Working Paper Series 759
(ISSN 2788-0443)

Instrumental Variable Estimation with Many Instruments Using Elastic-Net IV

Alena Skolkova

CERGE-EI
Prague, July 2023

ISBN 978-80-7343-566-0 (Univerzita Karlova, Centrum pro ekonomický výzkum a doktorské studium)
ISBN 978-80-7344-686-4 (Národohospodářský ústav AV ČR, v. v. i.)

Instrumental Variable Estimation with Many Instruments Using Elastic-Net IV*

Alena Skolkova†

Abstract

Instrumental variables (IV) are commonly applied for identification of treatment effects and subsequent policy evaluation. The use of many informative instruments improves the estimation accuracy. However, dealing with high-dimensional sets of instrumental variables of unknown strength may be complicated and requires model selection or regularization of the first stage regression. Currently, lasso is established as one of the most popular regularization techniques relying on the assumption of approximate sparsity. I investigate the relative performance of the lasso and elastic-net estimators for fitting the first-stage as part of IV estimation. As elastic-net includes a ridge-type penalty in addition to a lasso-type penalty, it generally improves upon lasso in finite samples when correlations among the instrumental variables are not negligible. I show that IV estimators based on the lasso and elastic-net first-stage estimates can be asymptotically equivalent. Via a Monte Carlo study I demonstrate the robustness of the sample-split elastic-net IV estimator to deviations from approximate sparsity, and to correlation among possibly high-dimensional instruments. Finally, I provide an empirical example that demonstrates potential improvement in estimation accuracy gained by the use of IV estimators based on elastic-net.

*I am grateful to my advisor, Stepan Jurajda, Stanislav Anatolyev, Nikolas Mittag and CERGE-EI faculty for their feedback. I thank the participants at the 24th International Conference on Computational Statistics and the 11th Nordic Econometric Meeting for their comments. This study was supported by Charles University, GAUK project No. 304620

†CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences, 111 21 Politickych veznu 7, Prague, Czech Republic. Email: alena.skolkova@cerge-ei.cz

1 Introduction

The instrumental variables (IV) regression is a common tool for identification of treatment effects under regressor endogeneity. From the theoretical perspective, researchers would like to utilize as much exogenous variation in the explanatory variables as possible, as it increases the precision of IV estimates: Newey (1990), Amemiya (1974) and Chamberlain (1987) motivate the use of many instruments for the purpose of nonparametric estimation of optimal instruments. However, conventional GMM-type estimators, such as 2SLS, tend to be substantially biased when the number of instrumental variables is not small relative to the sample size: see Bekker (1994) and Newey and Smith (2004).

The problem of many instruments may be circumvented in various ways. The use of statistical methods with imbedded regularization is increasingly popular among economists. Regularization techniques allow one to deal with ill-posed inverse problems, and date back to Tikhonov (1943). Such methods include the ridge regression (Hoerl and Kennard, 1970), lasso (Tibshirani, 1996), the penalized maximum likelihood estimation (Friedman, Hastie and Tibshirani, 2001), and boosting (Buhlmann, 2006), among others. There are several alternative regularization procedures used as part of IV estimators: ridge and James-Stein type shrinkage applied to the first stage by Hansen and Kozbur (2014) and Spiess (2017), respectively; lasso for estimation of both the first stage and the reduced form by Belloni, Chen, Chernozhukov and Hansen (2012, hereafter BCCH); applications of random forests and deep neural networks by Wager and Athey (2018) and Farrell et al (2021), respectively. In this list, BCCH stands out due to the extreme popularity of lasso as a regularization technique that is often employed under sparsity. In sparse models, there is a small number of variables[1] that convey most of the impact of all covariates in the response variable. Lasso represents the simplest sparse modeling approach that allows simultaneous variable selection and coefficient estimation.

The key assumption needed for lasso to produce a meaningful solution is the sparsity of the underlying model (see Section 2.1 for the definition of sparsity). The sparsity assumption may

[1] $s = o(n)$,where n is the sample size.

be justified in structural economic equations, where few variables participate in determinating an outcome variable. However, the lasso estimator is also promoted as a universal workhorse for pure prediction tasks. Despite the popularity of the sparse modeling framework, the adequacy of the sparsity (or approximate sparsity) assumption is often questionable. For example, Giannone et al. (2021) find evidence against sparsity for a collection of empirical applications from macroeconomics, microeconomics, and finance, where sparsity is routinely assumed without pretesting.

Furthermore, the simplicity of the lasso approach has its costs even under sparsity. For example, Zou and Hastie (2005, hereafter ZH) stress three limitations of classical lasso:
(1) if predictors are highly correlated as a whole, the prediction performance of the ridge regression dominates that of lasso (first observed in Tibshirani, 1996), as with highly correlated predictors the lasso solution paths tend to be unstable ; (2) in the $p > n$ case, when the number of variables p exceeds the number of observations n, lasso selects at most n variables; (3) if there are groups of predictors within which pairwise correlations are high, lasso generally selects only one variable from each group. ZH propose an alternative estimator – elastic-net (EN) – that successfully eliminates these shortcomings of lasso.[2] Through a simulation study and empirical examples they show that elastic-net often outperforms lasso in terms of prediction accuracy. In addition, EN essentially combines the properties of lasso and ridge , thus being able to accommodate some DGP's deviations from sparsity.

Of the three above-mentioned conditions under which the performance of lasso may be improved, at least the first two directly relate to IV estimation. Economists often estimate a causal effect based on a dataset at hand with many characteristics available for every unit (possibly $p > n$), where many serve as potential instruments (including the basic instrumental variables, their interactions and transformations). These instruments, however, tend to be moderately or highly correlated, leading to unstable lasso solution paths.[3] Thus, by using lasso to tackle the first-stage prediction problem, one faces exactly the scenario under which the performance of lasso may be improved via an additional ridge-type regularization, therefore justifying the use of the elastic-net technique.

[2]Elastic-net reduces to lasso in an orthogonal design, where lasso is optimal, see Donoho et al (1995).

[3]Under high dimensionality of the problem, even when the instrumental variables are independent, there might be large sample correlations, see Fan & Lv (2008).

This paper contributes to the literature on IV estimation with many instruments by considering the use of the elastic-net approach for estimating the first-stage regression. While the lasso (and post-lasso) IV estimator by BCCH and the ridge jackknife IV estimator by Hansen and Kozbur (2014) stem from the sparsity and the density of the first-stage relationship, respectively, I propose the elastic-net IV estimator (ENIV), which fits between those two. Similarly to lasso, elastic-net with a properly selected penalty parameter is shown to have oracle properties[4] under sparsity. Consequently, the results of BCCH on consistency and asymptotic normality (under possible non-Gaussianity and heteroskedasticity of the error term) of a generic sparsity-based IV estimator can be applied to the proposed elastic-net IV estimator. At the same time, in the case of no sparsity, elastic-net is by construction capable of acting like a ridge regression. Thus, for elastic-net with data-driven parameters (a penalty level, and a weighting parameter reflecting the degree of DGP sparsity), the proposed estimator should be robust to the unknown degree of sparsity of first-stage relationships.

To address the issue of overfitting (see, for example Chernozhukov et al, 2018), I consider sample-split and cross-fit versions of the basic elastic-net IV estimator (SS-ENIV and CF-ENIV, respectively), and compare them with the lasso-based analogues. I study the relative performance of the proposed estimators via simulations. Specifically, I compare the resulting IV estimates in terms of the median absolute bias, median absolute deviation and rejection rate. The SS-ENIV and CF-ENIV estimators perform well relative to the lasso-based alternatives, regardless of the signal's sparsity.

