Received: 6 July 2022 | Revised: 8 April 2023 | Accepted: 1 June 2023

DOI: 10.1002/asmb.2787

RESEARCH ARTICLE

WILEY

Deep partial least squares for instrumental variable regression

Maria Nareklishvili[1] | **Nicholas Polson**[1] | **Vadim Sokolov**[2]

[1]Booth School of Business, University of Chicago, Chicago, Illinois, USA

[2]Department of Systems Engineering and Operations Research, George Mason University, Fairfax, Virginia, USA

Correspondence
Nicholas Polson, Booth School of Business, University of Chicago, Chicago, IL, USA.
Email: ngp@chicagobooth.edu

Funding information
Norges Forskningsråd

Abstract

In this paper, we propose deep partial least squares for the estimation of high-dimensional nonlinear instrumental variable regression. As a precursor to a flexible deep neural network architecture, our methodology uses partial least squares for dimension reduction and feature selection from the set of instruments and covariates. A central theoretical result, due to Brillinger (2012) Selected Works of Daving Brillinger. 589-606, shows that the feature selection provided by partial least squares is consistent and the weights are estimated up to a proportionality constant. We illustrate our methodology with synthetic datasets with a sparse and correlated network structure and draw applications to the effect of childbearing on the mother's labor supply based on classic data of Chernozhukov et al. Ann Rev Econ. (2015b):649–688. The results on synthetic data as well as applications show that the deep partial least squares method significantly outperforms other related methods. Finally, we conclude with directions for future research.

KEYWORDS

dimensionality reduction, deep learning, instrumental variables, partial least squares

1 | INTRODUCTION

Nonlinear instrumental variable (IV) regression is a vital tool for estimating the causal effect of exposure on certain outcomes. In fact, virtually all legitimate techniques for causal inference can be seen as manifestations of instrumental variables, including but not limited to randomized clinical trials (considered a perfect instrument), intention-to-treat analysis (stemming from random incentive allocation), natural experiments, and regression discontinuity.[1] A valid instrumental variable meets the following assumptions: it has a significant effect on the treatment (i.e., relevance). It does not influence the outcome directly, through channels other than the treatment (i.e., the exclusion restriction), and a valid instrument is not associated with the unobserved characteristics that affect the outcome (i.e., exogeneity). Under these assumptions, the IV approach recovers consistent coefficients of interest. Traditionally, the techniques to estimate the causal parameter in the IV approach are well-suited for low and middle-scale data.[2,3] These techniques most often fail with high-dimensional instruments[4] and struggle to extract meaningful insights when instruments sparsely relate to each other. Several papers investigate the approach with many, possibly correlated and sparse instrumental variables.[5-7]

For predictive tasks, contemporary research has predominantly centered on narrow deep neural networks,[8-11] which are distinguished by their "self-featurizing" basis functions. In other words, feature extraction and dimension reduction are integrated into the pattern-matching algorithm. However, there are limitations to this approach, including a lack of theoretical understanding, difficulty in quantifying uncertainty, and limited capacity for probabilistic reasoning. To enhance the efficiency of deep neural networks, some authors introduce linear methods as a precursor to implementing a deep learning model.[12,13] use partial least squares (PLS) to extract features before learning the output with deep neural networks. PLS is a linear method that selects relevant dimensions that are particularly relevant for predicting the outcome.

To mitigate the issues inherent to high-dimensional instrumental variables, we propose deep partial least squares for IV regression. One can view our approach as merging two cultures of machine learning tools and statistical inference.[14-16] Specifically, we employ linear techniques, such as partial least squares to identify meaningful instruments and form predictors that are ensemble averages of deep learners. We show that, under several data-generating process assumptions, the method exhibits desirable large sample properties.

A traditional approach to estimating the IV regression is a two-stage least-squares (2SLS) technique. In the first stage, valid instrumental variables predict the treatment. In the second stage, we estimate the effect of the predicted treatment (and possibly other control variables) on the outcome of interest.[17] When the relation of the treatment and instruments is nonlinear,[3] show that the 2SLS method is inconsistent.[18] uses a re-centered and re-scaled outcome variable and proves the consistency of the method for censored and truncated data[3,19] use predicted treatment residuals in addition to the treatment and other exogenous variables in the second-stage regression. Controlling for residuals in the outcome regression has a long history.[20-22] The methods introduced by References 3,19 can also be identified as control function methods.

Our goal in this article is to extend the deep partial least squares (DPLS) method to extract relevant instruments in the first stage. More importantly, when the observed outcome variables consist of errors, we show that the method is consistent under a re-centered and re-scaled outcome described by Reference 18 or after controlling for the predicted residuals in the second stage regression as illustrated by Reference 21. The first layer of DPLS consists of a partial least squares method to extract features via hyperplanes in a high-dimensional setting. The subsequent layers efficiently post-process them based on ReLU networks as a deep learner.[23] PLS is an attractive method for feature extraction. However, as an alternative, the generalized method of moments (GMM[24]) can also be applied to our feature selection stage to provide variance stabilized estimators. This results in a statistical improvement on traditional stochastic descent estimators that are commonplace in machine learning (see, e.g., Reference 25).

Several related articles apply instrumental variables to address challenges and research questions in business and industry.[26] examine the effect of job training program participation on earnings. By comparison,[27] investigate the effect of 401(k) participation on wealth.[28] mine text sentiment from user-generated comments on an online service platform, and subsequently estimate the impact of (predicted) sentiment on buyers' purchasing decisions.[29] estimate the effect of the airline ticket price on demand. Another related article[30] builds a boosted tree model to identify minimum wage workers based on their demographics, and then examines the effect of minimum wage policies on labor market outcomes for these workers. In this article, we revisit the effect of childbearing on women's labor supply.[31] The instruments consist of the second and third kids being twins and the interactions with parental characteristics that are relevant for predicting the mother's labor supply.

In addition, we conduct two simulation experiments to demonstrate the predictive performance of DPLS in IV regression. The aim of these experiments is to mimic high-dimensional and sparse data, most often prevalent in business, finance, and economics. In the first experiment, we consider a high-dimensional IV space, where some of the instruments are redundant for predicting the treatment. The instruments are uncorrelated with each other. In the second experiment, we extend the first one by introducing a network structure among the instruments, as proposed by Reference 32. In this setup, the instruments are sparsely correlated, and some of them are not related to the treatment. Our aim is to illustrate whether DPLS can identify relevant instruments and improve prediction accuracy in both scenarios.

Statistical properties of deep neural networks are sparse but growing.[33] discuss theoretical foundations of sparse rectified-linear-unit (ReLU) networks and the advantage of using Spike-and-Slab prior as an alternative to Dropout. They show that the resulting posterior prediction of ReLU networks with Spike-and-Slab regularization converge to a true function at a rate of $log^{\delta}(n)/n^{-K}$ for $\delta > 1$ and a positive constant K (with n number of observations).[34] provide a theoretical connection between the Spike-and-Slab priors and L_0 norm regularization. They demonstrate that the regularized estimators can result in improved out-of-sample prediction performance. To emphasize the advantages of regularization in deep neural networks,[35] discuss theoretical and empirical justifications (as well as challenges) for the horseshoe prior.

By comparison,[36] derive the sharp upper bound for the number of activation function regions in ReLU neural networks. They find that this number in practice is far from the maximum possible and depends on the number of

neurons in the network, rather than the depth. A recent paper by Reference 37 provides groundbreaking results on the asymptotic theory of deep neural networks. Under the assumption that the number of hidden layers, that is, the depth of the network, grows with the sample size, they provide a high probability convergence rate for ReLU neural networks.

Our paper sheds light on the dimension reduction importance in sparse ReLU networks. Specifically, DPLS provides a simple, interpretable framework for modeling instrumental variables when the policy (treatment) and outcome are measured with errors. DPLS consists of a system of equations that can theoretically be viewed as an infinite sequential generalization of 2SLS. Beyond the consistency of this method, we demonstrate that the shrinkage in the first layer can substantially improve prediction performance in an instrumental variable regression with many instruments.

The results based on simulated experiments and the application show that DPLS significantly outperforms other related methods. Specifically, the first stage prediction performance improvement of DPLS relative to OLS, LASSO,[38] and DeepIV[8] is more than 79%, 57%, and 22%, respectively. We find that a deep, multi-layered structure of DPLS significantly increases predictive performance and representation learning ability. To incorporate the uncertainty of the model coefficients, we also consider the extension of DPLS to a Bayesian framework and discuss implications.

One area for future research is the study of full uncertainty quantification. It is well known that posterior distributions for IV regression require careful assessment dating back to Reference 39 (see also References 40-42 that discuss parameter uncertainty and variable selection in linear factor models).[43] provide a fully Bayesian model for posterior uncertainty for causal inference using Bayesian Additive Regression Trees (BART) to model the nonlinearity in the outcome equation. This provides a gold standard for comparison to a Bayesian DPLS method.

The rest of the paper is outlined as follows. Section 2 describes our general nonlinear IV model and specific Tobit variation. In Section 3, we discuss dimensionality reduction with partial least squares. In Section 4, we introduce deep partial least squares and examine the asymptotic theory. Section 5 illustrates the applications of the method. Finally, Section 6 concludes with directions for future research.

2 | NONLINEAR IV MODEL

One goal in this article is to predict the outcome $y^\star$ that possibly nonlinearly depends on a policy (treatment) $p^\star$ and predictors x for each observation $i = 1 \ \dots \ , n$:

$$y^\star = p^\star \beta + x\beta_x + u. \tag{1}$$

We assume, neither $y^\star$ nor $p^\star$ are directly observable. Instead, they consist of errors. Define the following variables

$$\begin{aligned}
(y^\star, y) \in \mathbb{R} &= \text{ the potential and observed outcome variables,} \\
(p^\star, p) \in \mathbb{R} &= \text{ the potential and observed policy variables,} \\
x \in \mathbb{R}^k &= \text{ observable features,} \\
z \in \mathbb{R}^m &= \text{ instrumental variables,} \\
(u, w, v, \varepsilon) \in \mathbb{R} &= \text{ latent/error variables that affect } (y^\star, p^\star, p, y), \text{ respectively,} \\
m + k &< n.
\end{aligned}$$

Instead of (1), we have access to the following structural equation model:

$$\begin{aligned}
p &= p^\star + v, \\
p^\star &= g(z\alpha + x\alpha_z) + w, \\
y &= \tau(y^\star) + \varepsilon.
\end{aligned} \tag{2}$$

$g(\cdot)$ and $\tau(\cdot)$ are potentially non-linear continuous functions, possibly deep learners.[8] $\tau(\cdot)$ is a known transformation. In this study, we define the Tobit model $\tau(y^\star) = 1(y^\star > 0) \cdot y^\star$. The errors v and ε are assumed to be uncorrelated with any other latent error. However, policy $p^\star$ is allowed to correlate with u. Covariates x are typically independent of w, v and u. β and β_x represent the effects of $p^\star$ and x on y, while α and α_z are the effects of instruments z and covariates x on

the policy $p^\star$, correspondingly. The errors w and u are correlated. For example, consider, y is a customer's decision to buy an airline ticket (observed for a particular group of customers), and $p^\star$ is the price of this ticket. In that case, policy $p^\star$ is said to be endogenous when, conditional on x, $(p^\star, u)$ correlate, that is, $\mathbb{E}(u|x, p^\star) \neq 0$. For example, ticket prices might increase during conferences that are unobservable to a researcher, and in that case, $\mathbb{E}(p^\star u|x) \neq 0$. Classical estimation methods, such as OLS, will lead to a spurious positive relation between prices and sales.

