term is larger when these objects are poorly conditioned, which magnifies the covariate noise.

The variance is the third term. Similar to the weighting bias, the denominator is the effective strength of the first stage of canonical correlation regression. Now, the numerator is the effective rank of canonical correlation regression: $\text{rank}(\Delta)$. This term is large when the effective first stage is weak, or when the complexity of the estimator is high.

To complete our analysis of mean square error, we now bound its remaining components.

Theorem 2 (Upper bound on remaining mean square error)**.** *Under Assumptions 1, 4, and 5,*

$$\|(\Pi^*_{null} - \Pi_{null})\beta^*\|_2^2 + \|(\Pi^*_\perp - \Pi_\perp)\beta^*\|_2^2 \lesssim \left\{ \frac{\kappa(W)\kappa(X)}{\sigma_k(\widetilde{U}_*^\top U_*)} \right\}^2 (\texttt{NSR}_W + \texttt{NSR}_X) + \frac{\kappa(W,X)^2 \|\Sigma_* - A_R\|_2^2}{\sigma_\ell(W)^2 \sigma_k(\widetilde{U}_*^\top U_*)^2}.$$

Corollary A.1 shows how the bound in Theorem 2 is generally dominated by bound in Theorem 1.

4.4 Practical guidance

To conclude our discussion of the upper bound on estimation error, we provide practical guidance for economists choosing among the estimators within the family we call canonical correlation regression. We simplify the bounds in Theorems 1 and 2 by focusing on leading regimes. Under these simplifications, we summarize which estimators perform best via phase diagrams.

To begin, we articulate a condition under which the bound in Theorem 1 dominates the bound in Theorem 2. Specifically, under this condition, Corollary A.1 confirms that the weighting bias in Theorem 1 dominates the quantities in Theorem 2.

Assumption 6 (Well conditioned weights). *The weights satisfy* $\frac{\kappa(A_L)}{\sigma_{\min}(A_R)} = O\left\{\frac{\kappa(W)}{\sigma_{\min}(X)}\right\}$.

Having established that Theorem 1 dominates Theorem 2, we now establish which terms dominate within Theorem 1. Under the following condition, the weighting bias dominates the conditioning bias. The condition essentially means that the covariates have a low enough noise to signal ratio.

Assumption 7 (Low noise to signal). *The noise to signal ratios satisfy* $\frac{\sigma_{\max}(A_R)}{\sigma_{\max}(X)} \lesssim \sqrt{1 + \frac{\mathtt{NSR}_W}{\mathtt{NSR}_X}}$.

Finally, to simplify the diagrams, we assume the weighting bias is never too extreme.

Assumption 8 (Limited weighting bias). *Within the weighting bias,* $\|I - A_L\|_2^2 \sigma_{\max}(X)^2 + \|A_L\|_2^2 \|\Sigma_* - A_R\|_2^2$ *is at most constant order.*

Under these simplifications, there are two remaining scenarios: either the weighting bias in Theorem 1 dominates, or the variance in Theorem 1 dominates.

Corollary 1 (Bias versus variance). *Suppose the conditions of Theorems 1 and 2 hold, as well as Assumptions 6–8. Then the bias dominated region is given by* $\mathtt{NSR}_X + \mathtt{NSR}_W \gtrsim \frac{\bar{\sigma}^2 r}{\kappa(X,W)^2}$. *The variance-dominated region is given by* $\mathtt{NSR}_X + \mathtt{NSR}_W \lesssim \frac{\bar{\sigma}^2 r}{\kappa(X,W)^2} - \frac{1}{\kappa(X,W)^2}$.

These scenarios can be visualized in Figure 1. The horizontal axis quantifies how poorly the covariate factors and the instrument factors are aligned, via the cross conditioning $\kappa(X,W)$. A larger value on this axis means less signal alignment and less signal strength. The vertical axis quantifies how poorly behaved the noise to signal ratios are, via $\mathtt{NSR} = \mathtt{NSR}_X + \mathtt{NSR}_W$. A larger value on this axis means more noise relative to the signal. Intuitively, bias dominates if the signals are poorly aligned and weak relative to the noise.

Finally, we present our practical contribution. Within the bias-dominated region, we characterize which estimator performs best. This depends on $\sigma_{\min}(X)$ and $\kappa(W)$. Within

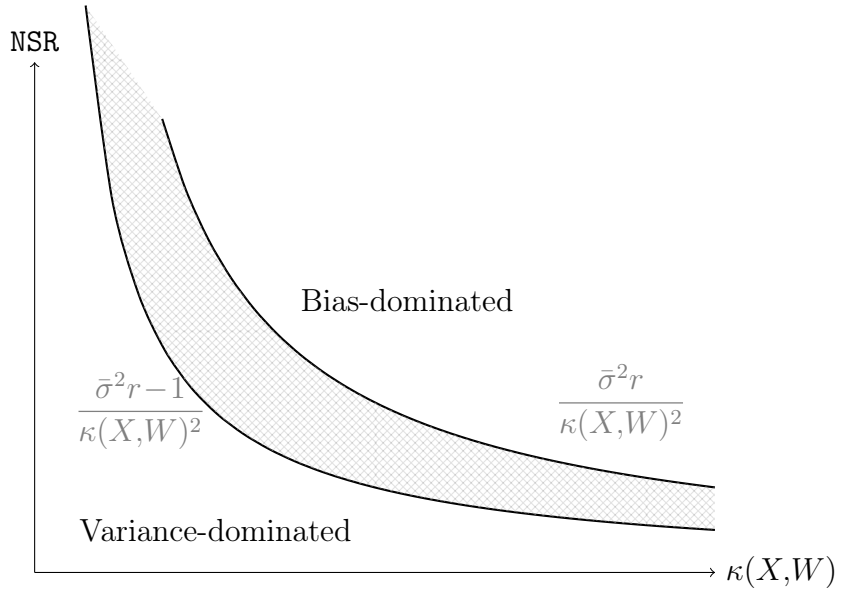


Figure 1: Phase diagram of Corollary 1

the variance-dominated region, we characterize which estimator performs best. This depends on $\sigma_{\min}(X)$ and $\sigma_{\max}(W)$.

Corollary 2 (Estimator comparison)**.** *Suppose the conditions of Corollary 1 hold. When bias dominates, Figure 2 illustrates which estimator performs best among Examples 2, 3, and 4. When variance dominates, Figure 3 illustrates which estimator performs best among Examples 2, 3, and 4.*

5 Lower bound on estimation error

Having established an upper bound on the estimation error of canonical correlation regression, we now turn to a lower bound on the estimation error for any possible estimator that uses noisy covariates and instruments (Z_X, Z_W). We show that canonical correlation regression is essentially rate optimal among possible estimators, in certain regimes. Specifically, in regimes with balanced spectra, the variance term of our upper bound (Theorem 1) matches

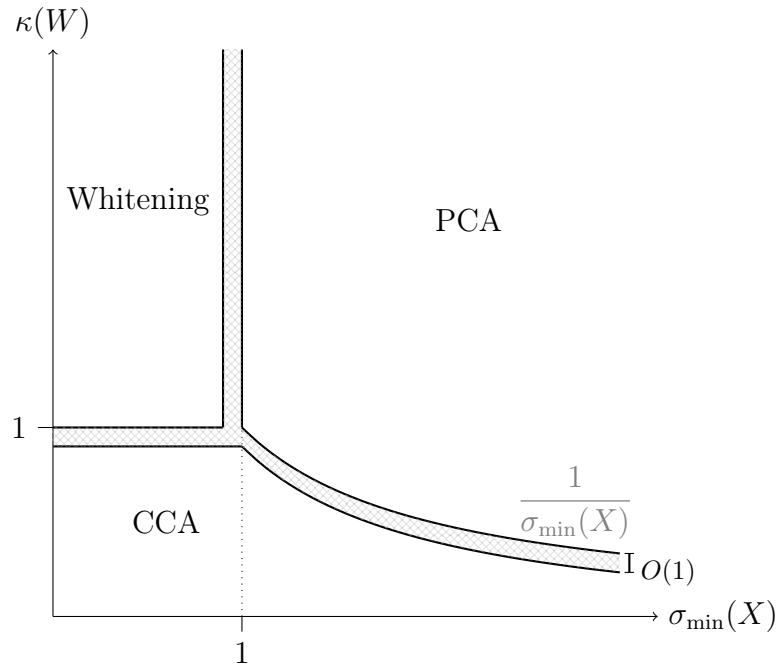


Figure 2: Which variation of canonical correlation regression to use when bias dominates

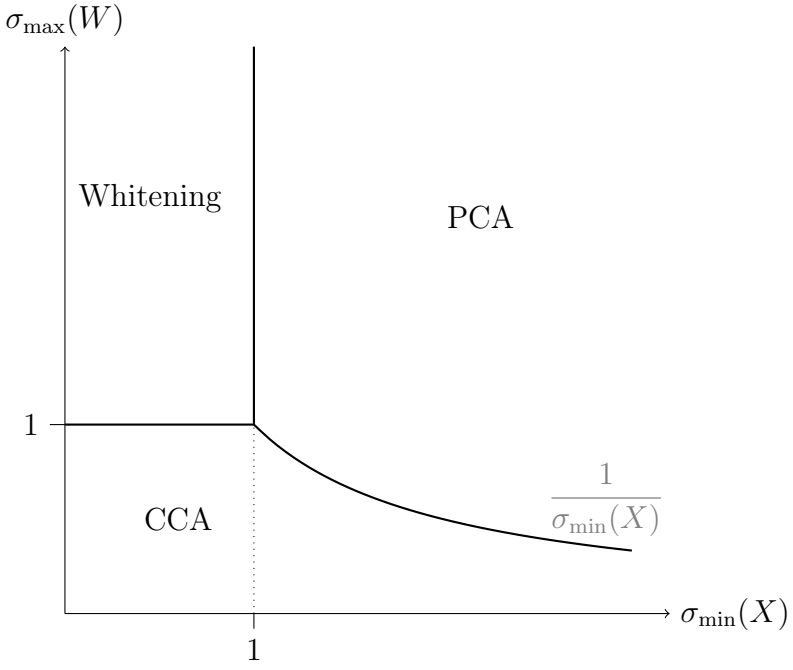


Figure 3: Which variation of canonical correlation regression to use when variance dominates

the minimax lower bound (Theorem 3) up to constants.

Our analysis also sheds light on the weak alignment regime, where canonical correlations are small due to poor alignment of the covariate factors and instrument factors, and therefore estimation breaks down. We interpret this negative result in light of classic work on weak instruments and recent work on high dimensional canonical correlation analysis.

5.1 Second main result

Throughout this section, we work conditionally on (X,W). We view the structural disturbance and measurement error (ε^*,H_X,H_W) as Gaussian with known variances $(\sigma_\varepsilon^2,\sigma_X^2,\sigma_W^2)$. Even with this additional structure, the lower bound will match our upper bound. The lower bound is obtained by a Fano-type packing argument over (X,W,β), where the different hypotheses are separated in the parameter β but induce probability distributions on (Y,Z_X,Z_W) that are difficult to distinguish.