Additionally, I demonstrate the potential gains of the EN-based IV estimation based on the classic empirical investigation from Angrist and Krueger (1991), who look at the causal effect of schooling on earnings. The identification strategy and data from Angrist and Krueger (1991) provide many available instrumental variables for schooling. While employing as many of them as possible potentially leads to higher accuracy of the estimated causal effect, it also leads to biases and inferential problems. Therefore, the use of instrument selection or regularization techniques is justified, thus

[4] i.e. to achieve the rate of convergence that is very close to the oracle rate $\sqrt{s/n}$ achievable when the true model is known.

making the example suitable for testing the performance of the EN-based IV estimators.

The plan of this paper is as follows. In Section 2 I describe an instrumental variables setup and overview the regularization-based methods for estimation of optimal instruments. In Section 3 I present and discuss the results of a simulation study that examines the performance of the proposed estimator relative to its closest competitors. Section 4 provides an empirical example to demonstrate potential improvement in estimation accuracy gained by the use of IV estimators based on elastic-net.

2 The Instrumental Variables Model

The problem setup is similar to that from BCCH, simplified to the case of a scalar endogenous variable. The model is $y_i = d_i'\delta_0 + e_i$, where y_i is a scalar outcome, d_i is a k_d-vector of variables, and δ_0 denotes the true value of a vector-valued parameter δ. The first element of d_i is endogenous, while the remaining elements of d_i constitute a vector of exogenous covariates w_i. The disturbance term e_i is such that $E[e_i|z_i] = 0$, where z_i is a k_z-vector of instrumental variables.

As a motivation, suppose the disturbance term is conditionally homoskedastic, $E[e_i^2|z_i] = \sigma^2$. For a k_d-vector of instruments $A(z_i)$, the standard IV estimator of δ_0 takes the form

$$\hat{\delta} = \left(\mathbb{E}_n\left[A(z_i)\, d_i'\right]\right)^{-1} \mathbb{E}_n\left[A(z_i)\, y_i\right],$$

where $\{(z_i, d_i, y_i), i = 1, ..., n\}$ is i.i.d. sample, $\mathbb{E}_n[f] \coloneqq \mathbb{E}_n[f(z_i)] \coloneqq \sum_{i=1}^n f_i/n$. Given instruments $A(z_i)$,

$$\sqrt{n}\left(\hat{\delta} - \delta_0\right) \to^d \mathcal{N}\left(0, Q_0^{-1}\Omega_0 Q_0^{-1}\right),$$

where $Q_0 = E[A(z_i)\, d_i']$ and $\Omega_0 = \sigma^2 E\left[A(z_i)\, A(z_i)'\right]$. Employing the optimal instrument $A(z_i) = D(z_i) = E[d_i|z_i]$ achieves the semiparametric efficiency bound for estimating δ_0, with the asymptotic variance $\Lambda^* = \sigma^2\left\{E\left[D(z_i)\, D(z_i)'\right]\right\}^{-1}$ (see Chamberlain, 1987).

2.1 Regularized Estimation Methods for Optimal Instruments

In practice, the optimal instrument $D(z_i)$ is not known, and many ways to estimate it exist in the literature. Suppose there is a large set of instruments,

$$f_i := (f_{i1}, ..., f_{ip})' := (f_1(z_1), ..., f_p(z_1))'$$

available for estimation of conditional expectation $D(z_i)$, and the number of instruments p is possibly larger than the sample size n. In BCCH, the optimal instrument $D(z_i)$ is assumed to be approximately sparse, i.e. a function $D(z_i)$ is deemed to be well-approximated by a function of unknown $1 \leq s \ll n$ instruments:

$$\begin{aligned} D(z_i) &= f_i'\beta_0 + a(z_i), \\ \|\beta_0\|_0 \leq s &= o(n), \ \left[\mathbb{E}_n a(z_i)^2\right]^{1/2} \leq c_s \lesssim_P \sqrt{s/n}. \end{aligned}$$

The identities of s relevant instruments, i.e. $T = \text{support}(\beta_0) = \{j \in \{1, ..., p\} : |\beta_{0j}| > 0\}$, are meant to be a priori unknown. The sparsity assumption requires that at most s instruments approximate the conditional expectation $D(z_i)$ so that the approximation error $a(z_i)$ does not exceed the conjectured size $\sqrt{s/n}$ of the error of the infeasible estimator that "knows" the identity of these s relevant instruments (the "oracle estimator").

Lasso

The first stage regression equation is

$$d_i = D(z_i) + v_i, \quad E[v_i|z_i] = 0.$$

For the sample $\{(z_i, d_i), i = 1, ..., n\}$, consider estimators of the optimal instrument $D(z_i)$ of the form

$$\widehat{D}_i := \widehat{D}(z_i) = f_i'\widehat{\beta},$$

where $\widehat{\beta}$ is the sparse estimator based on regressors f_i and d_i as the dependent variable. The sparse

estimator sets all but a small fraction of the coefficient estimates $\hat{\beta}_j$ to 0. Let $Q(\beta)$ denote the least squares criterion function, $\widehat{Q}(\beta) := \mathbb{E}_n\left[(d_i - f_i'\beta)^2\right]$, then the lasso estimator employed in BCCH is defined as a solution to

$$\widehat{\beta}_L \in \arg\min_{\beta \in R^p} \widehat{Q}(\beta) + \lambda^L \left\| \widehat{\Upsilon}\beta \right\|_1,$$

where λ^L is the penalty level and $\widehat{\Upsilon} = \text{diag}(\widehat{\gamma}_1, ..., \widehat{\gamma}_p)$ is a diagonal matrix with data-dependent weights, also called penalty loadings. The basic lasso estimator, with all penalty loadings set to 1, was introduced by Tibshirani (1996) as a technique for simultaneous estimation and variable selection. Basically, lasso shrinks the coefficients toward 0 as λ^L increases, and some coefficient estimates are set to 0 for large enough λ^L.

Lasso has been shown to be variable selection consistent, i.e. to be able to discover the correct model specification, under suitable conditions (see Meinshausen and Bühlmann, 2004). Initially, the weighted/adaptive version of lasso (with data-dependent penalty loadings) was proposed in Zou (2006) in response to debates about whether the lasso estimator is an oracle procedure (Fan and Li, 2001; Meinshausen and Bühlmann, 2004). For the data-dependent and cleverly chosen loadings[5], the adaptive lasso estimator is shown to enjoy oracle properties. Relatively recently, BCCH have proposed novel penalty loadings that result in sharp convergence rates for the lasso estimator under possible non-Gaussianity and heteroskedasticity.

Having estimated the optimal instrument via lasso, let $\widehat{D}_i$ be a vector of instruments that also includes the vector of exogenous covariates w_i

$$\widehat{D}_i = \left(\widehat{D}(z_i), w_i'\right)'.$$

Then the resulting lasso-IV estimator

$$\widehat{\delta}^L = \mathbb{E}_n\left[\widehat{D}_i d_i'\right]^{-1} \mathbb{E}_n\left[\widehat{D}_i y_i\right] \tag{1}$$

[5]Zou (2006) suggests the weight vector $\hat{w} = 1/\left|\hat{\beta}\right|^{\gamma}$, where $\hat{\beta}$ is a root-n consistent estimator for β, and $\gamma > 0$.

is shown to achieve the efficiency bound asymptotically, $\sqrt{n}\left(\widehat{\delta}^{L}-\delta_0\right) =_d N\left(0,\Lambda^*\right)+o_P\left(1\right)$. The IV estimator with the lasso-based optimal instrument is root-n consistent and asymptotically normal (see Theorem 3 of BCCH). Moreover, consistency and asymptotic normality continues to hold for any generic sparsity-based method achieving specific near-oracle performance bounds (see Theorem 4 of BCCH), and I exploit this result in the next section.

