Many Weak Instruments in Time Series Econometrics

By Anna Mikusheva[1]

Abstract

This paper studies linear instrumental variables (IV) estimation in time series settings where many instruments are available. Motivation comes from GMM estimation of rational expectation models, including the New Keynesian Phillips curve, Euler equations, and Taylor rules. The paper surveys and summarizes ideas from the cross-sectional literature on many weak instruments, establishes new results for a split-sample approach, and discusses extensions and adaptations of the cross-sectional results to time series settings. The main challenge of estimation with many weak instruments comes from endogeneity of the estimated instrument, which can be solved using sample splitting, cross-fitting, jackknifing and deleted diagonal approaches. This paper shows that the split-sample approach is agnostic to the method used to estimate the optimal instrument, allowing for a variety of machine learning estimators to be employed, and produces easy-to-implement, asymptotically reliable statistical inferences under both weak and strong identification.

Keywords: Weak Identification, Many Instruments, Time Series

JEL Codes: C14, C22, C26, C55

This draft: February 2021.

1 Introduction

Many structural macroeconometric relations, including the New Keynesian Phillips Curve, Euler equations, and Taylor rules, are known to be weakly identified when estimated by GMM using aggregated macro-data, see i.e. Mavroeidis (2004). One important and probably underused feature of these models is that they are formulated as conditional moment restrictions, leading to potentially many unconditional moment equations which may be used for estimation. Specifically, all lags of any available macro variable can serve

[1] Department of Economics, M.I.T., 50 Memorial Drive, E52-526, Cambridge, MA, 02142. Email: amikushe@mit.edu. National Science Foundation support under grant number 1757199 is gratefully acknowledged. I benefitted from discussions with Stanislav Anatolyev, Isaiah Andrews, Josh Angrist, Whitney Newey, Liyang Sun, David Hughes and Sylvia Klosin.

as a valid instrument. The potential of exploiting this wealth of available information to produce more accurate inferences about structural parameters makes usage of many weak instruments very promising. There have been many recent advances in understanding the statistical issues and developing reliable methods to exploit many weak instruments in cross-sectional settings (i.g. Hausman et al (2012), Belloni at all (2012) among many) – an adaption of these methods to time series is lagging behind.

The main goals of this paper are: to survey and systematize recent advances in cross-sectional studies of many weak instruments; to establish some missing results; and to investigate how these tools can be adapted to empirical macroeconometric applications. The paper advocates for the use of split-sample IV estimation as the easiest and most versatile approach for extracting information from an abundant set of instruments, and delivering clean statistical inferences on the structural coefficient. This approach, as we argue, is very adaptable to the additional challenges posed by time series data. We establish new results about the consistency and asymptotic distributions of the split-sample estimator, and discuss weak identification robust inferences. The paper also surveys machine learning (ML) approaches popular in time series settings that can be freely combined with the sample splitting idea to select/ estimate the optimal instrument.

We frame the central issue of using many instruments as the problem of endogeneity of the estimated instrument. The optimal instrument in a model with homoskedastic martingale-difference errors coincides with the best predictor of the endogenous regressor given the available set of instruments. A variety of ML techniques can be used to select the best predictive model and to construct the optimal instrument. The challenge, however, is that fitting the endogenous regressor with a very flexible model also fits the endogenous part of the regressor in a flexible way. Similarly, selecting an instrument out of a large set of available instruments based on its predictive power for the endogenous regressor favors the instruments showing larger in-sample correlation with the endogenous first-stage error term. As we argue, flexible estimation/selection of instruments leads to the constructed optimal instrument being endogenous, despite each original instrument being exogenous. When the instruments contain a strong signal about the regressor, the problem of endogenous selection is reflected in large finite-sample biases, while in the case where the information is limited, we may end up with an inconsistent estimator and misleading inferences.

The idea of the sample splitting approach (Angrist and Krueger (1995)) is straightforward. The researcher splits the sample into two parts: the first part of the sample is used to estimate the best predictor for the endogenous regressor based on available instruments; this estimated instrument is then applied to the second subsample to conduct regular, just-identified linear IV inferences using weak identification robust approaches. The sample splitting makes the estimated instrument exogenous both in cross-sectional and, with some small adaptations, in time series settings. The approach can be combined with any ML technique used on the first subsample. Other options to construct an exogenous instrument in cross-sectional applications include cross-fitting, jackknife (Angrist et al. (1999)) and deleted diagonal (Hansen et al. (2008)) ideas. This paper surveys the existing asymptotic results in cross-section for these approaches. Unfortunately, as we show, their applicability to time series data is currently more limited and we discuss the potential pitfalls.

The time series nature of macroeconometric applications raises additional challenges. We pay special attention to the problem of autocorrelated errors, as well as the distinction between weak exogeneity assumptions, which are common in time series, and the strict exogeneity typical for cross-sectional data. We point out ways of obtaining reliable asymptotic statements without assuming consistency of the estimation/selection of optimal instruments, and acknowledging the implied non-ergodicity of the estimated averages. This is achieved by making statistical inferences conditional on the first subsample and using stable martingale central limit theorems.

The paper is structured in the following way. Section 2 motivates the importance of many weak instruments in time series data by highlighting a series of empirical examples. Section 3 surveys the cross-sectional literature, explains the issue of endogeneity in the estimated instrument, and lists some available solutions. Section 4 surveys cross-sectional asymptotic results established in the literature for the jackknife and deleted diagonal approaches, and obtains new results for the split-sample estimator. Section 5 is devoted to the additional challenges of time series data and discusses the ways cross-sectional methods may be adapted and their potential pitfalls. We also pay attention to Factor IV methods.

2 Empirical examples

In structural macroeconometrics it is common to have a large number of potential instruments. For example, when estimating rational expectations models the exclusion restriction is often formulated as a conditional expectation, where the conditioning is on all information available at the time the expectation is taken. This makes all lags of any macro variables valid instruments. Despite the seeming abundance of potential instruments, structural estimation using aggregate data often suffers from weak identification, at least when a relatively small number of carefully chosen instruments is used.

Example 1. New Keynesian Phillips Curve. The NKPC is a rational expectation model capturing a trade-off between the rate of inflation and the level of economic activity. A theoretical justification of the NKPC comes from the Calvo model. There exists a diverse range of empirical specifications, but the most common is the following:

$$\pi_t = \lambda x_t + \gamma_f \mathbb{E}_t \pi_{t+1} + \gamma_b \pi_{t-1} + u_t. \tag{1}$$

Here π_t is inflation in period t, x_t is a proxy for marginal costs (often the labor share or output gap), u_t is unpredictable structural error, and $\mathbb{E}_t$ is a rational expectation formed at time t. Gali and Gertler (1999) proposed GMM-IV estimation of the NKPC by forming the moment condition

$$\mathbb{E}[(\pi_t - \lambda x_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1}) Z_{t-1}] = 0,$$

where one can use any variable observed at time $t-1$ in the instrument set Z_{t-1}. Kleibergen and Mavroeidis (2009) show that 'weak instrument problems arise if marginal costs have limited dynamics or if their coefficient is close to zero, that is, when the NKPC is flat, since in those cases the exogenous variation in inflation forecasts is limited.' A survey paper by Mavroeidis et al (2014) reports that 'estimation of the NKPC using macro data is subject to a severe weak instruments problem. Consequently, seemingly innocuous specification changes lead to big differences in point estimates.' □

Example 2. Euler equation. Euler equations for consumption or output are an important part of many macroeconomic models. There are multiple specifications used in empirical work – below is a formulation suggested by Fuhrer and Rudenbusch (2004):

$$c_t = \alpha + \phi \mathbb{E}_t c_{t+1} - \varphi r_t + \sum_{j=1}^{J} \alpha_j c_{t-j} + u_t,$$

where c_t and r_t are the logs of consumption (output) and the real interest rate. GMM estimation of the Euler equation was proposed by Hansen and Singleton (1982), who suggested using lags of available variables as instruments. Yogo (2004) raised the issue of weak identification of the coefficient of intertemporal substitution. Acsari et al (2020) provides a comprehensive survey of different specifications and estimation approaches to the Euler equation. □

Example 3. Taylor rule. A Taylor rule is a policy reaction function that describes how a Central bank conducts monetary policy. One common specification is

$$r_t = \bar{r} + \beta(\mathbb{E}_t \pi_{t+1} - \bar{\pi}) + \gamma \mathbb{E}_t x_{t+1} + \epsilon_t,$$

where r_t is the Federal Funds rate, π_t the inflation rate, x_t the output gap, and $\bar{r}$ and $\bar{\pi}$ are equilibrium rates. Clarida et al (1998) suggested a GMM approach to estimating the Taylor rule that allows the researcher to use any lagged variables as instruments. Mavroeidis (2004) draws attention to weak identification of the Taylor rule. □

Example 4. Factor pricing. Factor pricing models assume that the expected excess return on a stock or a portfolio of assets is equal to the price of risk (or risk premia) λ, for some risk factor F_t, multiplied by the portfolio's quantity of risk β_i:

$$\mathbb{E} r_{it} = \lambda \beta_i, \quad \beta_i = (Var(F_t))^{-1} cov(F_t, r_{it}).$$

One commonly used estimation procedure is the Fama-MacBeth approach (Fama and MacBeth (1973), Shanken (1992)) that first estimates β_i by running time series regressions of excess returns r_{it} on the realization of risk factor F_t for each asset separately. Then, one runs a cross-sectional regression of the average return for each asset on its estimated β_i to obtain an estimate of the risk premia λ. This procedure can be interpreted as a classical TSLS estimator with a large number of instruments. The number of instruments here equals to the number of assets used for estimation multiplied by the number of factors, and the β_i play the role of the first-stage coefficients. When factors F_t are only weakly correlated with excess returns, we are faced with a problem of weak instruments (see, Anatolyev and Mikusheva (2020)). □

Simplified setting. In this paper we address the issue of statistical inference on a structural parameter β, the coefficient on the single endogenous regressor X_t, in the presence of many potential instruments Z_t. We pay special attention to issues of weak identification and the time series nature of the data. We discuss only models linear in β and correspondingly the linear IV formulation. Hansen and Singleton (1982) introduced the GMM approach for the estimation of non-linear Euler equations. However, issues of many weak instruments and weak identification are significantly more complicated in the non-linear setting, and our understanding of them is very limited; they are left out of the current paper.

Our setting abstracts away from several complications that may arise in practice. Firstly, we assume that the structural equation has no included controls. This assumption is not very restrictive if the number of potential controls is small as we can assume that the equation of interest is obtained after partialling them out. However, the need to partial out many (or an increasing number of) controls poses a separate and very hard problem. An excellent survey is proveded by Anatolyev (2019). Secondly, we assume that we deal with a single endogenous regressor. The results can be easily generalized to multiple endogenous regressors as long as we are interested in joint inferences on all structural coefficients. Inference under weak identification on each structural coefficient separately is a complicated econometric issue (see Kleibergen and Mavroeidis (2009)). Finally, we assume away any complications that may arise from the persistence (unit root behavior) of some regressors or instruments. Specifically, we assume that all variables are stationary enough for some forms of the law of large numbers and central limit theorem to hold.

3 Cross-Section: statement of the problem

3.1 Many Instruments: constructing optimal instrument

In this section we concentrate our attention on cross-sectional data and assume that we observe an i.i.d. sample (Y_i, X_i, Z_i) for $i = 1, .., n$. We consider a linear IV model with one-dimensional endogenous regressor X and K-dimensional instrument Z

$$Y_i = \beta X_i + e_i, \quad \mathbb{E}[e_i|Z_i] = 0.$$

Chamberlain (1987) derived the optimal instrument f_i that minimizes the variance of the IV estimate: $f_i = \frac{\mathbb{E}[X_i|Z_i]}{\mathbb{E}[e_i^2|Z_i]}$. Many papers in this literature aim to find an estimation and inference procedure that achieves semi-parametric efficiency under homoskedasticity, while at the same time delivering valid results under heteroscedasticity (heteroscedasticity-robust). In accordance with this goal, we look for an optimal instrument of the form

$$f_i = \mathbb{E}[X_i|Z_i]. \tag{2}$$

In practice the optimal instrument is not known and has to be estimated – Newey (1990) suggested estimating f_i non-parametrically.

In this paper we consider only two-step estimators, which covers vast majority of available estimators. In the first step one constructs a model of the best predictor for X_i based on the potential predictors/features Z_i using some regularized non-parametric estimation and selection methods. Denote this estimated optimal instrument as $\widehat{f}_i = \widehat{E}[X_i|Z_i]$. In the second step, the estimated optimal instrument is employed in the just identified linear IV:

$$\widehat{\beta} = \frac{\sum_{i=1}^n \widehat{f}_i Y_i}{\sum_{i=1}^n \widehat{f}_i X_i}. \tag{3}$$

Let us mention several prominent approaches for first-step estimation. Donald and Newey (2001) proposed an instrument selection procedure based on a Mallows criteria. Belloni et al (2010) and Belloni et al (2012) suggest using LASSO estimation on the first step to construct the optimal instrument. Okui (2011) proposes a shrinkage estimator, assuming that there is a known set of strong instruments that delivers a consistent estimator of β. Carrasco (2012) suggests several regularization procedures based on the spectral decomposition of the conditional expectation operator. Among her proposals are the principal components approach and Tikhonov's regularization of the conditional expectation operator. There have been also recent suggestions to use other Machine Learning techniques for the optimal instrument construction, such as the random forest (Ash et al (2018)).