Theorem 3 (Minimax lower bound)**.** *Under Assumption 1, with* $r_* = \mathrm{rank}(\Delta_*)$,

$$\inf_{\widehat{\beta}} \sup_{\beta} \mathbb{E}[\|\widehat{\beta}-\beta\|_2^2] \gtrsim \max\left\{ \frac{\sigma_\varepsilon^2 r_* \sigma_{\min}(W)^2}{\sigma_{\max}(\widetilde{U}_*^\top U_*)^2 - \frac{r_*\sigma_X^2\sigma_{\min}(W)^2}{8}}, \frac{\sigma_\varepsilon^2 r_* \sigma_{\min}(W)^2}{\sigma_{\max}(\widetilde{U}_*^\top U_*)^2} \right\}.$$

Remark 1 (Simplification)**.** The two terms in the lower bound closely resemble each other, up to an extra subtraction in the denominator of the first term. This subtraction is negligible as long as we are in a low-rank regime:

$$r_* < \frac{8\sigma_{\max}(\widetilde{U}_*^\top U_*)^2}{\sigma_X^2\sigma_{\min}(W)^2} = \frac{8}{\sigma_X^2}\kappa(\widetilde{U}_*^\top U_*, W).$$

Then, the lower bound simplifies to one term: $\frac{\sigma_\varepsilon^2 r_* \sigma_{\min}(W)^2}{\sigma_{\max}(\widetilde{U}_*^\top U_*)^2}$.

The lower bound fundamentally depends on the three quantities: the largest canonical correlation via $\sigma_{\max}(\widetilde{U}_*^\top U_*)^2$, the weakest instrument factor via $\sigma_{\min}(W)^2$, and the dimension of the oracle weighting via $r_* = \mathrm{rank}(\Delta_*)$. The fraction $\frac{\sigma_{\max}(\widetilde{U}_*^\top U_*)^2}{\sigma_{\min}(W)^2}$ may be viewed as the

effective strength of the first stage. Theorem 3 shows that even with the best possible estimator, the squared error must be at least $\sigma_\varepsilon^2 r_*$ divided by this fraction. The simplified lower bound closely resembles the variance term in Theorem 1.

Remark 2 (Optimality of canonical correlation regression)**.** Consider the estimator in Example 4 with weights $A_L = \widetilde{\Sigma}_\ell^{-1}$ and $A_R = I$. When the canonical correlations are balanced in the sense that $\sigma_{\max}(\widetilde{U}_*^\top U_*) \asymp \sigma_{\min}(\widetilde{U}_*^\top U_*)$ and the instrument factors are balanced in the sense that $\sigma_{\max}(W) \asymp \sigma_{\min}(W)$, the variance term in Theorem 1 matches the lower bound in Theorem 3 up to constants. Therefore, in balanced spectral regimes where the variance term dominates the bias terms in Theorem 1, canonical correlation regression attains the minimax rate (up to constants) among all estimators based on the noisy measurements (Z_X, Z_W).

Recall that the oracle estimator (Example 1) contained the true canonical correlations $\widetilde{U}_*^\top U_*$, and their weighted versions Δ_*. We denote the number of these true canonical correlations by r_*. These oracle quantities appear in our mimimax lower bound (Theorem 3).

Meanwhile, the feasible estimators (Example 2, 3, 4) must be computed from empirical canonical correlations $\widetilde{U}_\ell^\top U_k$, weighted as Δ. We denote the number of the (weighted) empirical canonical correlations by r. The hyperparameters (ℓ, k), together with positive definite weights (A_L, A_R), imply r. These empirical quantities appear in our upper bound of canonical correlation regression (Theorems 1 and 2).

Our upper bounds and lower bounds match when the empirical quantities match the oracle quantities. In practice, this amounts to "oracle tuning" of the hyperparameters (ℓ, k): they are correctly chosen, perhaps based on scree plots or some data driven procedure. As long as they are well chosen, and the factors are aligned well enough, the empirical canonical correlations will estimate the true canonical correlations well with high probability, implying that the upper and lower rates match.

By preserving the generality on non-oracle tuning in our upper bound analysis (Theorems 1 and 2), our upper rates for canonical correlation regression continue to describe its behavior under general forms of tuning.

5.2 Weak instruments as weak alignment

In this work, we view weak instruments as the setting where the covariate factors and instrument factors are weakly aligned, as measured by the canonical correlations $\widetilde{U}_*^\top U_*$. With weak enough canonical correlations, our lower bound explodes, and it becomes impossible to learn the instrumental variable regression coefficient β^*. This negative result echoes classic results on weak instruments (Staiger and Stock, 1997). For the remainder of this section, we relate this negative result to the more modern asymptotic theory on high dimensional canonical correlation regression (Baik et al., 2005; Benaych-Georges and Nadakuditi, 2012; Bao et al., 2019).

In a high-dimensional canonical correlation setting with $p,q,n \to \infty$ at proportions, the random-matrix literature shows that there is a critical canonical correlation ρ_c such that: (i) if a population canonical correlation $\rho_j \leq \rho_c$ then the associated spike is spectrally indistinguishable from noise and the sample canonical directions are asymptotically orthogonal to the truth; (ii) if $\rho_j > \rho_c$ then the sample canonical direction has a nontrivial alignment with the population one, but its variability increases as $\rho_j \downarrow \rho_c$.

Our nonasymptotic, minimax lower bound is conditional upon the realized signal; in this sense, (X,W) are fixed. In proportional asymptotics where (X,W) are random, existing theory describes their spectral behavior and, in particular, separates weak-alignment directions (below ρ_c) from strong-alignment directions (above ρ_c). Reading our lower bound through this lens suggests a refined understanding of the fundamental difficulty: there is an irreducible error associated with components of β^* that lie in weakly-aligned canonical directions—which

cannot be stably recovered when the corresponding effective canonical correlations do not separate from noise—and a variance inflation phenomenon for the estimable directions as the effective canonical correlations approach the critical level from above.

6 Simulations

We corroborate our theoretical results via simulations. The simulations document impressive finite-sample behavior of the canonical correlation regression estimators defined in Section 3.

We compare four estimators. As a benchmark, we implement 2SLS with the noisy data, without any spectral regularization. Then, we implement three variations of canonical correlation regression: PCA-2SLS (Example 2), Whiten-2SLS (Example 3), and CCA-2SLS (Example 4). These spectrally regularized estimators use the same truncation hyperparameters (k,ℓ) for comparability.

We implement several synthetic low-rank data generating processes, each with many noisy covariates and instruments, and each with correlated measurement error in covariates and instruments. Figures 4 and 5 consider high and moderate dimensional regimes, respectively. Within each figure, the subfigures consider regimes with very weak, weak, or strong alignment between the instrument factors and covariate factors. Within each subfigure, we consider different sample sizes and report mean square error. See Appendix B for details on the data generating processes.

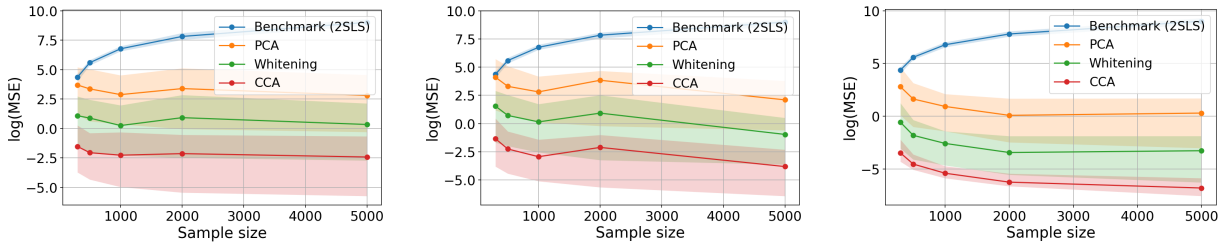
For each configuration, we run 250 independent replications with fixed base seed and independent replication seeds, and report $\mathrm{MSE}(\hat{\beta}) = \frac{1}{p}\|\hat{\beta} - \beta^*\|_2^2$, together with pointwise 2.5%–97.5% quantile bands across replications.

In these data generating processes, the dominant contribution to error is the variance term, rather than bias from weighting or conditioning. This is exactly the regime in which

our phase-diagram comparison is informative: performance differences are driven primarily by how each weighting shapes the spectrum of Δ.

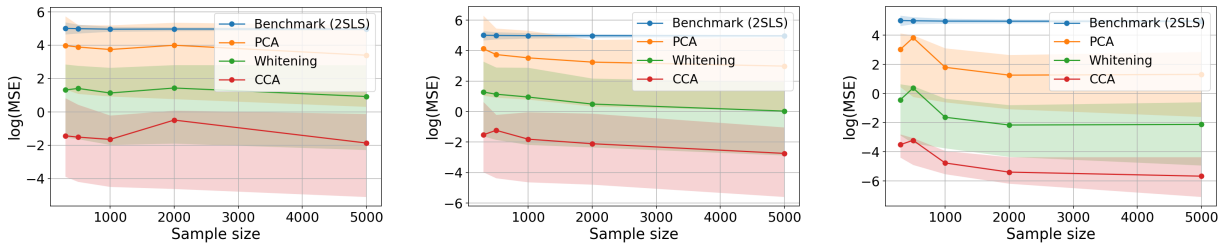
In this variance-dominated region, the phase diagram predicts that CCA-2SLS should outperform PCA-2SLS because the CCA weighting effectively stabilizes the contribution of poorly conditioned instrument directions. Equivalently, it avoids inflating the variance term by weighting towards the instruments.

The empirical curves match our theoretical prediction: CCA-2SLS yields the smallest MSE, followed by Whiten-2SLS and then PCA-2SLS. Because we simulate matched power-law spectra for X and W (a "balanced-spectrum" design), our minimax lower bound matches the variance term up to constants in this regime. Therefore, improvements beyond the CCA-2SLS variance level are information-theoretically limited.



(a) Very weak alignment (b) Weak alignment (c) Strong alignment

Figure 4: High dimensional regime



(a) Very weak alignment (b) Weak alignment (c) Strong alignment

Figure 5: Moderate dimensional regime

7 Discussion

This paper provides a unified analysis of low rank instrumental variable estimators with noisy, high dimensional data. Our main contributions are sharp upper and lower bounds on estimation error when covariate factors and instrument factors differ from each other. Theoretically, our results clarify how measurement error as well as alignment between the covariate factors and instrument factors govern the achievable accuracy. Practically, our results distinguish the regimes when an economist should choose an estimator based on PCA, whitening, or CCA. We conclude that canonical correlation regression is a natural and sometimes optimal method for exploiting low-rank structure in the presence of measurement error.

References

Agarwal, A., Shah, D., and Shen, D. (2025). On model identification and out-of-sample prediction of pcr with applications to synthetic controls. *Journal of Machine Learning Research*, 26(117):1–58.

Agarwal, A. and Singh, R. (2021). Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. arXiv:2107.02780.

Andrews, I. (2016). Conditional linear combination tests for weakly identified models. *Econometrica*, 84(6):2155–2182.

Andrews, I. (2018). Valid two-step identification-robust confidence sets for GMM. *Review of Economics and Statistics*, 100(2):337–348.

Bach, P., Chernozhukov, V., Klaassen, S., Spindler, M., Teichert-Kluge, J., and Vijaykumar, S. (2024). Adventures in demand analysis using AI. *arXiv:2501.00382*.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Bai, J. and Ng, S. (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory*, 26(6):1577–1606.

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8(1):53–80.

Baik, J., Ben Arous, G., and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697.

Bao, Z., Hu, J., Pan, G., and Zhou, W. (2019). Canonical correlation coefficients of high-dimensional gaussian vectors: Finite rank case. *Annals of Statistics*, 47(1):612–640.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.

Benaych-Georges, F. and Nadakuditi, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135.

