Fourier studies

ARTICLE

Developing Robust Inference Methods for Instrumental Variables in the Context of Machine Learning and Big Data Environments

Milosz Kowalczyk[1] **and Anzhela Draganova**[2]

[1] University of Podlasie, Department of Computer Systems and Security, 14 Nowomiejska Street, Białystok, Poland
[2] Veliko Tarnovo Technical University, Department of Intelligent Systems, 8 Sveta Gora Boulevard, Veliko Tarnovo, Bulgaria

Abstract

Instrumental variables (IV) estimation has emerged as a useful methodology for causal inference in econometrics, addressing the persistent challenge of endogeneity in observational data where unobserved confounders bias traditional regression estimates. The integration of machine learning techniques with instrumental variables estimation presents both unprecedented opportunities and significant methodological challenges, particularly in high-dimensional settings where the number of potential instruments may exceed sample sizes and where traditional asymptotic theory may not apply. This paper develops a comprehensive framework for robust instrumental variables inference in machine learning environments, introducing novel regularization techniques that simultaneously address the problems of weak instruments, many instruments, and high-dimensional confounding. We establish theoretical foundations for our proposed estimators by deriving finite-sample concentration inequalities and asymptotic normality results under heteroskedastic and potentially non-Gaussian error structures. Our methodology incorporates advanced matrix completion techniques and sparse regularization methods to handle missing data patterns commonly encountered in big data applications. Through extensive theoretical analysis involving sophisticated tools from empirical process theory and high-dimensional probability, we demonstrate that our proposed estimators achieve optimal rates of convergence while maintaining valid statistical inference properties. The practical implementation of our methods is illustrated through comprehensive simulation studies that demonstrate substantial improvements in both bias reduction and confidence interval coverage compared to existing approaches, with particular emphasis on scenarios involving weak identification and high-dimensional nuisance parameters.

1 Introduction

The problem of causal inference from observational data represents one of the most fundamental challenges in empirical research, spanning disciplines from economics and epidemiology to machine learning and social sciences [1]. Traditional regression-based approaches for estimating causal effects suffer from the critical limitation that correlation does not imply causation, particularly when unobserved confounding variables simultaneously influence both the treatment and outcome variables of interest. This endogeneity problem renders ordinary least squares and other standard estimation techniques inconsistent, leading to biased and potentially misleading conclusions about causal relationships.

Instrumental variables estimation has evolved as the primary methodological solution to address endogeneity concerns, providing a framework for consistent causal inference under specific identifying assumptions. The instrumental variables approach exploits exogenous variation in an instrumental variable that affects the outcome only through its

influence on the endogenous explanatory variable, thereby isolating the causal effect of interest from confounding influences. However, the classical instrumental variables framework faces significant challenges when applied to modern big data environments characterized by high-dimensional covariate spaces, complex dependency structures, and massive sample sizes that strain traditional asymptotic approximations.

The intersection of machine learning and causal inference has generated substantial methodological innovation in recent years, with researchers developing sophisticated algorithms that can handle high-dimensional settings while maintaining statistical rigor. Machine learning techniques offer powerful tools for modeling complex relationships and handling large-scale data, but their application to causal inference requires careful consideration of identification assumptions and statistical inference properties. The challenge becomes particularly acute in instrumental variables settings, where the performance of traditional estimators can deteriorate rapidly in high-dimensional environments due to the curse of dimensionality and the prevalence of weak instruments.

Modern applications of instrumental variables estimation frequently encounter scenarios where the number of potential instruments is large relative to the sample size, leading to the many instruments problem that can severely compromise the finite-sample performance of traditional two-stage least squares estimators [2]. Simultaneously, the presence of high-dimensional confounding variables necessitates sophisticated regularization techniques to avoid overfitting and ensure reliable inference. These challenges are further compounded by the frequent occurrence of weak instruments, where the correlation between instruments and endogenous variables is sufficiently small that standard asymptotic theory provides poor approximations to finite-sample behavior.

This paper addresses these interconnected challenges by developing a unified framework for robust instrumental variables inference in machine learning environments. Our approach integrates recent advances in high-dimensional statistics, empirical process theory, and machine learning to create estimators that simultaneously handle weak instruments, many instruments, and high-dimensional confounding while maintaining valid statistical inference properties. The theoretical foundation of our methodology rests on sophisticated concentration inequalities and uniform convergence results that extend classical instrumental variables theory to high-dimensional settings.

Our contribution to the literature is multifaceted, encompassing both theoretical innovations and practical methodological advances. From a theoretical perspective, we establish novel finite-sample concentration bounds for our proposed estimators under minimal distributional assumptions, extending beyond traditional Gaussian settings to accommodate the heavy-tailed distributions frequently encountered in real-world applications. We derive asymptotic normality results that enable valid confidence interval construction and hypothesis testing, even in challenging scenarios involving weak identification and high-dimensional nuisance parameters.

The methodological innovations presented in this paper include the development of adaptive regularization techniques that automatically adjust to the strength of available instruments and the dimensionality of the problem. Our approach incorporates matrix completion methods to handle missing data patterns and employs sophisticated cross-validation procedures to select tuning parameters in a data-driven manner [3]. The resulting estimators demonstrate superior performance compared to existing methods across a wide range of simulation scenarios, with particular advantages in settings involving weak instruments and high-dimensional confounding.

2 Theoretical Framework and Mathematical Foundations

The instrumental variables problem in high-dimensional settings requires a sophisticated mathematical framework that can accommodate the complex dependency structures and statistical challenges inherent in modern big data applications. We begin by establishing the fundamental setup and notation that will be used throughout our theoretical development.

Consider a structural equation model of the form $Y = D\beta + X'\gamma + U$, where $Y \in \mathbb{R}$ represents the outcome variable of interest, $D \in \mathbb{R}$ is the endogenous explanatory variable, $X \in \mathbb{R}^{p_x}$ is a vector of exogenous control variables, and U represents the unobserved error term that is potentially correlated with D. The parameter β represents the causal effect

of interest, while $\gamma \in \mathbb{R}^{p_x}$ captures the effects of the control variables. The endogeneity problem arises when $\mathbb{E}[DU] \neq 0$, rendering ordinary least squares estimation of β inconsistent.

To address this endogeneity concern, we assume access to a vector of instrumental variables $Z \in \mathbb{R}^{p_z}$ that satisfy the fundamental instrumental variables assumptions: relevance ($\mathbb{E}[ZD] \neq 0$) and exogeneity ($\mathbb{E}[ZU] = 0$). The first-stage relationship between the endogenous variable and instruments is characterized by the equation $D = Z'\pi + X'\delta + V$, where $\pi \in \mathbb{R}^{p_z}$ represents the first-stage coefficients, $\delta \in \mathbb{R}^{p_x}$ captures the effects of control variables, and V is the first-stage error term with $\mathbb{E}[ZV] = 0$.

The high-dimensional setting introduces significant complications to this classical framework, particularly when the dimensions p_x and p_z are large relative to the sample size n. In such settings, traditional two-stage least squares estimation becomes infeasible or exhibits poor finite-sample performance due to overfitting and the curse of dimensionality. Our theoretical framework addresses these challenges by incorporating sparsity assumptions and regularization techniques that enable consistent estimation and valid inference.

We assume that the true parameter vectors γ and π exhibit approximate sparsity, meaning that most elements are zero or sufficiently small that they can be effectively treated as zero for estimation purposes. Formally, we assume that $\|\gamma\|_0 \leq s_\gamma$ and $\|\pi\|_0 \leq s_\pi$, where $\|\cdot\|_0$ denotes the ℓ_0 norm (number of non-zero elements) and s_γ, s_π represent the sparsity levels. This sparsity assumption is crucial for enabling consistent estimation in high-dimensional settings and reflects the common belief that only a subset of available variables are truly relevant for the relationship of interest.

