Semiparametric Bayes Instrumental Variable Estimation with Many Weak Instruments

Ryo KATO
Takahiro HOSHINO

May 15, 2018

Semiparametric Bayes Instrumental Variable Estimation with Many Weak Instruments

Ryo Kato[*]

Research Institute for Economics and Business Administration,
Kobe University, 2-1 Rokkodai-cho, Nada-ku, Kobe, Japan

and

Takahiro. Hoshino[†]

Department of Economics, Keio University, 2-15-45 Mita, Minato-ku,
Tokyo, Japan /
RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi,
Chuo-ku, Tokyo, Japan

SUMMARY.We develop a new semiparametric Bayes instrumental variables estimation method. We employ the form of the regression function of the reduced-form equation and the disturbances are modelled nonparametrically to achieve better preditive power of the endogenous variables, whereas we use parametric formulation in the structural equation, which is of interest in inference. Our simulation studies show that under small sample size the proposed method obtains more efficient estimates and very precise credible intervals compared with existing IV methods. The existing methods fail to reject the null hypothesis with higher probability, due to larger variance of the estimators. Moreover, the mean squared error in the proposed method may be less than 1/30 of that in the existing procedures even in the presence of weak instruments. We applied our proposed method to a Mendelian randomization dataset where a large number of instruments are available and semiparametric specification is appropriate. This is a weak instrument case; hence, the non-Bayesian IV approach yields inefficient estimates. We obtained statistically significant results that cannot be obtained by the existing methods, including standard Bayesian IV.

KEY WORDS: Instrumental variable; Mendelian Randomization; Semiparametric Bayes model; Probit stick-breaking process mixture

[*]*email:* kato.ryo@keio.jp
[†]*email:* bayesian@jasmine.ocn.ne.jp

1. Introduction

This article presents a new semiparametric Bayes model for instumental variables problems. We treat the reduced-form equation (or the "first-stage" regression model) and the joint distribution of the error terms as nonparametric and potentially changing in form, corresponding to the values of the instrumental variables. In addition, our emphasis is on the semiparametric model formulation. Structural equation models (or the "second-stage" regression models), which are of interest in inference, are formulated parametrically, whereas the reduced-form equation and the disturbance terms are formulated nonparametrically.

Instrumental variables (IV) methods have become increasingly important in many empirical fields. Initially, IV estimation and its application were restricted to empirical economics. However, IV methods have been recently applied in other fields, such as epidemiology for causal inference (Ramsahai and Lauritzen, 2011, Baiocchi et al., 2014, Wang et al., 2017). In particular, IV approaches are employed when it is not feasible to carry out randomized controlled trials (RCT) or standard causal inference methodology, which assumes that no unobserved confounding exists. If it is possible to find sufficient IV that are predictive of endogenous variables, have no direct effect on the outcome, and are independent of the unobserved confounders, then the effect of the unobserved confounders can be controlled.

However, inference based on IV is prone to be imprecise when the instruments explain only a small portion of the variation in the endogenous variable (weak instrument case). This problem is inherent in samples of small size, as the data is insufficient to identify the parameters of interest (Conley et al., 2008). In general, frequentist methods depend on asymptotics and this property can be a hindrance when the sample size is small. Therefore they are not occasionally suitable for IV problems. By contrast, since Bayesian methods do not rely on asymptotics, applying these methods to IV problems is a reasonable choice. Even though there are no direct incentives for adopting Bayesian methods, Conley et al. (2008) showed that they incur smaller mean squared error (MSE) and provide better interval estimation compared with non-Bayesian methods when the structural equation and the reduced-from equation are properly specified. Moreover, Bayesian methods allow for more flexible modeling of the structural equation, since Bayesian inference depends only on the joint model of the structural and the reduced-form equation, whereas classical methods require the development of different estimation procedure according to whether we have discrete, clustered, or panel data.

One of the disadvantages of Bayesian methods is that they impose strong

distributional assumptions on the parameters. This is the case in IV problems, since Bayesian IV generally assumes that the joint distribution of disturbances is bivariate normal. Conley et al. (2008) proposed another Bayesian IV method that uses a Dirichlet process mixture (DPM) model for the error terms. It moderates the assumption on the disturbances by using DPM nonparametric specification.

Our proposed procedure is a semiparametric Bayesian IV method and is more flexible than Conley's method. We also assume that disturbances have nonparametric structure. Moreover, we use nonparametric formulation in the reduced-from equation. In general, the true functional form of the reduced-from equation is unknown, and its parameters are not of interest. In addition, if we use many instruments, parametric modeling of a large number of variables may result in misspecification bias (Chib, 2007). The approaches assuming that the reduced-form equation has some specific functional form (linear regression is often assumed), including frequentist and Conley's IV methods, yield unbiased estimates of the structural equation. However, they are less efficient. By contrast, if, for example, the reduced-from equation is not a simple linear combination with additive disturbances, our semiparametric Bayes model fits the data better and yields efficiency gains compared with classical parametric method and Conley's method.

Since the parameters of the structural equation are important in applied research, we assume that the structural equation regression model has parametric structure. Moreover, our model is different from other frequentist nonparametric IV approaches in that these approaches use nonparametric specification in the structural equation and parametric specification in the reduced-form equation.

We employ a probit stick-breaking process mixture (PSBPM) model proposed by Chung and Dunson (2009) to realize more flexible semiparametric representations for IV. Nonparametrics based on Dirichlet process makes it possible to represent a distribution by infinite mixture of well-known "base" distributions. Whereas the mean regression structure of the DPM is reduced to a linear regression model, PSBPM is more flexible than DPM since it enables us to make a probability weight of the components change by predictors in the regression model. Hence, we can treat reduced-form equation and the joint distribution of error terms as potentially changing in shape as the value of instruments vary. Even in the case that the reduced-form equation and the error terms are truly linear and bivariate normal, respectively, our procedure has small efficiency loss, since it is not necessary to prespecify the number of components. The optimal number of components, which is needed for a

3

finite mixture of regression models, is defined by the data.

We conduct a Monte Carlo simulation study in order to evaluate the performance of the proposed method. We investigate the finite sample performance of the estimators when the reduced-from model is not a simple linear combination. The proposed method may incur as little as 1/30 of the MSE incurred by existing procedures. Moreover, the coverage of nominal 95% confidence (or credible) intervals of the proposed method is very close to 0.95, whereas the other methods provide significantly narrower or wider interval estimates.

The proposed method is applied to a real Mendelian randomization dataset. In general, the number of instrumental variables in Mendelian randomization (*i.e.* information of the genotype) is large, and correct specification of the reduced-form regression model is difficult. In addition, the instruments may not explain the endogenous variables satisfactorily, and non-Bayesian methods that rely on the asymptotic approximation yield biased results. Therefore, the proposed Bayesian nonparametric formulation for the reduced-form equation is appropriate and results in obtaining efficient endogenous (causal) parameters. In fact, we provide statistically significant results that are not obtained by the standard Bayesian IV approach.

The remainder of the paper is organized as follows. In the next section, we propose and formulate a new semiparametric Bayes IV (SB-IV) method that can overcome the drawbacks of the existing IV methods. MCMC algorithm for our procedure is also presented in Section 2. The simulation studies illustrating the performance of the proposed method compared with the existing nonBayesian and Bayesian IV approaches are presented in Section 3. In Section 4, we apply our proposed method to the real dataset for the Mendelian randomization. Section 5 concludes after providing a short discussion.

2. Semiparametric Bayes estimation for instrumental variable

2.1 *The linear model and the existing approaches*

We consider a dataset consisting of N $(i = 1, ..., N)$ cases. We further consider a case with one linear structural equation and one corresponding reduced-form equation as follows:

$$\begin{cases} y_i = x_i\beta + \boldsymbol{w}_i\boldsymbol{\gamma} + \epsilon_{1,i} \\ \quad x_i = \boldsymbol{z}_i\boldsymbol{\delta} + \epsilon_{2,i} \end{cases} \tag{1}$$

where y_i is the outcome variable, x_i is the endogenous regressor $(E(\epsilon_{1,i}x_i) \neq 0)$, $\boldsymbol{w}_i$ is the k-dimensional vector of exogenous regressors, and $\boldsymbol{z}_i$ is the

4

m-dimensional vector of instruments ($E(\epsilon_{1,i}\boldsymbol{z}_i) = 0$) that includes $\boldsymbol{w}_i$. We consider the situation that $\epsilon_{1,i}$ and $\epsilon_{2,i}$ are correlated, which causes the endogeneity problem. We call the first line in (1) structural equation and the second line reduced-form equation.

