BAYESIAN NONPARAMETRIC MODEL FOR NONSEPARABLE INSTRUMENTAL VARIABLE REGRESSION *

By Simone Martinalli[‡] and Andriy Norets[†]

Brown University

We propose a Bayesian nonparametric model for nonseparable instrumental variable regression.

*We thank ...

[†]Corresponding Author, Professor, Department of Economics, Brown University

[‡] PhD student, Department of Economics, Brown University

Keywords and phrases: Nonseparable IV model, Bayesian nonparametrics, mixtures of normal distributions, smoothly mixing regressions, mixtures of experts.

1. Introduction. Instrumental variable (IV) models have become a cornerstone of empirical research in economics, especially in the context of causal inference. Most commonly, applied researchers estimate linear IV regression models using the two stage least squares (TSLS) method. While the classical econometric theory provides results on identification and estimation of nonlinear and nonseparable IV models (see, for example, Chesher (2003), Chesher (2005), Matzkin (2008), and Imbens and Newey (2009)), they are less frequently used in applied work due to the complexity of available inferential procedures. Moreover, discrete variables are ubiquitous in applications, and in such settings, causal effects are typically set identified (see Kitagawa (2021) and references therein) with very wide confidence intervals, which further diminishes the appeal of these procedures among practitioners.

Bayesian literature on nonparametric IV models is not large. Conley et al. (2008) proposed to use Dirichlet process mixtures to nonparametrically model the distribution of errors in a Bayesian linear IV model; McCulloch et al. (2021) extended this approach by using Bayesian additive regression trees to nonparametrically model the regression functions. To the best of our knowledge, nonseparable IV models have not been considered in the Bayesian framework.

In this paper, we propose a flexible Bayesian model for estimation of nonseparable IV regressions with a univariate endogenous covariate and a univariate unobservable responsible for endogeneity. Our model uses nonparametric priors for conditional distributions from Norets and Pelenis (2022a) for modeling the conditional distribution of the outcome given the endogenous covariate, the exogenous covariates, and the unobservable and the conditional distribution of the endogenous covariate given the instruments, the exogenous covariates, and the unobservable. The unobservable is assumed to be independent of the instruments conditional on the exogenous covariates and its distribution can be set to an arbitrary distribution without a loss of generality. The distribution of exogenous covariates and instruments is not modeled.

The conditional distribution model in Norets and Pelenis (2022a) is based on covariate dependent mixtures with a prior on the number of mixture components. The model can accommodate discrete and continuous dependent and independent variables. Norets and Pelenis (2022a) show that in their model the posterior contracts at an optimal (up to a log factor) rate that is adaptive to unknown underlying anisotropic smoothness in the continuous variables and also possible anisotropic smoothness and sparsity in the discrete variables as defined in Norets and Pelenis (2022b). The fact that the model exploits possible smoothness in discrete variables is essential for good model performance in handling discrete variables nonparametrically. Norets and Pe-

lenis (2022a) and Norets and Pelenis (2022b) also demonstrate that their model outperforms cross-validated kernel estimators and parametric models in simulations and applications. At this point, we conjecture that at least some of these excellent theoretical properties will be inherited by our IV model.

From the applied researcher's perspective, our approach has a number of appealing features. First of all, the model can be set up with a fixed number of mixture components and when only one mixture component is used the model becomes a linear system with normal errors; equations with discrete dependent variables are set up as ordered probit in this case. Thus, our approach provides a simple parametric baseline model that can be gradually extended to more flexible parametric models by increasing the number of mixture components; a truly nonparametric model can be obtained by setting a prior on the number of mixture components. Thus, robustness to nonlinearities, nonseparabilty, heteroskedasticity and other deviations from the baseline can be gradually introduced in a computationally tractable way. We illustrate the model performance in simulations and an application. All the usual practical advantages of the Bayesian approach apply. For example, posterior distributions provide more detailed and informative uncertainty descriptions than classical set estimators, model based predictions and decision making take account of parameter uncertainty, and available prior information can be easily incorporated into analysis.

The rest of the paper is organized as follows. Section 2 describes the data generating process (DGP) in nonseparable IV settings. Section 3.1 sets up a nonparametric conditional distribution model that is used in Section 3.2 as a building block for our IV model. Simulation exercises and an application to estimation of the return to education on data from Card (1995) are presented in Section 4. Section 5 concludes with directions for future work.

2. Data Generating Process. Suppose Y_1 is a scalar outcome variable, Y_2 is a scalar endogenous covariate, W is a vector of exogenous covariates, Z is a vector of instrumental variables, and U stands for continuously distributed scalar unobserved heterogeneity. Let $p_0(y_1, y_2, u|w, z)$ denote the conditional data generating density with respect to an appropriate product measure $\nu_1 \times \nu_2 \times \nu_u$. We do not model the marginal distribution of (W, Z). Let us assume that $Y_1 \perp Z | Y_2, W, U$ and $U \perp Z | W$ so that we have the following decomposition for the conditional density

$$p_0(y_1, y_2, u|w, z) = p_{01}(y_1|y_2, w, u) \cdot p_{02}(y_2|z, w, u) \cdot p_{0u}(u|w).$$

In what follows, we model $p_{01}(y_1|y_2,w,u)$ and $p_{02}(y_2|z,w,u)$ flexibly or nonparametrically and hence the density for the unobserved heterogeneity $p_{0u}(u|w)$ can be normalized without a loss of generality to an arbitrary density such as standard normal.

