Overfitting Inference with Instrumental Variables

Ruei-Chi Lee*

October 1, 2026

[Most recent version]

Abstract

Overfitting is traditionally viewed as a problem to be mitigated in empirical economics. This paper shows that it can instead be a blessing for inference in high-dimensional instrumental variables (IV) models. We revisit the factor IV framework, in which the instruments and the endogenous variable share common latent factors, and show that the first-stage signal is dense, providing a theoretical justification for recent empirical evidence against sparsity. Moreover, the implied first-stage coefficients shrink toward zero as the number of instruments diverges, exhibiting "local-to-zero" behavior commonly used to describe settings with many weak instruments. Building on this characterization, we propose an IV estimator based on ridgeless regression combined with sample splitting. This procedure remains well defined when the number of first-stage regressors exceeds the sample size. We derive the asymptotic distribution of the estimator under regimes with (very) many weak instruments, and show that, when the number of instruments is sufficiently large, its asymptotic variance coincides with that of an oracle factor IV estimator that directly observes the latent factors. Monte Carlo simulations indicate favorable finite-sample performance relative to existing high-dimensional IV estimators. Empirically, we apply the proposed procedure to estimating the elasticity of substitution between immigrants and natives in the United States.

*Department of Economics, Rutgers University; email: rl824@rutgers.edu

1 Introduction

Overfitting has long been a central concern in empirical economics. This concern underlies the familiar bias–variance trade-off: greater model complexity can reduce bias, while increased overfitting leads to additional variance, so improvements in fit need not translate into more accurate predictions or estimates. This overfitting arises across many areas of empirical economics, such as forecasting (Bańbura et al. (2010)), program evaluation (Cattaneo et al. (2018); Abadie (2021)), instrumental variables (IV) regression (Bekker (1994); Bound et al. (1995)). It has motivated a longstanding emphasis on controlling model complexity.

To mitigate overfitting, empirical researchers commonly rely on variable selection, dimension reduction, or regularization. Variable selection procedures are valid only under (approximate) sparsity assumption, that is, when the relationship of interest is captured by a parsimonious subset of the available variables. Under this assumption, Lasso-based procedures select controls or instruments for estimating causal effects (Belloni et al. (2012); Belloni et al. (2014)). Dimension reduction procedures such as principal component analysis (PCA) parsimoniously summarize large sets of variables with a finite number of latent factors, which can serve as predictors in macroeconomic forecasting or as instruments in IV regression (Stock and Watson (2002); Bai and Ng (2010)). In synthetic control, restricting donor weights to be nonnegative and to sum to one provides a form of regularization, which yields a sparse combination of donor units (Abadie et al. (2010); Abadie (2021)).

Recent research in statistics and machine learning, however, has challenged the view that extremely complex models that fit the data perfectly necessarily lead to poor prediction. Belkin et al. (2019) describe a benign overfitting/double descent phenomenon: as model complexity exceeds the sample size and continues increasing, it leads to a gradual reduction in variance and a second descent of the prediction risk occurs in the extremely overparametrized regime. Evidence of benefits from highly complex models has also emerged in the economic. In financial forecasting, Kelly et al. (2024) expand standard return predictors into thousands of features, far more than the number of observations, and show that out-of-sample market-timing performance increases with model complexity. In program evaluation, Spiess et al. (2023) show that regression models with more covariates than observations can outperform simple ones, and synthetic control keeps improving counterfactual predictions as donor units are added, even beyond the point of perfect pretreatment fit.

Why should an applied economist need overfitting, instead of mitigating it? Because economic data sets are rarely sparse. Giannone et al. (2021) present empirical evidence that sparsity is often not a realistic assumption across widely used data sets in macroeconomics, microeconomics, and finance. Kolesár et al. (2026) documented that the sparsity assumption is fragile in empirical studies and rejected it in three empirical applications. As they observe,

it is rarely possible to give a theoretical or institutional reason why a small number of controls should absorb most of the confounding. In IV settings, Angrist and Frandsen (2022) compared IV estimators that use Post-Lasso in the first stage with OLS and 2SLS in two empirical applications. The IV estimators using Post-Lasso generated biases of similar magnitude to those from OLS and 2SLS. These evidence suggest that imposing sparsity is often a poor approximation in empirical work.

While overfitting has been studied extensively as a prediction problem, overfitting inference is much less developed. In particular, although inference procedures for IV models with many instruments are well established (e.g., Hansen et al. (2008); Chao et al. (2012); Hausman et al. (2012)), much less is known for more complex models. In this paper, we study overfitting inference in high-dimensional IV models in which the number of instruments is large relative to the sample size, or even exceeds it. To model how the first-stage signal can be dense rather than sparse, we revisit the factor IV framework studied by Bai and Ng (2010) and Kapetanios and Marcellino (2010), in which the instruments and the endogenous variable share common latent factors. Instead of using PCA to estimate latent factors as instruments, we use the observed instruments directly as first-stage regressors. Our first contribution is to provide theoretical justification for recent empirical evidence that the first-stage signal is dense. That is, the underlying factor structure implies that the first-stage signal is distributed across many instruments rather than concentrated on a few.

We further show that the effective strength of the instruments declines as the underlying factors become stronger, which may appear counterintuitive at first. The reason is that, when many observed instruments reflect the same latent factors, the first-stage signal is spread across those instruments, so each one needs only a small coefficient to help recover the common factor structure. Because the coefficients are diluted among an increasing number of instruments, the instruments exhibit a local-to-zero behavior: as the number of instruments diverges, the corresponding first-stage coefficients shrink toward zero. Thus, our first-stage regression falls into the scenario of "many weak instruments", in the sense that there are many instruments, but each is only weakly correlated with the endogenous variable.

In the many weak IV literature, instrument strength is usually measured by the concentration parameter, μ^2, signal-to-noise ratio of the first-stage regression, and the number of instruments, p. The definition of weak instrument from Mikusheva and Sun (2021) is a regime where the concentration parameter divided by the square root of the number of instruments stays bounded in large samples, i.e., $\mu^2/\sqrt{p} \xrightarrow{p} C < \infty$. If the identification strength is insufficiently large relative to the instrument dimension, the model suffers from weak instrument problems; in fact, Mikusheva and Sun (2021) establish that when $\mu^2/\sqrt{p}$ is bounded, many weak IV estimators are inconsistent, including the jackknife IV estimator (JIVE) and jackknife versions of LIML and Fuller estimators (e.g., Chao et al.

(2012); Hausman et al. (2012)). In the high-dimensional IV setting where the number of instruments exceeds the sample size, our proposed estimator remains consistent even when $\mu^2/\sqrt{p}$ is bounded. That is, even when a large number of instruments collectively provide only limited information about the endogenous variable, our estimator remains consistent.

Given that the first-stage regression is inherently non-sparse, we develop an estimation and inference procedure that remains valid even when the number of first-stage regressors exceeds the sample size. Our key observation is that, once we include a sufficiently large set of observed instruments, the IV estimator that uses these observed instruments directly in the first stage can attain the same oracle efficiency as if the latent factors were observed and used as instruments. This result arises from what we refer to as a blessing of overfitting in the first stage.

To implement this idea, we estimate the first stage using Ridgeless regression, which is the minimum norm least squares solution obtained by using the Moore–Penrose pseudoinverse in place of the usual matrix inverse. This estimator is well defined regardless of the dimensionality of the first-stage regression. An important finding from recent studies of ridgeless regression is that once the number of regressors exceeds the sample size, variance can decline as additional regressors are included, which is the embodiment of double descent in the machine learning literature (e.g., Bartlett et al. (2020); Hastie et al. (2022)). Consequently, with a large enough number of instruments, the variance of our IV estimator is effectively diversified.

It is well-known that extreme overfitting in the first stage generally induces bias in the second-stage IV estimation. To address this, we adopt a sample-splitting strategy: we estimate the first-stage predictors for the endogenous variable on one subsample, and on the other subsample we estimate the structural parameter of interest using the predicted endogenous variable from the first stage. Combining sample splitting with Ridgeless regression allows us to sufficiently avoid bias while still extracting sufficient signal in the first stage. This delivers a consistent, asymptotically normal estimator of the structural parameter of interest.

In practice, we may fall short of the ideal situation in which we have enough instruments for our estimator to attain oracle efficiency. We provide practical guidance. First, we allow many exogenous covariates, since the presence of a large number of instruments may come from the interaction of basic instruments with many exogenous controls. For example, Angrist and Krueger (1991) used the quarter of birth as basic instruments, interacting the quarter of birth with the state of birth and year of birth to generate 180 instruments. A fully interacted specification can generate 1,530 instruments. Second, we suggest K-fold sample-splitting procedure in practice, to intentionally reduce the effective sample size. By splitting the sample, it becomes more plausible to turn an underparametrized first stage into an overparametrized one, enabling the double descent phenomenon to arise for the

sample-splitting estimator.

Our approach is closely related to Hansen and Kozbur (2014), who propose a ridge-regularized jackknife instrumental variables estimator (RJIVE) that allows the number of instruments to exceed the sample size and the first-stage relationship to be non-sparse. By coupling the jackknife with ridge regularization, RJIVE avoids extreme overfitting: the ridge penalty stabilizes the projection matrix and prevents the interpolation threshold from driving the variance to infinity. The ridge regularization nonetheless precludes the double descent phenomenon, so that RJIVE's variance does not decrease as the number of instruments increases. By contrast, our proposed estimator exploits the blessing of overfitting, attaining not only consistency when $\mu^2/\sqrt{p}$ is bounded but also oracle efficiency when the number of instruments is sufficiently large. Relatedly, while Carrasco and Tchuente (2015) and Carrasco and Doukali (2017) propose regularized LIML and JIVE estimators that also allow the number of instruments to exceed the sample size by using alternative regularization techniques, they do not consider weak instruments defined in Mikusheva and Sun (2021).

This paper connects to the burgeoning statistical and machine learning literature on the theoretical properties of extremely overparametrized models. Recent studies on high-dimensional Ridgeless regression in linear prediction problems establish the double descent phenomenon in prediction risk (Belkin et al. (2020); Bartlett et al. (2020); Mei and Montanari (2022); Hastie et al. (2022)). More closely related to our framework is Liao et al. (2023), which studies economics prediction risk in non-sparse regression models driven by underlying latent factors. In that setting, regressors are highly correlated, yet the prediction risk still diversifies away with a sufficient number of regressors.

The rest of the paper is organized as follows. Section 2 describes the model and introduces the estimation method for the structural parameter. Section 3 provides the asymptotic analysis results of the estimator. Section 4 provides practical guidance on constructing an instrument set that is sufficiently large relative to the sample size. In Section 5 we report the simulation results to evaluate the finite sample performance. Section 6 presents an empirical application that estimates the inverse elasticity of substitution between immigrants and natives in the United States. Section 7 concludes. The Appendix contains the proofs of the theoretical results.

We adopt the following notation. Let $\|A\|$ denote the ℓ_2 norm if A is a vector, and the operator norm if A is a matrix. Let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ denote the minimum and maximum nonzero singular values of a matrix A, respectively. For two sequences $a_{n,p}$ and $b_{n,p}$, we denote $a_{n,p} \ll b_{n,p}$ (or $b_{n,p} \gg a_{n,p}$) if $a_{n,p} = o(b_{n,p})$, and $a_{n,p} \asymp b_{n,p}$ if $a_{n,p} = O(b_{n,p})$ and $b_{n,p} = O(a_{n,p})$. For random sequences, o and O are denoted as o_P and O_P, respectively.

2 The Model and Estimation

2.1 The Oracle Model

We consider the following factor IV model as true data generating process (DGP):

$$\begin{aligned} y_i &= d_i\alpha + \delta' x_i + \epsilon_{y,i} \\ d_i &= \rho' f_i + \theta' x_i + \epsilon_{d,i} \end{aligned} \tag{1}$$

where y_i is the outcome variable and d_i is the endogenous variable. The structural parameter α represents the treatment effect, which is our main interest to infer. The endogeneity problem of d_i arises since idiosyncratic errors $\epsilon_{y,i}$ and $\epsilon_{d,i}$ are correlated. The vector $x_i \in \mathbb{R}^r$ denotes many exogenous controls, where r may grow with the sample size n but remains smaller than n, and $f_i \in \mathbb{R}^J$ denotes a finite number of latent factors. The controls x_i and factors f_i satisfy the exclusion restriction $E(\epsilon_{y,i}|\{f_j, x_j\}_{j=1}^n) = 0$.

In addition, we also observe the high-dimensional instrumental variables $z_i \in \mathbb{R}^p$ follow factor model:

$$z_i = \Lambda f_i + u_i \tag{2}$$

where Λ is a $p \times J$ matrix that denotes the corresponding factor loadings. The instruments z_i satisfy the relevance condition for the endogenous variable d_i since they share the common latent factors. The idiosyncratic error u_i satisfies $E(u_i|\{f_j, x_j, \epsilon_{y,j}, \epsilon_{d,j}\}_{j=1}^n) = 0$.

Moreover, we consider weak factor structure on the instruments z_i, under which a non-trivial proportion of the available instruments is uninformative in the sense that they do not load on the underlying latent factors. The factor strength is diluted across the full set of instruments. Specifically,

$$z_i = \begin{pmatrix} z_{I,i} \\ z_{N,i} \end{pmatrix} = \begin{pmatrix} \Lambda_I \\ 0 \end{pmatrix} f_i + u_i$$

where Λ_I is a $p_0 \times J$ matrix, $\dim(z_{I,i}) = p_0$ and $\dim(z_{N,i}) = p - p_0$. Thus, the latent factors affect only the subset of instruments $z_{I,i}$, while the remaining instruments $z_{N,i}$ are pure noise.

Since the latent factors f_i satisfy both the exclusion restriction and the relevance condition, they are ideal instruments in the factor IV model. The difficulty is that these factors are unobserved. Bai and Ng (2010) propose to estimate the latent factors using principal component analysis (PCA) and then use the estimated factors as instruments. Their approach is motivated by the fact that replacing a large number of observed instruments with a finite number of factor estimates can alleviate the overfitting bias that arises in GMM estimators that use the high-dimensional instrument.

However, PCA-based factor estimators may fail to recover the space spanned by the latent factors when the factor structure is weak or when many uninformative instruments dilute factor strength; see, for example, Chao and Swanson (2022) and Bai and Ng (2023). Kapetanios and Marcellino (2010) further examine the consequences of weak factors and show that asymptotic inference for factor IV can break down when the factors are insufficiently strong. This naturally raises the question of whether one should rely on estimated factors as instruments or instead work directly with the observed instruments.

Kapetanios and Marcellino (2010) also consider exogenous controls in the factor IV model, but only a finite number, and require these controls to follow factor model where they share the same latent factor with instruments. By contrast, we allow the number of exogenous controls to grow with the sample size without imposing factor structure on them. Our setting is therefore more closely aligned with the IV literature that allows many exogenous controls (e.g., Kolesár (2013); Chao et al. (2023)).

Let Y, D, $\boldsymbol{\epsilon_y}$ and $\boldsymbol{\epsilon_d}$ denote the $n \times 1$ vectors of y_i, d_i, $\epsilon_{y,i}$ and $\epsilon_{d,i}$, respectively. Let Z be the $n \times p$ matrix where each row is z_i', X be the $n \times r$ matrix where each row is x_i', F be the $n \times J$ matrix where each row is f_i', U be the $n \times p$ matrix where each row is u_i'. Then (1) and (2) can be written in matrix form:

$$\begin{aligned} Y &= D\alpha + X\delta + \boldsymbol{\epsilon}_y \\ D &= F\rho + X\theta + \boldsymbol{\epsilon}_d \\ Z &= F\Lambda' + U \end{aligned} \tag{3}$$

2.2 The Working Model: Linear Projection

Given the data $\{y_i, d_i, z_i, x_i\}_{i=1}^n$ generated from the oracle model, we work directly with these observed variables in the following working model[1]:

$$\begin{aligned} y_i &= d_i\alpha + \delta' x_i + \epsilon_{y,i} \\ d_i &= \beta' z_i + \xi' x_i + e_i \end{aligned} \tag{4}$$

which is a standard linear IV specification. This working model replaces the latent factors in the first stage with the observed instruments and covariates. To see how this representation

[1] Working directly with observed variables in the presence of a latent factor structure has precedent in economics. For example, in program evaluation, Hsiao et al. (2012) estimate the average treatment effect on the treated using panel data in which the observed outcomes share common latent factors. They express the counterfactual outcome of the treated unit as a linear function of the observed outcomes of the control group, relying on a mean independence assumption. Li and Bell (2017) derives the asymptotic distribution of this estimator while relaxing the stringent mean independence assumption via linear projection.

arises from the factor IV model, consider the linear projection of d_i on (z_i, x_i):

$$\min_{\beta,\xi} E((d_i - \beta' z_i - \xi' x_i)^2)$$

The corresponding coefficients satisfy

$$\begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} E(z_i z_i') & E(z_i x_i') \\ E(x_i z_i') & E(x_i x_i') \end{pmatrix}^{-1} \begin{pmatrix} E(z_i d_i) \\ E(x_i d_i) \end{pmatrix}$$

The linear projection of d_i on (z_i, x_i) is therefore $L(d_i|z_i, x_i) = \beta' z_i + \xi' x_i$, and we obtain the first-stage equation

$$\begin{aligned} d_i &= \beta' z_i + \xi' x_i + (d_i - L(d_i|z_i, x_i)) \\ &= \beta' z_i + \xi' x_i + e_i. \end{aligned}$$

Through linear projection, the endogenous variable d_i is expressed as a linear function of the observed covariates (z_i, x_i) rather than the latent factors f_i, because the information in f_i is embedded in z_i (and in x_i if they are correlated). In this sense, the factor component of d_i is represented through the observed covariates. Furthermore, because we are recovering a low-dimensional factor signal using high-dimensional instruments that are strongly correlated, the coefficient vector β is generally dense rather than sparse. That is, the signal is spread across instruments: each individual component of β contributes only a small magnitude, rather than being concentrated on a few instruments.

The endogeneity problem in the oracle model also carries over to the working model. In particular, the error term e_i contains the idiosyncratic error $\epsilon_{d,i}$ in the oracle model, which is correlated with the structural error $\epsilon_{y,i}$. Moreover, e_i only satisfies $E\left[(z_i', x_i')' e_i\right] = 0$ by construction. Its conditional mean, however, may not be zero. We will show that valid asymptotic inference for the structural parameter α does not require $E(e_i|z_i, x_i) = 0$. In addition, because the projection does not include an intercept, it need not imply $E(e_i) = 0$ when the means of (z_i, x_i) are nonzero. If desired, one can include an intercept in the projection, $L(d_i \mid 1, z_i, x_i)$, which ensures $E(e_i) = 0$. Since our focus is to conduct inference on α, including an intercept in the linear projection is not necessary.

Let $\boldsymbol{e}$ be the $n \times 1$ vector of e_i. Then (4) can be written in matrix form:

$$\begin{aligned} Y &= D\alpha + X\delta + \boldsymbol{\epsilon}_y \\ D &= Z\beta + X\xi + \boldsymbol{e} \end{aligned} \tag{5}$$

2.3 Issues with High-Dimensional Instruments and Many Exogenous Controls

High-dimensional instruments and many exogenous controls create distinct but related statistical and computational challenges. To clarify their respective roles, we first discuss the difficulties posed by high-dimensional instruments. We then explain why extending existing high-dimensional IV procedures to accommodate many exogenous controls is nontrivial.

In high-dimensional IV, the number of instruments is allowed to grow with the sample size. It is useful to distinguish between two regimes. The first is the underparametrized regime, where the number of instruments grows with the sample size but remains smaller than the number of observations. This is the setting usually studied in the many instruments literature; see, for example, Chao and Swanson (2005), Hansen et al. (2008), and Mikusheva and Sun (2021). The second is the overparametrized regime, where the number of instruments exceeds the sample size, i.e., $\dim(z_i) = p > n$, which raises distinct difficulties for estimation and computation. In particular, when $p > n$, the matrix $Z'Z$ is singular by construction, so the conventional many IV estimators are not well-defined. Therefore, in the overparametrized regime, additional regularization procedures are generally required to ensure that the estimator is well behaved in finite samples.

There are many regularization procedures that are available to perform to mitigate the overfitting problem. Perhaps the most popular regularization procedure is to reduce the dimension of instruments. Belloni et al. (2012) propose Post-Lasso estimator, which first selects a low-dimensional set of instruments using Lasso and then implements usual 2SLS with the selected instruments. This estimator can achieve consistency and semi-parametric efficiency when Lasso consistently selects the relevant instruments under an (approximate) sparsity condition. However, when instruments are driven by latent factors, variable selection procedures may become problematic for two reasons. First, latent factors introduce strong cross-sectional correlation among the instruments, Lasso may fail to achieve variable selection consistency in the presence of the severe multicollinearity among instruments (Fan et al. (2020)). Second, the variable selection procedure relies on (approximate) sparsity to choose the correct set of informative instruments (Belloni et al. (2012)). That is, sparsity requires that the signal can be concentrated among a small set of instruments. Since instruments and endogenous variables share common latent factors, it makes the signal densely distributed among instruments. Therefore, consistent variable selection is not feasible since the sparsity requirement is violated.

PCA provides an alternative approach to dimension reduction. Bai and Ng (2010) propose estimating latent factors by PCA and using them as instruments. By replacing many observed instruments with a small number of estimated factors, this approach can mitigate first-stage overfitting. However, factor recovery and subsequent inference can deteriorate

when the factors are weak or their signal is diluted by many uninformative instruments; see Chao and Swanson (2022) and Bai and Ng (2023). Specifically, valid inference based on the PCA estimator requires at least semi-strong factor:

$$\sigma_{\min}(\Lambda_I'\Lambda_I) \asymp p_0^c, \text{ for some } c \in (1/2, 1]$$

Another popular regularization procedure is ridge regression. Hansen and Kozbur (2014) propose ridge-regularized JIVE (RJIVE), which combines the leave-one-out estimator with ridge-regularized first stage. Specifically, the first-stage coefficient is estimated with ridge regression using all observations except the ith observation:

$$\hat{\beta}_{(-i)}^{\lambda} = \left(Z'Z - z_i z_i' + \lambda I_p\right)^{-1} \left(Z'D - z_i d_i\right),$$

where $\lambda > 0$ is the ridge penalty. The penalty is designed so that the ridge regression favors models with small coefficients, thereby controlling overfitting and stabilizing the first-stage regression. The corresponding fitted value is $\hat{d}_i^{\lambda} = z_i'\hat{\beta}_{(-i)}^{\lambda}$. The resulting RJIVE estimator is

$$\hat{\alpha}_{RJIVE} = (\sum_{i=1}^{n} \hat{d}_i^{\lambda} d_i)^{-1} \sum_{i=1}^{n} \hat{d}_i^{\lambda} y_i \tag{6}$$

Thus, the use of JIVE coupled with ridge regularization allows sufficient signal extraction from a large, potentially non-sparse set of instruments while mitigating first-stage overfitting.

Incorporating many exogenous controls into JIVE-type estimator, however, is challenging, since correlations between the two stages cannot be eliminated by conventional procedures. That is, once we partial out exogenous controls X using an orthogonal projection matrix $M_X = I - X(X'X)^{-1}X'$, each residualized observation is a linear combination of observations from the full sample. Consequently, this residualization step reintroduces the dependence between the first-stage and structural errors.

Several JIVE-type estimators have been developed to accommodate both many instruments and many exogenous controls; see, for example, Kolesár (2013) and Chao et al. (2023). In general, these estimators take the form:

$$\hat{\alpha}_{JIVE} = (D'GD)^{-1}D'GY$$

where G is an $n \times n$ matrix that takes several form that remove the diagonal elements. For example, Chao et al. (2023) proposed fixed effect JIVE (FEJIVE) that allow fixed effects and clustered sample settings with $G = P_{Z^\perp} - M_{(Z,X)}D_{\hat{\vartheta}}M_{(Z,X)}$ where $P_{Z^\perp} = Z^\perp(Z^{\perp\prime}Z^\perp)^{-1}Z^{\perp\prime}$, $D_{\hat{\vartheta}}$ is the diagonal matrix with elements such that $P_{Z^\perp,ii} = [M_{(Z,X)}D_{\hat{\vartheta}}M_{(Z,X)}]_{ii}$. The second term is the correction term that remove the diagonal element of $P_{Z^\perp}$ while remaining

orthogonal to (Z, X).

While FEJIVE is an effective procedure for many instruments and many controls, it still becomes problematic in the overparametrized regime since projection matrices are ill-defined when $p + r > n$. Combining FEJIVE with ridge regularization does not resolve this. For example, let $P^{\lambda}_{Z^{\perp}} = Z^{\perp}(Z^{\perp\prime}Z^{\perp} + \lambda I_p)^{-1}Z^{\perp\prime}$. If we consider $G = P^{\lambda}_{Z^{\perp}} - M_X D_{\hat{\vartheta}} M_X$, even if we can find $D_{\hat{\vartheta}}$ such that the diagonal elements of G are zero, the second term of G doesn't partial out Z will introduce a nonnegligible term in FEJIVE. Thus, FEJIVE become inconsistent when combined with ridge regularization. Therefore, constructing JIVE-type estimator that simultaneously accommodates high-dimensional instruments and many exogenous controls is far less straightforward.

2.4 The Estimation

In light of the issues existing procedures face with high-dimensional instruments and many exogenous controls, we propose a new estimation procedure that remains valid for statistical inference on the structural parameter of interest.

We begin by partialling out the exogenous variables X using an orthogonal projection matrix. Let $M_X = I_n - X(X'X)^{-1}X'$. We denote $Y^{\perp} = M_X Y$, $D^{\perp} = M_X D$, and $Z^{\perp} = M_X Z$ as the outcome variable, endogenous variable, and instrumental variables, respectively, with the exogenous covariates partialled out. This formulation requires $\dim(x_i) = r < n$, ensuring that $X'X$ is full column rank and M_X is well defined.

For the first-stage regression, we consider the Ridgeless regression of $D^{\perp}$ on $Z^{\perp}$, the estimator is defined by

$$\hat{\beta} = (Z^{\perp\prime}Z^{\perp})^{+}Z^{\perp\prime}D^{\perp} \tag{7}$$

where $(Z^{\perp\prime}Z^{\perp})^{+}$ denotes the Moore–Penrose pseudoinverse of $Z^{\perp\prime}Z^{\perp}$. Equivalently, $\hat{\beta}$ is the minimum ℓ_2 norm least-squares solution:

$$\hat{\beta} = \arg\min\{\|\beta\| : \beta \text{ minimize } \|D^{\perp} - Z^{\perp}\beta\|^2\}$$

The term "ridgeless" is motivated by the fact that $\hat{\beta} = \lim_{\lambda \to 0^+} \hat{\beta}_{\lambda}$, where $\hat{\beta}_{\lambda}$ denotes the ridge regression estimator

$$\hat{\beta}_{\lambda} = (Z^{\perp\prime}Z^{\perp} + \lambda I)^{-1}Z^{\perp\prime}D^{\perp}$$

When $p+r < n$, the estimator $\hat{\beta}$ reduces to the ordinary least squares, $\hat{\beta} = (Z^{\perp\prime}Z^{\perp})^{-1}Z^{\perp\prime}D^{\perp}$. When $p+r = n$, a unique least squares solution still exists, but the estimator interpolates the in-sample data perfectly, i.e., $D^{\perp} = Z^{\perp}\hat{\beta}$. When $p + r > n$, the model is overparametrized, and multiple solutions perfectly fit the data. Among these solutions, $\hat{\beta}$ is the one with the smallest norm (Hastie et al. (2022)).

This minimum norm property explains the manifestation of the double descent phe-

nomenon in extremely high-dimensional models. Specifically, as model complexity expands beyond the interpolation threshold, the variance of the estimator gradually decreases, yielding improved MSE performance. Intuitively, when $p + r > n$, increasing the number of instruments provides the ridgeless estimator with a larger set of perfect-fit solutions to search through. It can therefore find a β that perfectly fits the data while achieving a smaller ℓ_2 norm. Consequently, the estimator becomes more regularized as the number of regressors increases because the β estimates shrink toward zero and are distributed across more dimensions, which simultaneously diversifies and reduces the estimator's variance, thereby improving MSE.

Therefore, unlike conventional regularization methods such as Lasso or Ridge, which control overfitting by explicitly penalizing model complexity, the ridgeless estimator can exploit benign overfitting in sufficiently overparametrized settings. In this sense, overfitting may become a blessing rather than a curse when the number of instruments is sufficiently large.

However, if Ridgeless regression is implemented using the full sample in the first stage, the resulting fitted values still suffer from bias, since the same observations are used in both stages of estimation. Therefore, we combine Ridgeless regression with a sample-splitting procedure. Specifically, we estimate the first-stage coefficient on one subsample and use the resulting fitted values on the other subsample to estimate the structural parameter. Importantly, the exogenous controls are partialled out separately within each subsample rather than using the full-sample projection matrix M_X. This procedure preserves the independence between the first-stage estimation sample and the second-stage estimation sample. As a result, the proposed procedure avoids the bias that arises from using the same observations in both stages, while Ridgeless regression keeps the first stage well defined even when the number of first-stage regressors exceeds the subsample size. The estimation procedure is summarized as follows.

Algorithm (Sample-Splitting Ridgeless IV).