Elastic-Net IV Estimator

Although lasso is aimed at high-dimensional problems, its performance may be deteriorated by the correlation among predictors, which often takes place in high-dimensional settings. Zou & Hastie (2005, hereafter ZH) point out that the lasso solution paths are unstable (i.e. not smooth) when predictors are highly correlated. The relevance of this issue is stressed by Fan & Lv (2007) who show that even with the independent predictors the maximum sample correlation can be large, as long as the dimensionality is high. In addition, ZH notice that for high-dimensional problems with $p \gg n$, lasso is incapable of selecting more than p variables into the model. Consequently, they propose an alternative penalized estimator, elastic-net (EN),

$$\hat{\beta}^{EN} = \arg\min_{\beta}\left\{\sum_{i=1}^{N}\left(d_i-\sum_{j=1}^{p}f_{ij}\beta_j\right)^2+\lambda^{EN}\sum_{j=1}^{p}\left(\alpha\left|\beta_j\right|+\left(1-\alpha\right)\beta_j^2\right)\right\},$$

which involves an l_2-penalty in addition to lasso's l_1-penalty. The first term of the penalty, $\lambda^{EN}\sum_{j=1}^{p}\alpha\left|\beta_j\right|$ encourages a sparse solution, as does the lasso penalty, while the second term, $\lambda^{EN}\sum_{j=1}^{p}\left(1-\alpha\right)\beta_j^2$, regularizes the covariance matrix, and encourages equality of the coefficients on highly correlated predictors. ZH shows that elastic-net may be interpreted as a stabilized[6] version of lasso (p. 308, Theorem 2), and can therefore
improve upon lasso.

In the statistical literature, the performance of the elastic-net estimator is usually analyzed under a restrictive assumption of the Gaussian and homoskedastic error term. For example, when Gaussian and homoskedastic noise is assumed, Jia and Yu (2010) study the model selection properties of the elastic-net estimator in the asymptotic framework where the number of variables p grows with the

[6]Stabilization is achieved via replacement of the sample covariance matrix $\widehat{\Sigma}$ with its shrunken (towards the identity matrix) version.

sample size n. They provide sufficient conditions for elastic-net to be model selection consistent[7], as well as theoretical and simulation examples demonstrating when elastic-net can consistently select the true model, while lasso fails to do so.[8] Further, Ghosh (2011) considers adaptive elastic-net that generalizes elastic-net in the same way that adaptive lasso generalizes lasso, thus expanding the set of conditions under which elastic-net performs consistent variable selection. The adaptive elastic-net estimator uses a more flexible l_1-penalty for consistent variable selection, while the ridge-type penalty term stays unchanged[9] and continues to regularize the solution path:

$$\hat{\beta}_{ada}^{EN} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(d_i - \sum_{j=1}^{p} f_{ij}\beta_j \right)^2 + \lambda_1^{EN} \sum_{j=1}^{p} w_j \left|\beta_j\right| + \lambda_2^{EN} \sum_{j=1}^{p} \beta_j^2 \right\},$$

where the weight estimate $\widehat{w}_j = 1/|\widehat{\beta}_j|^{\gamma}$, $j = 1, ..., p$, for some $\gamma > 0$, with the ordinary least squares estimator $\widehat{\beta}^{OLS}$ being a possible choice of $\widehat{\beta}$. Under suitable conditions, the adaptive elastic-net estimator is shown to have oracle properties (variable selection consistency and asymptotic normality, see Theorem 3.2).

However, the breakthrough results of Theorem 4 in BCCH on root-n consistency and asymptotic normality apply to a wide class of sparsity-based methods that encompasses the elastic-net estimator. Consequently, to get the desired asymptotic properties of the elastic-net estimator under possible non-Gaussianity and heteroskedasticity of the error term, it is enough to establish the near-oracle bounds that are required by BCCH's Theorem 4. I use the result from Zou and Hastie (2006) about transformation of the elastic-net problem into an equivalent lasso problem on augmented data to show that the elastic-net estimator performs closely enough to the oracle under sparsity, in the sense of meeting sufficient conditions of BCCH's Theorem 4.

Proposition 1. For $\left(\lambda_1^{EN}, \lambda_2^{EN}\right)$ such that $\gamma = \lambda_1^{EN}/\sqrt{1+\lambda_2^{EN}} = \lambda_{opt}^{L}$, where λ_{opt}^{L} denotes the optimal penalty for the lasso-estimator, the elastic-net estimator obeys the near-oracle performance

[7] Jia and Yu (2010) also state a specific condition for the inconsistency of the elastic-net estimator.

[8] See also Yuan and Lin (2007) for a similar study for fixed p.

[9] In principle, adaptive weights can also be placed on an l_2 penalty, but it is not necessary to guarantee the oracle properties of the adaptive elastic-net estimator examined in Ghosh (2011).

bounds:

$$\left\|\widehat{D}_i^{EN} - D_i\right\|_{2,n} \leq_p \sqrt{\frac{s\log(n+p)}{n+p}}$$
$$\left\|\widehat{\delta}^{EN} - \delta_0\right\|_1 \leq_p \sqrt{\frac{s^2\log(n+p)}{n+p}}$$

Therefore, the elastic-net estimator can perform a variable selection and estimation similarly to the lasso estimator. Once the sufficient conditions of Theorem 4 in BCCH continue to hold, one can rely on the existing results regarding consistency and asymptotic normality of generic sparsity-based IV estimators obtained in BCCH. In other words, the IV estimators based on elastic-net and lasso can be asymptotically equivalent under sparsity and the appropriate choice of the penalty parameters $\left(\lambda_1^{EN}, \lambda_2^{EN}\right)$. At the same time, ridge regularization often leads to finite-sample improvement, so the relative finite-sample performance of the IV estimators based on elastic-net (with a ridge-type penalty) and lasso (without a ridge-type penalty) is of interest, and is investigated in Section 3 of this paper.

Sample-Split and Cross-Fit Elastic-Net IV Estimator

In principle, one could employ $\widehat{D}_i = f_i'\widehat{\beta}^{EN}$ for $\widehat{D}_i$ in (2.1) to define an IV estimator with a EN-regularized first stage. However, as noted in Hansen and Kozbur (2014), among others, this direct approach would typically introduce a so-called regularization bias (similar to other methods involving regularization).[10] In general, the least shrunk coefficients correspond to the instruments that are most highly correlated with the first stage noise, thus contaminating the exclusion restriction. The use of sample-splitting or jackknifing is a common way of lowering the regularization bias. I employ the sample-splitting technique to preserve the exclusion restriction, thus defining

$$\hat{\beta}_{I_1}^{EN} = \arg\min_{\beta}\left\{\sum_{i\in I_1}\left(d_i - \beta_0 - \sum_{j=1}^{p} f_{ij}\beta_j\right)^2 + \lambda^{EN}\sum_{j=1}^{p}\left(\alpha\left|\beta_j\right| + (1-\alpha)\beta_j^2\right)\right\},$$

which is the elastic-net estimate from an elastic-net regression of d on f with regularization parameters $\left(\lambda^{EN}, \alpha\right)$ using the random subset of observations I_1 (a half of the sample, in the simplest

[10] See Chernozhukov et al (2018) for an extended discussion of the regularization bias and de-biased estimation.

case). The estimator $\widehat{D}_i$ for the i^{th} unit is then defined as $\widehat{D}_i = f_i' \hat{\beta}_{I_1}^{EN}$. Finally, I define the sample-split ENIV estimator as

$$\widehat{\delta}^{SS-ENIV} = \left(\sum_{i \in I_1^c} f_i' \hat{\beta}_{I_1}^{EN} d_i' \right)^{-1} \sum_{i \in I_1^c} f_i' \hat{\beta}_{I_1}^{EN} y_i,$$

where $I_1^c \cap I_1 = \emptyset$.

By splitting the sample into halves, I break the correlation between $\widehat{D}_i$ and e_i that is not asymptotically negligible. Although the elastic-net regularization causes some loss of signal due to coefficient shrinkage (similar to other regularization methods), a data-driven choice of $\left(\lambda^{EN}, \alpha\right)$ is expected to result in quality signal extraction from a high-dimensional set of instruments, whether sparse or dense. For example, for $\alpha = 0$ and positive λ^{EN}, the elastic-net IV estimator reduces to the ridge IV estimator. I suggest choosing the shrinkage parameter based on the optimization of a first stage cross-validation criterion due to popularity and availability of cross-validation tools in R, Python, Stata, etc.[11] In general, for not very large datasets one can replace a sample-splitting approach with a jackknifing procedure to fit the first stage, thus generalizing the sample-split ENIV estimator to the jackknife ENIV estimator.