The statistical analysis of our proposed estimators relies heavily on concentration inequalities and uniform convergence results from empirical process theory [4]. Let $\mathcal{F}$ denote a function class of interest, and define the empirical process $\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(Z_i, X_i, D_i, Y_i) - \mathbb{E}[f(Z, X, D, Y)])$ for $f \in \mathcal{F}$. Our theoretical results require establishing uniform bounds on $\sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)|$, which necessitates careful analysis of the complexity of the function class $\mathcal{F}$.

The metric entropy of the function class $\mathcal{F}$ plays a crucial role in our theoretical analysis. For a given norm $\|\cdot\|$ and radius $\epsilon > 0$, the covering number $N(\epsilon, \mathcal{F}, \|\cdot\|)$ represents the minimum number of balls of radius ϵ needed to cover $\mathcal{F}$. The metric entropy is defined as $H(\epsilon, \mathcal{F}, \|\cdot\|) = \log N(\epsilon, \mathcal{F}, \|\cdot\|)$. Our theoretical results establish that under appropriate conditions on the metric entropy, the empirical process $\mathbb{G}_n$ converges uniformly to zero at the optimal rate.

A key component of our theoretical framework involves the analysis of the Gram matrices associated with the instrumental variables and control variables. Define the matrices $\Sigma_{ZZ} = \mathbb{E}[ZZ']$, $\Sigma_{XX} = \mathbb{E}[XX']$, and $\Sigma_{ZX} = \mathbb{E}[ZX']$. The eigenvalue properties of these matrices, particularly their minimum and maximum eigenvalues, play a crucial role in determining the performance of our estimators. We assume that these matrices satisfy restricted eigenvalue conditions that ensure stable inversion in high-dimensional settings.

The restricted eigenvalue condition for the matrix Σ_{ZZ} requires that for some constant $\kappa > 0$ and all vectors $v \in \mathbb{R}^{p_z}$ with $\|v\|_0 \leq s$, we have $v'\Sigma_{ZZ}v \geq \kappa\|v\|_2^2$. This condition ensures that the Gram matrix remains well-conditioned when restricted to sparse subspaces, enabling consistent estimation of sparse parameters. Similar conditions are imposed on Σ_{XX} and the joint covariance structure.

Our theoretical analysis also requires careful treatment of the error terms U and V in the structural and first-stage equations. We allow for heteroskedastic error structures where $\mathbb{E}[U^2|Z, X, D]$ and $\mathbb{E}[V^2|Z, X]$ may depend on the covariates in complex ways. This generalization beyond the classical homoskedastic setting is crucial for practical applications where error variances typically vary across observations.

The moment conditions that define our estimators can be expressed in terms of the sample analogues of the population orthogonality conditions. Define the moment function $m(Z_i, X_i, D_i, Y_i; \theta) = Z_i(Y_i - D_i\beta - X_i'\gamma)$ for the parameter vector $\theta = (\beta, \gamma')'$. The population moment condition is $\mathbb{E}[m(Z, X, D, Y; \theta_0)] = 0$, where θ_0 represents the true parameter value. Our estimators are defined as solutions to regularized versions of the sample moment conditions $\frac{1}{n}\sum_{i=1}^{n} m(Z_i, X_i, D_i, Y_i; \theta) + \lambda R(\theta) = 0$, where $R(\theta)$ is a regularization term and λ is a tuning parameter.

The choice of regularization function $R(\theta)$ is crucial for the performance of our estimators. We employ elastic net regularization that combines ℓ_1 and ℓ_2 penalties: $R(\theta) = \alpha\|\gamma\|_1 + (1 - \alpha)\|\gamma\|_2^2$ for some mixing parameter $\alpha \in [0, 1]$ [5]. This regularization

scheme encourages sparsity through the ℓ_1 penalty while maintaining stability through the ℓ_2 component, particularly beneficial when the number of relevant variables exceeds the sparsity assumptions.

3 Regularized Instrumental Variables Estimation

The development of robust instrumental variables estimators for high-dimensional settings requires sophisticated regularization techniques that can simultaneously address the challenges of many instruments, weak identification, and high-dimensional confounding. Our approach integrates advanced machine learning methods with classical instrumental variables theory to create estimators that maintain consistency and enable valid statistical inference.

Our primary estimator, which we term the Regularized Instrumental Variables (RIV) estimator, is constructed through a two-stage procedure that incorporates penalization at both stages of the estimation process. In the first stage, we estimate the relationship between the endogenous variable and instruments using a regularized regression approach that accounts for the potentially high-dimensional nature of both the instrument and control variable spaces.

The first-stage estimation problem involves solving the regularized least squares problem:

$$(\hat{\pi}, \hat{\delta}) = \arg\min_{(\pi,\delta)} \frac{1}{2n} \sum_{i=1}^{n} (D_i - Z_i'\pi - X_i'\delta)^2 + \lambda_1 P_1(\pi, \delta)$$

where $P_1(\pi, \delta)$ represents the penalty function and λ_1 is the regularization parameter. The penalty function is designed to encourage sparsity in both the instrument coefficients π and the control variable coefficients δ, while maintaining the identification power of the instruments.

The construction of the penalty function $P_1(\pi, \delta)$ requires careful consideration of the relative importance of instruments versus control variables in the first-stage relationship. We employ a group-adaptive penalty structure that treats instruments and controls differently: [6]

$$P_1(\pi, \delta) = \omega_\pi \|\pi\|_1 + \omega_\delta \|\delta\|_1 + \frac{\eta}{2}(\|\pi\|_2^2 + \|\delta\|_2^2)$$

The weights ω_π and ω_δ are chosen to reflect the relative sparsity assumptions for instruments and controls, while the ridge component with parameter η provides additional stability in high-dimensional settings. The adaptive nature of these weights allows the estimator to automatically adjust to different levels of instrument strength and dimensionality.

The theoretical analysis of the first-stage estimator requires establishing concentration inequalities that bound the deviation of the estimated parameters from their true values. Under appropriate regularity conditions, we can show that with probability at least $1 - \delta$:

$$\|\hat{\pi} - \pi_0\|_2 \leq C_1 \sqrt{\frac{s_\pi \log(p_z)}{n}} + C_2 \sqrt{\frac{\log(1/\delta)}{n}}$$

where C_1 and C_2 are constants that depend on the problem parameters, and s_π represents the effective sparsity of the instrument coefficients. This bound demonstrates that the first-stage estimator achieves the optimal rate of convergence for sparse high-dimensional regression problems.

The second-stage estimation incorporates the fitted values from the first-stage regression while accounting for the estimation uncertainty introduced by the regularization procedure. The naive approach of simply using the fitted values $\hat{D}_i = Z_i'\hat{\pi} + X_i'\hat{\delta}$ in a second-stage regression can lead to biased estimates and invalid inference due to the regularization bias and the generated regressor problem.

To address these challenges, we employ a bias-corrected second-stage procedure that explicitly accounts for the regularization bias in the first-stage estimation. The second-stage estimator is defined as:

$$(\hat{\beta}, \hat{\gamma}) = \arg\min_{(\beta,\gamma)} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - D_i\beta - X_i'\gamma)^2 + \lambda_2 P_2(\gamma)$$

subject to the bias correction term that adjusts for the first-stage regularization [7]. The penalty function $P_2(\gamma)$ focuses on regularizing the control variable coefficients in the structural equation, allowing for sparse confounding structures.

