NBER WORKING PAPER SERIES

THE INCREDIBLE SHRINKING INSTRUMENTS: USING EMPIRICAL BAYES TO INCREASE EFFICIENCY IN IV DESIGNS WITH MANY INSTRUMENTS

Brigham R. Frandsen
Lars J. Lefgren
Emily C. Leslie
Samuel P. McIntyre

Working Paper 34634
http://www.nber.org/papers/w34634

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2026

We thank Ammon Larsen for excellent research assistance. We are grateful to participants at the California Econometrics Conference for useful feedback. We acknowledge generous research support from the College of Family, Home, and Social Sciences of Brigham Young University. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Incredible Shrinking Instruments: Using Empirical Bayes to Increase Efficiency in IV Designs with Many Instruments

Brigham R. Frandsen, Lars J. Lefgren, Emily C. Leslie, and Samuel P. McIntyre

NBER Working Paper No. 34634
January 2026
JEL No. C1, C11, C21, K14

ABSTRACT

In instrumental variables (IV) estimation with many instruments, such as judge fixed effects designs, the precision of the jackknife-estimated first stage can vary widely across observations. When such variability exists, we show that the precision of JIVE second-stage estimates is meaningfully improved by shrinking judge propensities towards their conditional means, where the shrinkage factor depends on the precision of the first-stage fitted value. Doing so requires no further assumptions and identifies the same local average treatment effect as the usual (unshrunken) JIVE estimator. We illustrate the precision gains from using a Shrunken JIVE estimator (SJIVE) in an application from the literature studying pre-trial detention of defendants in criminal cases.

Brigham R. Frandsen
Brigham Young University
Department of Economics
frandsen@byu.edu

Emily C. Leslie
Brigham Young University
Department of Economics
emily.leslie@byu.edu

Lars J. Lefgren
Brigham Young University
Department of Economics
and NBER
lars_lefgren@byu.edu

Samuel P. McIntyre
samcintyre@ucsd.edu

1 Introduction

Many-instruments bias poses a challenge for estimation and inference in instrumental variables (IV) designs. Many recent applied studies have turned to some variation of Angrist et al.'s (1999) jackknife instrumental variables estimator (JIVE) as a practical solution to this problem. By constructing first-stage predictions that exclude an observation's own treatment realization, JIVE eliminates the contribution of an observation's own treatment realization to its predicted treatment status, overcoming the many-instruments bias that plagues two-stage least squares. JIVE is consistent even when the number of instruments is large, performs better than approaches that rely on instrument selection in many settings (e.g. Donald and Newey, 2001; Bai and Ng, 2010; Belloni et al., 2012; Chernozhukov et al., 2018), and can outperform LIML when LIML's assumptions are violated (Angrist and Frandsen, 2022). Despite these advantages, however, the statistical precision of JIVE may be limited when instruments are weak and many relative to the number of observations.

This paper introduces a shrunken jackknife instrumental variables estimator (SJIVE) that improves the precision of JIVE while preserving its consistency.[1] SJIVE applies empirical Bayes–inspired shrinkage to JIVE's leave-one-out first-stage predictions, reducing their variability without reintroducing bias. We derive the optimal shrinkage, characterize the limiting distribution of the resulting estimator, and show that SJIVE is asymptotically more efficient than unshrunken JIVE. In simulations, SJIVE not only eliminates many-instruments bias, like JIVE, but also delivers substantial improvements in precision. SJIVE also outperforms LIML, even under homoskedastic and normally distributed data generating processes.[2] Importantly, when shrinkage of-

[1] SJIVE is available as a Stata package, available from SSC via `ssc install sjive`.

[2] This surprising finding does not contradict LIML's optimality properties in Chioda and Jansson (2009), as SJIVE lies outside the rotation-invariant class considered there.

fers no benefit, SJIVE is asymptotically equivalent to JIVE, ensuring that the method is generally harmless when the additional structure is uninformative. Furthermore, one need make only the usual IV assumptions in order for SJIVE to deliver a consistent estimate of a proper weighted average of treatment effects.

While SJIVE applies generally, including in the case of many continuous instruments, we motivate it in the framework of the judge fixed effects design, one of the most common applications of a many-instruments framework. Since Kling's (2006) pioneering use of judge assignment as an instrument for incarceration length, at least 136 studies have employed similar designs to study topics ranging from criminal justice to foster care and disability (Chyn et al., 2024). These designs instrument treatment using a mutually exclusive set of judge indicators, or interactions of judge indicators with covariates such as subject characteristics. Because the number of cases per judge is often limited, first-stage estimates of judge propensities can be noisy, and the resulting treatment effect estimates imprecise.

Researchers have adopted two common strategies to address these limitations, each with important drawbacks that SJIVE helps to overcome. First, researchers frequently drop judges whose caseloads fall below an arbitrary threshold. For example, Aizer and Doyle (2015) include only judges with at least 10 cases, while Leslie and Pope (2017) require a minimum of 500 cases. Because propensities for judges with few cases are estimated imprecisely, including such judges when using JIVE can attenuate the first-stage relationship between estimated judge propensities and subject treatment status, sometimes resulting in less precise treatment effect estimates than excluding them entirely. Yet this approach is imperfect: even judges with only a few cases contain useful information for identification and should, in principle, contribute to the analysis. SJIVE offers a principled alternative by shrinking each judge's propensity according to the precision with which it is estimated. As a result,

all judges—including those with very small caseloads—contribute to identification in a way that reflects their information content. When additional judge-level covariates predict treatment propensities, SJIVE can shrink toward the conditional mean, pooling information across similar judges and further increasing first-stage power and second-stage precision.

Second, researchers often construct separate judge propensities by time period or subject characteristics to capture heterogeneity in judge behavior. For example, Arnold et al. (2018) estimates year-specific leniency measures, while Stevenson (2018) constructs distinct leniency measures based on crime type, criminal history, demographics, and time period. These approaches implicitly interact judge fixed effects with covariates, leveraging temporal or contextual variation in judge behavior and allowing for rich non-monotonicities in treatment assignment. While such interacted instruments may strengthen the first stage, each subgroup propensity is estimated using fewer observations and is therefore noisier than the corresponding judge-level propensity. Researchers can use SJIVE to retain the information contained in the interacted instruments while leveraging the greater precision of the judge-level aggregates by shrinking the subgroup-specific propensities toward each judge's overall propensity.

In many judge fixed effects applications, there is substantial variability in the number of cases per judge and information on subject, case, or judge characteristics. Consequently, SJIVE is likely to be of broad utility in such research designs. In simulations that mirror these empirical conditions, SJIVE meaningfully improves upon JIVE. Precision gains are largest when many judges have small caseloads and a few judges have large caseloads, and when judge propensities are imprecisely estimated and observable characteristics are predictive of latent propensities.

To illustrate the method's practical value, we apply it in the empirical setting of

Stevenson (2018), which studies the effect of pretrial detention on subsequent conviction outcomes. Defendants in this setting are randomly assigned to one of eight judges, and judges exhibit different propensities to detain defendants based on observed defendant characteristics. Interacting judge indicators with these characteristics generates a large set of instruments, implicating the usual challenges associated with many-instruments. We compare estimates from JIVE and SJIVE using both simple judge indicators and a high-dimensional set of interacted instruments. Consistent with the theoretical and simulation results, the gains from SJIVE are particularly pronounced when the number of instruments is large relative to the sample size.

