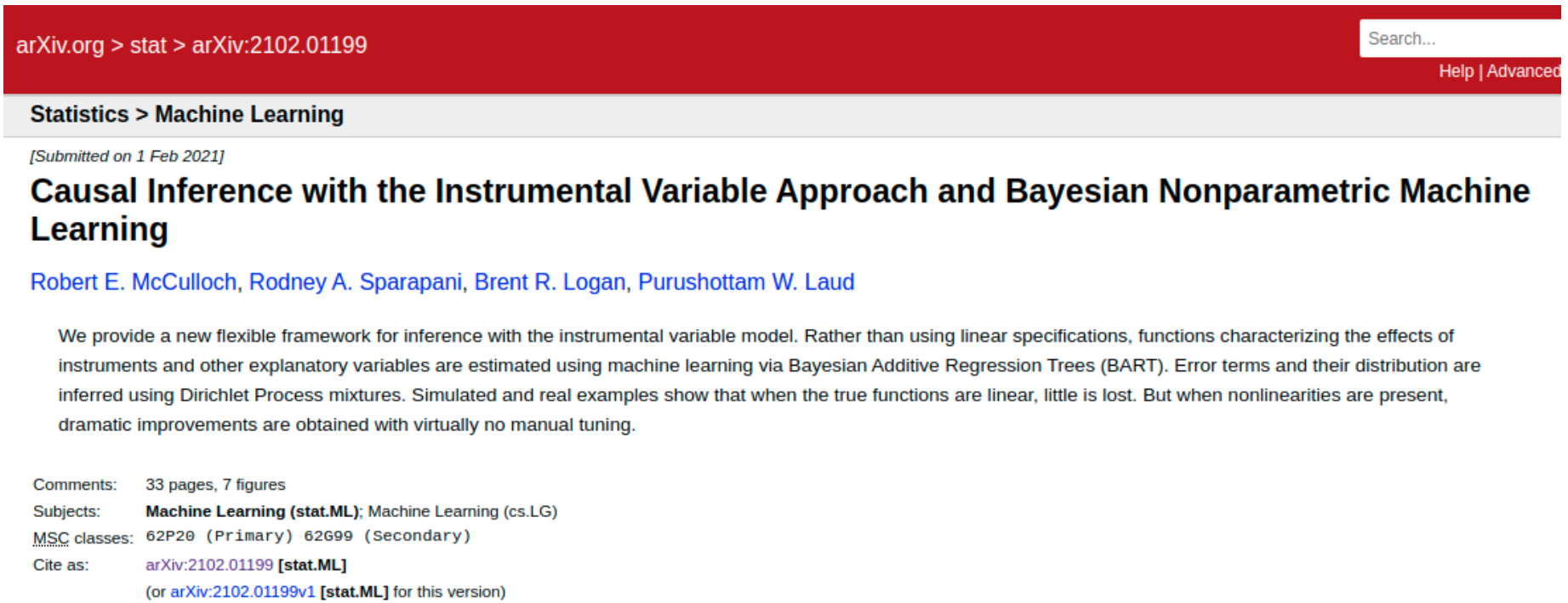
Causal Inference with the Instrumental Variable Approach and Bayesian Nonparametric Machine Learning

Robert E. McCulloch, School of Mathematical and Statistical Sciences, Arizona State University, Rodney Sparapani, Brent Logan, Purushottam Laud, Division of Biostatistics, Medical College of Wisconsin.

1. Basic References
2. Basic and Flexible IV
3. Prior Choice
4. Gibbs Sampler
5. Simulated Examples
6. Card Example

1. Basic References

There is a paper on arxiv and the paper has instructions for installing an `R` package from github.



arXiv.org > stat > arXiv:2102.01199

Search...

Help | Advanced

Statistics > Machine Learning

[Submitted on 1 Feb 2021]

Causal Inference with the Instrumental Variable Approach and Bayesian Nonparametric Machine Learning

Robert E. McCulloch, Rodney A. Sparapani, Brent R. Logan, Purushottam W. Laud

We provide a new flexible framework for inference with the instrumental variable model. Rather than using linear specifications, functions characterizing the effects of instruments and other explanatory variables are estimated using machine learning via Bayesian Additive Regression Trees (BART). Error terms and their distribution are inferred using Dirichlet Process mixtures. Simulated and real examples show that when the true functions are linear, little is lost. But when nonlinearities are present, dramatic improvements are obtained with virtually no manual tuning.