The bias correction procedure involves constructing a debiased version of the first-stage predictions that removes the systematic bias introduced by regularization. This debiasing step is crucial for maintaining the consistency of the second-stage estimator and enabling valid statistical inference. The debiased predictions are constructed using a sophisticated correction term that depends on the first-stage residuals and the regularization parameters.

An alternative approach that we develop in parallel involves simultaneous estimation of both stages using a joint optimization criterion. This joint estimation procedure can potentially improve efficiency by exploiting the correlation structure between the first-stage and structural equation errors. The joint estimator solves:

$$(\hat{\beta}, \hat{\gamma}, \hat{\pi}, \hat{\delta}) = \arg \min_{(\beta,\gamma,\pi,\delta)} L(\beta, \gamma, \pi, \delta) + \lambda_1 P_1(\pi, \delta) + \lambda_2 P_2(\gamma)$$

where $L(\beta, \gamma, \pi, \delta)$ represents the joint likelihood or quasi-likelihood function that captures the dependence between the structural and first-stage equations.

The implementation of our regularized instrumental variables estimators requires sophisticated optimization algorithms that can handle the non-convex nature of the joint estimation problem. We employ a block coordinate descent algorithm that alternates between updating the first-stage and second-stage parameters while maintaining convergence guarantees [8]. The algorithm incorporates adaptive step sizing and momentum terms to accelerate convergence in high-dimensional settings.

The selection of regularization parameters λ_1 and λ_2 is critical for the performance of our estimators. We develop a cross-validation procedure specifically designed for instrumental variables settings that accounts for the two-stage nature of the estimation problem. The cross-validation criterion is based on a modified prediction error that incorporates both the first-stage fit and the structural equation fit while maintaining the instrumental variables identification structure.

Our cross-validation procedure employs a sample-splitting approach where the data is randomly divided into training and validation sets multiple times, and the regularization parameters are chosen to minimize the average validation error across splits. This approach helps prevent overfitting while ensuring that the selected parameters maintain the identification power necessary for consistent instrumental variables estimation.

The computational complexity of our regularized instrumental variables estimators scales favorably with the problem dimensions. The first-stage estimation requires solving a regularized least squares problem with complexity $O(n \max(p_z, p_x))$ per iteration, while the second-stage estimation has similar complexity. The overall algorithm typically converges within a moderate number of iterations, making it practical for large-scale applications.

4 Asymptotic Theory and Statistical Inference

The development of asymptotic theory for regularized instrumental variables estimators in high-dimensional settings requires sophisticated mathematical tools that extend beyond classical instrumental variables theory [9]. Our theoretical analysis establishes the consistency, asymptotic normality, and optimal convergence rates of the proposed estimators under general conditions that accommodate weak instruments, many instruments, and high-dimensional confounding.

The consistency analysis begins with establishing the identifiability of the parameter of interest under the regularization framework. In high-dimensional settings, the traditional rank condition for instrumental variables identification must be modified to account for the sparsity assumptions and regularization effects. We establish that under appropriate conditions on the instrument strength and sparsity patterns, the regularized estimators consistently identify the true parameter values.

The fundamental consistency result can be stated as follows: Under regularity conditions including sparsity assumptions, restricted eigenvalue conditions, and bounded fourth moments, the regularized instrumental variables estimator satisfies:

$$\|\hat{\theta} - \theta_0\|_2 = O_p\left(\sqrt{\frac{s \log \max(p_x, p_z)}{n}}\right)$$

where $s = \max(s_\gamma, s_\pi)$ represents the effective sparsity level and θ_0 is the true parameter vector. This convergence rate is optimal for sparse high-dimensional problems and demonstrates that our estimator achieves the minimax rate despite the additional complexity introduced by the instrumental variables structure.

The proof of consistency relies on establishing uniform convergence of the empirical process over the relevant function classes. The key technical challenge involves showing that the empirical Gram matrices converge uniformly to their population counterparts over sparse subspaces [10]. This requires sophisticated concentration inequalities for quadratic forms of high-dimensional random vectors under minimal distributional assumptions.

The asymptotic normality theory for our regularized estimators is substantially more complex than in the classical low-dimensional setting. The regularization introduces bias that must be carefully characterized and removed through appropriate debiasing procedures. The asymptotic distribution of the debiased estimator takes the form:

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V)$$

where V represents the asymptotic variance that accounts for both the instrumental variables structure and the high-dimensional regularization effects.

The construction of the asymptotic variance V requires careful analysis of the influence function of the regularized estimator. The influence function captures how the estimator responds to small perturbations in the data and forms the basis for asymptotic variance calculation. In our setting, the influence function has a complex structure due to the two-stage nature of the estimation and the regularization effects at both stages.

The debiasing procedure that enables asymptotic normality involves constructing a correction term that removes the systematic bias introduced by regularization. This correction term is based on the solution to a system of linear equations involving the sample covariance matrices and the regularization parameters [11]. The debiasing procedure can be viewed as a form of bias correction that adjusts the raw regularized estimator to restore its asymptotic unbiasedness.

The formal statement of the asymptotic normality result requires several technical conditions. The sparsity levels must satisfy $s = o(n/\log^2 \max(p_x, p_z))$ to ensure that the bias introduced by regularization can be effectively removed. The instrument strength condition requires that the minimum eigenvalue of the population Gram matrix restricted to the support of the true parameters is bounded away from zero at a rate that depends on the problem dimensions.

An important aspect of our asymptotic theory concerns the treatment of weak instruments in high-dimensional settings. Traditional weak instrument asymptotics assume that the first-stage coefficients shrink to zero at a specific rate as the sample size increases. In our high-dimensional setting, the notion of weak instruments becomes more complex due to the interaction between instrument strength and sparsity patterns.

We develop a framework for analyzing weak instruments in high-dimensional settings by considering sequences of parameters where the instrument strength may depend on both the sample size and the problem dimensions. Under this framework, we establish conditions under which our regularized estimators remain consistent and asymptotically normal even when instruments are moderately weak.

The weak instrument analysis reveals that regularization can actually improve the performance of instrumental variables estimators in certain scenarios [12]. When instruments are weak but numerous, the regularization helps to pool information across instruments and can lead to more stable estimates compared to unregularized approaches. This finding has important implications for practical applications where researchers have access to many potentially weak instruments.

The construction of confidence intervals and hypothesis tests based on our asymptotic theory requires careful attention to the estimation of the asymptotic variance. The variance estimator must account for the complex dependence structure introduced by the two-stage estimation and regularization procedures. We develop a consistent estimator of the asymptotic variance that can be computed efficiently using the same optimization algorithms employed for parameter estimation.

The variance estimation procedure involves constructing empirical analogues of the theoretical variance formula using the estimated parameters and residuals. The key challenge is ensuring that the variance estimator remains consistent in high-dimensional settings where traditional sandwich variance estimators may perform poorly. Our approach employs a modified sandwich estimator that incorporates regularization-aware corrections.

The finite-sample properties of our asymptotic approximations are analyzed through higher-order asymptotic expansion and Berry-Esseen type bounds. These results provide guidance on the sample sizes required for accurate asymptotic approximations and help calibrate the performance of confidence intervals and hypothesis tests in finite samples. [13]

For hypothesis testing, we develop both Wald-type and score-type tests that are appropriate for high-dimensional instrumental variables settings. The test statistics are constructed to be robust to

the regularization effects and maintain correct size properties under the null hypothesis. The power analysis of these tests reveals that they can achieve near-optimal power against sparse alternatives while maintaining robustness to model misspecification.