There is rich literature in econometrics that develops non-Bayesian and Bayesian methods for IV model estimation. Two stage least squares (TSLS) (Theil, 1953, Basmann, 1957) and limited information maximum likelihood (LIML) (Anderson and Rubin, 1949, Hood and Koopmans, 1953) are the most frequently used non-Bayesian approaches to IV problems. These methods do not specify the distribution of $\boldsymbol{\epsilon} \equiv (\epsilon_1, \epsilon_2)^t$ and are robust with respect to the error distribution. However, these frequentist methods rely on asymptotics. This may be a problem when the sample size is not large enough. Non-Bayesian methods for IV problems include nonparametric IV methods. Newey and Powell (2003) proposed IV estimation of nonparametric models that considers the structural equation to be the conditional expectation of the endogenous variables. Later, Dreze (1976), and Kleibergen (1997) developed IV estimation methods from the Bayesian perspective. One of the differences between non-Bayesian and Bayesian methods for IV problems is that non-Bayesian approaches do not rely on the assumption of some specific joint distribution regarding the error terms. By contrast, Bayesian approaches generally assume bivariate normal distribution (Chao and Phillips, 2011, Geweke, 1996, Kleibergen and Zivot, 2003, Rossi et al., 2005, Hoogerheide et al., 2007).

However, explicit distributional assumptions can cause misspecification due to heterogeneity or outliers of the data. To overcome this, Conley et al. (2008) considered a more flexible semiparametric Bayesian model for instrumental variables problems. They employed a Dirichlet process prior for the error terms in the structural and reduced-form equations. More specifically, they considered $\boldsymbol{\epsilon}_i \equiv (\epsilon_{1,i}, \epsilon_{2,i})^t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ instead of $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_i$ are i.i.d. following a discrete random distribution G, and G is modeled with concentration parameter α and base distritbution G_0, that is, $G \sim DP(\alpha, G_0)$; namely, an infinite mixture of normal distributions where the number of components and the shape parameters are determined by the specified prior and the observed data. Their Monte Carlo simulation study showed that when the error terms do not follow the normal distribution, their method is more efficient than existing Bayesian and non-Bayesian approaches.

Even though the Bayesian method in Conley et al. (2008) moderates the Bayesian strong assumption on the distribution of the error terms, existing

Bayesian and non-Bayesian methods, as well as Conley's method, specify some functional form, usually a linear combination, for the reduced-form equation. In fact, the true relationship between endogenous variables and instrumental variables is unknown. Therefore, existing methods impose very strong assumptions on IV regression. In addition, we are not generally interested in the parameters of reduced-form models, since these models are just "first-step" regression models.

We develop a new more flexible semiparametric Bayes IV model where we use parametric formulation in the structural equation, whereas we use nonparametric formulation in the reduced-form equation and the joint distribution of the error terms. We note that we do not use nonparametric representation in the structural equation, since its parameters, such as the gradient coefficient of x_i or $\boldsymbol{w}_i$, are usually of interest.

2.2 *Semiparametric model formulation*

We express the full model as the product of three submodels.

$$p(y_i, x_i, \boldsymbol{\Sigma} | \boldsymbol{w}_i, \boldsymbol{z}_i) \propto p(y_i | x_i, \boldsymbol{w}_i, \boldsymbol{\vartheta}_s, \boldsymbol{\Sigma}) p(x_i | \boldsymbol{z}_i, \boldsymbol{\vartheta}_r, \boldsymbol{\Sigma}) p(\boldsymbol{\Sigma})$$

where $p(y_i | x_i, \boldsymbol{w}_i, \boldsymbol{\vartheta}_s, \boldsymbol{\Sigma})$ represents the structural equation, $p(x_i | \boldsymbol{z}_i, \boldsymbol{\vartheta}_r, \boldsymbol{\Sigma})$ represents the reduced form equation, and $p(\boldsymbol{\Sigma})$ represents the joint (prior) distribution of error terms covariance matrix. $\boldsymbol{\vartheta}_s$ and $\boldsymbol{\vartheta}_r$ denote the parameters of the structural equation and the reduced form equation except for $\boldsymbol{\Sigma}$, respectively. We assume the parametric model for the structural equation $p(y_i | x_i, \boldsymbol{w}_i, \boldsymbol{\vartheta}_s, \boldsymbol{\Sigma})$ since the researchers' concern generally lies in $\boldsymbol{\vartheta}_s$. On the other hand, we consider nonparametric structure for the reduced form equation $p(x_i | \boldsymbol{z}_i, \boldsymbol{\vartheta}_r, \boldsymbol{\Sigma})$ and $p(\boldsymbol{\Sigma})$ because they are not generally interested in $\boldsymbol{\vartheta}_r$ or $\boldsymbol{\Sigma}$. Specifically,

$$\left\{ \begin{array}{l} y_i = x_i \beta + \boldsymbol{w}_i \boldsymbol{\gamma} + \epsilon_{1,i} \\ x_i = g\left(\boldsymbol{z}_i; \boldsymbol{\delta}\right) + \epsilon_{2,i} \end{array} \right. , \boldsymbol{\epsilon}_i \equiv \left(\begin{array}{c} \epsilon_{1,i} \\ \epsilon_{2,i} \end{array} \right) \sim f \qquad (2)$$

We use DPM representation for the reduced-form equation and the joint error distribution. According to Sethuraman (1994), when the Dirichlet process prior is assumed for a random distribution G, denoted by $G \sim DP(\eta, G_0)$, G is expressed as follows:

$$G = \sum_{l=1}^{\infty} \kappa_l \delta_{\theta_l}, \ \theta_l \sim G_0$$

where η and G_0 are the concentration parameter and the base distribution, respectively. In this notation, δ_θ is a point mass at θ and $\kappa_l = \xi_l \prod_{h<l} (1 - \xi_h)$,

6

where $\xi_1, \xi_2, ...$ are independent draw from a beta distribution $Be(1, \eta)$ (refer to Walker et al. (1999) for a detailed description of DPM). These representations indicate that any distribution can be expressed as a mixture of well-known distributions, such as the Gaussian distribution.

Fortunately, the DPM model can be estimated with a relatively simpler MCMC algorithm by applying blocked Gibbs sampling (Ishwaran and James, 2011). Therefore, DPM modeling has been employed to avoid distributional assumptions on the parameters within the Bayesian framework. In addition to Conley et al. (2008), Hirano (2002) developed dynamic panel data models where the error terms are represented by a DP prior. DPM models not only treat disturbances of nonparametric classes but also afford a good deal of flexibility for a wide range of statistical models. Miyazaki and Hoshino (2009) proposed a Bayesian semiparametric item response model with DP prior. Kunihama and Dunson (2016) constructed a method for variable selection within Bayesian nonparametric DPM. The theoretical properties of DPM were investigated in Shen et al. (2013).

Even though DPM is often employed to avoid assumptions on parameter distribution, it is restricted due to the fact that the probability weight κ_l is constant across observations (Dunson et al., 2007). If the stick-breaking weights κ_l are constant and independent of the predictor $\boldsymbol{\chi}_i$, the mean regression structure is reduced to be linear, namely, $\sum_{l=1}^{\infty} \kappa_l \boldsymbol{\beta}_l^T \boldsymbol{\chi}_i = \overline{\boldsymbol{\beta}}^T \boldsymbol{\chi}_i$, where $\overline{\boldsymbol{\beta}} = \left(\sum_{l=1}^{\infty} \kappa_l \boldsymbol{\beta}_l^T \right)$. Therefore, instead of DPM, we apply the PSBPM model proposed by Chung and Dunson (2009) since it can change probability weights depending on the value of predictors $\boldsymbol{\chi}$ in the regression model. In addition, since PSBPM uses Gaussian distribution as a base distribution, the models affords a conjugate structure and simpler posterior calculation. PSBPM uses the probability weight $\pi_l(\boldsymbol{\chi})$ instead of κ_l,

$$\pi_l(\boldsymbol{\chi}) = \Phi(\boldsymbol{\xi}_l(\boldsymbol{\chi})) \prod_{h<l} \{1 - \Phi(\boldsymbol{\xi}_h(\boldsymbol{\chi}))\}. \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian distribution. PSBPM allows the probability weight $\pi_l(\boldsymbol{\chi})$ to vary as the predictors $\boldsymbol{\chi}$ vary.