Define the joint error $\eta = w + v$. Moreover, from (2) we see that $p^\star = p - v$. In that case, (1) becomes

$$y^\star = p\beta + x\beta_x + \xi, \tag{3}$$

where $\xi = u - v\beta$. Then by combining (3) and (2), we get:

$$\begin{aligned} y &= \tau(p\beta + x\beta_x + \xi) + \varepsilon, \\ p &= g(z\alpha + x\alpha_z) + \eta, \end{aligned} \tag{4}$$

where the latent errors ξ and η are correlated through w. The structural equation model in (4) defines the nonlinear IV framework. In a traditional nonlinear IV setup (when $\tau(\cdot)$ is an identity function), the presence of valid instruments z that satisfy Assumptions 1–3 allows us to predict the unbiased mean outcome.

Assumption 1 (Relevance). Instruments z strongly relate to policy p, that is, the density of p, $F(p|z, x)$, is not constant in z.

Assumption 2 (Exclusion Restriction). z is conditionally orthogonal to outcome y:

$$z \perp y | \ (x, p, \xi).$$

Assumption 3 (Exogeneity). z is conditionally orthogonal to the latent error term ξ:

$$z \perp \xi | x.$$

Specifically, when $y = p\beta + x\beta_x + \epsilon$ with $\epsilon = \xi + \varepsilon$, valid instruments efficiently separate information in p that is unrelated to ξ, and result in the consistent estimate of the effect of the policy on the outcome (β). For example, if the fuel cost represents an instrument for ticket prices, and is unrelated to conferences, it can recover the exogenous variation in ticket prices and correctly estimate the negative relation between prices and ticket sales.

Under Assumptions 1–3, a standard approach to estimating β is a 2SLS method. The method entails predicting $\hat{p}$ in the first stage ("treatment network"). Then the predicted policy $\hat{p}$ replaces p in the outcome equation, and we estimate a second stage regression ("outcome network") with another consistent method.[8,17,31]

Define the predicted mean outcome:

$$\mathbb{E}(y|x, z) = \mathbb{E}[f(p, x)|x, z] + \mathbb{E}[\xi|x]. \tag{5}$$

Then we can evaluate the effect of a marginal change in policy (e.g., prices) from p_0 to p_1 on the outcome of interest (treatment effect):

$$\mathbb{E}(y|p_1, x) - \mathbb{E}(y|p_0, x) = \mathbb{E}[f(p_1, x)|x, z] - \mathbb{E}[f(p_0, x)|x, z].$$

The issue is that $\tau(.)$ typically is not an identity function. In such nonlinear regression models, latent errors are no longer additively separable from the true regressors ($\hat{p}$ and u are still related), and hence, the true relationship breaks down with errors in variables. The orthogonality condition of the instruments and the outcome error is violated: $\mathbb{E}\left(z(y - p\beta)|x\right) \neq 0$. As a result, the 2SLS estimator fails to be consistent for nonlinear errors-in-variables models.[44] Moreover, when instruments and/or covariates are close to the number of observations and nonlinearly relate to the policy and outcomes, OLS is no longer an efficient solution of (4).

The goal of this paper is to predict outcome $y^\star$ with the policy that depends on many instruments z and regressors x. To do so, we extend deep partial least squares with a recentered and rescaled outcome[18] and additionally illustrate it with a control function approach.[3]

2.1 | Prediction in low-dimensional data

In this section, our attention is directed towards predicting the outcome using a Tobit model, wherein $\tau(y^\star) = 1(y^\star > 0)y^\star$. We demonstrate that by recentering and rescaling the outcome, two-stage least squares (2SLS) can yield a consistent estimator of β. Alternatively, the control function approach can provide a valid prediction of the outcome.

2.1.1 | 2SLS with the Tobit model

To establish consistency, we impose an essential assumption regarding the underlying process responsible for generating the observed data.

Assumption 4 (Elliptical distribution). $(y^\star, z)$ have a joint elliptical (or Gaussian in the simplest setting) distribution.

Assumption 4 requires that the instrumental variables and the outcome are jointly elliptically distributed. The assumption can be strong in many settings. Though References 14 and 18 show that the results are robust to significant deviations from the assumption. This assumption can also be relaxed and generalized to other distributions (see e.g., Reference 45).

Define the recentered and rescaled output variable that mimics the unobserved outcome $y^\star$,

$$\tilde{y} = \psi_1^{-1}(y - \psi_2), \quad \text{where} \quad \psi_1 = \text{cov}(y, y^\star)/\text{var}(y^\star), \psi_2 = \mathbb{E}(y) - \psi_1 \mathbb{E}(y^\star).$$

Specifically, ψ_1 is the ratio of the covariance between y and $y^\star$ to the variance of $y^\star$, while ψ_2 is the intercept term when we consider a linear projection of y on $y^\star$. Next, consider a linear projection of z on $y^\star$

$$\mathbb{E}(z|y^\star) = \mathbb{E}(z) + \frac{\text{cov}(z, y^\star)}{\text{Var}(y^\star)}\left[y^\star - \mathbb{E}(y^\star)\right].$$

Hence,

$$\mathbb{E}(z|y^\star) - \mathbb{E}(z) = \gamma(y^\star - \mathbb{E}(y^\star)), \tag{6}$$

where $\gamma = \text{cov}(z, y^\star)/\text{var}(y^\star)$ [1]. By Stein's lemma,[46] we can calculate the covariance of instruments and rescaled output as

$$\text{cov}(z, \tilde{y}) = \psi_1^{-1} cov(z, y) = \psi_1^{-1} \mathbb{E}_{y^\star}\left(\mathbb{E}\left[(z - \mathbb{E}(z))y|y^\star\right]\right). \tag{7}$$

Conditional expectations are linear under elliptical contours. Additionally, $\mathbb{E}(y|y^\star) = \tau(y^\star)$, $\gamma = \text{cov}(z, y^\star)/var(y^\star)$, and replacing the last equality from (7) with (6) gives

$$\begin{aligned}
\text{cov}(z, \tilde{y}) &= \psi_1^{-1} \mathbb{E}_{y^\star}\left(\mathbb{E}(\gamma(y^\star - \mathbb{E}(y^\star))y|y^\star\right) \\
&= \gamma\psi_1^{-1}\mathbb{E}((y^\star - \mathbb{E}(y^\star))y) = \gamma\psi_1^{-1}\text{cov}(y^\star, y) \\
&= \psi_1^{-1}\frac{cov(z, y^\star)}{var(y^\star)}\text{cov}(y, y^\star) = \psi_1^{-1} cov(z, y^\star)\psi_1 = cov(z, y^\star).
\end{aligned}$$

For additional details, see Reference 18 2SLS first regresses p on z and x to get the predicted policy $\hat{p}$. Define $\bar{Z} = [z, x]$ and $\hat{p} = P_z\bar{Z}$ where $P_z = \bar{Z}(\bar{Z}^T\bar{Z})^{-1}\bar{Z}^T$ is the projection matrix onto the instrumental variable space (including covariates). In addition, define $\bar{P} = [\hat{p}, x]$ and let $e = \tilde{y} - p\beta$ be the residual from the re-scaled regression, then

$$\text{cov}(z, e|x) = \text{cov}(z, \tilde{y} - p\beta|x) = \text{cov}(z, y^\star - p\beta|x) = \text{cov}(z, u|x) = 0.$$

The re-scaled instrumental variable estimator is given by

$$\hat{\beta}^{GMM} = (\bar{Z}^T P_z \bar{Z})^{-1} P_z \bar{Z}^T \tilde{y} = (\bar{P}^T\bar{P})^{-1}\bar{P}^T\tilde{y} \tag{8}$$

[1] Elliptical contours allow for discrete outcomes.

From the above, we then have $\mathbb{E}(\hat{\beta}^{GMM}) = \beta_p$ with $\beta_p = [\beta, \beta_x]$ and $\mathrm{E}(\bar{y}) = \bar{P}\beta_p$. Note that $\hat{\beta}^{GMM}$ is a $(m+k)\times 1$ vector, where the first element is the policy effect on outcome.

2.1.2 | Estimating the proportionality constants

In the Tobit case, the constants ϕ_1, ϕ_2 can be calculated theoretically from the model as

$$\phi_1 = \Phi(\delta) \quad \text{and} \quad \phi_2 = \sigma_y^\star \phi(\delta),$$

where $\Phi(\delta)$ and $\phi(\delta)$ are the cumulative and probability density functions of a normally distributed random variable, respectively, and

$$\delta = \mu_X^T \beta / \sigma_y^\star \quad \text{and} \quad (\sigma_y^\star)^2 = \sigma_{y^\star}^2 + \beta^T \Sigma_{xx} \beta.$$

Reference 18 shows that, with the truncated outcome variable, we can obtain their estimators:

$$\hat{\psi}_1 = \hat{\Phi} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y > 0),$$
$$\hat{\psi}_2 = \hat{\sigma}_y^\star \hat{\phi} \ \text{ where } \hat{\phi} = \phi\left(\Phi^{-1}(\hat{\psi}_1)\right).$$

The estimator of the variance of the outcome variable is given as

$$\hat{\sigma}_y^{\star 2} = \frac{1}{n\ c(\hat{k})}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

where

$$c(\hat{k}) = \hat{\Phi} - (\hat{\phi} - \Phi^{-1}(\hat{\Phi})(1-\hat{\Phi}))(\hat{\phi} + \Phi^{-1}(\hat{\Phi})\hat{\Phi}),$$

and $\hat{\Phi} = \hat{\psi}_1$ and ϕ are the cumulative and probability density functions of a normally distributed random variable, respectively. $\mathbb{1}(y > 0)$ is a binary variable and equals one if the outcome is positive[18] shows that the scaling constant $\widehat{\psi}_1$ and the estimator $\hat{\beta}^{GMM}$ defined in (8) are asymptotically jointly normally distributed, with

$$\sqrt{n}(\hat{\psi}_1 - \psi_1) \to \mathcal{N}(0, \hat{\psi}_1(1-\hat{\psi}_1)), \tag{9}$$

$$\sqrt{n}(\hat{\beta}^{GMM} - \beta) \to \mathcal{N}\left(0, \Sigma_\star - \Phi(1-\Phi)\beta\beta^T\right). \tag{10}$$

Based on Reference 24, $\Sigma_\star$ is the variance of the standard GMM estimator and can be computed as follows[2]:

$$\widehat{\Sigma}_\star = n(\bar{P}^T\bar{Z}\widehat{G}\bar{Z}^T\bar{P})^{-1}\bar{P}^T\bar{Z}\widehat{G}(n\widehat{A})\widehat{G}\bar{Z}^T\bar{P}(\bar{P}^T\bar{Z}\widehat{G}\bar{Z}^T\bar{P})^{-1},$$
$$\widehat{A} = \frac{1}{n}\sum_i \hat{e}_i^2 \bar{Z}_i \bar{Z}_i^T, \ \text{ with } \ \hat{e}^2 = \hat{\bar{y}} - \bar{P}\hat{\beta}^{GMM},$$
$$\widehat{G} = \widehat{H}\left(\widehat{H}^{-1} - \bar{Z}^T\bar{Z}/\bar{Z}^T\widehat{H}\bar{Z}\right)\widehat{H},$$
$$\widehat{H} = \widehat{A}^{-1}.$$

[2]See also Reference 47 for the discussion of a GMM estimator for the IV regression.