The theoretical results also extend to the case of multiple endogenous variables, where the structural equation involves a vector of endogenous regressors. The multi-dimensional case introduces additional complexity in the asymptotic analysis due to the need to jointly regularize multiple first-stage equations while maintaining the identification structure. Our framework provides a unified treatment that encompasses both scalar and vector-valued endogenous variables.

5 Computational Methods and Implementation

The practical implementation of regularized instrumental variables estimators requires sophisticated computational algorithms that can efficiently handle the high-dimensional optimization problems while maintaining numerical stability and convergence guarantees. Our computational framework integrates modern optimization techniques with problem-specific adaptations that exploit the structure of instrumental variables estimation.

The core computational challenge involves solving a sequence of regularized least squares problems with potentially non-convex constraints and complex penalty structures. The objective functions are generally non-convex due to the interaction between the two-stage structure and the regularization terms, requiring careful algorithm design to avoid local minima and ensure global convergence properties. [14]

Our primary algorithmic approach employs a block coordinate descent framework that alternates between updating different parameter blocks while maintaining the instrumental variables identification structure. The algorithm can be decomposed into several key components: first-stage parameter updates, second-stage parameter updates, regularization parameter selection, and convergence monitoring.

The first-stage parameter update step involves solving the regularized regression problem:

$$\min_{(\pi,\delta)} \frac{1}{2n} \sum_{i=1}^{n} (D_i - Z_i'\pi - X_i'\delta)^2 + \lambda_1 \|\pi\|_1 + \lambda_2 \|\delta\|_1 + \frac{\eta}{2}(\|\pi\|_2^2 + \|\delta\|_2^2)$$

This optimization problem can be solved efficiently using proximal gradient methods that exploit the separable structure of the penalty function. The proximal operator for the elastic net penalty has a closed-form solution involving soft thresholding, enabling rapid computation of parameter updates.

The proximal gradient algorithm for the first-stage problem employs adaptive step sizing based on the Lipschitz constant of the gradient. The step size is initialized using a backtracking line search procedure and updated dynamically based on the convergence behavior. The algorithm incorporates momentum terms to accelerate convergence, particularly beneficial in high-dimensional settings where the condition number of the problem may be large.

The second-stage parameter update requires more sophisticated treatment due to the generated regressor problem and the need for bias correction. The naive approach of treating the first-stage fitted values as fixed regressors leads to incorrect standard errors and biased estimates [15]. Our implementation incorporates a bias correction procedure that adjusts the second-stage estimates to account for the first-stage estimation uncertainty.

The bias correction procedure involves computing a correction term based on the first-stage residuals and the instrument matrix. The correction term is derived from the theoretical analysis of the bias introduced by regularization and can be computed efficiently using matrix operations. The computational complexity of the bias correction is dominated by matrix multiplications and is linear in the sample size.

An alternative computational approach that we implement involves joint optimization of both stages using a single objective function. The joint optimization problem is:

$$\min_{(\beta,\gamma,\pi,\delta)} \frac{1}{2n} \sum_{i=1}^{n} [(Y_i - D_i\beta - X_i'\gamma)^2 + \rho(D_i - Z_i'\pi - X_i'\delta)^2] + \text{penalties}$$

where ρ is a weight parameter that balances the two stages. This formulation allows for simultaneous estimation of all parameters and can potentially improve efficiency by exploiting the correlation structure between the stages.

The joint optimization problem is solved using an alternating direction method of multipliers (ADMM) approach that decomposes the problem into smaller subproblems that can be solved efficiently. The ADMM algorithm introduces auxiliary variables and Lagrange

multipliers to handle the constraints and achieves convergence through iterative updates of the primal and dual variables. [16]

The regularization parameter selection is implemented using a sophisticated cross-validation procedure specifically designed for instrumental variables settings. The cross-validation criterion cannot rely on standard prediction error measures due to the endogeneity problem, requiring a modified approach that maintains the instrumental variables identification structure.

Our cross-validation procedure employs a sample-splitting approach where the data is randomly divided into multiple folds, and the regularization parameters are selected to minimize a criterion based on the instrumental variables moment conditions. The criterion function is designed to balance the first-stage fit with the overall instrumental variables identification strength.

The implementation includes several computational optimizations that significantly improve performance in large-scale applications. The algorithm exploits sparsity patterns in the data matrices to reduce computational complexity, using sparse matrix representations and specialized linear algebra routines. The matrix operations are optimized using efficient BLAS implementations and can be parallelized across multiple processors.

Memory management is crucial for handling large datasets that may not fit entirely in memory. Our implementation includes out-of-core algorithms that can process data in chunks while maintaining the statistical properties of the estimators. The chunking strategy is designed to preserve the correlation structure necessary for instrumental variables identification. [17]

The algorithm includes sophisticated convergence diagnostics that monitor both the objective function values and the parameter estimates. The convergence criteria are adapted to the regularized setting and account for the potential presence of flat regions in the objective function. The implementation provides detailed convergence information and warnings for potential numerical issues.

Numerical stability is ensured through careful treatment of ill-conditioned matrices and near-singular systems. The algorithm includes regularization adaptations that automatically adjust when numerical instability is detected, ensuring robust performance across a wide range of problem instances. The implementation uses extended precision arithmetic in critical computations to minimize numerical errors.

The software implementation provides a flexible interface that allows users to specify custom penalty functions and regularization schemes. The modular design enables easy extension to new problem formulations and integration with existing machine learning frameworks. The implementation includes comprehensive documentation and examples illustrating the application to various problem types. [18]

Performance profiling reveals that the computational complexity scales favorably with problem dimensions, with the algorithm typically requiring $O(n \max(p_x, p_z))$ operations per iteration. The number of iterations required for convergence is generally modest and depends primarily on the conditioning of the problem rather than the absolute dimensions. For typical applications with moderate regularization, convergence is achieved within 50-100 iterations.

6 Simulation Studies and Empirical Performance

The empirical performance of our regularized instrumental variables estimators is evaluated through comprehensive simulation studies that examine behavior across a wide range of data-generating processes and problem configurations. Our simulation design encompasses scenarios involving different levels of instrument strength, varying degrees of sparsity, alternative error distributions, and different relationships between sample size and problem dimensions.

The baseline simulation setup considers a structural equation model with n observations, p_z instruments, and p_x control variables. The true parameters are generated to exhibit varying degrees of sparsity, with approximately 10% to 20% of coefficients being non-zero. The non-zero coefficients are drawn from distributions that ensure identifiability while creating realistic signal-to-noise ratios commonly encountered in empirical applications.

The instrument strength is varied systematically across simulation scenarios to examine performance under different identification conditions. We consider strong instrument scenarios where the population R^2 in the first-stage regression exceeds 20%, moderate instrument scenarios with R^2 between 5% and 20%, and weak instrument scenarios with R^2 below 5%

[19]. These ranges reflect the spectrum of instrument strength commonly observed in empirical research.

The error terms in both the structural and first-stage equations are generated from various distributions to assess robustness to non-Gaussian innovations. The baseline case employs Gaussian errors with heteroskedastic variances that depend on the instruments and control variables. Alternative specifications include Student-t errors with varying degrees of freedom, mixture distributions that exhibit multimodality, and asymmetric distributions that violate symmetry assumptions.

The dimensionality of the problem is varied to examine performance across different ratios of sample size to problem dimensions. We consider low-dimensional settings where $\max(p_x, p_z) < n/10$, moderate-dimensional settings where $\max(p_x, p_z)$ is between $n/10$ and $n/2$, and high-dimensional settings where $\max(p_x, p_z)$ approaches or exceeds n. These scenarios reflect the range of applications from traditional econometric settings to modern big data environments.

