Bayesian instrumental variables: Likelihoods and priors[1]

Hedibert F. Lopes

Booth School of Business
University of Chicago

[1] Based on Lopes and Polson (2012) Bayesian instrumental variables: priors and likelihoods. *Econometric Reviews* (to appear).

Standard regression: $cov(x, u) = 0$

Graphical model

$$\begin{array}{ccc} x & \rightarrow & y \\ & \nearrow & \\ u & & \end{array}$$

or, in the linear case,

$$y = \beta x + u.$$

OLS estimate

$$\hat{\beta} = (x'x)^{-1}x'y.$$

Example I: Confounding

y: earnings
x: years of schooling
u: all other factors that determine earnings, such as *ability*.

High u due to high (unobserved) *ability*:

- Leads to higher earnings since $u \to y$;
- May lead to higher levels of schooling, since schooling is likely to be higher for those with high ability.

Therefore,

$$\begin{array}{ccc} x & \rightarrow & y \\ \uparrow & \nearrow & \\ u & & \end{array}$$

Non-standard regression: $cov(x, u) \neq 0$

Dependencies

$$\begin{tikzcd} x \arrow[r] & y \\ u \arrow[u] \arrow[ur] & \end{tikzcd}$$

or, in the linear case,

$$\begin{aligned} y &= \beta x + u \\ u &= \gamma x + v \end{aligned}$$

where $cov(u, v) = 0$.

Therefore

$$y = (\beta + \gamma)x + v$$

and OLS estimate is biased!

Example II: Omitted variable bias

Suppose the "correct" model specification is

$$y = \beta x + \gamma z + \epsilon,$$

while the "fitted" model is

$$y = \beta^* x + u,$$

with z and x related via

$$z = \delta x + v.$$

Therefore,

$$y = (\beta + \delta\gamma)x + (\epsilon + \gamma v)$$

If there is no association
between treatment (x) and confounder (z), $\delta = 0$, or
between outcome (y) and confounder (z), $\gamma = 0$, then
x is not a confounder and there is no omitted variable bias.

Solution

What is needed is a method to generate only exogenous variation in x.

Obvious solution: controlled experiment.

For most economics applications experiments are too expensive or even infeasible.

Definition of an Instrument[2]

A crude experimental or treatment approach is still possible using observational data, provided there exists an instrument z that has the property that changes in z are associated with changes in x but do not led to change in y (aside from the indirect route via x).

This leads to the following path diagram

$$\begin{array}{ccccc} z & \rightarrow & x & \rightarrow & y \\ & & \uparrow & \nearrow & \\ & & u & & \end{array}$$

which introduces a variable z that is associated with x but not u. It is still the case that z and y will be correlated, but the only source of such correlation is the indirect path of z being correlated with x which in turn determines y.

[2] Taken from A.C. Cameron, Department of Economics, UC Davis.

Recursive system

Let y_i be the response variable and x_i the (endogenous) regressor obeying the system of recursive equations:

Seletion equation: $x_i = \delta z_i + \varepsilon_{1i}$

Outcome equation: $y_i = \beta x_i + \varepsilon_{2i}$

Not a standard regression: $cov(\varepsilon_{1i}, \varepsilon_{2i}) \neq 0$

Instrumental variable: z_i is related to x_i but independent of ε_{2i}.

Causal effect parameter: β.

Literature review

IV regression:

- Bayesian approach
- Bayes-Stein shrinkage
- Decision-theoretic methods
- Method of moments
- Dirichlet process mixtures
- Monte Carlo simulation

Bayesian versus classical:

- Lindley and El-Sayyad (1968)
- Kleibergen and Zivot (2003)

Related problems:

- Errors-in-the-variables models
- Co-integration
- Reduced rank regression

Bayesian advances

Drèze (1976) Bayesian limited information analysis of the simultaneous equations model. *Econometrica*, 44, 1045-1075.

Drèze and Richard (1983) Bayesian analysis of simultaneous equations systems. In *Handbook of Econometrics*, volume 1, edited by Z. Griliches and M.D. Intrilligator. Amsterdam: Elsevier Science.

Geweke (1996) Bayesian reduced rank regression. *Journal of Econometrics*, 75, 121-146.

Kleibergen and van Dijk (1998) Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory*, 14, 701-743.

Chao and Phillips (2002) Posterior distribution in limited information analysis of the simultaneous equations model using Jeffreys prior. *Journal of Econometrics*, 87, 49-86.

Kleibergen and Zivot (2003) Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics*, 114, 29-72.

Chamberlain and Imbens (2004) Random effects estimators with many instrumental variables. *Econometrica*, 72, 295-306.

Hoogerheide, Kaashoek and van Dijk (2007) On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank. *Journal of Econometrics*, 139, 154-180.

Sims (2007) Thinking about instrumental variables. *Technical report*, Princeton University.

Conley, Hansen, McCulloch and Rossi (2008) A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144, 276-305.

Textbook material

Lancaster (2004).
An Introduction to Modern Bayesian Econometrics.
Chapter 8

Rossi, Allenby and McCulloch (2005).
Bayesian Statistics and Marketing.
Chapter 7

Koop, Poirier and Tobias (2007).
Bayesian Econometric Methods.
Chapter 14

Greenberg (2008).
Introduction to Bayesian Econometrics.
Chapter 11

Li and Tobias (2011).
Bayesian Methods in Microeconometrics.
Handbook of Bayesian Econometrics,
Chapter 6, Section 2.5.