Bykhovskaya, A. and Gorin, V. (2025a). Canonical correlation analysis: review. arXiv:2411.15625.

Bykhovskaya, A. and Gorin, V. (2025b). High-dimensional canonical correlation analysis. arXiv:2306.16393.

Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60 – 89.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398. Thirtieth Anniversary of Generalized Method of Moments.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6:5633–5751.

Carrasco, M. and Tchuente, G. (2016). Regularization based Anderson–Rubin tests for many instruments. Technical Report 1608, School of Economics, University of Kent.

Chen, X. and Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

Ding, J., Guo, X., Shi, Y., and Wang, Y. (2025). Inference of high-dimensional weak instrumental variable regression models without ridge-regularization. arXiv preprint arXiv:2504.20686.

Dobriban, E. (2017). Sharp detection in PCA under correlations: All eigenvalues matter. *Annals of Statistics*, 45(4):1810–1833.

Gautier, E. and Rose, C. (2011). High-dimensional instrumental variables regression and confidence sets. arXiv:1105.2454.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Lim, D., Wang, W., and Zhang, Y. (2024a). A conditional linear combination test with many weak instruments. *Journal of Econometrics*, 238(2):105602.

Lim, D., Wang, W., and Zhang, Y. (2024b). A valid Anderson–Rubin test under both fixed and diverging number of weak instruments. Working paper.

Meunier, D., Moulin, A., Wornbard, J., Kostic, V. R., and Gretton, A. (2025). Demystifying spectral feature learning for instrumental variable regression. arXiv preprint arXiv:2506.10899.

Mikusheva, A. and Sun, L. (2022). Inference with many weak instruments. *The Review of Economic Studies*, 89(5):2663–2686.

Onatski, A., Moreira, M. J., and Hallin, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Annals of Statistics*, 41(3):1204–1231.

Onatski, A., Moreira, M. J., and Hallin, M. (2014). Signal detection in high dimension: The multispiked case. *Annals of Statistics*, 42(1):225–254.

Petersen, K. B. and Pedersen, M. S. (2008). The matrix cookbook. Version 20081110.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Sun, H., Moulin, A., Ren, T., Gretton, A., and Dai, B. (2024). Spectral representation for causal estimation with hidden confounders. arXiv preprint arXiv:2407.10448.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer.

van der Laan, L., Kallus, N., and Bibaut, A. (2025). Nonparametric instrumental variable inference with many weak instruments. arXiv:2505.07729.

Appendix

Table of Contents

A Proofs **28**

A.1 Additional notation 28

A.2 Subspace decomposition 29

A.3 Singular vector alignment 30

A.4 Projected coefficient consistency 31

A.5 Remaining coefficient consistency 35

A.6 Comparison of estimators 37

A.7 Lower bound 39

B Simulation details **44**

B.1 Concentrated signal 44

B.2 Diffuse noise 44

B.3 Instrument validity 44

B.4 Dimension 45

A Proofs

A.1 Additional notation

To shorten the proofs, we define some additional notation. We write the subspace alignment matrices and their singular value decompositions (SVDs) as

$$M_* = \widetilde{U}_*^\top U_* = Q_* S_* R_*^\top, \quad M = \widetilde{U}^\top U = Q S R^\top, \quad \widehat{M} = \widetilde{U}_\ell^\top U_k = Q_\ell S R_k^\top.$$

Formally, let $Q_\ell := E_\ell^\top Q$ and $R_k := F_k^\top R$ denote the row truncations that select the first ℓ and k rows, respectively. Note that Q_ℓ and R_k need not have orthonormal columns; they are row-truncations of orthonormal matrices for M.

For a matrix A, we write its row, column, and null spaces by $\mathrm{Row}(A)$, $\textsc{Col}(A)$, and $\textsc{Null}(A)$, respectively.

A.2 Subspace decomposition

Lemma A.1 (Subspace decomposition)**.** *Let $Y \in \mathbb{R}^n$ and let $\widehat{\beta} = \check{P}^\dagger Y$ be the minimal-norm least-squares solution to*

$$\min_{\beta \in \mathbb{R}^p} \|Y - \check{P}\beta\|_2^2.$$

1. *$\mathbb{R}^p$ decomposes orthogonally as*

$$\mathbb{R}^p = \underbrace{V_k \textsc{Col}\big(\Delta^\top\big)}_{\mathrm{Row}(\check{P})} \oplus \underbrace{V_k \textsc{Null}(\Delta)}_{\textit{uninformative within span}(V_k)} \oplus \underbrace{\textsc{span}(V_k)^\perp}_{\textit{outside } X\textit{-span}},$$

with corresponding orthogonal projectors $\Pi_{row}, \Pi_{null}, \Pi_\perp$.

2. *We have that $\widehat{\beta} \in \textsc{span}\{V_k \textsc{Col}\big(\Delta^\top\big)\}$. Equivalently,*

$$\Pi_{null}\,\widehat{\beta} = 0 \quad \textit{and} \quad \Pi_\perp \widehat{\beta} = 0.$$

3. *These imply that*

$$\|\hat{\beta} - \beta^*\|_2^2 = \|\Pi_{row}(\hat{\beta} - \beta^*)\|_2^2 + \|(\Pi^*_{null} - \Pi_{null})\beta^*\|_2^2 + \|(\Pi^*_\perp - \Pi_\perp)\beta^*\|_2^2.$$

Proof of Lemma A.1. We proceed in steps.

1. Since $\widetilde{U}_\ell^\top$ has full row rank ℓ, $\textsc{Col}\Big(\check{P}^\top\Big) = \textsc{Col}\big(V_k \Delta^\top\big) = V_k \textsc{Col}\big(\Delta^\top\big)$. Thus

$$\mathrm{Row}(\check{P}) = \textsc{Col}\Big(\check{P}^\top\Big) = V_k \textsc{Col}\big(\Delta^\top\big).$$

Now, take a (thin) SVD $\Delta = LST^\top$ with rank r. Then

$$\mathrm{proj}_\Delta = \Delta^\top\big(\Delta\Delta^\top\big)^\dagger \Delta = TT^\top,$$

the orthogonal projector onto $\textsc{Col}\big(\Delta^\top\big)$. Therefore

$$\Pi_{\mathrm{row}} = V_k \mathrm{proj}_\Delta V_k^\top = V_k T T^\top V_k^\top$$

is the orthogonal projector in $\mathbb{R}^p$ onto $V_k \textsc{Col}\big(\Delta^\top\big)$.

Set $\Pi_{\mathrm{null}} = V_k(I_k - \mathrm{proj}_\Delta)V_k^\top$ and $\Pi_\perp = I_p - V_k V_k^\top$. Symmetry and idempotence are immediate; mutual orthogonality follows from $V_k^\top V_{k,\perp} = 0$ and $\mathrm{proj}_\Delta(I_k - \mathrm{proj}_\Delta) = 0$. Their ranges are precisely the three subspaces stated above. Since $\mathrm{proj}_\Delta + (I_k - \mathrm{proj}_\Delta) = I_k$, we have $\Pi_{\mathrm{row}} + \Pi_{\mathrm{null}} + \Pi_\perp = I_p$. This proves the first result.

2. All least-squares solutions have the form

$$\beta = \check{P}^\dagger Y + \Big(I_p - \check{P}^\dagger \check{P}\Big) z, \quad z \in \mathbb{R}^p$$

and $I_p - \breve{P}^\dagger \breve{P}$ is the orthogonal projector onto $\text{NULL}(\breve{P})$. The minimal-norm solution is obtained with $z = 0$, hence $\widehat{\beta} = \breve{P}^\dagger Y \in \text{Row}\left(\breve{P}^\top\right) = V_k \text{COL}\left(\Delta^\top\right)$. Thus $\Pi_{\text{null}}\,\widehat{\beta} = \Pi_\perp \widehat{\beta} = 0$. This proves the second result.

3. The previous decomposition allows us to break down the estimation error. Let $\Pi^*_{\text{null}} = V_*(I_k - \text{proj}_{\Delta_*})V_*^\top$, $\Pi^*_\perp = I_p - V_* V_*^\top$. Observe that

$$\begin{aligned}
\hat{\beta} - \beta^* &= \Pi_{\text{row}}\left(\hat{\beta} - \beta^*\right) + \Pi_{\text{null}}\left(\hat{\beta} - \beta^*\right) + \Pi_\perp\left(\hat{\beta} - \beta^*\right) \\
&= \Pi_{\text{row}}\left(\hat{\beta} - \beta^*\right) + \Pi_{\text{null}}(-\beta^*) + \Pi_\perp(-\beta^*) \\
&= \Pi_{\text{row}}\left(\hat{\beta} - \beta^*\right) + (\Pi^*_{\text{null}} - \Pi_{\text{null}})\beta^* + (\Pi^*_\perp - \Pi_\perp)\beta^*
\end{aligned}$$

since by the result above,

$$\Pi_{\text{null}}\hat{\beta} = 0, \quad \Pi_\perp \hat{\beta} = 0, \quad \Pi^*_{\text{null}}\beta^* = 0, \quad \Pi^*_\perp \beta^* = 0.$$

Finally, by the Pythagorean theorem

$$\|\hat{\beta} - \beta^*\|_2^2 = \|\Pi_{\text{row}}(\hat{\beta} - \beta^*)\|_2^2 + \|(\Pi^*_{\text{null}} - \Pi_{\text{null}})\beta^*\|_2^2 + \|(\Pi^*_\perp - \Pi_\perp)\beta^*\|_2^2.$$

□

A.3 Singular vector alignment

From now on, denote the thin SVD of $\Delta = U_\Delta \Sigma_\Delta V_\Delta^\top$, then we can verify that $\Pi_{\text{row}} = V_k V_\Delta V_\Delta^\top V_k^\top$.

A fundamental quantity that appears repeatedly in our bounds is the operator norm of the deviation of the right singular vectors of Z_X and left singular vectors of Z_W. To have proper bounds, we require the singular vectors to be properly aligned (Procrustes rotations).

Let

$$Q_X \in \arg\min_{Q \in O(k)} \|V_* - V_k Q\|_2, \quad R_W \in \arg\min_{R \in O(\ell)} \left\|\widetilde{U}_* - \widetilde{U}_\ell R\right\|_2.$$

Define the aligned variables

$$V'_k := V_k Q_X, \quad \widetilde{U}'_\ell := \widetilde{U}_\ell R_W, \quad \Delta' := R_W^\top \Delta Q_X.$$

Then $\widetilde{U}_\ell \Delta V_k^\top = \left(\widetilde{U}'_\ell\right)\Delta'(V'_k)^\top$, so the design operator $\breve{P}$ (and hence $\breve{P}^\dagger$ and Π_{row}) is unchanged. Write an SVD of Δ as $\Delta = U_\Delta \Sigma_\Delta V_\Delta^\top$. Then the SVD of Δ' is $\Delta' = \left(R_W^\top U_\Delta\right)\Sigma_\Delta\left(Q_X^\top V_\Delta\right)^\top$ so $\Sigma_{\Delta'} = \Sigma_\Delta, U_{\Delta'} = R_W^\top U_\Delta, V_{\Delta'} = Q_X^\top V_\Delta$. Finally, after alignment (Cai and Zhang, 2018, Lemma 1):

$$\begin{aligned}
\|V_* - V'_k\|_2 &\le \sqrt{2}\sin\Theta_{\max}(\text{SPAN}(V_*), \text{SPAN}(V_k)) \\
\left\|\widetilde{U}_* - \widetilde{U}'_\ell\right\|_2 &\le \sqrt{2}\sin\Theta_{\max}\left(\text{SPAN}(\widetilde{U}_*), \text{SPAN}(\widetilde{U}_\ell)\right).
\end{aligned}$$

Wedin's sin-Θ theorem upper bounds these quantities.