Comments:	33 pages, 7 figures
Subjects:	**Machine Learning (stat.ML)**; Machine Learning (cs.LG)
MSC classes:	62P20 (Primary) 62G99 (Secondary)
Cite as:	arXiv:2102.01199 **[stat.ML]**
	(or arXiv:2102.01199v1 **[stat.ML]** for this version)

Key References:

Conley (**CHMR**) does linear IV with nonparametric error distributions using DPM (Dirichlet Process Mixtures).

Chipman (Bayesian Additive Regression Trees, **BART**) does Bayesian Machine Learning in the spirit of boosting.

George does BART with DPM errors.

H. Chipman, E. George, and R. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

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

M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. doi: 10.1080/01621459.1990.10476213.

E. George, P. Laud, B. Logan, R. McCulloch, and R. Sparapani. Fully nonparametric Bayesian Additive Regression Trees. *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part B (Advances in Econometrics)*, 40:89–110, 2019.

We are simply trying to get rid of linear and normal errors for IV!!

2. Basic and Flexible IV

The classic linear approach to IV modeling is expressed by the following two equations:

The **treatment equation**

$$T_i = \mu_T + \gamma' z_i + \alpha' x_i + \epsilon_{Ti}$$

The **outcome equation**

$$Y_i = \mu_Y + \beta\, T_i + \delta' x_i$$

In the *treatment equation* where we try to understand how much treatment (T) is being given to each subject (i^{th} subject).

Both z and x are things we can measure (mostly about the treated subject) that potentially affect T.

$$
\begin{aligned}
T_i &= \mu_T + \gamma' z_i + \alpha' x_i + \epsilon_{Ti} \\
Y_i &= \mu_Y + \beta\, T_i + \delta' x_i + \epsilon_{Yi}.
\end{aligned}
$$

The second equation is *the outcome equation*.

Our fundamental goal is to understand how the treatment (T) affects the outcome (Y).

This is captured by the single parameter β.

We consider the possibility that the variables in x may also affect Y.

The x variables are often called "confounders".

$$
\begin{aligned}
T_i &= \mu_T + \gamma' z_i + \alpha' x_i + \epsilon_{Ti} \\
Y_i &= \mu_Y + \beta\, T_i + \delta' x_i + \epsilon_{Yi}.
\end{aligned}
$$

Because we did not assign T, but merely observe it, we cannot assume that the errors in the two equations are independent.

There may be unobserved variables affecting both T and Y.

Thus, we cannot simply regress Y on T and x to understand the causal effect of T.

The *"causal effect"* is the change in Y due to an intervention in the system where we actively change T.

$$
\begin{aligned}
T_i &= \mu_T + \gamma' z_i + \alpha' x_i + \epsilon_{Ti} \\
Y_i &= \mu_Y + \beta\, T_i + \delta' x_i + \epsilon_{Yi}.
\end{aligned}
$$

This is the fundamental *causal inference* problem.

We want to predict a process (intervening to change T) we have not actually observed.

We love experiments because we actually intervene to change T so we are observing what we want to predict.

$$
\begin{aligned}
T_i &= \mu_T + \gamma' z_i + \alpha' x_i + \epsilon_{Ti} \\
Y_i &= \mu_Y + \beta\, T_i + \delta' x_i + \epsilon_{Yi}.
\end{aligned}
$$

The variables z are the *instruments*.

We assume that variation in z is comparable to us "exogenously" intervening in the system to cause changes in T.

Variation in T due to z is *good variation* in that it is comparable to us running an experiment.

Card Example

A classic IV example.

We will use it in our examples.

- Y: wages
- T: years of schooling
- z: how close to a two year college, how close to a four year college
- x: years of experience, race, lives in standard metropolitan area, live in south

Somewhat plausibly, z could be in the treatment equation but not in the outcome equation.

We assume that variation in T due to z, is comparable to variation in T due to an experiment.
This identifies the causal effect of T on Y.

Flexible IV

Classic Bayesian IV assume linearity and bivariate iid normal errors.

Then applied work has to worry about the "specification".

That is, how do we transform or feature engineer the x and z to enter linearly in our two equations.

Of course, inference about β may depend on the specification.