2 Econometric Framework

We are interested in the causal effect, δ, of a potentially endogenous regressor D_i (for example, an indicator for pre-trial release) on outcome Y_i (conviction), expressed in the following equation:

$$Y_i = \delta D_i + \varepsilon_i, \tag{1}$$

where ε_i reflects unobserved factors that determine the outcome and satisfies $E[\varepsilon_i] = 0$. For clarity, we initially consider additional covariates (including a constant) to have been partialed out of all variables, but we discuss their inclusion below. If ε_i is correlated with D_i, that is, $E[D_i\varepsilon_i] \neq 0$, then least squares will fail to consistently estimate δ. Identification relies on a $p \times 1$ vector of excluded instruments, Z_i (for example, indicators for the bail judge assigned to each person), that are related to D_i via the following first-stage regression equation:

$$D_i = Z_i'\pi + \nu_i. \tag{2}$$

We take Z_i to be nonstochastic and assume $E[\nu_i] = 0$. The number of instruments p may be large, a possibility we account for by allowing p to grow with the sample size n in our asymptotic approximations. Specifically, we adopt a many-instrument sequence such that $p/n \to \kappa$.

The standard exclusion and relevance conditions for Z_i to serve as valid instruments in our setting are the following:

Condition 1. *(Instrument Validity)*

(a) Exclusion: $\sum_{i=1}^{n} Z_i'\pi\varepsilon_i/n \to_p 0$

(b) Relevance: $\sum_{i=1}^{n} \pi' Z_i Z_i'\pi/n \to \sigma_P^2$ *for some* $\sigma_P^2 > 0$.

Under textbook conditions the ideal IV estimator uses $P_i := Z_i'\pi$ as an instrumental variable:

$$\hat{\delta}^{oracle} = \frac{\sum_{i=1}^{n} P_i Y_i}{\sum_{i=1}^{n} P_i D_i}.$$

We term this the oracle estimator because it requires knowledge of the population first-stage parameters π. In a judge fixed effects setting, π would reflect each judge's true propensity to assign treatment. Under standard conditions, $\hat{\delta}^{oracle}$ is consistent and asymptotically normal, with a finite-sample variance that can be approximated by:

$$V^{oracle} = \frac{1}{n}\frac{\sigma_\varepsilon^2}{\sigma_P^2},$$

where σ_ε^2 is the variance of ε_i under homoskedasticity.

In practice the oracle estimator is infeasible because π is unknown. The conventional alternative is two-stage least squares (2SLS), which replaces P_i by the least-squares projection of D_i on Z_i (e.g., the in-sample fraction of people treated by the

judge assigned to person i), $\hat{P}_i := Z_i' \left(\sum_{j=1}^n Z_j Z_j'\right)^{-1} \sum_{j=1}^n Z_j D_j$:

$$\hat{\delta}^{2SLS} = \frac{\sum_{i=1}^n \hat{P}_i Y_i}{\sum_{i=1}^n \hat{P}_i D_i}.$$

2SLS suffers from the well-known many-instruments bias and is inconsistent when the number of instruments grows with the sample size:

$$\hat{\delta}^{2SLS} \to_p \delta + \frac{\kappa \sigma_{\varepsilon\nu}}{\sigma_p^2 + \kappa \sigma_\nu^2},$$

where σ_ν^2 is the variance of the first-stage error term ν_i and $\sigma_{\varepsilon\nu}$ is the covariance between the ε_i and ν_i.[3]

Jackknife instrumental variables (JIVE) solves the many-instruments bias of 2SLS by replacing $\hat{P}_i$ with a jackknifed fitted value from a regression that leaves out observation i (e.g., the in-sample fraction of people treated by person i's judge, excluding person i), $\hat{P}_i^{JK} := Z_i' \left(\sum_{i' \neq i} Z_{i'} Z_{i'}'\right)^{-1} \sum_{i' \neq i}^n Z_{i'} D_{i'}$:

$$\hat{\delta}^{JIVE} = \frac{\sum_{i=1}^n \hat{P}_i^{JK} Y_i}{\sum_{i=1}^n \hat{P}_i^{JK} D_i}.$$

Under standard conditions JIVE is consistent and asymptotically normal, with a finite-sample variance that can be approximated by

$$V^{JIVE} = c^{JIVE} V^{oracle}, \quad c^{JIVE} \geq 1,$$

where c^{JIVE} is defined in the appendix and depends on the instrument matrix Z, σ_ε^2, σ_ν^2, and $\sigma_{\varepsilon\nu}$. Although JIVE is consistent even under many-instrument asymptotics,

[3]The expression for many-instruments bias here, derived in the appendix, makes the simplifying assumption of homoskedasticity and independence across observations.

it can be substantially noisier than the oracle estimator and 2SLS.

JIVE's imprecision stems from the noise with which $\hat{P}_i^{JK}$ predicts the ideal instrument P_i. Unlike in the case of 2SLS, the prediction noise is uncorrelated with ε_i, so consistency is uncompromised, but the noise does inflate the variance of the IV estimator. Furthermore, the noise in $\hat{P}_i^{JK}$ has a variance that differs across observations i, and depends on the instrument design matrix, Z. For example, suppose the columns of Z are a set of mutually exclusive group indicators, as in a judge fixed effects setting. Then $\hat{P}_i^{JK}$ will be noisier for observations i that belong to smaller groups, e.g., people assigned to judges with smaller caseloads. Consequently, the precision of JIVE can be improved by shrinking $\hat{P}_i^{JK}$ by a factor $\lambda_i \leq 1$ that depends on the observation-specific variance of the noise in $\hat{P}_i^{JK}$.

Our proposed IV estimator, SJIVE, improves the precision of JIVE by shrinking $\hat{P}_i^{JK}$ toward a regression-adjusted mean, $S_i := E[D_i|W_i]$, where $W_i = W(Z_i)$ is a low-dimensional function of the instruments. For example, if Z_i is a set of judge dummies, then W_i may be a vector of characteristics of the judge to which observation i is assigned. As another example, if Z_i contains many interactions, then W_i may contain the underlying main effects: if the instruments are a set of judge-by-year or judge-by-case-type dummies, then W_i could be the vector of judge dummies. To be helpful, the vector W_i should be low-dimensional relative to p (the number of instruments in Z_i) but also be predictive of P_i. W_i itself constitutes a set of valid instruments, but we shall see that it is better used to generate targets for shrinking $\hat{P}_i^{JK}$ than as a set of instruments. If no suitable set of variables W_i is available, then S_i can be set to the grand mean of $\hat{P}_i^{JK}$ (a constant), so that we can define without loss of generality the shrunken instrument $\tilde{P}_i$ as follows:

$$\tilde{P}_i = \lambda_i \hat{P}_i^{JK} + (1 - \lambda_i) S_i,$$

where λ_i depends only on constants and the instrument matrix Z. The corresponding SJIVE estimator is

$$\hat{\delta}^{SJIVE} = \frac{\sum_{i=1}^{n} \tilde{P}_i Y_i}{\sum_{i=1}^{n} \tilde{P}_i D_i}.$$

Under standard conditions, SJIVE is consistent and asymptotically normal, with a finite-sample variance that can be approximated by

$$V^{SJIVE} = c(\lambda) V^{oracle}, \quad c(\lambda) \geq 1.$$

Our main theoretical result is the variance-minimizing choice of shrinkage factor, given below. It depends on how well P_i can be predicted from W_i, and on how much each observation affects both its own fitted value and those of others. Judges with higher caseloads will tend to have smaller optimal shrinkage factors, because each individual case exerts less influence on the fitted values. Specifically, as shown in the appendix, the variance minimizing choice of shrinkage factor is

$$\lambda_i = \frac{(1-\alpha)\sigma_P^2}{(1-\alpha)\sigma_P^2 + \sigma_\nu^2 A_i + \frac{\sigma_{\varepsilon\nu}^2}{\sigma_\varepsilon^2} B_i}, \tag{3}$$

where α is the R^2 from a regression of P_i on W_i, $A_i = \frac{h_i}{1-h_i}$,

$$B_i = \sum_{i' \neq i} \frac{h_{ii'}^2}{(1-h_i)(1-h_{i'})},$$

$h_{ii'}$ is the (i,i')-th element of the projection matrix $H = Z(Z'Z)^{-1}Z'$ and h_i is the i-th diagonal element of H. A corollary is that for this choice of λ,

$$V^{SJIVE} \leq V^{JIVE}.$$

The inequality is weak because there are special cases where shrinking has no benefit (for example, the instruments are dummies for groups all of identical size), but in general SJIVE improves on the precision of JIVE, and the reduction in variance can be substantial.