Lemma A.2 (Procrustes analysis)**.** *Under Assumption 5, and the optimal alignment from above, we have the upper bounds*

$$\|V_* - V_k'\|_2 \le \sqrt{2}\frac{\|H_X\|_2}{\sigma_k(X)}, \quad \left\|\widetilde{U}_* - \widetilde{U}_\ell'\right\|_2 \le \sqrt{2}\frac{\|H_W\|_2}{\sigma_\ell(W)}.$$

Proof. By a direct application of Wedin's Theorem given the bound obtained above. □

Henceforth, we work with these aligned representatives; without loss of generality, we may assume that we are already in the aligned gauge and we can upper bound the quantities according to the previous Lemma A.2.

A.4 Projected coefficient consistency

In this section we provide the proof of Theorem 1.

Lemma A.3 (Decomposition of projected mean square error)**.** *Under Assumptions 3, and 2, then*

$$\mathbb{E}\left\|\Pi_{row}\left(\widehat{\beta} - \beta^*\right)\right\|_2^2 \lesssim \sigma_{\min}(\Delta)^{-2}\|\beta^*\|_2^2\left(\|\Delta_* - \Delta\|_2^2 + \left\|V_*^\top - V_k^\top\right\|_2^2\sigma_{\max}(\Delta)^2 + \left\|\widetilde{U}_\ell - \widetilde{U}_*\right\|_2^2\sigma_{\max}(\Sigma_*)^2\right) + \bar{\sigma}^2\sum_{i=1}^{\operatorname{rank}(\Delta)}\frac{1}{\sigma_i(\Delta)^2}.$$

Proof. Since the estimator is given by $\widehat{\beta} = \check{P}^\dagger Y$ and the true outcome model is $Y = X\beta^* + \varepsilon^*$, we can write:

$$V_k^\top\widehat{\beta} = V_k^\top V_k\Delta^\dagger\widetilde{U}_{ell}^\top Y = \Delta^\dagger\widetilde{U}_{ell}^\top(X\beta^* + \varepsilon^*). \tag{1}$$

Using the SVD $\Delta = U_\Delta\Sigma_\Delta V_\Delta^\top$, note that

$$\Pi_{\text{row}} = V_kV_\Delta V_\Delta^\top V_k^\top = V_kV_\Delta U_\Delta^\top U_\Delta V_\Delta^\top V_k^\top = L^\top L,$$

where we define $L := U_\Delta V_\Delta^\top V_k^\top$. Consequently,

$$\left\|\Pi_{\text{row}}\left(\widehat{\beta} - \beta^*\right)\right\|_2 = \left\|L\left(\widehat{\beta} - \beta^*\right)\right\|_2 = \left\|U_\Delta V_\Delta^\top V_k^\top\left(\widehat{\beta} - \beta^*\right)\right\|_2.$$

Since U_Δ has orthonormal columns, then

$$\left\|\Pi_{\text{row}}\left(\widehat{\beta} - \beta^*\right)\right\|_2 \le \left\|V_\Delta^\top\left(V_k^\top\widehat{\beta} - V_k^\top\beta^*\right)\right\|_2. \tag{2}$$

From (1) and $\Delta^\dagger = V_\Delta\Sigma_\Delta^{-1}U_\Delta^\top$,

$$V_\Delta^\top V_k^\top\widehat{\beta} = V_\Delta^\top\Delta^\dagger\widetilde{U}_{ell}^\top(X\beta^* + \varepsilon^*) = \Sigma_\Delta^{-1}U_\Delta^\top\widetilde{U}_{ell}^\top(X\beta^* + \varepsilon^*).$$

Plugging this into (2) yields

$$\begin{aligned}\left\|\Pi_{\text{row}}\left(\widehat{\beta}-\beta^*\right)\right\|_2 &= \left\|\Sigma_\Delta^{-1}U_\Delta^\top\widetilde{U}_\ell^\top X\beta^* - V_\Delta^\top V_k^\top\beta^* + \Sigma_\Delta^{-1}U_\Delta^\top\widetilde{U}_\ell^\top\varepsilon^*\right\|_2 \\ &\le \underbrace{\left\|\left(\Sigma_\Delta^{-1}U_\Delta^\top\widetilde{U}_\ell^\top X - V_\Delta^\top V_k^\top\right)\beta^*\right\|_2}_{\text{Bias}} + \underbrace{\left\|\Sigma_\Delta^{-1}U_\Delta^\top\widetilde{U}_\ell^\top\varepsilon^*\right\|_2}_{\text{Variance}}.\end{aligned}$$

We now focus on the bias term:

$$\begin{aligned}\left\|\left(\Sigma_\Delta^{-1}U_\Delta^\top\widetilde{U}_\ell^\top X - V_\Delta^\top V_k^\top\right)\beta^*\right\|_2 &= \left\|\Sigma_\Delta^{-1}U_\Delta^\top\left(\widetilde{U}_\ell^\top X - \Delta V_k^\top\right)\beta^*\right\|_2 \\ &\le \sigma_{\min}(\Delta)^{-1}\left\|\widetilde{U}_\ell^\top X - \Delta V_k^\top\right\|_2\|\beta^*\|_2 \\ &\le \sigma_{\min}(\Delta)^{-1}\|\beta^*\|_2\left(\left\|\widetilde{U}_*^\top X - \Delta V_k^\top\right\|_2 + \left\|(\widetilde{U}_\ell - \widetilde{U}_*)^\top X\right\|_2\right) \\ &= \sigma_{\min}(\Delta)^{-1}\|\beta^*\|_2\left(\left\|\widetilde{U}_*^\top U_*\Sigma_*V_*^\top - \Delta V_k^\top\right\|_2 + \left\|(\widetilde{U}_\ell - \widetilde{U}_*)^\top X\right\|_2\right) \\ &= \sigma_{\min}(\Delta)^{-1}\|\beta^*\|_2\left(\left\|\Delta_*V_*^\top - \Delta V_k^\top\right\|_2 + \left\|(\widetilde{U}_\ell - \widetilde{U}_*)^\top X\right\|_2\right) \\ &= \sigma_{\min}(\Delta)^{-1}\|\beta^*\|_2\left(\left\|\Delta_*V_*^\top - \Delta V_k^\top\right\|_2 + \left\|(\widetilde{U}_\ell - \widetilde{U}_*)^\top U_*\Sigma_*V_*^\top\right\|_2\right) \\ &= \sigma_{\min}(\Delta)^{-1}\|\beta^*\|_2\left(\left\|\Delta_*V_*^\top - \Delta V_k^\top\right\|_2 + \left\|\widetilde{U}_\ell^\top U_*\Sigma_*V_*^\top - \widetilde{U}_\ell^\top U_*\Sigma_*V_*^\top\right\|_2\right) \\ &\le \sigma_{\min}(\Delta)^{-1}\|\beta^*\|_2\left(\left\|\Delta_*V_*^\top - \Delta V_k^\top\right\|_2 + \left\|\widetilde{U}_\ell - \widetilde{U}_*\right\|_2\sigma_{\max}(\Sigma_*)\right) \\ &\le \sigma_{\min}(\Delta)^{-1}\|\beta^*\|_2\left(\|\Delta_* - \Delta\|_2 + \left\|V_*^\top - V_k^\top\right\|_2\sigma_{\max}(\Delta) + \left\|\widetilde{U}_\ell - \widetilde{U}_*\right\|_2\sigma_{\max}(\Sigma_*)\right).\end{aligned}$$

The variance term is bounded as

$$\begin{aligned}\mathbb{E}\left[\left\|\Sigma_\Delta^{-1}U_\Delta^\top\widetilde{U}_\ell^\top\varepsilon^*\right\|_2^2\right] &= \text{Tr}\left(\Sigma_\Delta^{-1}U_\Delta^\top\widetilde{U}_\ell^\top\mathbb{E}[\varepsilon^*\varepsilon^{*\top}|X,Z_X,W,Z_W]\widetilde{U}_\ell U_\Delta\Sigma_\Delta^{-1}\right) \\ &\le \bar{\sigma}^2\text{Tr}\left(\Sigma_\Delta^{-1}U_\Delta^\top\widetilde{U}_\ell^\top\widetilde{U}_\ell U_\Delta\Sigma_\Delta^{-1}\right) \\ &= \bar{\sigma}^2\left\|\Sigma_\Delta^{-1}\right\|_{\text{Fr}}^2 \\ &= \bar{\sigma}^2\sum_{i=1}^{\text{rank}(\Delta)}\frac{1}{\sigma_i(\Delta)^2}.\end{aligned}$$

□

The error term stemming from the singular vectors are bounded using Lemma A.2. It remains to bound $\|\Delta_* - \Delta\|_2^2$ for a complete treatment of the bias term.

Lemma A.4 (Weighting bias bound)**.** *Consider the following* estimator *for* Δ_**:*

$$\Delta(A_L, A_R) = A_L(\widetilde{U}_\ell^\top U_k)A_R,$$

where $A_L \in \mathbb{R}^{\ell\times\ell}, A_R \in \mathbb{R}^{k\times k}$ diagonal and positive definite matrices. Under Assumption 5 then,

$$\|\Delta_* - \Delta(A_L,A_R)\|_2 \lesssim \|(I-A_L)\|_2\sigma_{\max}(\Sigma_*) + \|A_L\|_2\left(\frac{\|H_X\|_2}{\sigma_k(\Sigma_*)} + \frac{\|H_W\|_2}{\sigma_\ell(\widetilde{\Sigma}_*)}\right)\sigma_{\max}(\Sigma_*) + \|A_L\|_2\|\Sigma_* - A_R\|_2.$$

Proof. We write

$$\begin{aligned}
\|\Delta_* - \Delta(A_L,A_R)\|_2 &= \|M_*\Sigma_* - A_L M A_R\|_2 \\
&= \|(I-A_L)M_*\Sigma_* + A_L(M_*-M)\Sigma_* + A_L M(\Sigma_* - A_R)\|_2 \\
&\le \|(I-A_L)M_*\Sigma_*\|_2 + \|A_L(M_*-M)\Sigma_*\|_2 + \|A_L M(\Sigma_*-A_R)\|_2 \\
&\le \|(I-A_L)\|_2\sigma_{\max}(\Sigma_*) + \|A_L\|_2\|M_*-M\|_2\sigma_{\max}(\Sigma_*) + \|A_L\|_2\|\Sigma_*-A_R\|_2 \\
&\le \|(I-A_L)\|_2\sigma_{\max}(\Sigma_*) + \sqrt{2}\|A_L\|_2\left(\frac{\|H_X\|_2}{\sigma_k(\Sigma_*)} + \frac{\|H_W\|_2}{\sigma_\ell(\widetilde{\Sigma}_*)}\right)\sigma_{\max}(\Sigma_*) + \|A_L\|_2\|\Sigma_*-A_R\|_2
\end{aligned}$$

where in the last line we bound

$$\|M_* - M\|_2 = \|\widetilde{U}_*^\top U_* - \widetilde{U}_\ell^\top U_k\|_2 \le \|\widetilde{U}_*^\top - \widetilde{U}_\ell^\top\|_2 + \|U_* - U_k\|_2$$

and apply Lemma A.2. □

In what follows we make the following assumption needed for stability.