Step 1: *Sample splitting.* Randomly partition the sample $\{1, \ldots, n\}$ into two disjoint subsets I_1 and I_2 of equal size $n_1 = n_2 = n/2$. Let (Y_1, D_1, Z_1, X_1) and (Y_2, D_2, Z_2, X_2) be the corresponding subsamples. For each $k \in \{1, 2\}$, define the orthogonal projection matrix

$$M_{X_k} = I_{n_k} - X_k(X_k'X_k)^{-1}X_k',$$

and variables after partialling out X_k:

$$Y_k^{\perp} = M_{X_k}Y_k, \quad D_k^{\perp} = M_{X_k}D_k, \quad Z_k^{\perp} = M_{X_k}Z_k.$$

Using subsample I_2, estimate β by Ridgeless regression:

$$\hat{\beta}_2 = (Z_2^{\perp\prime} Z_2^{\perp})^+ Z_2^{\perp\prime} D_2^{\perp}.$$

Using subsample I_1, construct the fitted value:

$$\hat{D}_1^{\perp} = Z_1^{\perp} \hat{\beta}_2.$$

Step 2: *Estimation of* α. Following the IV estimator, we use the fitted value $\hat{D}_1$ and I_1 observations:

$$\hat{\alpha}_1 = (\hat{D}_1^{\perp\prime} D_1^{\perp})^{-1} \hat{D}_1^{\perp\prime} Y_1^{\perp}.$$

Step 3: *Sample-splitting.* Repeat Steps 1-2 with I_1 and I_2 exchanged to obtain $\hat{\alpha}_2$. Define the estimator for α as

$$\hat{\alpha} = \frac{1}{2}(\hat{\alpha}_1 + \hat{\alpha}_2).$$

One may also consider combining Ridgeless regression with a JIVE-type estimator by running leave-one-out estimation using the Moore–Penrose inverse. However, this approach is not useful in the overparametrized regime. To see this, consider a naive Moore–Penrose extension of the FEJIVE weighting matrix. When (Z, X) has full row rank and $p+r > n$, The FEJIVE weighting matrix $G = P_{Z^\perp} - M_{(Z,X)} D_{\hat{\vartheta}} M_{(Z,X)}$ reduces to M_X[2], and the resulting estimator becomes

$$\hat{\alpha}_{JIVE} = (D' M_X D)^{-1} D' M_X Y$$

which is simply the OLS estimator after partialling out the exogenous controls X. Since D is endogenous, this estimator is inconsistent for α.

2.5 Comparison with RJIVE

The ridge-regularized JIVE (RJIVE) estimator proposed by Hansen and Kozbur (2014) is closely related to our procedure. Both methods allow for non-sparse first-stage signal and avoid using the same observation in both stages of estimation. Their main difference lies in how they regularize first-stage estimation: RJIVE applies explicit ridge shrinkage, while our estimator uses the Moore–Penrose pseudoinverse and relies on overparameterization for implicit regularization.

[2]Let $W = (Z, X)$. If $p + r > n$ and W has full row rank, its Moore–Penrose pseudoinverse is given by $W^+ = W'(WW')^{-1}$. The projection $P_{(Z,X)} = W(W'W)^+W' = WW'(WW')^{-1}(WW')^{-1}WW' = I_n$. Let $W^\perp = (Z^\perp, X)$, it follows similarly that $P_{W^\perp} = I_n$. Because X is orthogonal to $Z^\perp$ by construction ($X'Z^\perp = 0$), the projection matrix can be decomposed into the individual projections: $P_{W^\perp} = P_{Z^\perp} + P_X$, where $P_X = X(X'X)^{-1}X'$. Rearranging terms yields $P_{Z^\perp} = I_n - P_X = M_X$ whenever $p + r > n$.

To isolate this difference, we focus the analysis on the component directly affected by overfitting, that is, the first-stage estimation error involving the first-stage noise e_i.

For RJIVE estimator, by standard recursive residual[3], the component of $\hat{\alpha}_{RJIVE}$ attributable to first-stage noise is

$$\frac{1}{n}\sum_{i\neq j} e_i P_{ij}^{\lambda}(1-P_{jj}^{\lambda})^{-1}\epsilon_{y,j} = \begin{cases} O_P(\frac{\sqrt{p}}{n}), & \text{if } n > p \\ O_P(\frac{1}{\sqrt{n}}), & \text{if } p > n \end{cases}$$

where P_{ij}^{λ} is the (i,j) element of $P^{\lambda} = Z(Z'Z+\lambda I)^{-1}Z'$.

This result follows from the fact that the variance depends on the number of non-zero eigenvalues of P^{λ}, which is $\min\{n,p\}$. Thus, in the underparameterized regime ($n > p$), the variance suffers from overfitting that increases with p. Once p exceeds n, however, the variance of RJIVE no longer grows with p because P^{λ} has only n non-zero eigenvalues. RJIVE can therefore avoid the overfitting problem in the overparametrized regime. Importantly, this argument is not specific to the Factor IV structure; the same conclusion holds within the framework of Hansen and Kozbur (2014)

For our Ridgeless estimator, consider the corresponding component in one sample-splitting estimator $\hat{\alpha}_1$:

$$\frac{1}{n_1}\boldsymbol{e}_2' Z_2^{\perp}(Z_2^{\perp\prime}Z_2^{\perp})^{+}Z_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp} = \begin{cases} O_P(\sqrt{\frac{p}{n(n-r)}}), & \text{if } n_2 > p+r \\ O_P(\frac{1}{\sqrt{p}}), & \text{if } p+r > n_2 \end{cases} \tag{8}$$

Thus, in the underparameterized regime, the magnitude of this term also increases with p. After the interpolation threshold is crossed ($p+r > n_2$), however, it decreases at the rate $1/\sqrt{p}$, whereas the corresponding RJIVE term remains of order $1/\sqrt{n}$. Therefore, our estimator not only avoids the overfitting problem like RJIVE but also benefits from adding more instruments in the overparametrized regime. This represents the double descent phenomenon exhibited in the Ridgeless estimator (Hastie et al. (2022)).

Specifically, its variance conditional on (Z_2, X_2) can be shown[4] that

$$\mathrm{Var}(\frac{1}{n_1}\boldsymbol{e}_2' Z_2^{\perp}(Z_2^{\perp\prime}Z_2^{\perp})^{+}Z_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp}|Z_2,X_2) = \sigma_u^2\sigma_{\epsilon_d}^2\sigma_{\epsilon_y}^2\frac{1}{n_1}\mathrm{tr}((Z_2^{\perp\prime}Z_2^{\perp})^{+}) + o_P(1)$$

where this variance depends on $\mathrm{tr}((Z_2^{\perp\prime}Z_2^{\perp})^{+})$. Since Z_2 follows factor model, the leading J eigenvalues of $Z_2^{\perp\prime}Z_2^{\perp}$ are factor-driven, of order $n_2\psi_{n,p}$, while the remaining non-spiked eigenvalues (the $(J+1)$-th through $\min\{n_2-r,p\}$-th) follow those of $U_2^{\perp\prime}U_2^{\perp}$ and are of

[3]The detail can be found in the Appendix of Hansen and Kozbur (2014)

[4]The detail can be found in the Appendix

order $(n_2 - r) + p$ by random matrix theory. This yields

$$\operatorname{tr}((Z_2^{\perp\prime} Z_2^{\perp})^{+}) = \sum_{i=1}^{\min\{n_2-r,p\}} \frac{1}{\sigma_i(Z_2^{\perp\prime} Z_2^{\perp})} = \begin{cases} O_P(\frac{p}{n-r}), & \text{if } n_2 > p + r \\ O_P(\frac{n-r}{p}), & \text{if } p + r > n_2 \end{cases}$$

Consequently, (8) exhibits the double descent phenomenon. That is, when the model is underparametrized, i.e., $p + r < n_2$, its variance is amplified as the number of instruments diverge; but once we cross the interpolation threshold ($p + r > n_2$), the variance decays as the number of instruments grows.

Therefore, there are two fundamental differences between RJIVE and the Ridgeless estimator. First, the term related to first-stage noise decay faster than that of the corresponding RJIVE term in the overparametrized regime. For RJIVE, the ridge penalty ($\lambda > 0$) shrinks the eigenvalues of P^{λ} matrix to be less than one, so the associated variance is of order $1/n$. By contrast, the pseudo-inverse $(Z_2^{\perp\prime} Z_2^{\perp})^{+}$ directly invert the eigenvalues of $Z_2^{\perp\prime} Z_2^{\perp}$, so the associated variance is of order $1/p$. Second, as discussed above, incorporating many exogenous controls into RJIVE is not straightforward. By contrast, the Ridgeless estimator can naturally accommodate many controls and remains consistent in the overparametrized case.

3 Theoretical results

This section justifies the oracle model (1) implies the working model (2), and establishes the asymptotic properties of $\hat{\alpha}$.

Assumption 1. *(DGP on d_i and z_i). The variables d_i and z_i are:*

$$d_i = \rho' f_i + \theta' x_i + \epsilon_{d,i}$$
$$z_i = \begin{pmatrix} z_{I,i} \\ z_{N,i} \end{pmatrix} = \begin{pmatrix} \Lambda_I \\ 0 \end{pmatrix} f_i + u_i = \Lambda f_i + u_i$$

where $\dim(f_i) = J$, $\dim(x_i) = r$, $\dim(\Lambda_I) = p_0$, *and* $\dim(\Lambda) = p$

(i) The factor strength is determined by $\psi_{n,p}$. There is a sequence $\psi_{n,p} \to \infty$, such that

$$\sigma_{\min}(\Lambda_I' \Lambda_I) \asymp \sigma_{\max}(\Lambda_I' \Lambda_I) \asymp \psi_{n,p}$$

(ii) With probability approaching one (as $n \to \infty$),

$$\sigma_{\min}(X'X) \asymp \sigma_{\max}(X'X) \asymp n$$

(iii) $E(f_i f_i') = \Sigma_f$ *for some* $J \times J$ *positive definite matrix* Σ_f, $E(x_i x_i') = \Sigma_x$ *for some* $r \times r$ *positive definite matrix* Σ_x, $rank([F, X]) = J + r$. *Furthermore,* $\frac{1}{n} F' M_X F = (1 - \frac{r}{n})\Sigma_{f|x} + O_P(\frac{1}{\sqrt{n}})$, *where* $\Sigma_{f|x} = \Sigma_f - \Sigma_{fx}\Sigma_x^{-1}\Sigma_{xf}$, $\Sigma_{xf} = E(x_i f_i')$

(iv) $\|\rho\| \leq C$ *and* $\|\theta\| \leq C$ *for some constants* $C > 0$

Assumption 1 specifies the DGP for (d_i, z_i) in the oracle model. The instruments z_i follow a factor model with potentially weak factors. Weakness can arise for two reasons. First, only an unknown subset of instruments are informative: the components $z_{I,i}$ load nontrivially on the underlying factors, while the remaining components $z_{N,i}$ are irrelevant to the factors and thus constitute pure noise. In practice, variable selection (e.g., Chao and Swanson (2022); Giglio et al. (2025)) is usually implemented to prune the irrelevant instruments and enhance the factor strength, thereby enabling the consistent estimation of the factors. Second, even among informative instruments, loadings may be weak. Assumption 1 (i) imposes a relatively weak factor strength. To make this explicit, consider

$$\sigma_{\min}(\frac{\Lambda_I'\Lambda_I}{p_0^a}) \asymp \sigma_{\max}(\frac{\Lambda_I'\Lambda_I}{p_0^a}) \asymp C$$

for some constants $C > 0$ and $a \in (0, 1]$, so that $\psi_{n,p} \asymp p_0^a$. This allows factor strength only slightly stronger than when $a = 0$, a "weak factor" definition in Onatski (2012). By contrast, asymptotic inference based on the conventional PCA-based IV estimator is established under the sufficient conditions $a > 1/2$ and $p/(np_0^a) \to 0$ (Kapetanios and Marcellino (2010); Bai and Ng (2023)). Thus, for the PCA-based factors to be valid instruments, the factor loadings must be sufficiently strong and the number of informative instruments must be comparatively large.

Assumption 1 (ii) ensures that the controls X are well conditioned so that $X'X/n$ is invertible in large sample. The rank restriction in Assumption 1(iii) excludes perfect collinearity among first-stage regressors. Assumption 1 (iv) imposes that the l_2-norms of the nuisance parameters are bounded.

Assumption 2. *(idiosyncratic errors). Let* $\mathcal{Z} = (F, X, U)$. *Assume the following conditions are satisfied*

(i) Conditional on $\mathcal{Z}$, $(\epsilon_{y,i}, \epsilon_{d,i})$ *are i.i.d across* i, u_{ij} *is i.i.d across both* (i, j), *and* $E(u_{ij}|F, X, \boldsymbol{\epsilon_y}, \boldsymbol{\epsilon_d}) = 0$, $E(\epsilon_{y,i}|\mathcal{Z}) = 0$, $E(\epsilon_{d,i}|\mathcal{Z}) = 0$

(ii) There are some constants $C > 0$ *such that* $E(u_{ij}^4) \leq C$, $E(\epsilon_{y,i}^4|\mathcal{Z}) \leq C$, $E(\epsilon_{d,i}^4|\mathcal{Z}) \leq C$

(iii) $\{u_i\}_{i=1}^n$ *is independent of* $\{x_i\}_{i=1}^n$

Assumption 2 (i) imposes standard conditions on the error terms, similar to those in Hansen and Kozbur (2014) and Chao et al. (2023), and formalizes the exclusion restrictions.

Specifically, because the instruments Z are a linear function of F and U, this condition implies that $E(\epsilon_{y,i}|Z,X)=0$ and $E(\epsilon_{d,i}|Z,X)=0$. It also assumes the independence of the error terms across observations. Assumption 2 (ii) places bounded moment conditions on the error terms, which are useful for establishing asymptotic normality and the consistency of the asymptotic variance. Assumption 2 (iii) requires the error terms u_i to be independent of the exogenous controls x_i. Together with the assumption that the entries u_{ij} are i.i.d. with bounded fourth moments, it enables us to characterize the asymptotic eigenvalue properties of the residualized noise matrix $U'M_XU$ via random matrix theory.

We acknowledge that requiring the idiosyncratic errors u_{ij} to be i.i.d. is a stringent condition, as approximate factor model typically allows weak correlation; see Bai and Ng (2002). Dependence can be introduced through $U=V\Sigma_u^{1/2}$, where V has i.i.d. entries with bounded fourth moments and Σ_u is a deterministic positive definite covariance matrix, as studied by Hastie et al. (2022). However, the variance of the estimator would depend on the limiting spectral distribution of Σ_u and would generally be characterized through an implicit integral equation. We therefore maintain the i.i.d. assumption to keep the closed-form of the variance of the estimator, and leave this extension for future research.

Theorem 1. *Suppose oracle model* (3) *and Assumptions 1 and 2 hold. Then the following working model holds:*

$$\begin{aligned} y_i &= d_i\alpha + x_i'\delta + \epsilon_{y,i} \\ d_i &= z_i'\beta + x_i'\xi + e_i \end{aligned} \tag{9}$$

where

$$\begin{aligned} &\beta = E(z_i^{\perp}z_i^{\perp\prime})^{-1}E(z_i^{\perp}d_i^{\perp}), \quad \|\beta\| = O(\psi_{n,p}^{-1/2}) \\ &z_i^{\perp} = z_i - E(z_ix_i')E(x_ix_i')^{-1}x_i, \\ &d_i^{\perp} = d_i - E(d_ix_i')E(x_ix_i')^{-1}x_i, \end{aligned}$$

Theorem 1 justifies the linear IV working model. It characterizes the coefficient β is dense: signal is spread across all instruments, with each component being small in magnitude. This supports the empirical findings of Angrist and Frandsen (2022) that instrument signal in the first-stage regression is typically diffuse rather than sparse, i.e., no small subset of instruments provides a strong forecast for the endogenous variable. As a result, variable selection methods (e.g., Post-Lasso) that rely on (approximate) sparsity preform poor fit and exhibit substantial bias.

In addition, $\|\beta\| = O(\psi_{n,p}^{-1/2})$ indicates that stronger factors imply a weaker correlation between instruments and the endogenous variable. In the oracle model, the endogenous variable and the instruments share the same factors; when those loadings are strong, the

same first-stage signal can be produced with much smaller coefficients in the working model, so the correlation is weaker. As $\psi_{n,p} \to \infty$, β shrinks toward zero, exhibiting the local-to-zero behavior commonly assumed in the many weak instruments literature (Staiger and Stock (1997), Chao and Swanson (2005)).

Theorem 1 shows that each individual instrument is only weakly correlated with the endogenous variable. This raises the question of how much identification information the instruments provide collectively. In the many weak instruments literature, a standard measure of identification strength is the concentration parameter, defined as

$$\mu^2 = \frac{\beta' Z^{\perp\prime} Z^{\perp} \beta}{\sigma_e^2}$$

which is the signal-to-noise ratio of the first-stage regression. Because we allow many exogenous controls, we follow Bekker (1994) and Kolesár et al. (2015) and measure the signal using the residualized instruments $Z^{\perp}$. This differs from the conventional many weak instruments literature Stock and Yogo (2002); Chao and Swanson (2005); Mikusheva and Sun (2021), which typically does not consider many exogenous controls. If there are no exogenous controls, it reduces to the usual concentration parameter

$$\mu^2 = \frac{\beta' Z' Z \beta}{\sigma_e^2}.$$

As emphasized by Mikusheva and Sun (2021), the magnitude of μ^2 alone is not a reliable measure of identification strength when the number of instruments is large. What matters is the size of μ^2 relative to the number of instruments. In the many weak instrument literature, there have been several definitions of weak instruments. The regimes can be ordered from strongest to weakest as follows:

(i) $\mu^2/p \xrightarrow{p} \infty$,

(ii) $\mu^2/\sqrt{p} \xrightarrow{p} \infty$,

(iii) $\mu^2/\sqrt{p} \xrightarrow{p} C < \infty$.

Regime (i) is a necessary condition for 2SLS consistency (Chao and Swanson (2005)); when $\mu^2/p \xrightarrow{p} C < \infty$, instruments are "weak" in the sense of Stock and Yogo (2002). Regime (ii) is a necessary condition for the JIVE estimator to be consistent (Chao et al. (2012)). Regime (iii) corresponds to the case that conventional many weak IV estimators are inconsistent (Mikusheva and Sun (2021)).

To understand which identification regime arise from the working model, it is important to characterize the order of the concentration parameter:

Theorem 2. *Suppose Assumptions 1 and 2 hold. Then the concentration parameter in the working model (5) satisfies*

$$\mu^2 = \frac{\beta' Z^{\perp\prime} Z^{\perp} \beta}{\sigma_e^2} \asymp (n - r).$$

In particular, if there are fixed (or no) exogenous controls, then $\mu^2 \asymp n$.

Theorem 2 shows that the first-stage signal grows at the order $n-r$ where each individual instrument alone only offers weak identification power. Because the identification strength is between μ^2 and the instrument dimension p, the working model can still fall into a weak-identification regime when p grows sufficiently fast. In particular, since $\mu^2 \asymp (n - r)$, the ratio $\mu^2/\sqrt{p}$ may be bounded or even vanish when p is large. This is different from the many weak instruments because in the high-dimensional instruments, we allow the number of instruments to be larger than the sample size.

This observation motivates the analysis of consistency of our Ridgeless estimator under $\mu^2/\sqrt{p} \xrightarrow{p} C < \infty$. Recall from (8) that, in the overparametrized regime ($p + r > n/2$), the key component of the estimation error of $\hat{\alpha}_1$ attributable to first-stage noise is

$$\frac{1}{n_1} \boldsymbol{e}_2' Z_2^{\perp} (Z_2^{\perp\prime} Z_2^{\perp})^{+} Z_1^{\perp\prime} \boldsymbol{\epsilon}_{y,1}^{\perp} = O_P(\frac{1}{\sqrt{p}})$$

which vanishes asymptotically, allowing the estimator to achieve consistency under $\mu^2/\sqrt{p} \xrightarrow{p} C < \infty$.

Assumption 3. *(Number of instruments and exogenous regressors).*
As $n, r, p \to \infty$, $\frac{r}{n/2} \to \tau \in [0, 1)$ *and* $\frac{p}{n/2-r} \to \gamma \in (0, \infty] \setminus \{1\}$.

Assumption 3 extends the (very) many instruments asymptotic framework (e.g., Bekker (1994); Belloni et al. (2012)) by allowing many exogenous controls. It characterizes the relative growth of the number of first-stage regressors $p + r$ and the subsample size $n/2$. More precisely, the two asymptotic conditions imply that

$$\frac{p + r}{n/2} \to \chi := \gamma + (1 - \gamma)\tau \in (0, \infty] \setminus \{1\}$$

This accommodates two cases: (i) the number of first-stage regressors grows proportionally with the subsample size (i.e., $\chi \in (0, \infty) \setminus \{1\}$), and (ii) it grows much faster than the subsample size (i.e., $\chi = \infty$)

Furthermore, while Assumption 3 allows the number of instruments to exceed the subsample size, it restricts the dimension of the exogenous covariates to be smaller than the subsample size ($r < n/2$) so that the projection matrix M_{X_k} can be constructed using the standard inverse. If one instead defines M_{X_k} using the Moore-Penrose inverse, then when

$r \geq n/2$ and X_k has full row rank, $M_{X_k} = 0$, residualization removes all variation. Consequently, standard partialling-out cannot be directly applied when the number of controls exceeds the subsample size.

Alternatively, one may consider the following procedure: run Ridgeless regression of D_2 on (Z_2, X_2) in one subsample as the first-stage regression, and another subsample to construct the IV estimator. This approach, however, introduces terms involving the high-dimensional controls that need not vanish. In particular, when the exogenous controls load significantly on the structural equation (i.e., $\|\delta' X' X \delta\| = O(n)$), this IV estimator is inconsistent.

Assumption 4. *(Approximation).*

(i) $\max_{1\leq i\leq n} \|F' m_i\|/\sqrt{n} = o_P(1)$, *where* m_i *is the* i*-th column of* M_X.

(ii) $p = o(\psi_{n,p} n)$

Assumption 4 (i) is similar to Assumption 3 of Cattaneo et al. (2018), where many exogenous controls are also considered. This condition restricts the distributional relationship between the finite-dimensional factors f_i and the high-dimensional nuisance covariate x_i, and is close to a minimal requirement for the central limit theorem to hold. Assumption 4 (ii) guarantees the bias of first-stage estimator for β go to zero in the overparametrized regime. It requires that the informative instruments carry enough information. For example, if all informative instruments are strong, i.e., $p_0 \asymp \psi_{n,p}$, then the bias should vanish asymptotically, provided that $p = o(p_0 n)$.

Let $\Sigma_{Z^\perp} := \Lambda \Sigma_{f|x} \Lambda' + \sigma_u^2 I_p$. The following result establishes the asymptotic normality of the Ridgeless estimator.

Theorem 3. *Suppose Assumptions 1 to 4 hold. As* $n, r, p \to \infty$, $\frac{p}{n/2-r} \to \gamma \in (0,\infty) \setminus \{1\}$,

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, \Sigma)$$

where $\Sigma = \sigma_{\epsilon_y}^2 V^{-1} + V^{-1}(\Gamma + \Pi)V^{-1}$, $V = (1-\tau)\beta' \Sigma_{Z^\perp} \beta$, $\Gamma = \begin{cases} \sigma_{\epsilon_d}^2 \sigma_{\epsilon_y}^2 (1-\tau)\frac{\gamma}{1-\gamma}, & 0 < \gamma < 1 \\ \sigma_{\epsilon_d}^2 \sigma_{\epsilon_y}^2 (1-\tau)\frac{1}{\gamma - 1}, & \gamma > 1 \end{cases}$,

$\Pi = \begin{cases} \sigma^2(1-\tau)\gamma, & 0 < \gamma < 1 \\ \sigma^2(1-\tau)\frac{1}{\gamma}, & \gamma > 1 \end{cases}$

In particular, as $\frac{p}{n/2-r} \to \infty$, $\Sigma = \sigma_{\epsilon_y}^2 V^{-1}$

The asymptotic variance exhibits differently depending on the relative growth of the number of first-stage regressor $p + r$ and the subsample size $n/2$. When $p + r \asymp n/2$ (i.e., $\gamma \in (0,\infty)$), the asymptotic variance comprises two terms. The first term, $\sigma_{\epsilon_y}^2 V^{-1}$,

corresponds to the usual asymptotic variance, and the second term, $V^{-1}(\Gamma+\Pi)V^{-1}$, is an adjustment for the presence of (very) many instruments. The boundary case, $\gamma = 1$, is excluded because the adjustment term in asymptotic variance explodes at the interpolation threshold, reflecting the double descent behavior of the first-stage estimator for β; variance peaks at the interpolation threshold and then disperses once we cross it. Therefore, when we have substantial instruments so that $p + r \gg n/2$ (i.e., $\gamma = \infty$), the adjustment term vanishes and the asymptotic variance collapses to $\sigma^2_{\epsilon_y} V^{-1}$

To conduct asymptotically valid inference using the asymptotic normality result in Theorem 3, we next construct a consistent estimator of Σ. In the following theorem, we use subscripts to index the two subsamples: $k \in \{1, 2\}$ and $k^c \in \{1, 2\}\backslash\{k\}$. For example, Z_k denotes an $(n/2) \times p$ matrix where each row is z_i' for $i \in I_k$.

Theorem 4. *Suppose Assumptions 1 to 4 hold. As* $n, r, p \to \infty$,

$$\hat{\Sigma} \xrightarrow{p} \Sigma$$

where $\hat{\Sigma} = \hat{V}^{-1}(\frac{1}{2}\hat{\Omega}_1+\frac{1}{2}\hat{\Omega}_2+\hat{\Pi})\hat{V}^{-1}$, $\hat{V} = \frac{1}{n}(\hat{D}_1^{\perp\prime} D_1^{\perp}+\hat{D}_2^{\perp\prime} D_2^{\perp})$, $\hat{\Omega}_k = \frac{1}{n_k}\hat{\sigma}^2_{\epsilon_y}\hat{\beta}'_{k^c} P_{k^c} Z_k^{\perp\prime} Z_k^{\perp} P_{k^c}\hat{\beta}_{k^c}$, $\hat{\Pi} = \frac{1}{n_k}\hat{\sigma}^2 tr(P_1 P_2)$, $\hat{\sigma}^2_{\epsilon_y} = \frac{1}{n-r}(Y - D\hat{\alpha})' M_X (Y - D\hat{\alpha})$, $\hat{\sigma}^2 = \frac{1}{n-r}(Y - D\hat{\alpha})' M_X D$, $P_{k^c} = Z_{k^c}^{\perp\prime} Z_{k^c}^{\perp}(Z_{k^c}^{\perp\prime} Z_{k^c}^{\perp})^+$

For consistency of the variance estimator, Assumption 4 (ii), $p = o(n\psi_{n,p})$, is required to ensure that the bias of the first-stage estimator of β is asymptotically negligible. If too many uninformative instruments are included, however, $\hat{V}$ may become small relative to $\hat{\Omega}_k$, thereby inflating $\hat{\Sigma}$. Thus, although the asymptotic variance Σ declines in the sufficiently overparameterized regime because of the double descent phenomenon, the estimated variance need not exhibit the same behavior when the number of uninformative instruments grows too rapidly.

To evaluate the asymptotic efficiency of our proposed estimator, we consider a theoretical benchmark: an oracle estimator that could be constructed if the latent factors, F, are directly observable. This "Factor IV" (FIV) estimator, denoted $\hat{\alpha}_{FIV}$, uses the observable factors F as instruments for the endogenous variable D:

$$\hat{\alpha}_{FIV} = (\hat{D}_{FIV}^{\perp\prime} D^{\perp})^{-1} \hat{D}_{FIV}^{\perp\prime} Y^{\perp}$$

where $\hat{D}^{\perp}_{FIV}$ is the fitted value from the regression of $D^{\perp}$ on $F^{\perp}$, with $D^{\perp} = M_X D$, $F^{\perp} = M_X F$.

Theorem 5. *Suppose Assumptions 1 to 4 hold.*

(i) As $n, r \to \infty$,

$$\sqrt{n}(\hat{\alpha}_{FIV} - \alpha) \xrightarrow{d} N(0, \Sigma_{FIV})$$

where $\Sigma_{FIV} = \sigma^2_{\epsilon_y}((1-\frac{\tau}{2})\rho'\Sigma_{f|x}\rho)^{-1}$.

(ii) *As* $n, r, p \to \infty$, $\frac{r}{n/2} \to 0$, $\frac{p}{n/2-r} \to \infty$,

$$\Sigma = \Sigma_{FIV} + o(1)$$

Theorem 5 establishes the asymptotic distribution of the oracle estimator. It shows that, with a sufficient number of instruments and a finite (or zero) number of exogenous controls, the asymptotic variance Σ of our sample-splitting estimator is equivalent to the oracle variance Σ_{FIV}. This result demonstrates that, even without directly observing the latent factors, our estimator achieves the same efficiency as the oracle benchmark.

In fact, when the number of exogenous controls is fixed, Σ_{FIV} coincides with the semi-parametric efficiency bound for α. To see this, note that by the Frisch–Waugh–Lovell theorem, $\hat{\alpha}_{FIV}$ is identical to the coefficient on D from the 2SLS regression of Y on (D, X), using (F, X) as instruments. Under conditional homoskedasticity of $\epsilon_{y,i}$, the optimal instrument for estimating structural parameters $(\alpha, \delta')'$ is

$$A_i = E\left[\begin{pmatrix} d_i \\ x_i \end{pmatrix} \middle| f_i, x_i\right] = \begin{pmatrix} \rho' f_i + \theta' x_i \\ x_i \end{pmatrix}$$

This choice minimizes the asymptotic variance

$$\sigma^2_{\epsilon_y}\left\{E(A_i A_i')\right\}^{-1}$$

which is the semi-parametric efficiency bound; see Chamberlain (1987), Newey (1990) and Belloni et al. (2012). The component of this bound corresponding to α is Σ_{FIV}. Therefore, under conditional homoskedasticity and fixed number of exogenous controls, the oracle FIV estimator is semi-parametrically efficient for α. Because our estimator attains the same variance, it also attains the semi-parametric efficiency bound.

However, in the presence of many exogenous controls, the oracle variance Σ_{FIV} is smaller than Σ, with their asymptotic ratio bounded between 0 and 1. This is not surprising because exogenous controls are partialled out within subsamples of size $n/2$, the variance of our estimator is necessarily inflated.

4 Practical recommendations

In practice, the number of instruments may be insufficient relative to the effective sample size. We therefore offer the following practical recommendations:

4.1 Interaction with many exogenous variables

Empirical applications with many instruments often also include many exogenous controls; see, for example, Angrist and Krueger (1991), Kolesár et al. (2015), and Dovì et al. (2024). A common way to expand the instrument set is to interact baseline instruments with exogenous covariates. Provided that the interaction instruments remain valid, this construction offers a natural way to increase the dimension of the instrument set. For example, Dovì et al. (2024) interact 38 baseline instruments with exogenous controls, yielding 342 instruments for 124 observations. Thus, by interacting baseline instruments with exogenous controls, it is possible to place the first stage in the overparameterized regime.