2.2 | Control function approach

An alternative approach to estimate the parameters is a control function approach.[3] Instead of substituting the predicted policy in the second stage regression, we control for the predicted residuals of the first stage:

$$\hat{\eta} = p - \hat{p}, \tag{11}$$

$$y = \tau(p\beta + \hat{\eta}\beta_\eta) + \varepsilon^{2SRI}, \tag{12}$$

where the residuals of the treatment network $\hat{\eta}$ can be predicted by any consistent method. Note that ε^{2SRI} is not identical to ε due to the substitution of ξ with $\hat{\eta}$. The control function approach can be viewed as a special case of 2SLS. Specifically, the inclusion of $\hat{\eta}$ in the outcome equation of (4) allows us to control for the correlation of η and ξ, and predict the outcome.

3 | DIMENSIONALITY REDUCTION

The estimation of $\hat{\beta}^{GMM}$ in (8) rests on the assumption that the number of instruments (as well as independent variables) is strictly lower than the number of observations. Throughout the proof, we maintain this assumption.

Assumption 5. The number of covariates is strictly smaller than the number of observations, $m + k < n$, where m is the dimension of the instruments and k is the number of covariates.

Nevertheless, when the dimension of the independent variables is close to N, $\hat{\beta}^{GMM}$ captures the unwanted variation reflected in the predicted policy $\hat{p}$.[48] Since $\bar{Z}$ are independently and identically distributed elliptical random variables with covariance Σ_{zz},[14] shows that

$$\text{cov}(\bar{Z}, p) = \text{cov}(\bar{Z}, f(U)) = \text{cov}(\bar{Z}, U)\text{cov}(f(U), U)/\text{var}(U) = k\Sigma_{zz}\alpha_z, \tag{13}$$

where $\bar{Z} = [z, x]$, $U = \bar{Z}\alpha_z$ with $\alpha_z = [\alpha, \alpha_x]$, and the constant $k = \text{cov}(f(U), U)/\text{var}(U)$. Based on this result,[14] shows that the OLS coefficient is a consistent estimate of α up to a proportionality constant. In this article, we show that the deep partial least squares method has the same property.

3.1 | Partial least squares

Partial least squares is a dimensionality reduction method that generalizes and combines features from the principal component analysis and multiple regression.[49]

Consider the augmented instrumental variable $\bar{Z} = [z, x]$ with the dimension $n \times (m + k)$, where m and k are the dimensions of instruments and covariates, respectively. The policy $p = P$ is a $n \times 1$ vector as before. PLS can be summarized by the following relationship:

$$\begin{aligned} \bar{Z} &= TV + F, \\ P &= UQ + E, \end{aligned} \tag{14}$$

where T and U are $n \times L$ projections (scores) of $\bar{Z}$ and P, respectively. V and Q are orthogonal projection matrices (loadings). Maximizing the covariance between the augmented instruments and the policy leads to the first PLS projection pair (v_1, q_1):

$$\begin{aligned} &\max_{v,q} (\bar{Z}v_1)^T(Pq_1) \\ &\text{subject to } ||v_1|| = ||q_1|| = 1. \end{aligned}$$

The corresponding scores are $t_1 = \bar{Z}v_1$ and $u_1 = Pq_1$. It is clear that the directions (loadings) for the policy P and the augmented instruments $\bar{Z}$ are the right and left singular vectors of $\bar{Z}^T P$, respectively. PLS in the next step performs an ordinary

regression of U on T, namely $U = T\beta$. Then the next projection pair (v_2, q_2) is found by calculating the singular vectors of the residual matrix $(\bar{Z} - t_1 v_1^T)^T (P - T\beta q^T)$. Lastly, the final regression of interest is $U = T\beta$ (the tutorial[50] contains further details).

The key property of PLS is that it is consistent for estimating parameters of interest even in the presence of nonlinearity via a sequence of covariance calculations[51] first observed this in the Probit regression[52] show the consistency of PLS based on the result of Reference 14. One desirable property of PLS is that it has a closed-form solution. The PLS estimator of α is given as follows:[53,54]

$$\hat{\alpha}^{PLS} = \hat{R}(\hat{R}^T S_{zz} \hat{R})^{-1} \hat{R}^T s_{zp}, \tag{15}$$

where $\hat{R} = (s_{zp}, S_{zz} s_{zp}, \ldots, S_{zz}^{q-1} s_{zp})$ is the $(m+k) \times q$ matrix of the Krylov sequence with a $(m+k) \times (m+k)$ matrix S_{zz} and a $(m+k) \times 1$ vector s_{zp} defined as follows:

$$S_{zz} = \frac{\bar{Z}^T (I - 11^T/n) \bar{Z}}{n-1},$$
$$s_{zp} = \frac{(\bar{Z} - \mathbb{E}(\bar{Z}))^T (P - \mathbb{E}(P))}{n-1},$$

where I is an identity matrix and 1 is a matrix of ones. Intuitively, PLS searches for factors that capture the highest variability in $\bar{Z}$, and at the same time maximizes the covariance between $\bar{Z}$ and P. If the number of factors equals the dimension of instruments, $q = m + k$, the method is equivalent to OLS.[53]

4 | DEEP PARTIAL LEAST SQUARES FOR IV REGRESSION

A useful generalization of PLS is to consider a deep-layered feed-forward neural network structure:

$$\begin{aligned} \hat{p}^{(1)} &= f(\bar{Z}\hat{\alpha}^{PLS}), \\ \hat{p}^{(2)} &= f(\hat{p}^{(1)}\hat{\alpha}^{(2)}), \\ &\vdots \\ \hat{p}^{(L)} &= f(\hat{p}^{(L-1)}\hat{\alpha}^{(L)}), \end{aligned} \tag{16}$$

where $f(\cdot) = \max(\cdot, 0)$ is a rectified linear unit (ReLU) activation function. Note that $\bar{Z} = [z, x]$ as before and $\hat{\alpha}^{PLS}$ is a $q \times 1$ vector where q is the number of PLS factors in the treatment network. To reduce the dimensionality of the instruments, we predict the treatment in the first layer by PLS. The parameters $\alpha^{(\ell)}$ in subsequent layers $\ell = 2, \ldots, L$ can be identified by OLS or PLS. The following proposition shows that the parameters are consistent in each layer up to a proportionality constant.

Proposition 1. *Let S_{zz} and s_{zp} converge in probability to Σ_{zz} (the population variance of z) and σ_{zp} (the population covariance of z and p) when $n \to \infty$. Moreover, let there exist a pair of eigenvectors and eigenvalues (v_j, λ_j) for which $\sigma_{zp} = \sum_{j=1}^{M} \gamma_j v_j$ (with γ_j non-zero for each $j = 1, \ldots, M$). Assume also $\mathbb{E}(|g(U)|) < \infty$ and $\mathbb{E}(U|g(U)|) < \infty$ and $q = M$. Then $\hat{\alpha} = \{\hat{\alpha}^{PLS}, \hat{\alpha}^{(2)}, \ldots, \hat{\alpha}^{(L)}\}$ are consistent up to a proportionality constant.*

Proof. We follow the approach by Reference 52. Let $\alpha^\star = \Sigma_{zz}^{-1} \sigma_{zp}$. Define $R = (\sigma_{zp}, \Sigma_{zz}\sigma_{zp}, \ldots, \Sigma_{zz}^{q-1}\sigma_{zp})$. Then, based on the assumption that $S_{zz} \to \Sigma_{zz}$ and $s_{zp} \to \sigma_{zp}$ when n $\to \infty$, we have:

$$\hat{\alpha}^{PLS} \to R(R^T \Sigma_{zz} R)^{-1} \Sigma_{zz} \alpha^\star \text{ in probability when } n \to \infty.$$

The assumptions $q = M$ and $\sigma_{zp} = \sum_{j=1}^{M} \gamma_j v_j$ imply that $\alpha^\star$ is contained in the space spanned by R. Consequently, $\Sigma_{zz}^{1/2} \alpha^\star$ is contained in the space spanned by $R^\star = \Sigma_{zz}^{1/2} R$. Therefore,

$$R^\star (R^{\star^T} R^\star)^{-1} R^{\star^T} \Sigma_{zz}^{1/2} \alpha^\star = \Sigma_{zz}^{1/2} \alpha^\star,$$

and

$$R(R^{\star^T}R^{\star})^{-1}R^T\Sigma_{zz}\alpha^{\star} = \alpha^{\star}.$$

Hence, $\hat{\alpha}^{PLS} \to \alpha^{\star}$. Equation (13) implies that $\sigma_{zp} = k\Sigma_{zz}\alpha$, and therefore, $\hat{\alpha}^{PLS} \to \Sigma_{zz}^{-1}\sigma_{zp} = k\alpha$. k is defined in (13). This proves that the PLS in the first layer is consistent.