Lopes, McCulloch and Carvalho (2012).
A First Course in Bayesian Econometrics.
Wiley. Chapter ?

Goals of the paper

We revisit and discuss key formulation, identification and estimation aspects when performing Bayesian inference in the IV regression model based on several of the above alternative representations.

Sims (2007), for instance, highlights the many issues with a priori assumptions in modeling the IV system as inferences can be very sensitivity to these specifications in the reduced form model.

Ill-behaved posteriors

Acommon feature of many of these IV problem is that you can obtain unexpected "sharp" (and possibly ill-behaved) posterior distributions from "weak" prior distributions, mainly due to the heavy tails of the likelihood or the nonlinearity of the parameters of interest or both[3].

Rossi *et al.* (2005) highlight the importance of priors on the error covariance-matrix Σ as one goes fromt he structural model to the reduced form.

[3] Maddala (1976), Zellner (1971), Hoogerheide, Kaashoek and van Dijk (2007), Hoogerheide and van Dijk (2008a,b), Hoogerheide, Kleibergen and van Dijk (2008).

Cholesky-based prior

We propose, as an alternative to the inverted Wishart prior, a new Cholesky-based prior for the covariance matrix of the errors in IV regressions.

Rossi *et al.* (2005) point out to several of the drawbacks of the Wishart distribution:

> *"The most important is that the Wishart has only one tightness parameter. This means that we cannot be very informative on some elements of the covariance matrix and less informative on others."*

We argue that our Cholesky-based prior is more flexible and avoids such drawbacks.

IV regression

Recall the system of recursive equations:

$$
\begin{aligned}
x_i &= z_i'\delta + \varepsilon_{1i} \\
y_i &= \gamma + \beta x_i + \varepsilon_{2i},
\end{aligned}
$$

with z_i a p-dimensional vector of instruments, related to x_i but independent of ε_{2i}.

We assume that $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i})'$ are i.i.d. $N(0, \Sigma)$, i.e. a bivariate normal distribution with zero mean vector and Σ variance-covariance matrix, where Σ has diagonal components σ_{11} and σ_{22} and off-diagonal component $\sigma_{12} = \rho(\sigma_{11}\sigma_{22})^{1/2}$.

The key distinction between the above system of equations to a standard bivariate regression is the possible correlation between the error terms ε_{1i} and ε_{2i} (and, therefore, between x_i and ε_{2i}).

This leads to the well-known endogeneity' bias when learning β; that is the information of x_i that is correlated with ε_{2i} should not be used when learning about the regression parameter β.

Disentangling ε_{1i} and ε_{2i}

$$\Sigma = \begin{pmatrix} \sigma_{11} & \rho\sqrt{\sigma_{11}\sigma_{12}} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

$$\Sigma = L \cdot D \cdot L'$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{12}}{\sigma_{11}} & 1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22}(1-\rho^2) \end{pmatrix} \begin{pmatrix} 1 & \frac{\sigma_{12}}{\sigma_{11}} \\ 0 & 1 \end{pmatrix}$$

Independence between ε_{1i} and u_i

Therefore,

$$L^{-1} = \begin{pmatrix} 1 & 0 \\ -\frac{\sigma_{12}}{\sigma_{11}} & 1 \end{pmatrix}$$

and

$$L^{-1}\varepsilon_i \sim N(0, D),$$

such that

$$\begin{aligned} \varepsilon_{1i} = x_i - z_i'\delta \quad &\sim \quad N(0, \sigma_{11}) \\ \varepsilon_{2i} = y_i - \gamma - \beta x_i \quad &\sim \quad N\left(\frac{\sigma_{12}}{\sigma_{11}}\varepsilon_{1i}, \sigma_{22}(1-\rho^2)\right), \end{aligned}$$

or

$$y_i = \gamma + \beta x_i + \frac{\sigma_{12}}{\sigma_{11}}(x_i - z_i'\delta) + u_i$$

for u_i iid $N(0, \sigma_{22}(1-\rho^2))$.

Endogeneity bias

The recursive system of equations

$$
\begin{aligned}
x_i &= z_i'\delta + \varepsilon_{1i} \\
y_i &= \gamma + \beta x_i + \varepsilon_{2i},
\end{aligned}
$$

can now be rewritten as

$$
\begin{aligned}
x_i &= z_i'\delta + \varepsilon_{1i} \\
y_i &= \gamma + \left(\beta + \frac{\sigma_{12}}{\sigma_{11}}\right) x_i + z_i'\left(-\frac{\sigma_{12}}{\sigma_{11}}\delta\right) + u_i
\end{aligned}
$$

with u_is iid $N(0, \sigma_{22}(1-\rho^2))$ and independent of ε_{1i}s.

"Endogeneity" bias:

> *Information of x_i that is correlated with ε_{2i} should not be used when learning about β.*

$p(x|z)p(y|x,z)$

Structural parameters: $(\delta, \sigma_{11}, \gamma, \sigma_{12}, \beta, \sigma_{22})$

Step 1: Learn (δ, σ_{11}) from

$$x_i|z_i \sim N(z_i'\delta, \sigma_{11}).$$

Step 2: Learn $(\gamma, \beta^*, \delta^*, \sigma_{2|1})$ from

$$y_i|x_i, z_i \sim N(\gamma + \beta^* x_i + z_i'\delta^*, \sigma_{2|1}).$$

Step 3: Let

$$\begin{aligned} \sigma_{12} &= -\sigma_{11}\delta_j^*/\delta_j \qquad \text{(for any } j\text{)} \\ \sigma_{22} &= \sigma_{2|1} + \sigma_{12}^2/\sigma_{11} \\ \beta &= \beta^* - \sigma_{12}/\sigma_{11} \end{aligned}$$

Identification I

The model is not identified in the limiting case of $\delta = 0$.