All procedures mentioned above deliver semi-parametric efficient estimators under some set of assumptions. Typically this involves some assumption placed on the form of the optimal instrument, which allows for its consistent estimation. For example, the LASSO procedure of Belloni et al (2010) delivers the desired results if the first-stage regression is approximately sparse, that is, a relatively small number of the instruments successfully approximates the optimal instrument. Donald and Newey (2001) assumes a known ordering among instruments (or groups of instruments) by strength/informativeness. Another type of assumption often needed is a regularity condition placed on the conditional expectation operator. For example, Belloni et al (2012) restrict eigenvalues of empirical Gram matrix, while Carrasco (2012) assumes that the conditional expectation operator is a Hilbert-Schmidt operator. All papers mentioned above assume strong identification of the IV model.

3.2 Weak Instruments

In this section we provide a very brief summary of known facts about weak identification in a just identified case. Specifically, we consider the identification strength of the optimal instrument as if it is known, and take the 'optimal' instrument to be the one defined in (2). We acknowledge the limitation of this definition and recognize that the weak IV literature has an unresolved debate about the choice of a powerful test, and direction of power, for over-identified linear IV models. However, we intend to stay away from this debate and solve a somewhat simpler problem, maintaining definition (2) as a goalpost. Let us write the (infeasible) first-stage regression as:

$$X_i = \mathbb{E}[X_i|Z_i] + v_i = f_i + v_i, \tag{4}$$

where v_i is the prediction error with $\mathbb{E}[v_i|Z_i] = 0$. Weak identification arises when the uncertainty coming from the prediction error v_i is empirically important; that is, cases with low signal-to-noise ratio in equation (4). This ratio is captured by the concentration parameter

$$\mu^2 = \frac{n\left(\mathbb{E}[f_i^2]\right)^2}{\sigma^2}, \tag{5}$$

where σ^2 is the asymptotic variance of $\frac{1}{\sqrt{n}}\sum_{i=1}^n f_i v_i$. When the concentration parameter is not very large, the finite-sample distribution of the (infeasible) optimal IV estimate $\widehat{\beta}_o$ and its t-statistic are not well approximated by the gaussian distribution. We can write the error of the optimal IV estimate as

$$\widehat{\beta}_o - \beta = \frac{\sum_{i=1}^n f_i Y_i}{\sum_{i=1}^n f_i X_i} - \beta = \frac{\sum_{i=1}^n f_i e_i}{\sum_{i=1}^n f_i^2 + \sum_{i=1}^n f_i v_i}.$$

Asymptotic gaussianity of this quantity is usually based on the normalized numerator being well approximated by a gaussian distribution, while the normalized denominator is close to a constant. The latter approximation fails when the signal-to-noise ratio is small, as the stochastic mean-zero term $\sum_{i=1}^n f_i v_i$ remains empirically important in comparison to the term $\sum_{i=1}^n f_i^2$, which is taken by standard asymptotic theory to be the only important term.

When the signal-to-noise ratio is low, the denominator is not just noisy, it is endogenously noisy. Indeed, in most cases we seek an instrumental variables estimate because we suspect the regressor X_i to be endogenous, or equivalently, that the prediction error v_i is correlated with the structural error e_i. This implies that the near gaussian terms $\sum_{i=1}^n f_i e_i$ and $\sum_{i=1}^n f_i v_i$ are correlated. This endogeneity makes the infeasible optimal IV estimate biased (towards the OLS limit) and standard t-statistic based inferences misleading (of incorrect size), when the concentration parameter is small.

Stock and Yogo (2005) suggested a pre-test, commonly know as the first-stage F-test, that assesses empirically whether the usual TSLS estimator, and corresponding t-statistic, provides reliable inferences in a given application. Such a pre-test usually guides a researcher in choosing between identification-robust tests/confidence sets (if identification is deemed weak) or standard TSLS t-statistics. In the just-identified heteroscedastic case one can use a heteroscedasticity-robust form of this pre-test. In a recent survey, Andrews et al (2019) explain why there is no generally acceptable pre-test for weak identification in an over-identified linear IV model under conditional heteroscedasticity, although there is one for the just-identified case.

A wide literature is devoted to identification-robust testing and confidence set construction. The most well known and often used tests are the Anderson-Rubin statistic, Kleibergen's (2002) KLM and the conditional likelihood ratio of Moreira (2003). They are justified in settings with a small number of instruments and are equivalent in the just-identified case. The recommendation for the just-identified setting is to always use identification-robust tests rather than employing the weak identification pre-test. This recommendation follows from a statement that the robust tests mentioned above are asymptotically efficient in a just-identified homoskedastic case, if identification is strong.

3.3 Main problem: endogeneity of the estimated instrument

Problematic simulations. A recent thought-provoking paper by Angrist and Frandsen (2020) assesses the utility of machine learning techniques in modern applied labor economics applications. Discussed in great detail is the use of machine learning techniques for instrument selection. The authors create simulation exercises based on two applications: identification of the return to education using quarter of birth as instruments (Angrist and Krueger (1991)); and the effect of a movie's opening-weekend viewership on subsequent sales, with instruments generated by weather indicators (Gilchrist and Sands (2016)). The authors diligently design the simulation settings to match the empirical examples along many directions, including the heterogeneity of the first-stage effects and heteroscedasticity.

The amazing conclusion of Angrist and Frandsen (2020) is that the use of machine learning techniques for construction of the optimal instrument in these two applications does not deliver the results many hope for. The authors explored the performance of IV regression using both LASSO and random forest estimators for the first stage and contrasted it with OLS, TSLS and several jackknife and split-sample estimators we will discuss below. In almost all cases the IV estimators using LASSO and random forest estimates delivered large biases, comparable to that of TSLS and OLS without much improvement in terms of variance. The performance of the both machine learning methods depends significantly on the choice of regularization parameter (the cross-validation or plug-in penalties for the LASSO, or the leaf-size for the random forest) with none of the standard choices being totally satisfactory in these applications. These results rhyme well with the simulation evidence in Hansen and Kozbur (2014), where the authors report less than stellar performance of IV estimators using LASSO first stage when the signal on the first stage is weak.

Essence of the problem. Estimating the optimal instrument in a very flexible way, or selection from among many instruments, may lead to over-fitting the endogenous part of regressor X. This makes the estimated optimal instrument endogenous, $\mathbb{E}[\widehat{f}_i e_i] \neq 0$, even though each individual instrument is exogenous, and leads to the bias of the IV estimator. We explain this phenomenon below in the context of the TSLS estimator, following Bekker (1994).

Assume the optimal instrument $f_i = \pi' Z_i$ is a linear combination of the available K instruments and assume that $Z'Z$ is a matrix of rank K. The TSLS estimator uses the following estimated optimal instrument:

$$\widehat{f}_i = X'Z(Z'Z)^{-1}Z_i = f_i + v'Z(Z'Z)^{-1}Z_i,$$

where the estimation error is correlated with the structural error, since the prediction error v_i is endogenous. Under conditional homoskedasticity we find the following formula for the endogeneity of the estimated instrument:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_i - f_i)e_i\right] = \frac{\mathbb{E}[v_i e_i]}{n}\mathbf{tr}(Z(Z'Z)^{-1}Z') = \frac{K}{n}\sigma_{ev}.$$

As we see, the endogeneity of the estimated optimal instrument is increasing in the number of available instruments as the endogenous part of regressor X is being fitted more flexibly. A larger number of instruments, or a more flexible first stage, may result in a large finite-sample bias of the two-step estimator. Under asymptotics in which the number of instruments K is growing, this may even lead to inconsistency.

Similar observations can be made about other IV estimators that have relatively simple form (e.g. see Hansen and Kozbur (2014) for ridge-first stage). Unfortunately, the exact form of the bias is unavailable for first-stage estimators that involve simultaneous variable selection and estimation, and hence do not have a simple analytical form. Nevertheless, the logic behind the endogeneity of the estimated instrument is somewhat similar. Among many instruments that have similar explanatory power, the ones that are most correlated with the endogenous part of the regressor X, in sample, are more likely to be selected as they deliver better fit. Again the more flexible first-step prediction model leads to a more endogenous estimated instrument $\widehat{f}_i$. Similar arguments are stated in Dovi (2020), who showed in simulations large size distortions of the weak identification robust tests applied after selection of instruments from a large set of available instruments (endogenous instrument selection).

Most machine learning techniques fight the over-fitting bias with a regularization scheme; however, this approach may fail when the signal is low or is not of the form that the proposed technique was created to recover (e.g., not sparse in the case of LASSO). In such cases we may again end up with an endogenous estimated optimal instrument, which results in large bias of the IV estimator. This was clearly demonstrated in simulations conducted in Hansen and Kozbur (2014) which showed that a weak signal (dense or sparse) cannot be well recovered in a first-stage estimation done by ridge or LASSO.

3.4 Solutions to endogeneity problem

There are several econometric ideas proposed that effectively solve the problem of correlation between the estimated optimal instrument and the structural error.

Sample splitting. Angrist and Krueger (1995) proposed a split-sample estimator, which randomly splits the sample in two subsamples, the first of which is used to estimate the form of the optimal instrument, while the second is used to estimate the structural parameter β. In the case of TSLS this means we estimate the first-stage coefficients $\widehat{\pi}_1 = (Z_1'Z_1)^{-1}Z_1'X_1$, where the subscript 1 denotes estimation using only observations in the first subsample. The estimation of the structural coefficient employs only observations from the second subsample and uses $\widehat{f}_i = \widehat{\pi}_1'Z_i$. This approach guarantees the exogeneity of the estimated optimal instrument.

The idea of sample splitting can be applied to any method of instrument selection or optimal instrument estimation. Assume that a researcher randomly splits the data set into two subsamples indexed by I_1 and I_2. Denote $\mathcal{A}_1$ and $\mathcal{A}_2$ the full set (X_i, Y_i, Z_i) of random variables in the two subsamples. The researcher uses some regression or ML approach to estimate the optimal instrument $\widehat{f}_i = \widehat{f}(\mathcal{A}_1, Z_i)$ using data in $\mathcal{A}_1$, evaluates the estimated instrument for each observation $i \in I_2$, and then runs just-identified IV on the second subsample with instrument $\widehat{f}_i$. We denote $\widehat{\beta}_{SS}$, the split-sample estimator of β, defined as:

$$\widehat{\beta}_{SS} = \frac{\sum_{i \in I_2} \widehat{f}(\mathcal{A}_1, Z_i) Y_i}{\sum_{i \in I_2} \widehat{f}(\mathcal{A}_1, Z_i) X_i}. \tag{6}$$

Whenever the method used for the construction of the optimal instrument is specified, we add this to the name of the estimator. For example, SS-LASSO is the estimator that estimates $\widehat{f}_i$ using LASSO regression of X on the instruments in the I_1 sample only, and then runs just identified IV on the subsample I_2. We usually assume that the number of observations in the second subsample I_2 increases to infinity with the sample size, but it is not required for the two subsamples to be of the same size.

Cross-fit. To salvage the efficiency loss from using only part of the sample for structural estimation, the researcher may use the two subsamples symmetrically: fitting the first stage on each subsample separately and producing the estimated optimal instrument in each subsample by using the fitted model estimated on the opposite one. Specifically, for $i \in I_1$ the estimated instrument $\widehat{f}_i = \widehat{f}(\mathcal{A}_2, Z_i)$ is a function of $\mathcal{A}_2$ and Z_i only, while for $i \in I_2$ the corresponding $\widehat{f}_i = \widehat{f}(\mathcal{A}_1, Z_i)$ is a function of $\mathcal{A}_1$ and Z_i. The *cross-fit split-sample* estimator of the structural parameter β is defined as:

$$\widehat{\beta}_{CFSS} = \frac{\sum_{i \in I_1} \widehat{f}(\mathcal{A}_2, Z_i) Y_i + \sum_{i \in I_2} \widehat{f}(\mathcal{A}_1, Z_i) Y_i}{\sum_{i \in I_1} \widehat{f}(\mathcal{A}_2, Z_i) X_i + \sum_{i \in I_2} \widehat{f}(\mathcal{A}_1, Z_i) X_i}. \tag{7}$$

The formulation above make sense if the two subsamples are of equal size, but we may also consider other alternatives such as weighted averages of the two individual split-sample estimates.