Practical implementation

Implementation of SJIVE raises two practical considerations: the incorporation of covariates, and construction of the shrinkage parameters λ_i. We take each in turn.

Many IV designs, including our empirical example below, incorporate covariates, either to make the exclusion restriction more plausible, or to improve precision. Further, often the number of covariates is non-negligible relative to the sample size. In this case we suggest the UJIVE method of Kolesár (2013) to account for the covariates in the estimation of the jackknifed propensity, $\hat{P}_i^{JK}$. Let the vector of covariates for observation i be denoted as X_i. The UJIVE procedure first constructs a jackknifed projection of D_i on (Z_i, X_i), denoted $\hat{D}_i^{JK}(Z_i, X_i)$ and a jackknifed projection of D_i on X_i alone, denoted $\hat{D}_i^{JK}(X_i)$. The UJIVE covariate-adjusted instrument is $\hat{P}_i^{UJIVE} = \hat{D}_i^{JK}(Z_i, X_i) - \hat{D}_i^{JK}(X_i)$. In the special case of no covariates, UJIVE simplifies to standard JIVE.

The optimal shrinkage parameters given in equation (3) depend on four unknown parameters: the underlying variance of instrument propensities, σ_P^2; the variance of the first-stage error term, σ_ν^2; the covariance between the first- and second-stage errors, $\sigma_{\varepsilon\nu}$; and the variance of the second-stage error term, σ_ε^2. We estimate σ_P^2 as the covariance between $\hat{P}_i^{JK}$ and D_i. We estimate σ_ν^2 as the degrees-of-freedom adjusted residual variance from the first stage regression of D_i on (Z_i, X_i). We estimate $\sigma_{\varepsilon\nu}$ as the covariance between the estimated residual from UJIVE and the estimated residual

from the first-stage regression of D_i on (Z_i, X_i). Finally, we estimate σ_ε^2 as the degrees-of-freedom adjusted residual variance from two-stage least squares regression of Y_i on D_i and X_i, using $\hat{P}_i^{UJIVE}$ as an excluded instrument for D_i.

3 Power and Finite Sample Performance

3.1 Performance of SJIVE Relative to Alternative Approaches

To compare the finite-sample performance of SJIVE to alternative estimation strategies, we conduct a series of Monte Carlo simulations designed to be comparable to a setting a practitioner using judge fixed effects might encounter. Each simulation proceeds as follows:

We generate 100 instruments (interpretable as judge indicators) indexed by j, where judges differ in their caseload sizes following a geometric distribution that ranges from 2 to over 100 cases. This produces approximately $n = 5,000$ observations per replication.

For each instrument z_j, we draw a variable $W_j \sim \mathcal{N}(0, \alpha\sigma_P^2)$, capturing the component of judge leniency explained by observable judge characteristics, and an orthogonal error term $e_j \sim \mathcal{N}(0, (1-\alpha)\sigma_P^2)$. The true judge-specific treatment propensity is given by $p_j = W_j + e_j$, with $\sigma_P = 0.2$ and $\alpha = 0.5$, so that half of the variation in propensities is predictable by W.

For each individual, the observed treatment variable is

$$D_i = p_{j(i)} + \nu_i,$$

where the correlation between the first-stage error ν_i and the second-stage error ε_i is 0.5. Outcomes are generated as $Y_i = \varepsilon_i$, implying that the true treatment effect is

zero.

We estimate this effect using several methods: OLS, an oracle version of 2SLS that uses the true propensities as instruments, standard 2SLS, LIML, and JIVE. We also include an infeasible version of SJIVE that uses known parameters, a feasible version that estimates them, and a 2SLS specification using only the vector W as the instrument.

Table 1 reports results based on 1,000 simulation replications. The OLS estimator shows the lowest sampling variability but suffers from severe bias due to endogeneity. The oracle estimator—using the true propensities—achieves negligible bias and near-nominal coverage but is infeasible in practice. Conventional 2SLS reduces bias relative to OLS but does not entirely eliminate it, while LIML eliminates nearly all bias at the cost of higher variance. JIVE performs similarly to LIML, showing little bias but somewhat greater dispersion.

Both versions of SJIVE perform well. The infeasible version is essentially unbiased with lower variance than LIML or JIVE, and the feasible version—using estimated shrinkage parameters—performs almost identically, indicating that estimation error in the shrinkage factor has minimal impact. Moreover, SJIVE maintains appropriate coverage and delivers the smallest standard errors among the unbiased estimators.

Finally, the 2SLS specification using only W as an instrument is unbiased but considerably less precise than SJIVE. This comparison highlights that SJIVE effectively combines the information in both Z and W, improving upon JIVE (which ignores W) and upon the W-only IV approach (which ignores Z).

Table 1: Monte Carlo Simulations Comparing IV Estimators

Estimator	Bias	Std. dev.	Avg. SE	Coverage
OLS	0.482	0.012	0.012	0.000
Oracle	−0.002	0.075	0.074	0.953
2SLS	0.174	0.061	0.055	0.151
LIML	−0.001	0.092	0.089	0.943
JIVE	−0.006	0.107	0.097	0.936
SJIVE (infeasible)	−0.003	0.087	0.085	0.950
SJIVE (feasible)	0.000	0.086	0.085	0.945
2SLS (W only)	0.001	0.112	0.109	0.950

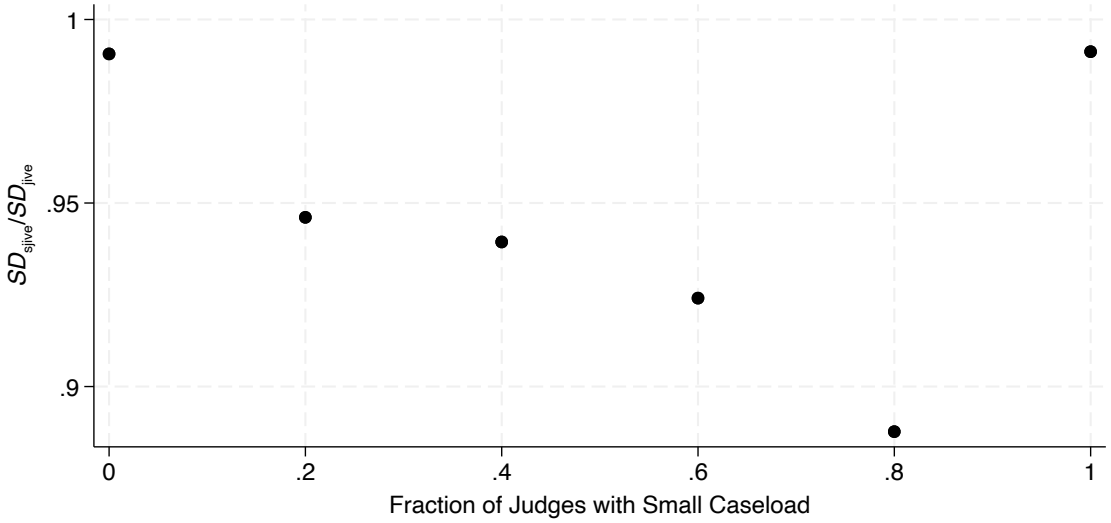
Notes: Based on 1,000 simulations each with sample size $n = 5{,}000$, number of instruments $p = 100$, and ratio of largest to smallest caseload $r = 100$. Standard errors are traditional heteroskedasticity-robust, except for LIML, which uses Bekker standard errors. Coverage is for nominal 95-percent confidence intervals.