If the instruments z explain only a small portion of the variability of x (weak instrument case), then the likelihood function is concentrated around

$$\beta^* = \beta + \frac{\sigma_{12}}{\sigma_{11}}$$

for some estimable constant β^*.

Identification II

The recursive system of equations

$$\begin{aligned} x_i &= z_i'\delta + \varepsilon_{1i} \\ y_i &= \gamma + \beta x_i + \varepsilon_{2i}, \end{aligned}$$

can also be written in the reduced form:

$$\begin{aligned} x_i &= z_i'\delta + \nu_{1i} \\ y_i &= \gamma + z_i'\beta\delta + \nu_{2i}, \end{aligned}$$

where ν_i are iid $N(0,\Omega)$ and

$$\Omega = \begin{pmatrix} \sigma_{11} & \beta\sigma_{11} + \sigma_{12} \\ \beta\sigma_{11} + \sigma_{12} & \beta^2\sigma_{11} + 2\beta\sigma_{12} + \sigma_{22} \end{pmatrix},$$

depending on 4 parameters $(\sigma_{11}, \sigma_{12}, \sigma_{22}, \beta)$.

Again, the model is not identified in the limiting case of $\delta = 0$.

Prior specification

β and Σ are intertwined in the reduced form and independent priors for both parameters would be counter-intuitive.

Rossi *et al.* (2005) suggests

$$\begin{aligned} \delta &\sim N(d_0, D_0) \\ (\gamma, \beta)' &\sim N(b_0, B_0) \\ \Sigma &\sim IW(v_0, \Sigma_0) \end{aligned}$$

for known hyperparameters d_0, D_0, b_0, B_0, v_0 and Σ_0.

Posterior inference is available via a Gibbs sampler.

Full conditional of the structural covariance Σ

Let data $= \{(x_i, y_i, z_i); i = 1, \ldots, n\}$.

Then,

$$(\Sigma|\gamma, \beta, \delta, \text{data}) \sim IW(v_0 + n, \Sigma_0 + S)$$

where

$$S = \sum_{i=1}^{n} \varepsilon_i \varepsilon_i'$$

Full conditional of the regression parameters (γ, β)

$$(\gamma, \beta | \delta, \Sigma, \text{data}) \sim N(b_1, B_1),$$

where

$$\begin{aligned}
B_1^{-1} &= B_0^{-1} + \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i' \\
B_1^{-1} b_1 &= B_0^{-1} b_0 + \sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i \\
\tilde{x}_i &= \frac{1}{\sqrt{\sigma_{2|1}}} (1, x_i)' \\
\tilde{y}_i &= \frac{1}{\sqrt{\sigma_{2|1}}} \left(y_i - \frac{\sigma_{12}}{\sigma_{11}} (x_i - z_i' \delta) \right) \\
\sigma_{2|1} &= \sigma_{22}(1 - \rho^2).
\end{aligned}$$

Full conditional of the regression parameter δ

$$(\delta|\gamma, \beta, \Sigma, \text{data}) \sim N(d_1, D_1),$$

where

$$\begin{aligned}
D_1^{-1} &= D_0^{-1} + \sum_{i=1}^n \tilde{z}_i \tilde{z}_i' \\
d_1 &= D_1(D_0^{-1} d_0 + \sum_{i=1}^n \tilde{z}_i \tilde{x}_i) \\
\tilde{x}_i &= \frac{1}{\sqrt{\sigma_{1|2}}} \left(x_i - \frac{\sigma_{12}}{\sigma_{22}} (y_i - \gamma - \beta x_i) \right) \\
\tilde{z}_i &= \frac{1}{\sqrt{\sigma_{1|2}}} z_i \\
\sigma_{1|2} &= \sigma_{11}(1 - \rho^2).
\end{aligned}$$

Illustration I

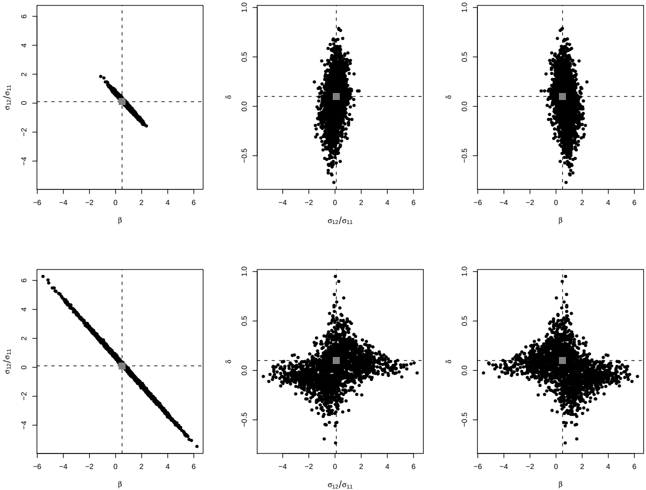
$n = 200$ simulated observations based on $\gamma = 1.0$, $\beta = 0.5$, $\sigma_{11} = \sigma_{22} = 1$, $\delta = (1.0, 0.1)'$ (one weak instrument) and $\rho = 0.1$ (low degree of endogeneity).