For example, Hoshino (2013) proposed a semiparametric Bayesian model for causal inference where the irrelevant parameters are modeled using PSBPM. The statistical properties of PSBPM are described in Pati et al. (2013).

We apply PSBPM modeling to the reduced-form equation and the joint error distribution. We rewrite the models as an infinite mixture model of simultaneous equations.

$$f\left(\widetilde{\boldsymbol{y}}_i|\widetilde{\boldsymbol{X}}_i\right) = \sum_{l=1}^{\infty} \pi_l(\boldsymbol{z}) N\left(\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{\beta}}_l, \boldsymbol{\Sigma}_l\right)$$

where

$$\widetilde{\boldsymbol{y}}_i = \begin{pmatrix} y_i \\ x_i \end{pmatrix}, \widetilde{\boldsymbol{X}}_i = \begin{pmatrix} x_i & \boldsymbol{w}_i & \boldsymbol{0} \\ 0 & \boldsymbol{0} & \boldsymbol{z}_i \end{pmatrix}, \widetilde{\boldsymbol{\beta}}_l = \begin{pmatrix} \beta \\ \boldsymbol{\gamma} \\ \boldsymbol{\delta}_l \end{pmatrix},$$

$$\widetilde{\boldsymbol{\epsilon}}_l = \begin{pmatrix} \epsilon_{1,l} \\ \epsilon_{2,l} \end{pmatrix}, \widetilde{\boldsymbol{\epsilon}}_l \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_l).$$

We note that since we assume parametric specification for the structural equation, β and $\boldsymbol{\gamma}$ are common parameters across classes, whereas $\boldsymbol{\delta}_l$ varies with l. Since we specified PSBPM formulation for the reduced-form equation, we let the probability weight be dependent on the instruments $\boldsymbol{z}$, $\pi_l(\boldsymbol{z})$. As in equation (3), $\pi_l(\boldsymbol{z})$ is expressed by the following "stick-breaking" process.

$$\pi_l(\boldsymbol{z}) = \Phi(\boldsymbol{\xi}_l(\boldsymbol{z})) \prod_{h<l} \{1 - \Phi(\boldsymbol{\xi}_h(\boldsymbol{z}))\}.$$

We employ the same structure of instruments depending weight $\pi_l(\boldsymbol{z})$ as in Chung and Dunson (2009). That is, we let $\boldsymbol{\xi}_l(\boldsymbol{z}) = \alpha_l + f_l(\boldsymbol{z})$, $\alpha_l \sim N(\mu_\alpha, 1)$, $\mu_\alpha \sim N(\mu_{\alpha_0}, \sigma_{\alpha_0}^2)$ and we introduce the following regression function:

$$f_l(\boldsymbol{z}) = -\sum_{q=1}^{m} \psi_{lq} |z_q - \Omega_{lq}|$$

with $\psi_{lq} \sim N(\mu_{\psi_q}, \sigma_{\psi_q}^2)\mathbf{1}_{\mathbb{R}^+}, \Omega_{lq} \sim \sum_{q=1}^{B_q} \frac{1}{B_q}\delta_{\Omega_{qb}^*}(\Omega_{lq})$, where $N(\mu_{\boldsymbol{U}}, \sigma_{\boldsymbol{U}}^2)\mathbf{1}_{\boldsymbol{U}}$ denotes a normal distribution with mean $\mu_{\boldsymbol{U}}$ and variance $\sigma_{\boldsymbol{U}}^2$ truncated to the set $\boldsymbol{U}$, and Ω_{qb}^* are discrete points over a reasonable range of the q-th covariate z_q.

Our semiparametric Bayes IV model in equation (2) is different from the recent nonparametric IV approaches that consider the nonparametric structural equation. Such examples are Newey and Powell (2003), Hall and Horowitz (2005) and Darolles et al. (2011), which developed nonparametric IV regression models from non-Bayesian perspective, and Liao and Jiang (2011), Florens and Simoni (2012), and Kato (2013), which employed quasi-Bayesian approach to nonparametric Bayesian IV regression problem. On the other hand, our semiparametric approach employ nonparametric reduced-form model and disturbances, which are difficult to correctly prespecify, and

parametric structural model, which is of main interest. Therefore, our semi-parametric method is practical when we can use many instruments and especially interested in the parameters of the structural equation.

Some of non-Bayesian and quasi-Bayesian IV methods employ GMM (generalized method of moments) based approach, and it is well-known that GMM attains semiparametric efficiency bounds. However, this efficiency is obtained when the sample size is large enough. On the other hand, the proposed method is based on Bayesian modeling, and therefore expected to be more efficient than exisitng GMM-based IV methods relying on asymptotic theory when the sample size is not large enough, as indicated by Conley et al. (2008). In fact, we will show that our proposed method obtains more efficient and smaller MSE (mean squared error) estimates than GMM-based IV approach in the simulation study section when the sample size is not large.

2.3 MCMC algorithm

We apply the blocked Gibbs sampler (Ishwaran and James, 2011) to the posterior computation of the PSBPM parameters. Ishwaran and James (2011) proved that the case with finite number of classes l can be used to approximate the inference that is based on infinite classes with satisfactory accuracy when the maximum number of classes L is large enough. Therefore, blocked Gibbs sampler considers the case for truncation of the number of classes (e.g. $L = 20$), hence the simpler posterior computation. It is very similar to the Gibbs sampler except for the assignment of samples to each class.

We provide the detailed posterior computation using the MCMC estimation as follows.

1. Conditional for K_i $(i = 1, ..., N)$

   Let $\boldsymbol{K}_i$ be the indicator denoting where case i belongs, and $\boldsymbol{K}_i = l$ if case i belongs to class l. To assign samples to each class, generate K_i by $\sum_{l=1}^{L} \pi_{li}\delta_l(\cdot)$, where π_{li} is

   $$\pi_{li} = \frac{\pi_l(\boldsymbol{z}_i)N(\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{\beta}}_l, \boldsymbol{\Phi}_l)}{\sum_{l=1}^{L} \pi_l(\boldsymbol{z}_i)N(\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{\beta}}_l, \boldsymbol{\Phi}_l)},$$

   with $\pi_l(\boldsymbol{z}_i) = \Phi(\boldsymbol{\xi}_l(\boldsymbol{z}_i)) \prod_{h<l} \{1 - \Phi(\boldsymbol{\xi}_h(\boldsymbol{z}_i))\}$.

2. Conditional for Z_{il}^*

9

The latent variable Z_{il}^* is introduced in this step where $Z_{il} = \mathbf{1}(Z_{il}^* > 0)$ and

$$
Z_{il}^* \sim
\begin{cases}
N\left(\alpha_l - \sum_{k=1}^{q} \psi_{lq}\left|z_{iq} - \Omega_{lq}\right|, 1\right) \mathbf{1}_{\mathbb{R}^+} & \text{for } l = K_i \\
N\left(\alpha_l - \sum_{k=1}^{q} \psi_{lq}\left|z_{iq} - \Omega_{lq}\right|, 1\right) \mathbf{1}_{\mathbb{R}^-} & \text{for } l < K_i.
\end{cases}
$$

3. Conditional for $\alpha_l (l = 1, ..., L-1)$

   Simulate α_l from the following normal distribution.

$$
\alpha_l \sim N\left(\frac{\sum_{i:K_i \geq l} W_{il}^* + \mu_v}{N_l + 1}, \frac{1}{N_l + 1}\right),
$$

4. Conditional for $\psi_{lq}(l = 1, ..., L-1, q = 1, ..., m)$

   Simulate ψ_{lq} from the following left-truncated normal distribution.