The objects of interest in applications could be the distribution of the outcome given the covariates

$$\int p_{01}(y_1|y_2,w,u)\cdot p_{0u}(u|w)du$$

that has a causal interpretation and various related functionals such as an average causal effect of changing Y_2 by a Δ

$$\mu(y_2,\Delta,w)=\int\int y_1[p_{01}(y_1|y_2+\Delta,w,u)-p_{01}(y_1|y_2,w,u)]d\nu_1(y_1)\cdot p_{0u}(u|w)du. \tag{2.1}$$

3. Bayesian Model. To model the conditional densities p_{01} and p_{02} flexibly we adapt a mixture model with covariate dependent mixing weights from Norets and Pelenis (2022a), who show that under anisotropic smoothness conditions on the data generating conditional distribution and a possibly increasing number of the support points for the discrete part of the distribution, the posterior in their model contracts at adaptive optimal rates up to a log factor. The model optimally takes advantage of smoothness in discrete variables if it is present in the DGP. The lower bounds on the estimation rates and posterior contraction rates for related joint distribution models were previously established in Norets and Pelenis (2022b).

In Section 3.1 below, we describe this general model for conditional distributions, its prior, and review appropriate MCMC algorithms for posterior simulation. Section 3.2 uses this model as a building block for our IV model.

3.1. *Nonparametric Conditional Density Model for Discrete-Continuous Variables.* Let us first consider a continuous dependent variable y. In a vector of independent variables, $x=(x_1,\dots,x_d)$, coordinates $x_k\in\mathbb{R}$, $k=1,\dots,d_c$ are continuous and coordinates $x_k\in\mathbb{X}_k\subset\mathbb{R}$, $k=d_c+1,\dots,d$ are discrete. Let us map each possible value of discrete $x_k\in\mathbb{X}_k$ into an interval $(a_{x_k},b_{x_k}]$ so that the intervals form a partition of $\mathbb{R}$

$$\bigcup_{x_k\in\mathbb{X}_k}(a_{x_k},b_{x_k}]=\mathbb{R} \text{ and } (a_{x_k},b_{x_k}]\cap(a_{\tilde{x}_k},b_{\tilde{x}_k}]=\emptyset \text{ for } \tilde{x}_k\neq x_k.$$

For a fixed positive integer m and $\theta=(\alpha_j,\beta_j,\mu_j,\sigma_j,\sigma_{j1},\mu_{j1},\dots,\sigma_{jd},\mu_{jd},\ j=1,2,\dots)$, the density of y conditional on x is modeled by

$$p(y|x,\theta,m)=\sum_{j=1}^{m}\gamma_j(x,\theta,m)\phi(y;\mu_j+x'\beta_j,\sigma_j), \tag{3.1}$$

where $\phi(y;\mu,\sigma)$ denotes a normal density with mean μ and standard deviation σ,

$$\gamma_j(x,\theta,m) \propto \alpha_j \exp\left\{-\frac{1}{2}\sum_{k=1}^{d_c}\left(\frac{x_k-\mu_{jk}}{\sigma_{jk}}\right)^2\right\}\cdot\prod_{k=d_c+1}^{d}\left[\Phi\left(\frac{b_{x_k}-\mu_{jk}}{\sigma_{jk}}\right)-\Phi\left(\frac{a_{x_k}-\mu_{jk}}{\sigma_{jk}}\right)\right], \tag{3.2}$$

and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal random variable.

When y is discrete, we map its values into intervals $(a_y, b_y]$ forming a partition of $\mathbb{R}$ similarly to the treatment of the discrete covariates above. The probability mass function of y conditional on x is modeled by

$$p(y|x,\theta,m) = \sum_{j=1}^{m}\gamma_j(x,\theta,m)\left[\Phi\left(\frac{b_y-\mu_j-x'\beta_j}{\sigma_j}\right)-\Phi\left(\frac{a_y-\mu_j-x'\beta_j}{\sigma_j}\right)\right]. \tag{3.3}$$

This model for conditional distributions can be motivated as follows. Suppose every discrete coordinate has an underlying continuous latent variable and the discrete values are mapped into intervals for the latent variable. Let the joint distribution of the continuous variables and the latent variables be a finite mixture of multivariate normal distributions with diagonal variances. The implied joint density-point mass function for continuous and discrete variables is a finite mixture of discrete-continuous distributions. Conditional distributions of interest can then be obtained from this joint distribution and they have the form of (3.1) for continuous y and (3.3) for discrete y (with β_j's set to zeros). If β_j's a set to zeros and scale parameters (σ_j, σ_{jk}) are forced to be the same across mixture components $j = 1,\dots,m$, the optimal bounds on the posterior contraction rates will still hold (a proof for the continuous case can be found in Norets and Pati (2014)), however, in finite samples, the model without these restrictions perform better.

A nonparametric Bayesian model for conditional distribution is completed with a prior $\Pi(\theta_{1m}|m)\Pi(m)$, where $\theta_{1m} = (\alpha_j, \beta_j, \mu_j, \sigma_j, \sigma_{j1}, \mu_{j1}, \dots, \sigma_{jd}, \mu_{jd},\ j = 1,\dots,m)$.

3.1.1. *Prior Distributions for Parameters of Conditional Density Model.* We specify the prior as follows,

$$\begin{aligned}
&\beta_j \overset{iid}{\sim} N(\underline{\beta}, \underline{H}_\beta^{-1}),\ \mu_j \overset{iid}{\sim} N(\underline{\mu}, \underline{h}_\mu^{-1}),\ \mu_{jk} \overset{iid}{\sim} N(\underline{\mu}_k, \underline{h}_{\mu_k}^{-1}),\\
&(\alpha_1,\dots,\alpha_m)|m \overset{iid}{\sim} D(\underline{a}/m,\dots,\underline{a}/m),\\
&\Pi\big(m=k\big) = (e^{\underline{A}_m}-1)e^{-\underline{A}_m\cdot k},
\end{aligned}$$

where D stands for a Dirichlet distribution. The log of prior density for $(\sigma_1,\dots,\sigma_m)$ conditional on m is equal up to an additive constant to

$$\underline{A}\sum_{j=1}^{m}\log\sigma_j - (\underline{A}_y + m\underline{A})\log\left(\underline{B}_y + \underline{B}\sum_{j=1}^{m}\sigma_j\right). \tag{3.4}$$