The prior hyperparameters are $d_0 = b_0 = 0$ and $D_0 = B_0 = 25I_2$, suggesting relatively low prior information about δ, β and γ.

For Σ two priors are entertained:
$v_0 = 3.00000$ and $\Sigma_0 = 3.00000I_2$ (relatively vague prior)
$v_0 = 0.00001$ and $\Sigma_0 = 0.00001I_2$ (non-informative prior).

The distinction between "relatively vague" and "non-informative" is arbitrary and its only purpose is to distinguish two prior specifications for Σ.

Top: relatively vague prior. Bottom: non-informative prior. MCMC run for 200000 draws and keeping every 100th.



See Hoogerheide and van Dijk (2008a).

Findings

The marginal posterior distributions of β and σ_{12}/σ_{11} becomes flatter as the values of δ approach zero.

The relatively vague prior turns out to be rather informative when compared to the non-informative prior:

Relatively vague prior: posterior of β roughly between $(-2, 2)$
Non-informative prior: posterior of β roughly between $(-6, 6)$.

The variability of σ_{12}/σ_{11} is also greatly affected by the choice of the prior on Σ.

To sum up, the trade-off between more precise (and less accurate) and less precise (and more accurate) posterior inference is an important, problem-specific part of the modeling process that needs to be dealt with on a case-by-case basis.

More on IV regression

A number of authors have proposed the use of Jeffreys prior in the IV regression.

From the reduced form, let $\pi = (\pi_x, \pi_y) = (\delta, \beta\delta)$ a $(p \times 2)$ matrix of rank one.

The Jeffreys prior is then given by (Chao and Phillips, 1998, and Zims, 2007)

$$\left| \frac{\partial \pi}{\partial(\beta,\delta)} \left(\frac{\partial \pi}{\partial(\beta,\delta)} \right)' \right| = ||\delta||(1+\beta^2)^{\frac{1}{2}},$$

which is suggestive of using priors with polynomial tails, such as a Cauchy. An extension of this is a reference prior.

Prior sensitivity

Sims (1997) makes a number of important remarks about prior sensitivity in IV problems.

Two of them are as follows.

> *As the likelihood does not go to zero as $\beta \to \infty$, with Σ fixed, no matter what the sample size is there is an issue of prior sensitivity.*

and

> *In a sample where the posterior is highly non-Gaussian and has substantial, slowly declining tails, even apparently weak prior information $(\delta, \beta) \sim N(0, 100I)$, can substantially affect the inference.*

Illustration II

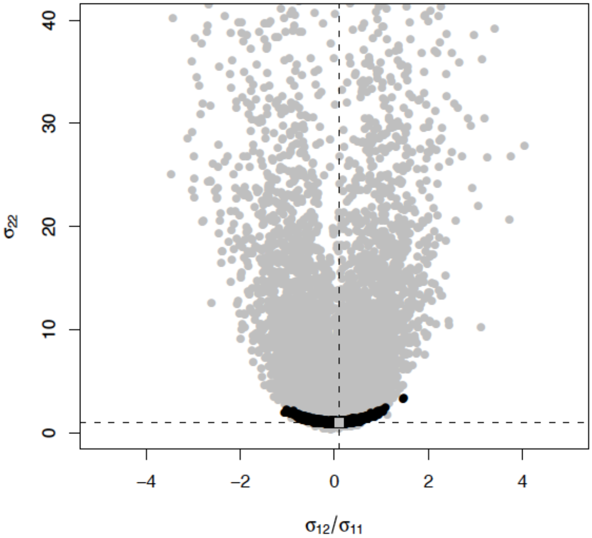
True values are:

- $\sigma_{22} = 1$ and $\sigma_{12}/\sigma_{11} = 0.1$ (low degree of endogeneity),
- $\delta = 0.1$ (weak instrument),
- $\beta = 0.5$.

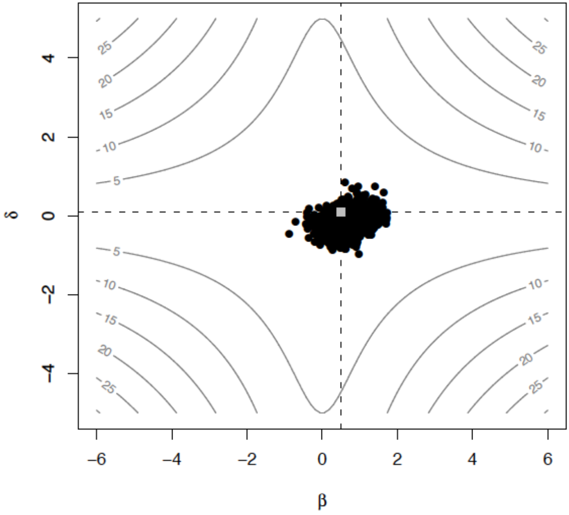
The prior of Σ is $IW(3.1, 0.6I_2)$, such that $E(\Sigma) = I_2$.
Most of the prior of σ_{12}/σ_{11} falls in $(-5, 5)$.

The prior of (δ, β) is Jeffreys, $p(\beta, \delta) \propto [\delta^2(1+\beta^2)]^{1/2}$.
The prior diverging for large values of either or both δ and β.