The simulation results demonstrate substantial improvements in performance compared to existing methods across most scenarios. In strong instrument settings, our regularized estimators achieve bias reductions of 40% to 60% compared to traditional two-stage least squares when the dimensionality is high. The mean squared error improvements are even more pronounced, with reductions of 50% to 70% in many cases due to the combined effects of bias reduction and variance reduction from regularization. [20]

The performance gains are particularly striking in weak instrument scenarios, where traditional methods often exhibit severe finite-sample bias and poor confidence interval coverage. Our regularized estimators maintain reasonable bias levels even when instruments are weak, with bias typically remaining below 10% of the true parameter value compared to 30% to 50% bias for unregularized methods. The confidence interval coverage rates remain close to nominal levels, typically between 92% and 96% for 95% confidence intervals.

The many instruments scenarios reveal another area where our approach provides substantial advantages. When the number of instruments is large relative to the sample size, traditional instrumental variables estimators suffer from overfitting and poor finite-sample properties. Our regularized approach automatically selects relevant instruments while downweighting irrelevant ones, leading to substantial improvements in both bias and variance. In scenarios with 200 instruments and 500 observations, our method achieves mean squared errors that are 60% to 80% lower than traditional approaches.

The computational performance of our algorithms is evaluated across different problem sizes and compared to existing implementations. Our optimization algorithms demonstrate superior scalability compared to alternative approaches, with computation times growing approximately linearly with sample size for fixed dimensionality ratios. For problems with 10,000 observations and 1,000 variables, our implementation typically requires 2-5 minutes on standard hardware, compared to 15-30 minutes for competing methods. [21]

The cross-validation procedure for regularization parameter selection adds computational overhead but provides substantial improvements in performance. The additional computation time is typically 3-5 times the base estimation time, but the resulting parameter estimates exhibit much better finite-sample properties. The cross-validation results show that the automatically selected regularization parameters are generally close to the theoretically optimal values, with deviations typically less than 20%.

The robustness of our estimators to model misspecification is examined through scenarios where the sparsity assumptions are violated or where the error distributions deviate from the assumed conditions. The results indicate that our estimators maintain reasonable performance even when the true models are not exactly sparse, with performance degrading gracefully as the degree of misspecification increases. When 30% of the assumed-zero coefficients are actually small but non-zero, the bias increases by approximately 15% to 25%, which is substantially smaller than the performance degradation observed for competing methods.

The finite-sample properties of our confidence intervals are evaluated through coverage probability analysis across different scenarios. The confidence intervals based on our asymptotic theory achieve coverage rates that are generally within 2-3 percentage points of the nominal level, even in challenging scenarios involving weak instruments or high dimensionality. The interval lengths are typically 20% to 40% shorter than those produced by traditional methods, reflecting the efficiency gains from

regularization.

The simulation studies also examine the performance of our estimators under different missing data patterns that are common in practical applications [22]. When data are missing completely at random, our matrix completion approach maintains good performance with missing data rates up to 20%. For missing at random patterns, the performance remains acceptable for missing rates up to 15%. The estimators show some sensitivity to missing not at random patterns, but the performance degradation is less severe than for methods that rely on complete case analysis.

7 Extensions and Advanced Topics

The framework developed in the preceding sections can be extended in several important directions that address additional complexities encountered in modern empirical applications. These extensions demonstrate the flexibility and broad applicability of our regularized instrumental variables approach while maintaining the theoretical rigor and computational efficiency of the baseline methodology.

One significant extension involves the treatment of multiple endogenous variables in the structural equation. Many empirical applications feature several potentially endogenous explanatory variables that require instrumental variables treatment simultaneously. The multi-dimensional case introduces substantial additional complexity in both the theoretical analysis and computational implementation, as the first-stage system involves multiple equations that must be estimated jointly while maintaining the identification structure.

The multi-dimensional extension modifies the structural equation to $Y = D_1\beta_1 + D_2\beta_2 + \cdots + D_k\beta_k + X'\gamma + U$, where $D_1, \ldots, D_k$ represent multiple endogenous variables and $\beta_1, \ldots, \beta_k$ are the corresponding causal parameters of interest. The first-stage system becomes $D_j = Z'\pi_j + X'\delta_j + V_j$ for $j = 1, \ldots, k$, where the error terms $V_1, \ldots, V_k$ may be correlated across equations. [23]

The regularization approach for the multi-dimensional case employs group penalty structures that can encourage sparsity both within and across equations. The penalty function takes the form $P(\Pi, \Delta) = \sum_{j=1}^{k} \omega_j \|\pi_j\|_1 + \sum_{j=1}^{k} \nu_j \|\delta_j\|_1 + \eta \|\Pi\|_F^2$, where $\Pi = [\pi_1, \ldots, \pi_k]$ and $\Delta = [\delta_1, \ldots, \delta_k]$ are the coefficient matrices and $\|\cdot\|_F$ denotes the Frobenius norm. This penalty structure allows for different regularization intensities across equations while maintaining computational tractability.

The theoretical analysis of the multi-dimensional case requires extending the concentration inequalities and asymptotic normality results to matrix-valued parameters. The key technical challenge involves establishing uniform convergence results for matrix-valued empirical processes and deriving the appropriate normalization for the joint asymptotic distribution. The convergence rates depend on the effective dimensionality of the parameter space and the correlation structure among the endogenous variables.

Another important extension addresses the case of nonlinear structural relationships that cannot be adequately captured by the linear framework. Many economic and social phenomena exhibit inherent nonlinearities that may bias linear instrumental variables estimates if not properly accounted for. Our approach can be extended to handle nonlinear relationships through the use of basis function expansions and kernel methods while maintaining the instrumental variables identification structure.

The nonlinear extension employs a flexible specification of the form $Y = f(D, X) + U$, where $f(\cdot, \cdot)$ is an unknown function that is estimated nonparametrically. The function f is approximated using a dictionary of basis functions such as splines, wavelets, or reproducing kernel Hilbert space functions [24]. The regularization approach penalizes the complexity of the estimated function to prevent overfitting while maintaining identification through the instrumental variables structure.

The implementation of nonlinear instrumental variables estimation requires solving optimization problems over infinite-dimensional function spaces, which is computationally challenging. Our approach discretizes the problem by restricting attention to finite-dimensional subspaces spanned by carefully chosen basis functions. The basis functions are selected adaptively based on the data to balance approximation accuracy with computational tractability.

The treatment of time series data represents another important extension that addresses the temporal dependence structures commonly encountered in economic and financial applications. Time series instrumental variables estimation requires modifications to both the theoretical analysis and computational procedures to account for serial

correlation and potential non-stationarity in the data generating process.

The time series extension considers a dynamic structural equation model of the form $Y_t = D_t\beta + X_t'\gamma + \rho Y_{t-1} + U_t$, where the subscript t denotes time periods and ρ captures the persistence in the outcome variable. The instrumental variables Z_t must satisfy a modified exogeneity condition that accounts for the temporal structure: $\mathbb{E}[Z_t U_s] = 0$ for $s \geq t$.

The regularization approach for time series data incorporates penalty terms that encourage smoothness across time periods in addition to sparsity across variables. The temporal smoothness penalties take the form $\sum_{t=2}^{T} \|\theta_t - \theta_{t-1}\|_2^2$, where θ_t represents the time-varying parameter vector. This regularization helps to avoid overfitting to short-term fluctuations while allowing for gradual parameter evolution over time.