This makes a mess.

Our goal is to eliminate the need to assume that the relationships are linear and to make minimal assumptions about the nature of the errors.

Flexible IV

Can we get the causal inference without making assumptions??

$$
\begin{aligned}
T_i &= f(z_i, x_i) + \epsilon_{Ti} \\
Y_i &= \beta\, T_i + h(x_i) + \epsilon_{Yi}.
\end{aligned}
$$

$$
\epsilon_i = (\epsilon_{Ti}, \epsilon_{Yi})' \sim N(\mu_i, \Sigma_i)
$$

$$
f, h \sim BART, \;\; (\mu_i, \Sigma_i) \sim DPM.
$$

We use BART to flexibly model the functions f and h.

We use the tried-and-true Dirichlet process mixture model with a bivariate normal base to model the errors.

Intercepts go into the errors.

Note:

In this talk and the currently available `R` package we are still linear in T in the outcome equation.

We intend to relax this, but we think this version may be very appealing to practitioners!!

BART

BART: Bayesian Additive Regression Trees.

- BART is able to learn high-dimensional, complex, non-linear relationships
- BART is a fully Bayesian procedure with an effective MCMC algorithm that inherently provides an assessment of uncertainty.
- BART often obtains an adequate fit with minimal tuning.
- Multiple additive BART models can be embedded in a larger model (as in our model !!).
- Simple prior: $f(x) \sim N(0, \sigma_f^2)$, you can choose σ_f.

DPM

Each observation gets to have its own (μ_i, Σ_i).

But, the DPM machinery allows us to uncover a set of (μ_j^*, Σ_j^*), $j = 1, 2, \ldots, J$ such that each

$$\text{for each } i,\ (\mu_i, \Sigma_i) = (\mu_j^*, \Sigma_j^*), \quad \text{for some } j.$$

J is much smaller than n so we partition the observations into joint subsets so that within a subset all the observations have the same (μ_j^*, Σ_j^*).

Connection to Mixture of Normals

Given (μ_i, Σ_i), $i = 1, 2, \ldots, n$, let

$$\{(\mu_j^*, \Sigma_j^*)\},\ j = 1, 2, \ldots, I$$

be the unique (μ, Σ) pairs.

Let

$$p_j = \frac{\#\left[(\mu_i, \Sigma_i) = (\mu_j^*, \Sigma_j^*)\right]}{n}$$

Then

$$\epsilon \approx \sum_{j=1}^{I} p_j\, N(\mu_j^*, \Sigma_j^*)$$

3. Prior Choice

While our examples will suggest that our results are not *too* sensitive to the prior specifications, we do see some sensitivity.

We are still working on this.

You can explain (T, Y) with the DPM or the BARTs !!

We first rescale both T and Y by subtracting off the sample mean and then dividing by the sample standard deviation.

The error DPM prior is then designed to be informative, but flexible enough to cover the full range of the data.

For the DPM we follow Conley, Hansen, McCulloch, and Rossi. This may be too spread out.

As a base case we use

$$f(x,z) \sim N(0,\sigma_f^2), \quad h(x) \sim N(0,\sigma_h^2)$$

with $\sigma_f = \sigma_h = 1.2$.

We will consider the sensitivity of the results to the choice of σ_f and σ_h.

These should be be chosen as in DPMBART.

4. Gibbs Sampler

$$
\begin{aligned}
T_i &= f(z_i, x_i) + \epsilon_{Ti} \\
Y_i &= \beta\, T_i + h(x_i) + \epsilon_{Yi}.
\end{aligned}
$$

$\theta_i = (\mu_i, \Sigma_i)$.

We use the obvious Gibbs sampler:

$$
\begin{aligned}
&f \mid h, \beta, \{\theta_i\}, D \\
&h \mid f, \beta, \{\theta_i\}, D \\
&\beta \mid f, h, \{\theta_i\}, D \\
&\{\theta_i\} \mid f, h, \beta, D
\end{aligned}
$$

Only the f draw is a little tricky, but they all reduce to standard conjugate Bayes or weighted BART.

$$
Y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, w_i\, \sigma^2).
$$

5. Simulated Examples

$$
\begin{aligned}
T_i &= f(z_i, x_i) + \epsilon_{Ti} \\
Y_i &= \beta\, T_i + h(x_i) + \epsilon_{Yi}.
\end{aligned}
$$