Jackknife. An extreme form of the sample splitting idea is the *jackknife* or *leave-one-out* (Angrist et al. (1999)) estimator, where in order to estimate the optimal instrument for observation i one uses the sample containing all observations except i. In the case where the first stage is estimated using OLS, this means (3) using $\widehat{f}_i = \widehat{\pi}'_{(-i)} Z_i$, with $\widehat{\pi}_{(-i)} = (Z'_{(-i)} Z_{(-i)})^{-1} Z'_{(-i)} X_{(-i)}$, where the index $(-i)$ indicates the matrix including all observations but i.

This idea may be applied to any other ML approach applied on the first step. For jackknife estimators let us for any index i denote $\mathcal{A}_{-i}$ the full set (X_j, Y_j, Z_j) of random variables in the samples excluding observation (X_i, Y_i, Z_i). Then the jackknife IV estimator is

$$\widehat{\beta}_{JIVE} = \frac{\sum_i \widehat{f}(\mathcal{A}_{-i}, Z_i) Y_i}{\sum_i \widehat{f}(\mathcal{A}_{-i}, Z_i) X_i}.$$

It is worth pointing out that for a complicated ML algorithm this estimator is computationally demanding as it requires re-running the ML first-stage estimation on each data set $\mathcal{A}_{-i}$ separately. For clarity, in this paper we name these estimators as JIVE-combined with the name of the first-stage estimator, e.g JIVE-LASSO runs a separate LASSO estimation for each observation i on the first stage. We refer to the estimator introduced in Angrist et al. (1999), in which the first stage is estimated using least squares, as JIVE-OLS.

Deleted diagonal estimators. Direct implementation of the jackknife form described above can be numerically demanding. However, Angrist et al. (1999) showed that the jackknife IV estimator with the OLS first step (JIVE-OLS) can be calculated as $\widehat{\beta}_{JIVE} = \frac{X'\widetilde{P}Y}{X'\widetilde{P}X}$, where the $n \times n$ matrix of weights $\widetilde{P}$ can be calculated from projection matrix $P = Z(Z'Z)^{-1}Z'$ by eliminating diagonal elements and re-scaling rows: $\widetilde{P}_{ij} = \frac{P_{ij}}{1-P_{ii}}$ if $i \neq j$, and $\widetilde{P}_{ii} = 0$. Notice that the JIVE-OLS estimate is the solution to the optimization problem that minimizes the quadratic form $(Y - \beta X)'\widetilde{P}(Y - \beta X)$, using the deleted diagonal weights $\widetilde{P}$. Han and Phillips (2006) show that, when the number of the moment conditions is large, the minimizer to the theoretical GMM objective function is not the true parameter. For TSLS, the value of the theoretical objective function at the true parameter value is equal to $\mathbb{E}\left[e'Pe\right] = \sum_i P_{ii}\mathbb{E}e_i^2 \neq 0$. As argued in Han and Phillips (2006), this leads to the bias of TSLS we discussed above. The JIVE-OLS estimator solves this problem, since $\widetilde{P}_{ii} = 0$.

Based on this idea, the JIVE-OLS formulation has inspired another class of estimators that is also called jackknife, though may or may not be associated with a direct jackknifing procedure. To distinguish these alternative estimators, in this paper we use the term *deleted diagonal.* Assume we have an estimator that is defined as the optimizer of some objective function that is either a quadratic form or ratio of two quadratic forms, say

TSLS, ridge, or LIML. In order to correct for the endogeneity bias, we instead solve a slightly corrected optimization problem which deletes diagonal elements of the quadratic form in the numerator. This idea works successfully for many estimators (TSLS, LIML, ridge) and results in DD-LIML (Hausman et al (2012), referred to in the literature as HLIM), and the DD-ridge (Hansen and Kozbur(2014)). Each one of these estimators with the deleted diagonal has superior finite-sample performance in comparison to the original estimators, as demonstrated in multiple simulation setups, and are consistent in larger number of settings.

Simulation evidence. It is worth noticing that in Angrist and Frandsen (2020) the JIVE-OLS and SS-OLS estimators systematically outperformed estimators with sophisticated ML approaches (LASSO and random forest) used on the first stage and without corrections for the endogeneity. This suggests that the endogeneity of the estimated instrument, in presence of a weak signal, may be the main challenge in the application of modern regularized methods to instrument selection. Another interesting observation is that the last proposed solution, 'deleting the diagonal', does not solve the problem of pre-test or model selection. For example, assume we employ the LASSO on the first step to select the parsimonious prediction model. Using the deleted diagonal approach for the second step estimation of the structural parameter with the selected instruments, does not account for the fact that the selected instruments may include some instruments that have high in sample correlation with the endogenous part of the regressor. This was demonstrated by Hansen and Kozbur (2014) in simulations, where the LASSO with deleted diagonal showed unimpressive performance.

4 Asymptotic inferences with many instruments

This section discusses asymptotic inference in the cross-sectional setting. We attempt to answer the important empirical question of when reliable statistical procedures exist for estimation with many available instruments in a low signal environment. The first subsection addresses the question of what theoretical restrictions on the concentration parameter and the number of instruments allow for an estimator to be consistent. The next subsection asks a similar question about asymptotic gaussianity of the estimator and its standard errors. Then we ask what empirical criteria a practitioner may check to pre-test if the signal is strong enough for gaussian inferences. The final subsection discusses identification robust testing as a default approach under very low signal.

4.1 When does a consistent estimator exist?

The signal strength of the optimal instrument that is required to achieve a consistent estimator, as measured by the concentration parameter (5), depends crucially on how much is known and how much we are willing to assume about the form of the optimal instrument. Let us start with the simplest case in which the optimal instrument is fully known and available, that is, f_i is a part of our data set and its identity is known. In this case, the optimal IV estimator $\widehat{\beta}_o$ is consistent as long as $\mu^2 \to \infty$ as the sample size increases. Under very mild assumptions $\sqrt{\mu^2}$ is the convergence rate of the optimal estimator (Stock et al. (2002)).

Conversely, assume that nothing is known about the form of the optimal instrument and that we search among all linear combinations of the available K instruments (for this result we assume that $K < n$). That is, assume the optimal instrument is linear $f_i = \pi' Z_i$, but that no information about the direction of π is available. Then, a necessary and sufficient condition for consistency of the IV estimator is $\frac{\mu^2}{\sqrt{K}} \to \infty$. The strength of identification should not just be large, but large in comparison to the complexity of the first-stage estimation, measured by the square root of the number of instruments. This necessary condition comes from a result in Mikusheva and Sun (2020) which states that, in the best possible circumstances (such as a linear, gaussian, homoskedastic model with known reduced form covariance of the error term), if $\frac{\mu^2}{\sqrt{K}}$ is bounded asymptotically and the direction of π is completely unknown, then one cannot consistently distinguish any two values of β. There are a number of estimators that are consistent under heteroscedasticity and some relatively minor technical assumptions when $\frac{\mu^2}{\sqrt{K}} \to \infty$. These include JIVE-OLS, DD-LIML and DD-Fuller (Hausman et al (2012)), with earlier results for the homoskedastic case obtained by Chao and Swanson(2005). Notice, that all these estimators are agnostic about the direction of the optimal instrument.

Consistency of split-sample IV with ML first stage. The negative result of Mikusheva and Sun (2020) critically relies on the direction of the optimal instrument being completely unrestricted and unknown. If a researcher has some knowledge about the optimal instrument and may adjust her first-stage estimation accordingly, then the requirement on the strength of identification is less stringent and depends on the rate of consistency of the first-step estimator. The following statement characterises the consistency of the estimator under such conditions.

Theorem 1 *Assume that the data is i.i.d.,* $\mathbb{E}[e_i|Z_i] = \mathbb{E}[v_i|Z_i] = 0$ *and* $\mathbb{E}[e_i^2|Z_i] < C$, $0 < c < \mathbb{E}[v_i^2|Z_i] < C$ *almost surely. Assume that the prediction error for the estimation approach used on the first step is such that* $\mathbb{E}\left[(\widehat{f}_i - f_i)^2\right] = O(\frac{r_n}{n})$ *and* $\mathbb{E}[\widehat{f}_i f_i] \geq c\mathbb{E}[f_i^2]$. *For the sample split estimator assume that the number of observations in* I_2 *is at least* $[\alpha n]$ *for* $\alpha > 0$. *For the cross-fit assume that the number of observations is equal in the both subsamples. If* $\frac{\mu^2}{\sqrt{r_n}} \to \infty$ *, then* $\widehat{\beta}_{CFSS}$ *and* $\widehat{\beta}_{SS}$ *are consistent for* β.

Theorem 1 shows that if one has information about the optimal instrument and can use it to improve the optimal instrument estimation rate, characterized by r_n, then the requirements on the strength of the optimal instrument may be weakened. If nothing is known about the instruments and the search is done among all linear combinations of K available instruments (with the assumption that $K < n$), then the approriate rate for the optimal instrument estimation is $r_n = K$, returning us to the earlier condition $\frac{\mu^2}{\sqrt{K}} \to \infty$.

If one is willing to impose assumptions on the first stage and use them, then better rates for optimal instrument estimation can be achieved. One such potential restriction is sparsity or approximate sparsity (Belloni et al, 2010) that allows the optimal instrument to be well approximated by a linear combination of a small number of the available instruments. In this case we may allow the number of available instruments, K, to be (much) larger than the sample size n. Let us assume that

$$f_i = Z_i'\pi_0 + R_i, \quad \|\pi_0\|_0 \leq s,$$

where the approximation error is small in the following sense $\sqrt{\frac{1}{n}\sum_i R_i^2} \leq C\sqrt{\frac{s}{n}}$, and the number of important terms is small $s = o(n/\log(K))$. The leading proposal to estimate the predictive sparse regression is via LASSO (Tibshirani, 1996). Under appropriate moment assumptions and assumptions on sub-matrices of $Z'Z$, Belloni et al (2010, Theorem

1) obtain the following rate of convergence for LASSO estimation of the first stage:

$$\|\widehat{\pi} - \pi_0\|^2 = O_p\left(\frac{s\log(K \vee n)}{n}\right).$$

The amazing feature of LASSO is that it allows the number of instruments K be much larger than n, but this number appears in the rate $r_n = s\log(K \vee n)$ only as a logarithm. The main impact on r_n is the sparsity of the first stage s, which reflects the number of important instruments needed to approximate the optimal instrument well. This number may grow very slowly with the sample size if the first stage is sparse. Thus, according to Theorem 1 the split-sample IV estimator of β with LASSO estimation of the sparse first stage, is consistent as long as $\frac{\mu^2}{\sqrt{s\log(K\vee n)}} \to \infty$. It should be noted that, unlike SS-OLS, the SS-LASSO is not invariant or agnostic to linear transformations of the instruments.

4.2 Are inferences standard?

In linear weak IV with a small number of instruments, once the TSLS estimator becomes consistent it is also asymptotically gaussian and standard TSLS formulas provide valid confidence sets. This is not the case with many available instruments. Two related phenomena concerning the asymptotic normality of many instrument estimators can be seen in the literature (e.g. Hansen et al (2008), Chao et al (2012)). Firstly, novel asymptotic statements are often used, such as a central limit theorem for quadratic forms, that are not used in standard TSLS asymptotics. Secondly, in many weak IV settings the usual TSLS standard errors formulas are incorrect.

Explanation of the problem. In order to explain the complication that arises from flexible first-stage estimation, let us consider the prototypical estimator $\widehat{\beta}$ defined in equation (3), that uses estimated instrument $\widehat{f}_i$:

$$\widehat{\beta} - \beta = \frac{\sum_{i=1}^n f_i e_i + \sum_{i=1}^n (\widehat{f}_i - f_i)e_i}{\sum_{i=1}^n f_i X_i + \sum_{i=1}^n (\widehat{f}_i - f_i)X_i}. \tag{8}$$

The consistency result requires proving that term $\sum_{i=1}^n f_i X_i = O_p(\mu^2)$ dominates the other three sums in (8). According to Theorem 1, this is true for the split-sample estimator whenever $\frac{\mu^2}{\sqrt{r_n}} \to \infty$.

For gaussian inference with regular standard errors, we also require that the numerator term $\sum_{i=1}^n f_i e_i = O_p(\mu)$ asymptotically dominates term $\sum_{i=1}^n (\widehat{f}_i - f_i)e_i$. In the case of the split-sample estimator described in Theorem 1, the latter term has asymptotic order $O_p(\sqrt{r_n})$. The condition for the leading term in the numerator to dominate is $\frac{\mu^2}{r_n} \to \infty$.