Another possible approach is cross-fitting. Cross-fitting estimators are also based on the idea of sample-splitting. First, the sample is partitioned into I_1 and I_2, and only observations from I_1 are used to get $\hat{\beta}_{I_1}^{EN}$, whereas only observations from I_2 are used to produce $\widehat{\delta}_{12} = \left(\sum_{I_2} f_i' \hat{\beta}_{I_1}^{EN} d_i' \right)^{-1} \times \sum_{i \in I_2} f_i' \hat{\beta}_{I_1}^{EN} y_i$. Then the subsamples are swapped so that $\hat{\beta}_{I_2}^{EN}$ and $\widehat{\delta}_{21} = \left(\sum_{I_1} f_i' \hat{\beta}_{I_2}^{EN} d_i' \right)^{-1} \times \sum_{i \in I_1}^{N} f_i' \hat{\beta}_{I_2}^{EN} y_i$ are obtained in an analogous way. Consequently, the cross-fit elastic-net IV estimator is defined as $\widehat{\delta}^{CF-ENIV} = \left(\widehat{\delta}_{12} + \widehat{\delta}_{21} \right) / 2$. This way both subsamples (symmetrically) contribute to the resulting estimate, thus increasing its efficiency. I adopt the algorithm by Anatolyev and Mikusheva (2022, Section 3.2) to estimate the variance of $\widehat{\delta}^{CF-ENIV}$ in a way that accounts for the correlation between $\widehat{\delta}_{12}$ and $\widehat{\delta}_{21}$.[12] Finally, sample-split and cross-fit lasso-based IV estimators, which act as benchmarks in the following section, are defined analogously.

[11] The use of cross-validation is yet to be theoretically justified for elastic-net, despite being a widely spread practice. See Chetverikov, Liao and Chernozhukov (2021), which justifies the practice of using cross-validation to choose the penalty parameter for lasso.

[12] Anatolyev and Mikusheva (2022) propose the algorithms for constructing a four-split estimator. I use a version simplified to a case with only two splits.

3 Simulation study

The design of this simulation study closely follows that of Hansen and Kozbur (2014). I demonstrate the performance of the IV estimators employing elastic-net, and compare it with the performance of lasso-based IV estimators, and the ridge jackknife IV estimator (RJIVE) by Hansen and Kozbur (2014). Let the data generating process be

$$\begin{aligned} y_i &= x_i\delta_0 + e_i \\ x_i &= Z_i'\Pi + u_i \end{aligned}$$

with

$$(e_i, u_i) \sim N\left(0, \begin{pmatrix} \sigma_e^2 & \sigma_{eu} \\ \sigma_{eu} & \sigma_u^2 \end{pmatrix}\right),$$

where x_i is the scalar treatment variable, and $\delta_0 = 1$ is the parameter of interest. The sample size $n = 100$, $\sigma_e^2 = 2$, and $\text{corr}(e_i, u_i) = 0.6$. The remaining parameters are varied within the simulation study.

I consider two instrument designs: binary and continuous (Gaussian). Real datasets typically employ very different combinations of both binary and continuous instruments, thus motivating examination of the two extreme cases: (i) all instruments are binary, and (ii) all instruments are continuous. The continuous instrument design considers correlated Gaussian instruments drawn with mean 0 and variance $\text{var}(Z_{ij}) = 0.3$. The correlation between Gaussian instruments is given by $\text{corr}(Z_{ij}, Z_{ik}) = 0.8^{|j-k|}$. The binary design is motivated by the presence of many categorical variables, which often takes place in practice. In this design, all instruments are drawn from $Z_{ij} \in \{0, 1\}$ with $\Pr(Z_{ij} = 1) = 0.8$ such that the pairwise correlations are close to $\text{corr}(Z_{ij}, Z_{ik}) = 0.8^{|j-k|}$.[13] For each design, the number of instruments is set to $K = 95$ or $K = 190$.

In addition to alternation of the instrument design, I also vary the first-stage coefficients Π to generate dense, sparse, and mixed first-stage signal structures. In the dense scenario, $\Pi = (\iota_{0.4K}, 0_{0.6K})'$,

[13] First, I make draws from the standard normal distribution, and apply Cholecky's decomposition to generate the Gaussian instruments Z_{ij}^0 with correlations $\text{corr}\left(Z_{ij}^0, Z_{ik}^0\right) = 0.8^{|j-k|}$. Then I set $Z_{ij} = \mathbb{I}_{\left\{Z_{ij}^0 > 0.8\right\}}$.

where ι_p is a $1 \times p$ vector of ones, and 0_q is a $1 \times q$ vector of zeros. In the sparse scenario $\Pi = (3\iota_5, 0_{K-5})'$, so only five instruments are relevant. Finally, in the mixed scenario, $\Pi = (3\iota_5, \iota_{0.4K}, 0_{0.6K-5})'$. By varying the noise σ_u^2 in the first-stage regression, I control the strength of the instrument set measured by the concentration parameter $\mu^2 = N\Pi' E[Z_i' Z_i]\Pi/\sigma_u^2$. To model the cases of the weak and strong signal provided by the instruments, I set $\mu^2 = 30$ and $\mu^2 = 150$, respectively.

I consider three IV estimators based on elastic-net: elastic-net IV estimator (ENIV), sample-split elastic-net IV estimator (SS-ENIV), and cross-fit elastic-net IV estimator (CF-ENIV). Their lasso-based counterparts are Lasso-IV, SS-Lasso-IV, and CF-Lasso-IV. I also report the results for RJIVE and the 2SLS estimator. In addition, I present the results for the post-Lasso-IV estimator described in BCCH[14], as well as its sample-split version (SS-post-Lasso-IV). The penalty levels for ENIV, SS-ENIV, and CF-ENIV is chosen through cross-validation.

The reported results are obtained by averaging across 1500 draws for each setting. For each estimator, I present the median bias (Med. Bias), the median absolute deviation (MAD), and the rejection rate for a 5%-level test of $H_0 : \delta_0 = 1$ (RP 5%). For the post-Lasso estimator with lasso sometimes selecting no instruments into the first stage regression, I calculate the median bias and the median absolute deviation conditional on the lasso estimator selecting at least one variable. In such a case, a failure to reject the null is recorded.

Table 1 shows the results for $K = 95$. Panels A and B focus on the results for weak instruments ($\mu^2 = 30$), Panels C and D report the results for a stronger signal ($\mu^2 = 150$). For the weak sparse signal, Lasso-IV, post-Lasso-IV, RJIVE, SS-ENIV, and CF-ENIV result in reasonable rejection frequencies, with RJIVE and SS-ENIV being among the most accurate. However, for the dense weak signal, only RJIVE, SS-ENIV and CF-ENIV continue to have approximately the correct size (CF-ENIV tends to over-reject but not as much as the Lasso-based estimators).

[14]BCCH recommend the penalty level to be proportional to $\sqrt{n \log K}$. We employ the same penalty as in Hansen and Kozbur (2014), namely $2.2\sqrt{2n \log(2K)}\sigma_u\sigma_e$.