$$
\psi_{lq} \sim N\left(\frac{\sigma_{\psi_q}^2 \mu_{\psi_q} + \sum_{i:K_i \geq l}\left|z_{iq} - \Omega_{lq}\right| U_{il}^*}{\sigma_{\psi_q}^2 + \sum_{i:K_i \geq l}\left|z_{iq} - \Omega_{lq}\right|^2}, \frac{1}{\sigma_{\psi_q}^2 + \sum_{i:K_i \geq l}\left|z_{iq} - \Omega_{lq}\right|^2}\right) \mathbf{1}_{\mathbb{R}^+},
$$

5. Conditional for $\Omega_{lq}(l = 1, ..., L-1, q = 1, ..., m)$

   Simulate Ω_{lq} from the following probability.

$$
\Pr\left(\Omega_{lq} = \Omega_{qb}^*\right) = \frac{\frac{1}{B_q} \prod_{i:K_i \geq l} N\left(Z_{il}^*; \mu_{\Omega,1}, 1\right)}{\sum_{q=1}^{B_q} \frac{1}{B_q} \prod_{i:K_i \geq l} N\left(Z_{il}^*; \mu_{\Omega,2}, 1\right)}.
$$

   where $\mu_{\Omega,1} = \alpha_l - \sum_{s=1,s\neq k}^{q} \psi_{ls}\left|z_{is} - \Omega_{ls}\right| - \psi_{lq}\left|z_{iq} - \Omega_{qb}^*\right|$ and $\mu_{\Omega,2} = \alpha_l - \sum_{s=1,s\neq k}^{q} \psi_{ls}\left|z_{is} - \Omega_{ls}\right| - \psi_{lq}\left|z_{iq} - \Omega_{qb}^*\right|$.

6. Conditional for μ_α

   Simulate μ_α from the following normal distribution.

$$
\mu_\alpha \sim N\left((L - 1 + \sigma_{\alpha_0}^{-2})^{-1}\left[\sum_{l=1}^{L-1} \alpha_l + \sigma_{\alpha_0}^{-2}\mu_{\alpha_0}\right], (L - 1 + \sigma_{\alpha_0}^{-2})^{-1}\right).
$$

7. Conditional for $\widetilde{\boldsymbol{\beta}}_l$

Let $n_l = \sum_{i=1}^{N} \mathbf{1}(\boldsymbol{K}_i = l)$. Simulate $\widetilde{\boldsymbol{\beta}}_l$ from the following multivariate normal and inverted Wishart distribution given $K_i = l$.

$$\widetilde{\boldsymbol{\beta}}_l \sim N\left(\boldsymbol{D}_{\boldsymbol{\beta}_l} d_{\boldsymbol{\beta}_l}, \boldsymbol{D}_{\boldsymbol{\beta}_l}\right)$$

where

$$\boldsymbol{D}_{\boldsymbol{\beta}_l} = \left(\sum_{i=1}^{n_l} \widetilde{\boldsymbol{X}}_{il}^t \boldsymbol{\Phi}_l^{-1} \widetilde{\boldsymbol{X}}_{il}\right)^{-1}, d_{\boldsymbol{\beta}_l} = \sum_{i=1}^{n_l} \left(\widetilde{\boldsymbol{X}}_{il}^t \boldsymbol{\Phi}_l^{-1} \widetilde{\boldsymbol{y}}_{il}\right),$$

and

$$\boldsymbol{\Phi}_l \sim IW\left(\left[\sum_{i=1}^{n_l} \left(\widetilde{\boldsymbol{y}}_{il} - \widetilde{\boldsymbol{X}}_{il}\widetilde{\boldsymbol{\beta}}_l\right)\left(\widetilde{\boldsymbol{y}}_{il} - \widetilde{\boldsymbol{X}}_{il}\widetilde{\boldsymbol{\beta}}_l\right)^t\right], n_l\right)$$

8. Conditional for β and $\boldsymbol{\gamma}$

Consistent with Rossi et al. (2005), the structural model parameters β and $\boldsymbol{\gamma}$ can be simulated by applying Bayesian linear regression draw with $N(0,1)$ error terms

$$\frac{y_i - E\left[\epsilon_{1,i}|\epsilon_{2,i}\right]}{\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}} = \frac{x_i\beta}{\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}} + \frac{\boldsymbol{w}_i\boldsymbol{\gamma}}{\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}} + u_i, \quad u_i \sim N(0,1)$$

where

$$\sigma_{\epsilon_{1,i}|\epsilon_{2,i}} = \sigma_{11,\omega_i} - \frac{\sigma_{12,\omega_i}^2}{\sigma_{11,\omega_i}^2}$$

and

$$E\left[\epsilon_{1,i}|\epsilon_{2,i}\right] = \frac{\sigma_{12,\omega_i}}{\sigma_{22,\omega_i}}\epsilon_{2,i}$$

3. Simulation study

We conduct two simulation studies in order to illustrate the performance of the proposed method when the reduced-form equation is not a linear combination with additive disturbance that follows Gaussian distribution. Simulation 1 considers the case where the reduced-form equation is a mixture model of two classes ($l = 2$). Simulation 2 considers the case of a non-linear reduced-form equation with skewly distributed (log-normally distributed) error terms. Each simulation is designed to cover the cases of weak instruments and strong instruments, namely, our simulation results in a experiment with four cells. We classify instruments as weak or strong according to the classical criteria proposed by Staiger and Stock (1997) and Stock and Yogo (2005).

These stipulate that when the F-statistics of the first stage regression is less than 10, we have a weak instrument case, otherwise we have a strong instrument case. We evaluate the proposed method by comparison with existing non-Bayesian and Bayesian IV methods.

In the two simulation studies, we evaluate the finite sample property of the estimators. We consider $1,000$ replications of the dataset and confirm if the true value of endogenous variable coefficient β can be recovered. Throughout the simulations, the sample size for each dataset is set to be $N = 200$. We calculate standard performance measures, such as empirical mean, standard deviation, the coverage of nominal 95% confidence (or credible) intervals (CI) of the estimate, and the deviation (MSE) from the true value. Moreover, we confirm, if the point estimate of β is positively statistically significant at 95% confidence level for non-Bayes approaches, and the 95% credible interval of β includes "zero" for Bayes approaches (denoted as "% of sig." in Table 1). Although it could be sometimes inappropriate, we take more practical perspective to compare the Bayesian credible intervals with the classical confidence intervals as Conley et al. (2008) did. The detailed experimental design is described in Supplementary Materials.

For comparison, we consider five non-Bayesian estimators and two Bayesian estimators. The first is the most widely used approach, namely, two-stage least squares (TSLS). The second alternative is limited information maximum likelihood (LIML). The third is a modification of the LIML estimator, namely, Fuller's procedure (F-LIML) (Fuller, 1977), which obtains better estimates when the instruments are weak. We also emplyed GMM based IV estimation (GMM-IV) as an alternative. Moreover, we compare the results from ordinary least squares (OLS). As for Bayesian alternatives, we employ the standard Bayesian IV method, which assumes that the joint distribution of the error terms is bivariate Gaussian (BayesIV), and Conley's semiparametric Bayesian IV method (C-BayesIV). For BayesIV and C-BayesIV, as well as the proposed method, burn-in of 5,000 iterations followed by 10,000 iterations were used for the posterior inference. All algorithms except for the proposed method and GMM-IV assume an ordinary linear regression model for the reduced-form equation, that is, $x_i = \boldsymbol{z}_i \boldsymbol{\delta} + \epsilon_{2,i}$. GMM-IV used the moment condition of $E(\epsilon_{1,i} \boldsymbol{z}_i) = \boldsymbol{0}$, therefore GMM-IV does not prespecify the functional form of the reduced-form equation. We used inverse variance of the moments as a weighting matrix, namely, the optimal weighting matrix. We calculated the intervals based on large sample approximations for the classical estimators.

This section presents the results of simulation study 1. The results of

simulation 2 are provided in Supplementary Materials.