3.2 The Relationship between Distribution of Caseloads and SJIVE Performance

One reason that SJIVE outperforms JIVE is its ability to leverage differences in the information content across instruments. In the context of a judge fixed effects research design, this information content is determined by the number of cases assigned to each judge. Variability in caseloads enables differential shrinkage, which increases the efficiency of SJIVE relative to JIVE. To illustrate this mechanism, we conduct a series of simulations using parameter values similar to those described earlier. There are two key differences from the earlier setup. First, because we are now focusing on caseload distributions, we shrink the instruments toward the grand mean rather than the conditional mean. Second, judges are assigned either a high caseload of 100 cases or a small caseload of 10. We vary the fraction of judges with small caseloads while keeping the total number of observations close to 5,000.

Figure 1 presents the results of this simulation. The vertical axis shows the ratio of the standard deviation of SJIVE estimates to that of JIVE estimates. The

Figure 1: SJIVE Relative Performance by Caseload Mix



Notes: Results are from Monte Carlo simulations in which each point represents the ratio of the standard deviation of SJIVE estimates to that of JIVE estimates. The x-axis shows the fraction of judges with small caseloads of 10 cases each while the remaining judges handle caseloads of 100. The standard deviation of latent judge propensities is 0.2, and shrinkage occurs toward the grand mean. Each estimate is based on approximately 5,000 observations. The standard deviations of both the outcome error and the first-stage error are set to 1, and the correlation between treatment and outcome is 0.5.

horizontal axis indicates the fraction of judges with small caseloads. Each point in the figure reflects results from 500 simulations. Because SJIVE exploits variation in instrument strength, it offers little to no advantage when the fraction of judges with small caseloads is either zero or one. The gains from differential shrinkage arise when there is heterogeneity in caseloads and are most pronounced when many judges have small caseloads while a few have large ones. In particular, when 80 percent of judges have small caseloads, the standard deviation of SJIVE estimates is approximately 11 percent lower than that of JIVE.

3.3 The Relationship between Informativeness of Judge Characteristics and SJIVE Performance

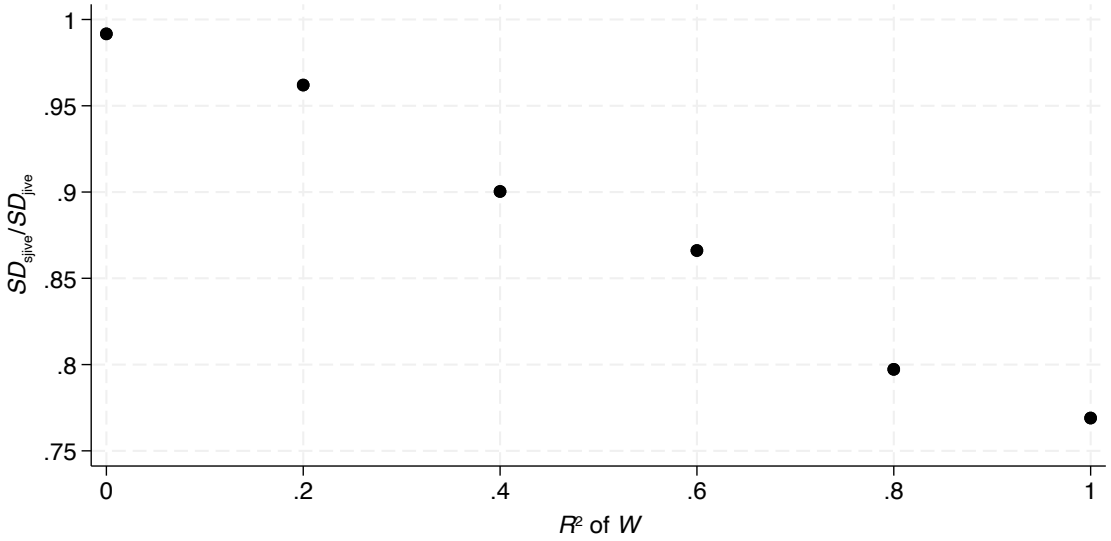
A second source of SJIVE's efficiency relative to JIVE is its use of shrinkage toward a conditional mean. In a typical judge fixed effects design, judge characteristics may be informative about judge propensities. In this next set of simulations, we examine the relative performance of SJIVE and JIVE as the R^2 from predicting judge propensities using a vector of covariates W increases. To isolate the role of conditional shrinkage and avoid confounding it with the differential information content across instruments, we fix each judge's caseload at 50. All other simulation parameters match those described earlier in the section.

Figure 2 shows that, as expected, SJIVE performs increasingly well as the informativeness of W improves. When the R^2 of W is zero, SJIVE offers essentially no advantage over JIVE. As the R^2 approaches one, the standard deviation of SJIVE estimates falls to approximately 78 percent of that of JIVE, illustrating the gains from effectively incorporating covariate information into the shrinkage process.

3.4 The Relationship between the Strength of Identification and SJIVE Performance

The benefits of SJIVE depend on the strength of identification. When instruments are strong, their information content is high relative to the amount of noise, making the advantages of shrinkage comparatively small. To explore this, we examine the performance of SJIVE relative to JIVE as we increase the standard deviation of latent judge propensities. A higher standard deviation is associated with stronger identification. All simulation parameters match those described earlier in the section, except that half of the judges are assigned 10 cases and the other half are assigned

Figure 2: SJIVE Relative Performance by R^2 of W



Notes: Results are from Monte Carlo simulations in which each point represents the ratio of the standard deviation of SJIVE estimates to that of JIVE estimates. The x-axis shows the R^2 of W. Judge caseload is 100. The standard deviation of latent judge propensities is 0.2, and shrinkage occurs toward the grand mean. Each estimate is based on 5,000 observations. The standard deviations of both the outcome error and the first-stage error are set to 1, and the correlation between treatment and outcome is 0.5.