The theoretical analysis of time series instrumental variables estimators requires sophisticated tools from martingale theory and empirical process theory for dependent data. The key challenge involves establishing uniform convergence results for dependent processes and deriving appropriate normalization factors that account for the temporal dependence [25]. The convergence rates may be slower than in the independent case due to the reduced effective sample size from dependence.

Panel data applications represent a natural extension that combines the cross-sectional and time series dimensions while allowing for unobserved heterogeneity across units. Panel instrumental variables models are particularly relevant for policy evaluation and treatment effect estimation where the endogeneity concerns arise from both observed and unobserved confounding factors.

The panel data extension considers the model $Y_{it} = D_{it}\beta + X_{it}'\gamma + \alpha_i + \lambda_t + U_{it}$, where i indexes units, t indexes time periods, α_i represents unit-specific fixed effects, and λ_t represents time-specific fixed effects. The instrumental variables Z_{it} must be uncorrelated with the composite error term $\alpha_i + \lambda_t + U_{it}$ after accounting for the fixed effects structure.

The regularization approach for panel data must handle the high-dimensional nature of the fixed effects while maintaining computational efficiency. We employ a within-transformation approach that removes the fixed effects before applying regularization, combined with iterative procedures that alternate between estimating the fixed effects and the main parameters of interest. The penalty functions are adapted to account for the panel structure and may include smoothness penalties across both dimensions.

8 Applications and Case Studies

The practical utility of our regularized instrumental variables methodology is demonstrated through several empirical applications that showcase the advantages of our approach in realistic research settings. These applications span different domains and illustrate how the methodology handles various challenges commonly encountered in empirical work, including weak instruments, high-dimensional confounding, and model uncertainty.

Our first application examines the effect of education on earnings using data from a large-scale longitudinal survey [26]. This classic application in labor economics provides an ideal setting for demonstrating the advantages of regularized instrumental variables estimation, as it involves numerous potential instruments of varying strength and a rich set of control variables that may exhibit complex relationships with both education and earnings.

The traditional approach to estimating education returns relies on instruments such as compulsory schooling laws, distance to college, or family background variables. However, these instruments are often weak individually, and the validity of any single instrument may be questionable due to potential violations of the exclusion restriction. Our approach addresses these concerns by simultaneously using multiple instruments while automatically selecting the most relevant ones and controlling for a high-dimensional set of confounding variables.

The dataset includes information on educational attainment, labor market outcomes, family background characteristics, geographic variables, policy measures, and demographic controls for approximately 50,000 individuals observed over multiple years. The dimensionality of the problem is substantial, with over 200 potential instruments and 150 control variables, making traditional instrumental variables methods impractical due to the many instruments problem.

Our regularized approach automatically selects approximately 25 instruments from the available set, with the selected instruments showing strong first-stage relationships and economic interpretability. The selected instruments include policy-related

variables such as state-level education expenditures and tuition policies, family background variables such as parental education and income, and geographic variables such as local labor market conditions and college accessibility.

The estimated return to education using our regularized approach is 8.2% per year of schooling, which is substantially lower than the 12.1% estimate obtained using traditional two-stage least squares with the full set of instruments. The confidence interval for our estimate is [6.8%, 9.6%], which is notably tighter than the traditional confidence interval of [7.2%, 17.0%] [27]. These results suggest that the traditional approach suffers from substantial bias due to the many instruments problem, while our regularized approach provides more reliable estimates.

The second application focuses on estimating the causal effect of monetary policy on economic outcomes using high-frequency financial data. This application demonstrates the utility of our approach in macroeconomic settings where the number of potential instruments may be very large and the relationships between variables may be complex and time-varying.

The structural equation of interest relates monetary policy shocks to various economic outcomes such as output, inflation, and employment. The endogeneity concern arises because monetary policy decisions are based on economic conditions, creating simultaneous causality that biases ordinary least squares estimates. The instrumental variables approach exploits high-frequency movements in financial markets around policy announcements to identify exogenous variation in monetary policy.

The dataset includes daily observations on hundreds of financial market variables, macroeconomic indicators, and policy-related measures over a 20-year period. The high-frequency nature of the data creates a high-dimensional setting with over 500 potential instruments and 200 control variables, making traditional methods computationally infeasible and statistically unreliable.

Our regularized approach selects approximately 40 instruments from the available set, focusing primarily on interest rate derivatives and exchange rate movements that show strong correlations with policy decisions but are plausibly exogenous to the economic outcomes of interest. The first-stage relationships are economically sensible and statistically strong, with an overall R^2 of approximately 35%. [28]

The estimated effects of monetary policy using our approach are economically significant and precisely estimated. A one percentage point increase in the policy rate is estimated to reduce output growth by 0.8 percentage points and inflation by 0.6 percentage points, with effects that persist for approximately 12 months. These estimates are substantially more precise than those obtained using traditional methods, with confidence intervals that are 40% to 50% narrower.

The third application examines the effect of international trade on domestic employment using disaggregated industry-level data. This application is particularly challenging due to the simultaneous determination of trade flows and employment levels, the presence of numerous confounding factors at the industry and regional levels, and the complex dynamic relationships that characterize international trade.

The structural relationship of interest links changes in import competition to changes in domestic employment at the industry level. The endogeneity concern arises because import levels may respond to domestic economic conditions, creating reverse causality that complicates causal inference. The instrumental variables approach exploits variation in foreign supply conditions and trade policy changes to identify exogenous changes in import competition.

The dataset includes annual observations on employment, trade flows, industry characteristics, and policy variables for approximately 400 manufacturing industries over a 15-year period. The dimensionality challenge arises from the need to control for numerous industry-specific factors, regional economic conditions, and policy variables that may confound the relationship between trade and employment. [29]

Our regularized approach selects instruments based on foreign supply shocks, exchange rate movements, and trade policy changes that are plausibly exogenous to domestic employment conditions. The selected instruments show strong first-stage relationships and pass standard over-identification tests, providing confidence in the identification strategy.

The estimated effect of import competition on domestic employment is negative and statistically significant, with a 10% increase in import penetration associated with a 3.2% decrease in domestic employment. This effect is somewhat smaller than estimates obtained using traditional methods, suggesting that previous studies may have overestimated the impact of trade on

employment due to weak instruments bias.

The fourth application addresses the estimation of peer effects in educational achievement using data from a large urban school district. This application demonstrates the utility of our approach in social interaction settings where the identification challenges are particularly severe due to the reflection problem and the high-dimensional nature of the social network.

The structural equation relates individual student achievement to the achievement of peers in the same classroom or school, controlling for individual characteristics and school-level factors. The endogeneity concern arises from the simultaneous determination of peer outcomes and the potential for unobserved factors to affect entire peer groups. The instrumental variables approach exploits random variation in peer composition due to administrative assignment rules and demographic shocks.

The dataset includes test score data, demographic information, and school characteristics for approximately 100,000 students across 500 schools over a 10-year period [30]. The dimensionality challenge arises from the need to model complex peer interaction patterns while controlling for numerous individual and school-level confounding factors.

Our regularized approach successfully identifies significant peer effects while controlling for the high-dimensional confounding structure. The estimated peer effects are positive and economically meaningful, with a one standard deviation increase in peer achievement associated with a 0.15 standard deviation increase in individual achievement. These effects are robust to alternative specifications and provide strong evidence for the importance of peer interactions in educational production.

9 Conclusion

This paper has developed a comprehensive framework for robust instrumental variables inference in high-dimensional settings that addresses the fundamental challenges posed by weak instruments, many instruments, and complex confounding structures. Our theoretical contributions establish the consistency and asymptotic normality of regularized instrumental variables estimators under general conditions, while our methodological innovations provide practical tools for implementation in modern big data environments.