It corresponds to setting $\sigma_j = S \cdot S_j$, where $S_j \overset{iid}{\sim} G(\underline{A}, \underline{B})$ $j = 1, \ldots, m$, $S \sim InvG(\underline{A}_y, \underline{B}_y)$, $G(\underline{A}, \underline{B})$ stands for a Gamma distribution with shape $\underline{A}$ and rate $\underline{B}$ and $InvG$ stands for the corresponding inverse Gamma distribution. The prior distributions for $(\sigma_{1k}, \ldots, \sigma_{mk})$ for each k are also defined by (3.4) with hyperparameters $(\underline{A}_{x_k}, \underline{B}_{x_k})$ replacing $(\underline{A}_y, \underline{B}_y)$. Norets and Pelenis (2022b) and Norets and Pelenis (2022a) establish posterior contraction rates for the special case of $S_j = 1$ $j = 1, \ldots, m$. As was shown by Norets and Pati (2014) for the model with continuous variables only, the introduction of component specific scale parameters (variable S_j) improves model performance in finite samples and does not affect the established upper bounds on the posterior contraction rates.

Section 3.2.2 describes the selection of values of prior hyperparameters

$$(\underline{\beta}, \underline{H}_\beta, \underline{\mu}, \underline{h}_\mu, \underline{a}, \underline{A}_m, \underline{A}, \underline{B}, \underline{A}_y, \underline{B}_y, \underline{A}_{x_k}, \underline{B}_{x_k}, \underline{\mu}_k, \underline{h}_{\mu_k},\ k = 1, \ldots, d).$$

3.1.2. *Posterior Simulation in Conditional Density Model.* To simulate parameters of the conditional density model we use a version of the MCMC algorithm from Norets (2021). It combines the use of latent mixture allocation variables introduced by Diebolt and Robert (1994) to facilitate the simulation of mixture component parameters $(\mu_j, \sigma_j, \beta_j)$, a Metropolis-Hastings algorithm for simulating the mixing weights parameters $(\alpha_j, \mu_{jk}, \sigma_{jk}, k = 1, \ldots, d)$, and an approximately optimal reversible jump algorithm for simulating the number of mixture components m introduced by Norets (2021). We also experiment with Hamiltonian Monte Carlo and NUTS sampler (Hoffman and Gelman (2014)) as implemented in Stan[1] for posterior simulation with a fixed m.

3.2. *Nonseparable IV model and Posterior Distribution.* Suppose $(Y_{1i}, Y_{2i}, W_i, Z_i, U_i)$, $i = 1, \ldots, n$ is a random sample from the DGP outlined in Section 2. Let the DGP conditional densities $p_{or}(\cdot|\cdot)$ be modeled flexibly by $p(\cdot|\cdot, \theta^r, m_r)$, $r = 1, 2$, defined in (3.1) or (3.3) with the covariate coordinates resorted so that the continuous ones come first. Then, the posterior for the unobservables is given by

$$\begin{aligned} &\Pi\left(\theta^1_{1m_1}, m_1, \theta^2_{1m_2}, m_2, U_1, \ldots, U_n \middle| Y_{1i}, Y_{2i}, W_i, Z_i,\ i = 1, \ldots, n\right) \\ &\propto \prod_{i=1}^n p\left(Y_{1i} \middle| (Y_{2i}, W_i, U_i), \theta^1, m_1\right) \cdot p\left(Y_{2i} \middle| (Z_i, W_i, U_i), \theta^2, m_2\right) \cdot \phi(U_i; 0, 1) \end{aligned} \tag{3.5}$$

[1] https://mc-stan.org/

$$\cdot \Pi(\theta^1_{1m_1}|m_1)\Pi(m_1)\Pi(\theta^2_{1m_2}|m_2)\Pi(m_2),$$

where the priors are defined in Section 3.1.1.

3.2.1. *MCMC for Exploring the Posterior.* The MCMC algorithm is a Metropolis-within-Gibbs algorithm that combines blocks for simulating $\theta^r_{1m_r}|\ldots$, $m_r|\ldots$, $r = 1, 2$ as briefly reviewed in Section 3.1.2 and blocks for simulating U_i $i = 1, \ldots, n$. Specifically, to simulate from $U_i|\ldots$ we use a Metropolis-Hastings algorithm with a proposal distribution given by the standard normal. We also currently experiment with an implementation of the model in Stan language for a fixed m, where on each MCMC iteration all the unobservables are updated jointly by a Hamiltonian Monte Carlo algorithm.

3.2.2. *Setting Prior Hyperparameters.* The priors $\Pi(\theta^1_{1m_1}|m_1)$ and $\Pi(\theta^2_{1m_2}|m_2)$ in (3.5) are specified as described in Section 3.1.1. Here, we describe how the hyperparameters of these prior distributions are set based on appropriate data counterparts. Scricciolo (2015) shows that setting prior hyperparameters in such a fashion does not affect the established upper bounds on the posterior contraction rates in estimation of conditional densities.