Assumption A.1 (Restricted eigenvalue condition)**.** *Define $c_\ell := \lambda_{\min}\big(Q_\ell^\top Q_\ell\big)$, and $c_k := \lambda_{\min}\big(R_k^\top R_k\big)$. We assume that $c_\ell, c_k \ge C > 0$. This is equivalent to $Q_\ell := E_\ell^\top Q$ and $R_k := F_k^\top R$ being full column rank.*

The constants c_ℓ and c_k quantify whether the truncations to ℓ instrument directions and k regressor directions retain the singular directions of the overlap operator M. Equivalently, $E_\ell^\top$ and $F_k^\top$ act as stable embeddings on the relevant overlap subspaces $\mathrm{span}(Q)$ and $\mathrm{span}(R)$.

Geometrically, the eigenvalues of $Q_\ell^\top Q_\ell$ are the squared cosines of the principal angles between $\mathrm{span}(Q)$ and the coordinate subspace $\mathrm{span}(E_\ell)$, so

$$\lambda_{\min}(Q_\ell^\top Q_\ell) > 0 \iff \Theta_{\max}(\mathrm{span}(Q), \mathrm{span}(E_\ell)) < 90^\circ,$$

and analogously for R_k. Thus c_ℓ (or c_k) close to zero means that at least one overlap direction lies almost entirely in the discarded components, making the induced first-stage operator nearly rank-deficient and inflating the error bounds. In practice, this condition rules out overly aggressive truncation.

Moreover, since Q has orthonormal columns,

$$\lambda_{\min}(Q_\ell^\top Q_\ell) = \lambda_{\min}(Q^\top E_\ell E_\ell^\top Q) = \min_{x\in\mathrm{span}(Q)\setminus\{0\}} \frac{\|E_\ell^\top x\|_2^2}{\|x\|_2^2},$$

so a uniform lower bound on c_ℓ is exactly a norm-preservation condition for the projection $E_\ell^\top$ restricted to $\mathrm{span}(Q)$ (and analogously for c_k and $F_k^\top$ on $\mathrm{span}(R)$). This viewpoint is summarized by the following restricted-eigenvalue condition.

Remark A.1 (Restricted eigenvalue interpretation)**.** There exist constants $\kappa_\ell,\kappa_k\in(0,1]$ such that

$$\lambda_{\min}(Q_\ell^\top Q_\ell)\geq\kappa_\ell,\qquad \lambda_{\min}(R_k^\top R_k)\geq\kappa_k.$$

Equivalently, for all $z\in\mathbb{R}^{\mathrm{rank}(M)}$,

$$\|Q_\ell z\|_2\geq\sqrt{\kappa_\ell}\|z\|_2,\qquad \|R_k z\|_2\geq\sqrt{\kappa_k}\|z\|_2.$$

Since Q and R are isometries, these are equivalent to the restricted norm-preservation bounds

$$\|E_\ell^\top x\|_2^2\geq\kappa_\ell\|x\|_2^2\quad\forall x\in\mathrm{span}(Q),\qquad \|F_k^\top x\|_2^2\geq\kappa_k\|x\|_2^2\quad\forall x\in\mathrm{span}(R),$$

i.e., $(E_\ell^\top,F_k^\top)$ act as subspace embeddings on the relevant overlap subspaces.

Lemma A.5 (Variance bound)**.** *Under Assumption A.1, the variance term can be bounded as*

$$\bar\sigma^2\sum_{i=1}^{\mathrm{rank}(\Delta)}\frac{1}{\sigma_i(\Delta)^2}\leq\frac{\bar\sigma^2\mathrm{rank}(\Delta)}{c_\ell^2c_k^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}.$$

Proof. The spectrum for Δ can be bounded as

$$\begin{aligned}\sigma_{\min}(\Delta)&=\sigma_{\min}(A_LE_\ell^\top\widetilde U^\top UF_kA_R)\\&\geq\sigma_{\min}(A_L)\sigma_{\min}(E_\ell)\sigma_{\min}(\widetilde U^\top U)\sigma_{\min}(F_k)\sigma_{\min}(A_R)\\&=\sigma_{\min}(A_L)\lambda_{\min}\big(Q_\ell^\top Q_\ell\big)\sigma_{\min}(\widetilde U^\top U)\lambda_{\min}\big(R_k^\top R_k\big)\sigma_{\min}(A_R)\\&=c_\ell c_k\sigma_{\min}(A_L)\sigma_{\min}(S)\sigma_{\min}(A_R)\end{aligned}$$

since $Q_\ell:=E_\ell^\top Q$, $R_k:=F_k^\top R$ and recall that both Q and R are orthonormal.

Analogously,

$$\sigma_{\max}(\Delta)\leq\sigma_{\max}(S)\sigma_{\max}(A_L)\sigma_{\max}(A_R).$$

Finally,

$$\bar\sigma^2\sum_{i=1}^{\mathrm{rank}(\Delta)}\frac{1}{\sigma_i(\Delta)^2}\leq\frac{\bar\sigma^2}{c_\ell^2c_k^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}\sum_{i=1}^{\mathrm{rank}(\Delta)}\frac{1}{\sigma_i(S)^2}\leq\frac{\bar\sigma^2\mathrm{rank}(\Delta)}{c_\ell^2c_k^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}.$$

□

The proof of Theorem 1 follows immediately after we consolidate all bounds above.

A.5 Remaining coefficient consistency

Proof of Theorem 2. We write

$$
\begin{aligned}
&\|(\Pi^*_{\text{null}} - \Pi_{\text{null}})\|_2 + \|(\Pi^*_{\perp} - \Pi_{\perp})\|_2 \\
&= \|V_* V_*^\top - V_k V_k^\top - V_* \text{proj}_{\Delta_*} V_*^\top + V_k \text{proj}_{\Delta} V_k^\top\|_2 + \|V_k V_k^\top - V_* V_*^\top\|_2 \\
&\lesssim \|V_k \text{proj}_{\Delta_*} V_k^\top - V_k \text{proj}_{\Delta_*} V_k^\top + V_* \text{proj}_{\Delta_*} V_k^\top - V_* \text{proj}_{\Delta_*} V_k^\top - V_* \text{proj}_{\Delta_*} V_*^\top + V_k \text{proj}_{\Delta} V_k^\top\|_2 \\
&\qquad + \|V_* V_*^\top - V_k V_k^\top\|_2 \\
&\le \|V_k (\text{proj}_{\Delta_*} - \text{proj}_{\Delta}) V_k^\top\|_2 + \|V_* \text{proj}_{\Delta_*} (V_k^\top - V_*^\top)\|_2 + \|(V_k^\top - V_*^\top) \text{proj}_{\Delta_*} V_k^\top\|_2 \\
&\qquad + \|V_* V_*^\top - V_k V_k^\top\|_2 \\
&\le \|\text{proj}_{\Delta_*} - \text{proj}_{\Delta}\|_2 + \|\text{proj}_{\Delta_*} (V_k^\top - V_*^\top)\|_2 + \|(V_k^\top - V_*^\top) \text{proj}_{\Delta_*}\|_2 \\
&\qquad + \|V_* V_*^\top - V_k V_k^\top\|_2 \\
&\lesssim \|\text{proj}_{\Delta_*} - \text{proj}_{\Delta}\|_2 + \|V_k^\top - V_*^\top\|_2 + \|V_* V_*^\top - V_k V_k^\top\|_2.
\end{aligned}
$$

The second and third terms are bounded by Wedin's Theorem as

$$
\|V_k^\top - V_*^\top\|_2, \|V_* V_*^\top - V_k V_k^\top\|_2 \lesssim \frac{\|H_X\|_2}{\sigma_k(\Sigma_*)}.
$$

To analyze the first term, recall that proj_{Δ_*} and proj_{Δ} are orthogonal projections onto the row space of $\Delta_* = \widetilde{U}_*^\top U_* \Sigma_*$ and $\Delta = A_L \widetilde{U}_\ell^\top U_k A_R$ respectively.

Therefore,

$$\begin{aligned}
\|\mathrm{proj}_{\Delta_*} - \mathrm{proj}_{\Delta}\|_2 &= \sin\Theta_{\max}(\textsc{row}(\Delta_*), \textsc{row}(\Delta)) \\
&= \sin\Theta_{\max}\Big(\textsc{row}(\widetilde{U}_*^\top U_* \Sigma_*), \textsc{row}(A_L \widetilde{U}_\ell^\top U_k A_R)\Big)
\end{aligned}$$

since left multiplication by an invertible diagonal matrix does not change the row space:

$$\begin{aligned}
&= \sin\Theta_{\max}\Big(\textsc{row}(\widetilde{\Sigma}_*^{-1}\widetilde{U}_*^\top U_* \Sigma_*), \textsc{row}(\widetilde{\Sigma}_*^{-1}\widetilde{U}_\ell^\top U_k A_R)\Big) \\
&\le \frac{\|\widetilde{\Sigma}_*^{-1}\widetilde{U}_*^\top U_* \Sigma_* - \widetilde{\Sigma}_*^{-1}\widetilde{U}_\ell^\top U_k A_R\|_2}{\sigma_k(\widetilde{\Sigma}_*^{-1}\widetilde{U}_*^\top U_* \Sigma_*) - \sigma_{k+1}(\widetilde{\Sigma}_*^{-1}\widetilde{U}_*^\top U_* \Sigma_*)} && \text{(Wedin's Theorem)}
\end{aligned}$$

since $\mathrm{rank}(\widetilde{\Sigma}_*^{-1}\widetilde{U}_*^\top U_* \Sigma_*) = k$ by Assumption 4:

$$= \frac{\|\widetilde{\Sigma}_*^{-1}\widetilde{U}_*^\top U_* \Sigma_* - \widetilde{\Sigma}_*^{-1}\widetilde{U}_\ell^\top U_k A_R\|_2}{\sigma_k(\widetilde{\Sigma}_*^{-1}\widetilde{U}_*^\top U_* \Sigma_*)}$$

by Courant-Fischer $\sigma_k(\widetilde{\Sigma}_*^{-1}\widetilde{U}_*^\top U_* \Sigma_*) \ge \sigma_{\min}(\widetilde{\Sigma}_*^{-1})\sigma_k(S_*)\sigma_{\min}(\Sigma_*) = \sigma_k(S_*)\dfrac{\sigma_{\min}(\Sigma_*)}{\sigma_{\max}(\widetilde{\Sigma}_*)}$, then