4.2 Sample size reduction

Another practical approach is to intentionally reduce the sample size via sample splitting. This approach does not waste data by discarding observations. Instead, by swapping the samples for the estimation of first-stage regression and the estimation of the target parameter α, it asymptotically avoids losing efficiency compared to full-sample estimator. After splitting, it is more plausible to turn underparametrized first-stage into overparametrized, thereby double descent phenomenon can arise for the sample-splitting estimator.

We implement this idea using the K-fold cross-fitting procedure that reuses all observations across folds and averages the resulting estimates. The algorithm is as follows.

Algorithm (Cross-fitting Ridgeless IV).

Step 1: *Sample splitting.* Randomly partition the sample $\{1, \ldots, n\}$ into K disjoint folds $\{I_k\}_{k=1}^K$. Let I_{-k} denote the sample excluding subsample I_k. Using subsample I_k, estimate β by Ridgeless regression:

$$\hat{\beta}_k = (Z_k^{\perp\prime} Z_k^{\perp})^+ Z_k^{\perp\prime} D_k^{\perp}.$$

Using subsample I_{-k}, construct the fitted value:

$$\hat{D}_{-k}^{\perp} = Z_{-k}^{\perp} \hat{\beta}_k.$$

Step 2: *Estimation of* α. Following IV estimator,

$$\hat{\alpha}_k = (\hat{D}_{-k}^{\perp\prime} D_{-k}^{\perp})^{-1} \hat{D}_{-k}^{\perp\prime} Y_{-k}^{\perp}.$$

Step 3: *Cross-fitting.* Repeat Steps 1-2 K times by swapping subsample to obtain $\hat{\alpha}_1, ..., \hat{\alpha}_K$.

The estimator for α is

$$\hat{\alpha} = \frac{1}{K}\sum_{k=1}^{K}\hat{\alpha}_k.$$

We emphasize that our procedure differs from conventional K-fold cross-fitting (e.g., Chernozhukov et al. (2018)) in the sense that conventional cross-fitting estimates the first-stage regression using I_{-k}, which contains all observations except the kth fold, and estimates the target parameter on the I_k. By contrast, we deliberately reverse these roles: we estimate the first-stage regression using only the smaller fold I_k and estimate α using its complement I_{-k}. This reversal makes overparameterization more attainable in the first stage while preserving the essential sample splitting property.

5 Simulation

In this section, we provide simulation evidence to examine the finite sample performance of our proposed method under high-dimensional IV. We also compare the results of our estimator to other IV estimators that are available in the literature.

Our simulations are based on a simple factor IV model as oracle model for data-generating process (DGP):

$$y_i = d_i\alpha + \delta' x_i + \epsilon_{y,i}$$
$$d_i = \rho' f_i + \theta' x_i + \epsilon_{d,i}$$

where $\dim(f_i) = 3$, $\dim(x_i) = 10$, and the components $(f_i, x_i, \rho, \delta, \theta)$ are drawn from standard normal distributions. The parameter of interest is $\alpha = 1$.

The error terms $(\epsilon_{y,i}, \epsilon_{d,i})$ follow a bivariate normal distribution:

$$(\epsilon_{y,i}, \epsilon_{d,i}) \sim N(0, \begin{pmatrix} 1 & \sigma \\ \sigma & 1 \end{pmatrix})$$

where $\sigma = 0.4$.

For the instruments, we generate p_0 informative instruments and $p - p_0$ noise:

$$z_{ij} = \begin{cases} \lambda_j' f_i + u_{ij}, & j = 1, ..., p_0 \\ u_{ij}, & j = p_0 + 1, ..., p \end{cases}, \quad \text{where} \lambda_j = \lambda_{j,0} \times p_0^{-\upsilon}$$

where $(\lambda_{j,0}, u_{ij})$ are independently drawn from standard normal distributions. The factor strength is governed jointly by $\upsilon \in [0, 1/2)$ and the number of informative instruments p_0:

$$\Lambda_I'\Lambda_I \asymp \psi_{n,p} \asymp p_0^{1-2\upsilon}$$

The larger values of υ correspond to weaker factor loadings. The proportion of informative instruments is controlled by κ, where $p_0 = \kappa p$. A higher κ indicates a larger share of informative instruments, leading to stronger identification.

We consider the following simulation settings: sample size $n = \{100, 200\}$, number of instruments $p = \{30, 150, 500, 1000\}$, proportions of informative instruments $\kappa = \{1, 1/2\}$, and factor strength $\upsilon = \{0, 1/4, 3/8\}$. Specifically, the case $\upsilon = 0$ corresponds to strong factors, with $\Lambda_I'\Lambda_I \asymp p_0$. The case $\upsilon = 1/4$ corresponds to semi-strong factors, with $\Lambda_I'\Lambda_I \asymp p_0^{1/2}$. The case $\upsilon = 3/8$ corresponds to relatively weak factors, with $\Lambda_I'\Lambda_I \asymp p_0^{1/4}$.

Tables 1 and 2 report the coverage probabilities and average lengths of the 95% confidence intervals for the standardized estimator $\hat{\alpha}$ across different values of p, for sample sizes $n = 100$ and $n = 200$, respectively. Overall, the Ridgeless estimator performs well across almost all configurations, delivering coverage rates close to the nominal level while maintaining stable average lengths as p increases, which is consistent with our asymptotic theory. The only exception arises in the case $\upsilon = 3/8$, $\kappa = 1/2$, and $p = 1000$, where the coverage probability slightly deviates from the nominal level. This is not surprising, since Ridgeless estimator requires a moderate amount of informative signal to control the bias. Under relatively weak factors and when only half of the instruments are informative, the condition $p = o(\psi_{n,p} n)$ in Assumption 4 (ii) may fail to hold. Figure 1 presents histograms of the standardized estimates of the structural parameter α, together with the standard normal density, for the case $n = 100$. Consistent with the evidence in Table 1, the empirical distributions across all DGPs provide a good approximation to the standard normal distribution.

For comparison, in addition to the Ridgeless estimator, we consider four alternative estimators: Post-Lasso, RJIVE, PCA, oracle FIV. To accommodate exogenous controls, all estimators are implemented using variables partialled out with respect to the exogenous controls. The PCA estimator selects the number of factors according to Bai and Ng (2002). The oracle FIV serves as an infeasible benchmark that directly uses the true latent factors as instruments in 2SLS. To implement regularization methods, we use 10 fold cross validation to choose the penalty parameter for Post-Lasso and RJIVE estimators, respectively.

Figure 2 reports the MSEs, averaged over 100 replications, as the total number of instruments p increases, with sample size $n = 100$. The maximum number of informative instruments is $p_0 \in \{200, 1000\}$, which is indicated by the vertical dashed line in each panel. When $p \leq p_0$, all the instruments are informative. When $p > p_0$, the additional $p - p_0$ instruments are noise instruments drawn from a standard normal distribution. Thus, as p increases beyond p_0, the factor signal is increasingly diluted across the full set of instruments. We also consider three levels of factor strength for the informative instruments, $\upsilon \in \{0, 1/4, 3/8\}$.

Figure 2 demonstrates that the Ridgeless estimator exhibits the double descent phe-

nomenon across all cases. Specifically, when p is below the interpolation threshold, the usual bias-variance trade-off leads to an initial decrease (first descent) followed by an increase in the MSE. Once p exceeds the interpolation threshold, the MSE decreases again (second descent). This pattern highlights that Ridgeless estimator continues to benefit from variance reduction as additional instruments in the overparametrized regime. Under the (relatively) strong factor case ($\upsilon \in \{0, 1/4\}$), when the number of instruments is sufficiently large, the MSE of Ridgeless estimator becomes close to that of the oracle estimator, regardless of whether the additional instruments are informative or not. Conversely, under the relatively weak factor case ($\upsilon = 3/8$), Ridgeless estimation exhibits a U-shaped MSE curve in the overparametrized regime when $p_0 = 200$. It attains its minimum near $p = p_0$ and rises as p grows further. As implied by our theory, moderately informative signals cause the rising bias to dominate the decaying variance.

By contrast, the Post-Lasso estimator performs the worst among all estimators once many instruments are introduced. This is expected, as the underlying factor structure induces a non-sparse first-stage signal, violating the sparsity assumption required by sparsity-based methods. While RJIVE outperforms Post-Lasso when the number of instruments is relatively small, its performance becomes almost identical to Post-Lasso as more instruments are added. This result is not surprising because JIVE-type estimators that rely on partialled-out variables still suffer from bias that grows with the number of instruments.

PCA performs the best among all feasible estimators in the strong factor case ($\upsilon = 0$) and is nearly indistinguishable from the Oracle FIV, regardless of whether p is small or large. This occurs because PCA consistently estimates the latent factors when the factors are strong. In the semi-strong case ($\upsilon = 1/4$), however, the performance of PCA deteriorates as additional noise instruments are introduced (Figure 2 (c)), since the added noise dilutes the factor strength and thereby contaminates the PCA estimates. Under relatively weak factors ($\upsilon = 3/8$), PCA performs poorly even when p is small. Overall, the Ridgeless estimator remains robust across all cases and attains the smallest MSE among feasible estimators, provided a sufficiently large number of instruments is included.

		Coverage				Average Length			
υ	κ	$p = 30$	$p = 150$	$p = 500$	$p = 1000$	$p = 30$	$p = 150$	$p = 500$	$p = 1000$
0	1	0.9540	0.9480	0.9420	0.9500	3.8471	3.9185	3.9157	3.8937
	1/2	0.9510	0.9530	0.9520	0.9460	4.0443	3.9352	3.9342	3.9355
1/4	1	0.9490	0.9490	0.9510	0.9520	3.9695	3.7866	4.0543	3.7491
	1/2	0.9560	0.9440	0.9520	0.9440	3.8228	3.9724	3.9824	4.0566
3/8	1	0.9540	0.9580	0.9560	0.9600	3.7709	3.7339	3.6242	3.5902
	1/2	0.9570	0.9590	0.9570	0.9620	3.7482	3.5962	3.6342	3.4606

Table 1: Simulated coverage probabilities and average lengths of 95% confidence intervals for the standardized estimator $\hat{\alpha}$ across different values of p, factor strength υ, and proportion of informative instruments κ. Notes: the sample size is $n = 100$. Based on simulations with 1000 repetitions.

		Coverage				Average Length			
υ	κ	$p = 30$	$p = 150$	$p = 500$	$p = 1000$	$p = 30$	$p = 150$	$p = 500$	$p = 1000$
0	1	0.9450	0.9550	0.9540	0.9460	3.9655	3.9205	3.9513	3.9789
	1/2	0.9470	0.9430	0.9510	0.9510	3.8723	3.9249	3.7158	3.8198
1/4	1	0.9450	0.9480	0.9490	0.9520	3.9987	3.9787	3.8605	3.8408
	1/2	0.9520	0.9470	0.9540	0.9510	3.8615	3.9799	3.7970	4.0084
3/8	1	0.9460	0.9490	0.9530	0.9520	3.9445	3.9974	3.8456	3.7877
	1/2	0.9490	0.9540	0.9530	0.9560	3.8449	3.8933	3.7822	3.7823

Table 2: Simulated coverage probabilities and average lengths of 95% confidence intervals for the standardized estimator $\hat{\alpha}$ across different values of p, factor strength υ, and proportion of informative instruments κ. Notes: the sample size is $n = 200$. Based on simulations with 1000 repetitions.

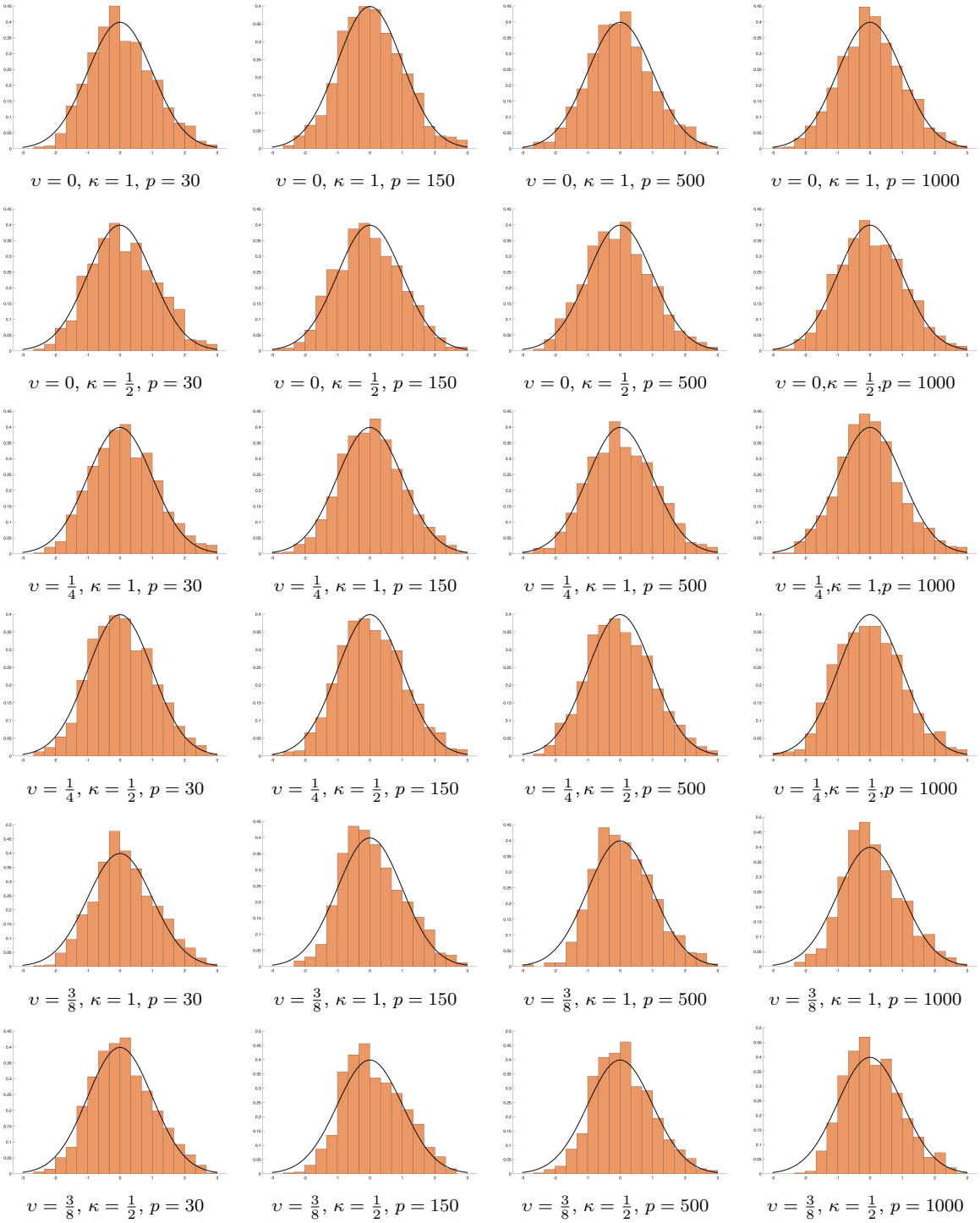
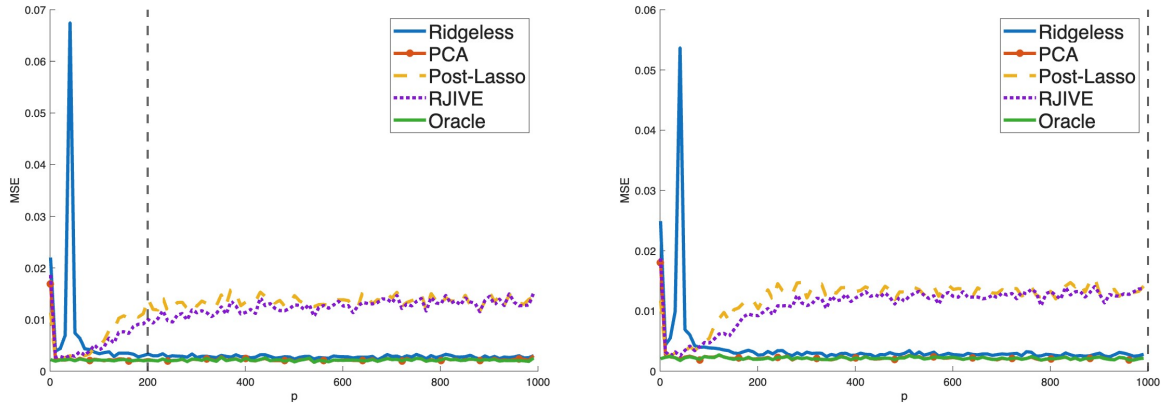
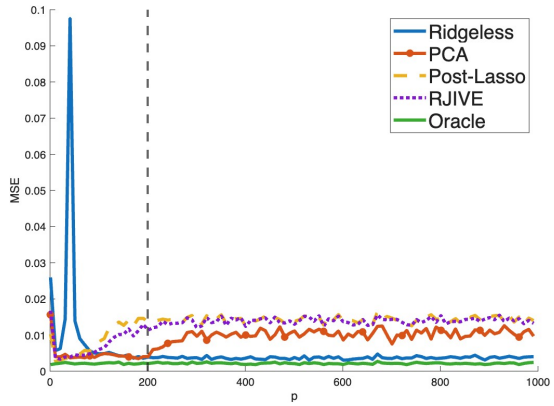


Figure 1: Histograms of the standardized estimates of the average treatment effect, $\sqrt{n}(\hat{\alpha} - \alpha)$, divided by the estimated asymptotic standard deviation. Each subplot corresponds to a different factor strength υ, proportion of informative instruments κ, and number of instruments p, respectively. All results are based on 1,000 simulation replications and sample size $n = 100$.

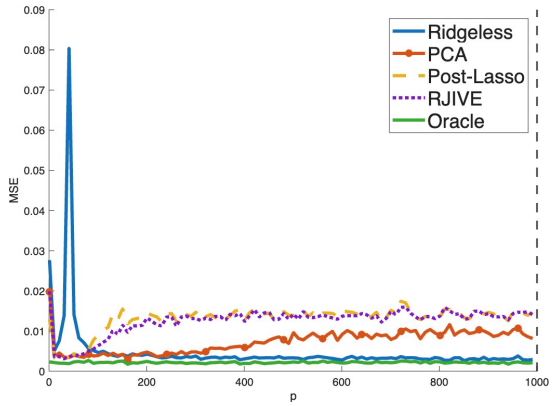


(a) $p_0 = 200$, $\upsilon = 0$

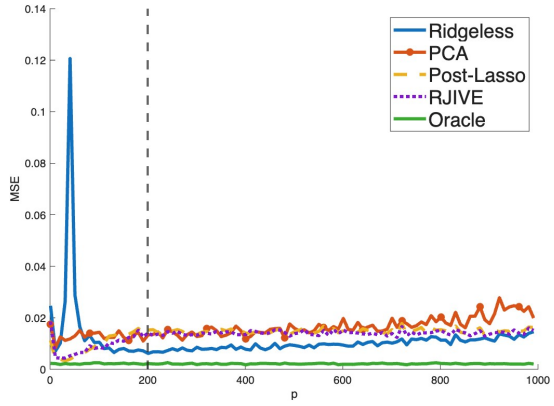
(b) $p_0 = 1000$, $\upsilon = 0$



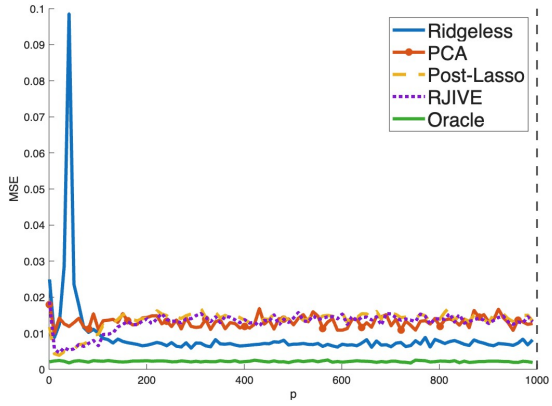
(c) $p_0 = 200$, $\upsilon = 1/4$



(d) $p_0 = 1000$, $\upsilon = 1/4$



(e) $p_0 = 200$, $\upsilon = 3/8$



(f) $p_0 = 1000$, $\upsilon = 3/8$

Figure 2: MSEs of $\hat{\alpha}$ as the number of instruments p increases. All estimation methods use the full set of instruments. The vertical dashed line indicates the maximum number of informative instruments p_0. Each subplot corresponds to a different factor strength υ and maximum number of informative instruments p_0, respectively. All results are based on 100 simulation replications and sample size $n = 100$.

6 Empirical Application

In this section, we present an empirical application to illustrate the usefulness of the proposed sample-splitting Ridgeless regression inference procedure. Following Card (2009), Goldsmith-Pinkham et al. (2020), and Dovì et al. (2024), we estimate the (negative) inverse elasticity of substitution between immigrant and native workers in the United States separately for high school and college workers. The analysis uses cross-sectional data from 2000 covering 124 cities.[5] Specifically, we consider the following model:

$$y_{is} = d_{is}\alpha_s + x_i'\delta + \epsilon_{y,is}, \tag{10}$$

where y_{is} denotes the difference in residual log wages between immigrants and natives within skill group $s \in \{h, c\}$ (high school or college equivalently) in city $i = 1, ..., 124$. Residual log wages refer to log wages adjusted for education, age, gender, race, and ethnicity, as discussed in Card (2009). The parameter of interest, α_s, can be interpreted as the negative of the inverse elasticity of substitution between immigrants and natives within skill group s. The vector $x_i \in \mathbb{R}^9$ represents control variables, which include log of city size, college shares, manufacturing shares in both 1980 and 1990, mean wage residuals for all natives and all immigrants in 1980, and a constant.

The endogenous variables d_{is} is defined as the log ratio of immigrant to native hours worked in skill group s in city i, including both male and female workers. As discussed in Card (2009), d_{is} is potentially endogenous because unobserved city-specific factors may shift the relative demand for immigrant workers, leading to higher relative wages and higher relative employment shares, thereby confounding the estimation of the inverse elasticity of substitution.

The identification strategy of Card (2009) is to use the ratio of the number of immigrants from country l in city i to the total number of immigrants from the foreign country l in the United States as instruments. The rationale for these instruments is that existing immigrant enclaves are likely to attract additional immigrant labor through social and cultural channels unrelated to labor market outcomes. We consider two instrument sets in our analysis. The first follows Card (2009)'s original specification that includes 38 instruments from different countries of origin of the immigrants. Following from Dovì et al. (2024), we consider the second setup where these 38 original instruments are interacted with 9 controls, yielding a total of 342 instruments.

Tables 3 and 4 report estimates of α_s obtained using different estimators and instrument sets for high school and college workers, respectively. The penalty parameters for Post-Lasso and RJIVE are selected using 10-fold cross-validation. For PCA, the number of factors is

[5]The dataset is available from the replication package of Goldsmith-Pinkham et al. (2020).

chosen using the information criteria of Bai and Ng (2002). The criteria select $J = 10$ for both instrument sets and both skill groups. This suggests that the instruments do not possess a strong factor structure.

Tables 3 and 4 show that, for both high-school and college workers, the Post-Lasso estimates have shifted toward the OLS estimates when the larger instrument set is used. In contrast, the PCA, RJIVE, and Ridgeless estimates remain stable across the original set of 38 instruments and the fully interacted set of 342 instruments. This pattern suggests that the first-stage relationship in the small instrument set may be sparse, which is partly consistent with the findings of Dovì et al. (2024), who observe that the first-stage relationship for college workers is highly sparse in the small instrument set.

The Ridgeless estimator has larger standard errors than the other estimators when the smaller instrument set is used. This pattern is consistent with the possibility that the first-stage signal is relatively sparse in the original instrument set. In contrast, when the fully interacted instrument set is used, Ridgeless estimator yields smaller standard errors than Post-Lasso, PCA, and RJIVE. This suggests that interacting the instruments creates a large number of small signal variables, making the first-stage relationship dense and thereby reducing the estimated variance of Ridgeless estimator. On the other hand, PCA and RJIVE produce relatively stable standard errors across the two instrument sets. Overall, these results suggest that Ridgeless regression may provide a useful complement to existing approaches when dealing with non-sparse high-dimensional IV.

Instruments	**OLS**	**Post-Lasso**	**PCA**	**RJIVE**	**Ridgeless**
38	-0.0259	-0.0394	-0.0300	-0.0308	-0.0311
	(0.0077)	(0.0094)	(0.0105)	(0.0113)	(0.0139)
342	-0.0259	-0.0329	-0.0304	-0.0316	-0.0320
	(0.0077)	(0.0088)	(0.0101)	(0.0116)	(0.0076)

Table 3: Estimates of the inverse elasticity of substitution between immigrants and natives in the United States for high school workers using different estimators and instrument sets. Standard errors are reported in parentheses. For comparison, the 2SLS estimate (standard error) using 38 instruments is -0.0362 (0.0077).

Instruments	OLS	Post-Lasso	PCA	RJIVE	Ridgeless
38	-0.0578	-0.0753	-0.0668	-0.0702	-0.0677
	(0.0110)	(0.0138)	(0.0138)	(0.0145)	(0.0197)
342	-0.0578	-0.0610	-0.0663	-0.0704	-0.0669
	(0.0110)	(0.0131)	(0.0129)	(0.0147)	(0.0118)

Table 4: Estimates of the inverse elasticity of substitution between immigrant and natives in the United States for college workers using different estimators and instrument sets. Standard errors are reported in parentheses. For comparison, the 2SLS estimate (standard error) using 38 instruments is -0.0623 (0.0111).

7 Conclusion

This paper revisits the Factor IV model from the perspective of high-dimensional IV regression. We show that an underlying factor structure implies a dense first-stage coefficient vector with many weak instruments, providing a theoretical justification for recent empirical evidence against sparsity. Building on this characterization, we propose a sample-splitting Ridgeless estimator that accommodates (very) many weak instruments and many exogenous controls. We establish its consistency and asymptotic normality, allowing valid inference on the causal parameter. With a sufficiently large number of instruments, the estimator exploits overparameterization and its asymptotic variance approaches oracle efficiency. Simulation and empirical results demonstrate that the proposed estimator performs well relative to other high-dimensional IV estimators.

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, *59*(2), 391–425.

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505.

Angrist, J. D., & Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, *40*(S1), S97–S140.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.

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

Bai, J., & Ng, S. (2023). Approximate factor models with weaker loadings. *Journal of Econometrics*, *235*(2), 1893–1916.

Bai, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica*, *9*(3), 611–662.

Bai, Z. D., & Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, *26*(1), 316–345.

Bai, Z., & Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices* (Vol. 20). Springer.

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*(1), 71–92.

Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, *117*(48), 30063–30070.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, *62*(3), 657–681.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854.

Belkin, M., Hsu, D., & Xu, J. (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, *2*(4), 1167–1180.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450.

Card, D. (2009). Immigration and inequality. *American Economic Review*, *99*(2), 1–21.

Carrasco, M., & Doukali, M. (2017). Efficient estimation using regularized jackknife iv estimator. *Annals of Economics and Statistics*, (128), 109–149.

Carrasco, M., & Tchuente, G. (2015). Regularized liml for many instruments [High Dimensional Problems in Econometrics]. *Journal of Econometrics*, *186*(2), 427–442.

Cattaneo, M. D., Jansson, M., & Newey, W. K. (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, *113*(523), 1350–1361.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, *34*(3), 305–334.

Chao, J. C., & Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, *73*(5), 1673–1692.

Chao, J. C., & Swanson, N. R. (2022). Consistent estimation, variable selection, and forecasting in favar models. *Variable Selection, and Forecasting in FAVAR Models (February 7, 2022)*.

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K., & Woutersen, T. (2012). Asymptotic distribution of jive in aheteroskedastic iv regression with manyinstruments. *Econometric Theory*, *28*(1), 42–86.

Chao, J. C., Swanson, N. R., & Woutersen, T. (2023). Jackknife estimation of a cluster-sample iv regression model with many weak instruments. *Journal of Econometrics*, *235*(2), 1747–1769.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Dovì, M.-S., Kock, A. B., & Mavroeidis, S. (2024). A ridge-regularized jackknifed anderson-rubin test. *Journal of Business & Economic Statistics*, *42*(3), 1083–1094.

Fan, J., Ke, Y., & Wang, K. (2020). Factor-adjusted regularized model selection [Annals Issue in honor of George Tiao: Statistical Learning for Dependent Data]. *Journal of Econometrics*, *216*(1), 71–85.

Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, *89*(5), 2409–2437.

Giglio, S., Xiu, D., & Zhang, D. (2025). Test assets and weak factors. *The Journal of Finance*, *80*(1), 259–319.

Goldsmith-Pinkham, P., Sorkin, I., & Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, *110*(8), 2586–2624.

Hansen, C., Hausman, J., & Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, *26*(4), 398–422.

Hansen, C., & Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, *182*(2), 290–308.

Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, *50*(2), 949.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C., & Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, *3*(2), 211–255.

Hsiao, C., Steve Ching, H., & Ki Wan, S. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, *27*(5), 705–740.

Kapetanios, G., & Marcellino, M. (2010). Factor-gmm estimation with large sets of possibly weak instruments [The Fifth Special Issue on Computational Econometrics]. *Computational Statistics & Data Analysis*, *54*(11), 2655–2675.

Kelly, B., Malamud, S., & Zhou, K. (2024). The virtue of complexity in return prediction. *The Journal of Finance*, *79*(1), 459–503.

Kolesár, M. (2013, November). *Estimation in an instrumental variables model with treatment effect heterogeneity* (Working Papers No. 2013-2). Princeton University. Economics Department.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., & Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, *33*(4), 474–484.

Kolesár, M., Müller, U. K., & Roelsgaard, S. T. (2026). The fragility of sparsity.

Li, K. T., & Bell, D. R. (2017). Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics*, *197*(1), 65–75.