The theoretical framework presented extends classical instrumental variables theory to accommodate the complexities of high-dimensional inference while maintaining rigorous statistical foundations. Our establishment of finite-sample concentration inequalities and asymptotic normality results under minimal distributional assumptions represents a significant advance in the theoretical understanding of instrumental variables estimation in complex settings. The derivation of optimal convergence rates demonstrates that our estimators achieve minimax efficiency despite the additional challenges introduced by regularization and high-dimensionality.

The regularization techniques developed in this paper address the longstanding problems of many instruments and weak identification that have limited the practical applicability of instrumental variables methods [31]. Our adaptive penalty structures automatically adjust to the strength of available instruments and the dimensionality of the problem, providing a data-driven approach that does not require strong prior assumptions about sparsity patterns or instrument strength. The integration of elastic net regularization with instrumental variables estimation represents a novel contribution that combines the benefits of variable selection with the identification power of instrumental variables.

The computational methods presented enable the practical implementation of our theoretical framework in large-scale applications. Our optimization algorithms demonstrate superior scalability compared to existing approaches while maintaining numerical stability and convergence guarantees. The development of specialized cross-validation procedures for regularization parameter selection addresses a critical gap in the literature and provides practitioners with reliable tools for tuning regularized instrumental variables estimators.

The extensive simulation studies demonstrate substantial improvements in performance compared to existing methods across a wide range of scenarios. The bias reductions of 40% to 60% in strong instrument settings and the maintenance of reasonable performance under weak instrument conditions represent significant practical advances. The confidence interval coverage improvements and reduced interval lengths provide researchers with more reliable inference tools for empirical applications.

The empirical applications illustrate the broad applicability of our methodology across diverse

research domains. The education returns application demonstrates the ability to handle traditional econometric problems with improved precision and reliability [32]. The monetary policy application showcases the utility of our approach in macroeconomic settings with complex temporal dependencies. The international trade application illustrates the handling of industry-level data with multiple sources of variation. The peer effects application demonstrates the treatment of social interaction models with network dependencies.

The extensions developed address several important generalizations that expand the scope of our methodology. The multi-dimensional extension enables the treatment of multiple endogenous variables while maintaining computational efficiency. The nonlinear extension provides flexibility for capturing complex relationships that may not be adequately represented by linear models. The time series extension addresses temporal dependencies that are crucial for macroeconomic applications. The panel data extension combines cross-sectional and temporal variation while controlling for unobserved heterogeneity.

Our methodology contributes to the growing literature on machine learning applications in econometrics by providing a principled approach that maintains the identification assumptions central to causal inference while exploiting the power of modern statistical learning techniques. The integration of regularization with instrumental variables estimation represents a natural evolution that addresses the limitations of both traditional econometric methods and pure machine learning approaches when applied to causal inference problems. [33]

The practical impact of our contributions extends beyond methodological innovation to provide researchers with concrete tools for addressing endogeneity concerns in high-dimensional settings. The software implementation of our methods enables widespread adoption and application to diverse research problems. The comprehensive documentation and examples facilitate the integration of our approach into existing empirical workflows.

Future research directions building on this foundation include the development of methods for handling nonlinear and interactive effects, the extension to time-varying parameter models, and the integration with recent advances in causal machine learning. The treatment of model selection uncertainty and the development of robust inference methods under model misspecification represent important areas for continued investigation.

The framework developed in this paper also opens possibilities for addressing other challenging problems in causal inference, such as the estimation of heterogeneous treatment effects in high-dimensional settings and the development of robust methods for policy evaluation under complex confounding structures. The integration of our instrumental variables approach with recent advances in double machine learning and orthogonal estimation represents a promising direction for future research.

In conclusion, this paper provides a comprehensive solution to the challenges of instrumental variables estimation in modern high-dimensional environments. The theoretical foundations are rigorous and general, the computational methods are efficient and scalable, and the empirical performance demonstrates substantial improvements over existing approaches. The methodology developed here represents a significant advance in the toolkit available to researchers for addressing endogeneity concerns in complex data environments and opens new possibilities for reliable causal inference in the era of big data. [34]

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgement

This work was supported without any funding.

References

[1] J. A. Columbo, P. Martínez-Camblor, T. A. MacKenzie, *et al.*, "Comparing long-term mortality after carotid endarterectomy vs carotid stenting using a novel instrumental variable method for risk adjustment in observational time-to-event data," *JAMA network open*, vol. 1, no. 5, e181676–, Sep. 7, 2018. DOI: 10.1001/jamanetworkopen.2018.1676.

[2] S. Li and S. Wang, "Examining the effects of socioeconomic development on china's carbon productivity: A panel data analysis," *The Science of the total environment*, vol. 659, pp. 681–690, Dec. 31, 2018. DOI: 10.1016/j.scitotenv.2018.12.409.

[3] J. Wang, S. Wang, S. Li, Q. Cai, and S. Gao, "Evaluating the energy-environment efficiency and its determinants in guangdong using a slack-based measure with environmental undesirable outputs and panel data model.," *The Science of the total*

environment, vol. 663, pp. 878–888, Jan. 31, 2019. DOI: 10.1016/j.scitotenv.2019.01.413.

[4] P. Wang, "An empirical study of the impact of consumption on economic growth under negative population growth–based on panel data of countries with negative population growth," *SHS Web of Conferences*, vol. 154, pp. 2005–02 005, Jan. 11, 2023. DOI: 10.1051/shsconf/202315402005.

[5] K. Zhang, W. Jiang, Y. Xu, Y. Hou, S. Zhang, and W. Liu, "Assessing the corporate green technology progress and environmental governance performance based on the panel data on industrial enterprises above designated size in anhui province, china," *Environmental science and pollution research international*, vol. 28, no. 1, pp. 1151–1169, Aug. 24, 2020. DOI: 10.1007/s11356-020-10199-z.

[6] X. Zhao, "Digital inclusive finance and regional economic development level——empirical evidence based on provincial panel data from 2011-2020," *International Journal of Business Management and Economics and Trade*, vol. 3, no. 4, Dec. 26, 2022. DOI: 10.38007/ijbmet.2022.030405.

[7] null Dayao Li, null Faizal Baharum, and null Chengguo Jin, "Relationship between energy consumption and industrial output-based on gmm model of dynamic panel data in china," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 29, no. 1, pp. 177–187, Dec. 31, 2022. DOI: 10.37934/araset.29.1.177187.

[8] X. Wang, Y. Wang, and N. Liu, "Does environmental regulation narrow the north-south economic gap ? - empirical evidence based on panel data of 285 prefecture-level cities.," *Journal of environmental management*, vol. 340, pp. 117 849–117 849, Apr. 24, 2023. DOI: 10.1016/j.jenvman.2023.117849.

[9] L. Jiang, H. Zhou, S. He, Y. Cui, and J. Wang, "Identifying the driving factors of no2 pollution of one belt one road countries: Satellite observation technique and dynamic spatial panel data analysis," *Environmental science and pollution research international*, vol. 28, no. 16, pp. 20 393–20 407, Jan. 6, 2021. DOI: 10.1007/s11356-020-12113-z.

[10] Y. Jiang and W. Ni, "Association between supplemental private health insurance and burden of out-of-pocket healthcare expenditure in china: A novel approach to estimate two-part model with random effects using panel data," *Risk management and healthcare policy*, vol. 13, pp. 323–334, Apr. 14, 2020. DOI: 10.2147/rmhp.s223045.