To make the notation more explicit let

$$\theta^r_{1m_r} = (\alpha^r_j, \beta^r_j, \mu^r_j, \sigma^r_j, \sigma^r_{j1}, \mu^r_{j1}, \ldots, \sigma^r_{jd}, \mu^r_{jd},\ j = 1, \ldots, m_r)$$

and the corresponding prior hyperparameters be denoted by

$$(\underline{\beta}^r, \underline{H}^r_\beta, \underline{\mu}^r, \underline{h}^r_\mu, \underline{a}^r, \underline{A}^r_m, \underline{A}^r, \underline{B}^r, \underline{A}^r_y, \underline{B}^r_y, \underline{A}^r_{x_k}, \underline{B}^r_{x_k}, \underline{\mu}^r_k, \underline{h}^r_{\mu_k},\ k = 1, \ldots, d),\ r = 1, 2.$$

The hyperparameter $\underline{\mu}^1$ and the components of $\underline{\beta}^1$ associated with the coefficients on (Y_2, W) in β^1_j are set to the TSLS estimates of the intercept and slope coefficients, respectively. The hyperparameter $\underline{\mu}^2$ and the components of $\underline{\beta}^2$ associated with coefficients on (Z, W) in β^2_j are set to the first-stage OLS estimates of the intercept and slope coefficients, respectively. Similarly, $0.01(\underline{h}^r)^{-1}_\mu$ and $0.01(\underline{H}^r)^{-1}_\beta$ for the observable variables are set, respectively, to the estimated variance of the TSLS intercept and a diagonal matrix with the variances of the TSLS slope estimator for $r = 1$ and to the first-stage OLS variances of the intercept and slope coefficients for $r = 2$. The multiplicative factor 0.01 makes the priors less informative and it can be changed in prior sensitivity checks.

To justify the choice of hyperparameters for the component of $\underline{\beta}^r$ associated with U, consider the following linear structural equation system for the outcome and first stage:

$$Y_1 = \beta' X_1 + \epsilon_1, \quad Y_2 = \pi' X_2 + \epsilon_2,$$

where $X_1 = (Y_2, W')'$ and $X_2 = (Z', W')'$. The structural and first-stage errors are decomposed into the common component U and independent equation-specific components v_1, v_2:

$$\epsilon_1 = \delta_1 U + v_1, \quad \epsilon_2 = \delta_2 U + v_2.$$

After normalizing $U \sim N(0,1)$, we set

$$(\delta_1, \delta_2, \sigma_{v_1}, \sigma_{v_2}) = \left(\sqrt{|\rho|}\sigma_{\epsilon_1}, \frac{\rho}{\sqrt{|\rho|}}\sigma_{\epsilon_2}, (1-|\rho|)\sigma_{\epsilon_1}, (1-|\rho|)\sigma_{\epsilon_2} \right)$$

and

$$\underline{\beta}^1_u = \sqrt{|\hat{\rho}|}\hat{\sigma}_{\epsilon_1}, \quad \underline{\beta}^2_u = \frac{\hat{\rho}}{\sqrt{|\hat{\rho}|}}\hat{\sigma}_{\epsilon_2},$$

where $\underline{\beta}^r_u$ is the component of $\underline{\beta}^r$ corresponding to the coefficient on U, $\hat{\sigma}_{\epsilon_1}$ and $\hat{\sigma}_{\epsilon_2}$ are the standard deviations of the TSLS residuals and the first-stage residuals, respectively, and $\hat{\rho}$ is their sample correlation coefficient. The diagonal elements of $(\underline{H}^r_\beta)^{-1}$ corresponding to the coefficients on U are set to 1.

The parameters $\underline{A}^r$ and $\underline{B}^r$ are chosen to fix the mode of the component-specific gamma prior at unity. Our simulations suggest that values $\underline{A}^r = 11, \underline{B}^r = 10$ perform well. The hyperparameters $\underline{A}^r_y$ and $\underline{B}^r_y$ are chosen to fix the mode of the inverse gamma prior to the sample equivalents of σ_{v_1} and σ_{v_2}

$$\hat{\sigma}_{v_1} = (1-|\hat{\rho}|)\hat{\sigma}_{\epsilon_1}, \quad \hat{\sigma}_{v_2} = (1-|\hat{\rho}|)\hat{\sigma}_{\epsilon_2}.$$

Our simulations suggest a reasonable choice is

$$\underline{A}^1_y = \underline{A}^2_y = 10, \quad \underline{B}^1_y = 11\hat{\sigma}_{v_1}, \quad \underline{B}^2_y = 11\hat{\sigma}_{v_2}.$$

The hyperparameters $(\underline{\mu}^r_k, \underline{h}^r_k)$ for the priors on the location parameters μ^r_{jk} in the mixing weights are set to the sample means and precisions of the respective variables (and 0 and 1 for U). The hyperparameters for the scale parameters, σ^r_{jk} are chosen such that the mode of the common inverse gamma distribution matches the standard deviation of the variables (1 for U). Our simulations suggest that the following values work well

$$\underline{A}^r_{x_k} = 10, \quad \underline{B}^r_{x_k} = 11\hat{\sigma}_{x_k}.$$

The parameter of the Dirichlet distribution, $\underline{a}^r$ is set to 10. So far we only estimated the Bayesian model with a fixed number of components, up to $m_1 = m_2 = m = 10$ and $\underline{A}^r_m$ is not used.

4. Applications.

4.1. *Card (1995).* Card (1995) used a linear IV model to estimate the return to education. The sample contains $n = 3010$ observations. In this application, the outcome variable is the logarithm of the wage, the endogenous covariate is the years of education, the instrument is an indicator variable for whether the individual grew up near a four year college, and the exogenous covariates or controls include the years of experience, a race indicator, an indicator for living in a standard metropolitan statistical area, and a South indicator variable.

We terat all the variables in this application except for the log of wage are discrete. Below, we present estimation results from our model with the number of mixture components fixed at $m_1 = m_2 = m \in \{1, 5, 10\}$. For comparison, the standard TSLS estimator of the coefficient on education in the linear IV system is equal to .13 with the standard error of .05.

Figure 1 shows the marginal prior and posterior of the ACE $\mu(y_2 = 12, \Delta = 1, w = (10, 0, 0, 0))$ defined in (2.1) of changing the level of education from 12 to 13 for an individual with 10 years of experience and the rest of control dummies equal to zeros.