Now, consider the second layer $\ell = 2$. If $\hat{\alpha}^{(2)}$ is estimated by either OLS or PLS, (13) directly imply that $\hat{\alpha}^{(2)} = k\alpha$. Next, to run OLS (or PLS) of p on $\hat{p}^{(1)}$, we construct

$$\hat{p}^{(1)} = \max(z\hat{\alpha}^{PLS}, 0) = \max(zk\alpha^{(1)}, 0) = k\max(z\alpha^{(1)}, 0) = kz^{(1)}.$$

The model then becomes

$$\hat{p}^{(2)} = max(kz^{(1)}k_2\alpha^{(2)}, 0) = k \cdot k_2 max(z^{(1)}\alpha^{(2)}, 0) = k^{(2)}z^{(2)}. \tag{17}$$

By induction, we can consistently estimate α in each subsequent layer, up to L. See Reference 55 for the detailed discussion of such a deep learning structure.

Lastly, we note that the consistent estimator of the effect of the treatment on the outcome is defined as

$$\hat{\beta}^{GMM} = \left((\bar{P}^{(L)})^T\bar{P}^{(L)}\right)^{-1}(\bar{P}^{(L)})^T\tilde{y}, \tag{18}$$

where $\bar{P}^{(L)}$ is the augmented treatment $[\hat{p}, x]$ predicted by layer L. Moreover, the predicted outcome is $\hat{y} = \bar{P}\hat{\beta}^{GMM}$.

Define $\hat{e}^L$ as the predicted residuals from the treatment network in layer L and consider, $\tilde{p} = [p, \hat{e}^L, x]$. Then, similarly for 2SRI:

$$\hat{\beta}^{GMM} = (\tilde{p}^T\tilde{p})^{-1}\tilde{p}^T y. \tag{19}$$

and $\hat{y} = \tilde{p}\hat{\beta}^{GMM}$. ■

4.1 | Prediction and Bayesian shrinkage

One possible extension of deep partial least squares is a quantification of uncertainty in the density of the outcome. Our probabilistic model takes the form:

$$\begin{aligned}\tilde{y}|f, \bar{P} &\sim p(\tilde{y}|f, \bar{P}),\\ f &= g(\bar{P}\beta_p) + \varepsilon,\end{aligned}$$

where $\tilde{y}$ is an $n \times 1$ rescaled and recentered outcome variable as before, $\bar{P} = [\hat{p}, x]$ is an $n \times (1 + k)$ augmented (predicted) policy. $\beta_p = [\beta, \beta_x]$ is a $(1 + k) \times 1$ vector of coefficients in the outcome network. Here g is a deep partial least squares method. To estimate parameters in the first layer, the method uses the SIMPLS algorithm.[56] Subsequent layers use the stochastic gradient descent (SGD) method[57] for optimizing and training the parameters.

The key result, due to References 14 and 52, is that β_p can be estimated consistently, up to a constant of proportionality using PLS, irrespective of the nonlinearity of g. Given a specification of g, the constant of proportionality can also be estimated consistently with $\sqrt{n}$-asymptotics. It is worth noting that typically, standard SGD methods will not yield asymptotically normally distributed parameters. However, they can substantially increase the precision of the coefficients of interest.

Suppose that we wish to predict the outcome at a new level $\bar{P}^{\star}$. Then, we can use the predictive distribution to make a forecast as well as provide uncertainty bounds:

$$y_{\star} \sim p\left(y \mid g(\bar{P}^{\star}\hat{\beta}_p^{DPLS})\right).$$

The advantage of modeling a probabilistic model is the flexibility and the possibility to incorporate uncertainty in the parameters of interest. We approximate the posterior distribution of $\hat{\beta}^{GMM}$ with its asymptotic distribution based on the Bernstein-Von Mises theorem (see e.g., References 58,59). Define Data $= (\tilde{y}, p, z, x)$, then the densities $P(\hat{\beta}^{GMM}|\text{Data})$ and $P(\hat{\psi}_1)$ come from a normal distribution with the mean and variance depicted in (9). To shrink the effect of redundant instruments in the treatment network, we consider a Ridge regression estimator:[60]

$$\hat{\alpha}^{Ridge} = (\bar{Z}^T\bar{Z} + \lambda I_m)^{-1}\bar{Z}^T p. \tag{20}$$

However, as pointed out by a referee, this still underestimates the uncertainty due to the estimation of $(\hat{\beta}_p^{DPLS}, \hat{\alpha}^{Ridge})$. A fully Bayesian model with a non-uniform prior density of the coefficients can increase the precision of uncertainty bounds.

5 | APPLICATIONS

In this section, we evaluate the predictive performance of the DPLS-IV method relative to benchmark methods. We generate synthetic data to mimic the high-dimensional and sparse nature of instruments as described by Reference 61. To illustrate how well the method captures the complex nonlinear and sparse relation of covariates, our design mimics their setup. In addition to synthetic designs, this section illustrates findings on data provided by Reference 31. Throughout the experiments, we split data into two partitions. To find optimal parameters in each method, we use one sub-sample (further partitioned into train and validation data) and illustrate prediction performance measures on the other. We implement the experiments in R and the code to replicate our findings is available upon request.

5.1 | Sparse uncorrelated instruments

The data-generating process is summarized by the following structural equation model:

$$\begin{aligned} y &= f(p\beta + x\beta_x + \xi) + \varepsilon, \\ p &= g(z\alpha + \text{sigmoid}(z^2)\gamma + x\alpha_x) + w, \\ w; u &\sim \mathcal{N}(0, \Sigma), \\ \varepsilon &\sim \mathcal{N}(0, \sigma_\varepsilon^2), \\ z &\sim \mathcal{N}(0, \Sigma_z),\ x \sim N(0, \Sigma_x), \\ \alpha; \alpha_x, \gamma, \beta, \beta_x &\sim \mathcal{N}(0, 1). \end{aligned} \tag{21}$$

(21) holds for each observation $i = 1, \ldots n$, where $n = 1000$. w and ξ are correlated and jointly normally distributed with a mean vector 0 and a variance-covariance matrix $\Sigma = \begin{pmatrix} 3.000 & -0.087 \\ -0.087 & 0.010 \end{pmatrix}$. In this simulation setup, g and f are modeled by ReLU (or leakyReLu).[62] In addition to a $n \times 50$ matrix of instruments z, the treatment p contains nonlinear transformations of z based on the sigmoid function. To introduce sparsity in the design, out of 50 instruments, 10 are redundant. In particular, only forty instruments are relevant for predicting the treatment p. x is an $n \times 25$ matrix of covariates, and 20 of them have no influence on outcome y. Σ_z and Σ_x represent covariance matrices of z and x, respectively. In this setting, the covariance between the instruments is small (0.001) and the features x are uncorrelated.

Figure 1 visualizes predictions of the policy p in the treatment network (Figure A1 in Appendix A.0.1 shows predicted outcome y). In each case, g and f represent a rectified linear unit function. A visual inspection of Figures 1 and A1 verifies that DPLS-IV results in more accurate predictions relative to the other methods. In this setting, DeepIV[8] is the second-best alternative. According to Figures 1 and A1, predictions of the treatment, $\hat{p}$, appear to have a higher variability compared to the outcome predictions. This is not surprising, as p contains instruments in addition to x. Note that, even though prediction performance measures of DeepIV are close to those of DPLS-IV, Table A1 in Appendix A.0.1 shows that, compared to DeepIV, DPLS-IV is more robust to changes in the activation function.

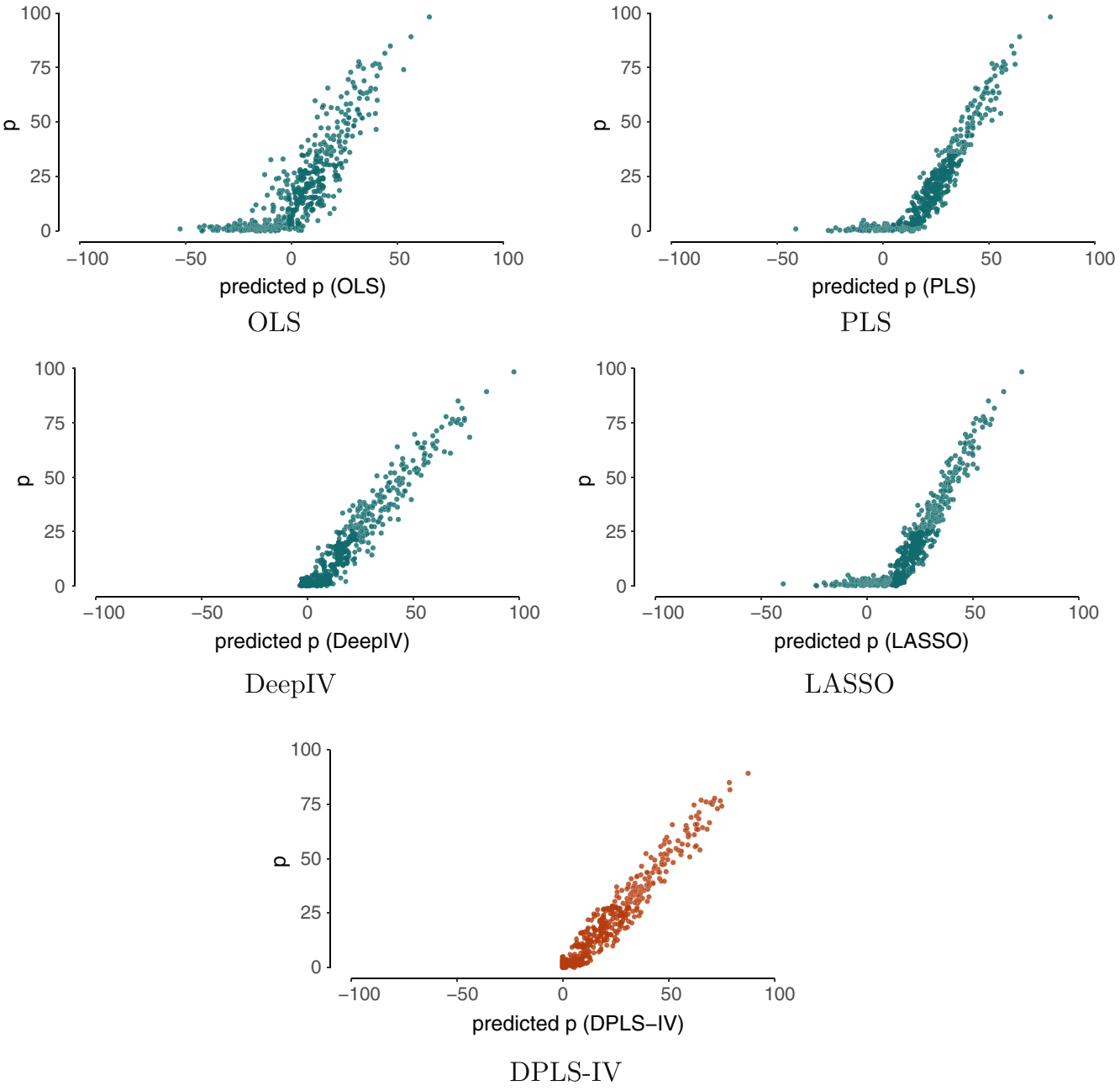

FIGURE 1 First stage prediction performance. The X-axis depicts predicted treatment $\hat{p}$, and the Y-axis represents true values of p. DPLS-IV denotes the method introduced in this study. We use test data for evaluating the methods.