Liao, Y., Ma, X., Neuhierl, A., & Shi, Z. (2023). Benign overfitting in economic forecasting via noise regularization. *arXiv e-prints*, arXiv–2312.

Mei, S., & Montanari, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, *75*(4), 667–766.

Mikusheva, A., & Sun, L. (2021). Inference with many weak instruments. *The Review of Economic Studies*, *89*(5), 2663–2686.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, *58*(4), 809–837.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, *168*(2), 244–258.

Spiess, J., imbens guido, g., & Venugopal, A. (2023). Double and single descent in causal inference with an application to high-dimensional synthetic control. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (pp. 63642–63659, Vol. 36). Curran Associates, Inc.

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

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

Stock, J. H., & Yogo, M. (2002). Testing for weak instruments in linear iv regression.

A Proofs for Section 3

Throughout the proof, we use C and C' to denote a generic positive constant. Let $\|A\|$ denote the ℓ_2 norm if A is a vector, and the operator norm if A is a matrix; for a matrix A, let $\|A\|_F$ and $\|A\|_*$ denote its Frobenius norm and nuclear norm, respectively. Given a function f defined on X, we let $\|f\|_\infty = \sup_{x\in X} |f(x)|$ denote its infinity norm. Let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ denote the minimum and maximum nonzero singular values of of a matrix A, respectively. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a square matrix A. We use A_{ij} to denote the (i,j)-th element in a matrix A. For two sequences $a_{n,p}$ and $b_{n,p}$, we denote $a_{n,p} \ll b_{n,p}$ (or $b_{n,p} \gg a_{n,p}$) if $a_{n,p} = o(b_{n,p})$, and $a_{n,p} \asymp b_{n,p}$ if $a_{n,p} = O(b_{n,p})$ and $b_{n,p} = O(a_{n,p})$. For random sequences, o and O are denoted as o_P and O_P, respectively. We partition the sample $\{1, ..., n\}$ into two disjoint subsets I_1 and I_2 with equal sample sizes, i.e, $n_1 = n_2 = n/2$.

This appendix contains the proofs of the main theoretical results and the technical lemmas used in their derivations.

Contents

A.1 Proof of Theorem 1 36
A.2 Technical lemmas 37
A.3 Proof of Theorem 2 49
A.4 Asymptotic normality 66
A.4.1 Proof of Theorem 3 72
A.5 Consistent estimator for the asymptotic variance Σ 73
A.5.1 Proof of Theorem 4 78
A.6 Proof of Theorem 5 78

A.1 Proof of Theorem 1

Proof of Theorem 1. Consider linear projection of d_i on (z_i, x_i), we optimize:

$$\min_{\beta,\xi} E((d_i - z_i'\beta - x_i'\xi)^2)$$

Then we have linear projection $L(d_i|z_i, x_i) = z_i'\beta + x_i'\xi$ with

$$\begin{pmatrix}\beta\\ \xi\end{pmatrix} = \begin{pmatrix}E(z_iz_i') & E(z_ix_i')\\ E(x_iz_i') & E(x_ix_i')\end{pmatrix}^{-1}\begin{pmatrix}E(z_id_i)\\ E(x_id_i)\end{pmatrix}$$

The oracle model (2) implies $E(z_iz_i') = \Sigma_Z = \Lambda\Sigma_f\Lambda' + \sigma_u^2 I_p$, $E(z_ix_i') = \Lambda E(f_ix_i') = \Lambda\Sigma_{fx}$, $E(x_ix_i') = \Sigma_x$, $E(z_id_i) = \Lambda\Sigma_f\rho + \Lambda\Sigma_{fx}\theta$, $E(x_id_i) = \Sigma_{xf}\rho + \Sigma_x\theta$. By solving the linear

equation, we have

$$\begin{pmatrix}\beta\\ \xi\end{pmatrix} = \begin{pmatrix}E(z_i^{\perp}z_i^{\perp\prime})^{-1}E(z_i^{\perp}d_i^{\perp})\\ E(x_ix_i')^{-1}E(x_i(d_i - z_i'\beta))\end{pmatrix} = \begin{pmatrix}(\Lambda\Sigma_{f|x}\Lambda' + \sigma_u^2 I_p)^{-1}\Lambda\Sigma_{f|x}\rho\\ \theta + \Sigma_x^{-1}\Sigma_{xf}(\rho - \Lambda'\beta)\end{pmatrix}$$

where $z_i^{\perp} = z_i - E(z_ix_i')E(x_ix_i')^{-1}x_i$, $d_i^{\perp} = d_i - E(d_ix_i')E(x_ix_i')^{-1}x_i$ and $\Sigma_{f|x} = \Sigma_f - \Sigma_{fx}\Sigma_x^{-1}\Sigma_{xf}$.

Next, by Woodbury matrix identity

$$(\Lambda\Sigma_{f|x}\Lambda' + \sigma_u^2)^{-1} = \sigma_u^{-2}I_p - \sigma_u^{-4}\Lambda(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\Lambda'$$

By substituting it into β, we have

$$\begin{aligned}\beta &= (\Lambda\Sigma_{f|x}\Lambda' + \sigma_u^2 I_p)^{-1}\Lambda\Sigma_{f|x}\rho\\ &= (\sigma_u^{-2} - \sigma_u^{-4}\Lambda(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\Lambda')\Lambda\Sigma_{f|x}\rho\\ &= \sigma_u^{-2}\Lambda(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\Sigma_{f|x}^{-1}\Sigma_{f|x}\rho\\ &= \sigma_u^{-2}\Lambda(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho\end{aligned} \tag{11}$$

Thus,

$$\begin{aligned}\|\beta\| &\le \sigma_u^{-2}\|\Lambda\|\|\rho\|\sigma_{\min}(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\\ &\le O(1)\|\Lambda\|\sigma_{\min}(\Lambda'\Lambda)^{-1}\\ &= O(\psi_{n,p}^{-1/2})\end{aligned}$$

□

A.2 Technical lemmas

Lemma A.1. *(Lemma 1, Bekker (1994)). Consider the quadratic form $Q = U'PU$, where $U = M + V$, $M \in \mathbb{R}^{n\times p}$ is deterministic, $V \in \mathbb{R}^{n\times p}$, and $V = (v_1, ..., v_n)'$ with $v_i \sim N(0, \Omega)$ i.i.d. P is an idempotent and symmetric matrix and its rank $r = \mathrm{tr}(P)$. Let the matrices U, M, V, P, Ω consist of columns $u_i, m_i, v_i, p_i, \omega_i$ respectively, and let $d = \mathrm{diag}(P)$. Then*

$$\begin{aligned}E(u_i'Pu_j) &= m_i'Pm_j + r\omega_{ij}\\ \mathrm{Var}(u_i'Pu_j) &= \omega_{ii}m_j'Pm_j + \omega_{jj}m_i'Pm_i + 2\omega_{ij}m_i'Pm_j + r(\omega_{ii}\omega_{jj} + \omega_{ij}^2)\\ &\quad + d'd(E(v_{1i}^2v_{1j}^2) - \omega_{ii}\omega_{jj} - 2\omega_{ij}^2) + 2d'Pm_i(E(v_{1i}v_{1j}^2))\\ &\quad + 2d'Pm_jE(v_{1i}^2v_{1j})\end{aligned}$$

In particular, if the distribution of v_i is normal, the last two lines of the variance expression equal zero.

In the next lemma, we focus on the limit of empirical spectral distribution of $U'M_XU$ via random matrix theory. Let $\lambda_1(A) \geq ... \geq \lambda_n(A)$ be the eigenvalues of a $n \times n$ matrix A. We define the empirical spectral distribution as following:

$$F_A(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\lambda_i(A) \leq x\}$$

Lemma A.2. *As $n, p \to \infty$ such that $\gamma_n := \frac{p}{n-r} \to \gamma \in (0, \infty) \setminus \{1\}$*

$$F_{U'M_XU/((n-r)\sigma_u^2)} \;\Rightarrow\; F_\gamma \;\; a.s.$$

with support $[a_\gamma, b_\gamma]$, where $a_\gamma = (1-\sqrt{\gamma})^2$ and $b_\gamma = (1+\sqrt{\gamma})^2$.

Proof. Since M_X is the orthogonal projector onto $col(X)^\perp$ with rank $n-r$, There exist an orthogonal matrix $Q = [Q_1, Q_2]$ with $Q_1 \in \mathbb{R}^{n\times r}, Q_2 \in \mathbb{R}^{n\times(n-r)}$. Then,

$$M_X = Q \begin{pmatrix} 0_{r\times r} & 0 \\ 0 & I_{n-r} \end{pmatrix} Q' = Q_2 Q_2'$$

Thus,

$$U'M_XU = U'Q_2Q_2'U = U_2'U_2$$

where U_2 is a $(n-r)\times p$ matrix.

Since M_X has $n-r$ eigenvalues equal one and zero eigenvalues equal zero, its empirical spectral distribution follows

$$F_{M_X} = \frac{r}{n}\mathbf{1}_{[0,\infty)} + \frac{n-r}{n}\mathbf{1}_{[1,\infty)} \to H = \tau\mathbf{1}_{[0,\infty)} + (1-\tau)\mathbf{1}_{[1,\infty)}$$

where the convergence follows from Assumption 3.

We first discuss the limit of empirical spectral distribution of $U_2U_2'/(p\sigma_u^2)$. Let $B_n = M_XUU'M_X/(p\sigma_u^2)$ and $m_B(z)$ denote the limiting Stieltjes transform of B_n. Using Bai–Silverstein's equation, we have

$$\begin{aligned} m_B(z) &= \int \frac{1}{s(1-c-czm_B(z))-z} dH(s) \\ &= \frac{-\tau}{z} + \frac{(1-\tau)}{1-c-czm_B(z)-z} \end{aligned} \tag{12}$$

Next, since Q is orthogonal matrix, $Q'B_nQ$ has same eigenvalues as B_n. Also,

$$Q'B_nQ = \begin{pmatrix} \frac{1}{p\sigma_u^2}U_2U_2', 0 \\ 0, 0_r \end{pmatrix}$$

which indicates the eigenvalues of B_n is exactly equal to $(n-r)$ eigenvalues of $\frac{1}{p\sigma_u^2}U_2U_2'$ plus r additional zero eigenvalues. Then we can show the relationship between empirical spectral distribution of B_n and $U_2U_2'/(p\sigma_u^2)$:

$$F_{B_n} = \frac{r}{n}\mathbf{1}_{[0,\infty)} + \frac{n-r}{n}F_{U_2U_2'/(p\sigma_u^2)}$$

Then its Stieltjes transform of F_{B_n} is

$$\begin{aligned} m_{B_n}(z) &= \int \frac{1}{x-z}dF_{B_n}(x) \\ &= -\frac{r}{nz} + \frac{n-r}{n}m_{U_2U_2'/(p\sigma_u^2)}(z) \end{aligned}$$

By taking the limit,

$$m_B(z) = \frac{-\tau}{z} + (1-\tau)m_C(z) \tag{13}$$

By combing (12) and (13), we have

$$m_C(z) = \frac{1}{1-\frac{1}{\gamma}-z-\frac{z}{\gamma}m_C(z)}$$

which can be recognized as the ordinary Marchenko–Pastur Stieltjes equation with parameter $\frac{1}{\gamma}$. Then by the uniqueness of Stieltjes transform

$$F_{U_2U_2'/(p\sigma_u^2)} \Rightarrow F_{1/\gamma} \quad \text{a.s.}$$

Lastly, we want to show $F_{U_2'U_2/((n-r)\sigma_u^2)} \Rightarrow F_\gamma$ a.s.. Let F_G be the limit of $F_{U_2'U_2/((n-r)\sigma_u^2)}$. Then since $U_2'U_2$ and U_2U_2' have $|p-(n-r)|$ different eigenvalues equal to zero, we have

$$F_{U_2'U_2/((n-r)\sigma_u^2)}(x) = (1-\frac{1}{\gamma_n})\mathbf{1}_{[0,\infty)} + \frac{1}{\gamma_n}F_{U_2U_2'/(p\sigma_u^2)}(x/\gamma_n)$$

By taking the limit,

$$F_G(x) = (1-\frac{1}{\gamma})\mathbf{1}_{[0,\infty)} + \frac{1}{\gamma}F_{1/\gamma}(x/\gamma)$$

which is the reciprocal Marchenko–Pastur identity. Thus,

$$F_{U_2'U_2/((n-r)\sigma_u^2)} \Rightarrow F_\gamma \text{ a.s.}$$

with support $[a_\gamma, b_\gamma]$, where $a_\gamma = (1-\sqrt{\gamma})^2$ and $b_\gamma = (1+\sqrt{\gamma})^2$.

□

Lemma A.3. *(i)* $\|U'M_XU\| = O_P((n-r)+p)$

(ii) $\|\frac{1}{n}Z'M_XZ\| = O_P(\psi_{n,p})$

(iii) $\|U'M_X\boldsymbol{\epsilon}_y\| = O_P(\sqrt{(n-r)+p})$, $\|U'M_X\boldsymbol{\epsilon}_d\| = O_P(\sqrt{(n-r)+p})$,
$\|\frac{1}{\sqrt{n}}Z'M_X\boldsymbol{\epsilon}_y\| = O_P(\sqrt{\psi_{n,p}})$, $\|\frac{1}{\sqrt{n}}Z'M_X\boldsymbol{\epsilon}_d\| = O_P(\sqrt{\psi_{n,p}})$,
$\|\frac{1}{\sqrt{n}}F'M_X\boldsymbol{\epsilon}_y\| = O_P(1)$, $\|\frac{1}{\sqrt{n}}F'M_X\boldsymbol{\epsilon}_d\| = O_P(1)$

(iv) $\|U\beta\|^2 = O_P(\frac{n}{\psi_{n,p}})$, $\|U\Lambda(\Lambda'\Lambda)^{-1}\rho\|^2 = O_P(\frac{n}{\psi_{n,p}})$

(v) $\|Z'M_XZ(Z'M_XZ)^+\| \leq 1, \quad a.s.$

Proof. (i) Let $\gamma = \lim \frac{p}{n-r}$. Following from Lemma A.2, we know

$$F_{U'M_XU/((n-r)\sigma_u^2)} \;\Rightarrow\; F_\gamma \text{ a.s.}$$

with support $[a_\gamma, b_\gamma]$, where $a_\gamma = (1-\sqrt{\gamma})^2$ and $b_\gamma = (1+\sqrt{\gamma})^2$.

Since M_X is a projection matrix, $\|M_X\| = 1$, which is the largest support in the limit of empirical spectral distribution of F_{M_X}. Therefore, by the corollary of Bai and Silverstein (1998), $\|U'M_XU/((n-r)\sigma_u^2)\|$ converges almost surely to the largest number in the support of F_γ:

$$\frac{1}{n-r}\|U'M_XU\| \to \sigma_u^2(1+\sqrt{\gamma})^2$$

Thus,

$$\|U'M_XU\| = O_P((n-r)+p)$$

(ii) Since Z follows factor model, after partialling out X,

$$M_XZ = M_XF\Lambda' + M_XU$$

where Z and U are $n \times p$ matrix, F is $n \times J$ matrix, and Λ is $p \times J$ matrix. Then

$$\frac{1}{n}Z'M_XZ = (1-\frac{r}{n})\Lambda\Sigma_{f|x}\Lambda' + \frac{1}{n}H + \frac{1}{n}U'M_XU \tag{14}$$

where $H = -\Lambda(F'M_XF - (n-r)\Sigma_{f|x})\Lambda' + \Lambda F'M_XU + U'M_XF\Lambda'$.

For the first term in H,

$$\|\Lambda(\frac{F'M_XF}{n} - (1-\frac{r}{n})\Sigma_{f|x})\Lambda'\| \leq \sigma_{\max}(\Lambda\Lambda')\|\frac{F'M_XF}{n} - (1-\frac{r}{n})\Sigma_{f|x}\| = O_P(\frac{\psi_{n,p}}{\sqrt{n}})$$

And the cross term

$$\begin{aligned}\|\frac{1}{n}U'M_XF\Lambda'\| &\leq \frac{1}{n}\sqrt{\sigma_{\max}(\Lambda F'F\Lambda')\sigma_{\max}(U'M_XU)} \\ &= \frac{1}{n}\sqrt{\sigma_{\max}(\Lambda'\Lambda)O_P(n)\sigma_{\max}(U'M_XU)} \\ &= O_P(\sqrt{\psi_{n,p}} + \sqrt{\frac{p\psi_{n,p}}{n}})\end{aligned}$$

Then we have

$$\|\frac{1}{n}H\| = O_P(\frac{\psi_{n,p}}{\sqrt{n}} + \sqrt{\psi_{n,p}} + \sqrt{\frac{p\psi_{n,p}}{n}}) \tag{15}$$

Also,

$$\begin{aligned}\|(1-\frac{r}{n})\Lambda\Sigma_{f|x}\Lambda'\| &\leq O(1)\sigma_{\max}(\Lambda\Lambda') = O_P(\psi_{n,p}) \\ \|\frac{1}{n}U'M_XU\| &= O_P(1+\frac{p}{n})\end{aligned}$$

From Assumption 4 (ii) that $p = o(\psi_{n,p}n)$, $(1-\frac{r}{n})\Lambda\Sigma_{f|x}\Lambda'$ is the dominant term such that

$$\frac{1}{n}Z'M_XZ = O_P(\psi_{n,p})$$

(iii) From Assumption 2 (i), the conditional expectation is $E(U'M_X\boldsymbol{\epsilon_y}|U,X) = 0$. For the conditional variance, from (i) we have

$$\text{Var}(U'M_X\boldsymbol{\epsilon}_y|Z,X) = \sigma^2_{\epsilon_y}U'M_XU = O_P((n-r)+p)$$

Therefore,

$$\|U'M_X\boldsymbol{\epsilon_y}\| = O_P(\sqrt{(n-r)+p})$$

Similarly,

$$\|U'M_X\boldsymbol{\epsilon}_d\| = O_P(\sqrt{(n-r)+p})$$

By similar arguments, it follows from Assumption 2(i) and (ii) that

$$\|\frac{1}{\sqrt{n}}Z'M_X\boldsymbol{\epsilon}_y\| = O_P(\sqrt{\psi_{n,p}}), \quad \|\frac{1}{\sqrt{n}}Z'M_X\boldsymbol{\epsilon}_d\| = O_P(\sqrt{\psi_{n,p}})$$

By similar arguments, it follows from Assumption 1 (iii) Assumption 2 (i) that

$$\|\frac{1}{\sqrt{n}}F'M_X\boldsymbol{\epsilon}_y\| = O_P(1), \quad \|\frac{1}{\sqrt{n}}F'M_X\boldsymbol{\epsilon}_d\| = O_P(1)$$

(iv)

$$\begin{aligned} E\|U\beta\|^2 &= E(\sum_{i=1}^{n}(\sum_{j=1}^{p}u_{ij}\beta_j)^2) \\ &= \sum_{i=1}^{n}(\sum_{j=1}^{p}E(u_{ij}^2)\beta_j^2) \\ &= n\sigma_u^2\|\beta\|^2 = O_P(\frac{n}{\psi_{n,p}}) \end{aligned}$$

where the second equality follows from Assumption 2 (i).

Thus,

$$\|U\beta\|^2 = O_P(\frac{n}{\psi_{n,p}})$$

By similar argument,

$$\|U\Lambda(\Lambda'\Lambda)^{-1}\rho\|^2 = O_P(\frac{n}{\psi_{n,p}})$$

(v) Define $P_{Z^\perp} = Z'M_XZ(Z'M_XZ)^+$. For the case $n-r > p$, $P_{Z^\perp} = I_p$. For the case $p > n-r$, we apply the singular value decomposition $M_XZ = U_nS_nV_n'$ where U_n is $n \times n-r$ matrix, S_n is $n-r \times n-r$ matrix, V_n is a $p \times n-r$ matrix. Then $Z'M_XZ = V_nS_n^2V_n'$ and $(Z'M_XZ)^+ = V_nS_n^{-2}V_n'$ such that

$$P_{Z^\perp} = Z'M_XZ(Z'M_XZ)^+ = V_nV_n'$$

Hence, for both cases, we can observe that $P_{Z^\perp}$ is symmetric and idempotent such that all eigenvalues of $P_{Z^\perp}$ are 0 or 1. Then,

$$\|P_{Z^\perp}\| \leq 1, \quad a.s.$$

□

Lemma A.4. *(i)* $\|(Z'M_XZ)^+\| = \begin{cases} O_P(\frac{1}{p}) & \text{if } p > n-r \\ O_P(\frac{1}{n-r}) & \text{if } n-r > p \end{cases}$

(ii) $\|\Lambda'(Z'M_XZ)^+\Lambda\| = O_P(\frac{1}{n})$

(iii) $\operatorname{tr}((Z'M_XZ)^+) = \begin{cases} O_P(\frac{n-r}{p}) & \text{if } p > n-r \\ O_P(\frac{p}{n-r}) & \text{if } n-r > p \end{cases}$,

$\operatorname{tr}(((Z'M_XZ)^+)^2) = \begin{cases} O_P(\frac{n-r}{p^2}) & \text{if } p > n-r \\ O_P(\frac{p}{(n-r)^2}) & \text{if } n-r > p \end{cases}$

Proof. (i) We first discuss the case when $p > n-r$,

$$\begin{aligned} \|(Z'M_XZ)^+\| &= \frac{1}{p}\sigma_{\max}((\frac{Z'M_XZ}{p})^+) \\ &= \frac{1}{p}\frac{1}{\sigma_{\min}(M_XZZ'M_X/p)} \end{aligned}$$

Let $\nu \in \mathbb{R}^n$ with $\|\nu\| = 1$ be the eigenvector of the $n \times n$ matrix $M_XZZ'M_X$ corresponding to its $n-r$-th eigenvalue. Then

$$\sigma_{\min}(M_XZZ'M_X/p) = \nu'\frac{M_XZZ'M_X}{p}\nu = \nu'M_XF\frac{\Lambda'\Lambda}{p}F'M_X\nu + 2\nu'M_XF\frac{\Lambda'U}{p}\nu + \nu'\frac{M_XUU'M_X}{p}\nu$$

Since M_X is a projection matrix, its $\sigma_{\min}(M_X) = 1$. Also, U is an $n \times p$ matrix whose elements are i.i.d and with finite fourth moment. Thus,

$$\sigma_{\min}(\frac{M_XUU'M_X}{p}) \geq \sigma_{\min}(M_X)\sigma_{\min}(\frac{UU'}{p}) > c_u$$

Then,

$$\begin{aligned} \sigma_{\min}(M_XZZ'M_X/p) &\geq \nu'M_XF\frac{\Lambda'\Lambda}{p}F'M_X\nu - |2\nu'M_XF\frac{\Lambda'U}{p}\nu| + c_u \\ &\geq \nu'M_XF\frac{\Lambda'\Lambda}{p}F'M_X\nu - 2\|(\frac{\Lambda'\Lambda}{p})^{1/2}F'M_X\nu\|\|(\frac{\Lambda'\Lambda}{p})^{-1/2}\frac{\Lambda'}{p}U'\nu\| + c_u \\ &\geq a'a - 2\|a\|\|M\| + c_u \end{aligned}$$

In addition,

$$\|\operatorname{Var}(M)\| = \|\sigma_u^2(\frac{\Lambda'\Lambda}{p})^{-1/2}\frac{\Lambda'\Lambda}{p^2}(\frac{\Lambda'\Lambda}{p})^{-1/2}\| \leq \frac{C}{p}$$

Therefore the event $\{\|M\| \leq \frac{\sqrt{c_u}}{2}\}$ holds with probability approaching one as $p, n \to \infty$. Under this event, we have

$$\begin{aligned} \sigma_{\min}(M_XZZ'M_X/p) &\geq \|a\|^2 - \sqrt{c_u}\|\alpha\| + c_u \\ &= (\|a\| - \frac{\sqrt{c_u}}{2})^2 + \frac{3}{4}c_u \geq \frac{3}{4}c_u \end{aligned}$$

Hence,

$$\|(Z'M_XZ))^+\| = \frac{1}{p}\frac{1}{\sigma_{\min}(M_XZ'ZM_X/p)} = O_P(\frac{1}{p})$$

For the case $n-r > p$, following a similar argument, we have $\sigma_{\min}(Z'M_XZ/(n-r)) \geq C$ such that

$$\begin{aligned}\|(Z'M_XZ)^+\| &= \frac{1}{n-r}\sigma_{\max}((\frac{Z'M_XZ}{n-r})^+) \\ &= \frac{1}{n-r}\frac{1}{\sigma_{\min}(Z'M_XZ/(n-r))} = O_P(\frac{1}{n-r})\end{aligned}$$

(ii) Since Z follows factor model. By partialling out X (multiply M_X), we have

$$M_XZ = M_XF\Lambda' + M_XU$$

Then the factor loading can be written as,

$$\Lambda = (Z-U)'M_XF(F'M_XF)^{-1}$$

Substituting it into $\Lambda'(Z'M_XZ)^+\Lambda$, we have

$$\begin{aligned}\|\Lambda'(Z'M_XZ)^+\Lambda\| &= \|\Lambda'(Z'M_XZ)^+(Z-U)'M_XF(F'M_XF)^{-1}\| \\ &= \|\Lambda'(Z'M_XZ)^+(Z-U)'M_XF/n\|O_P(1) \\ &\leq \|((Z'M_XZ)^+)^{1/2}\Lambda\|\|((Z'M_XZ)^+)^{1/2}(Z-U)'M_XF/n\|O_P(1)\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality. Then rearrange the above inequality,

$$\begin{aligned}\|\Lambda'(Z'M_XZ)^+\Lambda\| &\leq \|((Z'M_XZ)^+)^{1/2}(Z-U)'M_XF/n\|^2O_P(1) \\ &\leq 2\Big(\|((Z'M_XZ)^+)^{1/2}Z'M_XF/n\|^2 + \|((Z'M_XZ)^+)^{1/2}U'M_XF/n\|^2\Big)O_P(1) \\ &\leq 2\Big(\|M_XZ(Z'M_XZ)^+Z'M_X\| + \|(Z'M_XZ)^+\|\|U'M_XU\|)\Big)\|\frac{F'F}{n^2}\|O_P(1) \\ &= \frac{2}{n}\Big(1 + \|(Z'M_XZ)^+\|\|U'M_XU\|)\Big)O_P(1)\end{aligned} \tag{16}$$

where the last equality follows the fact that $M_XZ(Z'M_XZ)^+Z'M_X$ is symmetric and idempotent.

From (i) we know that $\|(Z'M_XZ)^+\| = O_P(\frac{1}{p})$ and $\|U'M_XU\| = O_P(p)$ when $p > n-r$ and $\|(Z'M_XZ)^+\| = O_P(\frac{1}{n-r})$ and $\|U'M_XU\| = O_P(n-r)$ when $n-r > p$

Therefore,

$$\|\Lambda'(Z'M_XZ)^+\Lambda\| = O_P(\frac{1}{n})$$

(iii) For $\mathrm{tr}((Z'M_XZ)^+)$, we first discuss the case $p > n-r$. From (i) we know $\sigma_{\min}(M_XZZ'M_X/p) \geq C$. Hence

$$\begin{aligned}\mathrm{tr}((Z'M_XZ)^+) &= \frac{1}{p}\mathrm{tr}((\frac{Z'M_XZ}{p})^+) = \frac{1}{p}\sum_{j=1}^{n-r}\frac{1}{\sigma_j(M_XZZ'M_X/p)} \\ &\leq \frac{1}{p}\frac{n-r}{\sigma_{n-r}(M_XZZ'M_X/p)} = O_P(\frac{n-r}{p})\end{aligned} \tag{17}$$

For the case $n-r > p$, from (i) we know $\sigma_{\min}(Z'M_XZ/(n-r)) \geq C$. Hence

$$\begin{aligned}\mathrm{tr}((Z'M_XZ)^+) &= \frac{1}{n-r}\mathrm{tr}((\frac{Z'M_XZ}{n-r})^+) = \frac{1}{n-r}\sum_{j=1}^{p}\frac{1}{\sigma_j(Z'M_XZ/(n-r))} \\ &\leq \frac{1}{n-r}\frac{p}{\sigma_p(Z'M_XZ/(n-r))} = O_P(\frac{p}{n-r})\end{aligned} \tag{18}$$

Therefore,

$$\mathrm{tr}((Z'M_XZ)^+) = \begin{cases} O_P(\frac{n-r}{p}) & \text{if } p > n-r \\ O_P(\frac{p}{n-r}) & \text{if } n-r > p \end{cases}$$

Following similar argument, we have

$$\mathrm{tr}(((Z'M_XZ)^+)^2) = \begin{cases} O_P(\frac{n-r}{p^2}) & \text{if } p > n-r \\ O_P(\frac{p}{(n-r)^2}) & \text{if } n-r > p \end{cases}$$

□

Lemma A.5. *As $n, p \to \infty$ such that $\gamma_n := \frac{p}{n-r} \to \gamma \in (0,\infty)\setminus\{1\}$*

$$\sigma_u^2\,\mathrm{tr}((Z'M_XZ)^+) \xrightarrow{p} \begin{cases} \frac{\gamma}{1-\gamma} & \text{if } 0 < \gamma < 1 \\ \frac{1}{\gamma-1} & \text{if } \gamma > 1 \end{cases}$$

Proof. Following from Lemma A.2 that $M_X = Q_2Q_2'$ where $Q_2 \in \mathbb{R}^{n\times(n-r)}$ is an orthogonal matrix and $U_2 = Q_2'U$ is an $(n-r)\times p$ matrix,

$$U'M_XU = U'Q_2Q_2'U = U_2'U_2$$

Then following from Lemma A.2, we have

$$F_{U'M_XU/((n-r)\sigma_u^2)} \;\Rightarrow\; F_\gamma \quad \text{a.s.}$$

with support $[a_\gamma, b_\gamma]$, where $a_\gamma = (1-\sqrt{\gamma})^2$ and $b_\gamma = (1+\sqrt{\gamma})^2$.