There is a gap between the rate required for consistency, and the stricter rate required for standard gaussian inference to be valid. Assume that the strength of identification is such that the estimator is consistent ($\frac{\mu^2}{\sqrt{r_n}} \to \infty$), but $\frac{\mu^2}{r_n}$ is asymptotically bounded. Then, the asymptotic distribution of the estimator depends more finely on the first-stage procedure and calls for asymptotic theorems for the term $\sum_{i=1}^n (\widehat{f}_i - f_i)e_i$. The biggest challenge in determining the asymptotic distribution of the last term is the complicated dependence between summands. Specifically, the first-stage estimation error $\widehat{f}_i - f_i$, will exhibit dependence over i if the first-stage estimation relies on common observations. For complicated ML procedures on the first stage, this dependence may be very intricate.

Some DD and JIVE estimators. This issue has been successfully solved for deleted diagonal style estimators and several JIVE estimators, including JIVE-OLS, JIVE-Ridge, DD-TSLS, DD-LIML and DD-Fuller (Chao et al (2012), Hansen et al (2008), Hausman et al (2012), Hansen and Kozbur (2014)). We discuss these results for the example of DD-TSLS. The DD-TSLS estimator equals to the ratio of two quadratic forms $\frac{X'\widetilde{P}Y}{X'\widetilde{P}X}$, where $\widetilde{P}$ equals to the projection matrix P with a deleted diagonal. It implicitly uses the estimated instrument

$$\widehat{f}_i = \mathbf{i}'\widetilde{P}X = \mathbf{i}'\widetilde{P}f + \sum_{j\neq i} \widetilde{P}_{ij}v_j,$$

with $\mathbf{i}$ denoting a selection vector with the ith component equal to 1 and all other elements equal to zero. For simplicity, let us ignore for a moment the distinction between $\mathbf{i}'\widetilde{P}f$ and f_i – this will be reasonably small for well chosen $\widetilde{P}$. The prediction mistake introduced in Theorem 1 is

$$\frac{1}{n}\sum_{i=0}^{n} \mathbb{E}(\widehat{f}_i - f_i)^2 \asymp \frac{1}{n}\sum_{i=0}^{n} \mathbb{E}\left[\sum_{j\neq i} \widetilde{P}_{ij}v_j\right]^2 \asymp \frac{K}{n}.$$

Thus, in this example we have the rate $r_n = K$, and by reasoning similar to Theorem 1, DD-TSLS (as well as the other DD estimators mentioned above) will be consistent when $\frac{\mu^2}{\sqrt{K}} \to \infty$. However, standard gaussian inferences require $\sum_{i=1}^n f_i e_i$ to dominate the numerator in equation (8), and hence $\frac{\mu^2}{K} \to \infty$. In the gap between these two rates, the rate needed for consistency and the rate needed for standard inferences, the asymptotic behavior of $\sum_{i=1}^n(\widehat{f}_i - f_i)e_i$ becomes the dominating one.

The first-stage estimation error of the optimal instrument $\widehat{f}_i - f_i = \sum_{j\neq i}\widetilde{P}_{ij}v_j$ is a weighted average of all but its own endogenous errors. This means that the estimation errors will be heavily correlated over i – notice also that $(\widehat{f}_i - f_i)$ is correlated with e_j (when $j \neq i$), as v_j is part of the first-stage estimation error and v_j is correlated with e_j. As a result, getting a central limit theorem for $\sum_{i=1}^n(\widehat{f}_i - f_i)e_i$ is highly non-trivial in general and calls for more structure to be put on $\widehat{f}_i - f_i$. Such structure exists in the above mentioned DD and JIVE estimators, and the leading term in $\sum_{i=1}^n(\widehat{f}_i - f_i)e_i$ is given by $\sum_{i=1}^n\sum_{j\neq i}\widetilde{P}_{ij}v_je_i$. Chao et al (2012) and Hansen et al (2008) establish a central limit theorem for quadratic forms of this type, that provides conditions for gaussianity of the leading term. Hausman et al. (2012) provides methods for estimating standard errors that work for several JIVE and DD-type estimators.

To summarize, once the identification is strong enough for a number of JIVE and DD estimators (including JIVE-OLS, JIVE-Ridge, DD-TSLS, DD-LIML) to become consistent ($\frac{\mu^2}{\sqrt{K}} \to \infty$), these estimators are also asymptotically gaussian (under mild additional assumptions). However, the standard errors needed for asymptotically valid inferences differ and require a quadratic form CLT to be used. It is worth pointing out that the standard errors proposed by Hausman et al. (2012) contain variance estimates for both terms appearing in the numerator of (8) and work well once the corresponding IV estimator is consistent.

Split-sample estimators. The theorem below establishes conditions for asymptotic gaussianity of the split-sample estimator $\widehat{\beta}_{SS}$. It shows that once the split-sample estimator is consistent, inference can be performed in a standard way, treating the estimated instrument $\widehat{f}_i$ as the only available instrument in a just-identified setting.

Theorem 2 *Assume that the data is i.i.d., $\mathbb{E}[\varepsilon_i|Z_i] = 0$, and $\mathbb{E}[|\varepsilon_i|^4|Z_i] < C$ for $\varepsilon_i = (e_i, v_i)'$ and $\mathbb{E}[f_i^4] < C$. Assume that the size of subsample I_2 is growing to infinity as $n \to \infty$. Let the following assumptions hold:*
(i) $\frac{1}{\left(\mathbb{E}[\widehat{f}_i^2|\mathcal{A}_1]\right)^2}\mathbb{E}\left[|\widehat{f}_i|^4|\mathcal{A}_1\right] < C$ *almost surely;*
(ii) $\frac{1}{a_n}\sum_{i\in I_2}\widehat{f}_if_i \to^p 1$ *for some $\mathcal{A}_1$-measurable sequence of random variables a_n;*

(iii) $\mu_n^2 = \frac{a_n^2}{\sum_{i\in I_2}\mathbb{E}\left[\widehat{f}_i^2|\mathcal{A}_1\right]} \to^p \infty.$

Then $\widehat{\beta}_{SS}$ *is consistent and is asymptotically mixed gaussian. Define a variance estimator* $\Sigma_n^2 = \frac{\sum_{i\in I_2}\widehat{f}_i^2\widehat{e}_i^2}{\left(\sum_{i\in I_2}\widehat{f}_iX_i\right)^2}$*, where* $\widehat{e}_i = Y_i - \widehat{\beta}_{SS}X_i$*. Then*

$$\Sigma_n^{-1}(\widehat{\beta}_{SS} - \beta) \Rightarrow N(0,1) \text{ as } n \to \infty.$$

An interesting feature of Theorem 2 is that it does not require or imply that the estimated variance Σ_n or the informativeness of the constructed instrument a_n converges to a constant. To the contrary, Theorem 2 allows the scale of the deviations $\widehat{\beta}_{SS} - \beta$ to be asymptotically random. This is important as the theorem does not require or assume that the estimation/selection of instruments in the first stage is consistent, but rather accommodates the asymptotic uncertainty of the first-stage estimator. The idea behind the proof is to do all inferences conditional on the first subsample. Notice that $\widehat{f}_i$ is exogenous in the conditional environment, and so the second stage is a standard IV estimation with one instrument. This estimate is consistent and (conditionally on the first subsample) asymptotically gaussian with the usual formula for standard errors as long as the signal in $\widehat{f}_i$ guarantees consistency (condition (iii)).

It is worth pointing out that Theorem 2 does not specify what technique is used in the first stage, only the prediction mistake rate. This makes the split-sample estimator suitable for use with a variety of ML approaches, which may be adapted to any information about the first stage the researcher possesses.

Cross-fit estimator. Theorem 1 is applicable to both split-sample and cross-fit estimators equally as they are either consistent or inconsistent under the same conditions. One may expect the cross-fit split-sample estimator to be more efficient asymptotically as it is effectively using both subsamples, while $\widehat{\beta}_{SS}$ uses only part of the sample in the second stage. Indeed, if $\frac{\mu^2}{r_n} \to \infty$ and the standard term $\sum_i f_ie_i$ dominates the numerator, $\widehat{\beta}_{CFSS}$ is asymptotically more efficient than $\widehat{\beta}_{SS}$. However, we cannot recommend $\widehat{\beta}_{CFSS}$ to wide use as its asymptotic behavior when $\frac{\mu^2}{\sqrt{r_n}} \to \infty$ but $\frac{\mu^2}{r_n} \not\to \infty$, is not well understood.

Unfortunately, the corresponding conditions for gaussianity of the cross-fit estimator $\widehat{\beta}_{CFSS}$ (or whether it is asymptotically gaussian at all) are unknown when $\frac{\mu^2}{\sqrt{r_n}} \to \infty$, but $\frac{\mu^2}{r_n} \nrightarrow \infty$. The challenge is that the term determining the distribution of the cross-fit estimator:

$$\sum_{i \in I_1} \widehat{f}(\mathcal{A}_2, Z_i) e_i + \sum_{i \in I_2} \widehat{f}(\mathcal{A}_1, Z_i) e_i$$

has a very complicated cross-term dependence. In particular, one can establish asymptotic gaussianity of $\sum_{i \in I_1} \widehat{f}(\mathcal{A}_2, Z_i) e_i$ conditional on $(\mathcal{A}_2, \mathcal{Z}_1)$ with the random asymptotic variance depending on $\mathcal{A}_2$ and $\mathcal{Z}_1$. A symmetric statement for the other sum can be obtained conditionally on $\mathcal{A}_1$ and $\mathcal{Z}_2$. However, the joint distribution of two sums is unclear as the conditioning variables in the first sum $\mathcal{A}_2$ are correlated with the error terms e_i forming the second sum over $i \in I_2$. It seems that more details about asymptotics of the first-step estimation, for example, the influence function representation of $\widehat{f_i} - f_i$, would be useful here.

4.3 Can we pre-test for weak identification?

An empirical researcher typically wants to know whether confidence sets or t-tests based on the gaussian approximation are reliable in a particular setting. A pre-testing procedure for weak identification may be useful for this purpose. The empirical plan is to use gaussian confidence sets and t-statistics if the pre-test suggests that the information in the instruments is strong enough to support asymptotically gaussian inferences, and to use some weak identification robust procedure otherwise.

Pre-tests for weak identification when the number of instruments is small are typically evaluated using the first-stage F statistic, or some robust version of it. However, the first-stage F-statistic is a less universally applicable pre-test than most researchers believe (see, Andrews et al (2019)). Specifically, it was designed for the TSLS estimator in a homoskedastic setting with a small number of instruments. Forms of the first-stage F pre-test that work for heteroscedastic/autocorrelated errors in a just identified linear IV also exist. Typically, the first-stage F does not work well outside of these two models (Andrews et al. (2019)) and it should not be used in settings with many instruments (Mikusheva and Sun (2020)).

This subsection discusses how to construct a pre-test, while the next section discusses construction of robust tests/confidence sets. The correct pre-test depends crucially on the estimator that will be used after strong identification is detected. As we saw in the previous section, we are aware of two classes of estimators that are asymptotically gaussian once the consistency condition is satisfied: (i) several JIVE and DD-type estimators; and (ii) split-sample estimators for a wide class of ML first stages.

The main requirement of a good pre-test for weak identification is that it should provide size guarantees to the combined procedure including the pre-test and the procedure used after the pre-test. One approach is to first derive a distribution for the $t-$statistics under the assumption of weak identification; that is, in a setting where the t- statistics are not asymptotically gaussian. We can then identify the parameter which governs the accuracy of the gaussian approximation, and determine what values of that parameter guarantee acceptable size distortions.

Pre-test for DD-TSLS estimator. Mikusheva and Sun (2020) derived the asymptotic approximation for the DD-TSLS Wald statistic, under assumptions that imply it is not consistent, and suggested a pre-test that can be used. Specifically, the pre-test statistic has the following form $\widetilde{F} = \frac{1}{\sqrt{K}\widehat{\Upsilon}} \sum_{i=1}^{n} \sum_{j \neq i} P_{ij} X_i X_j$, where $\widehat{\Upsilon}$ is an estimator of uncertainty in the first stage. Mikusheva and Sun (2020) show that, under mild regularity conditions, the DD-TSLS estimator $\widehat{\beta}_{DD}$ and an estimator of its variance $\widehat{V}$ (suggested in Chao et al (2012)) converge jointly:

$$\left(\frac{(\widehat{\beta}_{DD} - \beta_0)^2}{\widehat{V}}, \widetilde{F} \right) \Rightarrow \left(\frac{\xi^2}{1 - 2\varrho\frac{\xi}{\nu} + \frac{\xi^2}{\nu^2}}, \nu \right), \tag{9}$$

where ξ and ν are two normal random variables with means 0 and $\frac{\mu^2}{\sqrt{K}}$, unit variances, and a correlation coefficient ϱ that is related to the endogeneity of the structural error. Statement (9) holds for a wide range of values for $\frac{\mu^2}{\sqrt{K}}$ and shows the distortions from gaussianity of the DD-TSLS estimator when it is inconsistent. Notice that when $\frac{\mu^2}{\sqrt{K}}$ is large, the typical realization of ν is also large, and the the limit of the first component approaches ξ^2, so that the Wald statistic is approximately χ_1^2 distributed.