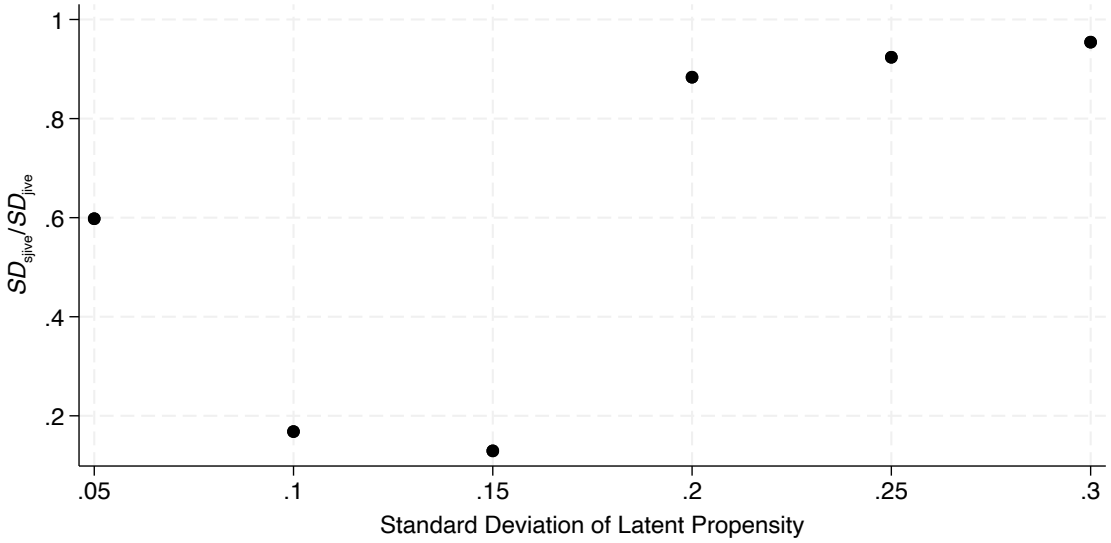
100.

Figure 3 shows that SJIVE consistently outperforms JIVE, but the relative benefit is non-monotonic with respect to the strength of identification. As we move from weak to moderate identification, the standard deviation of SJIVE relative to JIVE declines. However, as the standard deviation of judge propensities exceeds 0.15 and identification becomes strong, the relative benefit of SJIVE diminishes.

4 Empirical Application

To illustrate the performance of SJIVE in a real-world application, we show results using data from Stevenson (2018), which uses data from criminal cases in Philadelphia to study the effect of pretrial detention on case outcomes. Defendants during the sample period are assigned to one of eight judges. With so few judges, each

Figure 3: SJIVE Relative Performance by Standard Deviation of Latent Propensity



Notes: Results are from Monte Carlo simulations in which each point represents the ratio of the standard deviation of SJIVE estimates to that of JIVE estimates. The x-axis shows the standard deviation of latent propensities. Half of judges have a caseload of 10 and the other half have a caseload of 100. The R^2 of W is .5. Each estimate is based on 5,000 observations. The standard deviations of both the outcome error and the first-stage error are set to 1, and the correlation between treatment and outcome is 0.5.

judge propensity is estimated with a high degree of precision. Creating a finer set of instruments based on judge-by-defendant characteristic cells has the potential to increase precision by taking advantage of within-judge variation in propensity to treat different types of people, though these subgroup propensities are measured much less precisely since they are each based on smaller numbers of observations. SJIVE combines the advantages of using judge-level and judge-by-characteristic-level indicators as instruments.

Table 2 reports estimates of the effect of pretrial detention on conviction using four estimators: OLS, 2SLS, UJIVE, and SJIVE. In Panel A, all specifications except OLS instrument for pretrial detention using judge-by-offense category indicators. In the SJIVE specification, we shrink the judge-by-offense category propensities to the judge-level propensities. In Panel B, the instruments are the judge indicators, and the

Table 2: Comparison of Estimators for the Effect of Pretrial Detention on Conviction

	OLS	2SLS	UJIVE	SJIVE
	(1)	(2)	(3)	(4)
A.: 1315 instruments (judge x off. type x demogr.)				
Detained	0.020	0.094	0.128	0.105
	(0.002)***	(0.020)***	(0.030)***	(0.026)***
B.: 7 instruments (judge indicators)				
Detained	0.000	0.125	0.128	0.129
	(0.002)	(0.063)**	(0.065)**	(0.065)**

Notes: This table reports estimates of the effect of pretrial detention on conviction. Panel A uses 1,315 instruments which are interactions of judge, offense type, demographics, and criminal history indicators. Panel B uses 8 judge indicators as instruments. In the SJIVE specification, the high-dimensional propensities in Panel A are shrunk to judge-level propensities, while in Panel B judge propensities are shrunk to a constant. Both panels are based on 324,162 observations and include controls for day of week, shift of day (graveyard, morning, evening), and a cubic in day of year (1–365). Panel A additionally controls for offense type-by-demographics-by-criminal history interactions. Standard errors are in parentheses.

SJIVE estimator shrinks to the grand mean. Because there are only eight judges, each judge's propensity to treat is measured with a high level of precision, so shrinking to the grand mean yields virtually no reduction in the standard error. In contrast, the SJIVE estimate in Panel A has a standard error 13% smaller than that of the UJIVE estimate. The results illustrate that SJIVE can substantially improve precision when the instrument set is large and the individual propensities are estimated with greater noise.

5 Conclusion

This paper introduces the shrunken jackknife instrumental variables estimator (SJIVE), which enhances the precision of IV estimates in judge fixed effects designs and other many-instrument settings by applying an empirical Bayes shrinkage procedure in the first stage. By differentially shrinking jackknife-predicted first-stage values toward conditional or overall means, SJIVE leverages heterogeneity in the precision

of instrument estimates across observations. This approach retains the robustness of JIVE to many-instrument bias while improving statistical efficiency—particularly when there is substantial variation in caseload size or when judge propensities are partially predictable from observable characteristics.

Through simulations, we demonstrate that SJIVE offers lower variance than JIVE and other common estimators across a wide range of conditions. The gains are most pronounced when instruments vary in informativeness such as when some judges handle few cases or when the strength of the first-stage is limited. These benefits are borne out in an empirical application using data from Stevenson (2018), where SJIVE produces smaller standard errors than JIVE when using judge-by-covariates indicators as instruments. Given the prevalence of judge fixed effects designs in applied microeconomics, SJIVE provides a simple and widely applicable improvement that enhances precision without introducing additional assumptions.

References

Anna Aizer and Joseph J. Doyle, Jr. Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *Quarterly Journal of Economics*, 130(2):759–803, May 2015.

Joshua D. Angrist and Brigham R. Frandsen. Machine labor. *Journal of Labor Economics*, 40(S1):S97–S129, 2022. doi: 10.1086/717933.

Joshua D. Angrist, Guido W. Imbens, and Alan B. Krueger. Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67, 1999. doi: 10.1002/(SICI)1099-1255(199901/02)14:1⟨57::AID-JAE501⟩3.0.CO;2-D.

David Arnold, Will Dobbie, and Crystal S Yang. Racial bias in bail decisions*. *The*

Quarterly Journal of Economics, page qjy012, 2018. doi: 10.1093/qje/qjy012. URL http://dx.doi.org/10.1093/qje/qjy012.

Jushan Bai and Serena Ng. Instrumental variable estimation in a data rich environment. *Econometric Theory*, 26(6):1577–1606, 2010. doi: 10.1017/S0266466609990727.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012. doi: 10.3982/ECTA9626.

John C. Chao, Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen. Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory*, 28(1):42–86, 2012. doi: 10.1017/S0266466611000120.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1): C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Laura Chioda and Michael Jansson. Optimal invariant inference when the number of instruments is large. *Econometric Theory*, 25(3):793–805, 2009.

Eric Chyn, Brigham Frandsen, and Emily C Leslie. Examiner and judge designs in economics: A practitioner's guide. Working Paper 32348, National Bureau of Economic Research, April 2024. URL http://www.nber.org/papers/w32348.

Stephen G. Donald and Whitney K. Newey. Choosing the number of instruments. *Econometrica*, 69(5):1161–1191, 2001.

Jeffrey R. Kling. Incarceration length, employment, and earnings. *American Economic Review*, 96(3):863–876, June 2006. doi: 10.1257/aer.96.3.863. URL `http://www.aeaweb.org/articles?id=10.1257/aer.96.3.863`.