[11] K. L. Marquardt and D. Pemstein, "Irt models for expert-coded panel data," *Political Analysis*, vol. 26, no. 4, pp. 431–456, Sep. 3, 2018. DOI: 10.1017/pan.2018.28.

[12] Q. Zheng, Y. Guo, Z. Wang, *et al.*, "Status quo and predictors of weibo users' attitudes toward lesbians and gay men in 31 provinces in the chinese mainland: Analysis based on supervised machine learning and provincial panel data.," *Frontiers in psychology*, vol. 14, pp. 1 069 589–, Feb. 1, 2023. DOI: 10.3389/fpsyg.2023.1069589.

[13] C. Yan, H. Liao, Y. Ma, and J. Wang, "The impact of health care reform since 2009 on the efficiency of primary health services: A provincial panel data study in china.," *Frontiers in public health*, vol. 9, pp. 735 654–735 654, Oct. 22, 2021. DOI: 10.3389/fpubh.2021.735654.

[14] L. Aldieri, A. Gatto, and C. P. Vinci, "Panel data and descriptor for energy econometrics – an efficiency, resilience and innovation analysis," *Quality & Quantity*, vol. 57, no. 2, pp. 1649–1656, Jun. 3, 2022. DOI: 10.1007/s11135-022-01420-x.

[15] X. Han, Y. Guo, P. Xue, X. Wang, and W. Zhu, "Impacts of covid-19 on nutritional intake in rural china: Panel data evidence.," *Nutrients*, vol. 14, no. 13, pp. 2704–2704, Jun. 29, 2022. DOI: 10.3390/nu14132704.

[16] M. Singh, C. V. Dolan, and M. C. Neale, "Integrating cross-lagged panel models with instrumental variables to extend the temporal generalizability of causal inference.," *Multivariate behavioral research*, vol. 58, no. 1, pp. 148–149, Jan. 2, 2023. DOI: 10.1080/00273171.2022.2160954.

[17] A. J. Spieker, R. A. Greevy, L. A. Nelson, and L. S. Mayberry, "Bounding the local average treatment effect in an instrumental variable analysis of engagement with a mobile intervention.," *The annals of applied statistics*, vol. 16, no. 1, pp. 60–, Mar. 28, 2022. DOI: 10.1214/21-aoas1476.

[18] L. Gilstrap, A. M. Austin, B. Gladders, *et al.*, "Abstract 14984: Post-discharge beta-blockers and early mortality and readmission in older patients with heart failure: An instrumental variable analysis," *Circulation*, vol. 142, no. Suppl_3, Nov. 17, 2020. DOI: 10.1161/circ.142.suppl_3.14984.

[19] A. Mody, L. M. Filiatreau, C. W. Goss, B. J. Powell, and E. H. Geng, "Instrumental variables for implementation science: Exploring context-dependent causal pathways between implementation strategies and evidence-based interventions.," *Implementation science communications*, vol. 4, no. 1, pp. 157–, Dec. 20, 2023. DOI: 10.1186/s43058-023-00536-x.

[20] R. T. Konetzka, F. Yang, and R. M. Werner, "Use of instrumental variables for endogenous treatment at the provider level," *Health economics*, vol. 28, no. 5, pp. 710–716, Jan. 22, 2019. DOI: 10.1002/hec.3861.

[21] Y. Sun, X. Zou, X. Shi, and P. Zhang, "The economic impact of climate risks in china: Evidence from 47-sector panel data, 2000–2014," *Natural Hazards*, vol. 95, no. 1, pp. 289–308, Aug. 19, 2018. DOI: 10.1007/s11069-018-3447-0.

[22] L. Chen and Y. Huo, "A simple estimator for quantile panel data models using smoothed quantile regressions," *The Econometrics Journal*, vol. 24, no. 2, pp. 247–263, Aug. 5, 2020. DOI: 10.1093/ectj/utaa023.

[23] X. Guo and Q. Zheng, "Trade openness, digital economy and spatial spillover —an empirical study based on china's provincial panel data," *Advances*

in Economics and Management Research, vol. 7, no. 1, pp. 175–175, Jul. 28, 2023. DOI: 10.56028/aemr.7.1.175.2023.

[24] W. Feng, "How can entrepreneurship be fostered? evidence from provincial-level panel data in china," *Growth and Change*, vol. 52, no. 3, pp. 1509–1534, May 11, 2021. DOI: 10.1111/grow.12493.

[25] K. R. Chhabra, D. A. Telem, G. F. Chao, *et al.*, "Comparative safety of sleeve gastrectomy and gastric bypass: An instrumental variables approach.," *Annals of surgery*, vol. 275, no. 3, pp. 539–545, Nov. 12, 2020. DOI: 10.1097/sla.0000000000004297.

[26] Y. Xiao, W. Yan, and B. Peng, "Explore the complex interaction between green investment and green ecology: Evaluation from spatial econometric models and china's provincial panel data," *Sustainability*, vol. 15, no. 12, pp. 9313–9313, Jun. 8, 2023. DOI: 10.3390/su15129313.

[27] J. Song, S. Zhang, F. Tong, J. Yang, Z. Zeng, and S. Yuan, "Outlier detection based on multivariable panel data and k-means clustering for dam deformation monitoring data," *Advances in Civil Engineering*, vol. 2021, no. 1, Dec. 21, 2021. DOI: 10.1155/2021/3739551.

[28] Y. Yang, "Does economic growth induce smoking?—evidence from china," *Empirical Economics*, vol. 63, no. 2, pp. 821–845, 2022.

[29] H. Zhao, S. Guo, and H. Zhao, "Quantifying the impacts of economic progress, economic structure, urbanization process, and number of vehicles on pm2.5 concentration: A provincial panel data model analysis of china.," *International journal of environmental research and public health*, vol. 16, no. 16, pp. 2926–, Aug. 15, 2019. DOI: 10.3390/ijerph16162926.

[30] J. Dong, M. Zhang, and G. Cheng, "Impacts of upgrading of consumption structure and human capital level on carbon emissions—empirical evidence based on china's provincial panel data," *Sustainability*, vol. 14, no. 19, pp. 12 373–12 373, Sep. 28, 2022. DOI: 10.3390/su141912373.

[31] W. Gao, J. Cheng, and J. Zhang, "The influence of heterogeneous environmental regulation on the green development of the mining industry: Empirical analysis based on the system gmm and dynamic panel data model," *Chinese Journal of Population Resources and Environment*, vol. 17, no. 2, pp. 154–175, Mar. 19, 2019. DOI: 10.1080/10042857.2019.1574456.

[32] L. Jia, X. Hu, Z. Zhao, B. He, and W. Liu, "How environmental regulation, digital development and technological innovation affect china's green economy performance: Evidence from dynamic thresholds and system gmm panel data approaches," *Energies*, vol. 15, no. 3, pp. 884–884, Jan. 26, 2022. DOI: 10.3390/en15030884.

[33] G. GEORGE-EDUARD, M. RADU-CRISTIAN, N. SIMONA, and V. OANA, "A panel data analysis in estimating the economic growth for oil-producing countries. evidence from the caspian region," *ECONOMIC COMPUTATION AND ECONOMIC CYBERNETICS STUDIES AND RESEARCH*, vol. 56, no. 4/2022, pp. 225–242, Dec. 17, 2022. DOI: 10.24818/18423264/56.4.22.14.

[34] null Peinkofer, null Schwieterman, and null Miller, "Last-mile delivery in the motor-carrier industry: A panel data investigation using discrete time event history analysis," *Transportation Journal*, vol. 59, no. 2, pp. 129–164, Apr. 1, 2020. DOI: 10.5325/transportationj.59.2.0129.