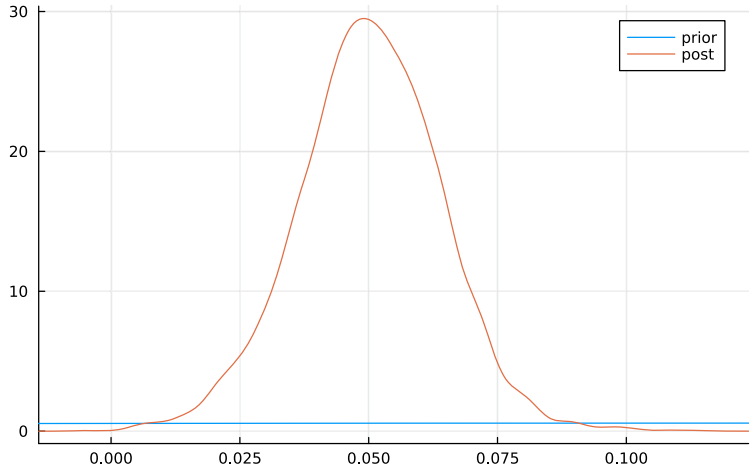

FIG 1. *Prior and posterior for* $\mu(y_2 = 12, \Delta = 1, w = (10, 0, 0, 0))$*, model with* $m = 5$.

As can be seen from the figure, the prior is completely flat relative to the posterior distribution, which confirms that the selection of the prior hyperparameters based on data as described in Section 3.2.2 does not lead to very informative priors.

Figures 2, 3 and 4 below show the posterior mean and quantiles for the ACE at different levels of education y_2, $w = (10, 0, 0, 0)$, and $m \in \{1, 5, 10\}$.

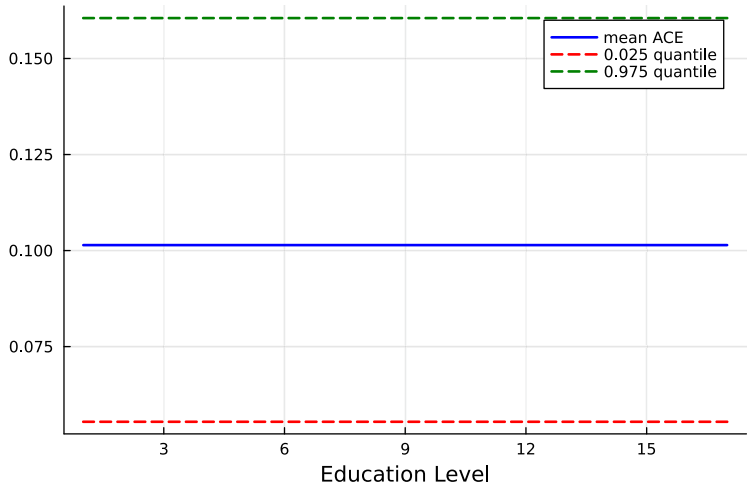

FIG 2. *Posterior mean and quantiles for* $\mu(y_2, \Delta = 1, w = (10, 0, 0, 0))$, $y_2 \in \{1, \ldots, 16\}$, *model with* $m = 1$.

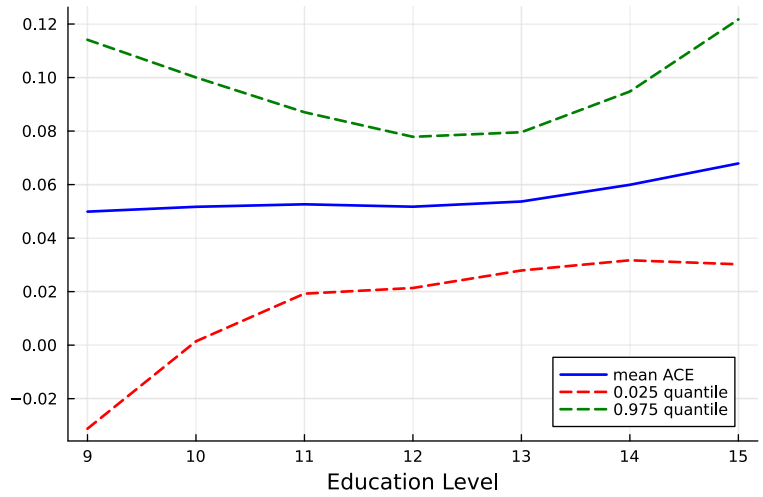

FIG 3. *Posterior mean and quantiles for* $\mu(y_2, \Delta = 1, w = (10, 0, 0, 0, 0))$, $y_2 \in \{9, \ldots, 15\}$, *model with* $m = 5$.

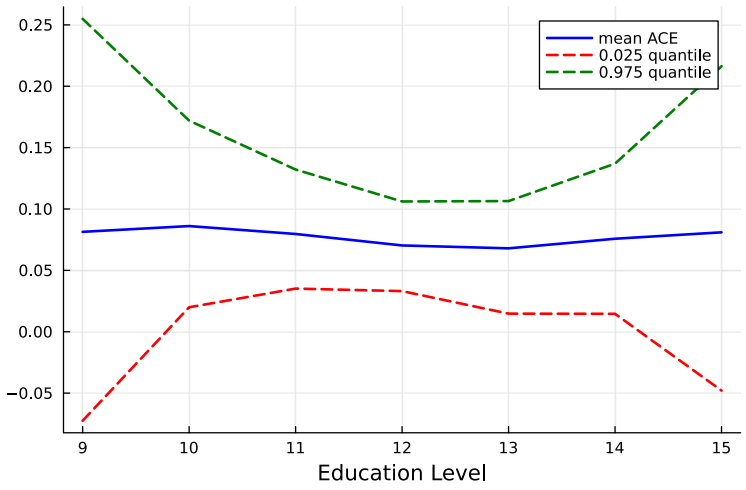

FIG 4. *Posterior mean and quantiles for* $\mu(y_2, \Delta = 1, w = (10, 0, 0, 0, 0))$, $y_2 \in \{9, \ldots, 15\}$, *model with* $m = 10$.