Table 1   Result of simulation 1

METHOD	Weak Instruments β = 1						Strong Instruments β = 0.1					
	Mean	(s.d.)	Cov	MSE	MSE Ratio	% of sig.	Mean	(s.d.)	Cov	MSE	MSE Ratio	% of sig.
OLS	1.056	(0.023)	0.344	0.004	0.397	1.000	0.190	(0.030)	0.119	0.009	8.380	1.000
TSLS	1.040	(0.400)	0.992	0.103	10.876	0.689	0.101	(0.086)	0.965	0.007	6.500	0.276
LIML	1.021	(0.534)	0.992	0.158	16.675	0.601	0.096	(0.088)	0.966	0.008	7.445	0.263
F-LIML	1.042	(0.236)	0.995	0.044	4.701	0.913	0.101	(0.086)	0.968	0.007	6.447	0.277
GMM-IV	1.041	(0.378)	0.995	0.090	9.528	0.723	0.104	(0.089)	0.986	0.007	6.945	0.284
BayesIV	1.046	(0.321)	0.773	0.351	37.080	0.861	0.089	(0.096)	0.937	0.020	18.408	0.243
C-BayesIV	0.994	(0.528)	0.975	0.082	8.685	0.641	0.103	(0.083)	0.977	0.008	7.188	0.300
proposed	1.007	(0.100)	0.942	0.009	1.000	1.000	0.101	(0.030)	0.928	0.001	1.000	0.922

Empirical mean, standard deviation, coverage of nominal 95% CIs, and mean squared error of the estimates, MSE Ratio of the proposed method basis, and "% of sig." are described. "% of sig." indicates the percentage of the estimates which is positively statistically significant (p<0.05) (for non-Bayesian method) and the percentage of the 95% credible intervals which do not include "zero" (for Bayesian method) from 1000 simulations are described. OLS, ordinal least squares; TSLS, two stage least squares; LIML, limited imformation maximum likelihood; F-LIML, Fuller's modiication of the LIML; BayesIV, standard Bayesian IV method; C-BayesIV, Conley's semiparametric Bayesian IV method
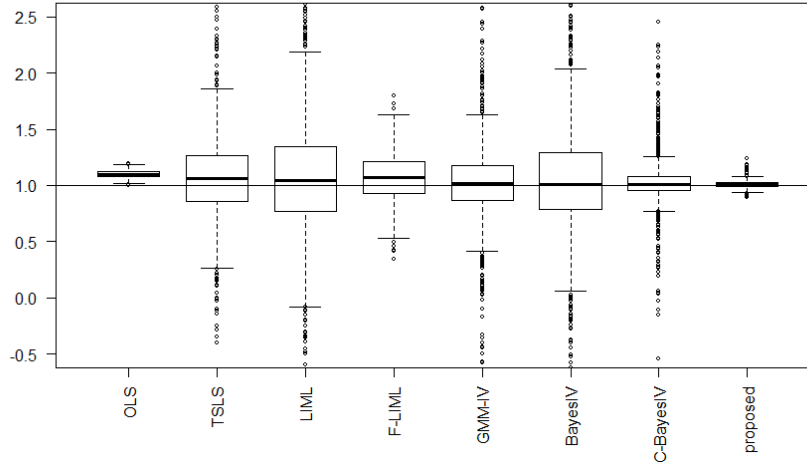

Fig.1. The solid horizontal line is the "true" coefficient value $\beta = 1$. The boxes span the range from the 25th to the 75th percentiles, and the whiskers extend to an area no more than 1.5 times the range from the 25th to the 75th percentiles from the box. The circles above and below the whiskers represent outliers. OLS, ordinary least squares; TSLS, two stage least squares; LIML, limited imformation maximum likelihood; F-LIML, Fuller's modiication of the LIML; BayesIV, standard Bayes IV method; C-BayesIV, Conley's semiparametric Bayes IV method

13

Table 1 shows the results of simulation 1, including the empirical mean, standard deviation, the coverage of nominal 95% confidence intervals (CI) of the estimate, the mean squared error (MSE) from the true value of β, and "*% of sig.*". Figure 1 shows the box plot of the estimates for β obtained by $1,000$ replications.

For both the weak and the strong instrument cases, OLS obtain overestimated results, since they ignore the existence of the instrumental variable. Therefore, the coverage of nominal 95% confidence intervals for OLS is very poor.

As for the weak instrument case, all estimation methods except for the proposed and C-BayesIV seem to be biased due to the weakness of the instruments. Moreover, the standard deviations of the estimates are totally different among the methods. The proposed method exhibits the best efficiency (the smallest standard deviation). Even though GMM-IV does not prespecify the reduced-from regression model and the distribution of the error terms, as is the case with the proposed, it shows the larger standard deviation than our semiparametric approach. The best coverage is obtained by the proposed method, which is very close to 0.95. For other Bayes models, BayesIV results in narrower credible intervals. Furthermore, C-BayesIV yields slightly wider intervals. Regarding non-Bayes methods, TSLS, LIML, F-LIML, and GMM-IV, show overly wider coverage and do not seem to obtain correctly the confidence intervals. These results indicate that weak instruments and misspecifications of IV models provide incorrect CI. Therefore, wrong parametric assumptions on the IV equations may lead to poor prediction power of the reduced-form equation and incorrect decision making in fields where results of statistical significance (or hypothesis testing) are crucial. The fifth and the sixth column shows the MSE from the true value of $\beta = 1$ and the MSE ratio when the MSE of the proposed method is set to be 1. It is seen that the existing methods yield from 4.7 to 37.1 times larger MSE than the proposed approach. All estimates from the proposed method do not contain zero for its credible intervals, by contrast, "*% of sig.*" of the estimates from the other methods are only 60% to 91%.

As far as the strong instrument case is concerned, all estimation methods except for OLS can recover the true value $\beta = 0.1$, since they have consistency, even when they misspecify the reduced-form equation. However, efficiency varies greatly among the methods. Hence, the MSE results are dispersed. Specifically, except for OLS estimates of inconsistency, the mean of the standard deviations of the proposed method is the smallest, indicating that misspecification of the reduced-form equation and the joint distribution

14

of the error terms results in less efficient estimates. GMM-IV has the property of semiparametric efficiency, but in this simulation study of finite sample with $N = 200$, GMM-based methodology seems to be less efficient even in the strong instruments case. Thus, our semiparametric specification is more suitable even in the strong instrument case. Existing methods yield at most 18.4 times larger MSE than the proposed. In addition, there is more than 92% chance that the estimates from the proposed method are "positively significant" (CIs do not cotain zero), whereas the estimates from the existing methods are significant with probability less than 30%. This indicates that even in the case of strong instruments, the proposed semiparametric approach can correctly detect significant variables.

4.   Real data analysis

Herein, we apply our methods to a real data example of Mendelian randomization. Mendelian randomization uses genotype information as an instrumental variable to correctly estimate the causal effect of a biomarker to a disease. It is an outstanding IV approach in that the appropriate genetic variant is independent of the confounders of the intermediate phenotype-outcome association. Moreover, it can affect the outcome only through the causal intermediate phenotype as long as it is related to the intermediate phenotype (Burgess et al., 2015).

The data used in this implementation are sourced from the Orkney Complex Disease Study (ORCADES) study, which is a genetic epidemiology study based on an isolated population in the north of Scotland (http://www.orcades. ed.ac.uk/orcades/). The data is open sourced and published in (http://www. homepages.ed.ac.uk/pmckeigu/instrumental/). It is based on the dataset described in McKeigue et al. (2010).

The aim of the analysis is to uncover the causal relationship between plasma urate level and the "metabolic syndrome". There exist two conflicting opinions on this issue. Feig et al. (2008) indicates that uric acid and the metabolic syndrome are positively related. Therefore, they are said to be have positive association (Lohsoonthorn et al., 2006, Onat et al., 2006, Choi and Ford, 2007, Hjortnaes et al., 2007, Sui et al., 2008). By contrast, there is indirect evidence in Waring et al. (2003) and Waring et al. (2006) supporting that uric acid is beneficial in avoiding the metabolic syndrome. Using Mendelian randomization with the dataset described in this section, McKeigue et al. (2010) showed the negative effect of uric acid on the metabolic syndrome. However, the effect was statistically not significant, with 95% credible interval of the logistic regression coefficient $(-2.91, 0.05)$.

15

In this example, plasma urate level is an endogenous variable and the metabolic syndrome is an outcome. As an instrumental variable for Mendelian randomization, we can use six SNPs (single nucleotide polymorphisms) in the urate transporter gene SLC2A9 which were typed using a TaqMan assay, rs737267, rs13129697, rs1014290, rs6449213, rs13131257, and rs4447863. In addition, the dataset include the information on age (average:53 years) and sex (40 % male). We include them in both the reduced-form and the structural equation. The plasma urate level is standardized to have mean 0 and standard deviation 1. The age is also scaled to have mean 0.