Michal Kolesár. Estimation in an instrumental variables model with treatment effect heterogeneity, 2013. Unpublished working paper.

Emily Leslie and Nolan G. Pope. The unintended impact of pretrial detention on case outcomes: Evidence from new york city arraignments. *The Journal of Law and Economics*, 60(3):529–557, 2017.

Megan T. Stevenson. Distortion of justice: How the inability to pay bail affects case outcomes. *Journal of Law, Economics, & Organization*, 34(4):511–542, 2018.

Appendix: Formal results

Notation

Denote the $n \times p$ matrix of instruments as $Z = (Z_1, \ldots, Z_n)'$. Let the associated projection matrix be $H = Z(Z'Z)^{-1}Z'$ and the vector of in-sample fitted values be $\hat{P} = \left(\hat{P}_1, \ldots, \hat{P}_n\right)'$. The $n \times 1$ vector of population first-stage fitted values is $P = (P_1, \ldots, P_n)'$, and likewise $D = (D_1, \ldots, D_n)$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ and $\nu = (\nu_1, \ldots, \nu_n)'$. Let h_i be the ith diagonal element of H, and let h_{ij} be the element in the ith row and jth column of H. Define $A_i = h_i/(1-h_i)$ and $B_i = \sum_{i' \neq i} h_{ii'}^2/(1-h_i)(1-h_{i'})$. Convergence in distribution is denoted by $\to_d$.

Assumptions

We maintain the following assumption (independence and bounded moments) throughout:

Assumption A1 (Independence and bounded moments)**.** •

a. $\{(\varepsilon_i, \nu_i)\}_i$ *are independent across* i*;*

b. (ε_i, ν_i) *has finite fourth moments for all* i*;*

c. $\lim Var\left(\nu' H \nu / n\right) = \lim Var\left(\nu' H \varepsilon / n\right) = \lim Var\left(\hat{P}^{JK\prime} D / n\right) = 0$*;*

d. $\lim Var\left(n^{-1/2} \sum_i \hat{P}_i^{JK} \varepsilon_i\right) = \Sigma^{JK}$ *for some finite positive-definite* Σ^{JK}*.*

We make the following assumption (homoskedasticity) only in deriving the expression for the many-instruments bias of 2SLS and to show the relative efficiency of SJIVE. The assumption is not required for consistency of SJIVE nor for the asymptotic distribution we derive below for heteroskedasticity.

Assumption A2 (Homoskedasticity)**.** $E\left[\varepsilon_i^2\right] = \sigma_\varepsilon^2$*,* $E\left[\nu_i^2\right] = \sigma_\nu^2$*, and* $E\left[\varepsilon_i \nu_i\right] = \sigma_{\varepsilon\nu}$ *for all* i*.*

We make the following high-level regularity conditions on the sequence of instrument design matrices. These hold for the leading case of dummy instruments in the standard many-instrument asymptotic sequence.

Assumption A3 (Regularity conditions)**.**

a. $\lim \frac{1}{n} \sum_{i=1}^n \frac{h_i}{1-h_i} = A$ *for some finite* $A \geq 0$*;*

b. $\lim \frac{1}{n} \sum_{i=1}^n \sum_{i' \neq i} \frac{h_{ii'}^2}{(1-h_i)(1-h_{i'})} = B$ *for some finite* $B \geq 0$*;*

c. $\lim n^{-1} \sum_{i=1}^n \lambda_i$ *exists, and with some abuse of notation is denoted* $E\left[\lambda_i\right]$*;*

d. $\lim n^{-1}\sum_{i=1}^{n} S_i P_i$ *exists. Define* $\alpha = \left(\lim n^{-1}\sum_{i=1}^{n} S_i P_i\right)/\sigma_P^2$*, so that* α *has the interpretation as the* R^2 *from a bivariate regression of* P_i *on* S_i*;*

e. $\lim n^{-1}\sum_{i=1}^{n} \lambda_i P_i^2 = E\left[\lambda_i\right]\sigma_P^2$*;*

f. $\lim n^{-1}\sum_{i=1}^{n} \lambda_i S_i P_i = E\left[\lambda_i\right]\alpha\sigma_P^2$*;*

g. $\lim \frac{1}{n}\sum_{i=1}^{n}\sum_{i'\neq i} \frac{h_{ii'}^2 \lambda_i \lambda_{i'}}{(1-h_i)(1-h_{i'})} = \lim \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 \sum_{i'\neq i} \frac{h_{ii'}^2}{(1-h_i)(1-h_{i'})}$ *(rich first stage);*

Finally, we state high-level assumptions on the weak convergence of various bilinear forms. Primitive conditions for results such as these for jackknife IV estimators can be found in, for example, Chao et al. (2012).

Assumption A4 (Weak convergence of bilinear forms)**.** *The bilinear form* $M'\varepsilon/\sqrt{n}$ *converges in distribution to a normal random variable with mean zero and finite variance (that depends on* M*) for* $M \in \left\{\hat{P}^{JK}, \tilde{P}\right\}$*.*

Theorems and proofs

Theorem 1 (Oracle IV limiting distribution under homoskedasticity)**.** *Suppose Condition 1 and Assumptions A1 and A2 hold. Then*

$$\sqrt{n}\left(\hat{\delta}^{oracle} - \delta\right) \to_d N\left(0, \frac{\sigma_\varepsilon^2}{\sigma_p^2}\right).$$

Proof. Recall that the oracle IV estimator is given by

$$\hat{\delta}^{oracle} = \frac{\sum_{i=1}^{n} P_i Y_i}{\sum_{i=1}^{n} P_i D_i},$$

where $P_i = Z_i'\pi$. Substituting for Y_i from equation (1), we have

$$\begin{aligned}\sqrt{n}\left(\hat{\delta}^{oracle} - \delta\right) &= \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n P_i\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n P_i D_i} \\ &\to_d N\left(0, \frac{\lim_{n\to\infty} Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n P_i\varepsilon_i\right)}{\left(\text{plim}\,\frac{1}{n}\sum_{i=1}^n P_i D_i\right)^2}\right),\end{aligned}$$

where the second line follows from Condition 1(a), Lyapunov's central limit theorem, and Slutsky's theorem. The numerator of the limiting variance is

$$\lim_{n\to\infty} Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n P_i\varepsilon_i\right) = \sigma_p^2\sigma_\varepsilon^2,$$

by Assumptions A1 and A2 and the definition of σ_p^2 in Condition 1. The term in the denominator, after substituting for D_i from first-stage equation (2), is:

$$\begin{aligned}\text{plim}\,\frac{1}{n}\sum_{i=1}^n P_i\left(P_i + \nu_i\right) &= \text{plim}\,\frac{1}{n}\sum_{i=1}^n P_i^2 + \text{plim}\,\frac{1}{n}\sum_{i=1}^n P_i\nu_i \\ &= \sigma_p^2,\end{aligned}$$

where the second line follows from the definition of σ_p^2 and the law of large numbers. Putting the numerator and denominator together we get the result:

$$\sqrt{n}\left(\hat{\delta}^{oracle} - \delta\right) \to_d N\left(0, \frac{\sigma_\varepsilon^2}{\sigma_p^2}\right).$$

∎

Theorem 2 (2SLS many-instruments bias under homoskedasticity)**.** *Suppose Condi-*

tion 1 and Assumptions A1 and A2 hold. Then

$$\hat{\delta}^{2SLS} \to_p \delta + \frac{\kappa\sigma_{\varepsilon\nu}}{\sigma_P^2 + \kappa\sigma_\nu^2}.$$