We will consider a nonlinear pair of function:

$$
\begin{aligned}
f(x, z) &= x_1 + .5\, x_1 x_2 + .5 x_2^2 + z_1 + z_2 x_1 + .5 z_2^2 \\
h(x) &= x_1 - .25 x_1 x_2^3 + x_3
\end{aligned}
$$

And a linear pair:

$$
\begin{aligned}
f(x, z) &= x_1 + x_2 + x_3 + z_1 + z_2 \\
h(x) &= x_1 - x_2 + .5 x_4 \ .
\end{aligned}
$$

True β: $\beta = 1$.

x and z

Each coordinate of both x and z are iid uniform on the interval $(-2, 2)$.

For x, we simulate x_j, $j = 1, 2, \ldots, 10$.

For z we simulated z_j, $j = 1, 2, \ldots 5$.

So, there are 10 x variables and 5 potential z instruments.

Errors:

For the error distribution we use:

$$\begin{aligned} \epsilon_T &= \sigma_T Z_T \\ \epsilon_Y &= \gamma Z_T + \sigma_Y Z_Y \end{aligned}$$

where (Z_T, Z_Y) are indepenent t_ν random variables with $\nu = 5$.

We let $\sigma_T = 1$ and $(\gamma, \sigma_Y) = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$.

With these choices, both errors have the same variance as the Z's and the correlation is $1/\sqrt{2} \approx .707$.

The pair of errors $(\epsilon_{Ti}, \epsilon_{Yi})$ are iid over observations.

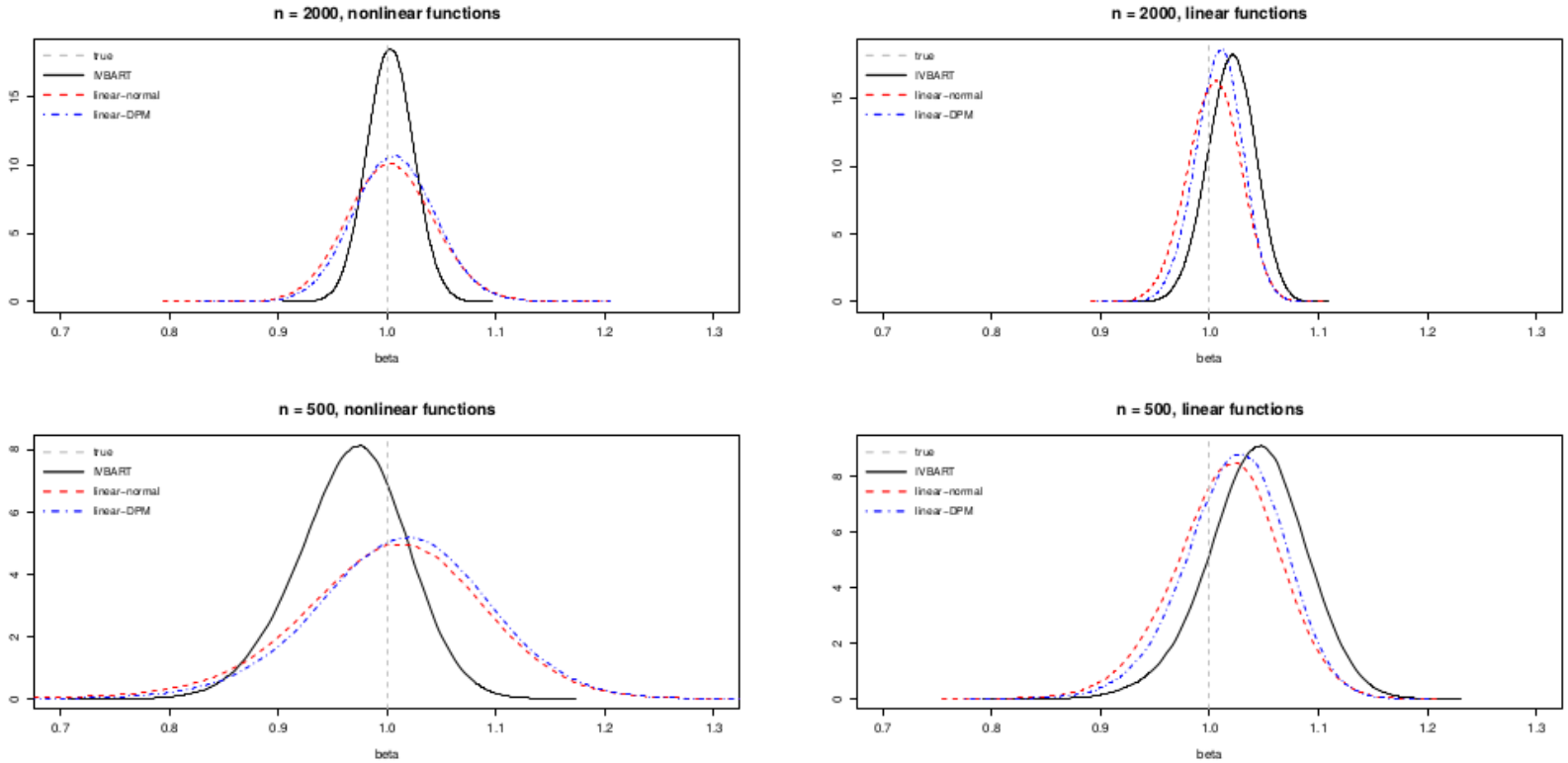
We consider four different simulation scenarios by letting the sample size n be 2,000 or 500 and letting the functions be nonlinear or linear.