Priors and posteriors



Priors and posteriors



Dirichlet process prior

Conley *et al.* (2008) use a normal-based Dirichlet process prior to jointly model structural and instrumental variable equations errors. That is,

$$\varepsilon_i \sim N(0, \Sigma_i)$$

replaces the standard

$$\varepsilon_i \sim N(0, \Sigma),$$

with Σ_i now i.i.d. from the discrete random distribution G modeled by a Dirichlet process

$$G \sim DP(\alpha, G_0).$$

with concentration parameter α and base distribution G_0.

The marginal distribution of Σ_i (integrating out G) is continuous and is called a mixture of Dirichlet Processes (Escobar and West, 1995, 1998).

They conclude by saying that

> *Our Bayesian semi-parametric procedure produces credibility regions which are dramatically shorter than confidence intervals based on the weak instrument asymptotics. The shorter intervals from our method are produced by more efficient use of sample information.*

They go on and finish the paper saying that

> *... our non-parametric Bayesian method dominates Bayesian methods based on normal errors and may be preferable to methods from the recent weak instruments literature if the investigator is willing to trade-off lower coverage for dramatically smaller intervals.*

Bayesian model averaging

Recent research has focussed on applying Bayesian model averaging across sets of instruments, exogeneity restrictions, the validity of identifying restrictions and the set of exogenous regressors (Eicher, Lenkoski and Raftery, 2009, Koop, Leon-Gonzalez and Strachan, 2011)[4].

Another avenue is to assume a flexible fat-tailed alternative such as a mixture of t_ν distributions (mixed over ν).

[4] Eicher, Lenkoski and Raftery (2009). Bayesian model averaging and endogeneity under model uncertainty: an application to development determinants. *Technical report*, University of Washington. Koop, Leon-Gonzalez and Strachan (2011). Bayesian Model averaging in the instrumental variable regression model. *Technical report*, University of Strathclyde.

Cholesky-based prior

Cholesky-based priors have been used in many contexts:

- Pourahmadi (1999) - longitudinal data studies
- Lopes, McCulloch and Tsay (2011) - multivariate SV.

Instead of modeling Σ via an inverted Wishart distribution with parameters v_0 and Σ_0, i.e. $\Sigma \sim IW(v_0, \Sigma_0)$, we will model the components of the recursive conditional regressions that arises form the Cholesky decomposition of Σ.

Recall that $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i})'$ are i.i.d. $N(0, \Sigma)$ and let

$$\Sigma = AHA'$$

be the Cholesky decomposition of Σ such that A is lower triangular with ones in the main diagonal and lower triangular component given by $a_{21} = \sigma_{12}/\sigma_{11}$ and $H = \text{diag}(\sigma_{11}, \sigma_{2|1})$.

Therefore

$$A^{-1}\varepsilon_i \sim N(0, H),$$

which leads to *triangular regressions*

$$\varepsilon_{1i} \sim N(0, \sigma_{11})$$

and

$$\varepsilon_{2i}|\varepsilon_{1i} \sim N(a_{21}\varepsilon_{1i}, \sigma_{2|1}).$$

We specify independent prior distributions for σ_{11}, a_{21} and $\sigma_{2|1}$.

The implied prior for Σ can be directly obtained, either analytically or via Monte Carlo simulation.

σ_{11} is learned from the first equation,
σ_{12} is learned from σ_{11} and a_{21} $(= \sigma_{12}/\sigma_{11})$,
σ_{22} is learned from σ_{11}, σ_{12} and $\sigma_{2|1}$ $(= \sigma_{22} - \sigma_{12}^2/\sigma_{11})$.

Illustration 1

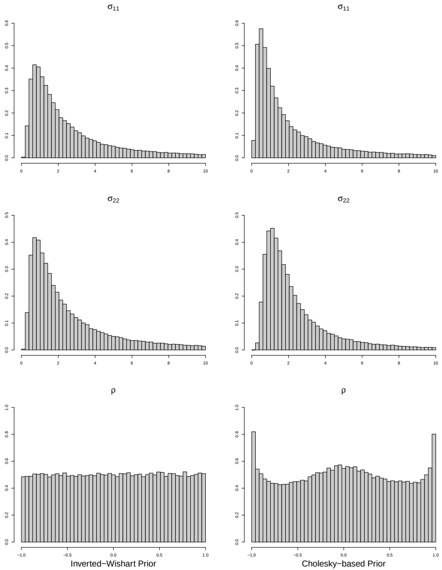
Weak instrument and low endogeneity context.

$\Sigma \sim IW(3, 3I_2)$
$(\sigma_{11}, \sigma_{2|1}, a_{21}) \sim IG(0.75, 0.75)IG(3,3)N(0, 0.7)$
or
$\Sigma \sim IW(0.00001, 0.00001I_2)$
$(\sigma_{11}, \sigma_{2|1}, a_{21}) \sim IG(10^{-7}, 10^{-7})IG(10^{-7}, 10^{-7})N(0, 1000)$

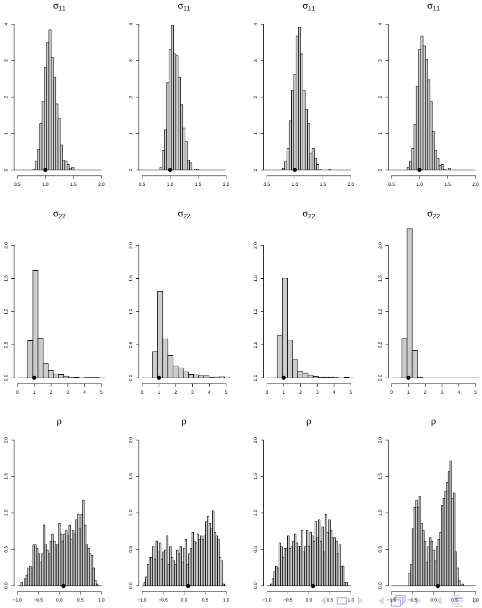
$(\beta, \delta) \sim N(0, 25I_3)$
or
$p(\beta, \delta) \propto ||\delta||(1+\beta^2)^{\frac{1}{2}}$