Proof. Using equation (1) to substitute for Y_i in the definition of $\hat{\delta}^{2SLS}$, we have

$$\begin{aligned}
p\lim \hat{\delta}^{2SLS} &= \delta + \frac{p\lim \frac{1}{n}\hat{P}'\varepsilon}{p\lim \frac{1}{n}\hat{P}'D} \\
&= \delta + \frac{p\lim \frac{1}{n}\left(P_Z\left(Z\pi+\nu\right)\right)'\varepsilon}{p\lim \frac{1}{n}\left(P_Z\left(Z\pi+\nu\right)\right)'\left(Z\pi+\nu\right)} \\
&= \delta + \frac{p\lim \frac{1}{n}P'\varepsilon + p\lim \frac{1}{n}v'P_Z\varepsilon}{\lim \frac{1}{n}P'P + 2p\lim \frac{1}{n}P'\nu + p\lim \frac{1}{n}v'P_Z\nu} \\
&= \delta + \frac{\lim E\left[\frac{1}{n}v'P_Z\varepsilon\right]}{\sigma_p^2 + \lim E\left[\frac{1}{n}v'P_Z\nu\right]} \\
&= \delta + \frac{\lim \frac{1}{n}tr\left(P_Z E\left[\varepsilon v'\right]\right)}{\sigma_p^2 + \lim \frac{1}{n}tr\left(P_Z E\left[\nu v'\right]\right)} \\
&= \delta + \frac{\kappa\sigma_{\varepsilon\nu}}{\sigma_P^2 + \kappa\sigma_\nu^2},
\end{aligned}$$

where the first equality follows from the continuous mapping theorem, the second line substitutes for D from first-stage equation (2), the third line follows from simple algebra, the fourth line follows from Condition 1, the law of large numbers, and Assumption A1. The fifth line follows from the invariance of the trace to cyclical permutations, and the final line uses Assumption A2 and the fact that the trace of a P_Z is equal to the rank of Z. ∎

Theorem 3 (JIVE limiting distribution under homoskedasticity)**.** *Suppose Condition 1 and Assumptions A1, A2, A3, and A4 hold. Then*

$$\sqrt{n}\left(\hat{\delta}^{JIVE} - \delta\right) \to_d N\left(0, c^{JIVE}\frac{\sigma_\varepsilon^2}{\sigma_P^2}\right),$$

for $c^{JIVE} = 1 + \frac{\sigma_\nu^2}{\sigma_P^2}A + \frac{\sigma_{\varepsilon\nu}^2}{\sigma_\varepsilon^2\sigma_P^2}B \geq 1$.

Proof. The centered and scaled JIVE estimator simplifies to:

$$\sqrt{n}\left(\hat{\delta}^{JIVE} - \delta\right) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{P}_i^{JK}\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n \hat{P}_i^{JK}D_i}.$$

The denominator converges in probability:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \hat{P}_i^{JK}D_i &\to_p \lim\frac{1}{n}\sum_{i=1}^n E\left[\hat{P}_i^{JK}D_i\right] \\ &= \lim\frac{1}{n}\sum_{i=1}^n P_i^2 \\ &= \sigma_P^2,\end{aligned}$$

where the first line follows from the law of large numbers and Assumption A1, the second line from the fact that $E[\nu_i] = 0$ for all i and Assumption A1, and the final line from the definition of σ_P^2. The numerator converges in distribution by Assumption A4, with limiting variance

$$\lim Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{P}_i^{JK}\varepsilon_i\right) = \lim\frac{1}{n}\left(\sum_{i=1}^n E\left[\left(\hat{P}_i^{JK}\right)^2\right]\sigma_\varepsilon^2 + \sum_{i=1}^n\sum_{i'\neq i} Cov\left(\hat{P}_i^{JK}\varepsilon_i, \hat{P}_{i'}^{JK}\varepsilon_{i'}\right)\right),$$

where the equality uses Assumption A2. First, note that

$$E\left[\left(\hat{P}_i^{JK}\right)^2\right] = P_i^2 + \frac{h_i}{1-h_i}\sigma_\nu^2,$$

by Assumption A2 and the fact that $\sum_{j=1}^n h_{ij}^2 = h_i$. Next, note that by Assumptions A1 and A2

$$Cov\left(\hat{P}_i^{JK}\varepsilon_i, \hat{P}_{i'}^{JK}\varepsilon_{i'}\right) = \frac{\sigma_{\varepsilon\nu}^2 h_{ii'}^2}{(1-h_i)(1-h_{i'})}.$$

Combining these two results, we have

$$\lim Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{P}_i^{JK}\varepsilon_i\right) = \left(1+\frac{\sigma_\nu^2}{\sigma_P^2}A+\frac{\sigma_{\varepsilon\nu}^2}{\sigma_\varepsilon^2\sigma_P^2}B\right)\sigma_\varepsilon^2\sigma_P^2.$$

By Slutsky's theorem we therefore have that

$$\sqrt{n}\left(\hat{\delta}^{JIVE}-\delta\right) \to_d N\left(0, c^{JIVE}\frac{\sigma_\varepsilon^2}{\sigma_P^2}\right),$$

where

$$c^{JIVE} = 1+\frac{\sigma_\nu^2}{\sigma_P^2}A+\frac{\sigma_{\varepsilon\nu}^2}{\sigma_\varepsilon^2\sigma_P^2}B \geq 1.$$

∎

Theorem 4 (Optimal shrinkage and SJIVE improvement over JIVE under homoskedasticity)**.** *Let $\hat{\delta}(\lambda)$ be a shrunken jackknife IV estimator with shrinkage factors λ:*

$$\hat{\delta}(\lambda) = \frac{\sum_{i=1}^{n}\tilde{P}_i Y_i}{\sum_{i=1}^{n}\tilde{P}_i D_i}$$

where $\tilde{P}_i = \lambda_i\hat{P}_i^{JK} + (1-\lambda_i)S_i$, and suppose Condition 1 and Assumptions A1, A2, A3, and A4 hold. Then the asymptotically efficient (limiting variance minimizing) choice of shrinkage factor is

$$\lambda_i = \frac{(1-\alpha)\sigma_P^2}{(1-\alpha)\sigma_P^2+\sigma_\nu^2 A_i+\frac{\sigma_{\varepsilon\nu}^2}{\sigma_\varepsilon^2}B_i},$$

and the corresponding SJIVE estimator converges in distribution:

$$\sqrt{n}\left(\tilde{\delta}-\delta\right) \to_d N\left(0, c^{SJIVE}\frac{\sigma_\varepsilon^2}{\sigma_P^2}\right),$$

for $c^{SJIVE} \leq c^{JIVE}$.

Proof. The centered and scaled shrunken jackknife IV estimator simplifies to:

$$\sqrt{n}\left(\hat{\delta}\left(\lambda\right)-\delta\right)=\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{P}_{i}\varepsilon_{i}}{\frac{1}{n}\sum_{i=1}^{n}\tilde{P}_{i}D_{i}}.$$

First, the denominator converges in probability:

$$\begin{aligned}
p\lim\frac{1}{n}\sum_{i=1}^{n}\tilde{P}_{i}D_{i} &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(1-\lambda_{i}\right)S_{i}P_{i}+p\lim\frac{1}{n}\sum_{i=1}^{n}\lambda_{i}\hat{P}_{i}^{JK}D_{i}\\
&= \left(1-E\left[\lambda_{i}\right]\right)\alpha\sigma_{P}^{2}+\lim\frac{1}{n}\sum_{i=1}^{n}\lambda_{i}E\left[\hat{P}_{i}^{JK}D_{i}\right]\\
&= \left(1-E\left[\lambda_{i}\right]\right)\alpha\sigma_{P}^{2}+\lim\frac{1}{n}\sum_{i=1}^{n}\lambda_{i}P_{i}^{2}\\
&= \left(\alpha+\left(1-\alpha\right)E\left[\lambda_{i}\right]\right)\sigma_{P}^{2},
\end{aligned}$$

where the first line follows from the definition of $\tilde{P}_i$ and the fact that $E\left[\nu_i\right]=0$ for all i; the second line follows from Assumption A3, the law of large numbers, and Assumption A2; the third line follows from Assumption A1 and the fact that $E\left[\nu_i\right]=0$ for all i; and the fourth line follows from Assumption A3.