Even for $m = 1$, the Bayesian model estimation results do not perfectly align with the TSLS 95% confidence set $(0.12, 0.14)$; the difference seems to stem from different treatment of the discrete variables. For more flexible models with $m = 5$ and $m = 10$, the estimated causal effects are considerably smaller. An advantage of our approach is that we can readily assess how the effects vary with the level of the endogenous and exogenous covariates. While in this particular application the effects do not seem to vary a lot with the level of education and experience (Figure 5), the uncertainty about them varies drastically as evident from the depicted posterior quantiles, especially for the more flexible model with $m = 10$.

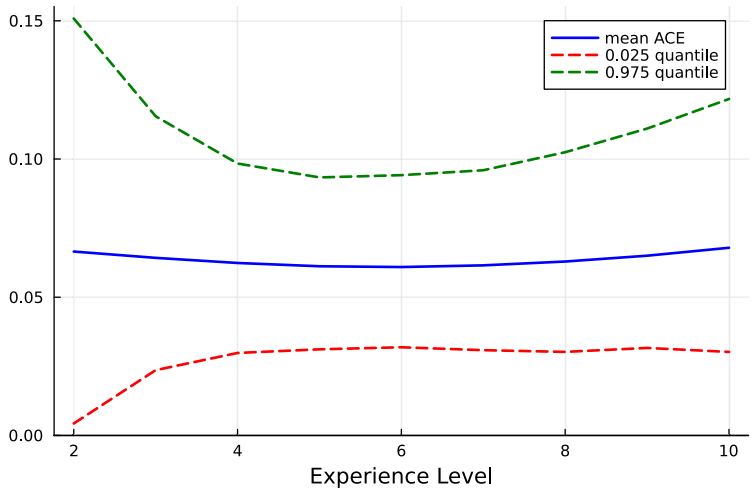

FIG 5. *Posterior mean and quantiles for the average effect of increasing education from 15 to 16 at different levels of experience, control dummies set to zeros, model with* $m = 5$.

4.1.1. *Evidence of MCMC Convergence.* Due to the possible lack of label switching for the mixture components, MCMC convergence in mixture models should be assessed using label invariant functions of parameters, see, for example, Geweke (2007). Average causal effects, unobserved heterogeneity U_i, the log of the likelihood times prior are label invariant and Figures 6-10 in this subsection depict their MCMC draws, so that MCMC convergence can be assessed. The number of MCMC iterations varies from 20 to 400 thousand. Figures below show MCMC draws from every 10th iteration. Overall, MCMC samplers appear to converge.

Figures 6 and 8 below show MCMC draws of $\mu\left(y_2 = 12, \Delta = 1, w = (10, 0, 0, 0)\right)$.

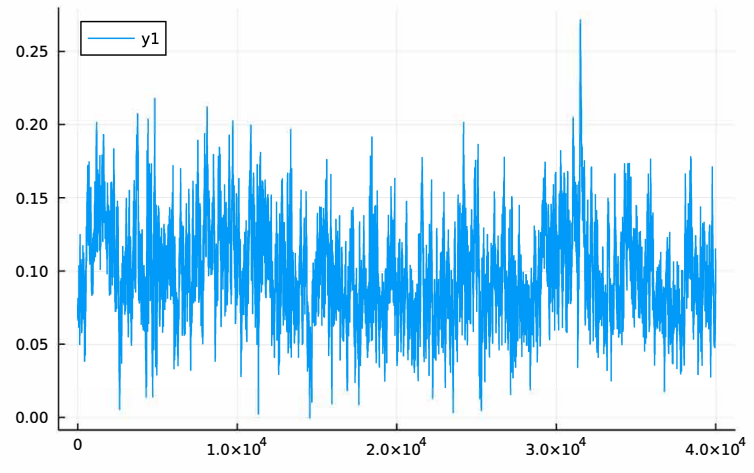

FIG 6. *MCMC draws of* $\mu\left(y_2 = 12, \Delta = 1, w = (10, 0, 0, 0)\right)$ *for model with* $m = 1$.

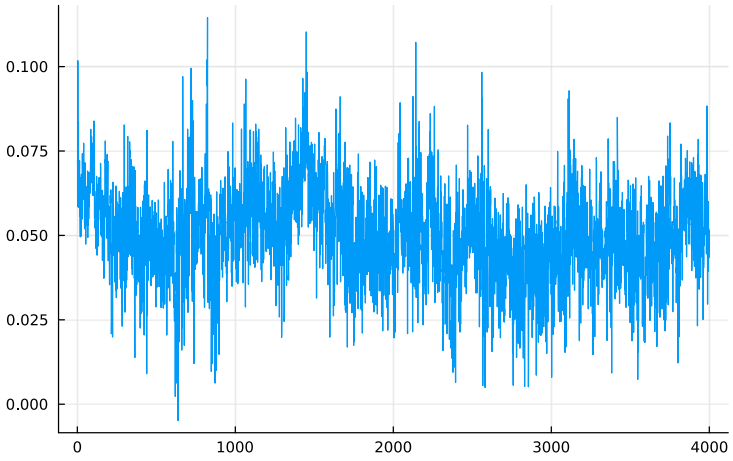

FIG 7. *MCMC draws of* $\mu\left(y_2 = 12, \Delta = 1, w = (10, 0, 0, 0)\right)$ *for model with* $m = 5$.