According to statement (9), the theoretical parameter $\frac{\mu^2}{\sqrt{K}}$ controls the size distortions of the DD-TSLS Wald statistic, while the statistic $\widetilde{F}$ provides empirical characterization of its size. Mikusheva and Sun (2020) suggest a pre-test for weak identification that compares $\widetilde{F}$ to a cut-off of 4. If one uses the 5%-size DD-TSLS Wald test when $\widetilde{F}$ is above the cut-off, and a 5%-size weak identification robust test otherwise, then the full procedure has asymptotic size less than 15%. The ideas and results of Mikusheva and Sun (2020) can most likely be extended to other DD-type estimators and several JIVE estimators that have DD-form.

Pre-test for the split-sample estimator. A distinct feature of the split-sample estimator introduced in equation (6) is that the selection and construction of the optimal instrument happens on a subsample I_1 that is not used in the estimation of the structural coefficient. This makes the estimated instrument $\widehat{f_i} = \widehat{f}(\mathcal{A}_1, Z_i)$ exogenous and the second step regression just identified. While the first-stage F pre-test does not produce reliable results in over-identified linear IV with heteroscedastic errors (Andrews et al., 2019), there is a form of the first-stage F pre-test that works reliably in a just identified linear IV.

We suggest the following procedure. First, estimate the first-stage model (predicting X_i using Z_i) using data $\mathcal{A}_1$ only – the researcher has full freedom in choosing any ML approach for this stage. Produce the prediction $\widehat{f_i} = \widehat{f}(\mathcal{A}_1, Z_i)$ on the sample $i \in I_2$. Then, run the regression of X_i on $\widehat{f_i}$ for the sample $i \in I_2$, and calculate the heteroscedasticity robust F-statistic. If the F statistic exceeds 10 the researcher can rely on gaussian confidence sets as stated in Theorem 2, otherwise she should use a weak identification robust test, some of which are suggested in the next subsection. The size of the two-step inference procedure is asymptotically less than 15%.

4.4 Robust tests when identification is weak

In empirical settings where the optimal instrument does not contain sufficient information for producing a consistent estimator, it is common to report statistical inferences through identification-robust confidence sets. Identification-robust confidence sets are constructed by inverting robust tests for point hypothesis about the structural parameter $H_0 : \beta = \beta_0$ (i.e. finding the set of null hypotheses that cannot be rejected). Robustness to weak identification means having asymptotically correct size uniformly over a wide range of identification strengths, including weak identification.

The null hypothesis $H_0 : \beta = \beta_0$ in the linear IV model is equivalent to testing many moment conditions:

$$\mathbb{E}[Z_{ij}(Y_i - \beta_0 X_i)] = 0 \text{ for } j = 1, ..., K.$$

It is easy to construct a test of asymptotically correct size for any single instrument by calculating the sample correlation between the instrument and the quasi-error $(Y_i - \beta_0 X_i)$. However, with multiple instruments, such a test will likely have close to trivial power. The biggest challenge is to aggregate information across an increasing number of instruments, while properly accounting for the uncertainty about the optimal instrument.

One way to use an increasing number of instruments is to aggregate all of them using the deleted diagonal idea. This results in the Anderson-Rubin type test proposed in Mikusheva and Sun (2020):

$$DD\text{-}AR(\beta_0) = \frac{1}{\sqrt{K\widehat{\Phi}}} \sum_{i \neq j} P_{ij}(Y_i - \beta_0 X_i)(Y_j - \beta_0 X_j),$$

where P_{ij} is the $n \times n$ projection matrix on the space of instruments and $\widehat{\Phi}$ is an estimator of the asymptotic variance (refer to Mikusheva and Sun (2020) for details). The test rejects the null if the statistic exceeds the $(1 - \alpha)$-quantile of the standard normal distribution. The test relies on the central limit theorem for quadratic forms with deleted diagonal as in Chao et al. (2012). The important feature of this test is its consistency against fixed alternatives whenever $\frac{\mu^2}{\sqrt{K}} \to \infty$, that is, whenever, estimators that are agnostic about the direction of the optimal instrument, such as JIVE-OLS or DD-LIML, are consistent. This test works well when no information about the direction of the optimal instrument is available.

If one expects that a relatively small number of instruments may capture most of the information (i.e., the first stage is sparse), then the maximum score statistic will have superior power properties. This idea was proposed in Belloni et al. (2012), where the test statistic has the form:

$$\Lambda(\beta_0) = \max_{j=1,\ldots,K} \frac{|\sum_i^n (Y_i - \beta_0 X_i) Z_{i,j}|}{\sqrt{\sum_i^n (Y_i - \beta_0 X_i)^2 Z_{i,j}^2}}.$$

See Belloni et al. (2012) for the details of how to construct critical values. The power of this test comes from the most informative instrument in the given set.

Sample splitting for testing. Here we propose a new split-sample testing idea that fits well with the previously introduced split-sample estimator. It allows the researcher to use any information about the form of optimal instrument by using the ML approach best fit for the application.

The proposed test proceeds as follows. As before we randomly split the data set into two subsamples indexed by I_1 and I_2. The researcher uses the method of her choice to estimate the optimal instrument using only observations $i \in I_1$, and then constructs the estimated optimal instrument on the second subsample: $\widehat{f}_i = \widehat{f}(\mathcal{A}_1, Z_i)$ for $i \in I_2$. The linear IV regression on the second subsample using the *exogenous* estimated optimal instrument $\widehat{f}_i$ is a just identified IV model. We suggest using the Anderson-Rubin test:

$$SS\text{-}AR(\beta_0) = \frac{\left(\frac{1}{\sqrt{|I_2|}} \sum_{i \in I_2} (Y_i - \beta_0 X_i) \widehat{f}_i\right)^2}{\widehat{\Sigma}},$$

where $\widehat{\Sigma}$ is a consistent heteroscedasticity-robust estimator of the variance of $\frac{1}{\sqrt{|I_2|}} \sum_{i \in I_2} \widehat{f}_i e_i$. The test rejects the null when the test statistic exceeds the $(1-\alpha)$-quantile of the χ^2_1 distribution. In the just identified case, the Anderson-Rubin statistic is equivalent to all commonly used weak identification robust tests (KLM, CLR) and is asymptotically efficient under strong identification. The common consensus in the literature is that in a just identified model no pre-test for weak identification is needed, and Anderson-Rubin confidences sets should always be used, independently of the identification status of the model.

The power trade-off between the proposed split-sample test and the two tests discussed above is a balance between two forces. The sample split is more flexible and adaptive to estimation of the optimal instrument, and may produce better power by using a more powerful estimate. On the other hand, the sample split uses only half of the sample for testing the structural parameter and thus may be less efficient than the other two tests discussed here.

5 Many Weak Instruments in Time Series

The cross-sectional literature on many weak instruments pointed towards one big obstacle arising from a very flexible first stage: the estimated optimal instrument may be endogenous. It also suggests several solutions that work successfully in cross-sectional settings. The successful approaches include sample splitting, cross fit, jackknife and deleting the diagonal. Another method, that is very popular in the time series literature but has had almost no traction in cross-section, is Factor model IV (FIV).

This section is structured in the following way. First, we point out two additional challenges posed by time series data. Next, we discuss the promises and pitfalls of the Factor IV approach. We then introduce split-sample procedures that can be used in many weak IV with autocorrelated data, paying special attention to the peculiarities of time series data. Finally, we describe theoretical challenges that prevent us from proposing modifications of cross-fit, jackknife or deleted diagonal approaches to time series data and suggest what type of theoretical results would be helpful for making these approaches possible.

5.1 Challenges of Time Series

The challenges that arise from times series data are related to two features: autocorrelation of error terms, and weak exogeneity of instruments.

Challenge 1: Autocorrelated errors. In the New Keynsian Phillips Curve example, the error in the structural equation

$$e_t = \pi_t - \lambda x_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1} = u_t - \gamma_f(\pi_{t+1} - \mathbb{E}_t \pi_{t+1}) \tag{10}$$

is a moving-average (MA(1)) process. Similarly, the errors in other rational expectations models tend to be autocorrelated. Autocorrelation of structural errors calls for using heteroscedasticity and autocorrelation (HAC) robust standard errors in all statistical inferences. At the same time, there are also opportunities to exploit the dependence in order to improve efficiency, for example, by choosing the optimal weighting matrix in GMM. These considerations are equivalent to the question of choosing the optimal instrument. If errors follow a martingale process, then the result of Chamberlain (1987) still holds and the optimal instrument is $f_t = \frac{\mathbb{E}[X_t|\mathcal{Z}_t]}{\mathbb{E}[e_t^2|\mathcal{Z}_t]}$. However, if the errors are autocorrelated, then the optimal instrument has a more complicated structure (Hansen, 1985, Hansen et al, 1988, and Anatolyev, 2003). The optimal instrument depends on the Wold decomposition of the structural error and suggests first filtering the structural equation in order to obtain structural equations with martingale-difference errors. Filtering is a delicate task according to Hayashi and Sims (1983), who warn against backward-filtering (known as GLS) and suggest a forward-filtering procedure. A survey by Anatolyev (2006) on methods for creating optimal instruments with time series data, shows that all methods of constructing the optimal instrument that exploit the autocorrelation structure, assume that a consistent estimator for β exists and the covariance structure of the error term can be estimated consistently. This is infeasible in our setting with a potentially weak signal, thus we will maintain the previous goal post and focus on estimating $f_t = \mathbb{E}[X_t|\mathcal{Z}_t]$.

Challenge 2: Weak exogeneity. In cross-section, the exogeneity assumption is typically formulated as $\mathbb{E}[e_i|Z_i] = 0$, and when combined with independence across i it implies that $\mathbb{E}[e_i|Z] = 0$, where Z is the full set of instruments for all observations. This statement has important implications for theory as it allows us to employ arguments that first condition on instruments and exogenous regressors, after which we may treat instruments as fixed.

In economic applications in time series, the typical exclusion restriction is formulated using a weak exogeneity condition

$$\mathbb{E}[e_t|Z_t, Z_{t-1}, ...] = 0,$$

which assumes that e_t is uncorrelated with all instruments (and functions of them) taken at the current and past periods. This allows the structural shock to have an effect on future values of the instrument, and it is very likely that such an effect exists as all variables used in macro estimation simultaneously develop as a dynamic process. This means that we typically cannot make a strict exogeneity assumption that is formulated as $\mathbb{E}[e_t|Z_s, s \in \{1, ..., T\}] = 0$. The absence of strict exogeneity puts additional restrictions on what can and should be done with time series data, as it warns against mixing observations from different periods. For example, Hansen and West (2002) provide a formal argument against using GLS-type procedures in this setting.

5.2 Existing method: Factor Model IV

Factor Models provide a very attractive paradigm for parsimonious description of macroeconomic data sets containing a large number of macroeconomic indexes. The underlying assumption is that the cross-series dependence between indexes (or instruments in our setting) is driven by a relatively small number of unobserved factors:

$$Z_t = \lambda F_t + \epsilon_t,$$

where Z_t is $K \times 1$ vector of instruments observed at time t, λ is $K \times r$ matrix of non-random loadings and F_t is a set of $r \times 1$ factors. It is commonly assumed that r is much smaller than K. In its strictest form, a factor model assumes that all components of the error terms ϵ_t are uncorrelated both cross-sectionally and over time. An approximate factor model allows for very weak correlation in ϵ_t. The discovery of factor structure is typically done via Principle Component Analysis (PCA). Bai (2003) and Bai and Ng (2002) establish asymptotic properties of PCA and suggest a method for selecting the number of factors.

There is a consensus in the empirical literature that factor models fit the typical macroeconomic data very well. Stock and Watson (2005) put together a widely used set of quarterly macro indicators that includes 132 individual series. About eight to nine factors explain the vast majority of variation contained in the data. Similar observations can be made about other modern large data sets of macroeconomic indexes, such as FRED-QD and FRED-MD put together by McCracken and Ng (2020). Since introduction, factor models have been widely used for forecasting (e.g., Stock and Watson (2002 a,b)) and in structural estimation (e.g. Stock and Watson (2005), Bernanke et al. (2005)).

Bai and Ng (2010) propose to use factor models as a mechanism for constructing higher quality instruments. Their idea is to extract a small number of principle components F_t and use them as a smaller set of informative instruments, rather than using all individual series Z_t. Namely,

$$\widehat{\beta}_{FIV} = \frac{X' P_{\widehat{F}} Y}{X' P_{\widehat{F}} X}, \quad P_{\widehat{F}} = \widehat{F}(\widehat{F}'\widehat{F})^{-1}\widehat{F}', \tag{11}$$

where $\widehat{F}$ are estimated factors obtained by PCA on the set of instruments. The hope here is that the use of a small number of factor-instruments alleviates the bias coming from using many instruments, while preserving most of economic information contained in the data.