$$\begin{aligned}
&\le \frac{\|\widetilde{\Sigma}_*^{-1}\widetilde{U}_*^\top U_* \Sigma_* - \widetilde{\Sigma}_*^{-1}\widetilde{U}_\ell^\top U_k A_R\|_2}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} \\
&\le \frac{\|\widetilde{\Sigma}_*^{-1}\|_2\Big(\|\widetilde{U}_*^\top U_* - \widetilde{U}_\ell^\top U_k\|_2\|\Sigma_*\|_2 + \|\widetilde{U}_\ell^\top U_k(\Sigma_* - A_R)\|_2\Big)}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} \\
&\le \frac{\|\widetilde{\Sigma}_*^{-1}\|_2\Big(\|\widetilde{U}_*^\top U_* - \widetilde{U}_\ell^\top U_k\|_2\|\Sigma_*\|_2 + \|\Sigma_* - A_R\|_2\Big)}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} \\
&\le \frac{\|\widetilde{\Sigma}_*^{-1}\|_2\Big(\Big\{\|\widetilde{U}_* - \widetilde{U}_\ell\|_2 + \|U_* - U_k\|_2\Big\}\|\Sigma_*\|_2 + \|\Sigma_* - A_R\|_2\Big)}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} \\
&\lesssim \frac{\|\widetilde{\Sigma}_*^{-1}\|_2\Big(\Big\{\frac{\|H_W\|_2}{\sigma_\ell(\widetilde{\Sigma}_*)} + \frac{\|H_X\|_2}{\sigma_k(\Sigma_*)}\Big\}\|\Sigma_*\|_2 + \|\Sigma_* - A_R\|_2\Big)}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} && \text{(Lemma A.2)} \\
&= \frac{\|\widetilde{\Sigma}_*^{-1}\|_2\|\Sigma_*\|_2\frac{\|H_W\|_2}{\sigma_\ell(\widetilde{\Sigma}_*)}}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} + \frac{\|\widetilde{\Sigma}_*^{-1}\|_2\|\Sigma_*\|_2\frac{\|H_X\|_2}{\sigma_k(\Sigma_*)}}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} + \frac{\|\widetilde{\Sigma}_*^{-1}\|_2\|\Sigma_* - A_R\|_2}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} \\
&= \frac{\kappa\Big(\Sigma_*, \widetilde{\Sigma}_*\Big)\frac{\|H_W\|_2}{\sigma_\ell(\widetilde{\Sigma}_*)}}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} + \frac{\kappa\Big(\Sigma_*, \widetilde{\Sigma}_*\Big)\frac{\|H_X\|_2}{\sigma_k(\Sigma_*)}}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} + \frac{\sigma_\ell(\widetilde{\Sigma}_*)^{-1}\|\Sigma_* - A_R\|_2}{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)^{-1}\sigma_k(S_*)} \\
&= \frac{\kappa\Big(\widetilde{\Sigma}_*\Big)\kappa(\Sigma_*)\|H_W\|_2}{\sigma_k(S_*)\sigma_\ell(\widetilde{\Sigma}_*)} + \frac{\kappa\Big(\widetilde{\Sigma}_*\Big)\kappa(\Sigma_*)\|H_X\|_2}{\sigma_k(S_*)\sigma_k(\Sigma_*)} + \frac{\kappa\Big(\widetilde{\Sigma}_*, \Sigma_*\Big)\|\Sigma_* - A_R\|_2}{\sigma_\ell(\widetilde{\Sigma}_*)\sigma_k(S_*)}.
\end{aligned}$$

□

Corollary A.1 (Theorem 1 dominates Theorem 2)**.** *The weighting bias from Theorem 1 dominates the bound in Theorem 2. In particular,*

$$\frac{\|I-A_L\|_2^2\sigma_{\max}(\Sigma_*)^2+\|A_L\|_2^2\|\Sigma_*-A_R\|_2^2}{c^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}\gtrsim\frac{\kappa\left(\widetilde{\Sigma}_*,\Sigma_*\right)^2\|\Sigma_*-A_R\|_2^2}{\sigma_\ell(\widetilde{\Sigma}_*)^2\sigma_k(S_*)^2}$$

and

$$\left[\frac{\|A_L\|_2\sigma_{\max}(\Sigma_*)}{c\sigma_{\min}(S)\sigma_{\min}(A_L)\sigma_{\min}(A_R)}\right]^2(\mathtt{NSR}_X+\mathtt{NSR}_W)\gtrsim\left[\frac{\kappa\left(\widetilde{\Sigma}_*\right)\kappa(\Sigma_*)}{\sigma_k(S_*)}\right]^2(\mathtt{NSR}_X+\mathtt{NSR}_W).$$

Proof. Observe that

$$\begin{aligned}
&\frac{\|I-A_L\|_2^2\sigma_{\max}(\Sigma_*)^2+\|A_L\|_2^2\|\Sigma_*-A_R\|_2^2}{c^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}-\frac{\kappa\left(\widetilde{\Sigma}_*,\Sigma_*\right)^2\|\Sigma_*-A_R\|_2^2}{\sigma_\ell(\widetilde{\Sigma}_*)^2\sigma_k(S_*)^2}\\
&\geq\frac{\|A_L\|_2^2\|\Sigma_*-A_R\|_2^2}{c^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}-\frac{\kappa\left(\widetilde{\Sigma}_*,\Sigma_*\right)^2\|\Sigma_*-A_R\|_2^2}{\sigma_\ell(\widetilde{\Sigma}_*)^2\sigma_k(S_*)^2}\\
&=\left[\frac{\kappa(A_L)^2}{c^2\sigma_{\min}(A_R)^2}-\frac{\kappa(\widetilde{\Sigma}_*)^2}{\sigma_{\min}(\Sigma_*)^2}\right]\frac{\|\Sigma_*-A_R\|_2^2}{\sigma_{\min}(S)^2} && \text{(Assumptions 3, 4)}\\
&\gtrsim 0. && \text{(Assumption 6)}
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\left[\frac{\|A_L\|_2\sigma_{\max}(\Sigma_*)}{c\sigma_{\min}(S)\sigma_{\min}(A_L)\sigma_{\min}(A_R)}\right]^2(\mathtt{NSR}_X+\mathtt{NSR}_W)-\left[\frac{\kappa\left(\widetilde{\Sigma}_*\right)\kappa(\Sigma_*)}{\sigma_k(S_*)}\right]^2(\mathtt{NSR}_X+\mathtt{NSR}_W)\\
&=\left[\frac{\kappa(A_L)^2}{c^2\sigma_{\min}(A_R)^2}-\frac{\kappa(\widetilde{\Sigma}_*)^2}{\sigma_{\min}(\Sigma_*)^2}\right]\frac{\sigma_{\max}(\Sigma_*)^2(\mathtt{NSR}_X+\mathtt{NSR}_W)}{\sigma_{\min}(S)^2} && \text{(Assumption 4)}\\
&\gtrsim 0. && \text{(Assumption 6)}
\end{aligned}$$

□

A.6 Comparison of estimators

Corollary A.2 (Comparing bias terms)**.** *Suppose the spectral regime satisfies Assumption 7. Then the weighting bias dominates the conditioning bias in Theorem 1.*

Corollary A.2. Denote the noise-driven part of the weighting bias as

$$(A_{\text{noise}})=\frac{\|A_L\|_2^2(\mathtt{NSR}_X+\mathtt{NSR}_W)\sigma_{\max}(\Sigma_*)^2}{c^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}.$$

The conditioning bias is

$$(B)=\mathtt{NSR}_X\frac{\kappa(S)^2\kappa(A_L)^2\kappa(A_R)^2}{c^2}.$$

Taking the ratio we have that

$$\frac{B}{A_{\text{noise}}} = \frac{\texttt{NSR}_X \frac{\sigma_{\max}(S)^2\sigma_{\max}(A_L)^2\sigma_{\max}(A_R)^2}{c^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}}{\frac{\sigma_{\max}(A_L)^2(\texttt{NSR}_X+\texttt{NSR}_W)\sigma_{\max}(\Sigma_*)^2}{c^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}} = \frac{\texttt{NSR}_X}{\texttt{NSR}_X+\texttt{NSR}_W} \cdot \frac{\sigma_{\max}(S)^2\sigma_{\max}(A_R)^2}{\sigma_{\max}(\Sigma_*)^2}.$$

Since S encodes the canonical correlations of the overlap $\widetilde{U}^\top U$, the condition for dominance of the weighting bias is:

$$\frac{\sigma_{\max}(A_R)}{\sigma_{\max}(\Sigma_*)} \leq \sqrt{1+\frac{\texttt{NSR}_W}{\texttt{NSR}_X}}.$$

□

Proof of Corollary 1. Given Corollary A.1, we only consider the weighting bias above. We write the bias and variance term as

$$(A) = \frac{\underbrace{\|I-A_L\|_2^2\sigma_{\max}(\Sigma_*)^2+\|A_L\|_2^2\|\Sigma_*-A_R\|_2^2}_{\text{weighting bias}}+\underbrace{\|A_L\|_2^2(\texttt{NSR}_X+\texttt{NSR}_W)\sigma_{\max}(\Sigma_*)^2}_{\text{measurement noise}}}{c^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}$$

$$(C) = \frac{\bar{\sigma}^2 r}{c^2\sigma_{\min}(S)^2\sigma_{\min}(A_L)^2\sigma_{\min}(A_R)^2}$$

with $r = \text{rank}(\Delta)$. Then the variance-dominated region is given by

$$\underbrace{\|I-A_L\|_2^2\sigma_{\max}(\Sigma_*)^2+\|A_L\|_2^2\|\Sigma_*-A_R\|_2^2}_{\text{weighting bias}}+\underbrace{\|A_L\|_2^2(\texttt{NSR}_X+\texttt{NSR}_W)\sigma_{\max}(\Sigma_*)^2}_{\text{measurement noise}} \lesssim \bar{\sigma}^2 r$$

and we simplify to,

$$\texttt{NSR} := \texttt{NSR}_X+\texttt{NSR}_W \lesssim \frac{\bar{\sigma}^2 r}{\kappa(X,W)^2} - \frac{\text{weighting bias}}{\kappa(X,W)^2}.$$

Note that we are comparing with the W-weighting factor since the CCA estimator enters later in the variance-dominated regime, and we can suppose, without loss of generality, $\sigma_{\min}(W)^{-1} \geq \|A_L\|_2$.

Similarly, the following condition is sufficient for a bias-dominated region:

$$\texttt{NSR} := \texttt{NSR}_X+\texttt{NSR}_W \gtrsim \frac{\bar{\sigma}^2 r}{\kappa(X,W)^2}.$$

□

Proof of Corollary 2. In the variance-dominated region the upper bound ratios between the estimators are:

$$\frac{\text{CCA}_{\sqrt{UB}}}{\text{Whitening}_{\sqrt{UB}}} = \sigma_{\max}(W), \quad \frac{\text{CCA}_{\sqrt{UB}}}{\text{PCA}_{\sqrt{UB}}} = \sigma_{\max}(W)\sigma_{\min}(X), \quad \frac{\text{Whitening}_{\sqrt{UB}}}{\text{PCA}_{\sqrt{UB}}} = \sigma_{\min}(X).$$

Hence, CCA is better whenever

$$\sigma_{\max}(W) \leq 1 \quad \text{and} \quad \sigma_{\max}(W)\sigma_{\min}(X) \leq 1.$$

PCA is better when

$$1 \leq \sigma_{\max}(W)\sigma_{\min}(X) \quad \text{and} \quad 1 \leq \sigma_{\min}(X).$$

Finally, the whitening is better when

$$1 \leq \sigma_{\max}(W) \quad \text{and} \quad \sigma_{\min}(X) \leq 1.$$

On the other hand, under Assumption 8 ignoring the weighting bias, the ratio for comparison in the bias-dominated region is

$$\frac{\text{CCA}_{\sqrt{UB}}}{\text{Whitening}_{\sqrt{UB}}} = \kappa(W), \quad \frac{\text{CCA}_{\sqrt{UB}}}{\text{PCA}_{\sqrt{UB}}} = \kappa(W)\sigma_{\min}(X), \quad \frac{\text{Whitening}_{\sqrt{UB}}}{\text{PCA}_{\sqrt{UB}}} = \sigma_{\min}(X)$$

and we get a similar partition from above.