We draw 90 samples and run MCMC estimation of each of the three models IVBART, linear-normal, and linear-DPM (CHMR) on each of the 90 samples.

We used the `R` package `bayesm` for the linear-normal and linear-DPM results and our `R` package for IVBART.

Each density represents all posterior draws of β over all MCMC iterations and all data simulations.

compare:
IVBART, linear, normal errors, linear, DPM errors, (CHMR)

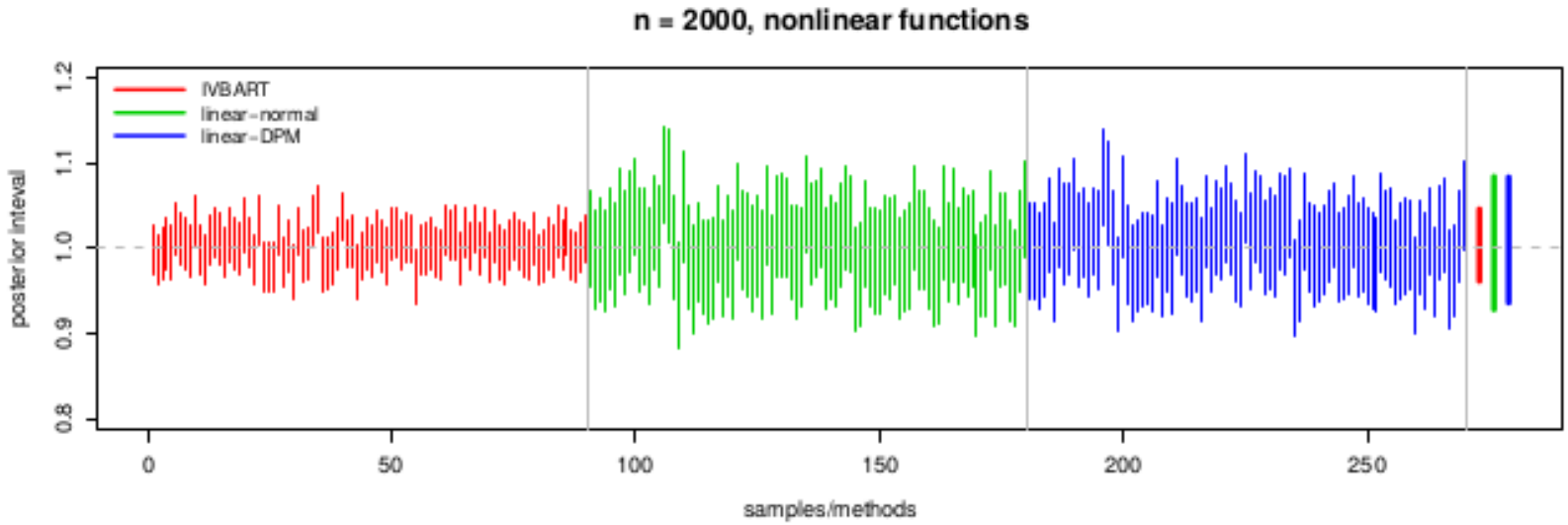


topleft: works !!; bottom left: bias with small n but still better than linear!!