One positive feature of FIV is that it constructs instruments using only data on Z but not on X or Y. This alleviates a lot of concerns about endogeneity of the selected instruments. Aside from the PCA estimation mistake, this makes $\widehat{F}_t$ weakly exogenous. Thus, the second step (11) IV regression is a standard one. If identification is strong, the usual TSLS inferences apply. In the case of weak identification, the standard robust tests deliver valid results as shown in Kapetanios and Marcellino (2010) and Kapetanios et al (2016).

The biggest issue with FIV is its efficiency. The positive result in Bai and Ng (2010) states that if the endogenous regressor also obeys a factor structure, in particular, if

$$X_t = \mu F_t + v_t, \tag{12}$$

where v_t is uncorrelated with instruments Z_t, then FIV delivers a semi-parametric efficient estimator. Indeed, if the regressor X is correlated with the instruments only through factors then it is clear that the optimal instrument is a linear combination of these factors, and PCA does a good job of estimating this. However, the question arises of how likely assumption (12) is to hold in a particular application. Indeed, the factors that best explain variation in Z are not always best in explaining X. Bai and Ng (2009) suggest to pair ideas of FIV with different instrument selection criteria or boosting, that use the regressor X to assess the informativeness of the instrument. These selection criteria bring back the concerns about endogeneity of the selected instrument, and as such can be considered as another ML approach among many other discussed below.

5.3 New suggestion: Sample splitting

Let us define $\mathcal{Z}_t$ to be the sigma-algebra generated by instruments observed up until time t. The main structural equation is

$$Y_t = \beta X_t + e_t, \quad \mathbb{E}[e_t|\mathcal{Z}_t] = 0$$

and we define the 'optimal' instrument as: $f_t = \mathbb{E}[X_t|\mathcal{Z}_t]$, which implies $X_t = f_t + v_t$ where $\mathbb{E}[v_t|\mathcal{Z}_t] = 0$.

It is common in applications to use lags of the endogenous regressor and dependent variable as instruments. Thus, we assume that e_t and v_t are measurable with respect to $\mathcal{Z}_{t+p}$, where p is the lag between the observed regressor/ dependent variable and the first time they enter the set of instruments. For example, in the NKPC example one of the regressors is future inflation π_{t+1}, while the instruments may (and often do) include lagged inflation π_{t-1}, thus $p = 2$. If $p = 1$, then (e_t, v_t) forms a martingale difference sequence with respect to the filtration $\mathcal{Z}_t$, and error terms are not autocorrelated. In the NKPC example, the error terms are not martingale differences but rather an autocorrelated process with MA(1) structure.

Split-sample estimator. Given the success in developing robust asymptotic inference in cross-section using the split-sample approach, we extend the idea to the time series setting as well. We divide the sample into two subsamples with a gap of $p-1$ observations in between them (to guarantee exogeneity). The first subsample is used to elicit information about the optimal instrument, while the second is used for running just identified IV regression with the estimated instrument. As before, we do not require the subsamples to be of equal size and are rather agnostic about the method used on the first subsample to construct the estimated optimal instrument.

To simplify the notation, we re-numerate the observations so the first subsample has non-positive indexes $t = -\tau_0, ..., -p + 1$, while the second subsample is indexed by $t = 1, ..., \tau$ with $T = \tau_0 + \tau + 1$. We assume that as the total sample size increases we have $\tau \to \infty$. Assume that the first subsample is used to estimate f_t, employing some regression or ML technique. The estimated instrument $\widehat{f}_t$ for $t > 0$ is constructed using the model estimated on the first subsample applied to the instrument Z_t. This implies that $\widehat{f}_t$ is measurable with respect to $\mathcal{Z}_t$ and thus is weakly exogenous for all $t > 0$. The split-sample estimator is defined as

$$\widehat{\beta}_{SS} = \frac{\sum_{t=1}^{\tau} \widehat{f}_t Y_t}{\sum_{t=1}^{\tau} \widehat{f}_t X_t}. \tag{13}$$

The properties of the split-sample estimator may be highly influenced by the randomness of the first stage, even when the sample size is large, due to inconsistent estimation/selection on the first stage. For example, assume that the estimated optimal instrument is constructed as $\widehat{f}_t = \widehat{\pi}' Z_t$, where $\widehat{\pi}$ is a $K \times 1$ vector of loadings estimated on the first sample. We want to avoid making the assumption that either $\widehat{\pi}$ converges to a non-random vector, or that $\widehat{f}_t - f_t$ converges to zero so fast that its randomness does not matter. Instead, since the vector of loadings $\widehat{\pi}$ is measurable with respect to $\mathscr{Z}_0$, we can derive asymptotic statements conditionally on the sigma-algebra $\mathscr{Z}_0$.

Theorem 3 *Assume there exist a sequence of almost sure positive-definite random matrix η_τ and sequences of random variables μ_τ both measurable with respect to $\mathscr{Z}_0$ such that the following two conditions are satisfied for $\varepsilon_t = (e_t, v_t)'$:*

$$\eta_\tau^{-1} \sum_{t=1}^{\tau} \widehat{f}_t \varepsilon_t \Rightarrow N(0, I_2), \quad \mathscr{Z}_0 - \textit{ stably as } \quad \tau \to \infty, \tag{14}$$

$$\frac{1}{a_\tau} \sum_{t=1}^{\tau} \widehat{f}_t f_t \to^p 1. \tag{15}$$

If $\mu_\tau = \frac{a_\tau}{\|\eta_\tau\|} \to^p \infty$ then $\widehat{\beta}_{SS}$ is consistent for β. In addition, $\widehat{\beta}_{SS}$ is asymptotically mixed gaussian:

$$\Sigma_T^{-1}(\widehat{\beta}_{SS} - \beta) \Rightarrow N(0,1) \quad \mathscr{Z}_0 - \textit{ stably as } \quad \tau \to \infty,$$

where $\Sigma_T = \frac{\eta_{\tau,11}}{a_\tau}$.

The concept of stable convergence, discussed in a textbook by Hausler and Luschgy (2015), is a middle ground between weak convergence and convergence in probability. It allows for a stochastic normalization in the central limit theorem. In simplified terms, the stable convergence can be thought as a weak convergence conditional on the sigma-algebra $\mathscr{Z}_0$ (the first subsample). As we will not assume the consistency of the model estimation/selection on the first subsample, it is critical to preserve the randomness introduced by it to the split-sample estimation. In order to construct reliable inferences, we consider asymptotic approximations ($\tau \to \infty$) on the second subsample only, while using finite-sample inferences (through conditioning) on the first subsample.

Martingale CLT. One may ask under what conditions the central limit theorem (14) would hold in the split-sample setting. Firstly, consider the case where $p = 1$ and thus $\widehat{f}_t \varepsilon_t$ is a martingale-difference sequence with respect to filtration $\mathscr{Z}_{t+1}$. Here we can appeal to a martingale central limit theorem (see, Hausler and Luschgy (2015) Theorem 6.1 and Remark 6.8). It implies that if $\max_t \mathbb{E}\left[\|\varepsilon_t\|^{2+2\delta}|\mathcal{Z}_t\right] < \infty$ a.s. for some $\delta > 0$ and

$$\eta_\tau^{-2}\sum_{t=1}^{\tau}\mathbb{E}[\varepsilon_t\varepsilon_t'|\mathcal{Z}_t]\widehat{f}_t^2 \to^p I_2 \text{ and } \frac{1}{\|\eta_\tau\|^{1+\delta}}\sum_{t=1}^{\tau}\widehat{f}_t^{2+2\delta} \to^p 0, \qquad (16)$$

then statement (14) holds. The second of the two conditions in (16) is a very typical Lyapunov condition of asymptotic negligibility.

The first condition in (16) may come from the law of large numbers for non-ergodic sequences. For example, assume that the errors are conditionally homoskedastic $\mathbb{E}[\varepsilon_t\varepsilon_t'|\mathcal{Z}_t] = \Sigma$ and the estimated optimal instrument is constructed as $\widehat{f}_t = \widehat{\pi}'Z_t$, where $\widehat{\pi}$ is a $K \times 1$ vector of random estimated loadings measurable with respect to $\mathcal{Z}_0$. Denote $\xi = \widehat{\pi}'\mathbb{E}[Z_tZ_t']\widehat{\pi}$, then

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\widehat{f}_t^2 - \xi = \widehat{\pi}'\left(\frac{1}{\tau}\sum_{t=1}^{\tau}Z_tZ_t' - \mathbb{E}[Z_tZ_t']\right)\widehat{\pi}.$$

Then, some combination of a large-dimensional law of large numbers for stationary variables Z_tZ_t' in some norm, as well as restrictions on some norm of $\widehat{\pi}$, will give the first statement in (16). Depending on the methods used on the first subsample, many different combinations of assumptions may fit the bill. Also, notice that randomness (inconsistency) of $\widehat{\pi}$ would make η_τ random as well through ξ.

Now let us allow for autocorrelated errors by considering the case of $p = 2$. Now we have that ε_t is measurable with respect to $\mathcal{Z}_{t+2}$ but may be not measurable with respect to $\mathcal{Z}_{t+1}$, while $\mathbb{E}[\varepsilon_t|\mathcal{Z}_t] = 0$ and $\widehat{f}_t$ is measurable with respect to $\mathcal{Z}_t$. This is the situation we encounter in the NKPC example. Let us define $u_{1t} = \mathbb{E}[\varepsilon_t|\mathcal{Z}_{t+1}]$ and $u_{2t} = \varepsilon_t - \mathbb{E}[\varepsilon_t|\mathcal{Z}_{t+1}]$. Then u_{1t} is measurable with respect to $\mathcal{Z}_{t+1}$ and $\mathbb{E}[u_{1t}|\mathcal{Z}_t] = 0$, while u_{2t} is measurable with respect to $\mathcal{Z}_{t+2}$ and $\mathbb{E}[u_{2t}|\mathcal{Z}_{t+1}] = 0$. In general $\widehat{f}_te_t$ is not a martingale difference sequence in this case. However, we can re-arrange the summands to give:

$$\frac{1}{\sqrt{\tau}}\sum_{t=1}^{\tau}\widehat{f}_t\varepsilon_t = \frac{1}{\sqrt{\tau}}\sum_{t=1}^{\tau}\widehat{f}_t(u_{1t}+u_{2t}) = \frac{1}{\sqrt{\tau}}\sum_{t=2}^{\tau}(\widehat{f}_tu_{1t}+\widehat{f}_{t-1}u_{2,t-1}) + o_p(1).$$

Notice that sequence $\widehat{f}_tu_{1t}+\widehat{f}_{t-1}u_{2,t-1}$ is both measurable with respect to $\mathcal{Z}_{t+1}$ and mean zero conditionally on $\mathcal{Z}_t$. Thus, it is a martingale difference sequence, and the previous discussion about the martingale central limit theorem applies. This argument sketches why statement (14) may hold for autocorrelated errors, though in practice it also implies that asymptotic variances should be calculated using HAC robust procedures.

Inference procedures based on sample splitting. As we have already noted, a key advantage of the split-sample idea is that the IV regression on the second subsample is just-identified. The theory of weak identification is well developed for this case, and also allows for HAC robust inferences. Many issues that plague the weak IV literature, such as difficulties constructing a good pre-test for weak identification (Andrews et al (2019)), power trade-offs for robust tests, and empty Anderson-Rubin confidence sets (Davidson and MacKinnon (2014)), all apply to over-identified linear IV only. There is a consensus that the robust AR test/confidence set should be the default inference procedure in a just-identified linear IV setting, independently on whether or not weak identification is an issue. Indeed, the Anderson-Rubin procedure in the just identified setting is of asymptotically correct size and is asymptotically efficient under both weak and strong identification. It is also very easy to implement.

Specifically, in order to test $H_0 : \beta = \beta_0$, one calculates the implied error term $e_t(\beta_0) = Y_t - \beta_0 X_t$ and tests whether or not variable $\xi_t = \widehat{f}_t e_t(\beta_0)$ has mean zero using a standard Wald test:

$$AR(\beta_0) = \frac{\left(\frac{1}{\sqrt{\tau}} \sum_{t=1}^{\tau} \xi_t\right)^2}{\widehat{\sigma}^2},$$

where $\widehat{\sigma}^2$ is a consistent proxy for the random normalization $(\eta\eta')_{11}$ appearing in condition (14). For example, if $p = 1$ and ξ_t is a martingale difference, then

$$\widehat{\sigma}^2 = \frac{1}{\tau} \sum_{t=1}^{\tau} \left(\xi_t - \frac{1}{\tau} \sum_{s=1}^{\tau} \xi_s \right)^2$$

is a consistent proxy under the relatively general assumptions that are needed for the stable martingale central limit theorem (see Lemma 6.5 in Hausler and Luschgy (2015)). For autocorrelated errors one simply needs to use a HAC robust estimator of the asymptotic variance. The test accepts the null hypothesis if the AR statistic does not exceed the $(1 - \alpha)$-quantile of the χ^2_1 distribution. The Anderson-Rubin confidence set is constructed by inverting the AR test, that is, by collecting all values of β_0 not rejected by the test. In this case, test inversion can be done analytically as the AR statistic is a ratio of two quadratic functions of β_0 (see Mikusheva (2010)).