The outcome, the metabolic syndrome, is a dichotomous variable (equals one if the sample indicates metabolic syndrome, otherwise zero; 20 % of the sample is regarded as metabolic syndrome). Therefore, we consider the probit regression model for the structural equation. The original sample size is $1,017$, and the SNP information is available for 706 of these individuals. We used a complete-case sample, that is, individuals for whom all variables, including the outcome (metabolic syndrome), the endogenous variable (plasma urate level), the covariates (age and sex), and the instrumental variables (six SNP) were observed. The sample size resulting from the complete-case analysis is 647.

For comparison, we first used OLS regression. The estimated coefficient of plasma urate level is 0.44 and is statistically significant ($p < 0.01$). Moreover, we considered the classical method for the IV probit model. In this case, the plasma urate level coefficient is -0.67. However, this is not statistically significant, with confidence interval $(-1.52, 0.18)$. In addition, the F-statistics of the reduced-form equation (regression model of plasma urate level against six SNP) is 3.56. Hence, this is considered to be a weak instrument case. Therefore, the results obtained from a non-Bayesian IV method, in which the precision of the estimates depends on asymptotic approximation, are not reliable.

Subsequently, we applied the standard Bayesian IV method to the dataset. We considered an ordinary linear regression model as the reduced-form model and assumed that the distribution of the error terms is bivariate normal. Burn-in of $5,000$ iterations followed by $10,000$ iterations were used for the posterior inference. The posterior mean of the coefficient of plasma urate level is -0.42, with credible interval of $(-0.98, 0.03)$.

Finally, we applied the proposed method. We considered a truncation of the components $L = 20$. The prespecified parameters were same as in the previous simulation study. Burn-in of $5,000$ iterations followed by $10,000$ iterations were used for the posterior inference. The posterior mean of the plasma

urate level coefficient is -0.48, with credible interval of $(-1.01, -0.02)$. The interval obtained from the proposed is shorter than the standard Bayesian IV method and do not contain 0 in the interval. The results obtained from the proposed method are the opposite of OLS results, which do not consider the presence of the instruments.

In line with McKeigue et al. (2010), this indicates that high level of uric acid is negatively related to the metabolic syndrome. However, the results may be controversial, since they not consistent with an established theory supported by a large number prior studies. This article focuses on statistical methods for IV rather than the relationship between the urate acid level and the metabolic syndrome. The results, as well as the plausibility of the assumptions in the IV approach, should be verified by additional studies.

5.   Conclusion and discusions

We proposed a semiparametric Bayes instrumental variable estimation method. Our method does not specify the distribution of the reduced-form and the disturbances and is robust in the weak instrument case, as shown in the simulation studies. Therefore, application in Mendelian randomization is beneficial, since the instruments of genetic information are numerous and occasionally explain only a small portion of the variation in the biomarker.

There are several directions for future research. We considered the case of the structural equation is an ordinary linear regression model or probit regression model. An extension to the logistic regression model or other regression models is natural. As stated in Section 1, Bayesian methods are able to develop these models straightforwardly. In addition, further study is required to improve the computational speed of the proposed algorithm. Since our inference is based on the MCMC algorithm with nonparametric Bayes parameters, the computation time required to obtain valid estimates is greater than that of the existing non-Bayesian and Bayesian methods.

<div align="center">Supplementary Materials</div>

1.   MCMC algorithm for binary outcome case

We can extend MCMC algorithm for binary outcome case. Consider the following representation of probit model by latent variable y_i^*:

$$
\begin{cases} y_i^* = x_i\beta + \boldsymbol{w}_i\boldsymbol{\gamma} + \epsilon_{1,i} \\ x_i = g\left(\boldsymbol{z}_i; \boldsymbol{\delta}\right) + \epsilon_{2,i} \end{cases}, \boldsymbol{\epsilon}_i \equiv \begin{pmatrix} \epsilon_{1,i} \\ \epsilon_{2,i} \end{pmatrix} \sim f
$$

and

<div align="center">17</div>

$$y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & y_i^* \leq 0 \end{cases}$$

Posterior simulation requires following 3 steps in addition to the main MCMC computation described above, and replace y_i in the algorithm by latent variable y_i^*.

1. Conditional for y_i^* $(i = 1, ..., N)$

   Simulate y_i^* from the following truncated normal distribution.

$$y_i^* \sim \begin{cases} TN_{(0,\infty)}\left(x_i\beta + \boldsymbol{w}_i\,\boldsymbol{\gamma} + \frac{\sigma_{12,l}}{\sigma_{22,l}^2}\left[x_i - \boldsymbol{z}_i\boldsymbol{\delta}_l\right], \left(1 - \frac{\sigma_{12,l}^2}{\sigma_{22,l}^2}\right)\right) \\ TN_{(-\infty,0)}\left(x_i\beta + \boldsymbol{w}_i\,\boldsymbol{\gamma} + \frac{\sigma_{12,l}}{\sigma_{22,l}^2}\left[x_i - \boldsymbol{z}_i\boldsymbol{\delta}_l\right], \left(1 - \frac{\sigma_{12,l}^2}{\sigma_{22,l}^2}\right)\right) \end{cases},$$

   where $\boldsymbol{\Sigma}_l = \begin{pmatrix} 1 & \sigma_{12,l} \\ \sigma_{12,l} & \sigma_{22,l}^2 \end{pmatrix}$.

2. Conditional for $\sigma_{12,l}$ $(l = 1, ..., L)$

   Simulate $\sigma_{12,l}$ from the following normal distribution.

$$\sigma_{12,l} \sim N(D_\sigma d_\sigma, D_\sigma),$$

   where $D_\sigma = \left(\sum_{i=1}^{n_l} \left(y_i^* - x_i\beta + \boldsymbol{w}_i\boldsymbol{\gamma}\right)^2 / \sigma_{v,l}^2 + V_{0,l}^{-1}\right)^{-1}$,

   $d_\sigma = \sum_{i=1}^{n_l} \left(y_i^* - x_i\beta + \boldsymbol{w}_i\boldsymbol{\gamma}\right)\left(x_i - \boldsymbol{z}_i\boldsymbol{\delta}_l\right)/\sigma_{v,l}^2 + V_{0,l}^{-1}\mu_{0,l}$, and $\sigma_{v,l}^2 = \sigma_{22,l}^2 - \sigma_{12,l}^2$, and the prior for $\sigma_{12,l} \sim N(\mu_{0,l}, V_{0,l})$.

3. Conditional for $\sigma_{v,l}^2 (l = 1, ..., L)$

   Simulate $\sigma_{v,l}^2$ from the following inverse Gamma distribution.

$$\sigma_{v,l}^2 \sim IG\left(\frac{n_l}{2} + a, b^{-1} + \frac{1}{2}\sum_{i=1}^{n_l} \{x_i - \boldsymbol{z}_i\boldsymbol{\delta}_l - \sigma_{12,l}\left(y_i^* - x_i\beta + \boldsymbol{w}_i\boldsymbol{\gamma}\right)\}\right),$$

   with the prior for $\sigma_{v,l}^2 \sim IG(a, b)$.

2. Detailed simulation design and results

Herein, we describe in detail the simulation study mentioned in Section 3.

2.1 *Experimental design*

Throughout the simulation, we consider the case of linear structural equation with one endogenous variable. This experimental design is justified by the survey conducted in Chernozhukov and Hansen (2008); they noticed that in the leading journals of economics from 1999 to 2004, 89 out of 129 articles regarded only one explanatory variable as endogenous.

Figure 2 shows boxplots of the F-statistics distribution of simulation 1 and simulation 2 for the weak and strong instrument case. We classify instruments as weak or strong according to the classical criteria proposed by Staiger and Stock (1997) and Stock and Yogo (2005) as noted in the main paper.