Table 1. Simulation Results many instruments K = 95

	Sparse Signal			Dense Signal			Mixed Signal		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
	A. Concentration parameter = 30. Binary Instruments								
Lasso-IV	0.009	0.015	0.091	0.017	0.018	0.237	0.012	0.013	0.201
SS-Lasso-IV	0.003	0.023	0.009	0.004	0.04	0.011	0.000	0.024	0.007
post-Lasso-IV	0.010	0.015	0.111	0.016	0.017	0.253	0.012	0.013	0.249
SS-post-Lasso-IV	0.003	0.023	0.008	0.004	0.038	0.013	0.000	0.024	0.009
CF-Lasso-IV	0.014	0.015	0.000	0.002	0.025	0.000	0.007	0.015	0
RJIVE	-0.001	0.020	0.047	-0.001	0.011	0.055	-0.001	0.010	0.052
ENIV	0.022	0.022	0.405	0.020	0.020	0.448	0.015	0.015	0.466
SS-ENIV	0.000	0.028	0.038	0.001	0.020	0.056	0.000	0.015	0.048
CF-ENIV	0.001	0.022	0.104	0.000	0.015	0.098	-0.001	0.012	0.095
	B. Concentration parameter = 30. Gaussian Instruments								
Lasso-IV	0.005	0.011	0.076	0.011	0.012	0.210	0.007	0.008	0.177
SS-Lasso-IV	0.002	0.016	0.031	0.002	0.030	0.005	0.002	0.015	0.012
post-Lasso-IV	0.007	0.011	0.104	0.011	0.012	0.224	0.008	0.009	0.215
SS-post-Lasso-IV	0.002	0.016	0.029	0.001	0.030	0.005	0.003	0.015	0.011
CF-Lasso-IV	0.004	0.010	0.001	0.004	0.022	0.000	0.006	0.009	0.000
RJIVE	-0.001	0.014	0.051	-0.002	0.010	0.041	0.000	0.008	0.053
ENIV	0.012	0.014	0.284	0.013	0.013	0.421	0.010	0.010	0.459
SS-ENIV	0.001	0.019	0.041	0.001	0.018	0.037	0.002	0.013	0.043
CF-ENIV	0.001	0.014	0.101	0.001	0.014	0.123	0.001	0.010	0.119
	C. Concentration parameter = 150. Binary Instruments								
Lasso-IV	0.005	0.014	0.065	0.012	0.014	0.133	0.008	0.010	0.130
SS-Lasso-IV	0.000	0.022	0.047	0.000	0.022	0.048	-0.001	0.016	0.043
post-Lasso-IV	0.005	0.014	0.068	0.013	0.014	0.155	0.009	0.010	0.144
SS-post-Lasso-IV	-0.001	0.022	0.047	0.001	0.020	0.047	0.000	0.016	0.047
CF-Lasso-IV	-0.001	0.016	0.000	-0.001	0.016	0.000	-0.001	0.013	0.000
RJIVE	-0.002	0.017	0.052	0.000	0.011	0.063	-0.001	0.009	0.063
ENIV	0.012	0.016	0.149	0.016	0.016	0.218	0.012	0.012	0.233
SS-ENIV	0.000	0.022	0.052	0.001	0.016	0.060	-0.001	0.013	0.057
CF-ENIV	-0.001	0.015	0.054	0.000	0.012	0.053	-0.001	0.010	0.071
	D. Concentration parameter = 150. Gaussian Instruments								
Lasso-IV	0.002	0.010	0.064	0.010	0.011	0.175	0.005	0.007	0.113
SS-Lasso-IV	0.000	0.015	0.058	0.000	0.02	0.045	-0.001	0.012	0.048
post-Lasso-IV	0.004	0.010	0.076	0.010	0.011	0.186	0.006	0.007	0.145
SS-post-Lasso-IV	0.000	0.015	0.057	0.001	0.018	0.045	-0.001	0.012	0.048
CF-Lasso-IV	-0.001	0.011	0.006	0.000	0.014	0.000	0.000	0.009	0.002
RJIVE	0.000	0.011	0.057	0.000	0.009	0.065	-0.001	0.006	0.055
ENIV	0.008	0.011	0.132	0.012	0.012	0.251	0.008	0.009	0.225
SS-ENIV	0.000	0.015	0.054	0.001	0.013	0.060	0.000	0.010	0.047
CF-ENIV	-0.001	0.011	0.056	0.000	0.010	0.079	0.000	0.007	0.070

Note: Results are based on 1500 simulation replications. I report Median Bias (Med. Bias), Median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for nine different estimators: the Lasso IV and post-Lasso IV estimators of Belloni et al. (2012, Lasso-IV and post-Lasso-IV), their sample-split versions (SS-Lasso-IV and SS-post-Lasso-IV), the cross-fit Lasso IV estimator, the RJIVE by Hansen and Kozbur (2014, RJIVE), and three estimators proposed in this paper: the elastic-net IV estimator (ENIV), the sample-split elastic-net IV estimator (SS-ENIV) and the cross-fit elastic-net IV estimator (CF-ENIV).

For the mixed design, only RJIVE and SS-ENIV deliver accurate test size. Overall, SS-ENIV tends to produce more precise rejection rates when the true non-zero coefficients on the instruments vary in magnitude (the case of a mixed signal), compared to the case of the equal coefficient magnitude[15], which is often examined as part of simulation exercises in the literature (e.g. in Hansen and Kozbur, 2014, among others). In practice, there is often no good reason to expect a signal to be evenly distributed across all instruments that explain a decent share of variance in x_i, the treatment variable. Whereas RJIVE tends to result in rejection frequencies slightly above the nominal test size, the opposite is true for SS-ENIV.

With a strong sparse signal, most Lasso-based estimators produce adequate rejection frequencies, as expected. RJIVE, SS-ENIV and CF-ENIV retain rather accurate test size irrespective of the data structure when the signal is strong. Notably, CF-ENIV performs better with strong signals (sparse, dense, or mixed) than with weak signals. The SS-ENIV estimator proves to be a good alternative to RJIVE when dealing with a strong mixed signal, similarly to the case of a weak mixed signal discussed above.

Table 2 shows the results for $K = 190$. Panels A and B again focus on the results for weak instruments ($\mu^2 = 30$), Panels C and D report the results for a stronger signal ($\mu^2 = 150$). For the weak sparse signal, some Lasso-based estimators have reasonable rejection frequencies, although RJIVE and SS-ENIV tend to be superior in terms of bias and rejection rate, irrespective of sparsity. With the weak signal and mixed data structure, RJIVE and SS-ENIV perform similarly, although the sample-split elastic-net IV estimator seems to be more prone to under-rejection. With the strong sparse signal, Lasso-based estimators (Lasso-IV, SS-Lasso-IV, post-Lasso-IV, SS-post-Lasso-IV) most often result in relatively adequate rejection frequencies, the same holds for RJIVE, SS-ENIV, and CF-ENIV. With the strong mixed signal, binary or Gaussian, SS-ENIV tends to produce slightly lower rejection frequencies than RJIVE, including the case of Gaussian instruments when both estimators slightly over-reject.

[15]All first-stage variables are standardized before ridge/lasso/elastic-net estimation is performed.