We suggest that, if a researcher decides to use the split-sample approach, she reports inferences mainly via the AR confidence set. The researcher may also choose to report the split-sample estimate as well, since it always lies inside the AR confidence set in a just-identified linear IV. The reliability of the split-sample estimate can be assessed by reporting the (HAC) robust first-stage F-statistic, applying a cut-off of roughly 10.

There is an alternative to the split-sample approach if identification is weak. A weak identification robust test for time series with many instruments is proposed in Dovi (2020). It uses ideas similar to the maximum score test proposed by Belloni et al (2012). Dovi (2020) applies it to the NKPC and shows that the power is superior in comparison to a random choice of instruments.

Machine Learning techniques to be used on the first subsample. One attractive feature of the split-sample approach is that it is agnostic about which procedure is used on the first subsample to estimate $f_t = \mathbb{E}[X_t|\mathcal{Z}_t]$. Here we want to mention several ML approaches that have been successfully used in applied macroeconomic research for such a forecasting task.

Firstly, there exist several methods that have strong theoretical foundations and results on the speed of convergence that are tailored to time series. Carrasco and Rossi (2016) consider multiple regularization approaches to forecasting using a large number of predictors including ridge, Landweber-Fridman estimation, and partial least squares. They obtained rates of convergence under a wide range of assumptions including approximate factor structure on one extreme and an ill-posed model on the other. They proposed a cross-validation method for choosing the tuning parameter and establish its optimality. Babii et al (2020) proposed LASSO selection in a time-series context, established consistency and asymptotic normality of the estimated coefficients under proper mixing conditions, and derived HAC-type estimators for standard errors.

Secondly, there are several methods that demonstrate great empirical results in applications but may miss formal theoretical statements on consistency or a speed of convergence. For example, Bai and Ng (2009) proposed several promising ideas on how to select good instruments. One of their suggestions, a procedure called boosting, has solid justification in cross-section (Luo and Spindler (2017)) but its theoretical properties in time series settings are unknown. Bai and Ng (2009) also propose selecting the instruments with highest t-statistics in one regressor at a time regressions.

Thirdly, there is also a very broad and mature literature on forecasting, which is hard to summarize in one short article. Methods proposed in that literature can certainly be used on the first subsample to form the efficient instrument. Here we refer to a paper by Stock and Watson (2012) making a comprehensive comparison of different methods for forecasting using many predictors, including Bayesian model averaging, empirical Bayes and bagging.

5.4 What about other approaches in Time Series?

In the cross-sectional setting there are several additional methods available to correct for estimated instrument endogeneity. These include cross-fit, jackknife, and deleted diagonal approaches. In their simplest forms none of these methods seem to be valid for time series. Both problems with time series data, autocorrelated errors and weakly exogenous instruments, play a role here.

The problem of autocorrelated errors may be relatively easily resolved by small modifications of the procedures when there is m-dependence (that is, errors at least m periods apart are independent) or when autocorrelation decays quickly. For example, the deleted diagonal method may be modified by deleting m-diagonals, that is, elements P_{ij} with $|i-j| \leq m$. The jackknife may be modified by excluding from the first-step estimation for t not just its own observation but also m observations before and m observations after.

Solving the issue of weak exogeneity is much more challenging. The cross-sectional deleted diagonal approach treats the projection matrix P (which is a transformation of all instruments for all observations) as fixed, which is possible due to the strict exogeneity typical for cross-sectional data. In time series, the conditioning on all instruments is invalid as endogenous variables (regressors or the dependent variable) are correlated with future values of the instruments or (often) become future instruments themselves. As such, the projection matrix P is not just random but is correlated with the endogenous variables. Similarly, the standard jackknife in time series uses in the first step future observations of the instruments that are correlated with (or even include) endogenous variables, making the estimated optimal instrument endogenous.

One potential fix for the jackknife is to estimate the first stage for observation t using only past observations but not the future ones (excluding the m most recent if there is an issue of the autocorrelated errors). That is, to estimate $\widehat{f_t}$ one uses some machine learning technique applied to the sample only including observations with indexes $s < t - m$, and then runs a just identified linear IV similar to (13). In this formulation, $\widehat{f_t}$ is measurable with respect to $\mathcal{Z}_t$ and thus is weakly exogenous. The exact equivalent of Theorem 3 holds for this modified jackknife. However, checking condition (14) may be more challenging. Specifically, while the stable martingale central limit theorem is still a powerful tool, establishing sufficient conditions for a non-ergodic law of large numbers (16) is more complicated. In the modified jackknife case we have $\widehat{f_t} = \widehat{\pi}_t' Z_t$ not $\widehat{f_t} = \widehat{\pi}' Z_t$ as in the split-sample case. The first-step estimates $\widehat{\pi}_t$ are slowly changing and highly correlated, but not the same.

References

Anatolyev, S. (2003): "The Form of the Optimal Nonlinear Instrument for Multiperiod Conditional Moment Restrictions," *Econometric Theory*, 19(4), pp. 602-609.

Anatolyev, S. (2007): "Optimal Instruments in Time Series: A Survey." *Journal of Economic Surveys*, 21, pp. 143-173.

Anatolyev, S. (2019): "Many Instruments and/or Regressors: A Friendly Guide," *Journal of Economic Surveys*, 33 (2), pp. 689-726.

Anatolyev, S. and Mikusheva, A. (2020): "Factor models with many assets: strong factors, weak factors, and the two-pass procedure," *Journal of Econometrics*, forthcoming, arXiv: 1807.04094

Andrews, I., Stock J.H. and Sun L. (2019): "Weak Instruments in IV Regression: Theory and Practice," *Annual Review of Economics*, 11, pp. 727 - 753.

Angrist, J.D., and Frandsen B. (2020): "Machine Labor," *NBER working paper* 26584.

Angrist, J.D., Imbens G.W. and Krueger A.B. (1999): "Jackknife Instrumental Variables

Estimation," *Journal of Applied Econometrics,* 14, pp. 57 – 67.

Angrist, J.D., and Krueger, A.B. (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, pp. 979 - 1014.

Angrist, J.D. and Krueger, A.B. (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics,* 13, pp. 225–235.

Ascari G. , Magnusson L.M. and S. Mavroeidis (2020): "Empirical evidence on the Euler equation for consumption in the US" *Journal of Monetary Economics*, forthcoming

Ash, E., D. Chen, X. Zhang, Z. Huang and R. Wang (2018): "Deep IV in Law: Analysis of Appellate Impacts on Sentencing Using High-Dimensional Instrumental Variables" mimeo.

Babii A., E. Ghysels and J. Striaukas (2020): "Estimation and HAC-based Inference for Machine Learning Time Series Regressions," UNC Working Paper

Bai, J., (2003): "Inferential theory for factor models of large dimensions," *Econometrica* 71, pp. 135-173.

Bai, J. and Ng, S. (2002): "Determining the number of factors in approximate factor models," *Econometrica*, 70, pp. 191-221.

Bai, J. and Ng, S. (2009): "Selecting Instrumental Variables in a Data Rich Environment," Journal of Time Series Econometrics, 1 (1), pp. 1-34.

Bai, J. and Ng, S. (2010): "Instrumental Variables Estimation in a Data Rich Environment," *Econometric Theory,* 26, 2010, pp. 1577 - 1606.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica,* 62, pp. 657–681.

Belloni, A., D. Chen, V. Chernozhukov and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an application to Eminent Domain, " *Econometrica*, Vol. 80, pp. 2369-2429.

Belloni A. and Chernozhukov V. (2011): "High Dimensional Sparse Econometric Models: An Introduction." In: Alquier P., Gautier E., Stoltz G. (eds) Inverse Problems and High-Dimensional Estimation. Lecture Notes in Statistics, vol 203. Springer, Berlin, Heidelberg

Belloni, A., V. Chernozhukov and C. Hansen (2010): "Inference Methods for High-Dimensional Sparse Econometric Models," Advances in Economics and Econometrics, ES World Congress

Bernanke, B., Boivin, J., and Eliasz, P.S. (2005): “Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach,” *Quarterly Journal of Economics* 120, pp. 387-422.

Carrasco, M. (2012): “A Regularization Approach to the Many Instruments Problem,” *Journal of Econometrics*, 170 (2), pp. 383-398.

Carrasco, M., and B. Rossi (2016): “In-sample inference and forecasting in misspecified factor models,” *Journal of Business and Economic Statistics*, 34(3), pp. 313-338.

Chamberlain, G. (1987): “Asymptotic Efficiency in Estimation With Conditional Moment Restrictions,” *Journal of Econometrics*, 34, pp. 305-334.

Chao, J.C. and Swanson, N.R. (2005): “Consistent Estimation with a Large Number of Weak Instruments,” *Econometrica*, 73 (5), pp. 1673-1692.

Chao, J.C., Swanson, N.R., Hausman, J.A., Newey, W.K., and Woutersen, T. (2012): “Asymptotic Distribution of JIV in a heteroscedastic IV Regression with Many Instruments,” *Econometric Theory* 28, pp. 42- 86.

Clarida, R., Galı, J. and Gertler, M. (1998): “Monetary policy rules in practice: some international evidence,” *European Economic Review*, Vol. 42, pp. 1033–1067.

Davidson, R. and MacKinnon, J. (2014): “Confidence sets based on inverting Anderson–Rubin tests,” *Econometrics Journal*, 17, pp. S39-S58.

Donald, S. G., and W. K. Newey (2001): “Choosing the Number of Instruments,” *Econometrica,* 69 (5), pp. 1161-1191.

Dovi, M. (2020):“Robust Inference with High-Dimensional Instrumental Variables and the New Keynesian Phillips Curve,” *unpublished manuscript.*

Fama, E.F. and MacBeth, J. (1973): “ Risk, Return and Equilibrium: Empirical Tests,” *Journal of Political Economy*, 81, pp. 607-636.

Fuhrer, J. C., Rudebusch, G. D., (2004): “Estimating the Euler Equation for output,” *Journal of Monetary Economics,* 51, pp. 1133-1153.

Galí, J., and M. Gertler (1999): “Inflation Dynamics: A Structural Econometric Analysis,” *Journal of Monetary Economics,* 44 (2), pp. 195–222.

Gilchrist, D.S., and E.G. Sands (2016): “Something to Talk About: Social Spillover in Movie Consumption,” *Journal of Political Economy*, 124(5), pp. 1339-1382.

Hansen, B. and West, K. D. (2002): "Generalized method of moments and macroeconomics." *Journal of Business and Economic Statistics*, 20, pp. 460–469.

Hansen C., J. Hausman and W. Newey (2008): "Estimation with Many Instruments," *Journal of Business & Economic Statistics*, 26, pp. 398-422.

Hansen C. and D. Kozbur (2014):"Instrumental variables estimation with many weak instruments using regularized JIVE," *Journal of Econometrics*, 182, pp. 290–308.

Hansen, L. P. (1985): "A method for calculating bounds on the asymptotic variance-covariance matrices of generalized method of moments estimators." *Journal of Econometrics*, 30, pp. 203–228.

Hansen, L.P., J. C. Heaton and M. Ogaki (1988): "Efficiency Bounds Implied by Multiperiod Conditional Moment Restrictions," *Journal of the American Statistical Association*, 83, pp. 863-871

Hansen, L. P., Singleton, K. J., (1982): " Generalized instrumental variables estimation of nonlinear rational expectations models," *Econometrica*, 50 (5), pp. 1269-1286.

Häusler, E. and Luschgy, H. (2015): *Stable Convergence and Stable Limit Theorems.* Springler. Probability Theory and Stochastic Modelling.

Hausman J., W. Newey, T. Woutersen, J. Chao and N. Swanson (2012): "Instrumental variable estimation with heteroskedasticity and many instruments" *Quantitative Economics*, 3 , pp. 211–255.

Hayashi, F., and Sims, C. (1983): "Nearly Efficient Estimation of Time Series Models with Predetermined, but not Exogenous, Instruments." *Econometrica*, 51(3), pp. 783-798.

Kapetanios G., L. Khalaf and M. Marcellino (2016): "Factor-based Identification-robust Inference in IV regressions" *J. Appl. Econ.* 31, pp. 821–842.

Kapetanios G. and M. Marcellino (2010): "Factor-GMM estimation with large sets of possibly weak instruments," *Computational Statistics and Data Analysis*, 54, pp. 2655-2675.

Kleibergen F. (2002): "Pivotal statistics for testing structural parameters in instrumental variables regression," *Econometrica*, 70, pp. 1781–1803.

Kleibergen F. and S. Mavroeidis (2009): "Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve," *Journal of Business & Economic Statistics,* Vol. 27, No. 3, pp. 293-339.