Through the simulation studies, we adopt the same default choices for hyperparameters as in Chung and Dunson (2009), namely, $L = 20$, $\mu_{\alpha_0} = 0$, $\sigma^2_{\alpha_0} = 1$, $\mu_{\psi_q} = 0$, $\sigma^2_{\psi_q} = 100$, and Ω^*_{qb} are 50 equally spaced grid points in $(-3.5, 3.5)$.
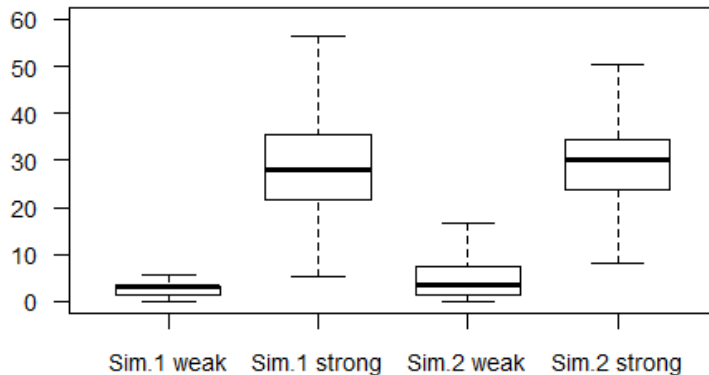

Fig.2. The boxes span the range from the 25th to the 75th percentiles, and the whiskers extend to an area no more than 1.5 times the range from the 25th to the 75th percentiles from the box. Sim.1 weak, simulation 1 for weak instruments; Sim.1 strong, simulation 1 for strong instruments; Sim.2 weak, simulation 2 for weak instruments; Sim.2 strong, simulation 2 for strong instruments

2.2 *Setup for simulation study 1*

In this study, we evaluate the finite sample property of the estimators when the true reduced-form equation is a finite mixture regression model. We consider the case with a sample size $N = 200$ and a mixture model with two classes ($l = 2$). We assume 1-dimensional endogenous regressor x_i, 3-dimensional exogeneous regressor $\boldsymbol{w}_i$, whose first elements are set to be 1 (i.e.

intercept), and 5-dimensional instruments $\boldsymbol{z}_i$, whose first elements are set to be 1 and are independent of $\epsilon_{2,i}$. The true relationship of the full model is as follows.

$$\begin{cases} y_i = x_i\beta + \boldsymbol{w}_i\boldsymbol{\gamma} + \epsilon_{1,i} \\ x_i = \sum_{l=1}^{2} \mathbf{1}_{i\in l}\boldsymbol{z}_i\boldsymbol{\delta}_l + \epsilon_{2,i} \end{cases}, \boldsymbol{\epsilon}_i \equiv \begin{pmatrix} \epsilon_{1,i} \\ \epsilon_{2,i} \end{pmatrix} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_l\right) \qquad (.1)$$

For weak instrument case, we set $(\pi_1, \pi_2) = (0.5, 0.5)$, $\boldsymbol{\delta}_1 = (-1, 2, 2, 0, 0)^t$, $\boldsymbol{\delta}_2 = (-1, -2, -2, 0, 0)^t$, $\boldsymbol{\gamma} = (1, 2, 3)^t$, $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \left(\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \right)$, and the interested parameter $\beta = 1$. We generate the exogeneous random variables from $\boldsymbol{w}_{-1,i} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{pmatrix} \right)$, where $\boldsymbol{w}_{-1,i}$ denote the components of $\boldsymbol{w}_i$ with the first component 1 removed. As for $\boldsymbol{z}_i$, we first generate $\boldsymbol{z}^*_{-1,i} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right)$, where $\boldsymbol{z}^*_{-1,i}$ denotes the components of $\boldsymbol{z}_i$ without the first component 1 and the component of $\boldsymbol{w}_i$, and set $\boldsymbol{z}_i$ to be $\boldsymbol{z}_i = \left(1, \boldsymbol{z}^*_{-1,i}, \boldsymbol{w}_i\right)$.

For the case of strong instruments with small coefficients, we replace $\boldsymbol{\delta}_l$ with $\boldsymbol{\delta}_1 = (-1, 2, 2, 0, 0)^t$ and $\boldsymbol{\delta}_2 = (-1, -1, 0, 0, 0)^t$ and the parameter of interest β with $\beta = 0.1$. The other true parameters remain unchanged.

In this simulation setup, the mean of the F-statistics of the reduced-form model for the weak and the strong instrument case is 2.40 and 29.11, respectively.

The results are provided in the main article.

2.3 *Setup for simulation study 2*

In this study, we consider the case with a sample size $N = 200$, a non-linear reduced-form equation, and skewly distributed (log-normally distributed) error terms. We assume 1-dimensional endogenous regressor x_i, 3-dimensional exogeneous regressor $\boldsymbol{w}_i$ whose first elements are set to be 1 (i.e., intercept), and 4-dimensional instruments $\boldsymbol{z}_i$ that are independent of $\epsilon_{2,i}$. The true relationship of the model is as follows.

$$\begin{cases} y_i = x_i\beta + \boldsymbol{w}_i\boldsymbol{\gamma} + \epsilon_{1,i} \\ x_i = 2\ln\left(\left|z^*_{-1,i} - 1\right| + 1\right) sgn\left(z^*_{-1,i} - 1\right) + \epsilon_{2,i} \end{cases}, \boldsymbol{\epsilon}_i \equiv \begin{pmatrix} \epsilon_{1,i} \\ \epsilon_{2,i} \end{pmatrix} = \boldsymbol{v} - \boldsymbol{c};\ \boldsymbol{v} \sim \ln N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$$
$$(.2)$$

where $sgn(\chi_i)$ takes the value $1, 0, -1$ if χ_i is positive, zero, and negative, respectively. $\boldsymbol{c}$ is taken so that the mean of $\boldsymbol{\epsilon}_i$ is zero. In this setting, there is S-shaped relationship between x_i and $z^*_{-1,i}$. Again, $z^*_{-1,i}$ denotes

the components of $\boldsymbol{z}_i$ with the first component 1 and the component of $\boldsymbol{w}_i$ removed. Hence, in this data generating process, $z_{-1,i}^*$ is one dimension.

For weak instrument case, we set $\boldsymbol{\gamma} = (1, 2, 3)^t$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2.5 \end{pmatrix}$, and the interested parameter $\beta = 1$. We generate the exogenous random variables from $\boldsymbol{w}_i \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{pmatrix} \right)$, $z_{-1,i}^* \sim N(0, 1)$. For strong instrument with small coefficient case, we replace $\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1.25 \end{pmatrix}$ and the interested parameter β with $\beta = 0.3$. The other true parameters are remained to be unchanged.

In this simulation setup, the mean of the F-statistics of the reduced-form model for the weak and the strong instrument case is 5.13 and 29.85, respectively.

2.4 Results of Simulation study 2

Table 2 shows the results of simulation 2. It is seen that for both cases, OLS obtains overestimated results. Therefore, the coverage of nominal 95% confidence intervals is very poor.

Regarding the weak instrument case, all estimation methods except for the proposed approach are unstable. Specifically, the standard deviation of the estimates is larger than that of the proposed approach. The best coverage is obtained by the proposed method. This is very close to 0.95. As for the other Bayes models, BayesIV and C-BayesIV obtain somewhat narrower credible intervals. By contrast, non-Bayes approaches again yield overly wider credible intervals. These results also indicate that weak instruments and misspecifications of IV models yield incorrect CI. The results on the MSE and the MSE ratio show that existing methods yield from 5.6 to 36.7 times larger MSE than the proposed approach. In addition, "% of sig." of the existing methods are only with probability 86% at minimum, whereas all estimates of β from the proposed method do not include zero for its credible intervals.

As for the strong instrument case, all estimation methods except for OLS can recover the true value $\beta = 0.3$. However, efficiency varies among the methods. More specifically, except for OLS, the mean of the standard deviations of the proposed model is the smallest. In the simulation 2 settings, which consider the non-normal disturbances with finite samples, GMM-IV does not outperform the proposed for both weak and strong instruments case even though it do not require the normality of the error terms. Additionally, "% of sig." of the proposed method is 96%, whereas those of the

competitors are only 39% to 59%.