$\gamma \sim N(0, 25)$

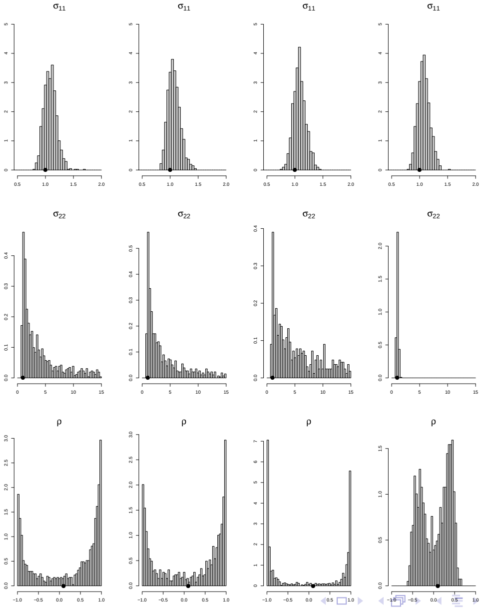
IW and Cholesky-based priors

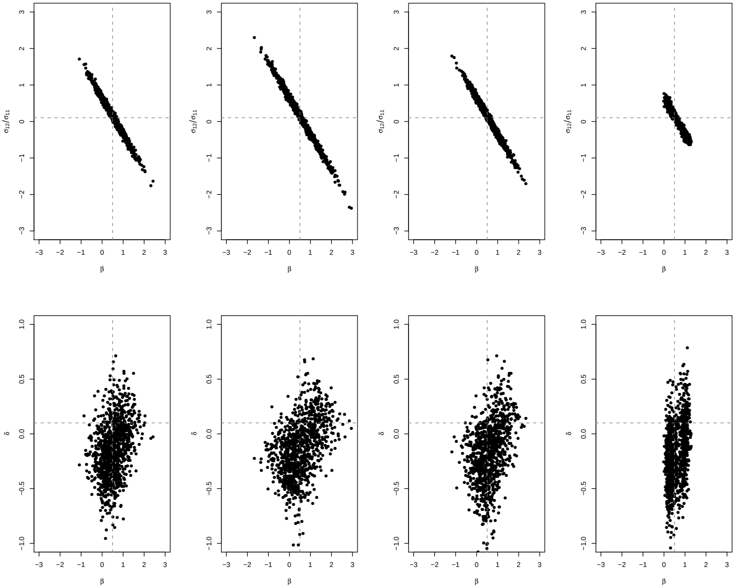


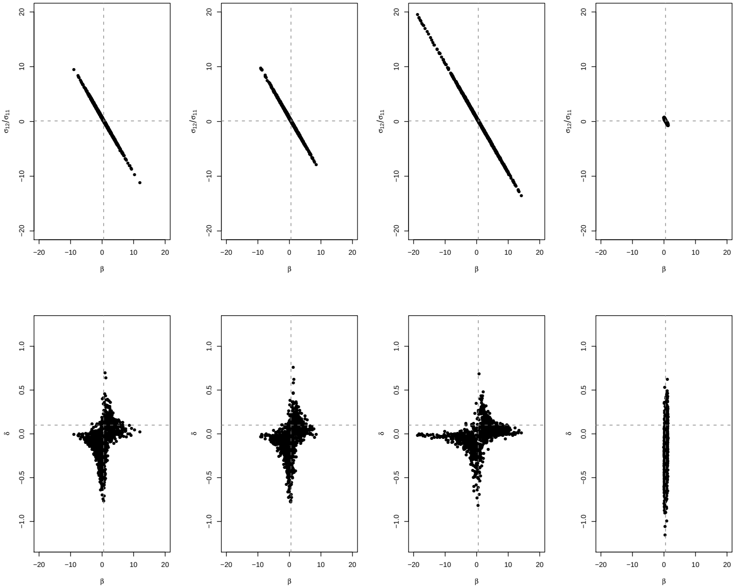
Marginal posteriors



Marginals: more diffuse priors




σ_{12}/σ_{11}
β
δ
−3 −2 −1 0 1 2 3
−1.0 −0.5 0.0 0.5 1.0



σ_{12}/σ_{11}
β
δ
−20
−10
0
10
20
−1.0
−0.5
0.0
0.5
1.0

Illustration 2

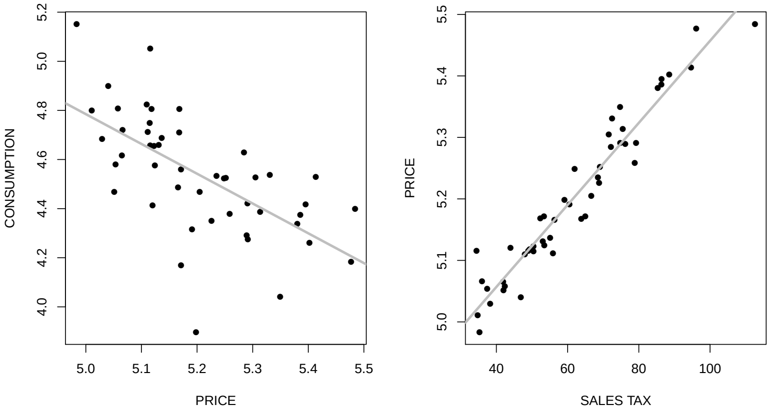
Demand for cigarettes[5]: Study the effect of price changes (x) on the demand for cigarettes (y).