Table 2. Simulation Results many instruments K = 190

	Sparse Signal			Dense Signal			Mixed Signal		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
	A. Concentration parameter = 30. Binary Instruments								
Lasso-IV	0.010	0.015	0.103	0.016	0.016	0.329	0.013	0.013	0.290
SS-Lasso-IV	0.001	0.032	0.009	0.007	0.038	0.003	0.004	0.025	0.001
post-Lasso-IV	0.010	0.015	0.120	0.015	0.015	0.359	0.013	0.013	0.333
SS-post-Lasso-IV	0.001	0.032	0.008	0.005	0.039	0.003	0.003	0.026	0.002
CF-Lasso-IV	0.025	0.025	0.000	0.009	0.027	0.000	0.006	0.012	0.000
RJIVE	0.000	0.025	0.042	0.000	0.009	0.043	0.000	0.008	0.048
ENIV	0.026	0.026	0.498	0.018	0.018	0.720	0.016	0.016	0.729
SS-ENIV	0.001	0.034	0.037	0.000	0.015	0.043	0.002	0.013	0.043
CF-ENIV	0.001	0.026	0.132	0.000	0.012	0.108	0.001	0.010	0.124
	B. Concentration parameter = 30. Gaussian Instruments								
Lasso-IV	0.005	0.011	0.074	0.010	0.01	0.275	0.007	0.008	0.231
SS-Lasso-IV	0.000	0.016	0.031	0.010	0.025	0.001	0.002	0.019	0.005
post-Lasso-IV	0.007	0.011	0.124	0.010	0.01	0.315		0.008	0.274
SS-post-Lasso-IV	0.000	0.016	0.030	0.011	0.023	0.001	0.002	0.019	0.005
CF-Lasso-IV	0.005	0.010	0.000	0.015	0.015	0.000	0.007	0.009	0.000
RJIVE	-0.001	0.017	0.052	0.000	0.008	0.050	-0.001	0.007	0.042
ENIV	0.013	0.015	0.348	0.011	0.011	0.688	0.009	0.009	0.641
SS-ENIV	0.001	0.019	0.044	0.002	0.013	0.045	0.001	0.013	0.025
CF-ENIV	0.001	0.015	0.110	0.002	0.011	0.129	0.001	0.010	0.141
	C. Concentration parameter = 150. Binary Instruments								
Lasso-IV	0.005	0.014	0.069	0.011	0.012	0.208	0.010	0.01	0.180
SS-Lasso-IV	-0.001	0.021	0.049	0.000	0.023	0.039	0.001	0.018	0.029
post-Lasso-IV	0.005	0.014	0.074	0.012	0.012	0.235	0.010	0.011	0.217
SS-post-Lasso-IV	0.000	0.021	0.051	0.001	0.020	0.042	0.001	0.016	0.032
CF-Lasso-IV	0.001	0.015	0.001	0.001	0.016	0.002	0.000	0.014	0.001
RJIVE	-0.001	0.018	0.053	0.000	0.008	0.059	0.000	0.007	0.049
ENIV	0.015	0.018	0.201	0.015	0.015	0.400	0.014	0.014	0.430
SS-ENIV	0.001	0.021	0.052	0.000	0.012	0.061	0.000	0.010	0.042
CF-ENIV	0.000	0.015	0.047	0.000	0.009	0.053	0.000	0.007	0.055
	D. Concentration parameter = 150. Gaussian Instruments								
Lasso-IV	0.003	0.010	0.055	0.009	0.010	0.239	0.007	0.008	0.207
SS-Lasso-IV	-0.001	0.016	0.043	0.002	0.018	0.021	0.000	0.014	0.036
post-Lasso-IV	0.004	0.010	0.060	0.010	0.010	0.274	0.008	0.008	0.259
SS-post-Lasso-IV	0.000	0.015	0.043	0.001	0.016	0.024	0.000	0.014	0.033
CF-Lasso-IV	-0.001	0.011	0.002	0.002	0.015	0.001	0.000	0.010	0.002
RJIVE	0.000	0.012	0.045	-0.001	0.006	0.053	0.000	0.006	0.064
ENIV	0.009	0.012	0.146	0.012	0.012	0.440	0.010	0.010	0.444
SS-ENIV	0.000	0.016	0.037	0.000	0.010	0.039	0.000	0.009	0.053
CF-ENIV	0.000	0.011	0.046	0.000	0.007	0.083	0.000	0.006	0.088

Note: Results are based on 1500 simulation replications. I report Median Bias (Med. Bias), Median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for nine different estimators: the Lasso IV and post-Lasso IV estimators of Belloni et al. (2012, Lasso-IV and post-Lasso-IV), their sample-split versions (SS-Lasso-IV and SS-post-Lasso-IV), the cross-fit Lasso IV estimator, the RJIVE by Hansen and Kozbur (2014, RJIVE), and three estimators proposed in this paper: the elastic-net IV estimator (ENIV), the sample-split elastic-net IV estimator (SS-ENIV) and the cross-fit elastic-net IV estimator (CF-ENIV).

To sum up the results of the simulation study, the IV estimators based on elastic-net constitute a safe alternative to those based on lasso under an unknown degree of sparsity. In particular, the sample-split elastic-net IV estimator tends to dominate its lasso-based counterpart, the sample-split lasso IV estimator, as well as other lasso-based IV estimators, in terms of bias and test accuracy. In addition, the performance of the sample-split elastic-net IV estimator is comparable to that of the ridge jackknife IV estimator. SS-ENIV tends to result in slightly lower rejection frequencies than RJIVE, thus being superior in the settings when both estimators over-reject. RJIVE shows minor over-rejection in most settings considered with the mixed signal, thereby motivating further investigation of the relative performance of RJIVE and SS-ENIV estimators in various settings with uneven distribution of explanatory power across the instrumental variables. Finally, data generating processes with alternative degrees of sparsity are also worth examining.

Figure 1 presents frequency plots for the penalty ratio from first-stage regressions estimated via elastic-net. The elastic-net penalty ratio is $a/(a+b)$ where a and b come from representing the elastic-net penalty term $\lambda\left(\alpha|\beta_j| + (1-\alpha)\beta_j^2\right)$ as $a|\beta_j| + b\beta_j^2$. The penalty ratio is chosen through cross-validation.[16] For the ratio 1.0 the penalty is an l_1-penalty (lasso-type), whereas for the ratio 0.0 it is an l_2-penalty (ridge-type). A combination of both l_1 and l_2 penalties is employed when the cross-validation procedure results in a value between 0 and 1. The results for a sparse, dense, and mixed DGP are shown in the first, second and third column of plots, respectively. As before, panels A, B, C and D correspond to various instrument designs. Only the case with $p = 95$ is presented, since the results for the case with $p = 190$ look very similar.

When fitting the right combination of both l_1 and l_2 penalties to a first-stage relationship, the elastic-net estimator is quite successful in detecting a sparse structure, and thus often sets the penalty ratio to 1 in this case. When dealing with non-sparse first-stage relationships, the distribution of the penalty ratio is more even, with massive point mass on 0 and 1, and also on the

[16] I use a Python package, sklearn.linear_model.ElasticNetCV, to fit the first-stage via elastic-net, with a prespecified grid [0.01, 0.03, .05, .07, .1, .2, .5, .8, .9, 0.93, .95, 0.97, .99, 1]. For each value of the penalty ratio, the grid for a parameter α, which is also estimated through cross-validation, consists of 100 values and is defined automatically as part of the ElasticNetCV package.