To evaluate the effectiveness of different methods, we examine their out-of-sample R^2 and root mean squared error (RMSE) as we increase the parameter values of Σ. The results, as depicted in Figure 2, demonstrate that DPLS-IV exhibits robustness to increasing errors and is capable of accounting for the endogeneity of p reflected in the covariance of error terms w and ξ. Furthermore, our analysis, as shown in Table 1, indicates that DPLS-IV outperforms OLS, PLS, LASSO, and DeepIV methods by a significant margin. These findings suggest that DPLS-IV holds considerable promise as a powerful and reliable tool for predicting outcomes in the presence of endogeneity and measurement error.

To illustrate the predictive power of DPLS-IV, Figure 3 compares coefficients estimated by the OLS and PLS methods in the treatment network. Parameters estimated by PLS are closer to their true values relative to OLS. Additionally, Figure 4 shows the absolute bias of these parameters. According to Figure 4, the cumulative distribution function of the absolute bias of the parameters recovered by PLS stochastically dominates the ones based on OLS and LASSO. Specifically, smaller values of the absolute bias of the coefficients are more likely under PLS, relative to OLS and LASSO. The sum of the

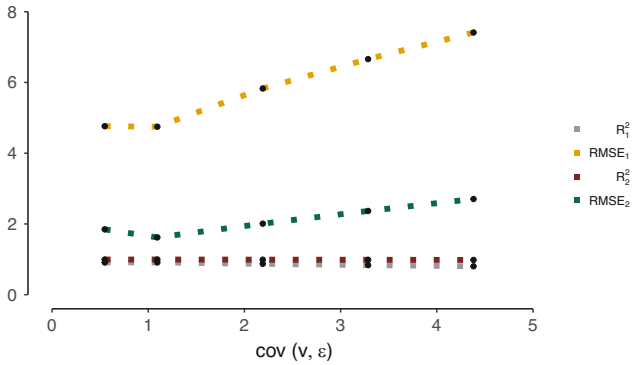

FIGURE 2 R^2 and RMSE for increasing values of cov(v, ε). The black circles represent the estimated values which are combined with the dotted lines. Each color corresponds to the corresponding R^2 and RMSE in the treatment and outcome networks. Specifically, R_1^2 and RMSE$_1$ denote prediction performance measures in a treatment network. R_2^2 and RMSE$_2$ are the prediction performance measures in the outcome network.

TABLE 1 Prediction performance of DPLS-IV relative to other methods.

Measures	PLS	OLS	DeepIV	LASSO	DPLS-IV
Treatment network					
R^2	0.751	0.688	0.932	0.753	**0.956**
RMSE	10.603	21.540	5.789	10.497	**4.508**
Outcome network					
R^2	0.932	0.933	0.889	0.933	**0.938**
RMSE	1.668	1.718	4.573	1.670	**1.622**

Note: We present out-of-sample R^2 and RMSE. To present the maximum prediction performance of OLS, PLS and LASSO in the outcome network, we use residuals predicted by DPLS-IV in the first stage.

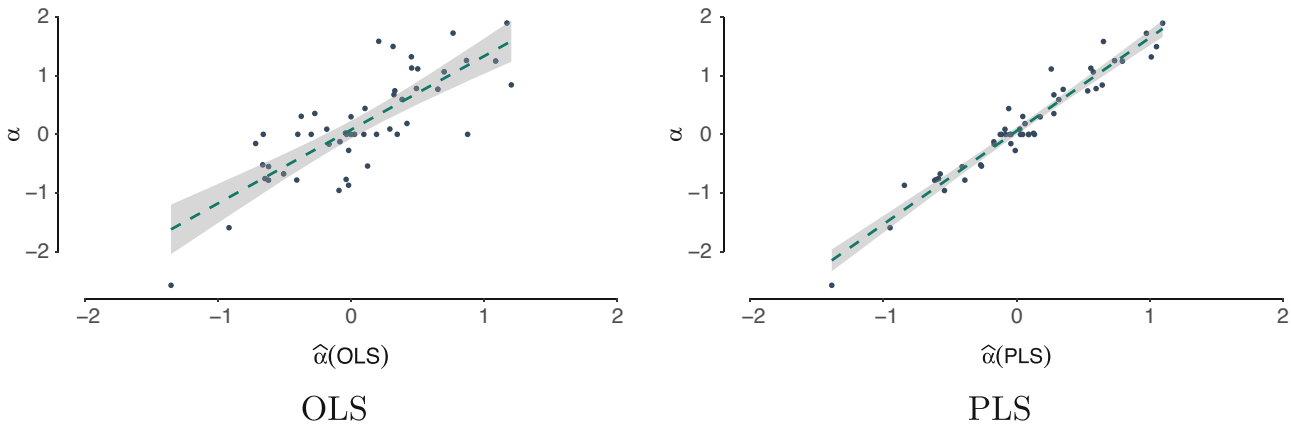

FIGURE 3 Estimated coefficients and their corresponding true values in the treatment network. The X-axis shows coefficients predicted by OLS (left) and PLS (right), and the Y-axis depicts the corresponding true values.

absolute bias is the smallest under PLS (14.406), followed by LASSO (15.233). OLS leads to the highest value of the sum of the absolute bias (21.610).

To introduce uncertainty in the parameters of interest, we extend DPLS-IV to a Bayesian setup and compare it to a classical Bayesian IV approach.

Table 2 shows that Bayesian DPLS-IV (with ReLU as an activation function and two hidden layers) significantly outperforms its' linear counterpart. Figure 5 verifies that the Bayesian DPLS-IV closely replicates the density of the original outcome variable.

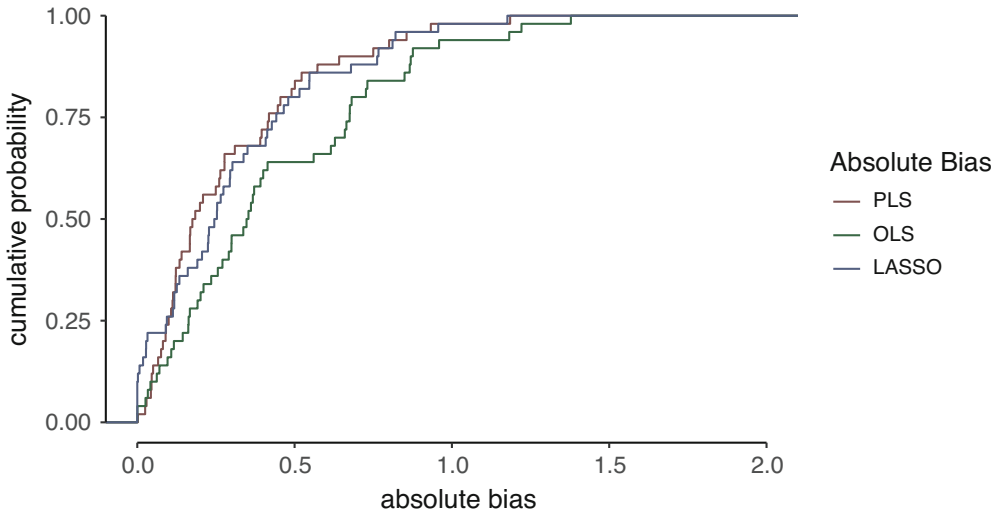

FIGURE 4 Empirical cumulative distribution functions (CDF) of the absolute bias of the parameters estimated by PLS, OLS, and LASSO.

TABLE 2 Prediction performance of Bayesian DPLS-IV and IV methods.

Measures	Bayesian IV	Bayesian DPLS-IV
Outcome network		
R^2	0.778	**0.838**
RMSE	4.375	**3.092**

Note: The measures are computed based on the mean prediction out of 10,000 predicted outcome variables.

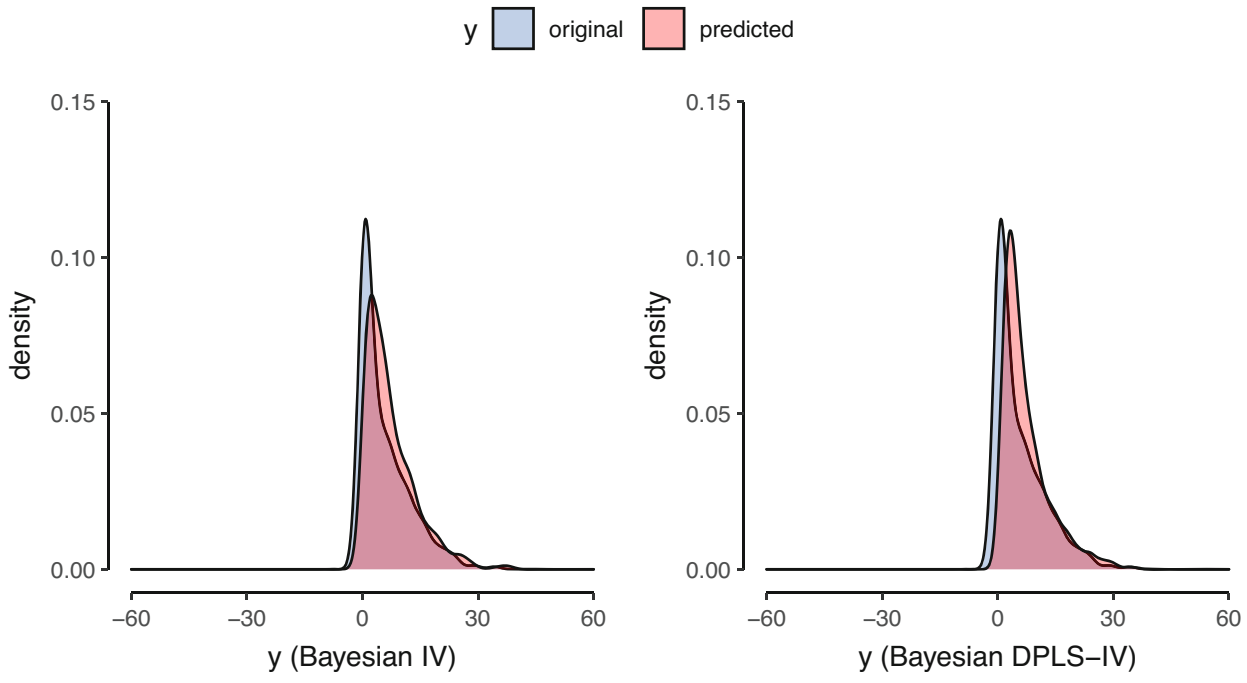

FIGURE 5 The density of the original and predicted outcome variables (on the test data).