Next, we use the rank inequality (Bai (1999), Lemma 2.6): for any $p \times m$ matrices A and B,
Under the factor model $Z = F\Lambda' + U$,

$$M_X Z = M_X F\Lambda' + M_X U$$

Taking $A = \frac{Z'M_X}{\sqrt{n}}$ and $B = \frac{U'M_X}{\sqrt{n}}$, we have

$$\left\|F_{Z'M_XZ/(n-r)} - F_{U'M_XU/(n-r)}\right\|_\infty \;\le\; \frac{1}{p}\operatorname{rank}(M_X F\Lambda') \;\le\; \frac{J}{p} \;\to\; 0, \tag{19}$$

Let x be any continuous point of F_γ,

$$\begin{aligned}|F_{Z'M_XZ/((n-r)\sigma_u^2)}(x) - F_\gamma(x)| &\le |F_{Z'M_XZ/((n-r)\sigma_u^2)}(x) - F_{U_2'U_2/((n-r)\sigma_u^2)}(x)| + |F_{U_2'U_2/((n-r)\sigma_u^2)}(x) - F_\gamma(x)| \\ &\le \|F_{Z'M_XZ/((n-r)\sigma_u^2)}(x) - F_{U_2'U_2/((n-r)\sigma_u^2)}(x)\|_\infty + |F_{U_2'U_2/((n-r)\sigma_u^2)}(x) - F_\gamma(x)|\end{aligned}$$

Since the first term converge to zero by (19) and the second term converge to zero by weak convergence of $F_{U'M_XU/((n-r)\sigma_u^2)}$ to F_γ. Therefore, $F_{U'M_XU/((n-r)\sigma_u^2)}(x) \to F_\gamma(x)$. Since this holds for all points x where F_γ is continuous, $F_{Z'M_XZ/((n-r)\sigma_u^2)}(x)$ weakly converge to $F_\gamma(x)$ almost surely.

By Portmanteau theorem, weak convergence is equal to convergence in expectation of all bounded and continuous function g. Taking $g(x) = \frac{1}{x}$ since lower bound support of F_γ is $a = (1-\sqrt{\gamma})^2 > 0$, the function g is continous and bounded. As $n, p \to \infty$, we have

$$\int \frac{1}{x} dF_{Z'M_XZ/((n-r)\sigma_u^2)}(x) \to \int \frac{1}{x} dF_\gamma(x)$$

From Bai and Silverstein (2010) lemma 3.11, $\int \frac{1}{x} dF_\gamma(x)$ is equal to the Stieltjes transform of $m(z)$ of Marchenkop-Pastur law F_γ at $z = 0$. The Stieltjes transform of $m(z)$ is

$$m(z) = \int \frac{1}{x - z} dF_\gamma(x) = \frac{-(1-\gamma+z) + \sqrt{(1-\gamma+z)^2 + 4\gamma z}}{2\gamma z}$$

Since it is indeterminate as $z = 0$, we use l'Hôpital's Rule

$$m(0) = \lim_{z \to 0} \frac{-1 + \frac{1+\gamma+z}{\sqrt{(1-\gamma+z)^2+4\gamma z}}}{2\gamma} = \frac{-1 + \frac{1+\gamma}{1-\gamma}}{2\gamma} = \frac{1}{1-\gamma}$$

Thus,

$$\int \frac{1}{x} dF_\gamma(x) = m(0) = \frac{1}{1-\gamma}$$

Therefore,

$$\begin{aligned}\sigma_u^2 \operatorname{tr}((Z'M_XZ)^+) &= \frac{1}{n-r} \operatorname{tr}((Z'M_XZ/((n-r)\sigma_u^2))^+) = \frac{1}{n-r} \sum_{i=1}^{p} \frac{1}{\lambda_i(Z'M_XZ/((n-r)\sigma_u^2))} \\ &= \frac{p}{n-r} \int \frac{1}{x} dF_{Z'M_XZ/((n-r)\sigma_u^2)}(x) \to \frac{\gamma}{1-\gamma}\end{aligned}$$

For the case $n-r<p$, let $(n-r)/p \to \tau = 1/\gamma < 1$, by the similar arguments, we have

$$\begin{aligned}\sigma_u^2 \operatorname{tr}((Z'M_XZ)^+) &= \frac{1}{p} \operatorname{tr}((Z'M_XZ/(p\sigma_u^2))^+) = \frac{1}{p} \sum_{i=1}^{n-r} \frac{1}{\lambda_i(M_XZZ'M_X/(p\sigma_u^2))} \\ &= \frac{n-r}{p} \int \frac{1}{x} dF_{M_XZZ'M_X/(p\sigma_u^2)}(x) \to \frac{\tau}{1-\tau} = \frac{1}{\gamma-1}\end{aligned}$$

□

Lemma A.6. *Suppose $p > n-r$, and define $P_{Z^\perp} = Z'M_XZ(Z'M_XZ)^+$. Then*

(i) $\|\Lambda'(P_{Z^\perp} - I_p)\Lambda\| = O_P(\sqrt{\frac{\psi_{n,p}p}{n}})$

(ii) $\|\beta' P_{Z^\perp}\Sigma_{Z^\perp}\beta - \beta'\Sigma_{Z^\perp}\beta\| = O_P(\sqrt{\frac{p}{\psi_{n,p}n}})$,
$\|\beta' P_{Z^\perp}\Sigma_{Z^\perp}P_{Z^\perp}\beta - \beta'\Sigma_{Z^\perp}\beta\| = O_P(\sqrt{\frac{p}{\psi_{n,p}n}})$

Proof. (i) We apply the singular value decomposition $M_XZ = U_nS_nV_n'$ where V_n is a $p \times n$ matrix. Thus $P_{Z^\perp} - I_p = V_nV_n' - I_p = -V_AV_A'$ where V_A is a $p \times (p-(n-r))$ matrix, columns being eigenvectors of the $p \times p$ matrix $Z'M_XZ$ corresponding to the $p-(n-r)$ eigenvalues. Thus, $V_A'V_n = 0$

Then since $M_XZ = M_XF\Lambda' + M_XU$,

$$Z'M_XZ = (n-r)\Lambda\Sigma_{f|x}\Lambda' + H + U'M_XU$$

where $H = \Lambda(F'M_XF - (n-r)\Sigma_{f|x})\Lambda' + \Lambda F'M_XU + U'M_XF\Lambda'$. Under Lemma A.3 (i), $\|U'M_XU\| = O_P(p)$ as $p > n-r$. Following a similar argument in the proof of Lemma A.3 (ii), we have

$$\|\frac{1}{n}H\| = O_P(\sqrt{\frac{p\psi_{n,p}}{n}})$$

Then by Davis-Kahan theorem, the top J eigenvalues of $Z'M_XZ/n$ and $(1-\frac{r}{n})\Lambda\Sigma_{f|x}\Lambda'$,

respectively denoted as $\sigma_{j,Z^\perp}$ and σ_j for $j \le J$, are bounded by

$$\max_{j\le J} |\sigma_{j,Z^\perp} - \sigma_j| \le \|\frac{H}{n}\| + \|\frac{U'M_XU}{n}\| = O_P(\sqrt{\frac{p\psi_{n,p}}{n}} + \frac{p}{n})$$

Hence $\psi_{n,p} \gg p/n$ implies $\sigma_{j,Z^\perp} \asymp \psi_{n,p}$ for $j \le J$. There is $J \times J$ matrix S so that columns of ΛS are eigenvectors of $\Lambda\Sigma_{f|x}\Lambda'$, and $\|S^{-1}\| = O_P(\psi_{n,p}^{1/2})$. Let V_J denote the first J eigenvectors of $Z'M_XZ/n$. By Sin-theta inequality,

$$\|V_J - \Lambda S\| \le O_P(\psi_{n,p}^{-1})(\|\frac{H}{n}\| + \|\frac{U'M_XU}{n}\|) = O_P(\sqrt{\frac{p}{\psi_{n,p}n}}) + O_P(\frac{p}{\psi_{n,p}n}) = O_P(\sqrt{\frac{p}{\psi_{n,p}n}})$$

Also note that $V_A'V_J = 0$ because of the orthogonality. Hence,

$$\|\Lambda'(P_{Z^\perp} - I_p)\Lambda\| = \|V_A'\Lambda\|^2 \le \|V_A'\Lambda SS'\Lambda V_A\|\|S^{-1}\|^2 \le O_P(\psi_{n,p})\|\Lambda S - V_J\| = O_P(\sqrt{\frac{p\psi_{n,p}}{n}})$$

(ii) Since $P_{Z^\perp} - I_p = -V_AV_A'$, we have

$$\beta'P_{Z^\perp}\Sigma_{Z^\perp}\beta - \beta'\Sigma_{Z^\perp}\beta = -\beta'V_AV_A'\Sigma_{Z^\perp}\beta$$

By substituting $\Sigma_{Z^\perp} = \Lambda\Sigma_{f|x}\Lambda' + \sigma_u^2 I_p$ and $\beta = \sigma_u^{-2}\Lambda(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho$ into it, we have

$$\begin{aligned}
\|\beta'V_AV_A'\Sigma_{Z^\perp}\beta\| &= \|\beta'V_AV_A'\Lambda\Sigma_{f|x}\Lambda'\beta + \sigma_u^2\beta'V_AV_A'\beta\| \\
&\le \|\sigma_u^{-2}\rho'(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\Lambda'V_AV_A'\Lambda\Sigma_{f|x}\Lambda'\beta\| \\
&\quad + \|\sigma_u^{-2}\rho'(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\Lambda'V_AV_A'\Lambda'(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho\| \\
&\le O(1)\sigma_{\min}(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\|V_A'\Lambda\|^2\|\Lambda\|\|\beta\| \\
&\quad + O(1)\sigma_{\min}(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-2}\|V_A'\Lambda\|^2 \\
&\le O(1)\sigma_{\min}(\Lambda'\Lambda)^{-1}\|V_A'\Lambda\|^2\|\Lambda\|\|\beta\| + O(1)\sigma_{\min}(\Lambda'\Lambda)^{-2}\|V_A'\Lambda\|^2 \\
&= O_P(\sqrt{\frac{p}{\psi_{n,p}n}}) + O_P(\frac{\sqrt{p}}{\psi_{n,p}\sqrt{n}}) = O_P(\sqrt{\frac{p}{\psi_{n,p}n}})
\end{aligned}$$

where the last equality follows from (i), Theorem 1.

Therefore,

$$\|\beta'P_{Z^\perp}\Sigma_{Z^\perp}\beta - \beta'\Sigma_{Z^\perp}\beta\| = O_P(\sqrt{\frac{p}{\psi_{n,p}n}})$$

Following similar argument, we have

$$\|\beta' P_{Z^\perp} \Sigma_{Z^\perp} P_{Z^\perp} \beta - \beta' \Sigma_{Z^\perp} \beta\| = O_P(\sqrt{\frac{p}{\psi_{n,p} n}})$$

□

A.3 Proof of Theorem 2

In this section, we provide a proof of Theorem 2 to measure the rate of concentration parameter.

Proof of Theorem 2. Following from Theorem 1 and Lemma A.3 (ii), an upper bound for μ^2 is

$$\mu^2 \leq \frac{1}{\sigma_e^2} \|\beta\|^2 \|Z' M_X Z\| = O_P(n - r)$$

For the lower bound of μ^2, first observe that for some constant c_0,

$$\begin{aligned} \frac{1}{n-r} \beta' \Lambda F' M_X F \Lambda' \beta &= \beta' \Lambda \Sigma_{f|x} \Lambda' \beta - \beta' \Lambda (\Sigma_{f|x} - \frac{F' M_X F}{n-r}) \Lambda' \beta \\ &\geq \beta' \Lambda \Sigma_{f|x} \Lambda' \beta - o_P(1) \\ &\geq c_0 - o_P(1) \end{aligned} \tag{20}$$

where the first inequality follows from Assumption 1 (iii) and last inequality follows from $\Sigma_{f|x}$ is positive definite.

Next,

$$\begin{aligned} \frac{1}{n-r} \beta' U' M_X U \beta &\leq \frac{1}{n-r} \|\beta' U\|^2 \|M_X\| \\ &= O_P(\frac{1}{\psi_{n,p}}) = o_P(1) \end{aligned} \tag{21}$$

where the second inequality follows from Lemma A.3 (iv).

Then, with (20) and (21),

$$\begin{aligned}
\frac{1}{n-r}\mu^2 &= \frac{1}{\sigma_e^2(n-r)}\|\beta' Z^{\perp\prime} Z^{\perp}\beta\| \\
&\geq \frac{1}{\sigma_e^2(n-r)}\left(\|\beta'\Lambda F' M_X F\Lambda'\beta\| - 2\|\beta'\Lambda F' M_X\|\|M_X U\beta\| + \|\beta' U' M_X U\beta\|\right) \\
&= \frac{1}{\sigma_e^2(n-r)}\left((\frac{1}{\sqrt{2}}\|M_X F\Lambda'\beta\| - \sqrt{2}\|M_X U\beta\|)^2 + \frac{1}{2}\|M_X F\Lambda'\beta\|^2 - \|M_X U\beta\|^2\right) \\
&\geq \frac{1}{\sigma_e^2(n-r)}\left(\frac{1}{2}\|\beta'\Lambda F' M_X F\Lambda'\beta\| - \|\beta' U' M_X U\beta\|\right) \\
&\geq \frac{c_0}{2\sigma_e^2} - o_P(1)
\end{aligned}$$

Therefore, the above results show that there are two constant $c, C \in (0,\infty)$ such that the event

$$c(n-r) \leq \mu^2 \leq C(n-r)$$

holds with probability approaching one.

□

Since the estimator is constructed using sample splitting, we introduce notation with subscripts to indicate the corresponding subsample: $k = \{1,2\}$ and $k^c = \{1,2\}\backslash\{k\}$. For example, Z_k denotes an $n_k \times p$ matrix where each row is z_i' for $i \in I_k$, and $P_k = Z_k' M_{X_k} Z_k (Z_k' M_{X_k} Z_k)^+$ denotes the subsample k counterpart of $P_{Z^\perp}$.

Lemma A.7. *(i)* $\operatorname{tr}((Z_{k^c}' M_{X_{k^c}} Z_{k^c})^+ \frac{Z_k' M_{X_k} Z_k}{n_k}) \xrightarrow{p} \begin{cases} (1-\tau)\frac{\gamma}{1-\gamma}, & 0<\gamma<1 \\ (1-\tau)\frac{1}{\gamma-1}, & \gamma>1 \end{cases}$

(ii) $\frac{1}{n_k}\operatorname{tr}(P_1 P_2) \xrightarrow{p} \begin{cases} (1-\tau)\gamma, & 0<\gamma<1 \\ (1-\tau)\frac{1}{\gamma}, & \gamma>1 \end{cases}$

(iii) $\operatorname{tr}((Z_{k^c}' M_{X_{k^c}} Z_{k^c})^+ \frac{Z_k' M_{X_k} Z_k}{n_k} (Z_{k^c}' M_{X_{k^c}} Z_{k^c})^+ \frac{Z_k' M_{X_k} Z_k}{n_k}) = o_P(1)$

(iv) $\|\frac{Z_k' M_{X_k} Z_k}{n_k} (Z_{k^c}' M_{X_{k^c}} Z_{k^c})^+ \frac{Z_k' M_{X_k} Z_k}{n_k}\| = O_P(\frac{\psi_{n,p}}{n_{k^c}-r})$

Proof. (i) Without loss of generality, we consider the case of $k=1$, i.e., $\operatorname{tr}((Z_2' M_{X_2} Z_2)^+ \frac{Z_1' M_{X_1} Z_1}{n_1})$.

Since Z_1 follow factor model, we can rewrite $Z_1' M_{X_1} Z_1$ as

$$\begin{aligned}\frac{1}{n_1} Z_1' M_{X_1} Z_1 &= \frac{1}{n_1}\Lambda F_1' M_{X_1} F_1 \Lambda' + (\frac{1}{n_1}\Lambda F_1' M_{X_1} U_1 + \frac{1}{n_1} U_1' M_{X_1} F_1 \Lambda') \\ &\quad + (\frac{1}{n_1} U_1' M_{X_1} U_1 - (1 - \frac{r}{n_1})\sigma_u^2 I_p) + (1 - \frac{r}{n_1})\sigma_u^2 I_p \\ &= \sum_{j=1}^{4} A_j\end{aligned}$$

Then,

$$\operatorname{tr}((Z_2' M_{X_2} Z_2)^+ \frac{Z_1' M_{X_1} Z_1}{n_1}) = \sum_{j=1}^{4} \operatorname{tr}((Z_2' M_{X_2} Z_2)^+ A_j) \tag{22}$$

For the first term in (22),

$$\begin{aligned}\operatorname{tr}((Z_2' M_{X_2} Z_2)^+ A_1) &= \operatorname{tr}(\Lambda'(Z_2' M_{X_2} Z_2)^+ \Lambda \frac{F_1' M_{X_1} F_1}{n_1}) \\ &\le J \|\Lambda'(Z_2' M_{X_2} Z_2)^+ \Lambda\| \|\frac{F_1' M_{X_1} F_1}{n_1}\| \\ &= O_P(\frac{1}{n_2})\end{aligned} \tag{23}$$

where the last equality follows from Lemma A.4 (ii).

For the second term in (22), first note that

$$\begin{aligned}\|(Z_2' M_{X_2} Z_2)^+ \Lambda\| &\le \|((Z_2' M_{X_2} Z_2)^+)^{1/2}\| \|((Z_2' M_{X_2} Z_2)^+)^{1/2} \Lambda\| \\ &= \|((Z_2' M_{X_2} Z_2)^+)^{1/2}\| O_P(\frac{1}{\sqrt{n_2}}) \\ &= \sigma_{\min}^{-1/2}(Z_2' M_{X_2} Z_2) O_P(\frac{1}{\sqrt{n_2}}) \\ &= O_P(\frac{1}{\sqrt{n_2(n_2 - r)}} + \frac{1}{\sqrt{n_2 p}})\end{aligned}$$

where the last equality follow from the proof in Lemma A.4 (i) that $\sigma_{\min}^{-1}(Z_2' M_{X_2} Z_2) = O_P(\frac{1}{n_2 - r} + \frac{1}{p})$. Then,

$$\begin{aligned}\operatorname{tr}((Z_2' M_{X_2} Z_2)^+ A_2) &= 2tr(F_1' M_{X_1} U_1 (Z_2' M_{X_2} Z_2)^+ \Lambda / n_1) \\ &\le 2J \frac{1}{\sqrt{n_1}} \|\frac{1}{n_1} F_1' F_1\|^{1/2} \|U_1' M_{X_1} U_1\|^{1/2} \|(Z_2' M_{X_2} Z_2)^+ \Lambda\| \\ &= O_P(\frac{1}{\sqrt{n_1 n_2}})\end{aligned} \tag{24}$$

where the last equality follow from Lemma A.3 (i).

For the third term in (22),

$$\operatorname{tr}((Z_2'M_{X_2}Z_2)^+A_3) = \operatorname{tr}\left((Z_2'M_{X_2}Z_2)^+(\frac{U_1'M_{X_1}U_1}{n_1} - (1-\frac{r}{n_1})\sigma_u^2 I_p)\right)$$

Let $B = (Z_2'M_{X_2}Z_2)^+$ and $S = U_1'M_{X_1}U_1$. We know that $E(S|X_1) = \sigma_u^2 \operatorname{tr}(M_{X_1})I_p = \sigma_u^2(n_1 - r)I_p$. Then we can rewrite the third term in (22) as

$$\begin{aligned}\operatorname{tr}((Z_2'M_{X_2}Z_2)^+A_3) &= \operatorname{tr}(\frac{1}{n_1}(Z_2'M_{X_2}Z_2)^+(U_1'M_{X_1}U_1 - \sigma_u^2(n_1-r)I_p)) \\ &= \operatorname{tr}(\frac{1}{n_1}B(S - E(S|X_1))\end{aligned}$$

Its conditional mean is

$$E(\operatorname{tr}(\frac{1}{n_1}B(S - E(S|X_1)))|X_1, X_2, Z_2) = 0$$

Let $\operatorname{Var}(S_{ij}|X_1)$ and $\operatorname{Var}(S_{ii}|X_1)$ denote, respectively, the conditional variances of the off-diagonal and diagonal entries of S. From Lemma A.1, we have $\operatorname{Var}(S_{ij}|X_1) = \operatorname{tr}(M_{X_1})\sigma_u^4 = (n_1 - r)\sigma_u^4$, $\operatorname{Var}(S_{ii}|X_1) = 2tr(M_{X_1})\sigma_u^4 + \sum_{k=1}^n M_{X_1,kk}^2(E(u_{ij}^4) - 3\sigma_u^4) \leq C(n_1 - r)$ since $\sum_{k=1}^n M_{X_1,kk}^2 \leq \operatorname{tr}(M_{X_1}^2) = (n_1 - r)$. Then the conditional variance is

$$\begin{aligned}\operatorname{Var}(\frac{1}{n_1}\operatorname{tr}(B(S - E(S|X_1)))|X_1, X_2, Z_2) &= \frac{1}{n_1^2}\sum_{i=1}^p\sum_{j=1}^p\sum_{k=1}^p\sum_{l=1}^p B_{ji}B_{lk}Cov(S_{ij}, S_{kl}|X_1) \\ &= \frac{1}{n_1^2}\sum_{i=1}^p B_{ii}^2 \operatorname{Var}(S_{ii}|X_1) + \frac{1}{n_1^2}\sum_{i=1}^{p-1}\sum_{j=i+1}^p (B_{ij} + B_{ji})^2 \operatorname{Var}(S_{ij}|X_1) \\ &\leq C\frac{n_1 - r}{n_1^2}(\sum_{i=1}^p B_{ii}^2 + \sum_{i=1}^{p-1}\sum_{j=i+1}^p (B_{ij} + B_{ji})^2) \\ &\leq C'\frac{n_1 - r}{n_1^2}\operatorname{tr}(B^2) \\ &= \begin{cases} O_P(\frac{(n_1-r)(n_2-r)}{n_1^2p^2}), & p > n_2 - r \\ O_P(\frac{(n_1-r)p}{n_1^2(n_2-r)^2}), & n_2 - r > p \end{cases}\end{aligned}$$

where the last equality follows from Lemma A.4 (iii)

Thus,

$$\operatorname{tr}((Z_2'M_{X_2}Z_2)^+A_3) = \begin{cases} O_P(\frac{\sqrt{(n_1-r)(n_2-r)}}{n_1p}), & p > n_2 - r \\ O_P(\frac{\sqrt{(n_1-r)p}}{n_1(n_2-r)}), & n_2 - r > p \end{cases} \tag{25}$$

For the fourth term in (22), following from Lemma A.5 and Assumption 3,

$$\operatorname{tr}((Z_2'M_{X_2}Z_2)^+A_4) = \sigma_u^2(1-\frac{r}{n_1})\operatorname{tr}((Z_2'M_{X_2}Z_2)^+) \xrightarrow{p} \begin{cases} (1-\tau)\frac{\gamma}{1-\gamma}, & 0<\gamma<1 \\ (1-\tau)\frac{1}{\gamma-1}, & \gamma>1 \end{cases} \tag{26}$$

Therefore, combining (22), (23), (24), (25) and (26), we have

$$\operatorname{tr}((Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}) = \operatorname{tr}((Z_2'M_{X_2}Z_2)^+A_4)+o_P(1) \xrightarrow{p} \begin{cases} (1-\tau)\frac{\gamma}{1-\gamma}, & 0<\gamma<1 \\ (1-\tau)\frac{1}{\gamma-1}, & \gamma>1 \end{cases}$$

(ii) We separately consider the cases that $p < n_k - r$ and $p > n_k - r$.

First, when $p < n_k - r$, $P_k = I_p$ for $k = 1, 2$. Then, following from Assumption 3

$$\frac{1}{n_k}tr(P_1P_2) = \frac{1}{n_k}tr(I_p) = \frac{n_k-r}{n}\frac{p}{n_k-r} \to (1-\tau)\gamma, \quad 0<\gamma<1$$

Next, we discuss the case that $p > n_k - r$. Let $Q_k = U_k'M_{X_k}U_k(U_k'M_{X_k}U_k)^+$ for $k = 1, 2$. We first show that

$$\frac{1}{n_k}tr(P_1P_2) - \frac{1}{n_k}tr(Q_1Q_2) = o_P(1)$$

Under the factor model $Z = F\Lambda' + U$,

$$M_XZ = M_XF\Lambda' + M_XU$$

Since $rank(M_XF\Lambda') \le J$, we have $rank(P_k - Q_k) \le 2J$. Then,

$$\|P_k - Q_k\|_* \le rank(P_k - Q_k)\|P_k - Q_k\| \le 2J$$

where $\|P_1 - Q_1\| \le 1$ since P_1 and Q_1 are projection matrix.

Then,

$$\begin{aligned} |tr(P_1P_2) - tr(Q_1Q_2)| &\le |tr((P_1-Q_1)P_2)| + |tr(Q_1(P_2-Q_2))| \\ &\le \|P_1-Q_1\|_*\|P_2\| + \|Q_1\|\|P_2-Q_2\|_* \le 4J \end{aligned}$$

Thus, $\frac{1}{n_k}tr(P_1P_2) - \frac{1}{n_k}tr(Q_1Q_2) = o_P(1)$

Then, we want to show the probability convergence of $\frac{1}{n_k}tr(Q_1Q_2)$. We first observe that the distribution of Q_k is invariant conditional on X, Since u_{ij} are i.i.d. and U

are independent to X, permuting the columns of U_k does not change its conditional distribution:

$$U_k S|X \stackrel{d}{=} U_k|X$$

where S be a $p \times p$ permutation orthogonal matrix. Then,

$$\begin{aligned} Q_k(U_k S) &= S'(U_k' M_{X_k} U_k) S (S'(U_k' M_{X_k} U_k) S)^+ \\ &= S'(U_k' M_{X_k} U_k)(U_k' M_{X_k} U_k)^+ S \\ &= S' Q_k(U_k) S \end{aligned}$$

Thus,

$$Q_k|X \stackrel{d}{=} S' Q_k S|X$$

Let $G_k = E(Q_k|X)$. Because two random matrices with the same conditional distribution have the same conditional expectation, then for every permutation matrix S,

$$G_k = E(Q_k|X) = E(S' Q_k S|X) = S' G_k S$$

For the diagonal element of G_k, there exist a permutation matrix such that $G_{k_{ii}} = G_{k_{jj}}$, which indicates all diaognal elements are equal. Similarly, for the off-diagonal element of G_k, there exist a permutation matrix such that $G_{k_{ij}} = G_{k_{st}}$ for $i \neq s$ and $j \neq t$, which indicates all off-diagonal element are equal. Thus,

$$G_k = (a_k - b_k)\mathbf{I}_p + b_k \mathbf{1}_p \mathbf{1}_p'$$

where a_k and b_k diagonal and off-diagonal element of G_k, respectively. Then since Q_k is a projection matrix with rank $n_k - r$,

$$p a_k = tr(G_k) = E(tr(Q_k)|X) = n_k - r$$

We have $a_k = \frac{n_k - r}{p}$

Also,

$$p a_k + p(p-1) b_k = \mathbf{1}_p' G_k \mathbf{1}_p = E(\mathbf{1}_p' Q_k \mathbf{1}_p | X)$$

Since Q_k is a projection matrix, $0 \leq E(\mathbf{1}_p' Q_k \mathbf{1}_p|X) \leq p$, we have $|b_k| \leq \frac{1}{p-1}$

Therefore, since $Q_1 \perp\!\!\!\perp Q_2|X$,

$$E(tr(Q_1 Q_2)|X) = tr(G_1 G_2) = p a_1 a_2 + p(p-1) b_1 b_2 = \frac{(n_1 - r)(n_2 - r)}{p} + O(1)$$

Thus, following from Assumption 3,

$$\frac{1}{n_k}E(tr(Q_1Q_2)|X) \xrightarrow{p} (1-\tau)\frac{1}{\gamma}, \quad \gamma > 1$$

For its conditional variance, we use conditional Efron-stein inequality since u_{ij} are i.i.d and independent to X.

Let u'_{ki} denote row i of U_k. Replace the row u_{ki} by an independent copy u^*_{ki} and denote the resulting projector by $Q_k^{(i)}$. Since $M_{X_k}e_i(u'_{ki} - u^{*'}_{ki})$ is an outer product, $rank(M_{X_k}e_i(u'_{ki} - u^{*'}_{ki})) \leq 1$ and $rank(Q_k - Q_k^{(i)}) \leq 2$. Then,

$$\|Q_k - Q_k^{(i)}\|_* \leq rank(Q_k - Q_k^{(i)})\|Q_k - Q_k^{(i)}\| \leq 2$$

It follows that

$$|tr(Q_kQ_{3-k}) - tr(Q_k^{(i)}Q_{3-k})| \leq \|Q_k - Q_k^{(i)}\|_*\|Q_{3-k}\| \leq 2$$

Conditional Efron–Stein then gives

$$\begin{aligned} Var(tr(Q_1Q_2)|X) &\leq \frac{1}{2}\sum_{k=1}^{2}\sum_{i=1}^{n_k} E[(tr(Q_kQ_{3-k}) - tr(Q_k^{(i)}Q_{3-k}))^2|X] \\ &\leq 2(n_1+n_2) \end{aligned}$$

Thus,

$$Var(\frac{1}{n_k}tr(Q_1Q_2)|X) \xrightarrow{p} 0$$

It follows that

$$\begin{aligned} \frac{1}{n_k}tr(P_1P_2) &= \frac{1}{n_k}tr(Q_1Q_2) + (\frac{1}{n_k}tr(P_1P_2) - \frac{1}{n_k}tr(Q_1Q_2)) \\ &= \frac{1}{n_k}tr(Q_1Q_2) + o_P(1) \xrightarrow{p} (1-\tau)\frac{1}{\gamma}, \quad \gamma > 1 \end{aligned}$$

Therefore,

$$\frac{1}{n_k}tr(P_1P_2) \xrightarrow{p} \begin{cases} (1-\tau)\gamma, & 0 < \gamma < 1 \\ (1-\tau)\frac{1}{\gamma}, & \gamma > 1 \end{cases}$$

(iii) Without loss of generality, we consider the case of $k = 1$, i.e., $\text{tr}((Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}(Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1})$.