In the linear case, not much worse than linear methods.

$n = 2,000$, nonlinear functions.

95% posterior intervals for each data simulation.



Prior Sensitivity

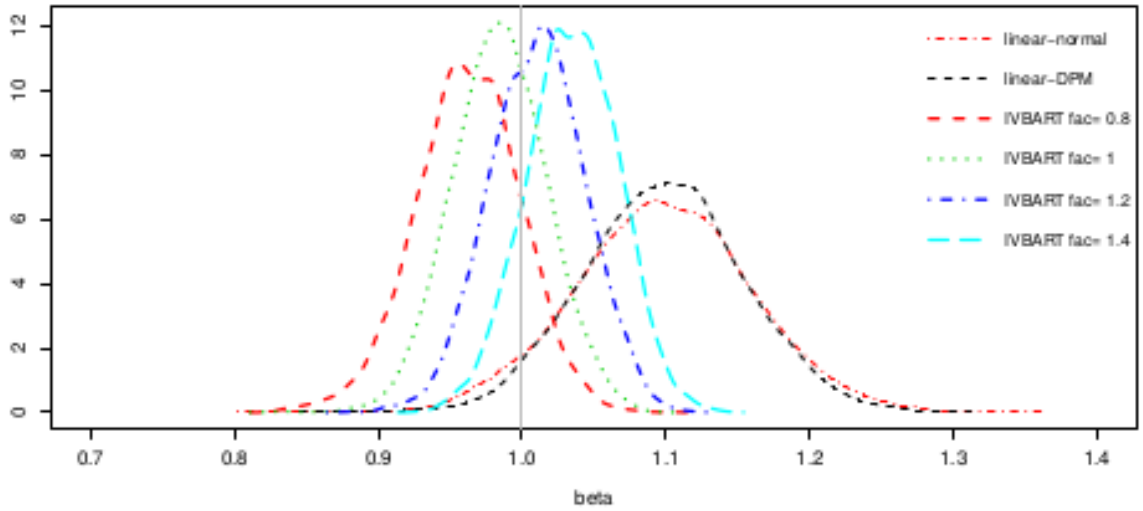
$n = 500$, nonlinear functions, a single data simulation.
σ_f, σ_h, $\in S_\sigma = \{.8, 1, 1.2, 1.4\}$

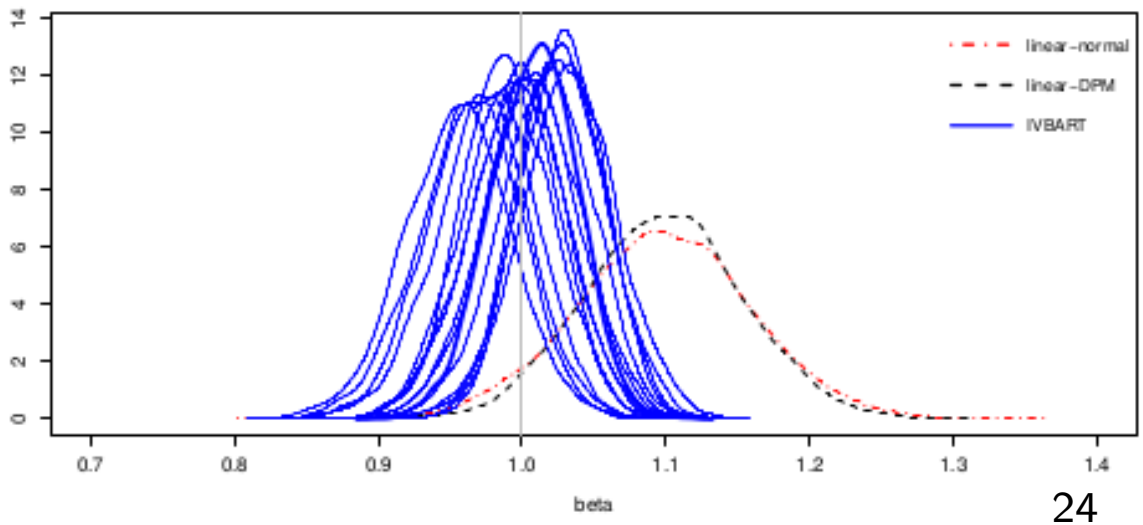
Top:
$\sigma_f = \sigma_h \in S_\sigma$.