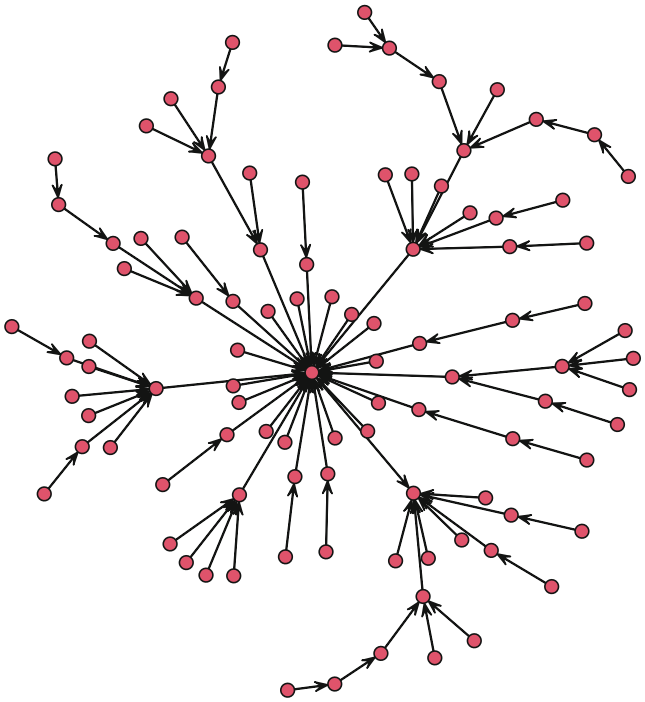

FIGURE 6 A network with 100 nodes.

5.2 | Instruments with the network structure

The second simulation closely follows the network data structure described by Reference 61. The data-generating process is the same as shown in (21). However, the instrumental variable network comes from the preferential attachment algorithm.[32] Each node of the network represents one feature. The resulting network follows a power-law degree distribution, and thus, is scale-free. That means, only a few instruments in the network have a relatively large number of "neighbors". The distance between two instruments is the shortest path between them in the network. We calculate a $p \times p$ ($p = 50$) pairwise distance matrix D. Next, this distance matrix is transformed into a covariance matrix $\Sigma_{z,(i,j)} = 0.7^{D_{(i,j)}}$, where (i, j) represents the element in each row i and column j of a matrix D $(i, j = 1, \ldots, p)$. Figure 6 shows an example of such a network with 100 nodes.

For comparability, DPLS-IV and DeepIV consist of the same number of layers and neurons in each layer. In particular, the input layer consists of 200 neurons; the hidden layers consist of 200, 100, and 50 neurons, respectively. We use ReLU as an activation function. Figure 7 shows cumulative distribution functions of the predicted outcome variable and the absolute bias of the estimated coefficients (right) in the treatment network. Based on the results in Figure 7, the outcome predicted by DPLS-IV is closer to the CDF of the simulated outcome variable. Moreover, DPLS-IV yields a CDF of the absolute bias of the parameters that stochastically dominates the CDF of the other benchmark methods.

We also investigate prediction performance in the outcome network. Table 3 shows R^2 and RMSE of DPLS-IV relative to other methods. In this setting, DPLS-IV outperforms other benchmark methods.

Additionally, we compare a Bayesian DPLS-IV to a Bayesian IV approach. Bayesian DPLS-IV consists of two hidden layers and a ReLU activation function. Note that, the RMSE based on the mean predicted outcome out of 10,000 predictions increases (2.262) relative to DPLS-IV. However, R^2 is unchanged (0.941). Figure 8 shows predicted outcome values against their true counterparts. Bayesian DPLS-IV leads to outcome predictions that are closer to the true values.

5.3 | Labor supply of women

Reference 31 examines the effect of childbearing on women's labor supply. They use a mixed sibling-sex composition and twins as instruments for the size of the family. To illustrate the method, our analysis uses 1980 U.S. Census data

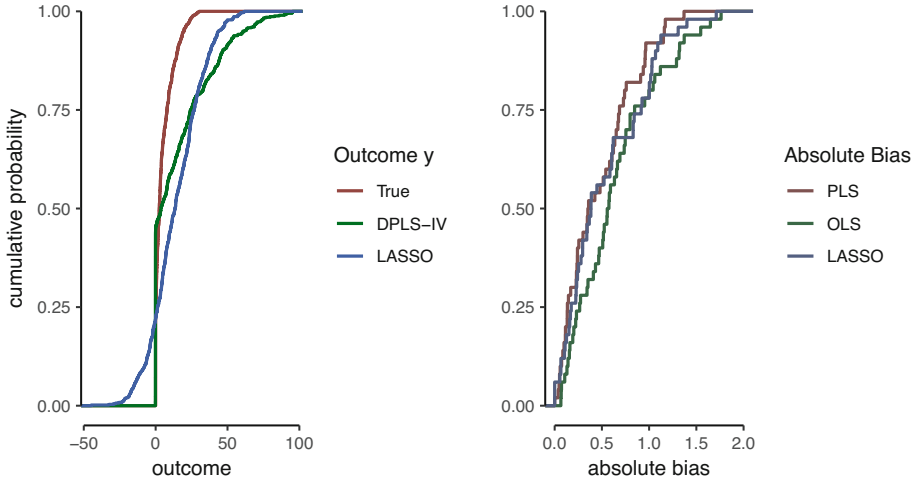

FIGURE 7 Left–CDF of the outcome variable; Right–CDF of the absolute value of the bias of the parameters (estimated by the corresponding method) in the treatment network.

TABLE 3 Prediction performance of DPLS-IV relative to other methods.

Measures	PLS	OLS	DeepIV	LASSO	DPLS-IV
Outcome network					
R^2	0.940	0.930	0.912	0.942	**0.944**
RMSE	1.631	1.846	2.108	1.609	**1.585**

Note: We present out-of-sample R^2 and RMSE. To present the maximum prediction performance of OLS, PLS, and LASSO in the outcome network, we use residuals predicted by DPLS-IV in the first stage.

that include all women with two or more children. The model that we are going to estimate is defined by the following structural equation model:

$$\begin{aligned} y &= f(\text{kids} \cdot \beta + x\beta_x + \xi) + \varepsilon. \\ \text{kids} &= g(\text{twins} \cdot \alpha + \text{twins} \cdot x\gamma + x\alpha_x) + w, \end{aligned} \tag{22}$$

where y is the outcome variable and measures the logarithm of hours worked per week by the mother. The outcome is observed when the age of the mother is more than the average age of the mothers in the population. In particular,

$$y = \log(\text{hourswm}) \times \mathbb{I}(agem > \mathbb{E}(agem)), \tag{23}$$

where $\mathbb{I}(agem > \mathbb{E}(agem))$ is an indicator variable and equals one if the age of the mother is more than the population mean, and zero otherwise. The treatment, kids, is the number of total kids in a family. twins represents an instrumental variable and equals one if the second and third children are twins, otherwise zero. Additionally, we use interactions of the instrument with covariates, twins $\cdot\, x$ as instruments for the number of kids. The covariates include the gender and age of the first and second child, the mother's age, marital status, race, education, and the age of the mother when she first gave birth. See Reference 31 for a detailed summary of the variables.

The prediction performance advantage of DPLS-IV is also clearly evident in Table 4. The first stage prediction performance of DPLS-IV is similar in each method, however, R^2 (RMSE) is considerably high (low) in the outcome network. It is worth noting that the measures become substantially worse in the outcome network compared to the ones in the treatment network. One of the potential reasons is that the outcome distribution is bimodal. Figure A2 presents the original density of the outcome variable and the density of the outcome predicted by Bayesian DPLS-IV. Figure A2 shows that Bayesian DPLS-IV successfully captures the bimodal nature of the outcome.

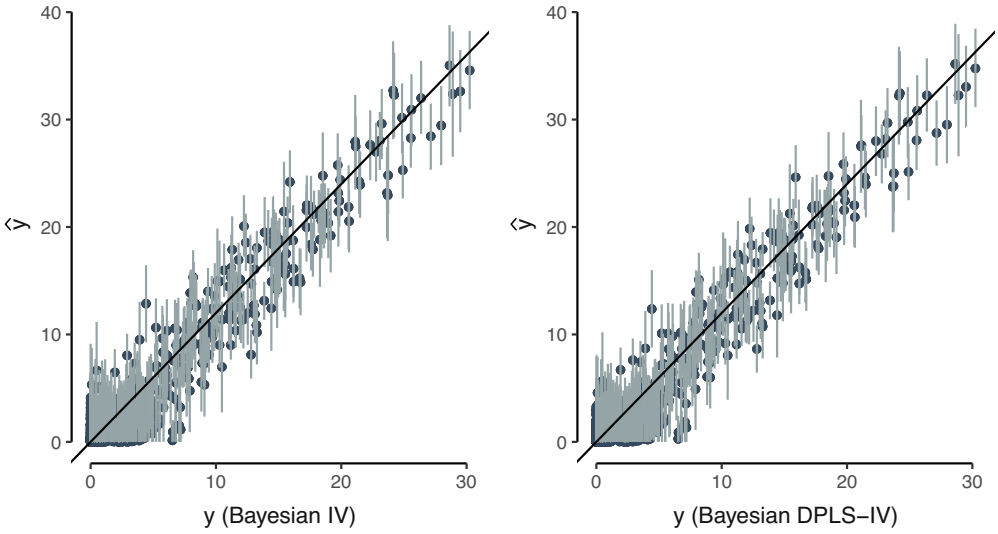

FIGURE 8 Predicted outcome y and the corresponding 95% confidence intervals.

TABLE 4 Prediction performance of DPLS-IV relative to other methods.

Measures	PLS	OLS	DeepIV	LASSO	DPLS-IV
Treatment network					
R^2	0.204	0.205	0.226	0.205	**0.233**
RMSE	0.655	0.655	**0.647**	**0.647**	**0.644**
Outcome network					
R^2	0.532	0.536	**0.771**	0.536	**0.772**
RMSE	12.838	12.782	8.980	12.780	**8.960**

Note: We present out-of-sample R^2 and RMSE. To present the maximum prediction performance of OLS, PLS, and LASSO in the outcome network, we use residuals predicted by DPLS-IV in the first stage.

We also investigate the layer-by-layer transformation of the proposed method. Figure 9 illustrates the feature representation of the original outcome y (on the test data), the score matrix (T), and the final feed-forward neural network layer in DPLS-IV. We consider the projection of features on two distinct classes of the outcome variable. In particular, when $y > 0$, we label it as a class 1, and when $y = 0$, it represents a class 0. Figure 9 shows that the intermediate layers in DPLS-IV significantly improve the representation of the original covariate space. The borders of the two classes become highly evident in the last layer of DPLS-IV.