The data set consists of annual data for the 48 continental U.S. states for the year of 1995.

Proxy for price: Log of average real price per pack of cigarettes including all taxes.
Proxy for consumption: Log of the number of packs of cigarettes sold per capita in the state.
Proxy for sales tax: Portion of the tax on cigarettes arising from the general sales tax, measured in dollars per pack in real dollars, deflated by the consumer price index.

---

[5] Stock and Watson's *Introduction to Econometrics*, 339-341.


CONSUMPTION
4.0
4.2
4.4
4.6
4.8
5.0
5.2
5.0
5.1
5.2
5.3
5.4
5.5
PRICE
PRICE
5.0
5.1
5.2
5.3
5.4
5.5
40
60
80
100
SALES TAX

OLS results

$$(\gamma_{ols}, \beta_{ols}) = (10.850, -1.213)$$
$$\delta_{ols} = (4.79006, 0.00667).$$

The sample coefficient of correlation between the residuals of both OLS fits is around -0.1775306.

These preliminary results suggest a scenario of a low degree of endogeneity and a relatively weak instrument, somewhat similar to the previous simulation exercise.

Hyperparameters:

$\Sigma \sim IW(a_0, a_0 I_2)$

$(\sigma_{11}, \sigma_{2|1}, a_{21}) \sim IG(a_0, a_0) IG(a_0, a_0) N(0, \sigma_0^2)$

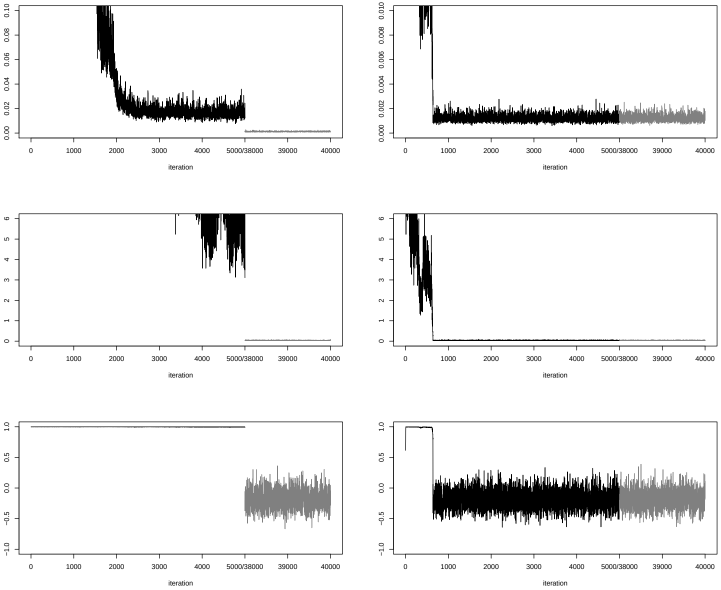
$(\beta, \delta) \sim N(0, \sigma_0^2 I_3)$

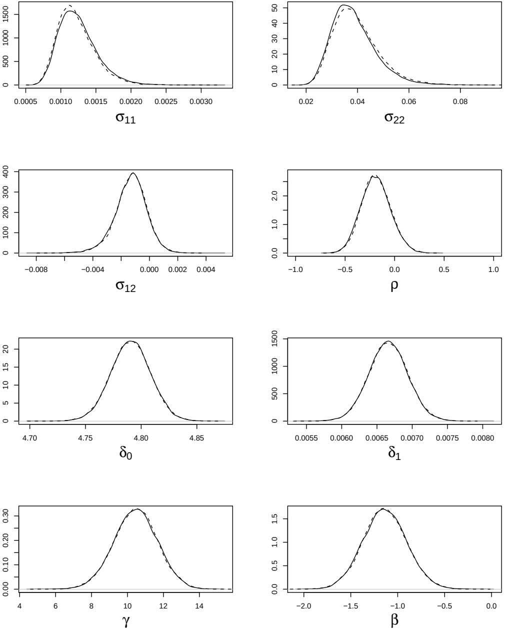
$\gamma \sim N(0, \sigma_0^2)$

$a_0 = 0.0000001$ and $\sigma_0^2 = 1000000$.

Initial values:

$\sigma_{11}^{(0)} = \sigma_{2|1}^{(0)} = 1$, $a_{21}^{(0)} = 0$, $\gamma^{(0)} = \beta^{(0)} = 0$ and $\delta^{(0)} = (0, 0)$.