Since Z follows factor model, we can rewrite $\frac{1}{n_1}Z_1'M_{X_1}Z_1$ as

$$\begin{aligned}\frac{1}{n_1}Z_1'M_{X_1}Z_1 &= \frac{1}{n_1}\Lambda F_1'M_{X_1}F_1\Lambda' + \frac{1}{n_1}U_1'M_{X_1}U_1 \\ &\quad + (\frac{1}{n_1}\Lambda F_1'M_{X_1}U_1 + \frac{1}{n_1}U_1'M_{X_1}F_1\Lambda') \\ &= \sum_{j=1}^{3} B_j\end{aligned}$$

Then, since $(Z_2'M_{X_2}Z_2)^+$ is p.s.d,

$$\begin{aligned}&\mathrm{tr}((Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}(Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}) \\ &\qquad = \|((Z_2'M_{X_2}Z_2)^+)^{1/2}\frac{Z_1'M_{X_1}Z_1}{n_1}((Z_2'M_{X_2}Z_2)^+)^{1/2}\|_F^2 \\ &\qquad \le 3\sum_{j=1}^{3}\|((Z_2'M_{X_2}Z_2)^+)^{1/2}B_j((Z_2'M_{X_2}Z_2)^+)^{1/2}\|_F^2\end{aligned} \tag{27}$$

For the first term in (27),

$$\begin{aligned}\|((Z_2'M_{X_2}Z_2)^+)^{1/2}B_1((Z_2'M_{X_2}Z_2)^+)^{1/2}\|_F^2 &\le \|((Z_2'M_{X_2}Z_2)^+)^{1/2}\Lambda\|^4\|\frac{1}{n_1}F_1'M_XF_1\|_F^2 \\ &\le J\|\Lambda'(Z_2'M_{X_2}Z_2)^+\Lambda\|^2\|\frac{1}{n_1}F_1'M_{X_1}F_1\|^2 \\ &= O_P(\frac{1}{n_2^2})\end{aligned}$$

where the last equality follows from Lemma A.4 (ii).

For the second term in (27),

$$\begin{aligned}\|((Z_2'M_{X_2}Z_2)^+)^{1/2}B_2((Z_2'M_{X_2}Z_2)^+)^{1/2}\|_F^2 &\le \|(Z_2'M_{X_2}Z_2)^+\|^2\|\frac{U_1'M_{X_1}U_1}{n_1}\|_F^2 \\ &\le p\|(Z_2'M_{X_2}Z_2)^+\|^2\|\frac{U_1'M_{X_1}U_1}{n_1}\|^2 \\ &= \begin{cases} O_P(\frac{1}{p}), \text{ if } p > n_2 - r \\ O_P(\frac{p}{(n_2-r)^2}), \text{ if } n_2 - r > p \end{cases}\end{aligned}$$

where the last equality follows from Lemma A.3 (i) and Lemma A.4 (i).

For the third term in (27),

$$\begin{aligned}
&\|((Z_2'M_{X_2}Z_2)^+)^{1/2}B_3((Z_2'M_{X_2}Z_2)^+)^{1/2}\|_F^2 \\
&\qquad \le 2\|((Z_2'M_{X_2}Z_2)^+)^{1/2}\frac{1}{n_1}\Lambda F_1'M_{X_1}U_1((Z_2'M_{X_2}Z_2)^+)^{1/2}\|_F^2 \\
&\qquad + 2\|((Z_2'M_{X_2}Z_2)^+)^{1/2}\frac{1}{n_1}U_1'M_{X_1}F_1\Lambda'((Z_2'M_{X_2}Z_2)^+)^{1/2}\|_F^2 \\
&\qquad \le 4\|((Z_2'M_{X_2}Z_2)^+)^{1/2}\Lambda\|^2\|((Z_2'M_{X_2}Z_2)^+)^{1/2}\|^2\|\|\frac{1}{n_1}F_1M_{X_1}U_1\|_F^2 \\
&\qquad \le 4J\|\Lambda'((Z_2'M_{X_2}Z_2)^+)\Lambda\|\|(Z_2'M_{X_2}Z_2)^+\|\|\|\frac{1}{n_1}F_1'F_1\|\|\frac{1}{n_1}U_1'M_{X_1}U_1\| \\
&\qquad = O_P(\frac{1}{n_1n_2})
\end{aligned}$$

where the last equality follows from Lemma A.3 (i) and Lemma A.4 (i)(ii).

Since $\|((Z_2'M_{X_2}Z_2)^+)^{1/2}B_j((Z_2'M_{X_2}Z_2)^+)^{1/2}\|_F^2 = o_P(1)$ for each j, we have

$$\mathrm{tr}((Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}(Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}) = o_P(1)$$

(iv) Without loss of generality, we consider the case of $k = 1$, i.e., $\|\frac{Z_1'M_{X_1}Z_1}{n_1}(Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}\|$. Since Z follows factor model, we can rewrite $\frac{1}{n_1}Z_1'M_{X_1}Z_1$ as

$$\begin{aligned}
\frac{1}{n_1}Z_1'M_{X_1}Z_1 &= \frac{1}{n_1}\Lambda F_1'M_{X_1}F_1\Lambda' + \frac{1}{n_1}U_1'M_{X_1}U_1 \\
&\quad + (\frac{1}{n_1}\Lambda F_1'M_{X_1}U_1 + \frac{1}{n_1}U_1'M_{X_1}F_1\Lambda') \\
&= \sum_{j=1}^{3}B_j
\end{aligned}$$

Then, since all B_j are symmetric and $(Z_2'M_{X_2}Z_2)^+$ is p.s.d,

$$\|\frac{Z_1'M_{X_1}Z_1}{n_1}(Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}\| \le 3\sum_{j=1}^{3}\|B_j(Z_2'M_{X_2}Z_2)^+B_j\| \tag{28}$$

For the first term and second term in (28), following from Lemma A.3 (i) and

Lemma A.4 (i)(ii),

$$\|B_1(Z_2'M_{X_2}Z_2)^+B_1\| \leq \|\Lambda\|^2\|\frac{F_1'M_{X_1}F_1}{n_1}\|^2\|\Lambda'(Z_2'M_{X_2}Z_2)^+\Lambda\| = O_P(\frac{\psi_{n,p}}{n_2})$$

$$\|B_2(Z_2'M_{X_2}Z_2)^+B_2\| \leq \|(Z_2'M_{X_2}Z_2)^+\|\|\frac{1}{n_1}U_1'M_{X_1}U_1\|^2 = \begin{cases} O_P(\frac{p}{n_2^2}), & p > n_2 - r \\ O_P(\frac{1}{n_2}), & n_2 - r > p \end{cases}$$

For the third term in (28),

$$\begin{aligned}
&\|B_3(Z_2'M_{X_2}Z_2)^+B_3\| \leq \|\frac{1}{n_1^2}\Lambda F_1'M_{X_1}U_1(Z_2'M_{X_2}Z_2)^+U_1'M_{X_1}F_1\Lambda'\| \\
&+ \|\frac{1}{n_1^2}U_1'M_{X_1}F_1\Lambda'(Z_2'M_{X_2}Z_2)^+\Lambda F_1'M_{X_1}U_1\| + 2\|\frac{1}{n_1^2}U_1'M_{X_1}F_1\Lambda'(Z_2'M_{X_2}Z_2)^+U_1'M_{X_1}F_1\Lambda'\| \\
&\leq \frac{1}{n_1^2}\|F_1'M_{X_1}U_1\|^2\left(\|\Lambda\|^2\|(Z_2'M_{X_2}Z_2)^+\| + \|\Lambda'(Z_2'M_{X_2}Z_2)^+\Lambda\| + 2\|\Lambda\|\|(Z_2'M_{X_2}Z_2)^+\Lambda\|\right) \\
&\leq \frac{1}{n_1}\|\frac{1}{n_1}F_1'M_{X_1}F_1\|\|U_1'M_{X_1}U_1\|\left(\|\Lambda\|^2\|(Z_2'M_{X_2}Z_2)^+\| + \|\Lambda'(Z_2'M_{X_2}Z_2)^+\Lambda\| + 2\|\Lambda\|\|(Z_2'M_{X_2}Z_2)^+\Lambda\|\right) \\
&= O_P(\frac{\psi_{n,p}}{n_2 - r})
\end{aligned}$$

where the last equality follows from Lemma A.3 (i) and Lemma A.4 (i)(ii). Therefore, following from Assumption 4 (ii) that $p = o(\psi_{n,p}n)$ such that

$$\|\frac{Z_1'M_{X_1}Z_1}{n_1}(Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}\| = O_P(\frac{\psi_{n,p}}{n_2 - r})$$

□

The following two lemmas derive the leading terms in the numerator and denominator of $\hat{\alpha}$. Before deriving the leading terms, we decompose $M_X\boldsymbol{e}$ that will be used below. Let $b = \Lambda(\Lambda'\Lambda)^{-1}\rho$. From the proof of Theorem 1, we have

$$\begin{aligned}
M_X\boldsymbol{e} &= M_X\boldsymbol{\epsilon}_d + M_XF\rho - M_XZ\beta \\
&= M_X\boldsymbol{\epsilon}_d - M_XUb + M_XZ\Lambda(\Lambda'\Lambda)^{-1}\rho - M_XZ\beta \\
&= M_X\boldsymbol{\epsilon}_d - M_XUb + M_XZ(\Lambda(\Lambda'\Lambda)^{-1}\rho - \beta)
\end{aligned}$$

where the second equality follows from $Z = F\Lambda' + U$.

Since $\beta = \sigma_u^{-2}\Lambda(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho$ and $\rho - \Lambda'\beta = \Sigma_{f|x}^{-1}(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho$ from (11)

in the proof of Theorem 1, we have

$$\begin{aligned}\Lambda(\Lambda'\Lambda)^{-1}\rho-\beta &= \Lambda(\Lambda'\Lambda)^{-1}\rho-\sigma_u^{-2}\Lambda(\Sigma_{f|x}^{-1}+\sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho\\ &= \sigma_u^{-2}\Lambda(\sigma_u^{-2}\Lambda'\Lambda)^{-1}(\Sigma_{f|x}^{-1}+\sigma_u^{-2}\Lambda'\Lambda)(\Sigma_{f|x}^{-1}+\sigma_u^{-2}\Lambda'\Lambda)^{-1}-\sigma_u^{-2}\Lambda(\Sigma_{f|x}^{-1}+\sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho\\ &= \Lambda(\Lambda'\Lambda)^{-1}\Sigma_{f|x}^{-1}(\Sigma_{f|x}^{-1}+\sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho\\ &= \Lambda(\Lambda'\Lambda)^{-1}(\rho-\Lambda'\beta)\end{aligned}$$

By substituting it into $M_X\boldsymbol{e}$,

$$M_X\boldsymbol{e}=M_X\boldsymbol{\epsilon}_d-M_XUb+M_XZ\Lambda(\Lambda'\Lambda)^{-1}(\rho-\Lambda'\beta) \tag{29}$$

Lemma A.8. *For each $k\in\{1,2\}$,*

$$\frac{1}{\sqrt{n_k}}\hat{D}_k^{\perp\prime}\epsilon_{y,k}^{\perp}=\begin{cases}\frac{1}{\sqrt{n_k}}\beta'Z_k'M_{X_k}\epsilon_{y,k}+\frac{1}{\sqrt{n_k}}\epsilon_{d,k^c}'M_{X_{k^c}}Z_{k^c}(Z_{k^c}'M_{X_{k^c}}Z_{k^c})^{+}Z_k'M_{X_k}\epsilon_{y,k}+o_P(1), & \text{if } n_{k^c}-r\asymp p,\\ \frac{1}{\sqrt{n_k}}\beta'Z_k'M_{X_k}\epsilon_{y,k}+o_P(1), & \text{if } p\gg n_{k^c}-r.\end{cases}$$

Proof. Without loss of generality, we consider the case of k=1, i.e. $\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}$. Let $P_2=Z_2'M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^{+}$. By substituting $\hat{D}_1^{\perp}=Z_1^{\perp}\hat{\beta}_2=M_{X_1}Z_1(Z_2'M_{X_2}Z_2)^{+}Z_2'M_{X_2}D_2$, $M_{X_2}D_2=M_{X_2}Z_2\beta+M_{X_2}\boldsymbol{e}_2$, and $M_{X_2}\boldsymbol{e}_2=M_{X_2}\boldsymbol{\epsilon}_{d,2}-M_{X_2}U_2b+M_{X_2}Z_2\Lambda(\Lambda'\Lambda)^{-1}(\rho-\Lambda'\beta)$ into it, we have

$$\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp}=\sum_{i=1}^{6}S_i \tag{30}$$

where

$$\begin{aligned}&S_1=\frac{1}{\sqrt{n_1}}\beta'Z_1'M_{X_1}\boldsymbol{\epsilon}_{y,1},\quad S_2=\frac{1}{\sqrt{n_1}}\beta'(P_2-I_n)\Lambda F_1'M_{X_1}\boldsymbol{\epsilon}_{y,1},\quad S_3=\frac{1}{\sqrt{n_1}}\beta'(P_2-I_n)U_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}\\ &S_4=\frac{1}{\sqrt{n_1}}\boldsymbol{\epsilon}_{d,2}'M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^{+}Z_1'M_{X_1}\boldsymbol{\epsilon}_{y,1},\quad S_5=\frac{-1}{\sqrt{n_1}}b'U_2'M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^{+}Z_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}\\ &S_6=\frac{1}{\sqrt{n_1}}(\rho-\Lambda'\beta)'(\Lambda'\Lambda)^{-1}\Lambda'P_2Z_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}\end{aligned}$$

For S_1, S_3, following from Lemma A.3 (iii) we have

$$\begin{aligned}\|S_1\|&\le\|\beta\|\|\frac{1}{\sqrt{n_1}}Z_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}\|=O_P(1)\\ \|S_3\|&\le\|\beta\|\|P_2-I_n\|\|\frac{1}{\sqrt{n_1}}U_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}\|=O_P(\frac{1}{\sqrt{\psi_{n,p}}}+\sqrt{\frac{p}{n_1\psi_{n,p}}})\end{aligned}$$

With $\beta = \sigma_u^{-2}\Lambda(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho$ in Theorem 1, we have

$$(\rho - \Lambda'\beta) = \Sigma_{f|x}^{-1}(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho$$

Substituting β and $\rho - \Lambda'\beta$ into S_2, we have

$$\begin{aligned}
\|S_2\| &\leq \sigma_u^{-2}\|\rho\|\|(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\|\|\Lambda'(P_2 - I_n)\Lambda\|\|\frac{1}{\sqrt{n_1}}F_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}\| \\
&\leq O(1)\sigma_{\min}(\Lambda'\Lambda)^{-1}\|\Lambda'(P_2 - I_n)\Lambda\|\|\frac{1}{\sqrt{n_1}}F_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}\| \\
&= O_P(\sqrt{\frac{p}{\psi_{n,p}n_2}})
\end{aligned}$$

where the last equality follows from Lemma A.3 (iii) and Lemma A.6 (i).

For S_4, its conditional expectation is

$$\begin{aligned}
E(S_4|Z,X) &= E(\frac{1}{\sqrt{n_1}}\boldsymbol{\epsilon}_{d,2}'M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^+Z_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}|Z,X) \\
&= E(\frac{1}{\sqrt{n_1}}\boldsymbol{\epsilon}_{d,2}'M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^+Z_1'M_{X_1}\underbrace{E(\boldsymbol{\epsilon}_{y,1}|Z,X,\boldsymbol{\epsilon}_{d,2})}_{0}|Z,X) = 0
\end{aligned}$$

And its conditional variance is

$$\begin{aligned}
&\mathrm{Var}(S_4|Z,X) \\
&= E(\frac{1}{n_1}\boldsymbol{\epsilon}_{y,1}'M_{X_1}Z_1(Z_2'M_{X_2}Z_2)^+Z_2'M_{X_2}\boldsymbol{\epsilon}_{d,2}\boldsymbol{\epsilon}_{d,2}'M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^+Z_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}|Z,X) \\
&= \sigma_{\epsilon_y}^2\sigma_{\epsilon_d}^2\,\mathrm{tr}((Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}) \\
&= \begin{cases} O_P(\frac{1}{n_2} + \frac{n_2-r}{p}) & \text{if } p > n_2 - r \\ O_P(\frac{p}{n_2-r}) & \text{if } n_2 - r > p \end{cases}
\end{aligned}$$

where the last equality follows from Lemma A.7 (i).

Thus,

$$\|S_4\| = \begin{cases} O_P(\frac{1}{\sqrt{n_2}} + \sqrt{\frac{n_2-r}{p}}) & \text{if } p > n_2 - r \\ O_P(\sqrt{\frac{p}{n_2-r}}) & \text{if } n_2 - r > p \end{cases}$$

In particular, when $p \asymp n_2 - r$, $\|S_4\| = O_P(1)$

For S_5, its conditional expectation is

$$E(S_5|Z,X,U) = \frac{-1}{\sqrt{n_1}}b'U_2'M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^+Z_1'M_{X_1}\underbrace{E(\boldsymbol{\epsilon}_{y,1}|Z,X,U)}_{0} = 0$$

For its conditional variance, by substituting $Z = F\Lambda' + U$ into it, we have

$$\begin{aligned}
\mathrm{Var}(S_5|Z,X,U) &= \sigma_{\epsilon_y}^2 b' U_2' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ \frac{Z_1' M_{X_1} Z_1}{n_1} (Z_2' M_{X_2} Z_2)^+ Z_2' M_{X_2} U_2 b \\
&\leq \sigma_{\epsilon_y}^2 b' U_2' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ \Lambda \frac{F_1' M_{X_1} F_1}{n_1} \Lambda' (Z_2' M_{X_2} Z_2)^+ Z_2' M_{X_2} U_2 b \\
&\quad + \sigma_{\epsilon_y}^2 b' U_2' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ \frac{U_1' M_{X_1} U_1}{n_1} (Z_2' M_{X_2} Z_2)^+ Z_2' M_{X_2} U_2 b \\
&\leq \sigma_{\epsilon_y}^2 \|U_2 b\|^2 \|\Lambda' (Z_2' M_{X_2} Z_2)^+ \Lambda\| \| \frac{F_1' M_{X_1} F_1}{n_1} \| + \sigma_{\epsilon_y}^2 \|U_2 b\|^2 \|(Z_2 M_{X_2} Z_2)^+\| \| \frac{U_1' M_{X_1} U_1}{n_1} \| \\
&= O_P(\frac{1}{\psi_{n,p}})
\end{aligned} \tag{31}$$

where the last equality follows from Lemma A.3 (i)(iv) and Lemma A.4 (i)(ii).

For S_6, its conditional expectation is

$$E(S_6|Z,X) = \frac{1}{\sqrt{n_1}} (\rho - \Lambda'\beta)' (\Lambda'\Lambda)^{-1} \Lambda' P_2 Z_1' M_{X_1} \underbrace{E(\boldsymbol{\epsilon}_{y,1}|Z,X)}_{0} = 0$$

For its conditional variance, by substituting $(\rho - \Lambda'\beta) = \Sigma_{f|x}^{-1} (\Sigma_{f|x}^{-1} + \sigma_u^{-2} \Lambda'\Lambda)^{-1} \rho$ into it, we have

$$\begin{aligned}
\mathrm{Var}(S_6|Z,X) &= (\rho - \Lambda'\beta)' (\Lambda'\Lambda)^{-1} \Lambda' P_2 \frac{Z_1' M_{X_1} Z_1}{n_1} P_2 \Lambda (\Lambda'\Lambda)^{-1} (\rho - \Lambda'\beta) \\
&\leq \sigma_{\epsilon_y}^2 \|\rho\|^2 \|(\Sigma_{f|x}^{-1} + \sigma_u^{-2} \Lambda'\Lambda)^{-1}\|^2 \|\Sigma_{f|x}^{-1}\|^2 \|(\Lambda'\Lambda)^{-1} \Lambda\|^2 \|P_2\|^2 \| \frac{Z_1' M_{X_1} Z_1}{n_1} \| \\
&= O_P(\frac{1}{\psi_{n,p}^2})
\end{aligned} \tag{32}$$

where the last equality follows from Lemma A.3 (ii)(v).

Thus,

$$\|S_6\| = O_P(\frac{1}{\sqrt{\psi_{n,p}}})$$

Therefore, when $p \asymp n_2 - r$, S_1, S_4 are the dominant term in (30) such that

$$\frac{1}{\sqrt{n_1}} \hat{D}_1^{\perp\prime} \boldsymbol{\epsilon}_{y,1}^{\perp} = \frac{1}{\sqrt{n_1}} \beta' Z_1' M_{X_1} \boldsymbol{\epsilon}_{y,1} + \frac{1}{\sqrt{n_1}} \boldsymbol{\epsilon}_{d,2}' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} \boldsymbol{\epsilon}_{y,1} + o_P(1)$$

When $p \gg n_2 - r$, S_1 is the dominant term such that

$$\frac{1}{\sqrt{n_1}} \hat{D}_1^{\perp\prime} \boldsymbol{\epsilon}_{y,1}^{\perp} = \frac{1}{\sqrt{n_1}} \beta' Z_1' M_{X_1} \boldsymbol{\epsilon}_{y,1} + o_P(1)$$

□

Lemma A.9. *As $n, r, p \to \infty$, for each $k \in \{1, 2\}$,*

$$\frac{1}{n_k}\hat{D}_k^{\perp\prime} D_k^{\perp} = \frac{1}{n_k}\beta' P_{k^c} Z_k' M_{X_k} Z_k \beta + o_P(1) \xrightarrow{p} (1-\tau)\beta'\Sigma_{Z^\perp}\beta$$

Proof. Without loss of generality, we consider the case of $k = 1$, i.e. $\frac{1}{n_1}\hat{D}_1^{\perp\prime} D_1^{\perp}$. Let $P_2 = Z_2' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+$. Then by substituting $\hat{D}_1^{\perp} = Z_1^{\perp}\hat{\beta}_2 = M_{X_1} Z_1 (Z_2' M_{X_2} Z_2)^+ Z_2' M_{X_2} D_2$ and $D_1 = Z_1\beta + \boldsymbol{e}_1$ into it, we have

$$\begin{aligned}
\frac{1}{n_1}\hat{D}_1^{\perp\prime} D_1^{\perp} &= \frac{1}{n_1}\beta' P_2 Z_1' M_{X_1} Z_1 \beta + \frac{1}{n_1}\boldsymbol{e}_2' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} Z_1 \beta \\
&\quad + \frac{1}{n_1}\beta' P_2 Z_1' M_{X_1} \boldsymbol{e}_1 + \frac{1}{n_1}\boldsymbol{e}_2' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} \boldsymbol{e}_1 \\
&= \sum_{i=1}^{4} W_i
\end{aligned} \tag{33}$$

For $W_1 = \frac{1}{n_1}\beta' P_2 Z_1' M_{X_1} Z_1 \beta$, we want to show that

$$W_1 \xrightarrow{p} (1-\tau)\beta'\Sigma_{Z^\perp}\beta$$

Given that

$$Z_1' M_{X_1} Z_1 = (n_1 - r)\Lambda\Sigma_{f|x}\Lambda' + H_1 + U_1' M_{X_1} U_1, \quad \text{and} \quad \Sigma_{Z^\perp} = \Lambda\Sigma_{f|x}\Lambda' + \sigma_u^2 I_p$$

where $H_1 = \Lambda(F_1' M_{X_1} F_1 - (n_1 - r)\Sigma_{f|x})\Lambda' + \Lambda F_1' M_{X_1} U_1 + U_1' M_{X_1} F_1 \Lambda'$.Then We have

$$\frac{Z_1' M_{X_1} Z_1}{n_1} - (1 - \frac{r}{n_1})\Sigma_{Z^\perp} = \frac{H_1}{n_1} + \frac{U_1' M_{X_1} U_1}{n_1} - \sigma_u^2(1 - \frac{r}{n_1})I_p$$

Then,

$$\begin{aligned}
\|\beta' P_2(\frac{Z_1' M_{X_1} Z_1}{n_1})\beta - (1 - \frac{r}{n_1})\beta' P_2 \Sigma_{Z^\perp}\beta\| &= \|\beta' P_2(\frac{H_1}{n_1} + \frac{U_1' M_{X_1} U_1}{n_1} - \sigma_u^2(1 - \frac{r}{n_1})I_p)\beta\| \\
&\le \|\beta\|^2 \|P_2\|(\|\frac{H_1}{n_1}\| + \|\frac{U_1' M_{X_1} U_1}{n_1}\| + O(1)) \\
&\le O_P(\frac{1}{\psi_{n,p}})\|(\|\frac{H_1}{n_1}\| + \|\frac{U_1' M_{X_1} U_1}{n_1}\| + O(1)) \\
&= O_P(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{\psi_{n,p}}} + \sqrt{\frac{p}{\psi_{n,p} n_1}}) = o_P(1)
\end{aligned} \tag{34}$$

where the last equality follows from the proof of Lemma A.3 (ii) that $\|\frac{1}{n_1}H_1\| = O_P(\frac{\psi_{n,p}}{\sqrt{n_1}} + \sqrt{\psi_{n,p}} + \sqrt{\frac{p\psi_{n,p}}{n_1}})$ and Lemma A.3 (i).

Next, if $n_2 - r > p$, we have $P_2 = I_p$ so that $\beta' P_2 \Sigma_{Z^\perp}\beta = \beta'\Sigma_{Z^\perp}\beta$. When $p > n_2 - r$, Lemma A.6 (ii) gives $\|\beta' P_2 \Sigma_{Z^\perp}\beta - \beta'\Sigma_{Z^\perp}\beta\| = o_P(1)$.

Thus,

$$\begin{aligned}\|\beta' P_2(\frac{Z_1' M_{X_1} Z_1}{n_1})\beta - (1-\frac{r}{n_1})\beta'\Sigma_{Z^\perp}\beta\| &\le \|\beta' P_2(\frac{Z_1' M_{X_1} Z_1}{n_1})\beta - (1-\frac{r}{n_1})\beta' P_2\Sigma_{Z^\perp}\beta\| \\ &\quad + (1-\frac{r}{n_1})\|\beta' P_2\Sigma_{Z^\perp}\beta - \beta'\Sigma_{Z^\perp}\beta\| \\ &= o_P(1)\end{aligned}$$

Therefore,

$$W_1 = \frac{1}{n_1}\beta' P_2 Z_1' M_{X_1} Z_1 \beta \xrightarrow{p} (1-\tau)\beta'\Sigma_{Z^\perp}\beta$$

Next, from (29)

$$\begin{aligned}\|M_X \boldsymbol{e} - M_X \boldsymbol{\epsilon}_d\| &= \| - M_X U b + M_X Z\Lambda(\Lambda'\Lambda)^{-1}(\rho - \Lambda'\beta)\| \\ &\le \|M_X U b\| + \|M_X Z\Lambda(\Lambda'\Lambda)^{-1}(\rho - \Lambda'\beta)\| \\ &= \|M_X U b\| + \|M_X Z\Lambda(\Lambda'\Lambda)^{-1}\Sigma_{f|x}^{-1}(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho\| \\ &\le \|M_X\|\|U b\| + \|Z M_X Z\|^{1/2}\|\Lambda\|\sigma_{\min}(\Lambda'\Lambda)^{-2}\|\Sigma_{f|x}^{-1}\|\|\rho\| \\ &= O_P(\sqrt{\frac{n}{\psi_{n,p}}})\end{aligned} \tag{35}$$

where the second equality follows from $(\rho - \Lambda'\beta) = \Sigma_{f|x}^{-1}(\Sigma_{f|x}^{-1} + \sigma_u^{-2}\Lambda'\Lambda)^{-1}\rho$, the last equality follows from M_X is a projection matrix and Lemma A.3 (ii).

By substituting $M_X \boldsymbol{e} - M_X \boldsymbol{\epsilon}_d = -M_X U b + M_X Z\Lambda(\Lambda'\Lambda)^{-1}(\rho - \Lambda'\beta)$ into W_2, W_3, W_4,

we have

$$
\begin{aligned}
W_2 &= \frac{1}{n_1}\boldsymbol{\epsilon}_{d,2}' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} Z_1 \beta + \frac{1}{n_1}(\boldsymbol{e}_2 - \boldsymbol{\epsilon}_{d,2})' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} Z_1 \beta \\
&= \sum_{j=1}^{2} W_{2,j} \\
W_3 &= \frac{1}{n_1}\beta' P_2 Z_1' M_{X_1} \boldsymbol{\epsilon}_{d,1} + \frac{1}{n_1}\beta' P_2 Z_1' M_{X_1}(\boldsymbol{e}_1 - \boldsymbol{\epsilon}_{d,1}) \\
W_4 &= \frac{1}{n_1}\boldsymbol{\epsilon}_{d,2}' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} \boldsymbol{\epsilon}_{d,1} + \frac{1}{n_1}\boldsymbol{\epsilon}_{d,2}' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1}(\boldsymbol{e}_1 - \boldsymbol{\epsilon}_{d,1}) \\
&\quad + \frac{1}{n_1}(\boldsymbol{e}_2 - \boldsymbol{\epsilon}_{d,2})' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} \boldsymbol{\epsilon}_{d,1} + \frac{1}{n_1}(\boldsymbol{e}_2 - \boldsymbol{\epsilon}_{d,2})' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1}(\boldsymbol{e}_1 - \boldsymbol{\epsilon}_{d,1}) \\
&= \sum_{j=1}^{4} W_{4,j}
\end{aligned}
$$

For the first term $W_{2,1}$ in W_2, its conditional expectation is

$$
E(W_{2,1}|X,Z) = \frac{1}{n_1}\underbrace{E(\boldsymbol{\epsilon}_{d,2}'|Z,X)}_{0} M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} Z_1 \beta = 0
$$

Then following from Lemma A.7 (iv), its conditional variance is

$$
\begin{aligned}
\mathrm{Var}(W_{2,1}|Z,X) &= \frac{1}{n_1^2}\beta' Z_1' M_{X_1} Z_1 (Z_2' M_{X_2} Z_2)^+ Z_2 M_{X_2} E(\boldsymbol{\epsilon}_{d,2}\boldsymbol{\epsilon}_{d,2}'|Z,X) M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} Z_1 \beta \\
&= \sigma_{\epsilon_d}^2 \|\beta\|^2 \|\frac{1}{n_1^2} Z_1' M_{X_1} Z_1 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} Z_1\| = O_P(\frac{1}{n_2 - r})
\end{aligned}
$$

Thus, $\|W_{2,1}\| = O_P(\frac{1}{\sqrt{n_2 - r}})$.