6 | DISCUSSION

In this article, we propose deep partial least squares for reducing the dimension of the instrumental variable space. The deep partial least squares method efficiently extracts features based on partial least squares and further processes the input with a feed-forward deep learner. The method is well-tailored for correlated instruments with sparse and nonlinear structures. More importantly, deep partial least squares are consistent, up to a proportionality constant.

The applications on synthetic data as well as the application to the effect of childbearing on the mother's labor supply[31] show that the deep partial least squares method outperforms other related methods. Moreover, a flexible number of layers allows us to efficiently capture nonlinearities embedded in the instrumental variable network.

An interesting extension of this work is to consider a Bayesian model with various priors on the coefficients of interest.[64] consider asymptotic properties of Bayes risk with the horseshoe prior. We believe, investigating different prior beliefs

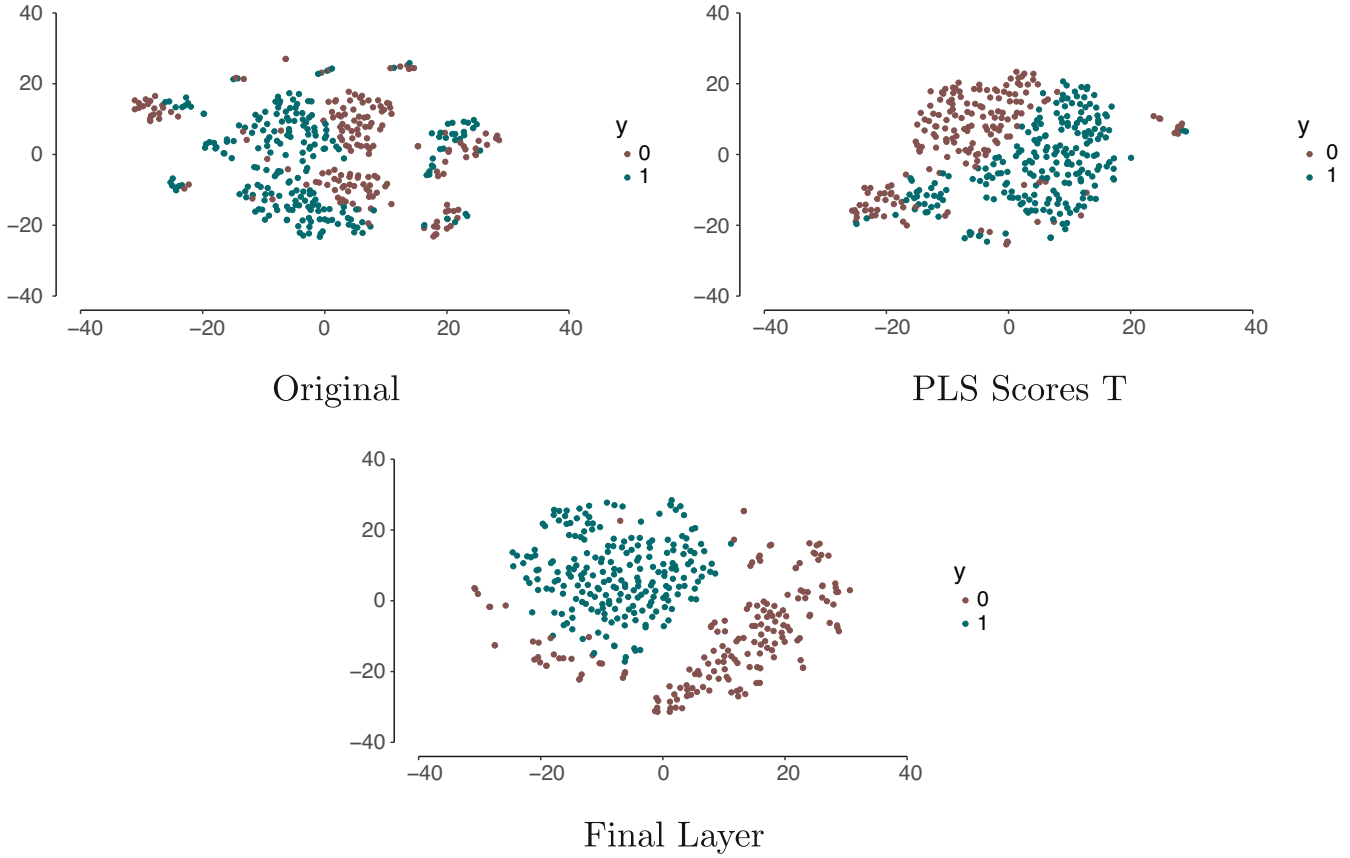

FIGURE 9 Feature representation of DPLS-IV based on the t-SNE algorithm,[63] with the perplexity parameter equal to 50. Features are reduced to two distinct classes of the outcome y, with a class 1 when $y > 0$, and 0 otherwise. The final layer represents the predicted y of individual neurons in the last layer of DPLS-IV. For computational benefits, we randomly sample 1000 observations from the original data.

can result in the increased predictive performance of deep partial least squares. Another useful extension is to draw applications to eminent domain with judge characteristics as instruments.[65] In addition to prediction problems, the consistency of DPLS allows us to address the estimation of the treatment effect. In the future, we plan to demonstrate the precision of the treatment effect and compare it to other benchmark algorithms.

ORCID

Vadim Sokolov https://orcid.org/0000-0002-6618-2965

REFERENCES

1. Heckman JJ. *Randomization as an Instrumental Variable*. National Bureau of Economic Research; 1995.
2. White H. Instrumental variables regression with independent observations. *Econometrica*. 1982;50(2):483-499.
3. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J Health Econ*. 2008;27(3):531-543.
4. McCullagh P, Polson NG. Statistical sparsity. *Biometrika*. 2018;105(4):797-814.
5. Belloni A, Chernozhukov V, Hansen C. Inference for high-dimensional sparse econometric models. arXiv preprint arXiv:1201.0220. 2011.
6. Chernozhukov V, Hansen C, Spindler M. Valid post-selection and post-regularization inference: an elementary, general approach. *Annu Rev Econ*. 2015b;7(1):649-688.
7. Chernozhukov V, Hansen C, Spindler M. Post-selection and post-regularization inference in linear models with many controls and instruments. *Am Econ Rev*. 2015a;105(5):486-490.
8. Hartford J, Lewis G, Leyton-Brown K, Taddy M. Counterfactual Prediction with Deep Instrumental Variables Networks. arXiv:1612.09596 [cs, stat]. 2016.
9. Polson NG, Sokolov VO. Deep learning for short-term traffic flow prediction. *Transp Res Part C: Emerg Technol*. 2017;79:1-17.
10. Liu R, Shang Z, Cheng G. On deep instrumental variables estimate. arXiv preprint arXiv:2004.14954. 2020.
11. Beise H-P, Da Cruz SD, Schröder U. On decision regions of narrow deep neural networks. *Neural Netw*. 2021;140:121-129.

12. Debiolles A, Oukhellou L, Aknin P. Combined use of partial least squares regression and neural network for diagnosis tasks. Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, volume 4, IEEE. 2004 573–576.
13. Jia W, Zhao D, Ding L. An optimized rbf neural network algorithm based on partial least squares and genetic algorithm for classification of small sample. *Appl Soft Comput*. 2016;48:373-384.
14. Brillinger DR. A generalized linear model with "gaussian" regressor variables. *Selected Works of David Brillinger*. Springer; 2012:589-606.
15. Sarstedt M, Hair JF, Pick M, Liengaard BD, Radomir L, Ringle CM. Progress in partial least squares structural equation modeling use in marketing research in the last decade. *Psychol Mark*. 2022;39(5):1035-1064.
16. Monis JB, Sarkar R, Nagavarun S, Bhadra J. Efficient net: identification of crop insects using convolutional neural networks. Paper presented at: 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), IEEE. 2022 1–7.
17. Mogstad M, Torgovitsky A, Walters CR. The causal interpretation of two-stage least squares with multiple instrumental variables. *Am Econ Rev*. 2021;111(11):3663-3698.
18. Iwata S. Recentered and rescaled instrumental variable estimation of Tobit and Probit models with errors in variables. *Econom Rev*. 2001;20(3):319-335.
19. Adkins LC. Testing parameter significance in instrumental variables probit estimators: some simulation. *J Stat Comput Simul*. 2012;82(10):1415-1436.
20. Florens J-P, Heckman J, Meghir C, Vytlacil E. Instrumental Variables, Local Instrumental Variables and Control Functions. Technical report, Cemmap Working Paper. 2002.
21. Heckman J, Navarro-Lozano S. Using matching, instrumental variables, and control functions to estimate economic choice models. *Rev Econ Stat*. 2004;86(1):30-57.
22. Navarro S. Control functions. *Microeconometrics*. Springer; 2010:20-28.
23. Yarotsky D. Error bounds for approximations with deep relu networks. *Neural Netw*. 2017;94:103-114.
24. Hansen LP, Singleton KJ. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*. 1982;50(5):1269-1286.
25. Iwata T, Ghahramani Z. Improving output uncertainty estimation and generalization in deep learning via neural network gaussian processes. arXiv preprint arXiv:1707.05922. 2017.
26. Abadie A, Angrist J, Imbens G. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*. 2002;70(1):91-117.
27. Chernozhukov V, Hansen C. The effects of 401 (k) participation on the wealth distribution: an instrumental quantile regression analysis. *Rev Econ Stat*. 2004;86(3):735-751.
28. Moreno A, Terwiesch C. Doing business with strangers: reputation in online service marketplaces. *Inf Syst Res*. 2014;25(4):865-886.
29. Hartford J, Lewis G, Leyton-Brown K, Taddy M. Deep iv: a flexible approach for counterfactual prediction. Paper presented at: International Conference on Machine Learning, PMLR, 2017; 2017:1414–1423.
30. Cengiz D, Dube A, Lindner A, Zentler-Munro D. Seeing beyond the trees: using machine learning to estimate the impact of minimum wages on labor market outcomes. *J Labor Econ*. 2022;40(1):S203-S247.
31. Angrist J, Evans WN. *Children and their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size*. National Bureau of Economic Research; 1996.
32. Jeong H, Néda Z, Barabási A-L. Measuring preferential attachment in evolving networks. *EPL (Europhys Lett)*. 2003;61(4):567.
33. Polson NG, Ročková V. Posterior concentration for sparse deep learning. *Adv Neural Inf Process Syst*. 2018;31:4-19.
34. Polson NG, Sun L. Bayesian l 0-regularized least squares. *Appl Stochast Models Bus Ind*. 2019;35(3):717-731.
35. Bhadra A, Datta J, Li Y, Polson N. Horseshoe regularisation for machine learning in complex and deep models 1. *Int Stat Rev*. 2020;88(2):302-320.
36. Hanin B, Rolnick D. Deep relu networks have surprisingly few activation patterns. *Adv Neural Inf Process Syst*. 2019;32:1-9.
37. Farrell MH, Liang T, Misra S. Deep neural networks for estimation and inference. *Econometrica*. 2021;89(1):181-213.
38. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Series B Stat Methodology*. 2011;73(3):273-282.
39. Zellner A. Bayesian analysis of regression error terms. *J Am Stat Assoc*. 1975;70(349):138-144.
40. Hoogerheide L, Kleibergen F, van Dijk HK. Natural conjugate priors for the instrumental variables regression model applied to the angrist–krueger data. *J Econom*. 2007;138(1):63-103.
41. Lopes HF, Polson NG. Extracting sp500 and nasdaq volatility: the credit crisis of 2007–2008. *Handbook of Applied Bayesian Analysis*, Oxford University Press; 2010:319-342.
42. Puelz D, Hahn PR, Carvalho CM. Variable selection in seemingly unrelated regressions with random predictors. *Bayesian Anal*. 2017;12(4):969-989.
43. Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal*. 2020;15(3):965-1056.
44. Amemiya Y. Instrumental variable estimator for the nonlinear errors-in-variables model. *J Econom*. 1985;28(3):273-289.
45. Adcock C. Extensions of stein's lemma for the skew-normal distribution. *Commun Stat—Theory Methods*. 2007;36(9):1661-1671.
46. Landsman Z, Nešlehová J. Stein's lemma for elliptical random vectors. *J Multivar Anal*. 2008;99(5):912-927.
47. Baum CF, Schaffer ME, Stillman S. Instrumental variables and gmm: estimation and testing. *Stat J*. 2003;3(1):1-31.
48. Gagnon-Bartsch JA, Jacob L, Speed TP. Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, Penn State University Press; 2013:1-112.

49. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev: Comput Stat*. 2010;2(4):433-459.
50. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Anal Chim Acta*. 1986;185:1-17.
51. Fisher RA. On the mathematical foundations of theoretical statistics. *Philos Trans R Soc London Ser A, Contain Papers Math Phys Char*. 1922;222(594-604):309-368.
52. Naik P, Tsai C-L. Partial least squares estimator for single-index models. *J R Stat Soc Series B Stat Methodology*. 2000;62(4):763-771.
53. Helland IS. Partial least squares regression and statistical models. *Scand J Stat*. 1990;17(2):97-114.
54. Stone M, Brooks RJ. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J R Stat Soc B Methodol*. 1990;52(2):237-258.
55. Polson N, Sokolov V, Xu J. Deep learning partial least squares. arXiv preprint arXiv:2106.14085. 2021.
56. De Jong S. Simpls: an alternative approach to partial least squares regression. *Chemom Intel Lab Syst*. 1993;18(3):251-263.
57. Bottou L. Stochastic gradient descent tricks. *Neural Networks: Tricks of the Trade*. Springer; 2012:421-436.
58. Van der Vaart AW. *Asymptotic Statistics*. Vol 3. Cambridge University Press; 2000.
59. Bhadra A, Datta J, Polson NG, Willard B. Lasso meets horseshoe: a survey. *Stat Sci*. 2019;34(3):405-427.
60. Marquardt DW, Snee RD. Ridge regression in practice. *Am Stat*. 1975;29(1):3-20.
61. Kong Y, Yu T. A deep neural network model using random forest to extract feature representation for gene expression data classification. *Sci Rep*. 2018;8(1):1-9.
62. Dubey AK, Jain V. Comparative study of convolution neural network's relu and leaky-relu activation functions. *Applications of Computing, Automation and Wireless Systems in Electrical Engineering*. Springer; 2019:873-880.
63. Wattenberg M, Viégas F, Johnson I. How to use t-sne effectively. *Distill*. 2016;1(10):e2.
64. Datta J, Ghosh JK. Asymptotic properties of bayes risk for the horseshoe prior. *Bayesian Anal*. 2013;8(1):111-132.
65. Belloni A, Chen D, Chernozhukov V, Hansen C. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*. 2012;80(6):2369-2429.

How to cite this article: Nareklishvili M, Polson N, Sokolov V. Deep partial least squares for instrumental variable regression. *Appl Stochastic Models Bus Ind*. 2023;1-21. doi: 10.1002/asmb.2787

APPENDIX

A.1 Prediction performance measures with LeakyReLu

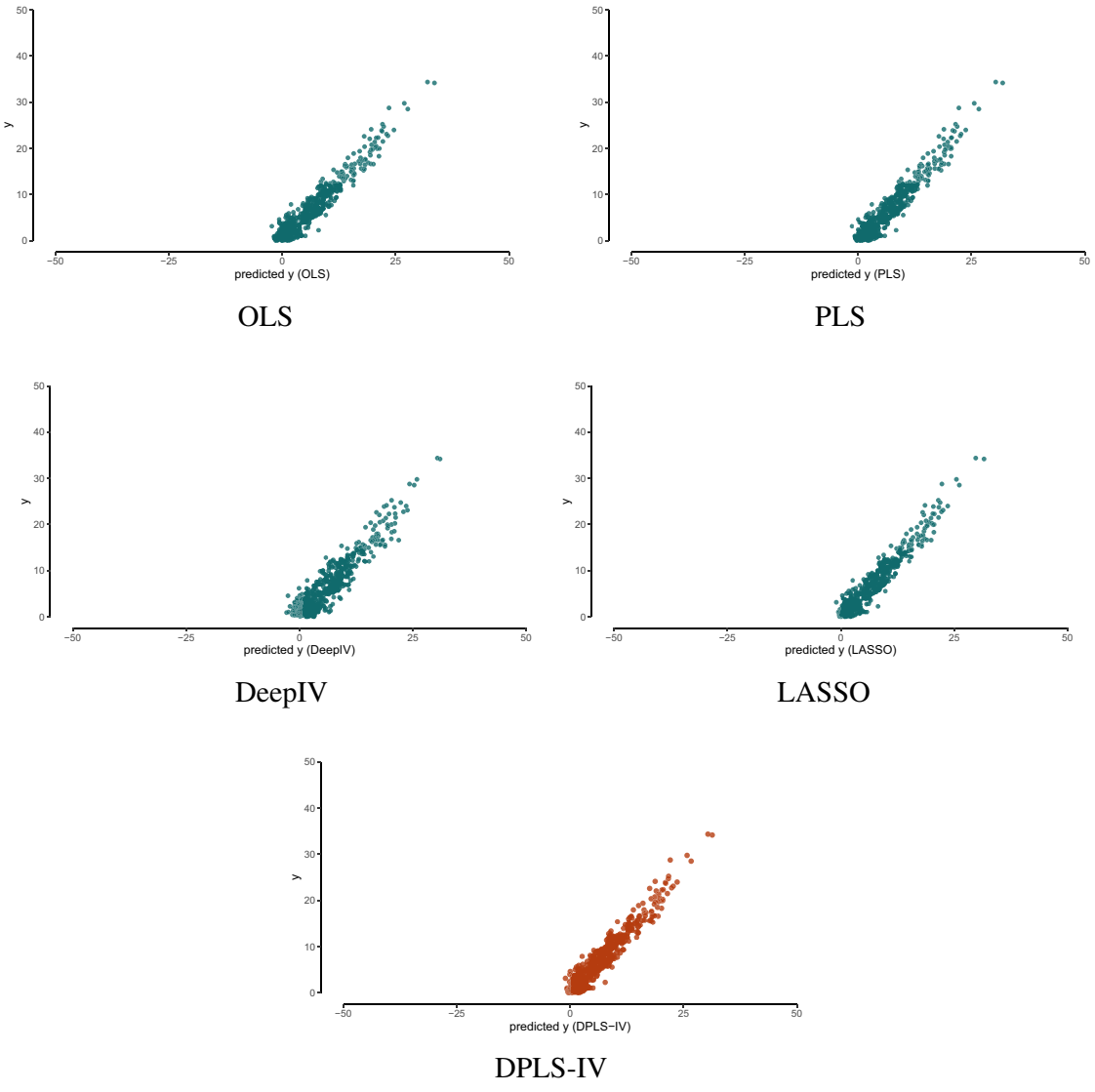

FIGURE A1 The second stage prediction performance. The X-axis depicts predicted outcome $\hat{y}$, and the Y-axis represents true values of y. DPLS-IV denotes the method introduced in this study. We use test data for evaluating the methods.

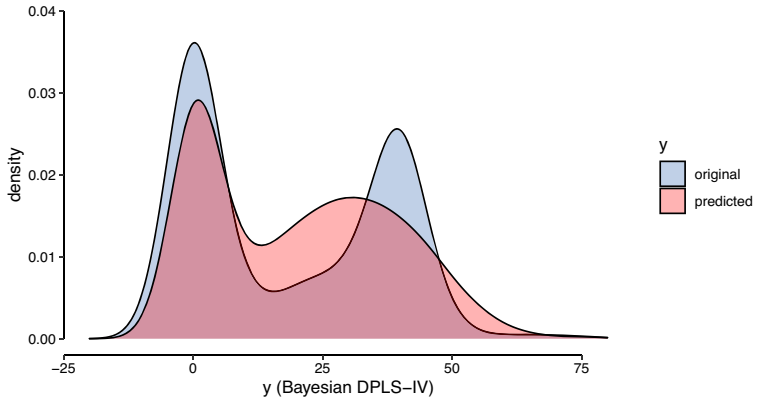

FIGURE A2 The original distribution of the outcome variable and the one predicted by Bayesian DPLS-IV. Bayesian DPLS-IV consists of two hidden layers and a ReLU activation function.

TABLE A1 Prediction performance of DPLS-IV relative to other methods.

Measures	PLS	OLS	LASSO	DeepIV	DPLS-IV
Treatment network					
R^2	0.757	0.696	0.760	0.939	**0.957**
RMSE	10.503	21.404	10.398	5.498	**4.449**
Outcome network					
R^2	0.933	0.933	0.934	0.878	**0.938**
RMSE	1.666	1.720	1.667	2.224	**1.623**

Note: We present out-of-sample R^2 and RMSE. To present the maximum prediction performance of OLS, PLS, and Tobit in the outcome network, we use residuals predicted by DPLS-IV in the first stage. We use LeakyReLU to simulate the outcome variables. DPLS-IV consists of an input layer with 50 neurons, and a hidden layer with 30 neurons. DeepIV consists of 100 neurons in the input layer, followed by three hidden layers with 100, 100, and 30 neurons, respectively. We use ReLU to activate neurons in DeepIV and DPLS-IV.