0.00
0.02
0.04
0.06
0.08
0.10
0.000
0.002
0.004
0.006
0.008
0.010
0
1
2
3
4
5
6
-1.0
-0.5
0.0
0.5
1.0
0
1000
2000
3000
4000
5000/38000
39000
40000
iteration


σ_{11}
σ_{22}
σ_{12}
ρ
δ_0
δ_1
γ
β

Parameter	Median	95% credible interval
σ_{11}	0.0012	(0.0008,0.0019)
σ_{22}	0.0374	(0.0254,0.0584)
σ_{12}	-0.0013	(-0.0038,0.0008)
ρ	-0.1934	(-0.4655,0.1111)
δ_0	4.7907	(4.7554,4.8251)
δ_1	0.0067	(0.0061,0.0072)
γ	10.4560	(8.0198,12.8142)
β	-1.1373	(-1.5904,-0.6694)

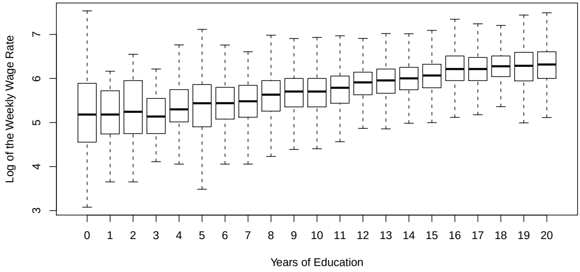
Illustration 3

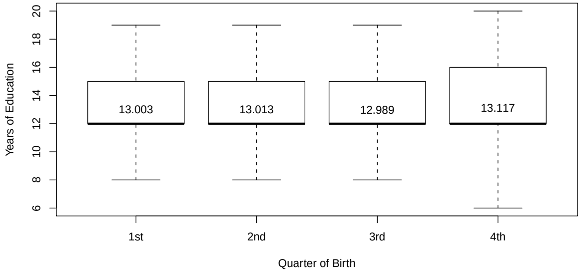
Here we revisited the return to education illustration presented in Chapter 8 of Lancaster's textbook *An Introduction to Modern Bayesian Econometrics*, pages 325-334.

The goal is to study the effect of (years of) education on (log) wages when using quarter of birth as an instrumental variable.

This is a fairly well known study and was first proposed by Angrist and Krueger (1991).

We follow Lancaster's simplification and focus only on men born in 1939 ($n = 35,805$), which is the last year of the 10-year study of Angrist and Krueger (1991).

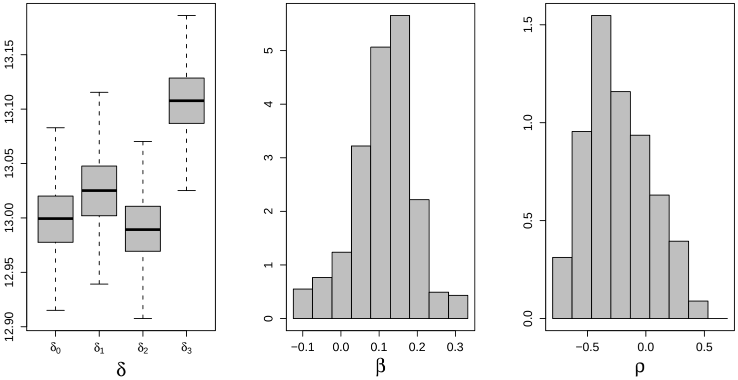

Log of the Weekly Wage Rate
3
4
5
6
7
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Years of Education


Years of Education
6
8
10
12
14
16
18
20
13.003
13.013
12.989
13.117
1st
2nd
3rd
4th
Quarter of Birth

They argued that quarter of birth might be related to years of education due to age-related regulations to both enter and leave school.

More precisely, children whose birthdays fall in the 4th quarter of the calendar year will enter (elementary) school the fall of that same year or within one or two months from their birthdays, while children whose birthdays fall, say, in the 1st quarter of the calendar year will enter school at least six months after their birthday.

In addition, compulsory schooling laws require students to remain in school until a predetermined age (usually sixteen or seventeen).



Posterior means of δs: $(12.999, 13.025, 12.990, 13.108)$.
$Pr(\beta < 0|\text{data}) = 7.7\%$
$Pr(\rho < 0|\text{data}) = 73.4\%$

As expected, apart form δ_3, all δs are quite similar, corroborating with the initial suggestion that quarter of birth is a weak instrument for this illustration.

The posterior mean and median of the percentage marginal return, 100β, are 11.2% and 11.9%, respectively, while the 95% credibility interval is $(6.57\%, 16.5\%)$.

The degree of endogeneity can be measure by ρ, whose posterior mean and median are -0.245 and -0.191, respectively, while $Pr(\rho < 0|\text{data}) = 73.4\%$.

Lancaster (page 333) says that

> *"the posterior suggests that the structural form errors are negatively correlated and this is a bit surprising on the hypothesis that a major element of both ε_1 and ε_2 is* ability *and this variable tends to affect positively both education and wages. But the evidence is very far from conclusive."*

We agree and claim that this example, as well as the previous ones, illustrates the inferential difficulties when combining weak instruments and low degree of endogeneity.

Discussion

IV likelihoods have challenging likelihood surfaces.
$\Rightarrow$ Sensitivity to prior specification.

Bayesian solution to the identifiability problem is attractive (Zellner, 1971, Conley *et al.*, 2008) and that, given a prior, inference can be based on marginal likelihoods.

We introduce the use of Cholesky-based priors, which are more flexible that the traditional normal inverse-Wishart regression priors, and more realistic than an “uninformative” Jeffreys prior.

Future directions

We show how prior-posterior inference can be formulated in a Gibbs sampler and compare its performance in the weak instruments case.

Given modern-day computational methods for Bayesian inference (Gamerman and Lopes, 2006, Lopes *et al.*, 2011), these complicated likelihoods seem ripe for more discussion of priors.

One growing area of research is that of many (potentially weak) instruments with factor-like reduction structures. The Cholesky-based prior for Σ seems a natural vehicle in this arena.