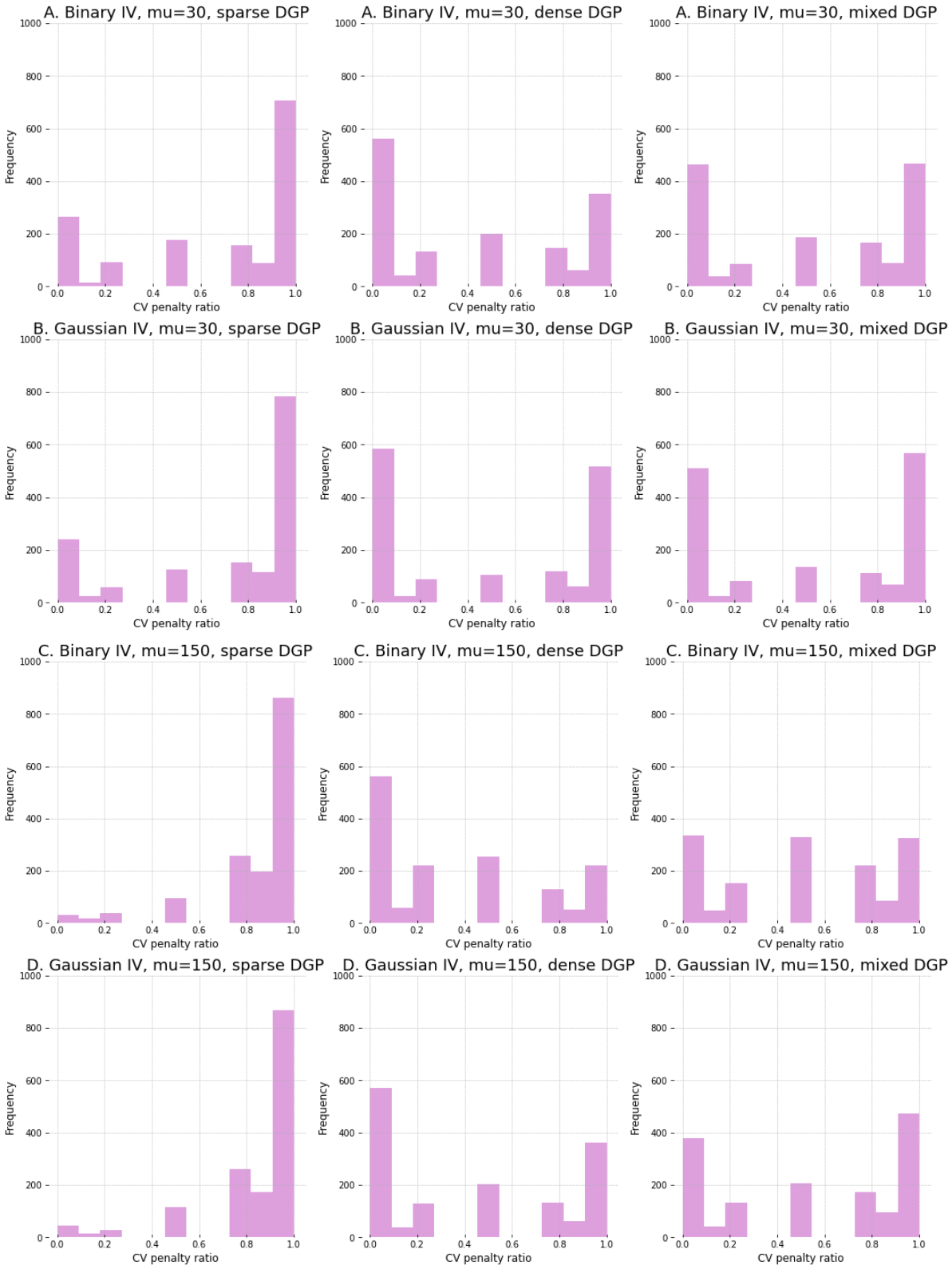

Figure 1: The penalty ratio chosen through cross-validation as part of the first-stage elastic-net regression. Cross-validation is performed on a grid from 0 to 1. Graphs show the frequency of each value being selected. For the penalty ratio 1 the penalty is an l_1-penalty; for the penalty ratio 0 it is an l_2-penalty; for the penalty ratio between 0 and 1 it is a combination of both. The case with $p = 95$ instruments and $n = 100$ observations is presented.

intermediate values if the signal is strong ($\mu = 150$). Thus, the elastic-net estimator is performing better in combining l_1 and l_2 penalties when facing a strong signal, whereas it tends to often converge to a corner solution (imposing no ridge-type penalty, or no lasso-type penalty at all) when dealing with a weak signal ($\mu = 30$). In addition, the graphs presented indicate the need for a finer grid to search over for the best penalty ratio (especially around the middle value), for a better fit to the unknown sparsity of the data at hand.

4 Empirical Example

In this section, I demonstrate the application of the EN-based IV estimators to the classic example from the many-instrument literature – Angrist and Krueger (1991). The coefficient of interest in this example is the causal effect of schooling on earnings, and the schooling endogeneity is addressed through the use of instrumental variables. The data from Angrist and Krueger (1991) potentially allow one to employ many instruments for identification of the treatment effect, and there is a rich literature on consequences of alternative IV-choice decisions, in terms of both point estimate's and inference quality, driven by the numerosity and weakness of the available instrumental variables (Bound, Jaeger and Baker, 1995; Angrist, Imbens and Krueger, 1999; Staiger and Stock, 1997; Hansen, Hausman and Newey, 2008).

The simple model under consideration is

$$\begin{aligned} \log(\text{wage}_i) &= \alpha\, \text{Schooling}_i + W_i'\gamma + \varepsilon_i \\ \text{Schooling}_i &= Z_i'\Pi_1 + W_i'\Pi_2 + u_i \end{aligned}$$

where ε_i and u_i satisfy $E[\varepsilon_i | W_i, Z_i] = E[u_i | W_i, Z_i] = 0$, $\log(\text{wage}_i)$ is a log of individual wage, Schooling_i is individual years of completed schooling, W_i is a vector of control variables and Z_i is a vector of instrumental variables that affect the wage only through the education channel. The data come from the 1980 U.S. Census and represent 329,509 men born between 1930 and 1939. The control set consists of 510 variables: a constant, 9 year-of-birth dummies, 50 state-of-birth dummies and 450 state-of-birth $\times$ year-of-birth cross-products. I employ three alternative sets of

instruments, varying from three quarter-of-birth dummies to a full set of interactions with state-of-birth and year-of-birth control variables W_i, i.e. a total of 1,527 instrumental variables. By the identification argument of Angrist and Krueger (1991), α, the IV coefficient on Schooling$_i$, is a causal effect of education on earnings.

I report the results for three instrument sets in Table 3. For each set of instrumental variables, I present the estimates from conventional 2SLS, post-Lasso, SS-post-Lasso, ENIV, SS-ENIV, and CF-ENIV. For the estimators involving sample-splitting, I report two estimates (separated by / in Table 3) that result from swapping the sample halves used for fitting the first stage. This way I demonstrate the sensitivity of the point estimates that takes place despite the large sample at hand.

Table 3.

	2SLS	post-Lasso	SS-post-Lasso	RJIVE	ENIV	SS-ENIV	CF-ENIV
			A. 3 instruments				
Coefficient	0.108	0.111	0.097 / 0.112	0.109	0.108	0.098 / 0.118	0.108
St. error	0.020	0.0205	0.034 / 0.039	0.020	0.020	0.027 / 0.029	0.020
			B. 180 instruments				
Coefficient	0.093	0.112	0.097 / 0.112	0.106	0.093	0.103 / 0.114	0.108
St. error	0.010	0.017	0.034 / 0.039	0.016	0.010	0.026 / 0.027	0.009
			C. 1527 instruments				
Coefficient	0.071	0.086	0.097	0.107	0.074	0.079 / 0.145	0.112
St. error	0.005	0.025	0.039	0.017	0.005	0.061 / 0.064	0.004

Panel A uses the three main quarter-of-birth dummies from Angrist and Krueger (1991). As expected, all estimators considered result in similar point estimates and standard errors. Due to the high strength of each of the small number of instrumental variables being used, the methods involving regularization impose a small regularization penalty, thus leading to nearly identical results as 2SLS.

Panel B employs 180 instruments including the three quarter-of-birth dummies and their cross-products with the 9 year-of-birth dummies and 50 state-of-birth dummies. This set is also used in Angrist and Krueger (1991), with the aim of increasing the efficiency of the estimates. As expected, the 2SLS estimate is biased toward the OLS estimate of 0.0673. The same applies to ENIV that actually employs approximately as many instruments as 2SLS does. Post-Lasso, SS-

post-Lasso, SS-ENIV, and CF-ENIV tend to deliver adequate estimates, though the instability of the estimators involving sample splitting is noticeable. The post-Lasso estimator does not have a downward bias, while CF-ENIV results in the smallest estimated standard error.

In Panel C, I show results based on the full set of 1527 instrumental variables. Even stronger bias of the 2SLS estimate towards the OLS estimate is observed. In this case, the SS-post-Lasso estimator tends to select no variables into the first stage regression (therefore, only a single number is provided). The post-Lasso, SS-post-Lasso, ENIV estimators now also result in a substantial downward bias. However, the CF-ENIV still delivers a reasonable point estimate, and also the smallest estimated standard error as well.