The offset in the phase diagram (gray region) is due to the error term from the weighting bias. □

A.7 Lower bound

We include the statement of Fano's inequality for completeness

Lemma A.6 (Fano's inequality)**.** *Suppose there exists a finite subset $\mathcal{M}$ of the full parameter space $\{\beta_1,...\beta_{|\mathcal{M}|}\}$ such that*

1. ***Well separation:*** *$\|\beta_i - \beta_j\|_2^2 \geq d^2$ for all $i \neq j$*

2. ***Statistical indistinguishability:*** *For any β_i and β_j, there exist corresponding probability distributions P_i,P_j over the observed data that have bounded KL divergence $D_{KL}(P_i \| P_j) \leq \alpha$.*

Then

$$\inf_{\widehat{\beta}} \sup_{\beta} \mathbb{E}\left[\|\widehat{\beta} - \beta\|_2^2\right] \geq \frac{d^2}{2}\left(1 - \frac{\alpha + \log 2}{\log|\mathcal{M}|}\right).$$

Below is the proof of the minimax lower bound.

Assumption A.2 (Simplifying rank condition)**.** *In addition to Assumption 3, suppose that* $8 \leq \text{rank}(\Delta_*) := r$.

Proof of Theorem 3. We proceed in steps.

1. **Data generating process.** For a single, fixed set of nonrandom matrices and parameter (X,W,β), the probability distribution for the observed data $(Y,\text{vec}(Z_X),\text{vec}(Z_W)) \sim P$ is characterized by the independent marginals

$$\begin{aligned} Y &= X\beta^* + \epsilon^* \sim N(X\beta^*, \sigma_\varepsilon^2) := P_Y \\ \text{vec}(Z_X) &= \text{vec}(X) + \text{vec}(H_X) \sim N(\text{vec}(X), \sigma_X^2 I) := P_{Z_X} \\ \text{vec}(Z_W) &= \text{vec}(W) + \text{vec}(H_W) \sim N(\text{vec}(W), \sigma_W^2 I) := P_{Z_W}. \end{aligned}$$

Then for any two set of fixed matrices and parameter values (X,W,β), (X',W',β'), we have

$$\begin{aligned} D_{KL}(P\|P') &= D_{KL}(P_Y \otimes P_{Z_X} \otimes P_{Z_W} \| P'_Y \otimes P'_{Z_X} \otimes P'_{Z_W}) \\ &= D_{KL}(P_Y\|P'_Y) + D_{KL}(P_{Z_X}\|P'_{Z_X}) + D_{KL}(P_{Z_W}\|P'_{Z_W}) \end{aligned}$$

since for independent product measures KL divergence is additive.

Now we use the KL expression between Gaussians:

$$\begin{aligned} D_{\text{KL}}(\mathcal{N}(\mu,\Sigma)\|\mathcal{N}(\mu',\Sigma)) &= \frac{1}{2}(\mu-\mu')^\top \Sigma^{-1}(\mu-\mu') \\ &= \frac{1}{2\sigma_\varepsilon^2}\|X\beta - X'\beta'\|_2^2 + \frac{1}{2\sigma_X^2}\|\text{vec}(X) - \text{vec}(X')\|_2^2 + \frac{1}{2\sigma_W^2}\|\text{vec}(W) - \text{vec}(W')\|_2^2 \\ &= \frac{1}{2\sigma_\varepsilon^2}\|X\beta - X'\beta'\|_2^2 + \frac{1}{2\sigma_X^2}\|X - X'\|_{\text{Fr}}^2 + \frac{1}{2\sigma_W^2}\|W - W'\|_{\text{Fr}}^2. \end{aligned}$$

2. **Construction of the packing.** The previous result shows that the KL divergence is pinned down by (X,W,β). Therefore, we will define a sequence of $\left(X^{(\omega)},W^{(\omega)},\beta^{(\omega)}\right)$, indexed by (ω), that ultimately defines the sequence of DGPs $(Y^{(\omega)},Z_X^{(\omega)},Z_W^{(\omega)})$.[1]

We start by choosing any two matrices $X^{(0)}, W^{(0)}$ such that $X^{(0)} \in \text{Col}(W^{(0)})$. Let $W^{(0)\top}X^{(0)} = Q_* S_* R_*^\top$, and $\sigma_{\min}(W^{(0)}) = \sigma_{\min}(\widetilde{\Sigma}_*)$. Denote the r-dimensional canonical directions as $\{e_j\}_{j=1}^r$.

By Tsybakov (2009, Lemma 2.9), there exists a set $\mathcal{M} \subseteq \{0,1\}^r$ such that (for $r \geq 8$):

(a) $\omega_0 = (0,...,0) \in \mathcal{M}$,

(b) $\|\omega - \omega'\|_0 \geq \frac{r}{8}$, and

(c) $|\mathcal{M}| \geq 2^{\frac{r}{8}}$.

This is known in the literature as the Gilbert-Varshamov (GV) pruned hypercube. The construction of the DGPs will be indexed by the elements of the GV hypercube. For each

[1] We drop the vectorized notation for simplicity.

$\omega=(\omega_1,...,\omega_r)\in\mathcal{M}$ define

$$\beta^{(\omega)}=h\sum_{j=1}^{r}\frac{\omega_j}{\sigma_j(S_*)}R_*e_j$$

where $\sigma_{\max}(S_*)\geq...\geq\sigma_r(S_*)>0$ are the cosine of principal angles between $W^{(0)}$ and $X^{(0)}$. Here, $h>0$ is an appropriate scaling to be chosen later.

We maintain that the instrumental variables are fixed: $W^{(\omega)}=W^{(0)}$. We now construct $X^{(\omega)}$ as the minimizer of $D_{KL}\big(P^{(\omega)}\|P^0\big)$ subject to the restriction that $X^{(\omega)}\in\text{Col}(W^{(0)})$, where P^0 is the distribution of $(Y^{(0)},Z_X^{(0)},Z_W^{(0)})\sim P^{(0)}$, and similarly for $P^{(\omega)}$. In particular,

$$\begin{aligned}
X^{(\omega)}&=\underset{X\in Col(W^{(0)})}{argmin}\ D_{KL}\big(P^{(\omega)}\|P^0\big)\\
&=\underset{X\in Col(W^{(0)})}{argmin}\ \frac{1}{2\sigma_\varepsilon^2}\big\|X\beta^{(\omega)}\big\|_2^2+\frac{1}{2\sigma_X^2}\big\|X-X^{(0)}\big\|_{\text{Fr}}^2\\
&=\underset{X\in Col(W^{(0)})}{argmin}\ \frac{1}{2\sigma_\varepsilon^2}\big\|\text{proj}_{W^{(0)}}X\beta^{(\omega)}+(I-\text{proj}_{W^{(0)}})X\beta^{(\omega)}\big\|_2^2+\frac{1}{2\sigma_X^2}\big\|X-X^{(0)}\big\|_{\text{Fr}}^2\\
&=\underset{X\in Col(W^{(0)})}{argmin}\ \frac{1}{2\sigma_\varepsilon^2}\big\|\text{proj}_{W^{(0)}}X\beta^{(\omega)}\big\|_2^2+\frac{1}{2\sigma_X^2}\big\|X-X^{(0)}\big\|_{\text{Fr}}^2
\end{aligned}$$

where $\text{proj}_{W^{(0)}}=W^{(0)}(W^{(0)\top}W^{(0)})^\dagger W^{(0)\top}$. Using Petersen and Pedersen (2008, Equation 82), the first order condition can be expressed as

$$X^{(\omega)}-X^{(0)}=-\frac{\sigma_X^2}{\sigma_\varepsilon^2}\big(\text{proj}_{W^{(0)}}X^{(\omega)}\beta^{(\omega)}\big)\beta^{(\omega)\top}. \tag{3}$$

Instead of solving for the entire matrix $X^{(\omega)}$, we solve for the resulting projected signal vector, $S^{(\omega)}=\text{proj}_{W^{(0)}}X^{(\omega)}\beta^{(\omega)}$, defining it implicitly as

$$X^{(\omega)}=X^{(0)}-\frac{\sigma_X^2}{\sigma_\varepsilon^2}\big(\text{proj}_{W^{(0)}}X^{(\omega)}\beta^{(\omega)}\big)\beta^{(\omega)\top}=X^{(0)}-\frac{\sigma_X^2}{\sigma_\varepsilon^2}S^{(\omega)}\beta^{(\omega)\top}.$$

Left-multiply the entire equation by the projection matrix $\text{proj}_{W^{(0)}}$ and right-multiply by $\beta^{(\omega)}$:

$$\text{proj}_{W^{(0)}}X^{(\omega)}\beta^{(\omega)}=\text{proj}_{W^{(0)}}X^{(0)}\beta^{(\omega)}-\text{proj}_{W^{(0)}}\left(\frac{\sigma_X^2}{\sigma_\varepsilon^2}S^{(\omega)}\beta^{(\omega)\top}\right)\beta^{(\omega)}.$$

Since $\text{proj}_{W^{(0)}}$ is a projection we simplify as

$$S^{(\omega)}=\text{proj}_{W^{(0)}}X^{(0)}\beta^{(\omega)}-\frac{\sigma_X^2}{\sigma_\varepsilon^2}\big(\text{proj}_{W^{(0)}}S^{(\omega)}\big)\big(\beta^{(\omega)\top}\beta^{(\omega)}\big)=\text{proj}_{W^{(0)}}X^{(0)}\beta^{(\omega)}-\frac{\sigma_X^2}{\sigma_\varepsilon^2}\|\beta^{(\omega)}\|_2^2S^{(\omega)}.$$

Therefore, $X^{(\omega)}$ is such that it lies in the subspace defined by $W^{(0)}$ and such that

$$S^{(\omega)}=\left(1+\frac{\sigma_X^2}{\sigma_\varepsilon^2}\|\beta^{(\omega)}\|_2^2\right)^{-1}\text{proj}_{W^{(0)}}X^{(0)}\beta^{(\omega)}.$$

Our packing is then characterized by $\big(X^{(\omega)},W^{(0)},\beta^{(\omega)}\big)_{\omega\in\mathcal{M}}$.

3. **Well-separation of the parameter.** Note that for any two distinct parameters in

the packing,

$$\|\beta^{(\omega)}-\beta^{(\omega')}\|_2^2 = h^2\left\|\sum_{j=1}^{r}\frac{\omega_j-\omega'_j}{\sigma_j(S_*)}R_*e_j\right\|_2^2 \overset{\text{Parseval}}{=} h^2\sum_{j=1}^{r}\frac{(\omega_j-\omega'_j)^2}{\sigma_j(S_*)^2} = h^2\sum_{j:\omega_j\neq\omega'_j}\frac{1}{\sigma_j(S_*)^2}$$
$$\geq \frac{h^2}{\sigma_{\max}(S_*)^2}\|\omega-\omega'\|_0 \geq \frac{h^2 r}{8\sigma_{\max}(S_*)^2} := d^2$$

where we are denoting the last quantity as the separation gap d^2.