Table 2   Result of simulation 2

METHOD	Weak Instruments β = 1						Strong Instruments β = 0.3					
	Mean	(s.d.)	Cov	MSE	MSE Ratio	% of sig.	Mean	(s.d.)	Cov	MSE	MSE Ratio	% of sig.
OLS	1.060	(0.020)	0.293	0.005	1.428	1.000	0.542	(0.051)	0.034	0.073	23.909	1.000
TSLS	0.991	(0.350)	0.990	0.119	32.783	0.870	0.292	(0.160)	0.963	0.031	10.129	0.490
LIML	0.990	(0.370)	0.990	0.133	36.651	0.861	0.282	(0.160)	0.965	0.033	10.780	0.480
F-LIML	1.016	(0.145)	0.990	0.022	5.957	0.995	0.296	(0.152)	0.963	0.022	7.324	0.527
GMM-IV	1.023	(0.341)	0.983	0.111	30.703	0.889	0.309	(0.159)	0.982	0.030	9.871	0.519
BayesIV	0.997	(0.154)	0.738	0.120	33.225	0.938	0.267	(0.169)	0.930	0.037	12.080	0.391
C-BayesIV	1.028	(0.124)	0.828	0.020	5.625	0.967	0.311	(0.139)	0.910	0.020	6.673	0.592
proposed	0.999	(0.072)	0.953	0.004	1.000	1.000	0.295	(0.073)	0.968	0.003	1.000	0.956

Empirical mean, standard deviation, coverage of nominal 95% CIs, and mean squared error of the estimates, MSE Ratio of the proposed method basis, and "% of sig." are described. "% of sig." indicates the percentage of the estimates which is positively statistically significant (p<0.05) (for non-Bayesian method) and the percentage of the 95% credible intervals which do not include "zero" (for Bayesian method) from 1000 simulations are described. OLS, ordinal least squares; TSLS, two stage least squares; LIML, limited imformation maximum likelihood; F-LIML, Fuller's modiication of the LIML; BayesIV, standard Bayesian IV method; C-BayesIV, Conley's semiparametric Bayesian IV method

REFERENCES

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Stat.* **20**, 46–63.

Baiocchi, M., Cheng, J. and Small, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33**, 2297–2340.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* **25**, 77–83.

Burgess, S., Scott, R. A., Timpson, N. J., Smith, G. D., Thompson, S. G. and Consortium, E.-I. (2015). Using published data in mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30**, 543–552.

Chao, J. and Phillips, P. (2011). Posterior distributions in limited information analysis of the simultaneous equations model using the jeffreys prior. *J. Econometrics* **87**, 49–86.

Chernozhukov, V. and Hansen, C. (2008). The reduced form: A simple approach to inference with weak instruments. *Econ. Lett.* **100**, 68–71.

Chib, S. (2007). Analysis of treatment response data without the joint distribution of potential outcomes. *J. Econometrics* **140**, 401–412.

Choi, H. K. and Ford, E. S. (2007). Prevalence of the metabolic syndrome in individuals with hyperuricemia. *Am. J. Med.* **120**, 442–447.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Am. Statist. Ass.* **104**, 1646–1660.

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

Darolles, S., Fan, Y., Florens, J. P. and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica* **79**, 1541–1565.

Dreze, J. H. (1976). Bayesian limited information analysis of the simultaneous equations model. *Econometrica* **44**, 1045–1075.

Dunson, D. B., Pillai, N. and Park, J. H. (2007). Bayesian density regression. *J. Roy. Stat. Soc. B Met.* **69**, 163–183.

Feig, D. I., Kang, D. H. and Johnson, R. J. (2008). Uric acid and cardiovascular risk. *N. Engl. J. Med.* **359**, 1811–1821.

Florens, J. P. and Simoni, A. (2012). Nonparametric estimation of an instrumental regression: A quasi-bayesian approach based on regularized posterior. *J. Econometrics* **170**, 458–475.

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica* **45**, 939–953.

Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *J. Econometrics* **75**, 121–146.

Hall, P. and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Stat.* **33**, 2904–2929.

Hirano, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica* **70**, 781–799.

Hjortnaes, J., Algra, A., Olijhoek, J., Huisman, M., Jacobs, J., van der Graaf, Y. and Visseren, F. (2007). Serum uric acid levels and risk for vascular diseases in patients with metabolic syndrome. *J. Rheumatol* **34**, 1882–1887.

Hood, W. and Koopmans, T. C. (1953). *Studies in econometric method.* John Wiley and Son, New York.

Hoogerheide, L., Kleibergen, F. and Van Dijk, H. (2007). Natural conjugate priors for the instrumental varaibles regression model applied to the angrist-krueger data. *J. Econometrics* **138**, 63–103.

Hoshino, T. (2013). Semiparametric Bayesian estimation for marginal para-

metric potential outcome modeling: Application to causal inference. *J. Am. Statist. Ass.* **108**, 1189–1204.

Ishwaran, H. and James, L. F. (2011). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Ass.* **96**, 161–173.

Kato, K. (2013). Quasi-bayesian analysis of nonparametric instrumental variables models. *Ann. Stat.* **41**, 2359–2390.

Kleibergen, F. and Zivot, E. (2003). Bayesian and classical approaches to instrumental variable regression. *J. Econometrics* **114**, 29–72.

Kleibergen, F. R. (1997). Equality restricted random variables: Densities and sampling algorithms. *Econometric Institute Research Papers, No. EI 9662-/A* .

Kunihama, T. and Dunson, D. B. (2016). Nonparametric bayes inference on conditional independence. *Biometrika* **103**, 35–47.

Liao, Y. and Jiang, W. (2011). Posterior consistency of nonparametric conditional moment restricted models. *Ann. Stat* **39**, 3003–3031.

Lohsoonthorn, V., Dhanamun, B. and Williams, M. A. (2006). Prevalence of hyperuricemia and its relationship with metabolic syndrome in thai adults receiving annual health exams. *Arch. Med. Res.* **37**, 883–889.

McKeigue, P. M., Campbell, H., Wild, S., Vitart, V., Hayward, C., Rudan, I., F., W. A. and Wilson, J. F. (2010). Bayesian methods for instrumental variable analysis with genetic instruments ('mendelian randomization'): example with urate transporter slc2a9 as an instrumental variable for effect of urate levels on metabolic syndrome. *Int. J. Epidemiol.* **39**, 375–393.

Miyazaki, K. and Hoshino, T. (2009). A bayesian semiparametric item response model with dirichlet process priors. *Psychometrika* **74**, 375–393.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–1578.

Onat, A., Uyarel, H., Hergenş, G., Karabulut, A., Albayrak, S., Sarı, I., Yazici, M. and Keleş, I. (2006). Serum uric acid is a determinant of metabolic syndrome in a population-based study. *Am. J. Hypertens.* **19**, 1055–1062.

Pati, D., Dunson, D. B. and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.* **116**, 456–472.

Ramsahai, R. R. and Lauritzen, S. L. (2011). Likelihood analysis of the binary instrumental variable model. *Biometrika* **98**, 987–994.

Rossi, P. E., Allenby, G. M. and McCulloch, R. (2005). *Bayesian statistics and marketing*. Wiley, NewYork.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Stat.*

Sinica **4**, 639–650.

Shen, W., Tokdar, S. T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with dirichlet mixtures. *Biometrika* **100**, 623–640.

Staiger, D. O. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65**, 557–586.

Stock, J. and Yogo, M. (2005). Testing for weak instruments in linear iv regression. *Andrews DWK Identification and Inference for Econometric Models* pages 80–108.

Sui, X., Church, T. S., Meriwether, R. A., Lobelo, F. and Blair, S. N. (2008). Uric acid and the development of metabolic syndrome in women and men. *Metabolism* **57**, 845–852.

Theil, H. (1953). *Estimation and simultaneous correlation in complete equation systems.* North-Holland Publishing Co., Amsterdam.

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. (1999). Bayesian nonparametric inference for random distributions and related functions. *J. Roy. Stat. Soc. B Met.* **61**, 485–527.

Wang, L., Robins, J. M. and Richardson, T. S. (2017). On falsification of the binary instrumental variable model. *Biometrika* **104**, 229–236.

Waring, W. S., Convery, A., Mishra, V., Shenkin, A., Webb, D. J. and Maxwell, S. R. J. (2003). Uric acid reduces exercise-induced oxidative stress in healthy adults. *Clin. Sci.* **105**, 425–430.

Waring, W. S., McKnight, J. A., Webb, D. J. and Maxwell, S. R. (2006). Uric acid restores endothelial function in patients with type 1 diabetes and regular smokers. *Diabetes* **55**, 3127–3132.