Conclusion

In this paper, I propose elastic-net instrumental variable estimators to deal with high-dimensional sets of instruments. The proposed estimators can be asymptotically equivalent to the lasso-based IV estimators but have better sampling properties if correlations among the instruments are not negligible. In addition, the IV estimators based on elastic-net are robust to deviations of the first-stage regression from sparsity. These features make the elastic-net IV estimators a valuable alternative to the lasso IV estimators for policy evaluation.

References

Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2(2), 105-110.

Anatolyev, S., & Mikusheva, A. (2022). Factor models with many assets: strong factors, weak factors, and the two-pass procedure. *Journal of Econometrics*, 229(1), 103-126.

Anderson, T. W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics*, 127(1), 1-16.

Anderson, T. W., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1), 46-63.

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

Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1), 57-67.

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

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

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

Belloni, A., Chernozhukov, V., & Hansen, C. (2011). Lasso methods for Gaussian instrumental variables models. *MIT Department of Economics Working Paper* No. 11-14. Available at SSRN: https://ssrn.com/abstract=1908409

Belloni, A., Chernozhukov, V., & Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4), 606-619.

Bickel, P. J., Ritov, Y. A., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4), 1705-1732.

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

Buehlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2), 559-583.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2), 383-398.

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

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K., & Woutersen, T. (2012). Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric*

Theory, 28(1), 42-86.

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

Chetverikov, D., Liao, Z., & Chernozhukov, V. (2021). On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3), 1300-1317.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., & Picard, D. (1995). Wavelet shrinkage: asymptopia?. *Journal of the Royal Statistical Society: Series B*, 57(2), 301-337.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348-1360.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5), 849-911.

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1), 181-213.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Ghosh, S. (2011). On the grouped selection and model complexity of the adaptive elastic net. *Statistics and Computing*, 21(3), 451-462.

Giannone, D., Lenza, M., & Primiceri, G. E. (2021) Economic Predictions with Big Data: The Illusion of Sparsity. *Econometrica*, 89(5), 2409-2437.

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

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

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Jia, J., & Yu, B. (2010). On model selection consistency of the elastic net when p>>n. *Statistica Sinica*, 595-611.

Meinshausen, N., and Bühlmann, P. (2004), Variable Selection and High-Dimensional Graphs With the Lasso. Technical report, ETH Zürich.

Newey, W. (1990). Efficient Instrumental Variables Estimation of Nonlinear Models. *Econometrica*, 58(4), 809-37.

Newey, W. K., & Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1), 219-255.

Spiess, J. (2017). Bias reduction in instrumental variable estimation through first-stage shrinkage. *arXiv preprint* arXiv:1708.06443.

Staiger, D., & Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3), 557-586.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *The Journal of Urology*, 141(5), 1076-1083.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR* (Vol. 39, pp. 195-198).

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.

Yuan, M., & Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 143-161.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association,* 101(476), 1418-1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 67(2), 301-320.

Appendix

Proof of Proposition 1.

Lemma 1 from Zou and Hastie (2006) shows that the naive elastic-net criterion

$$L\left(\lambda_1^{EN}, \lambda_2^{EN}, \beta\right) = |y - X\beta|^2 + \lambda_1^{EN} |\beta|_1 + \lambda_2^{EN} |\beta|_2$$

can be written as the lasso criterion

$$L\left(\gamma, \beta^*\right) = \left|y^* - X^*\beta^*\right|^2 + \gamma \left|\beta^*\right|_1,$$

where $\gamma = \lambda_1^{EN}/\sqrt{1+\lambda_2^{EN}}$, $\beta^* = \sqrt{1+\lambda_2^{EN}}\beta$, and an augmented data set (y^*, X^*) is defined by

$$X^*_{(n+p)\times p} = \left(1+\lambda_2^{EN}\right)^{-1/2} \begin{pmatrix} X \\ \sqrt{\lambda_2^{EN}} I \end{pmatrix}, \qquad y^*_{(n+p)} = \begin{pmatrix} y \\ 0 \end{pmatrix}.$$

Then, for $\widehat{\beta}^* = \arg\min_\beta L\left(\gamma, \beta^*\right)$,

$$\widehat{\beta}^{EN} = \frac{1}{\sqrt{1+\lambda_2^{EN}}} \widehat{\beta}^*.$$

Having the elastic-net problem represented as the lasso problem, we can directly apply the results from Corollary 1 by BCCH on lasso's convergence rates under non-Gaussian and heteroskedastic errors. For a properly chosen γ,

$$\left\| \widehat{D}_i^* - D_i^* \right\|_{2,n} \lesssim_P \sqrt{\frac{s \log\left(p \vee (n+p)\right)}{n+p}} = \sqrt{\frac{s \log\left(n+p\right)}{n+p}}$$

and therefore,

$$\left\| \widehat{D}_i^{EN} - D_i \right\|_{2,n} \lesssim_P \sqrt{\frac{s \log\left(n+p\right)}{n+p}}.$$

Similarly, using the second inequality from Corollary 1,

$$\left\| \widehat{\beta}^* - \beta^* \right\|_1 \lesssim_P \sqrt{\frac{s^2 \log\left(n+p\right)}{n+p}},$$

and it can be written as

$$\left\|\widehat{\beta}^{EN} - \beta\right\|_1 \lesssim_P \frac{1}{\sqrt{1+\lambda_2^{EN}}}\sqrt{\frac{s^2 \log(n+p)}{n+p}} \leq \sqrt{\frac{s^2 \log(n+p)}{n+p}},$$

thus giving us a sufficient condition for Theorem 4 by BCCH to hold.

Abstrakt

Instrumentální proměnné (IV) se běžně používají pro identifikaci účinků treatmentu a následné vyhodnocení politiky. Použití mnoha informativních instrumentů zlepšuje přesnost odhadu. Užití mnohorozměrných sad instrumentálních proměnných neznámé síly však může být komplikované a vyžaduje výběr modelu nebo regularizaci regrese v prvním stupni. V současné době je lasso zavedeno jako jedna z nejpopulárnějších regularizačních technik, která se opírá o předpoklad přibližné řídkosti. Zkoumám relativní výkon odhadů lassa a elastických sítí (elastic net) pro predikované hodnoty prvního stupně jako součást odhadu IV. Jelikož elastická síť obsahuje kromě penalizace typu lasso penalizaci hřebenového typu, obecně se oproti lassu v konečných vzorcích zlepšuje, když korelace mezi instrumentálními proměnnými nejsou zanedbatelné. Ukazuji, že IV odhady založené na odhadech lasa a elastické sítě v prvním stupni mohou být asymptoticky ekvivalentní. Prostřednictvím Monte Carlo studie demonstruji robustnost estimátoru elastic net IV s rozděleným vzorkem dat vůči odchylkám od přibližné řídkosti a vůči korelaci mezi potenciálně mnohorozměrnými instrumenty. Nakonec uvádím empirický příklad, který demonstruje potenciální zlepšení přesnosti odhadu získané použitím IV odhadů založených na elastické síti.

Working Paper Series
ISSN 2788-0443

Individual researchers, as well as the on-line version of the CERGE-EI Working Papers (including their dissemination) were supported from institutional support RVO 67985998 from Economics Institute of the CAS, v. v. i.

Specific research support and/or other grants the researchers/publications benefited from are acknowledged at the beginning of the Paper.

Published by
Charles University, Center for Economic Research and Graduate Education (CERGE)
and
Economics Institute of the CAS, v. v. i. (EI)
CERGE-EI, Politických vězňů 7, 111 21 Prague 1, tel.: +420 224 005 153, Czech Republic.
Phone: + 420 224 005 153
Email: office@cerge-ei.cz
Web: https://www.cerge-ei.cz/

Editor: Byeongju Jeong

The paper is available online at https://www.cerge-ei.cz/working-papers/.

ISBN 978-80-7343-566-0 (Univerzita Karlova, Centrum pro ekonomický výzkum a doktorské studium)
ISBN 978-80-7344-686-4 (Národohospodářský ústav AV ČR, v. v. i.)