4. **Bounding the KL divergence.** We have that

$$\begin{aligned}
D_{KL}\big(P^{(\omega)}\|P^0\big) &= \frac{1}{2\sigma_\varepsilon^2}\big\|\text{proj}_{W^{(0)}}X^{(\omega)}\beta^{(\omega)}\big\|_2^2 + \frac{1}{2\sigma_X^2}\big\|X^{(\omega)}-X^{(0)}\big\|_{\text{Fr}}^2 \\
&= \frac{1}{2\sigma_\varepsilon^2}\big\|S^{(\omega)}\big\|_2^2 + \frac{1}{2\sigma_X^2}\big\|X^{(\omega)}-X^{(0)}\big\|_{\text{Fr}}^2 && \text{(since } S^{(\omega)}=\text{proj}_{W^{(0)}}X^{(\omega)}\beta^{(\omega)}\text{)} \\
&= \frac{1}{2\sigma_\varepsilon^2}\big\|S^{(\omega)}\big\|_2^2 + \frac{1}{2\sigma_X^2}\left[\left(\frac{\sigma_X^2}{\sigma_\varepsilon^2}\right)^2\big\|S^{(\omega)}\beta^{(\omega)\top}\big\|_{\text{Fr}}^2\right] && \text{(Equation 3)} \\
&= \frac{1}{2\sigma_\varepsilon^2}\big\|S^{(\omega)}\big\|_2^2 + \frac{1}{2\sigma_X^2}\left[\left(\frac{\sigma_X^2}{\sigma_\varepsilon^2}\right)^2\big\|S^{(\omega)}\big\|_2^2\big\|\beta^{(\omega)}\big\|_2^2\right] && \text{(Frobenius norm of a rank-1 matrix)} \\
&= \frac{1}{2\sigma_\varepsilon^2}\big\|S^{(\omega)}\big\|_2^2\left(1+\frac{\sigma_X^2}{\sigma_\varepsilon^2}\|\beta^{(\omega)}\|_2^2\right).
\end{aligned}$$

Now substituting back $S^{(\omega)} = \left(1+\frac{\sigma_X^2}{\sigma_\varepsilon^2}\|\beta^{(\omega)}\|_2^2\right)^{-1}\text{proj}_{W^{(0)}}X^{(0)}\beta^{(\omega)}$ yields

$$D_{KL}\big(P^{(\omega)}\|P^0\big) = \frac{1}{2}\cdot\frac{\big\|\text{proj}_{W^{(0)}}X^{(0)}\beta^{(\omega)}\big\|_2^2}{\sigma_\varepsilon^2+\sigma_X^2\|\beta^{(\omega)}\|_2^2}.$$

Note that

$$W^{(0)\top}X^{(0)}\beta^{(\omega)} = W^{(0)\top}X^{(0)}\left(h\sum_{j=1}^{r}\frac{\omega_j}{\sigma_j(S_*)}R_*e_j\right) = h\sum_{j=1}^{r}\frac{\omega_j}{\sigma_j(S_*)}W^{(0)\top}X^{(0)}R_*e_j$$
$$= h\sum_{j=1}^{r}\frac{\omega_j}{\sigma_j(S_*)}\sigma_j(S_*)Q_*e_j = h\sum_{j=1}^{r}\omega_j Q_*e_j.$$

Therefore,

$$\big\|\text{proj}_{W^{(0)}}X^{(0)}\beta^{(\omega)}\big\|_2^2 = \big\|W^{(0)}(W^{(0)\top}W^{(0)})^\dagger W^{(0)\top}X^{(0)}\beta^{(\omega)}\big\|_2^2 \leq \big\|W^{(0)}(W^{(0)\top}W^{(0)})^\dagger\big\|_2^2\left\|h\sum_{j=1}^{r}\omega_j q_{*j}\right\|_2^2$$
$$\overset{\text{Parseval}}{=} \frac{1}{\sigma_{\min}(\widetilde{\Sigma}_*)^2}h^2\|\omega\|_2^2 = \frac{1}{\sigma_{\min}(\widetilde{\Sigma}_*)^2}h^2\|\omega\|_0 \leq \frac{h^2 r}{\sigma_{\min}(\widetilde{\Sigma}_*)^2}.$$

Consolidating, together with $\big\|\beta^{(\omega)}\big\|_2^2 \geq \frac{h^2 r}{8\sigma_{\max}(S_*)^2}$, we obtain an upper bound for the KL divergence between any two elements in the packing

$$D_{KL}\Big(P^{(\omega)}\|P^{(\omega')}\Big) \leq 2D_{KL}\big(P^{(\omega)}\|P^0\big) + 2D_{KL}\Big(P^0\|P^{(\omega')}\Big) \leq 2\cdot\frac{\frac{h^2 r}{\sigma_{\min}(\widetilde{\Sigma}_*)^2}}{\sigma_\varepsilon^2+\sigma_X^2\left(\frac{h^2 r}{8\sigma_{\max}(S_*)^2}\right)} := \alpha.$$

As an aside, observe that the KL divergence does not satisfy the triangle inequality. However, restricting to normal distributions with same covariance, the square-root of the KL divergence is indeed a metric. For the family $N(\mu,\Sigma)$ with fixed Σ, one has

$$D_{KL}(N(\mu_1,\Sigma)\|N(\mu_2,\Sigma))=\frac{1}{2}\|\mu_1-\mu_2\|^2_{\Sigma^{-1}},$$

so $\sqrt{2D_{KL}}=\|\mu_1-\mu_2\|_{\Sigma^{-1}}$ is a (Mahalanobis) distance and satisfies the triangle inequality.

5. **Application of Fano's Inequality (Lemma A.6).** In order to obtain a meaningful bound we require that

$$\frac{r}{8}\log 2=\log|\mathcal{M}|\gtrsim\alpha:=2\frac{\frac{h^2r}{\sigma_{\min}(\widetilde{\Sigma}_*)^2}}{\sigma_\varepsilon^2+\sigma_X^2\left(\frac{h^2r}{8\sigma_{\max}(S_*)^2}\right)}.$$

We choose the scaling h such that it satisfies the inequality above. Note that such h satisfies

$$h^2\asymp\max\left\{\frac{\sigma_{\max}(S_*)^2\sigma_\varepsilon^2\sigma_{\min}(\widetilde{\Sigma}_*)^2}{\sigma_{\max}(S_*)^2-\frac{r\sigma_X^2\sigma_{\min}(\widetilde{\Sigma}_*)^2}{8}},\sigma_\varepsilon^2\sigma_{\min}(\widetilde{\Sigma}_*)^2\right\}.$$

Therefore we can conclude that

$$\begin{aligned}\inf_{\widehat{\beta}}\sup_{\beta}\mathbb{E}\left[\|\widehat{\beta}-\beta\|_2^2\right]&\geq\frac{d^2}{2}\left(1-\frac{\alpha+\log 2}{\log|\mathcal{M}|}\right)\\&\gtrsim\max\left\{\frac{\sigma_{\max}(S_*)^2\sigma_\varepsilon^2\sigma_{\min}(\widetilde{\Sigma}_*)^2}{\sigma_{\max}(S_*)^2-\frac{r\sigma_X^2\sigma_{\min}(\widetilde{\Sigma}_*)^2}{8}},\sigma_\varepsilon^2\sigma_{\min}(\widetilde{\Sigma}_*)^2\right\}\frac{r}{\sigma_{\max}(S_*)^2}\\&=\max\left\{\frac{\sigma_\varepsilon^2 r\sigma_{\min}(\widetilde{\Sigma}_*)^2}{\sigma_{\max}(S_*)^2-\frac{r\sigma_X^2\sigma_{\min}(\widetilde{\Sigma}_*)^2}{8}},\frac{\sigma_\varepsilon^2 r\sigma_{\min}(\widetilde{\Sigma}_*)^2}{\sigma_{\max}(S_*)^2}\right\}.\end{aligned}$$

□

Remark A.2 (Relaxing Assumption A.2)**.** The constraint $r>8$ (Assumption A.2) is just a convenience tied to the Gilbert-Varshamov packing; it is not fundamental. We can remove it by a more careful construction of the GV hypercube:

1. $\omega_0=(0,...,0)\in\mathcal{M}$,
2. $\|\omega-\omega'\|_0\geq\lceil\frac{r}{8}\rceil$, and
3. $\log|\mathcal{M}|\geq cr$ for a universal constant $c>0$.

B Simulation details

B.1 Concentrated signal

For each replication we generate a rank-k latent regressor matrix

$$X = U_X \Sigma_X V_X^\top \in \mathbb{R}^{n \times p}, \qquad \Sigma_X = \mathrm{diag}((i+1)^{-\alpha})_{i=1}^{k},$$

where $U_X \in \mathbb{R}^{n \times k}$ and $V_X \in \mathbb{R}^{p \times k}$ have orthonormal columns (drawn via QR of i.i.d. Gaussian matrices). We generate a rank-ℓ latent instrument matrix

$$W = U_W \Sigma_W V_W^\top \in \mathbb{R}^{n \times p_w}, \qquad \Sigma_W = \mathrm{diag}((i+1)^{-\alpha})_{i=1}^{\ell},$$

with $V_W \in \mathbb{R}^{p_w \times \ell}$ orthonormal and with U_W constructed to have controlled overlap with the signal subspace of X. Specifically, letting $r = \min\{k, \ell\}$, we form the first r columns of U_W as

$$(U_W)_{:,1:r} = \delta (U_X)_{:,1:r} + \sqrt{1-\delta^2}\,(U_\perp)_{:,1:r},$$

where $U_\perp$ spans an orthogonal complement of $\mathrm{span}(U_X)$, and then orthonormalize. The parameter $\delta \in [0,1]$ controls the strength of instrument alignment: larger δ yields larger canonical correlations between the signal subspaces of X and W.

B.2 Diffuse noise

We observe noisy measurements

$$Z_X = X + H_X, \qquad Z_W = W + H_W,$$

where (H_X, H_W) are Gaussian measurement errors with a shared component. Concretely, on the first $p_{\text{common}} = \min\{p, p_w\}$ columns we draw $(H_{Wi,j}, H_{Xi,j})$ jointly normal with correlation ρ and marginal standard deviations $\sigma_{H_W}, \sigma_{H_X}$, and draw the remaining columns independently. We scale

$$\sigma_{H_X} = \frac{c_1}{(k+1)^\alpha \sqrt{p}}, \qquad \sigma_{H_W} = \frac{c_1}{(\ell+1)^\alpha \sqrt{p_w}},$$

so that the magnitude of corruption is comparable across dimensional regimes.

B.3 Instrument validity

We generate a disturbance that is correlated with X but orthogonal to the instrument space:

$$\varepsilon_0 = X\gamma + \eta, \qquad \varepsilon = (I - \mathrm{proj}_W)\varepsilon_0, \qquad Y = X\beta^* + \varepsilon,$$

where proj_W is the orthogonal projector onto $\text{span}(U_W)$, $\eta \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_n)$, and $\gamma \in \mathbb{R}^p$ is fixed within each $(n, \delta, \text{regime})$ block. By construction $\mathbb{E}[\varepsilon \mid W] = 0$ while allowing endogeneity $\mathbb{E}[\varepsilon \mid X] \neq 0$. We rescale ε to have variance σ_ε^2 for comparability across designs.

B.4 Dimension

We consider two regimes. In the moderate-dimensional regime we set $p = \lfloor n/2 \rfloor$ and $p_w = \lfloor n/3 \rfloor$. In the high-dimensional regime we set $p = n - 100$ and $p_w = n - 200$ (capped at 5000). We fix $(k, \ell) = (8, 10)$, $\alpha = 1.5$, $\rho = 0.9$, $\sigma_\varepsilon = 1.25$, and $c_1 = 2.0$. Instrument alignment is varied over $\delta \in \{0.001, 0.05, 0.65\}$, and we report results across $n \in \{300, 500, 1000, 2000, 5000\}$.