For the second and third term in W_2, following from Lemma A.7 (iv)

$$
\|W_{2,2}\| \le \|M_X \boldsymbol{e}_2 - M_X \boldsymbol{\epsilon}_{d,2}\| \|\frac{1}{n_1^2} Z_1' M_{X_1} Z_1 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} Z_1\|^{1/2} \|\beta\| = O_P(\frac{1}{\sqrt{\psi_{n,p}}})
$$

Hence, $\|W_2\| = o_P(1)$.

For W_3, following from Lemma A.3 (ii)(iii),

$$
\begin{aligned}
\|W_3\| &\le \|\beta\| \|P_2\| \Big(\frac{1}{\sqrt{n_1}} \|\frac{1}{\sqrt{n_1}} Z_1' M_{X_1} \boldsymbol{\epsilon}_{d,1}\| + \frac{1}{\sqrt{n_1}} \|\frac{Z_1' M_{X_1} Z_1}{n_1}\|^{1/2} \|M_{X_1} \boldsymbol{e}_1 - M_{X_1} \boldsymbol{\epsilon}_{d,1}\| \Big) \\
&= O_P(\frac{1}{\sqrt{n_1}}) + O_P(\frac{1}{\sqrt{\psi_{n,p}}})
\end{aligned}
$$

For the first term $W_{4,1}$ in W_4, its conditional expectation is

$$\begin{aligned} E(W_{4,1}|Z,X) &= E(\frac{1}{n_1}\boldsymbol{\epsilon}'_{d,2}M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^+Z_1'M_{X_1}\boldsymbol{\epsilon}_{d,1}|Z,X) \\ &= E(\frac{1}{n_1}\boldsymbol{\epsilon}'_{d,2}M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^+Z_1'M_{X_1}\underbrace{E(\boldsymbol{\epsilon}_{d,1}|Z,X,\boldsymbol{\epsilon}_{d,2})}_{0}|Z,X) = 0 \end{aligned}$$

And its conditional variance is

$$\begin{aligned} \mathrm{Var}(W_{4,1}|Z,X) &= \frac{1}{n_1}\sigma^4_{\epsilon_d}\,\mathrm{tr}((Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}) \\ &= o_P(1) \end{aligned}$$

where the last equality follows from Lemma A.7 (i).

Thus, $W_{4,1} = o_P(1)$.

For the second term $W_{4,2}$ in W_4, its conditional expectation is

$$E(W_{4,2}|Z,X,\boldsymbol{e}_1,\boldsymbol{\epsilon}_{d,1}) = \frac{1}{n_1}\underbrace{E(\boldsymbol{\epsilon}'_{d,2}|Z,X,\boldsymbol{e}_1,\boldsymbol{\epsilon}_{d,1})}_{0}M_{X_2}Z_2(Z_2'M_{X_2}Z_2)^+Z_1'M_{X_1}(\boldsymbol{e}_1-\boldsymbol{\epsilon}_{d,1}) = 0$$

And its conditional variance is

$$\begin{aligned} Var(W_{4,2}|Z,X,\boldsymbol{e}_1,\boldsymbol{\epsilon}_{d,1}) &= \frac{1}{n_1^2}\sigma^2_{\epsilon_d}(\boldsymbol{e}_1-\boldsymbol{\epsilon}_{d,1})'M_{X_1}Z_1(Z_2'M_{X_2}Z_2)^+Z_1'M_{X_1}(\boldsymbol{e}_1-\boldsymbol{\epsilon}_{d,1}) \\ &\le \frac{1}{n_1}\sigma^2_{\epsilon_d}\|\boldsymbol{e}_1-\boldsymbol{\epsilon}_{d,1}\|^2\|\frac{Z_1'M_{X_1}Z_1}{n_1}\|\|(Z_2'M_{X_2}Z_2)^+\| \\ &= \begin{cases} O_P(\frac{1}{p}), & \text{if } p > n_2 - r \\ O_P(\frac{1}{n_2-r}), & \text{if } n_2 - r > p \end{cases} \end{aligned}$$

where the last equality follows from Lemma A.3 (ii) and Lemma A.4 (i).

Thus, $\|W_{4,2}\| = o_P(1)$.

For $W_{4,3}$ and $W_{4,4}$ in W_4,

$$\begin{aligned}
\|W_{4,3}\| &\leq \frac{1}{\sqrt{n_1}}\|\boldsymbol{e}_2 - \boldsymbol{\epsilon}_{d,2}\|\|(Z_2'M_{X_2}Z_2)^+\|^{1/2}\|\frac{1}{\sqrt{n_1}}Z_1'M_{X_1}\boldsymbol{\epsilon}_{d,1}\| \\
&= \begin{cases} O_P(\frac{1}{\sqrt{p}}), \text{if } p > n_2 - r \\ O_P(\frac{1}{\sqrt{n_2-r}}), \text{if } n_2 - r > p \end{cases} \\
\|W_{4,4}\| &\leq \frac{1}{\sqrt{n_1}}\|\boldsymbol{e}_2 - \boldsymbol{\epsilon}_{d,2}\|\|(Z_2'M_{X_2}Z_2)^+\|^{1/2}\|\frac{Z_1'M_{X_1}Z_1}{n_1}\|^{1/2}\|\boldsymbol{e}_1 - \boldsymbol{\epsilon}_{d,1}\| \\
&= \begin{cases} O_P(\sqrt{\frac{n_2}{\psi_{n,p}p}}), & \text{if } p > n_2 - r \\ O_P(\frac{1}{\sqrt{\psi_{n,p}}}), & \text{if } n_2 - r > p \end{cases}
\end{aligned}$$

Hence, $\|W_4\| = o_P(1)$.

Therefore, since W_1 is the dominant term in (33), we have

$$\frac{1}{n_1}\hat{D}_1^{\perp\prime}D_1^{\perp} = \frac{1}{n_1}\beta'P_2Z_1'M_{X_1}Z_1\beta + o_P(1) \xrightarrow{p} (1-\tau)\beta'\Sigma_{Z^\perp}\beta$$

□

A.4 Asymptotic normality

In this section, we derive the asymptotic normality of $\sqrt{n}(\hat{\alpha} - \alpha)$ in Theorem 3. As we have shown the dominant terms of $\hat{\alpha}$ in Lemma A.8 and Lemma A.9, the following lemma is useful in establishing the asymptotic normality of $\sqrt{n}(\hat{\alpha} - \alpha)$.

Lemma A.10. *As $n, r, p \to \infty$, $\frac{p}{n/2-r} \to \gamma \in (0,\infty) \setminus \{1\}$,*

$$\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp} + \frac{1}{\sqrt{n_2}}\hat{D}_2^{\perp\prime}\boldsymbol{\epsilon}_{y,2}^{\perp} \xrightarrow{d} N(0, 2\Omega)$$

where $\Omega = \sigma_{\epsilon_y}^2 V + \Gamma + \Pi$, $V = (1-\tau)\beta'\Sigma_{Z^\perp}\beta$, $\Gamma = \begin{cases} \sigma_{\epsilon_y}^2\sigma_{\epsilon_d}^2(1-\tau)\frac{\gamma}{1-\gamma}, & 0 < \gamma < 1 \\ \sigma_{\epsilon_y}^2\sigma_{\epsilon_d}^2(1-\tau)\frac{1}{\gamma-1}, & \gamma > 1 \end{cases}$,

$\Pi = \begin{cases} \sigma^2(1-\tau)\gamma, & 0 < \gamma < 1 \\ \sigma^2(1-\tau)\frac{1}{\gamma}, & \gamma > 1 \end{cases}$

In particular, as $\frac{p}{n/2-r} \to \infty$, Ω reduces to $\sigma_{\epsilon_y}^2 V$

Proof. From Lemma A.8, we know

$$\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp}+\frac{1}{\sqrt{n_2}}\hat{D}_2^{\perp\prime}\boldsymbol{\epsilon}_{y,2}^{\perp}=\frac{1}{\sqrt{n_1}}\beta' Z_1' M_{X_1}\boldsymbol{\epsilon}_{y,1}+\frac{1}{\sqrt{n_1}}\boldsymbol{\epsilon}_{d,2}' M_{X_2}Z_2(Z_2' M_{X_2}Z_2)^{+}Z_1' M_{X_1}\boldsymbol{\epsilon}_{y,1}$$
$$+\frac{1}{\sqrt{n_2}}\beta' Z_2' M_{X_2}\boldsymbol{\epsilon}_{y,2}+\frac{1}{\sqrt{n_2}}\boldsymbol{\epsilon}_{d,1}' M_{X_1}Z_1(Z_1' M_{X_1}Z_1)^{+}Z_2' M_{X_2}\boldsymbol{\epsilon}_{y,2}+o_P(1)$$

Let $\mathcal{F}_i$ be the σ-field generated by $\{(Z,X,\epsilon_{d,j},\epsilon_{y,j}),j\le i\}$. Then,

$$\xi_i=\begin{cases}\frac{1}{\sqrt{n_1}}\beta' Z_1' m_i\epsilon_{y,i}, & 1\le i\le n_1,\\ \frac{1}{\sqrt{n_2}}\beta' Z_2' m_i\epsilon_{y,i}+\frac{1}{\sqrt{n_1}}\epsilon_{d,i}\, m_i' Z_2(Z_2' M_{X_2}Z_2)^{+}Z_1' M_{X_1}\boldsymbol{\epsilon}_{y,1} & \\ \quad+\frac{1}{\sqrt{n_2}}\boldsymbol{\epsilon}_{d,1}' M_{X_1}Z_1(Z_1' M_{X_1}Z_1)^{+}Z_2' m_i\epsilon_{y,i}, & n_1+1\le i\le n,\end{cases}$$

where m_i denotes the i-th column of M_{X_1} for $i\le n_1$, and the $(i-n_1)$-th column of M_{X_2} for $n_1+1\le i\le n$. It is straightforward to verify that $E(\xi_i|\mathcal{F}_{i-1})=0$, then ξ_i is a martingale difference array with respect to the filtration $\mathcal{F}$.

Step 1: We want to show that $\sum_{i=1}^n E(\xi_i^2|\mathcal{F}_{i-1})\xrightarrow{p}2(V+\Gamma+\Pi)$

where $V=\sigma_{\epsilon_y}^2(1-\tau)\beta'\Sigma_{Z^\perp}\beta$, $\Gamma=\begin{cases}\sigma_{\epsilon_y}^2\sigma_{\epsilon_d}^2(1-\tau)\frac{\gamma}{1-\gamma}, & 0<\gamma<1\\ \sigma_{\epsilon_y}^2\sigma_{\epsilon_d}^2(1-\tau)\frac{1}{\gamma-1}, & \gamma>1\end{cases}$, $\Pi=\begin{cases}\sigma^2(1-\tau)\gamma, & 0<\gamma<1\\ \sigma^2(1-\tau)\frac{1}{\gamma}, & \gamma>1\end{cases}$

$$\sum_{i=1}^n E(\xi_i^2|\mathcal{F}_{i-1})=\frac{\sigma_{\epsilon_y}^2}{n_1}\beta' Z_1' M_{X_1}Z_1\beta+\frac{\sigma_{\epsilon_y}^2}{n_2}\beta' Z_2' M_{X_2}Z_2\beta+\frac{\sigma_{\epsilon_d}^2}{n_1}\boldsymbol{\epsilon}_{y,1}' M_{X_1}Z_1(Z_2' M_{X_2}Z_2)^{+}Z_1' M_{X_1}\boldsymbol{\epsilon}_{y,1}$$
$$+\frac{\sigma_{\epsilon_y}^2}{n_2}\boldsymbol{\epsilon}_{d,1}' M_{X_1}Z_1(Z_1' M_{X_1}Z_1)^{+}Z_2' M_{X_2}Z_2(Z_1' M_{X_1}Z_1)^{+}Z_1' M_{X_1}\boldsymbol{\epsilon}_{d,1}+\frac{2\sigma}{\sqrt{n_1n_2}}\beta' P_2 Z_1' M_{X_1}\boldsymbol{\epsilon}_{y,1}$$
$$+\frac{2\sigma}{\sqrt{n_1n_2}}\boldsymbol{\epsilon}_{d,1}' M_{X_1}Z_1(Z_1' M_{X_1}Z_1)^{+}P_2 Z_1' M_{X_1}\boldsymbol{\epsilon}_{y,1}+\frac{2\sigma_{\epsilon_y}^2}{n_2}\beta' Z_2' M_{X_2}Z_2(Z_1' M_{X_1}Z_1)^{+}Z_1' M_{X_1}\boldsymbol{\epsilon}_{y,1}$$
$$=\sum_{i=1}^7 R_i$$

where $\sigma=E(\epsilon_{y,i}\epsilon_{d,i})$.

For R_1, by Lemma A.6 (ii) and the similar proof of (34) in Lemma A.9, we have

$$R_1\xrightarrow{p}\sigma_{\epsilon_y}^2 V=\sigma_{\epsilon_y}^2(1-\tau)\beta'\Sigma_{Z^\perp}\beta$$

By the same argument, $R_2\xrightarrow{p}\sigma_{\epsilon_y}^2 V$

For R_3, its conditional expectation is

$$E(R_3|Z,X) = \sigma^2_{\epsilon_y}\sigma^2_{\epsilon_d}\operatorname{tr}((Z_2'M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}) \xrightarrow{p} \Gamma = \begin{cases} \sigma^2_{\epsilon_y}\sigma^2_{\epsilon_d}(1-\tau)\frac{\gamma}{1-\gamma}, & 0<\gamma<1 \\ \sigma^2_{\epsilon_y}\sigma^2_{\epsilon_d}(1-\tau)\frac{1}{\gamma-1}, & \gamma>1 \end{cases}$$

where the probability convergence follows from Lemma A.7 (i).

Since $\epsilon_{y,i}$ are conditional i.i.d and has finite fourth moment, its conditional variance is

$$\begin{aligned}\operatorname{Var}(R_3|Z,X) &\le C\operatorname{tr}((Z_2M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}(Z_2M_{X_2}Z_2)^+\frac{Z_1'M_{X_1}Z_1}{n_1}) \\ &= o_P(1)\end{aligned}$$

where last equality follows from Lemma A.7 (iii). Thus,

$$R_3 \xrightarrow{p} \Gamma = \begin{cases} \sigma^2_{\epsilon_y}\sigma^2_{\epsilon_d}(1-\tau)\frac{\gamma}{1-\gamma}, & 0<\gamma<1 \\ \sigma^2_{\epsilon_y}\sigma^2_{\epsilon_d}(1-\tau)\frac{1}{\gamma-1}, & \gamma>1 \end{cases}$$

By similar argument, $R_4 \xrightarrow{p} \Gamma$.

For R_5, we have

$$R_5 \le 2\sigma\frac{1}{\sqrt{n_2}}\|\beta\|\|P_2\|\|\frac{1}{\sqrt{n_1}}Z_1'M_{X_1}\boldsymbol{\epsilon}_{y,1}\| = O_P(\frac{1}{\sqrt{n_2}})$$

For R_6, following from Lemma A.7 (ii), its conditional expectation is

$$E(R_6|Z,X) = \frac{2\sigma^2}{\sqrt{n_1n_2}}tr(P_1P_2) \xrightarrow{p} \begin{cases} 2\sigma^2(1-\tau)\gamma, & 0<\gamma<1 \\ 2\sigma^2(1-\tau)\frac{1}{\gamma}, & \gamma>1 \end{cases}$$

Since $\epsilon_{y,i}$ and $\epsilon_{d,i}$ are conditional i.i.d and have finite fourth moment, its conditional variance is

$$Var(R_6|Z,X) \le \frac{C}{n_1n_2}tr(P_1P_2P_1P_2) \le \frac{C}{n_1n_2}tr(P_1P_2) = o_P(1)$$

where the second inequality follows from P_1 and P_2 are projection matrices. The last equality follows from Lemma A.7 (ii). Thus,

$$R_6 \xrightarrow{p} \begin{cases} 2\sigma^2(1-\tau)\gamma, & 0<\gamma<1 \\ 2\sigma^2(1-\tau)\frac{1}{\gamma}, & \gamma>1 \end{cases}$$

For R_7, by the similar proof of $W_{2,1}$ in Lemma A.9, we have

$$R_7 = O_P(\frac{1}{\sqrt{n_2}})$$

Therefore,

$$\sum_{i=1}^{n} E(\xi_i^2|\mathcal{F}_{i-1}) \xrightarrow{p} 2(V + \Gamma + \Pi)$$

Step 2: By Lyapunov condition, we want to show for some $\delta > 0$,

$$\sum_{i=1}^{n} E(|\xi_i|^{2+\delta}|\mathcal{F}_{i-1}) \to 0$$

Taking $\delta = 2$, we have

$$\begin{aligned}
\sum_{i=1}^{n} E(|\xi_i|^4|\mathcal{F}_{i-1}) \leq C\Big(&\frac{1}{n_1^2}\sum_{i=1}^{n_1} E((\beta' Z_1' m_i \epsilon_{y,i})^4|\mathcal{F}_{i-1}) + \frac{1}{n_2^2}\sum_{i=n_1+1}^{n} E((\beta' Z_2' m_i \epsilon_{y,i})^4|\mathcal{F}_{i-1}) \\
&+ \frac{1}{n_1^2}\sum_{i=n_1+1}^{n} E((\epsilon_{d,i}\, m_i' Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} \boldsymbol{\epsilon}_{y,1})^4|\mathcal{F}_{i-1}) \\
&+ \frac{1}{n_2^2}\sum_{i=n_1+1}^{n} E((\boldsymbol{\epsilon}_{d,1}' M_{X_1} Z_1 (Z_1' M_{X_1} Z_1)^+ Z_2' m_i \epsilon_{y,i})^4|\mathcal{F}_{i-1})\Big) \\
= C&\sum_{i=1}^{4} D_i
\end{aligned}$$

For D_1, with $E(\epsilon_{y,i}^4|F, U, X) \leq C$ and $Z = F\Lambda' + U$, we have

$$\begin{aligned}
D_1 &\leq \frac{1}{n_1^2}\sum_{i=1}^{n_1} (\beta' Z_1' m_i)^4 E(\epsilon_{y,i}^4|Z, X) \\
&\leq C\frac{1}{n_1^2}\sum_{i=1}^{n_1} (\beta'\Lambda F_1' m_i)^4 + C\frac{1}{n_1^2}\sum_{i=1}^{n_1} (\beta' U_1' m_i)^4
\end{aligned} \tag{36}$$

For the first term in (36), we have

$$\begin{aligned}
\frac{1}{n_1^2}\sum_{i=1}^{n_1} (\beta'\Lambda F_1' m_i)^4 &\leq \|\beta\|^4\|\Lambda\|^4 \|\frac{1}{n_1} F_1' \sum_{i=1}^{n_1} m_i m_i' F_1\| \max_i \|\frac{1}{\sqrt{n_1}} F_1' m_i\|^2 \\
&= \|\beta\|^4\|\Lambda\|^4 \|\frac{1}{n_1} F_1' M_{X_1} F_1\| \max_i \|\frac{1}{\sqrt{n_1}} F_1' m_i\|^2 \\
&= O_P(1) o_P(1) = o_P(1)
\end{aligned} \tag{37}$$

where the last line follows from Assumption 4 (i).

For the second term in (36), let $d_i = U_1' m_i$, first observe that

$$\begin{aligned} E(d_{ij}^4|X) &= E((\sum_{l=1}^{n_1} m_{il}u_{lj})^4|X) \\ &\leq CE((\sum_{l=1}^{n_1} m_{il}^2 u_{lj}^2)^2|X) \\ &\leq C\Big(\sum_{l=1}^{n_1} m_{il}^4 E(u_{lj}^4) + \sum_{l\neq t} m_{il}^2 m_{it}^2 E(u_{lj}^2)E(u_{tj}^2)\Big) \\ &\leq C' m_{ii} \end{aligned}$$

where the last inequality follows from $E(u_{it}^4) \leq C$ and $m_{ij}^4 \leq m_{ij}^2$, $\sum_{j=1}^{n_1} m_{ij}^2 = m_{ii}$

Also,

$$\begin{aligned} E(d_{ij}^2 d_{ik}^2|X) &= E(d_{ij}^2|X)E(d_{ik}^2|X) \\ &= (\sum_{l=1}^{n_1} m_{il}^2 E(u_{lj}^2))(\sum_{t=1}^{n_1} m_{it}^2 E(u_{tk}^2)) \\ &= \sigma_u^4 (\sum_{l=1}^{n_1} m_{il}^2)^2 \\ &= \sigma_u^4 m_{ii}^2 \\ &\leq \sigma_u^4 m_{ii} \end{aligned}$$

where the first equality follows from u_{ij} is independent across (i, j).

Thus,

$$\begin{aligned} \frac{1}{n_1^2}\sum_{i=1}^{n_1} E((\beta' U_1' m_i)^4|X) &= \frac{1}{n_1^2}\sum_{i=1}^{n_1} E((\beta' d_i)^4|X) \\ &= \frac{1}{n_1^2}\sum_{i=1}^{n_1}\Big(\sum_{j=1}^{p}\beta_j^4 E(d_{ij}^4|X) + 3\sum_{j\neq k}\beta_j^4\beta_k^4 E(d_{ij}^4 d_{ik}^4|X)\Big) \\ &\leq C\frac{1}{n_1^2}\sum_{i=1}^{n_1} m_{ii}(\sum_{j=1}^{p}\beta_j^4 + \sum_{j\neq k}\beta_j^2\beta_k^2) \\ &= C\frac{1}{n_1^2}\sum_{i=1}^{n_1} m_{ii}\|\beta\|^4 \\ &= C\frac{n_1 - r}{n_1^2}\|\beta\|^4 \\ &= o_P(1) \end{aligned}$$

where the last second equality follows from $\sum_i^{n_1} m_{ii} = n_1 - r$.

Hence,

$$D_1 \leq C\frac{1}{n_1^2}\sum_{i=1}^{n_1}(\beta'\Lambda F_1' m_i)^4 + C\frac{1}{n_1^2}\sum_{i=1}^{n_1}(\beta' U_1' m_i)^4 \xrightarrow{p} 0$$

By the same argument,

$$D_2 \xrightarrow{p} 0$$

For D_3, let $\omega_{ij} = m_i' Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' m_j$, we have

$$\begin{aligned} D_3 &= \frac{1}{n_1^2}\sum_{i=n_1+1}^{n} E((\epsilon_{d,i}\sum_{j=1}^{n_1}\omega_{ij}\epsilon_{y,j})^4|\mathcal{F}_{i-1}) \\ &= \frac{1}{n_1^2}\sum_{i=n_1+1}^{n} E(\epsilon_{d,i}^4|Z,X)(\sum_{j=1}^{n_1}\omega_{ij}\epsilon_{y,j})^4 \\ &\leq C\frac{1}{n_1^2}\sum_{i=n_1+1}^{n}(\sum_{j=1}^{n_1}\omega_{ij}\epsilon_{y,j})^4 \end{aligned} \tag{38}$$

where the first equality and inequality follow from $(\epsilon_{d,i}, \epsilon_{y,i})$ are i.i.d and $E(\epsilon_{d,i}^4|Z,X) \leq C$. Then

$$\begin{aligned} \frac{1}{n_1^2}\sum_{i=n_1+1}^{n} E((\sum_{j=1}^{n_1}\omega_{ij}\epsilon_{y,j})^4|Z,X) &= \frac{1}{n_1^2}\sum_{i=n_1+1}^{n}\Big(\sum_{j=1}^{n_1}\omega_{ij}^4 E(\epsilon_{y,j}^4|Z,X) + \sum_{j=1}^{n_1}\sum_{k=1,j\neq k}^{n_1}\omega_{ij}^2\omega_{ik}^2 E(\epsilon_{y,j}^2|Z,X)E(\epsilon_{y,k}^2|Z,X) \\ &\leq C\frac{1}{n_1^2}\sum_{i=n_1+1}^{n}(\sum_{j=1}^{n_1}\omega_{ij}^2)^2 \\ &= C\frac{1}{n_1^2}\sum_{i=n_1+1}^{n}(m_i' Z_2(Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} Z_1 (Z_2' M_{X_2} Z_2)^+ Z_2' m_i)^2 \\ &\leq C\frac{1}{n_1^2}\|M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ Z_1' M_{X_1} Z_1 (Z_2' M_{X_2} Z_2)^+ Z_2' M_{X_2}\|_F^2 \\ &= C\,\mathrm{tr}((Z_2' M_{X_2} Z_2)^+ \frac{Z_1' M_{X_1} Z_1}{n_1}(Z_2' M_{X_2} Z_2)^+ \frac{Z_1' M_{X_1} Z_1}{n_1}) \\ &\xrightarrow{p} 0 \end{aligned} \tag{39}$$

where the first equality and inequality follow from $(\epsilon_{y,i}, \epsilon_{y,i})$ are i.i.d and $E(\epsilon_{y,i}^4|Z,X) \leq C$. The last line follows from Lemma A.7 (iii).

Hence,

$$D_3 \xrightarrow{p} 0$$

By the same argument,

$$D_4 \xrightarrow{p} 0$$

Step 3: From which we obtain the Lyapunov condition. Hence, by the Martingale Central Limit Theorem,

$$\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp} + \frac{1}{\sqrt{n_1}}\hat{D}_2^{\perp\prime}\boldsymbol{\epsilon}_{y,2}^{\perp} \xrightarrow{d} N(0, 2\Omega)$$

where $\Omega = \sigma_{\epsilon_y}^2 V + \Gamma + \Pi$, $V = (1-\tau)\beta'\Sigma_{Z^\perp}\beta$, $\Gamma = \begin{cases} \sigma_{\epsilon_y}^2\sigma_{\epsilon_d}^2(1-\tau)\frac{\gamma}{1-\gamma}, & 0 < \gamma < 1 \\ \sigma_{\epsilon_y}^2\sigma_{\epsilon_d}^2(1-\tau)\frac{1}{\gamma-1}, & \gamma > 1 \end{cases}$,

$\Pi = \begin{cases} \sigma^2(1-\tau)\gamma, & 0 < \gamma < 1 \\ \sigma^2(1-\tau)\frac{1}{\gamma}, & \gamma > 1 \end{cases}$

In particular, as $\frac{p}{n/2-r} \to \infty$, Ω reduces to $\sigma_{\epsilon_y}^2 V$

□

A.4.1 Proof of Theorem 3

Proof of Theorem 3. Since $\hat{\alpha} = \frac{1}{2}(\hat{\alpha}_1 + \hat{\alpha}_2)$, we can decompose $\sqrt{n}(\hat{\alpha} - \alpha)$ in terms of the subsample estimators:

$$\begin{aligned}
\sqrt{n}(\hat{\alpha} - \alpha) &= \sqrt{n}\frac{1}{2}(\hat{\alpha}_1 - \alpha) + \sqrt{n}\frac{1}{2}(\hat{\alpha}_2 - \alpha) \\
&= \sqrt{\frac{n}{n_1}}\frac{1}{2}\sqrt{n_1}(\hat{\alpha}_1 - \alpha) + \sqrt{\frac{n}{n_2}}\frac{1}{2}\sqrt{n_2}(\hat{\alpha}_2 - \alpha) \\
&= \frac{1}{\sqrt{2}}\sqrt{n_1}(\hat{\alpha}_1 - \alpha) + \frac{1}{\sqrt{2}}\sqrt{n_2}(\hat{\alpha}_2 - \alpha)
\end{aligned}$$

We now analyze the asymptotic properties of the subsample estimators. Without loss of generality, we focus on $\hat{\alpha}_1$. The arguments for $\hat{\alpha}_2$ follow analogously.

Recall that $\hat{\alpha}_1 = (\frac{1}{n_1}\hat{D}_1^{\perp\prime}D_1^{\perp})^{-1}\frac{1}{n_1}\hat{D}_1^{\perp\prime}Y_1^{\perp}$. Then we have

$$\sqrt{n_1}(\hat{\alpha}_1 - \alpha) = (\frac{1}{n_1}\hat{D}_1^{\perp\prime}D_1^{\perp})^{-1}\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp}$$

From Lemma A.9 that $\frac{1}{n_1}\hat{D}_1^{\perp}D_1^{\perp} \xrightarrow{p} V = (1-\tau)\beta'\Sigma_{Z^\perp}\beta$,

$$\sqrt{n_1}(\hat{\alpha}_1 - \alpha) = (\frac{1}{n_1}\hat{D}_1^{\perp\prime}D_1^{\perp})^{-1}\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp} = V^{-1}\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp} + o_P(1)$$

From Lemma A.10 that $\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp} + \frac{1}{\sqrt{n_2}}\hat{D}_2^{\perp\prime}\boldsymbol{\epsilon}_{y,2}^{\perp} \xrightarrow{d} N(0, 2\Omega)$ where $\Omega = \sigma_{\epsilon_y}^2 V + \Gamma + \Pi$.