Bottom:
$\sigma_f, \sigma_h \in S_\sigma$.

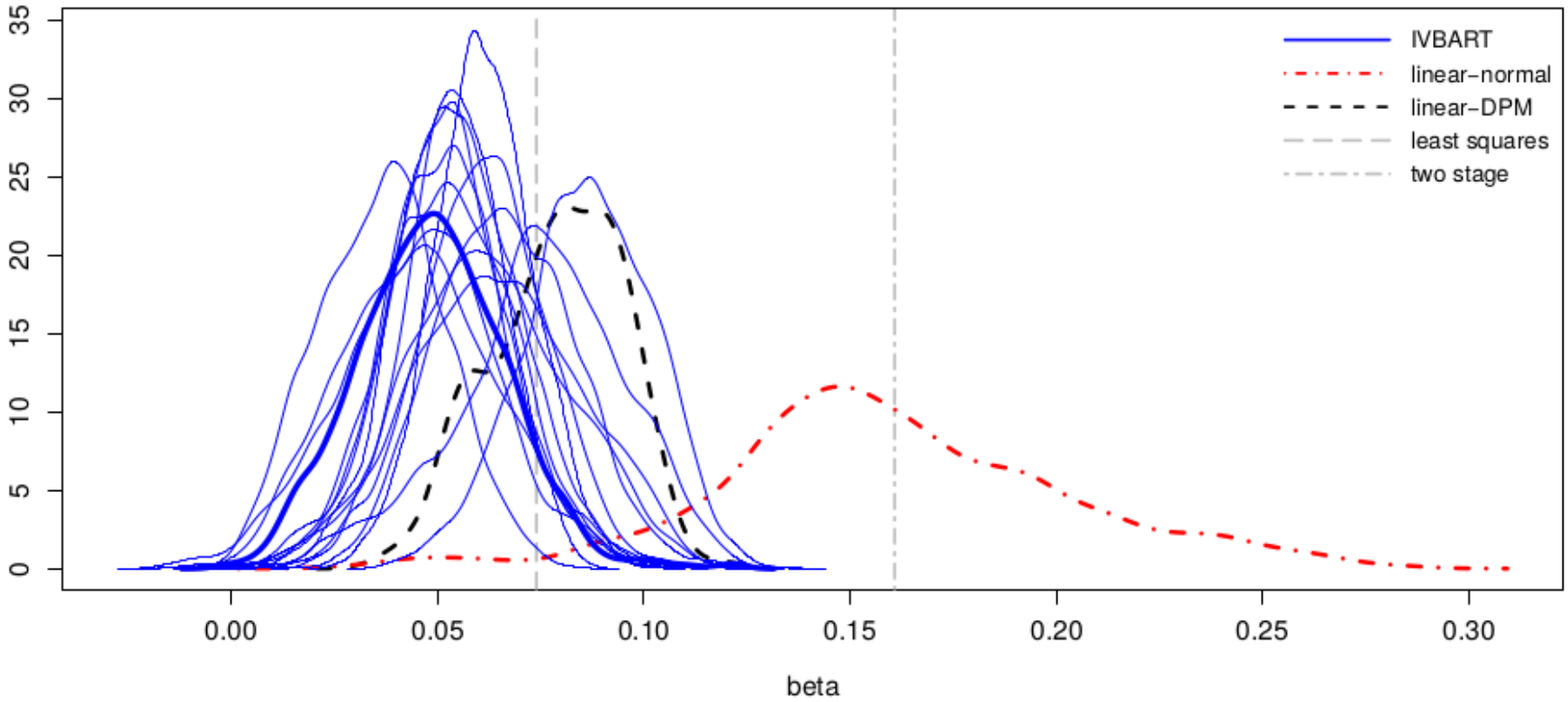
Pretty good!!

But we are still working on the prior.





6. Card Example



16 posteriors by varying σ_f and σ_h in $S_\sigma = \{.8, 1, 1.2, 1.4\}$.

Confirms CHMR result that β is much smaller than suggested by classic IV.

β is plausibly smaller than CHMR.