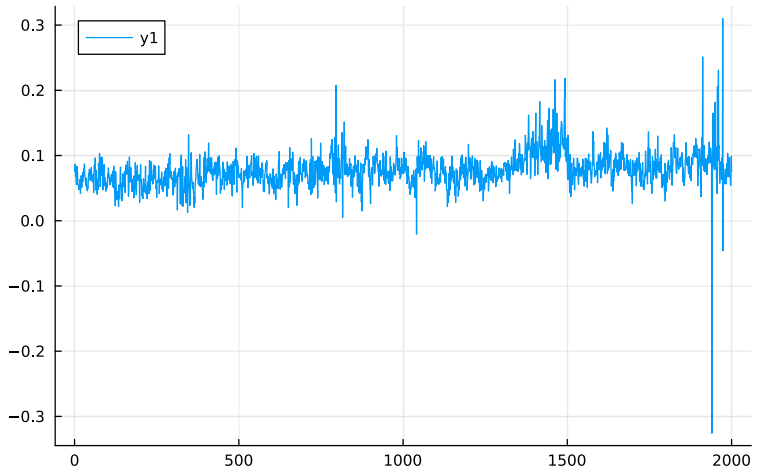

FIG 8. *MCMC draws of* $\mu\,(y_2 = 12, \Delta = 1, w = (10, 0, 0, 0))$ *for model with* $m = 10$.

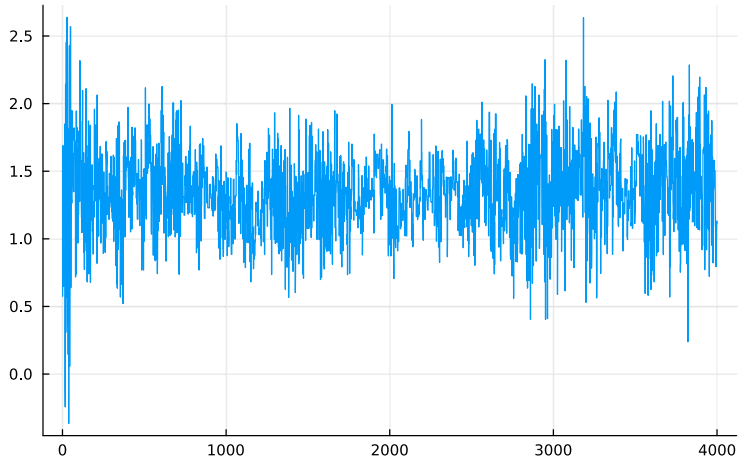

FIG 9. *MCMC draws of* U_i, $i = 1$ *for model with* $m = 5$.

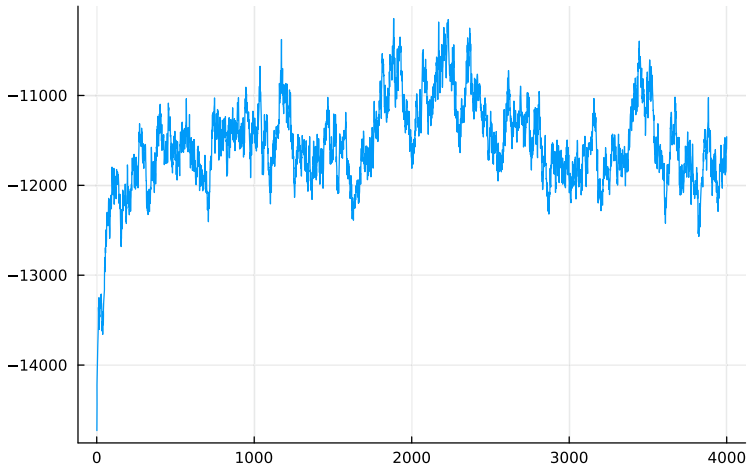

FIG 10. *MCMC draws of log of likelihood times prior for model with* $m = 5$.

4.2. *Experiment with Artificial Data* . To start exploring how our model can handle nonlinearities let us consider the following nonseparable and highly nonlinear DGP

$$Y_1 = 0.5e^{0.5(Y_2-5)}/(e^{0.5}-1) + 1 + U + \beta_{01}'W + e_1 \tag{4.1}$$
$$Y_2 = 0.5e^{-(0.1Z\cdot U-3)^2} + 1 + 2U + \beta_{02}'(W,Z) + e_2,$$

where $e_r \sim N(0,1)$, $Z = (Z_1, Z_2)$ and $W = (W_1, W_2)$, Z_1 and W_1 have $N(0,1)$ distribution, Z_2 and W_2 have a uniform distribution on $\{1,2,3,4,5,6\}$, and components of β_{0r} are set to draws from a uniform on $[-1,1]$. The simulated outcomes and endogenous covariates are depicted in Figure 11. The sample size is $n = 1000$.

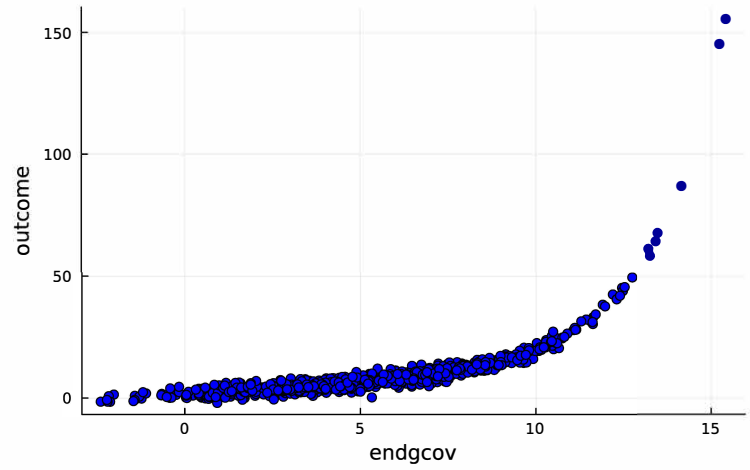

FIG 11. *Scatter plot of simulated outcomes and endogenous covariates.*

Figure 12 shows the DGP values and the posterior mean and quantiles for $\mu(y_2, \Delta = 1, w = (0,3))$

at different levels of y_2, along with the TSLS coefficient on y_2.