By Slutsky's theorem

$$\begin{aligned}\sqrt{n}(\hat{\alpha}-\alpha) &= \frac{1}{\sqrt{2}}(\frac{1}{n_1}\hat{D}_1^{\perp\prime}D_1^{\perp})^{-1}\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp}+\frac{1}{\sqrt{2}}(\frac{1}{n_2}\hat{D}_2^{\perp\prime}D_2^{\perp})^{-1}\frac{1}{\sqrt{n_2}}\hat{D}_2^{\perp\prime}\boldsymbol{\epsilon}_{y,2}^{\perp}\\ &= \frac{1}{\sqrt{2}}V^{-1}(\frac{1}{\sqrt{n_1}}\hat{D}_1^{\perp\prime}\boldsymbol{\epsilon}_{y,1}^{\perp}+\frac{1}{\sqrt{n_2}}\hat{D}_2^{\perp\prime}\boldsymbol{\epsilon}_{y,2}^{\perp})+o_P(1)\\ &\xrightarrow{d} N(0,\Sigma)\end{aligned}$$

where $\Sigma = V^{-1}\Omega V^{-1} = \sigma_{\epsilon_y}^2 V^{-1}+V^{-1}(\Gamma+\Pi)V^{-1}$, $V=(1-\tau)\beta'\Sigma_{Z^\perp}\beta$, $\Gamma = \begin{cases}\sigma_{\epsilon_y}^2\sigma_{\epsilon_d}^2(1-\tau)\frac{\gamma}{1-\gamma}, & 0<\gamma<1\\ \sigma_{\epsilon_y}^2\sigma_{\epsilon_d}^2(1-\tau)\frac{1}{\gamma-1}, & \gamma>1\end{cases}$,

$\Pi = \begin{cases}\sigma^2(1-\tau)\gamma, & 0<\gamma<1\\ \sigma^2(1-\tau)\frac{1}{\gamma}, & \gamma>1\end{cases}$

In particular, by Lemma A.10, as $\frac{p}{n/2-r}\to\infty$, Σ reduces to $\sigma_{\epsilon_y}^2 V^{-1}$

□

A.5 Consistent estimator for the asymptotic variance Σ

In this section, we establish the consistency of the proposed estimator $\hat{\Sigma}$ in Theorem 4 for the asymptotic variance of $\sqrt{n}(\hat{\alpha}-\alpha)$. The following lemmas show that each component of $\hat{\Sigma}$ consistently estimates its corresponding component in Σ.

Lemma A.11.

$$\frac{1}{n}\boldsymbol{\epsilon}_y'M_X\boldsymbol{\epsilon}_y \xrightarrow{p} \sigma_{\epsilon_y}^2(1-\frac{\tau}{2}),\quad \frac{1}{n}\boldsymbol{\epsilon}_d'M_X\boldsymbol{\epsilon}_d \xrightarrow{p} \sigma_{\epsilon_d}^2(1-\frac{\tau}{2}),\quad \frac{1}{n}\boldsymbol{\epsilon}_y'M_X\boldsymbol{\epsilon}_d \xrightarrow{p} \sigma^2(1-\frac{\tau}{2})$$

Proof. For $\frac{1}{n}\boldsymbol{\epsilon}_y'M_X\boldsymbol{\epsilon}_y$, following from Lemma A.1, its conditional expectation is

$$\frac{1}{n}E(\boldsymbol{\epsilon}_y'M_X\boldsymbol{\epsilon}_y|X) = \frac{1}{n}\sigma_{\epsilon_y}^2\operatorname{tr}(M_X) = \sigma_{\epsilon_y}^2(1-\frac{r}{n})\to\sigma_{\epsilon_y}^2(1-\frac{\tau}{2})$$

where the convergence follows from Assumption 3.

Let $\mu_4 = E(\epsilon_{y,i}^4|X)$. Then following from Lemma A.1, the conditional variance is

$$\begin{aligned}\operatorname{Var}(\frac{1}{n}\boldsymbol{\epsilon}_y'M_X\boldsymbol{\epsilon}_y|X) &= 2\sigma_{\epsilon_y}^4\frac{1}{n^2}\operatorname{tr}(M_X)+(\mu_4-3\sigma_{\epsilon_y}^4)\frac{1}{n^2}\sum_{j=1}^{n}M_{X_{jj}}^2\\ &\le 2\sigma_{\epsilon_y}^4\frac{1}{n^2}\operatorname{tr}(M_X)+(\mu_4-3\sigma_{\epsilon_y}^4)\frac{1}{n^2}\operatorname{tr}(M_X)\\ &= 2\sigma_{\epsilon_y}^4\frac{n-r}{n^2}+(\mu_4-3\sigma_{\epsilon_y}^4)\frac{n-r}{n^2}\\ &= O_P(\frac{1}{n})\end{aligned}$$

where the inequality follow from $\sum_{j=1}^{n} M_{X_{jj}}^2 \leq \text{tr}(M_X^2) = \text{tr}(M_X)$.

Therefore,

$$\frac{1}{n}\boldsymbol{\epsilon}_y' M_X \boldsymbol{\epsilon}_y \xrightarrow{p} \sigma_{\epsilon_y}^2 (1 - \frac{\tau}{2})$$

By the similar argument,

$$\frac{1}{n}\boldsymbol{\epsilon}_d' M_X \boldsymbol{\epsilon}_d \xrightarrow{p} \sigma_{\epsilon_d}^2 (1 - \frac{\tau}{2}), \quad \frac{1}{n}\boldsymbol{\epsilon}_y' M_X \boldsymbol{\epsilon}_d \xrightarrow{p} \sigma^2 (1 - \frac{\tau}{2})$$

□

The following lemma develops the consistent estimator for the components in the variance Σ.

Lemma A.12. *(i)* $\hat{\sigma}_{\epsilon_y}^2 = \frac{1}{n-r}(Y - D\hat{\alpha})' M_X (Y - D\hat{\alpha}) \xrightarrow{p} \sigma_{\epsilon_y}^2$, $\hat{\sigma} = \frac{1}{n-r}(Y - D\hat{\alpha})' M_X D \xrightarrow{p} \sigma$

(ii) $\hat{V} = \frac{1}{n}(\hat{D}_1^{\perp\prime} D_1^{\perp} + \hat{D}_2^{\perp\prime} D_2^{\perp}) \xrightarrow{p} V$

(iii) $\hat{\Omega}_k = \frac{1}{n_k}\hat{\sigma}_{\epsilon_y}^2 \hat{\beta}_{k^c}' P_{k^c} Z_k' M_{X_k} Z_k P_{k^c} \hat{\beta}_{k^c} \xrightarrow{p} \Omega = \sigma_{\epsilon_y}^2 V + \Gamma$ *for* $k = 1, 2$

(iv) $\hat{\Pi} = \frac{1}{n_k}\hat{\sigma}^2 tr(P_1 P_2) \xrightarrow{p} \Pi$

Proof. (i) First note that after partialling out X, we can write

$$M_X Y = M_X D\alpha + M_X \boldsymbol{\epsilon}_y$$

Its mean squares of residual is

$$\begin{aligned}
\check{\sigma}_{\epsilon_y}^2 &= \frac{1}{n}\hat{\boldsymbol{\epsilon}}_y' M_X \hat{\boldsymbol{\epsilon}}_y \\
&= \frac{1}{n}(Y - D\hat{\alpha})' M_X (Y - D\hat{\alpha}) \\
&= \frac{1}{n}(\boldsymbol{\epsilon}_y - D(\hat{\alpha} - \alpha))' M_X (\boldsymbol{\epsilon}_y - D(\hat{\alpha} - \alpha)) \\
&= \frac{1}{n}\boldsymbol{\epsilon}_y' M_X \boldsymbol{\epsilon}_y - 2\frac{1}{n}\boldsymbol{\epsilon}_y' M_X D(\hat{\alpha} - \alpha) + \frac{1}{n}(\hat{\alpha} - \alpha)' D' M_X D(\hat{\alpha} - \alpha)
\end{aligned} \tag{40}$$

For the first term in (40), by Lemma A.11, we have

$$\frac{1}{n}\boldsymbol{\epsilon}_y' M_X \boldsymbol{\epsilon}_y \xrightarrow{p} \sigma_{\epsilon_y}^2 (1 - \frac{\tau}{2}) \tag{41}$$

For the third term in (40), first note that by substituting $D = F\rho + \boldsymbol{\epsilon}_d$ into $\frac{1}{n}D'M_XD$

$$\begin{aligned}\|\frac{1}{n}D'M_XD\| &\leq 2\|\frac{1}{n}\rho'F'M_XF\rho\| + 2\|\frac{1}{n}\boldsymbol{\epsilon}_d'M_X\boldsymbol{\epsilon}_d\| \\ &\leq 2\|\rho\|^2\|\frac{1}{n}F'M_XF\| + 2\|\frac{1}{n}\boldsymbol{\epsilon}_d'M_X\boldsymbol{\epsilon}_d\| \\ &= O_P(1)\end{aligned}$$

where the last equality follows from Lemma A.11 and $\|\frac{1}{n}F'M_XF\| = O_P(1)$

Then, by Theorem 3 that $\hat{\alpha} - \alpha = O_P(\frac{1}{\sqrt{n}})$, we have

$$\|\frac{1}{n}(\hat{\alpha} - \alpha)'D'M_XD(\hat{\alpha} - \alpha)\| \leq \|\hat{\alpha} - \alpha\|^2\|\frac{1}{n}D'M_XD\| = O_P(\frac{1}{n}) \tag{42}$$

For the second term in (40), substituting $D = F\rho + \boldsymbol{\epsilon}_d$ into it,

$$\frac{1}{n}\boldsymbol{\epsilon}_y'M_XD(\hat{\alpha} - \alpha) = \frac{1}{n}\boldsymbol{\epsilon}_y'M_XF\rho(\hat{\alpha} - \alpha) + \frac{1}{n}\boldsymbol{\epsilon}_y'M_X\boldsymbol{\epsilon}_d(\hat{\alpha} - \alpha) \tag{43}$$

For the first term in (43),

$$\|\frac{1}{n}\boldsymbol{\epsilon}_y'M_XF\rho(\hat{\alpha} - \alpha)\| \leq \|\frac{1}{n}\boldsymbol{\epsilon}_y'M_XF\|\|\rho\|\|\hat{\alpha} - \alpha\| = O_P(\frac{1}{n})$$

where the equality follows from Lemma A.3 (iii).

For the second term in (43),

$$\|\frac{1}{n}\boldsymbol{\epsilon}_y'M_X\boldsymbol{\epsilon}_d(\hat{\alpha} - \alpha)\| \leq \|\frac{1}{n}\boldsymbol{\epsilon}_y'M_X\boldsymbol{\epsilon}_d\|\|\hat{\alpha} - \alpha\| = O_P(\frac{1}{\sqrt{n}})$$

Thus,

$$\frac{1}{n}\boldsymbol{\epsilon}_y'M_XD(\hat{\alpha} - \alpha) = O_P(\frac{1}{\sqrt{n}}) \tag{44}$$

Then, by combining (40), (41), (42), and (44), we have

$$\check{\sigma}^2_{\epsilon_y} = \frac{1}{n}\boldsymbol{\epsilon}_y'M_X\boldsymbol{\epsilon}_y + o_P(1) \xrightarrow{p} (1 - \frac{\tau}{2})\sigma^2_{\epsilon_y}$$

Therefore, since $\frac{r}{n} \to \frac{\tau}{2}$ from Assumption 3, it follows from Slutsky's theorem that

$$\hat{\sigma}^2_{\epsilon_y} = \frac{\check{\sigma}^2_{\epsilon_y}}{1 - r/n} \xrightarrow{p} \sigma^2_{\epsilon_y}$$

By the similar argument,

$$\hat{\sigma} = \frac{1}{n-r}(Y - D\hat{\alpha})M_X D \xrightarrow{p} \sigma$$

(ii) It follows from Lemma A.9 that $\frac{1}{n_k}\hat{D}_k^{\perp\prime} D_k^{\perp} \xrightarrow{p} (1-\tau)\beta'\Sigma_{Z^\perp}\beta$, we have

$$\hat{V} = \frac{1}{n}(\hat{D}_1^{\perp\prime} D_1^{\perp} + \hat{D}_2^{\perp\prime} D_2^{\perp}) = \frac{1}{2}\frac{1}{n_1}\hat{D}_1^{\perp\prime} D_1^{\perp} + \frac{1}{2}\frac{1}{n_2}\hat{D}_2^{\perp\prime} D_2^{\perp} \xrightarrow{p} V = (1-\tau)\beta'\Sigma_{Z^\perp}\beta$$

(iii) Without loss of generality, we consider the case of $k = 1$. We want to show that $\hat{\Omega}_1 \xrightarrow{p} \Omega = \sigma_{\epsilon_y}^2 V + \Gamma$.

It follows from (i) that $\hat{\sigma}_{\epsilon_y}^2 - \sigma_{\epsilon_y}^2 = o_P(1)$, we have

$$\begin{aligned}
\hat{\Omega}_1 &= \hat{\sigma}_{\epsilon_y}^2 \hat{\beta}_2 P_2 \frac{Z_1' M_{X_1} Z_1}{n_1} P_2 \hat{\beta}_2 \\
&= \sigma_{\epsilon_y}^2 \hat{\beta}_2 P_2 \frac{Z_1' M_{X_1} Z_1}{n_1} P_2 \hat{\beta}_2 + (\hat{\sigma}_{\epsilon_y}^2 - \sigma_{\epsilon_y}^2)\hat{\beta}_2 P_2 \frac{Z_1' M_{X_1} Z_1}{n_1} P_2 \hat{\beta}_2 \\
&= \sigma_{\epsilon_y}^2 \hat{\beta}_2 P_2 \frac{Z_1' M_{X_1} Z_1}{n_1} P_2 \hat{\beta}_2 + o_P(1)
\end{aligned}$$

Substituting $\hat{\beta}_2 = (Z_2' M_{X_2} Z_2)^+ Z_2' M_{X_2} D_2$ and $M_{X_2} D_2 = M_{X_2} Z_2 \beta + M_{X_2} e_2$ into it,

$$\begin{aligned}
\hat{\Omega}_1 &= \sigma_{\epsilon_y}^2 \beta' P_2 \frac{Z_1' M_{X_1} Z_1}{n_1} P_2 \beta + 2\sigma_{\epsilon_y}^2 \beta' P_2 \frac{Z_1' M_{X_1} Z_1}{n_1} (Z_2' M_{X_2} Z_2)^+ Z_2' M_{X_2} \boldsymbol{e}_2 \\
&\quad + \sigma_{\epsilon_y}^2 \boldsymbol{e}_2' M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ \frac{Z_1' M_{X_1} Z_1}{n_1} (Z_2' M_{X_2} Z_2)^+ Z_2' M_{X_2} \boldsymbol{e}_2 + o_P(1) \qquad (45) \\
&= V_n + 2W_n + \Gamma_n + o_P(1)
\end{aligned}$$

For V_n, by Lemma A.6 (ii) and the similar argument in the proof of (34) in Lemma A.9, we have

$$V_n \xrightarrow{p} \sigma_{\epsilon_y}^2 V = \sigma_{\epsilon_y}^2 (1-\tau)\beta'\Sigma_{Z^\perp}\beta$$

For W_n, by the similar argument in the proof of W_2 in Lemma A.9, we have

$$W_n \xrightarrow{p} 0$$

Let $M_{X_2} Z_2 (Z_2' M_{X_2} Z_2)^+ \frac{Z_1' M_{X_1} Z_1}{n_1} (Z_2' M_{X_2} Z_2)^+ Z_2' M_{X_2} = K_n$. Then $\Gamma_n = \sigma_{\epsilon_y}^2 \boldsymbol{e}_2' K_n \boldsymbol{e}_2$. By substituting $M_{X_2} \boldsymbol{e}_2 = M_{X_2} \boldsymbol{\epsilon}_{d,2} - M_{X_2} U_2 b + M_{X_2} Z_2 \Lambda (\Lambda'\Lambda)^{-1}(\rho - \Lambda'\beta)$ into Γ_n, we

have

$$\begin{aligned}\Gamma_n &= \sigma^2_{\epsilon_y}\boldsymbol{\epsilon}'_{d,2}K_n\boldsymbol{\epsilon}_{d,2} + \sigma^2_{\epsilon_y}b'U_2'K_nU_2b + \sigma^2_{\epsilon_y}(\rho-\Lambda'\beta)'(\Lambda'\Lambda)^{-1}\Lambda P_2\frac{Z_1'M_{X_1}Z_1}{n_1}P_2\Lambda'(\Lambda'\Lambda)^{-1}(\rho-\Lambda\beta)\\ &\quad - 2\sigma^2_{\epsilon_y}\boldsymbol{\epsilon}'_{d,2}K_nU_2b + 2\sigma^2_{\epsilon_y}\boldsymbol{\epsilon}'_{d,2}Z_2'(Z_2'Z_2)^{+}\frac{Z_1'M_{X_1}Z_1}{n_1}P_2\Lambda'(\Lambda'\Lambda)^{-1}(\rho-\Lambda\beta)\\ &\quad - 2\sigma^2_{\epsilon_y}b'U_2'Z_2'(Z_2'Z_2)^{+}\frac{Z_1'M_{X_1}Z_1}{n_1}P_2\Lambda'(\Lambda'\Lambda)^{-1}(\rho-\Lambda\beta)\\ &= \sum_{i=1}^{6}\Gamma_{n,i}\end{aligned}$$

For $\Gamma_{n,1}$, by the similar proof of R_3 in Lemma A.10, we have

$$\Gamma_{n,1} \xrightarrow{p} \Gamma = \begin{cases}\sigma^2_{\epsilon_y}\sigma^2_{\epsilon_d}(1-\tau)\frac{\gamma}{1-\gamma}, & 0<\gamma<1\\ \sigma^2_{\epsilon_y}\sigma^2_{\epsilon_d}(1-\tau)\frac{1}{\gamma-1}, & \gamma>1\end{cases}$$

For $\Gamma_{n,2}$ and $\Gamma_{n,3}$, by (31) and (32) in the proof of Lemma A.8 respectively, we have

$$\begin{aligned}\Gamma_{n,2} &= O_P(\frac{1}{\psi_{n,p}})\\ \Gamma_{n,3} &= O_P(\frac{1}{\psi^2_{n,p}})\end{aligned}$$

For $\Gamma_{n,4}$, $\Gamma_{n,5}$ and $\Gamma_{n,6}$, by Cauchy Schwarz inequality,

$$\begin{aligned}|\Gamma_{n,4}| &\le 2\|\Gamma_{n,1}\|^{1/2}\|\Gamma_{n,2}\|^{1/2} = O_P(\frac{1}{\sqrt{\psi_{n,p}}})\\ |\Gamma_{n,5}| &\le 2\|\Gamma_{n,1}\|^{1/2}\|\Gamma_{n,3}\|^{1/2} = O_P(\frac{1}{\psi_{n,p}})\\ |\Gamma_{n,6}| &\le 2\|\Gamma_{n,2}\|^{1/2}\|\Gamma_{n,3}\|^{1/2} = O_P(\frac{1}{\psi_{n,p}\sqrt{\psi_{n,p}}})\end{aligned}$$

Thus,

$$\Gamma_n = \Gamma_{n,1} + o_P(1) \xrightarrow{p} \Gamma = \begin{cases}\sigma^2_{\epsilon_y}\sigma^2_{\epsilon_d}(1-\tau)\frac{\gamma}{1-\gamma}, & 0<\gamma<1\\ \sigma^2_{\epsilon_y}\sigma^2_{\epsilon_d}(1-\tau)\frac{1}{\gamma-1}, & \gamma>1\end{cases}$$

Therefore,

$$\hat{\Omega}_1 \xrightarrow{p} \Omega = \sigma^2_{\epsilon_y}V + \Gamma$$

(iv) It follows from (i) and the continuous mapping theorem that $\hat{\sigma}^2 - \sigma^2 = o_P(1)$. Together

with Lemma A.7(ii), we have

$$\begin{aligned}\hat{\Pi} &= \frac{2}{n}\hat{\sigma}^2 tr(P_1 P_2) \\ &= \frac{2}{n}\sigma^2 tr(P_1 P_2) + \frac{2}{n}(\hat{\sigma}^2 - \sigma^2) tr(P_1 P_2) \\ &= \frac{2}{n}\sigma^2 tr(P_1 P_2) + o_P(1) \xrightarrow{p} \Pi = \begin{cases} \sigma^2(1-\tau)\gamma, & 0 < \gamma < 1 \\ \sigma^2(1-\tau)\frac{1}{\gamma}, & \gamma > 1 \end{cases}\end{aligned}$$

□

A.5.1 Proof of Theorem 4

Proof of Theorem 4. Following Lemma A.12, we know $\hat{V} \xrightarrow{p} V$, $\hat{\Pi} \xrightarrow{p} \Pi$ and $\hat{\Omega}_k \xrightarrow{p} \Omega$ for $k = 1, 2$. Therefore, from Slutsky's theorem,

$$\hat{\Sigma} = \hat{V}^{-1}(\frac{1}{2}\hat{\Omega}_1 + \frac{1}{2}\hat{\Omega}_2 + \hat{\Pi})\hat{V}^{-1} \xrightarrow{p} \Sigma$$

which completes the proof of consistency. □

A.6 Proof of Theorem 5

In this section, we establish the asymptotic properties of the oracle factor IV estimator. Part (i) establishes its asymptotic normality, while Part (ii) compares its asymptotic variance with that of the proposed estimator in Theorem 3 under the sufficiently overparameterized regime.

Proof of Theorem 5. (i) First note that $\hat{\alpha}_{FIV} = (\hat{D}_{FIV}^{\perp\prime} D^{\perp})^{-1} \hat{D}_{FIV}^{\perp\prime} Y^{\perp}$ and $\hat{D}_{FIV}^{\perp} = M_X F(F' M_X F)^{-1} F' M_X D$. Then

$$\sqrt{n}(\hat{\alpha}_{FIV} - \alpha) = (\frac{1}{n} D' M_X F(F' M_X F)^{-1} F' M_X D)^{-1} \frac{1}{\sqrt{n}} D' M_X F(F' M_X F)^{-1} F' M_X \boldsymbol{\epsilon}_y = S^{-1} R$$

Substituting $D = F\rho + \boldsymbol{\epsilon}_d$ into S,

$$\begin{aligned}\frac{1}{n} D' M_X F(F' M_X F)^{-1} F' M_X D = &\frac{1}{n}\rho' F' M_X F \rho + \frac{1}{n}\boldsymbol{\epsilon}_d' M_X F(F' M_X F)^{-1} F' M_X \boldsymbol{\epsilon}_d \\ &+ \frac{2}{n}\rho' F' M_X \boldsymbol{\epsilon}_d \end{aligned} \tag{46}$$

For the first term in (46), since $\frac{F' M_X F}{n} \xrightarrow{p} (1 - \frac{\tau}{2})\Sigma_{f|x}$ from Assumption 1 (iii), we

have

$$\frac{1}{n}\rho' F' M_X F \rho \xrightarrow{p} (1-\frac{\tau}{2})\rho' \Sigma_{f|x} \rho$$

For the second term and third term in (46), by Lemma A.3 (iii),

$$\|\frac{1}{n}\boldsymbol{\epsilon}_d' M_X F (F' M_X F)^{-1} F' M_X \boldsymbol{\epsilon}_d\| \leq \|\frac{1}{n} F' M_X \boldsymbol{\epsilon}_d\|^2 \|(\frac{F' M_X F}{n})^{-1}\| = O_P(\frac{1}{n})$$
$$\|\frac{1}{n}\rho' F' M_X \boldsymbol{\epsilon}_d\| \leq \|\rho\| \|\frac{1}{n} F' M_X \boldsymbol{\epsilon}_d\| = O_P(\frac{1}{\sqrt{n}})$$

Thus, $S = \frac{1}{n} D' M_X F (F' M_X F)^{-1} F' M_X D \xrightarrow{p} (1-\frac{\tau}{2})\rho' \Sigma_{f|x} \rho$

Substituting $D = F\rho + \boldsymbol{\epsilon}_d$ into R,

$$\frac{1}{\sqrt{n}} D' M_X F (F' M_X F)^{-1} F' M_X \boldsymbol{\epsilon}_y = \frac{1}{\sqrt{n}} \rho' F' M_X \boldsymbol{\epsilon}_y + \frac{1}{\sqrt{n}} \boldsymbol{\epsilon}_d' M_X F (F' M_X F)^{-1} F' M_X \boldsymbol{\epsilon}_y \tag{47}$$

For the second term in (47), by Lemma A.3 (iii),

$$\|\frac{1}{\sqrt{n}} \boldsymbol{\epsilon}_d' M_X F (F' M_X F)^{-1} F' M_X \boldsymbol{\epsilon}_y\| \leq \|\frac{1}{\sqrt{n}} F' M_X \boldsymbol{\epsilon}_d\| \|(\frac{F' M_X F}{n})^{-1}\| \|\frac{1}{n} F' M_X \boldsymbol{\epsilon}_y\| = O_P(\frac{1}{\sqrt{n}})$$

For the first term in (47), let $\mathcal{F}_i$ be the σ-field generated by $\{(F, X, \epsilon_{y,j}), j \leq i\}$, then we have,

$$\frac{1}{\sqrt{n}} \rho' F' M_X \boldsymbol{\epsilon}_y = \frac{1}{\sqrt{n}} \rho' F' \sum_{i=1}^{n} m_i \epsilon_{y,i} = \sum_{i=1}^{n} \varsigma_i$$

where m_i is the i-th column of M_X. We have $E(\varsigma_i | \mathcal{F}_{i-1}) = 0$, then ς_i is a martingale difference array with respect to the filtration $\mathcal{F}$. Also,

$$\begin{aligned}\sum_{i=1}^{n} E(\varsigma_i^2 | \mathcal{F}_{i-1}) &= \sigma_{\epsilon_y}^2 \frac{1}{n} \sum_{i=1}^{n} \rho' F' m_i m_i' F \rho \\ &\leq \sigma_{\epsilon_y}^2 \|\rho\|^2 \|\frac{1}{n} F' M_X F\| \\ &= O_P(1)\end{aligned}$$

By Lyapunov condition, we want to show for some $\delta > 0$,

$$\sum_{i=1}^{n} E(|\varsigma_i|^{2+\delta} | \mathcal{F}_{i-1}) \xrightarrow{p} 0$$

Taking $\delta = 2$, we have

$$\begin{aligned}\sum_{i=1}^{n} E(|\varsigma_i|^4|\mathcal{F}_{i-1}) &= \frac{1}{n^2}\sum_{i=1}^{n} E((\rho' F' m_i \epsilon_{y,i})^4|\mathcal{F}_{i-1}) \\ &\leq \frac{1}{n^2}\sum_{i=1}^{n} (\rho' F' m_i)^4 E(\epsilon_{y,i}^4|F, X) \\ &\leq C\|\rho\|^4 \|\frac{1}{n} F' M_X F\| \max_i \|\frac{F' m_i}{\sqrt{n}}\|^2 \\ &= O_P(1) o_P(1) = o_P(1)\end{aligned}$$

where the second inequality follows from Assumption 2 (ii), and the last equality follows from Assumption 4 (i).

From which we obtain the Lyapunov condition. Hence, by the Martingale Central Limit Theorem,

$$\frac{1}{\sqrt{n}} \rho' F' M_X \boldsymbol{\epsilon}_y \xrightarrow{d} N(0, \Psi)$$

where $\Psi = \sigma_{\epsilon_y}^2 (1 - \frac{\tau}{2}) \rho' \Sigma_{f|x} \rho$. Therefore,

$$R = \frac{1}{\sqrt{n}} \rho' F' M_X \boldsymbol{\epsilon}_y + o_P(1) \xrightarrow{d} N(0, \sigma_{\epsilon_y}^2 (1 - \frac{\tau}{2}) \rho' \Sigma_{f|x} \rho)$$

Combined with $S \xrightarrow{p} (1 - \frac{\tau}{2}) \rho' \Sigma_{f|x} \rho$, we have

$$\sqrt{n}(\hat{\alpha}_{FIV} - \alpha) = S^{-1} R = ((1 - \frac{\tau}{2}) \rho' \Sigma_{f|x} \rho)^{-1} \frac{1}{\sqrt{n}} \rho' F' M_X \boldsymbol{\epsilon}_y + o_P(1) \xrightarrow{d} N(0, \sigma_{\epsilon_y}^2 ((1 - \frac{\tau}{2}) \rho' \Sigma_{f|x} \rho)^{-1})$$

(ii) From (i) and Theorem 3, as $n_1, n_2, r, p \to \infty$, $\frac{r}{n_k} \to 0$, $\frac{p}{n_k - r} \to \infty$ for $k = 1, 2$,

$$\Sigma = \sigma_{\epsilon_y}^2 (\beta' \Sigma_{Z^\perp} \beta)^{-1}, \quad \Sigma_{FIV} = \sigma_{\epsilon_y}^2 (\rho' \Sigma_{f|x} \rho)^{-1}$$

Note that $\Sigma_{Z^\perp} = \Lambda \Sigma_{f|x} \Lambda' + \sigma_u^2 I_p$. By the Woodbury matrix identity,

$$\Sigma_{Z^\perp}^{-1} = \sigma_u^{-2} I_p - \sigma_u^{-4} \Lambda (\Sigma_{f|x}^{-1} + \sigma_u^{-2} \Lambda' \Lambda)^{-1} \Lambda'$$

Then from Theorem 1, we have $\beta = \sigma_u^{-2} \Lambda (\Sigma_{f|x}^{-1} + \sigma_u^{-2} \Lambda' \Lambda)^{-1} \rho = \Sigma_{Z^\perp}^{-1} \Lambda \Sigma_{f|x} \rho$. Thus,

$$\beta' \Sigma_{Z^\perp} \beta = \rho' \Sigma_{f|x} \Lambda \Sigma_{Z^\perp}^{-1} \Lambda \Sigma_{f|x} \rho$$

Let $A = \sigma_u^{-2}\Lambda'\Lambda$, we have

$$\Lambda'\Sigma_{Z^\perp}^{-1}\Lambda = A - A(\Sigma_{f|x}^{-1} + A)^{-1}A = A(A^{-1} - (\Sigma_{f|x}^{-1} + A)^{-1})A$$

Using $A^{-1} - (A+B)^{-1} = A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1}$ and with $B = \Sigma_{f|x}^{-1}$,

$$\Lambda'\Sigma_{Z^\perp}^{-1}\Lambda = (\sigma_u^2(\Lambda'\Lambda)^{-1} + \Sigma_{f|x})^{-1} = \Sigma_{f|x}^{-1} + O(\frac{1}{\psi_{n,p}})$$

Thus,

$$\beta'\Sigma_{Z^\perp}\beta = \rho'\Sigma_{f|x}\rho + o(1)$$

Therefore,

$$\Sigma = \sigma_{\epsilon_y}^2 (\beta'\Sigma_{Z^\perp}\beta)^{-1} = \sigma_{\epsilon_y}^2(\rho'\Sigma_{f|x}\rho)^{-1} + o(1) = \Sigma_{FIV} + o(1).$$

□