Luo, Ye, and Martin Spindler (2017): "L2-Boosting for Economic Applications." *American Economic Review*, 107 (5), pp. 270-73.

Maestas, N., Mullen, K.J., and Strand, A. (2013): "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *American Economic Review* 103, pp. 1797–1829.

Mavroeidis, S. (2004): "Weak Identification of Forward-looking Models in Monetary Economics," Oxford Bulletin of Economics and Statistics, 66, pp. 609-635.

Mavroeidis, S., M. Plagborg-Møller, and J. H. Stock (2014): "Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve," *Journal of Economic Literature,* , 52(1), pp. 124–188.

McCracken, M., and S. Ng (2020): "FRED-QD: A Quaterly Database for Macroeconomic Research," mimeo.

Mikusheva, A. (2010): "Robust Confidence Sets in the Presence of Weak Instruments," *Journal of Econometrics*, 157, pp. 236-247.

Mikusheva, A. and Sun, L. (2020):"Inference with Many Weak Instruments,"*working paper*

Moreira, M. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71(4), pp. 1027-1048.

Newey, W K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica,* 58, pp. 809-837.

Newey, W. and Smith, R. (2004): "Higher Order Properties of GMM and Generalized Empirical Likeliood Estimators," *Econometrica*, 71, pp. 219-255.

Newey W. and F. Windmeijer (2009) "Generalized Method of Moments with Many Weak Moment conditions" *Econometrica,* 77 (3), pp. 687–719.

Okui, R. (2011): "Instrumental Variable Estimation in the Presence of Many Moment Conditions," *Journal of Econometrics*, 165, pp. 70-86.

Shanken, J. (1992): " On the Estimation of Beta-Pricing Models," *Review of Financial Studies*, 5, pp. 1-33.

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

1167- 1179.

Stock, J.H., Watson, M.W. (2002b): " Macroeconomic forecasting using diffusion indices," *Journal of Business and Economic Statistics* 20, pp. 147-162.

Stock, J.H., Watson, M.W. (2005): "Implications of Dynamic Factor Models for VAR Analysis," Mimeo.

Stock J.H., and Watson M.W. (2012): "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business & Economic Statistics*, 30 (4), pp. 481-493.

Stock, J.H., J.H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic statistics*, 20(4), pp. 518-529.

Stock, J.H., and Yogo, M. (2005): "Testing for weak instruments in Linear Iv regression." In Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg, pp. 80–108.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via Lasso," *Journal of Statistical Society, Ser. B*, 58, pp. 267-288.

Wang, R. and Shao, X. (2020): ''Hypothesis Testing for High-Dimensional Time Series Via Self-Normalisation," *Annals of Statistics* 2020, 48(5), pp. 2728 - 2758.

Yogo, M. (2004): "Estimating the elasticity of intertemporal substitution when instruments are weak," *Review of Economics and Statistics*, 86 (3), pp. 797-810.

6 Appendix

Proof of Theorem 1. We prove the consistency of the sample split estimator, the proof for the cross-fit is similar.

$$\widehat{\beta}_{SS} - \beta = \frac{\sum_{i\in I_2} f_i e_i + \sum_{i\in I_2}(\widehat{f_i} - f_i)e_i}{\sum_{i\in I_2} \widehat{f_i} f_i + \sum_{i\in I_2} f_i v_i + \sum_{i\in I_2}(\widehat{f_i} - f_i)v_i}$$

We notice that $c\mathbb{E}[f_i^2] \leq \mathbb{E}[f_i^2 v_i^2] \leq C\mathbb{E}[f_i^2]$, thus $\mu^2 \asymp n\mathbb{E}[f_i^2]$. By the Law of Large Numbers

$$\frac{1}{|I_2|}\sum_{i\in I_2} \widehat{f_i} f_i \to^p \mathbb{E}[\widehat{f_i} f_i] \geq c\mathbb{E}[f_i^2].$$

Thus $\frac{1}{\mu^2}\sum_{i\in I_2} \widehat{f_i} f_i$ is asymptotically separated from zero by a constant. Calculating the second moments gives that $\sum_{i\in I_2} f_i v_i = O_p(\sqrt{\mu^2})$ and $\sum_{i\in I_2} f_i e_i = O_p(\sqrt{\mu^2})$. Denote

$\mathcal{Z}_2 = \{Z_j\}_{j \in I_2}$

$$\mathbb{E}\left[\left(\sum_{i\in I_2}(\widehat{f}_i - f_i)e_i\right)^2\right] = \mathbb{E}\mathbb{E}\left[\left(\sum_{i\in I_2}(\widehat{f}(\mathcal{A}_1, Z_i) - f_i)e_i\right)^2 |\mathcal{A}_1, \mathcal{Z}_2\right] =$$

$$= \mathbb{E}\left[\sum_{i\in I_2}(\widehat{f}(\mathcal{A}_1, Z_i) - f_i)^2 \mathbb{E}\left[e_i^2|Z_i\right]\right] \leq Cn\mathbb{E}\left[(\widehat{f}_i - f_i)^2\right] \leq Cr_n.$$

Here we used that conditional on $\mathcal{A}_1, \mathcal{Z}_2$ the errors e_i are uncorrelated and mean zero, since they are independent from $\mathcal{A}_2$ and the instruments $\mathcal{Z}_1$ are exogenous. Thus, $\sum_{i\in I_2}(\widehat{f}_i - f_i)e_i = O_p(\sqrt{r_n})$. We treat $\sum_{i\in I_2}(\widehat{f}_i - f_i)v_i$ in the same way. Finally, Assumption $\frac{\mu^2}{\sqrt{r_n}} \to \infty$ guarantees that $\widehat{f}_i f_i$ asymptotically dominates all other terms, and thus $\widehat{\beta}_{SS} - \beta \to^p 0$.

Proof of Theorem 2

$$\widehat{\beta}_{SS} - \beta = \frac{\sum_{i\in I_2} \widehat{f}_i e_i}{\sum_{i\in I_2} \widehat{f}_i f_i + \sum_{i\in I_2} \widehat{f}_i v_i}. \tag{17}$$

First, we prove that the assumptions of Theorem 2 guarantee that

$$\Gamma_n^{-1/2} \sum_{i\in I_2} \widehat{f}_i \varepsilon_i \Rightarrow N(0, I_2), \quad \mathcal{A}_1 - \text{ stably}, \tag{18}$$

with $\Gamma_n = \sum_{i\in I_2} \mathbb{E}\left[\varepsilon_i \varepsilon_i' \widehat{f}_i^2 |\mathcal{A}_1\right]$. Define $\mathcal{F}_i$ to be the sigma-algebra of all random variables with indexes not exceeding i. Then $\mathcal{A}_1$ is the sigma-algebra in intersection of $\mathcal{F}_i, i \in I_2$. Define $\xi_i = \frac{1}{\sqrt{\mathbb{E}\left[\widehat{f}_i^2|\mathcal{A}_1\right]}} \widehat{f}_i \varepsilon_i = w_i \varepsilon_i$, where w_i is measurable with respect to sigma algebra generated by $\mathcal{A}_1$ and Z_i. Then $\{(\xi_i, \mathcal{F}_i), i \in I_2\}$ is a martingale difference sequence. Indeed,

$$\mathbb{E}[\xi_i|\mathcal{F}_{i-1}] = \mathbb{E}[w_i\varepsilon_i|\mathcal{F}_{i-1}] = \mathbb{E}[w_i\varepsilon_i|\mathcal{A}_1] = \mathbb{E}[w_i \mathbb{E}[\varepsilon_i|\mathcal{A}_1, Z_i]|\mathcal{A}_1] = 0.$$

Here we used that due to i.i.d. nature ξ_i is independent from observations with index $j \in I_2$ such that $j \neq i$. Moment assumptions guarantee that $\mathbb{E}[\xi_i\xi_i'|\mathcal{A}_1] = \frac{\Gamma_n}{|I_2|\mathbb{E}\left[\widehat{f}_i^2|\mathcal{A}_1\right]}$ has bounded eigenvalues. The Law of Large numbers shows that $\frac{1}{|I_2|}\sum_{i\in I_2} \xi_i \xi_i' \to^p \mathbb{E}[\xi_i\xi_i'|\mathcal{A}_1]$. Assumption (i) is Lyapunov condition. Martingale Central Limit Theorem (see, Hausler and Luschgy (2015), Theorem 6.1 and condition (CLY$_p$)) gives

$$\frac{1}{\sqrt{|I_2|}} \left(\mathbb{E}[\xi_i\xi_i'|\mathcal{A}_1]\right)^{-1/2} \sum_{i\in I_2} \xi_i \Rightarrow N(0, I_2), \quad \mathcal{A}_1 - \text{ stably}.$$

The last implies (18). Since the eigenvalues of Γ_n is bounded by $|I_2|\mathbb{E}\left[\widehat{f}_i^2|\mathcal{A}_1\right]$, we have that $\frac{1}{\sqrt{|I_2|\mathbb{E}\left[\widehat{f}_i^2|\mathcal{A}_1\right]}}\sum_{i\in I_2}\widehat{f}_i v_i = O_p(1)$. Thus, Assumptions (ii) and (iii) guarantee that the denominator of (17) is dominated by the first sum only and the second sum can be neglected.

Denote $\gamma_n = \sum_{i\in I_2}\mathbb{E}\left[e_i^2\widehat{f}_i^2|\mathcal{A}_1\right]$ to be the upper-left element of Γ_n. Then

$$\sqrt{\frac{a_n^2}{\gamma_n}}\left(\widehat{\beta}_{SS}-\beta\right) = \frac{\frac{1}{\sqrt{\gamma_n}}\sum_{i\in I_2}\widehat{f}_i e_i}{\frac{1}{a_n}\sum_{i\in I_2}\widehat{f}_i f_i}(1+o_p(1)) \Rightarrow N(0,1) \quad \mathcal{A}_1 - \text{ stably.}$$

We already proved above that $\frac{1}{a_n}\sum_{i\in I_2}\widehat{f}_i X_i \to^p 1$, thus for the last statement of Theorem 2 we are left to prove that $\frac{1}{\gamma_n}\sum_{i\in I_2}\widehat{f}_i^2\widehat{e}_i^2 \to^p 1$.

Let us first show that $\frac{1}{a_n^2}\sum_{i\in I_2}\widehat{f}_i^2 X_i^2 \to^p 0$. Indeed,

$$\frac{1}{a_n^2}\sum_{i\in I_2}\mathbb{E}[\widehat{f}_i^2 X_i^2|\mathcal{A}_1] \leq \frac{1}{a_n^2}\sum_{i\in I_2}\sqrt{\mathbb{E}[\widehat{f}_i^4|\mathcal{A}_1]}\sqrt{\mathbb{E}[X_i^4]} \leq \frac{C}{a_n^2}\sum_{i\in I_2}\mathbb{E}[\widehat{f}_i^2|\mathcal{A}_1] \to^p 0.$$

Here we used Cauchy-Schwarz inequality, moment conditions and assumptions (i) and (iii). Similarly,

$$\frac{1}{a_n\sqrt{\gamma_n}}\sum_{i\in I_2}\mathbb{E}[\widehat{f}_i^2|e_iX_i||\mathcal{A}_1] \leq \frac{1}{\sqrt{\gamma_n}}\sqrt{\sum_{i\in I_2}\mathbb{E}[\widehat{f}_i^2 e_i^2|\mathcal{A}_1]}\sqrt{\frac{1}{a_n^2}\sum_{i\in I_2}\mathbb{E}[\widehat{f}_i^2 X_i^2|\mathcal{A}_1]} \to^p 0.$$

Now, we have

$$\sum_{i\in I_2}\widehat{f}_i^2\widehat{e}_i^2 = \sum_{i\in I_2}\widehat{f}_i^2 e_i^2 + 2(\widehat{\beta}_{SS}-\beta)\sum_{i\in I_2}\widehat{f}_i^2 e_i X_i + (\widehat{\beta}_{SS}-\beta)^2\sum_{i\in I_2}\widehat{f}_i^2 X_i^2.$$

Thus,

$$\frac{1}{\gamma_n}\sum_{i\in I_2}\widehat{f}_i^2\widehat{e}_i^2 = 1 + 2\sqrt{\frac{a_n^2}{\gamma_n}}(\widehat{\beta}_{SS}-\beta)\frac{1}{a_n\sqrt{\gamma_n}}\sum_{i\in I_2}\widehat{f}_i^2 e_i X_i + \frac{a_n^2}{\gamma_n}(\widehat{\beta}_{SS}-\beta)^2\frac{1}{a_n^2}\sum_{i\in I_2}\widehat{f}_i^2 X_i^2.$$

Putting all proven statements together we get that $\frac{1}{\gamma_n}\sum_{i\in I_2}\widehat{f}_i^2\widehat{e}_i^2 \to^p 1$.