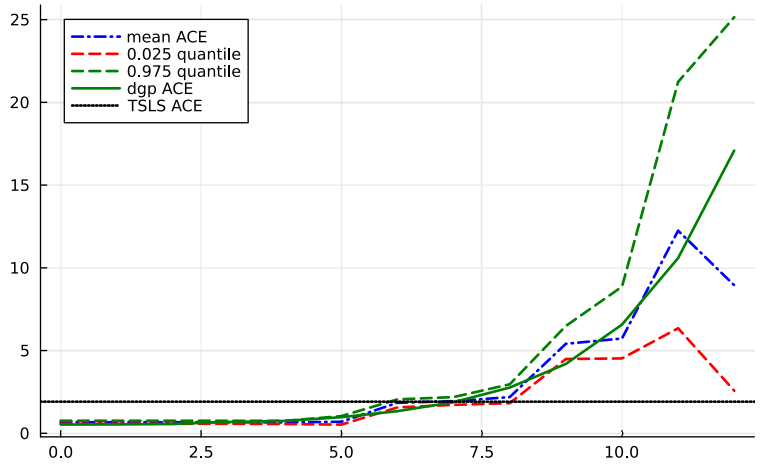

FIG 12. *Posterior mean and quantiles for* $\mu(y_2, \Delta = 1, w = (0,3))$, $y_2 \in [0,12]$, $m = 5$.

As can be seen from the figure, in this simulation exercise, the posterior mean of the ACE captures well the nonlinear behavior of the DGP, at least in the range where most observations of Y_2 are located. The spread of the posterior for $\mu(y_2, \Delta = 1, w = (0,3))$ increases with y_2, which is expected as there are relatively few observations with large values of y_2 in the sample. Figure 13 suggests that the posterior simulator converged.

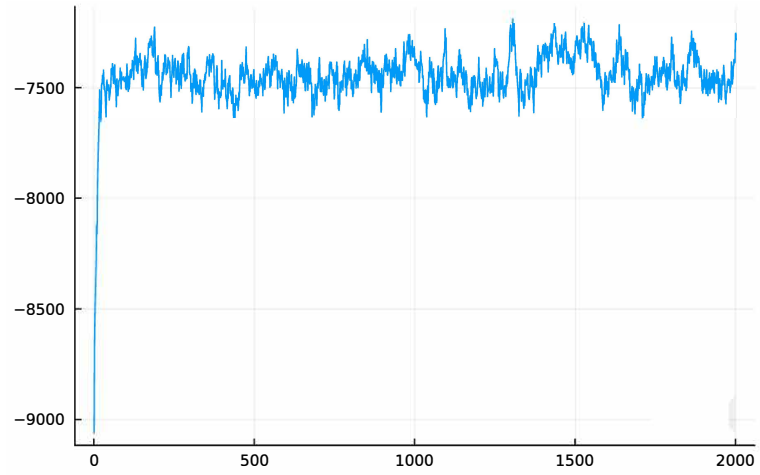

FIG 13. *MCMC draws of log of likelihood times prior for model with* $m = 5$.

5. Conclusions and Future Work. The proposed nonparametric Bayesian model for a nonseparable IV regression is being further assessed in a more extensive set of simulation exercises and applications. The asymptotic properties of the posterior distribution for this model

are of interest. We conjecture that the posterior contraction rates for estimation of the distribution of outcome and endogenous covariate conditional on the instruments and exogenous covariates should be possible to establish relying on the relevant results in Norets and Pelenis (2022a). A characterization of identified sets for objects with causal interpretation and posterior concentration on these identified sets is also on our research agenda.

References.

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, ed. by L. Christophides, E. Grant, and R. Swidinsky, Toronto: University of Toronto Press, 201–222.

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441.

——— (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73, 1525–1550.

Conley, T. G., C. B. Hansen, R. E. McCulloch, and P. E. Rossi (2008): "A semi-parametric Bayesian approach to the instrumental variable problem," *Journal of Econometrics*, 144, 276–305.

Diebolt, J. and C. Robert (1994): "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 363–375.

Geweke, J. (2007): "Interpretation and inference in mixture models: Simple MCMC works," *Computational Statistics and Data Analysis*, 51, 3529 – 3550.

Hoffman, M. D. and A. Gelman (2014): "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15, 1593–1623.

Imbens, G. W. and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481–1512.

Kitagawa, T. (2021): "The identification region of the potential outcome distributions under instrument independence," *Journal of Econometrics*, 225, 231–253.

Matzkin, R. L. (2008): "Identification in Nonparametric Simultaneous Equations Models," *Econometrica*, 76, pp. 945–978.

McCulloch, R. E., R. A. Sparapani, B. R. Logan, and P. W. Laud (2021): "Causal Inference with the Instrumental Variable Approach and Bayesian Nonparametric Machine Learning," .

Norets, A. (2021): "Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models," *Econometric Theory*, 37, 49–81.

Norets, A. and D. Pati (2014): "Adaptive Bayesian Estimation of Conditional Densities," Forthcoming in Econometric Theory.

Norets, A. and J. Pelenis (2022a): "Adaptive Bayesian estimation of conditional discrete-continuous distributions with an application to stock market trading activity," *Journal of Econometrics*, 230, 62–82.

——— (2022b): "Adaptive Bayesian Estimation of Discrete-Continuous Distributions Under Smoothness and Sparsity," *Econometrica*, 90, 1355–1377.

Scricciolo, C. (2015): "Empirical Bayes Conditional Density Estimation," *Statistica*, 75, 37–55.

Economics Department,
Brown University, Providence, RI 02912
E-mail: andriy_norets@brown.edu