Next, by Assumption A4, the numerator converges in distribution to a mean-zero normal random variable with the following variance:

$$\lim Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{P}_{i}\varepsilon_{i}\right)=\lim\frac{1}{n}\sum_{i=1}^{n}\left(Var\left(\tilde{P}_{i}\varepsilon_{i}\right)+\sum_{i'\neq i}Cov\left(\tilde{P}_{i}\varepsilon_{i},\tilde{P}_{i'}\varepsilon_{i'}\right)\right). \quad (4)$$

First, by Assumption A2, the variance term on the right hand side of the equation is

$$\begin{aligned}
Var\left(\tilde{P}_{i}\varepsilon_{i}\right) &= \sigma_{\varepsilon}^{2}E\left[\tilde{P}_{i}^{2}\right]\\
&= \sigma_{\varepsilon}^{2}\left(Var\left(\tilde{P}_{i}\right)+E\left[\tilde{P}_{i}\right]^{2}\right)\\
&= \sigma_{\varepsilon}^{2}\left(\lambda_{i}^{2}A_{i}\sigma_{\nu}^{2}+\left(\left(1-\lambda_{i}\right)S_{i}+\lambda_{i}P_{i}\right)^{2}\right).
\end{aligned}$$

Next, the covariance term in the double sum on the right side of (4) simplifies to:

$$Cov\left(\tilde{P}_i\varepsilon_i, \tilde{P}_{i'}\varepsilon_{i'}\right) = \sigma^2_{\varepsilon\nu} \sum_{i' \neq i} \frac{h^2_{ii'}\lambda_i\lambda_{i'}}{(1-h_i)(1-h_{i'})},$$

which uses the fact that by Assumptions A1 and A2 $Cov\left(\hat{P}_i\varepsilon_i, \hat{P}_{i'}\varepsilon_{i'}\right) = h^2_{ii'}\sigma^2_{\varepsilon\nu}$ and the following terms are all zero: $Cov\left(\varepsilon_i, \hat{P}^{JK}_{i'}\varepsilon_{i'}\right)$, $Cov\left(\hat{P}^{JK}_i\varepsilon_i, \varepsilon_{i'}\right)$, $Cov\left(\hat{P}_i\varepsilon_i, \varepsilon_{i'}\right)$, $Cov\left(\nu_i\varepsilon_i, \hat{P}_{i'}\varepsilon_{i'}\right)$. Putting these results together, equation (4) becomes

$$\begin{aligned}
&\lim Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{P}_i\varepsilon_i\right) \\
&= \lim \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2_\varepsilon\left(\lambda^2_i A_i\sigma^2_\nu + \left((1-\lambda_i)S_i + \lambda_i P_i\right)^2\right) + \sigma^2_{\varepsilon\nu}\sum_{i'\neq i}\frac{h^2_{ii'}\lambda_i\lambda_{i'}}{(1-h_i)(1-h_{i'})}\right) \\
&= \lim \frac{1}{n}\sum_{i=1}^{n}\left(\left((1-\lambda_i)S_i + \lambda_i P_i\right)^2 + \lambda^2_i\left(A_i\sigma^2_\nu + \frac{\sigma^2_{\varepsilon\nu}}{\sigma^2_\varepsilon}B_i\right)\right)\sigma^2_\varepsilon,
\end{aligned}$$

where the third line follows from Assumption A3. Putting the numerator and denominator together, the limiting variance of $\sqrt{n}\left(\hat{\delta}(\lambda) - \delta\right)$ is

$$c(\lambda)\frac{\sigma^2_\varepsilon}{\sigma^2_P},$$

where

$$c(\lambda) = \frac{\lim \frac{1}{n}\sum_{i=1}^{n}\left(\left((1-\lambda_i)S_i + \lambda_i P_i\right)^2 + \lambda^2_i\left(A_i\sigma^2_\nu + \frac{\sigma^2_{\varepsilon\nu}}{\sigma^2_\varepsilon}B_i\right)\right)}{\left(\alpha + (1-\alpha)E[\lambda_i]\right)^2\sigma^2_P}.$$

Minimizing $c(\lambda)$ with respect to $\{\lambda_i\}_{i=1}^n$, the asymptotically efficient shrinkage factors are:

$$\lambda_i = \frac{(1-\alpha)\sigma^2_p}{(1-\alpha)\sigma^2_p + A_i\sigma^2_\nu + \frac{\sigma^2_{\varepsilon\nu}}{\sigma^2_\varepsilon}B_i},$$

and the corresponding limiting variance is

$$\begin{aligned} & c^{SJIVE}\frac{\sigma_\varepsilon^2}{\sigma_P^2}, \\ c^{SJIVE} \quad = \quad & \frac{1}{\alpha + (1-\alpha)\, E\left[\lambda_i\right]}. \end{aligned}$$

Because JIVE is a special case of SJIVE with $\lambda_i = 1$ for all i, a corollary is that $c^{SJIVE} \leq c^{JIVE}$. ∎

Theorem 5 (SJIVE limiting distribution under heteroskedasticity)**.** *Suppose Condition 1 and Assumption A1, A3, and A4 hold. Then*

$$\sqrt{n}\left(\hat{\delta}^{SJIVE} - \delta\right) \to_p N\left(0, V^{robust}\right),$$

where

$$V^{robust} = \lim_{n\to\infty} \frac{\frac{1}{n}\left(\sum_{i=1}^n Var\left(\tilde{P}_i\varepsilon_i\right) + \sum_{i,i'} Cov\left(\tilde{P}_i\varepsilon_i, \tilde{P}_{i'}\varepsilon_{i'}\right)\right)}{\left(p\lim \frac{1}{n}\sum_{i=1}^n \tilde{P}_i D_i\right)^2}.$$

Proof. By equation (1) the centered and scaled estimator can be written:

$$\sqrt{n}\left(\hat{\delta}^{SJIVE} - \delta\right) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{P}_i\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n \tilde{P}_i D_i}.$$

By Assumption A3 the denominator converges in probability. By Assumption A4 the numerator converges in distribution to a mean-zero normal random variable with limiting variance $\lim Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{P}_i\varepsilon_i\right)$. The result therefore follows by Assumption A1 and Slutsky's theorem. ∎

The limiting variance may be consistently estimated by

$$\hat{V}^{robust} = \frac{\frac{1}{n}\left(\sum_{i=1}^{n}\left(\tilde{P}_i^2\hat{u}_i^2 + \sum_{i'\neq i}\lambda_i\lambda_{i'}\frac{h_{ii'}^2\hat{u}_i\hat{v}_i\hat{u}_{i'}\hat{v}_{i'}}{(1-h_i)(1-h_{i'})}\right)\right)}{\left(p\lim\frac{1}{n}\sum_{i=1}^{n}\tilde{P}_iD_i\right)^2},$$

where $\hat{u}_i = Y_i - \hat{\delta}^{SJIVE}D_i$ and $\hat{\nu}_i$ is the ith residual from the estimated first stage.